



UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA
INTERNATIONAL DOCTORATE IN STRUCTURAL BIOLOGY

CICLO XXXIV

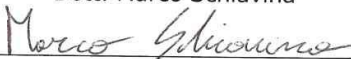
COORDINATOR Prof. Lucia Banci

***High-resolution tools to target intrinsically
disordered proteins***

Settore Scientifico Disciplinare CHIM/03

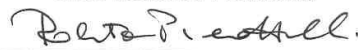
PhD student

Dott. Marco Schiavina



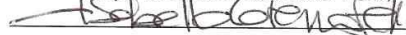
Tutor

Prof. Roberta Pierattelli



Tutor

Prof. Isabella Caterina Felli



Coordinator
Prof. Lucia Banci



November 2018 – October 2021

***This thesis has been approved by the University of Florence,
the University of Frankfurt and the Utrecht University***



Universiteit Utrecht



To my beloved niece

Acknowledgements:

My journey here at CERM started in August 2017 as a master student and since those days I fell in love with nuclear magnetic resonance. I really think that most of this passion has been infused me by my two supervisors and mentors, *Prof Isabella C. Felli* and *Prof. Roberta Pierattelli*. Since the first days I met them, they have been a guide for me, they have transferred me every knowledge they had as scientists and, probably even more important, as humans. The patience they had was incommensurable and even in the toughest situations they always ended the day with a smile or soothing words towards me. I profoundly believe that the working environment that was established in our research group is one of the best that a Ph.D. student can desire, where every idea is taken into consideration but also directed in a scientific and experienced way. Hopefully I will have the opportunity to play with spins and proteins under their guidance for a long time to come.

Every journey, to be pleasant, requires the perfect team-mates. This Ph.D. period gave me the opportunity to meet and work chick to chick with *Letizia Pontoriero*. Her never ending happiness, smiles, jokes were fundamental in the last three years. Her scientific intuitions are bright and clever, and her work has been essential in my scientific growth. Besides our working relationship what makes Letizia a special person to me is the friendship that we have established since the first days. We have supported each other in the gloomy days and rejoiced together in the joyful periods.

My journey obviously had many more members on the crew. Three of them, *Sara Matteucci*, *Domenico Rizzo* and *Giovanni Saudino* shared my same fate. Dividing worries, discussing ideas and exchanging opinions with them was an important part of my professional and human growth.

With all the young people in the lab I have built very kind and positive scientific and/or personal relationships and all of them would need a mention. Among all of them, I really want to acknowledge *Veronica Ghini* who has been a reference point in the lab from many points of view in the past three years.

Several other people were obviously involved in my scientific growth and my deepest gratitude goes to them. Thus, I would like to thank all the people I had the pleasure to work with: *Prof. Rober Konrat*, *Prof. Sonia Longhi*, *Prof. Vladimir N. Uversky*, *Prof*

Wiktor Koźmiński, Dr. Rainer Kümmerle, Dr. Wolfgang Bermel, Dr. Borja Mateos, Dr. Giancarlo Tria, Dr Maria Grazia Murrari and Dr. Edoardo Salladini.

Special thanks go to the entire CERM's staff for the quick and efficient technical assistance they provided me every time I was in need.

Few more guys deserve a special mention: the "kidnappers" *Giorgio Donati, Tommaso Pierini and Marco Trovato* that always remind me how young I am, the "coda di Fava" members *Federico Fava, Simone Ferretti, Lorenzo Mardegan and Marcello Ricci*, lifetime friends. *Luca Bondi*, is a true friend and the time spent together was always a source of light-heartedness.

The members of my family, *Alessandra Tosetto, Paolo Schiavina and Riccardo Schiavina* with their limitless encouragements, recommendations and support were always on my side and even if far in space they were always present.

Concluding I have to mention *Valentina Pancrazi*. She was always present in every single day since a couple of years. Her never-ending support, patience, smiles and her care are one of the main sources of my daily happiness. To her goes my love and my greatest gratitude.

Table of Content

List of abbreviations	ix
Abstract	xii
Keywords	xii
Aim of the project	xiii
1. Introduction	1
1.1 Intrinsically disordered proteins	3
1.2 Solution NMR in (un)structural biology	7
1.3 Heteronuclear direct detection	12
1.4 Non-Uniform Sampling	18
2. Carbon-13 direct detection as a tool for IDPs	21
2.1 Amide proton exchange process: DeCON to investigate the invisible	21
2.2 The “CON-Centric” approach: acquire two experiments in the time of one	29
2.3 My first steps in IDPs drug design: identify key motifs	38
2.4 Viral proteins challenge carbon detection (and lose)	49
3. Conclusions and perspectives	57
4. Bibliography	60
5. Chemical shift assignments	71
6. Cyclized NDGA modifies dynamic α -synuclein monomers preventing aggregation and toxicity	72
7. Taking simultaneous snapshots of Intrinsically disordered proteins in action	107
8. The Ambivalent Role of Proline Residues in an Intrinsically Disordered Protein: From Disorder Promoters to Compaction Facilitators.	141
9. Monitoring the Interaction of alpha-Synuclein with Calcium Ions through Exclusively Heteronuclear Nuclear Magnetic Resonance Experiments.	177

10. Ensemble description of the intrinsically disordered N-terminal domain of the Nipah virus P/V protein from combined NMR and SAXS.	206
11. Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications	239
12. The highly flexible disordered regions of the SARS-CoV-2 nucleocapsid N protein within the 1-248 residue construct: sequence-specific resonance assignments through NMR.	249

List of abbreviations

AP	– Anti-Phase
APSY	– Automated Projection Spectroscopy
BEST	– Band-selective Excitation Short-Transient
BMRB	– Biological Magnetic Resonance Bank
BT	– BEST-TROSY
CPMG	– Car-Purcell-Meiboom-Gill
CS	– Chemical Shift
DeCOR	– Decorrelation experiment
HSQC	– Heteronuclear Single Quantum Correlation
IDP	– Intrinsically Disordered Protein
IDR	– Intrinsically Disordered Region
IP	– In Phase
LRE	– Longitudinal Relaxation Enhancement
ncSSP	– Neighbor corrected Secondary Structure Propensity
ncSPC	– Neighbor corrected Structural Propensity Calculator
NDGA	– Nordihydroguaiaretic acid
NMR	– Nuclear Magnetic Resonance
NOE	– Nuclear Overhauser Effect
NTD	– N-Terminal Domain
NUS	– Non-Uniform Sampling
PDB	– Protein Data Bank

PED	– Protein Ensemble Database
PRE	– Paramagnetic Relaxation Enhancement
PTM	– Post-Translational Modification
RF	– Radio Frequency
RNP	– Ribonucleoprotein complex
SLiM	– Short Linear Motif
SSP	– Secondary Structure Propensity
TROSY	– Transverse Relaxation-Optimized Spectroscopy
UTOPIA	– Unified Time OPTimized Interleaved Acquisition

Abstract

Nuclear Magnetic Resonance (NMR) is one of the most powerful techniques exploited in structural biology. While the NMR strategies for the study of folded proteins are well established, Intrinsically Disordered Proteins (IDPs) pose new challenges to the technique. Special precautions must be taken due to the lack of 3D structure and to the high flexibility provided by these macromolecules. In this frame ^{13}C direct detection represents one of the best options to deal with the structural and dynamic features of IDPs. However, many steps forward have still to be taken to expand the NMR toolbox for the study of this class of biologically relevant proteins.

In this doctoral thesis, my contribution to the advancement of NMR spectroscopy for the study of IDPs is described. Mainly, I've focused on the development and application of new NMR methods. The new strategies enable the characterization of dynamic processes, facilitate the acquisition of spectra with high content of information in a reduced time and allow the identification of new key motifs in IDPs. Moreover, the outcomes from the study of two biologically relevant viral proteins are presented.

Through the proposed advances new motifs are identified in different IDPs, providing information of general interest and demonstrating the general applicability of these methods.

Keywords

Nuclear Magnetic Resonance · Intrinsically Disordered Proteins · ^{13}C detection · Amide proton exchange · Multiple receivers · Key-motifs · α -synuclein · Osteopontin · Viral proteins

Aim of the project

My Ph.D. project aimed at developing new NMR strategies tailored for IDPs to obtain spectra with high content of information (dynamical and/or structural) and in a vast range of conditions (T, pH, during chemical reactions etc...).

To accomplish this task, I have focused on the development of methods to investigate the interaction between IDPs and small molecules as well as metal ions, the exchange process between amide protons and water protons and the occurrence of some post-translational modifications in real-time.

Some of the well-known characteristics of IDPs were investigated with α -synuclein, a paradigmatic IDP. This protein was studied in its interaction with calcium ions to properly characterize the importance of side chains in the interaction between IDPs and metal ions as well as to better understand the exchange process between amide protons and water protons. The same protein was used to test novel experiments implementing the so-called “multiple receivers”, which enable the acquisition of more than one FID for each repetition, in a tailored way for IDPs to follow the phosphorylation of serine 129 by Fyn kinase. I also took part in the study of the interaction of α -synuclein with a putative drug: nordihydroguaiaretic acid (NDGA).

IDPs possess a peculiar aminoacidic composition, very different from the one of globular proteins, responsible of their unique features such as the presence of the so-called “compact states”. Nowadays, the driving forces of the compaction are still unclear. To better characterize this behaviour, I have worked, with the intrinsically disordered protein Osteopontin (qOPN) which revealed an interesting amino acid pattern that could be at the basis of the observed compact state present in qOPN.

Finally, the assignment and the study of different IDPs was carried out to investigate their peculiar features and to better understand the way they act. In particular, the study of NiV PNT, a 506 residue long intrinsically disordered protein, was exploited through NMR; SAXS was also used for the ensemble description of this protein. SARS-CoV2 nucleocapsid protein was also studied through NMR and its assignment was performed exploiting the ^{13}C direct detection strategy. This represents a first step in its biochemical characterization and opens the possibility for further studies to understand the role of flexible linkers for the function of the protein.

1. Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy is one of the most used techniques in structural biology thanks to its ability of providing key information to characterize biological macromolecules (proteins or nucleic acids).

Through a series of well-established NMR experiments it is indeed possible to achieve the complete characterization of a protein and to determine structural constraints (distances between atoms, angles between bonds) with a resolution in the order of the Å as well as dynamic features^[1-3]. The atomic resolution provided by this technique also enables the study of kinetic and thermodynamic processes involving proteins (e.g. folding/unfolding) as well as the interactions between macromolecules or other compounds as in the case of drug discovery^[4-6].

It took almost 40 years since the discovery of NMR principle (1946^[7]) before the first *de-novo* structure of a small globular protein obtained via NMR spectroscopy was published (bull seminal protease inhibitor, 1984^[8]). The possibility to obtain structural information in protein NMR got an extra push in 1990^[9], when the first triple resonance experiments were developed.

For structure determination, X-Ray crystallography still outplays NMR spectroscopy with thousands of structures published every year (Figure 1 red) because of the ease in solving protein structures once suitable crystals are obtained.

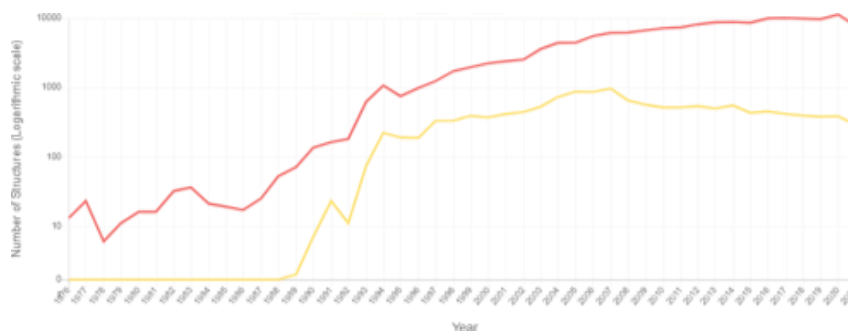


Figure 1.1: PDB statistics as obtained by <https://www.rcsb.org/stats/all-released-structures>. The orange line represents the number of protein structures deposited per year obtained by NMR data. The red line represents the number of protein structures deposited per year obtained by X-Ray data. The vertical scale is logarithmic.

However, NMR spectroscopy provides more information with respect to X-Ray crystallography. Some examples of the opportunities offered by NMR spectroscopy include the characterization of highly dynamic systems, the observation of the interaction between macromolecule and a small compound or between macromolecules or the investigation of biologically relevant reactions while they occur. Moreover, the possibility to exploit such experiments in solution and approaching/mimicking the physiological environment represents a unique feature of this technique with respect to X-Ray crystallography.

Indeed, X-Ray crystallography requires protein crystals that diffract an incident X-Ray beam in specific directions. The absence of these protein crystals renders this technique inefficient. One of the main sources that can inhibit crystallization is the high flexibility exhibited by several proteins or protein regions. An entire class of flexible, biologically relevant proteins whose atomic information can, thus, be grasped only by NMR is represented by intrinsically disordered proteins (IDPs). As described below, these proteins possess peculiar characteristics that must be considered when NMR spectroscopy is applied for their study. The development and application of optimal NMR strategies tailored for IDPs is the matter of study of this doctoral thesis.

1.1) Intrinsically Disordered Proteins

In 1999 Romero and co-workers predicted that about 15000 proteins in the Swiss Protein database contain intrinsically disordered regions^[10]. In 2012 more accurate screening of the data reported in the PDB demonstrate that about 68% of the crystal structures deposited contain unstructured regions^[11]. Nowadays, it is also known that IDPs and IDRs are much more abundant in eukaryotic proteins than in prokaryotic ones with 38% vs. 12-16% respectively and that this percentage is higher the more complex the organism is^[12].

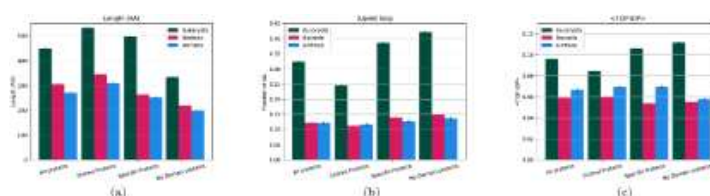


Figure 1.2 Three panels reporting different proteins' characteristics: average length (A) and the predicted content of intrinsic disorder from two different disorder predictors: IUPred long (B) and TOP-IDP (C). Proteins from Eukaryota, Bacteria and Archea (green, magenta, and blue respectively) were divided in: "shared proteins" consisting of protein containing at least one Pfam domain that exists in all the three kingdoms, "specific proteins" consisting of protein that contain Pfam domains that are exclusive of that kingdom, and "No domain proteins" which contains proteins without any annotated Pfam domain. "All protein" considers the sum of all the previously mentioned classes of proteins with a total of 26 million proteins. Figure from [13].

IDPs and IDRs are indeed a class of proteins/protein regions that lack a stable 3D structure in their native condition but rather sample a range of conformations with very low energy barriers that divide one conformational state from the other. What represented a breakthrough in structural biology was the observation that, despite the absence of a well-folded structure, these proteins were perfectly active in many different biological processes^[11,14-21].

From the biological point of view, structural flexibility represents an advantage in several mechanisms exploited by living organisms. IDRs could function as flexible linkers between domains, such as the case of FG Nups (a nuclear pore complex with Phenylalanine-Glycine repetitions), which acts as a size-dependent filtering device of the nuclear pore^[22]. The high flexibility of these proteins/protein regions leaves the backbone exposed to the environment facilitating the occurrence of important reactions such as post-translational modifications (PTMs). One case is represented by phosphorylation of Threonine 112 in synaptotagmin IDR1^[23]. This event results in the formation of salt bridges which inhibit the disorder-to-order transition. Phosphorylation is used as a key

mechanism to prevent the structural transition of the IDR and consequently regulating neuronal exocytosis, the main function carried out by synaptotagmin. The “Disorder-to-Order” transition is indeed one more feature of IDPs/IDRs that can occur, for example, in the so-called fold upon binding process. The latter case is represented, for example, by α -helical molecular recognition element (α -MoRE) of the intrinsically disordered C-terminal domain of the measles virus nucleoprotein (NTAIL) when it interacts with the X Domain (XD) of the measles virus phosphoprotein complex^[24].

In the vast range of interactions that IDPs can undertake, they could also experience “order-to-disorder” transition or be employed into “fuzzy complexes” to perform various functions^[25–29].

If the binding affects the partner’s activity, the protein is called “effector”. This is the case of p21^{cip1} that inhibits the cyclin-dependent kinase 2 (Cdk2) binding to CyclinA-Cdk2 complex which is the inhibitor of an important transcription factor in bacteria^[30]. IDPs can also bind metal ions as in the well documented case of α -synuclein^[31–34], or small molecules as in the case of the oncogenic transcription factor c-Myc (Myc)^[35].

Intrinsically disordered proteins are also involved in several human diseases ranging from cancer to neurodegenerative and cardiovascular diseases. Their involvement arises from malfunction that affects, for example, tumour suppressors, many of which contain IDRs. This is the case of human Osteopontin (OPN) whose role is crucial in regulating cellular functions^[36]. Its anomalous expression and/or splicing causes undesirable alterations in cancer, being implicated in promoting the progression of several carcinomas. One more example of medically relevant IDP is the protein Tau, one of the main players in Alzheimer’s disease, as well as the previously mentioned α -synuclein, involved in Parkinson’s disease^[17,37]. A novel, chronologically speaking, emerging topic in the field of IDPs is their capability of phase-separate as it is for FUS protein that forms liquid compartments. Aberrant liquid-like compartments formation seems to be associated with ALS and other age-related diseases^[38].

The extraordinary conformational plasticity exhibited by IDPs allows expanding the functional repertoire of these proteins, enabling them to achieve several biological tasks.

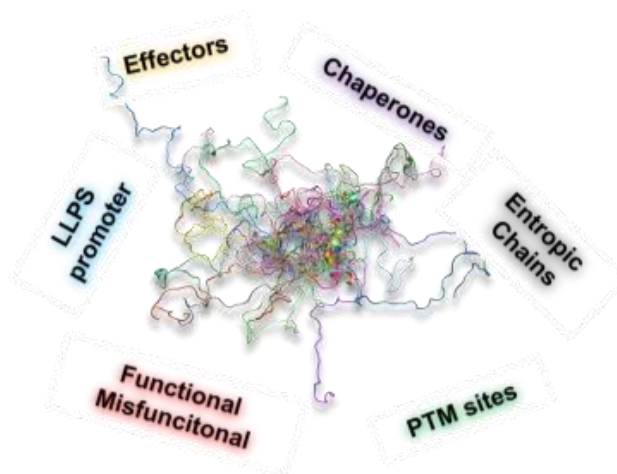


Figure 1.3: Schematic representation of examples of possible roles exerted by IDPs.

From the chemical point of view, IDP/IDRs are characterized by a peculiar amino-acid composition. Indeed, in contrast to folded proteins, these macromolecules usually deplete aromatics residues (tryptophane, phenylalanine and tyrosine) and hydrophobic residues (mainly isoleucine, leucine, methionine, cysteine) while they are usually rich in charged residues (aspartate, glutamate, arginine and lysine), polar residues (mainly glutamine and serine) as well as in glycine and proline residues^[39,40]. IDPs are frequently characterized by several repetition motifs (for example, AAKAAKAA motifs are found in the “KA-domain” of Tropoelastic proteins). Stretches composed by the same amino acid are frequent as well. These are the so-called “poly-X motifs”, one case is represented by the biologically relevant Poly-Glutamine case which is associated with “PolyQ diseases”^[41]. Thanks to their amino acidic composition, flexibility, and presence of low-complexity regions, IDPs bind to partners both with high specificity and high affinity or with high specificity but moderate affinity, extending the interactions landscape and posing new challenges in the field of drug design^[42–45].

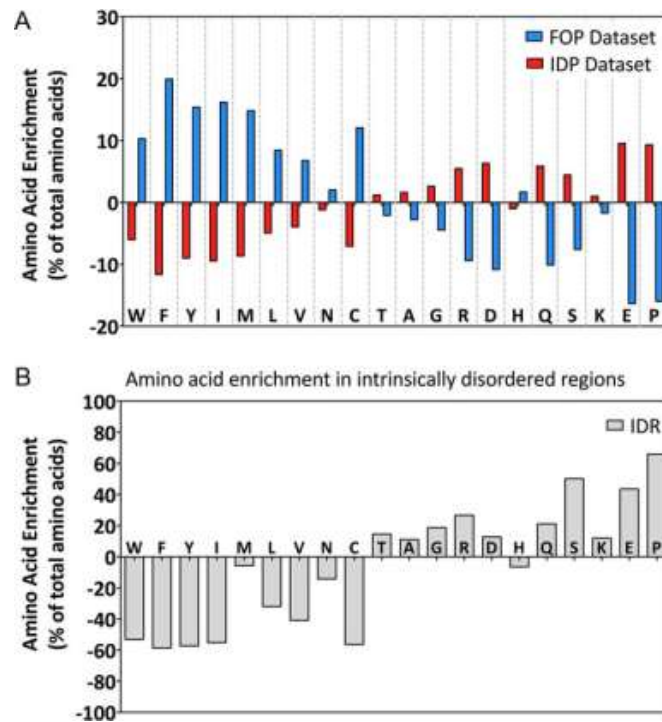


Figure 1.4: Panel A reports the relative percentage of the twenty natural amino acids in a dataset composed of folded proteins (FOP dataset, blue) and in a dataset composed of disordered proteins (IDP dataset, magenta). Panel B reports the relative percentage of twenty amino acids in the specific case of IDRs. Both panels provide the same qualitative picture dividing the amino acids in “disorder promoting” or “order promoting” residues. Figure from [40].

Thanks to the common features of IDPs’ primary sequence numerous algorithms and informatic tools were developed to predict the content of dis/order of a given protein and to identify the presence of peculiar linear motifs, binding regions, cleavage sites etc^[46-49]. Among all these predictors only a few databases exist nowadays for the collection of IDPs^[50,51].

1.2) Solution NMR in (un)structural biology

The lack of a well-folded structure renders the traditional NMR strategies for globular protein underperforming for disordered proteins/regions and NMR methods should be optimized for their investigation^[42,52-56].

One of the most challenging features of IDPs when solution NMR is applied is the very low chemical shift dispersion of the signals provided by these proteins with respect to globular ones. Comparing the 1D ¹H NMR spectrum of a well-folded small protein such as Ubiquitin with that of an IDP, α -synuclein (Figure 1.15), we can observe that the proton amide chemical shift distribution of the former protein is larger than that of the latter. The presence of sizeable ring currents, dipolar interactions, or H-bonds arising from the presence of defined structures, provides a significant contribution to the chemical shift of the various spins, resulting in well-resolved NMR spectra^[57]. On the contrary, the absence of a structure and the high flexibility average outs the previously mentioned effect, with the covalent structure that becomes the only source of contribution to the chemical shift. In this way an almost homogeneous environment is sensed by all the spins of IDPs' atoms that possess overall the same resonance frequencies. The result is spectra, even of high dimensionality, with several peaks falling in a very narrow spectral region. Resolution becomes one of the most limiting factors in studying IDPs.

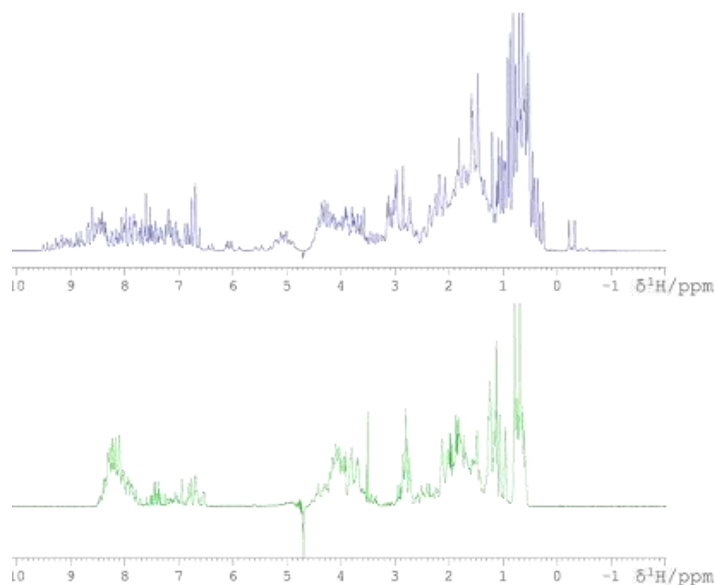


Figure 1.5: Two ^1H detected mono-dimensional spectra are reported. Blue spectrum reports Ubiquitin while α -synuclein is reported in green. As it can be seen from the comparison of the two spectra, the H^{N} region of ubiquitin (9.7 – 6.7 ppm) is much dispersed than the one of α -synuclein (8.5 – 7.9 ppm). The methyl signals present in the Ubiquitin spectrum (around 0 ppm) are absent in α -synuclein proving the absence of structure in this latter protein. Ubiquitin sample was 1.0 mM in 25 mM HEPES and 50 mM chloride phosphate. α -synuclein sample was 600 μM in 200 mM NaCl, 25 mM sodium phosphate. The spectra were acquired on a ^1H 1200 MHz spectrometer equipped with a 3 mm triple-resonance cryogenically cooled probehead (TCI).

A second problem that arises due to the lack of a 3D structure is that the backbone of the protein as well as the side chains are in close contact with the solvent (usually water) and thus the exchange process between the amide protons (and labile protons in general) and the solvent's protons is extremely fast. The exchange process renders amide proton detection inefficient in several conditions of temperature, pH and salt^[58].

One of the most common acquired 2D experiment in solution NMR is the HN-HSQC experiment. In this spectrum the connectivity between the amide proton (acquired directly) to its nitrogen (encoded in the indirect dimension) is provided and it displays a protein fingerprint.

Figure 1.6 reports, on the left, an HN-HSQC of α -synuclein performed at two different temperatures. High temperatures promote the exchange process and the exchange rate, k_{ex} , increases with it and thus with the increase of the temperature from 288 K (blue in figure 1.6) to 310 K (red in Figure 1.6) almost every signal is lost.

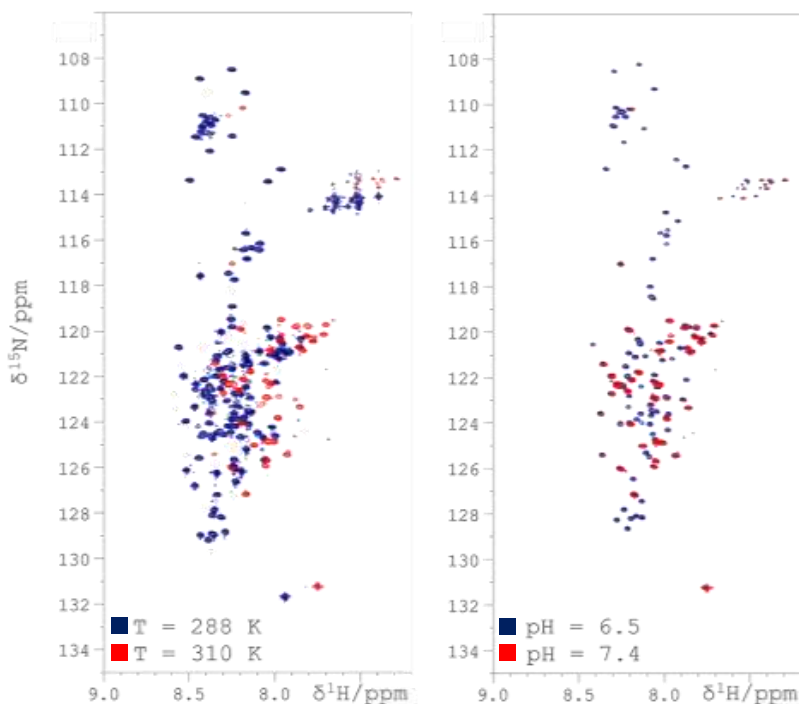


Figure 1.6: The effect of temperature becomes striking when comparing two 2D HN spectra acquired at two different temperatures on a model IDP as α -synuclein (A). Most of the peaks disappear when increasing the temperature from 288 K (blue spectrum) to 310 K (red spectrum). An even more striking effect is visible when the pH is increased (B). Most of the peaks, in the 2D HN spectra of α -synuclein are broad beyond detection when the pH is increased from 6.5 (blue spectrum) to 7.4 (red spectrum). α -synuclein sample was 600 μ M in 100 mM NaCl, 25 mM sodium phosphate. The spectra were acquired on a ^1H 700 MHz spectrometer equipped with a 5 mm triple resonance cryogenically cooled probehead optimized for ^{13}C -detection (TXO).

A very similar effect could be observed when the pH is increased. The higher is the pH the more H^+ atoms are present in solution and the exchange rate is raised^[59].

A similar, but less drastic effect is reported when the ionic strength in solution is increased and the anions that cause the enhancement follows the Hoffmeister series^[60]. For α -synuclein it was demonstrated that the increased ionic strength modulates the exchange process of amide protons with the solvent^[61].

The overall result is that the higher is the exchange rate values, the higher is the line broadening effect on amide proton signals. Studying IDPs through proton direct detection in conditions approaching the physiological ones in terms of temperature, pH and ionic strength (i.e. 310 K, 7.2 and 150 mM NaCl) might result in spectra with poor content of information.

Some strategies were carried out to try to overcome the intrinsic problem of the exchange process that affect H^N atoms in IDPs without renouncing to all the nice features provided by proton NMR spectra. Indeed, even if it's not possible to suppress the chemical exchange *in-toto*, it is possible to solve some of the problems related to it. The exchange process significantly affects the recovery of longitudinal magnetization for amide protons. Effective longitudinal recovery can be accelerated by perturbing the amide protons in a selective way with respect to those of the water. One of the most well-known ways to avoid any perturbation of water spins is to use the Band-selective Excitation Short-Transient (BEST) approach^[62]. Band selective pulses on a particular 1H spectral region ensure minimal perturbation of the proton spins resonating outside the excited spectral region. In this way, the longitudinal recovery is enhanced for the excited protons because of the dipolar interaction between the selective perturbed amide protons and the unperturbed ones ensuring efficient spin-lattice relaxation. As a result, the inter-scan delay from one repetition to the other in the experiment can be dramatically reduced and several scans can be achieved in the same amount of time resulting in an overall increased content of information. A similar approach is represented by band-Selective Optimized Flip-Angle Short-Transient heteronuclear multiple quantum coherence (SOFAST-HMQC) experiment^[63]. HMQC-type pulse sequences use fewer RF pulses with respect to HSQC-based ones. Fewer pulses reduce signal loss due to B1-field inhomogeneities and pulse imperfections. In addition, as for the BEST-type experiments, band selective pulses are used to perturb only the desired subset of spins without perturbing water. This experiment has been demonstrated to be very sensitive even with few scans (even one) per increment in particular if compared to other HN-HSQC-based experiments. However, as previously mentioned, the exchange process is not suppressed and the loss of information at high temperature/pH still broad signals, usually upon detection.

Thus, a common approach coupled to the BEST one, is the Transverse Relaxation Optimized Spectroscopy (TROSY)^[64,65]. The TROSY approach was firstly designed for the study of large proteins. These macromolecules possess long correlation times, resulting in very large transverse relaxation rates R_2 (peaks with broad line width). The TROSY idea relies on the possibility to select the narrowest component of the four relaxation components that are present in the case of two coupled spins. When applied to highly flexible IDPs the reduction in line broadening is modest due to the effective correlation time. However, as previously mentioned, one of the effects of the exchange process is to broad line widths. Thus, the exploitation of TROSY is also recommended in IDPs because, as described above, resolution matters and having this slight boost of resolution becomes fundamental if systems with several cross-peaks are studied. Moreover, RF decoupling during

acquisition of the signals is not applied and thus no limits for the detection of the FIDs are imposed by sample heating or decoupling RF power.

Another approach to avoid the loss of signals due to the exchange process consists in the detection of non-exchangeable nuclear spins. To avoid loss of the favourable properties offered by the ^1H nucleus such as its high gyromagnetic ratio, $^1\text{H}^\alpha$ direct detection has been used in the study of IDPs. $^1\text{H}^\alpha$ -based NMR experiments are also applicable in a range of temperature and pH due to the non-exchangeable nature of this nucleus. In contrast to $^1\text{H}^\text{N}$, $^1\text{H}^\alpha$ also provides information on proline residues, quite abundant in IDPs^[66-69]. A set of $^1\text{H}^\alpha$ -detected experiments has been proposed and provides the sequential connectivity in IDPs by correlating solely $^{15}\text{N}_i$ or $^{15}\text{N}_{i+1}$ chemical shifts with $^{13}\text{C}'_i$ and $^1\text{H}^\alpha_i$ spins in 3D H(CA)NCO and H(CA)CON spectra. Moreover, this set of experiments can be coupled with 3D (HCA)CON(CA)H and (HCA)NCO(CA)H which help in case of ^{15}N degeneracy. In these latter experiments the intensity of the sequential peaks can be maximized while the auto-correlated peaks can be minimized with minimal adjustments. $^1\text{H}^\alpha$ detection was also used, in a SemiConstantTime-HMQC-HA experiment to measure $^3\text{JN}^\text{H}^\alpha$, both in structured and disordered protein, as source of structural information^[70].

On the other hand, also the impact of homonuclear scalar coupling should be considered, in particular when studying IDP in which resolution matters. Furthermore, $^1\text{H}^\alpha$ direct detection still suffers from the non-optimal chemical shift dispersion typical of IDPs. As it can be noted in Figure 1.7 that reports the $\text{H}^\alpha\text{-C}^\alpha$ region of a 2D $^1\text{H}\text{-}^{13}\text{C}$ HSQC acquired on α -synuclein, the chemical shift dispersion is less than 2 ppm and the spectrum presents several crowded regions. In addition the presence of the water signal around 4.7 ppm might complicate the acquisition in case of imperfect water suppression.

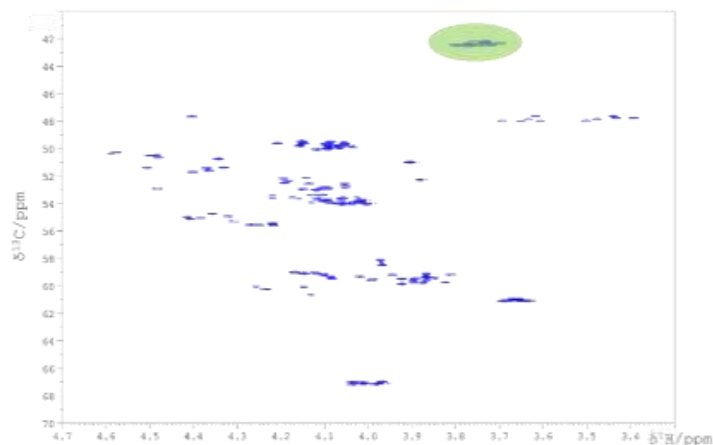


Figure 1.7: The spectrum reports the H^α-C^α region of a 2D ¹H-¹³C HSQC acquired on α-synuclein. As for every IDP, the peaks tend to cluster in few areas of the spectrum depending exclusively on the residue type, resulting in a spectrum with several crowded regions. Here the glycine region is highlighted, and 18 peaks are expected to be found here. The spectrum was collected on a 600 μM α-synuclein sample in 100 mM NaCl 25 mM NaPi. The spectra were acquired on a ¹H 1200 MHz spectrometer equipped with a TXO probe.

In conclusion, the absence of a 3D structure in IDPs, always seen as a drawback, also has some advantages when highly flexible IDPs are considered. In this case the effective short correlation time implies long transverse relaxation times^[71]. This represents a very favourable characteristic because it enables long magnetization transfer pathways with only minimal loss of magnetization resulting also in the possibility to exploit long acquisition times (in any dimensions) as well as multidimensional experiments (≥ 4D) up to 5 dimensions^[72,73].

1.3) Heteronuclear direct detection

As it can be observed in Figure 1.8, comparing ¹H^N, ¹H^α, ¹³C' and ¹⁵N^H 1D spectra, amide nitrogen provides the widest chemical shift dispersion followed by carbonyl carbon, α-proton and amide proton. However, despite the optimal chemical shift dispersion, ¹⁵N also possesses the lower gyromagnetic ratio among the listed nuclear spins. Indeed, ¹⁵N γ (-27.116 10⁶ rad·s⁻¹·T⁻¹) is indeed almost $-\frac{1}{10}$ of ¹H γ (267.522 10⁶ rad·s⁻¹·T⁻¹). For comparison, ¹³C nucleus has about $\frac{1}{4}$ of the gyromagnetic ratio of ¹H (67.283 10⁶ rad·s⁻¹·T⁻¹). Thanks to the extremely rapid technological advantages, carried out in recent years, both on the magnetic field strength (ultra-high field spectrometers are on field and operating) as well as on the probe design (¹³C optimized probes exist)

the low gyromagnetic ratio and thus the lower intrinsic sensitivity doesn't represent an insurmountable problem anymore and ^{13}C direct detection NMR experiments represent one of the best options for the study of IDPs.

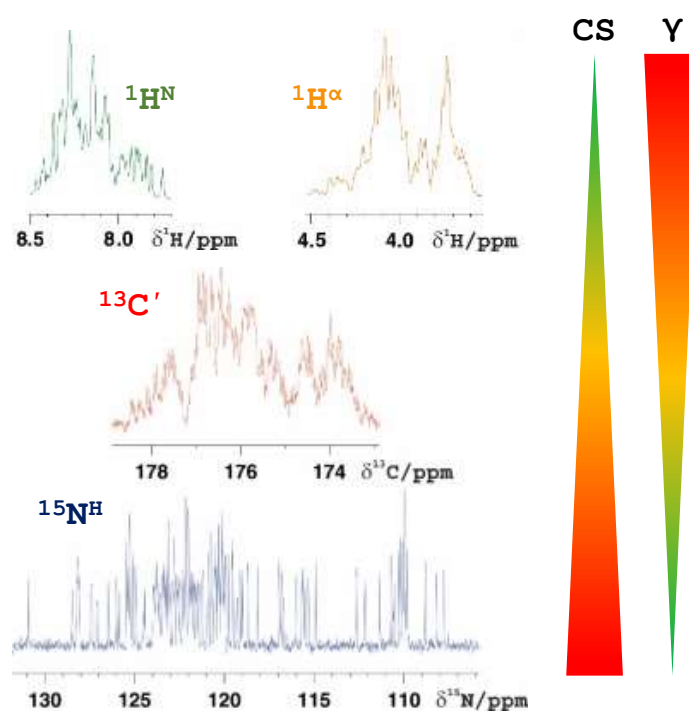


Figure 1.8: Chemical shift dispersion provided by different nuclei for α -synuclein. ^{15}N provides the highest chemical shift dispersion followed by carbonyl carbon, alpha protons and then amide protons. Gyromagnetic ratio increases in the opposite order of chemical shift dispersion. The spectra were collected on a 600 μM α -synuclein sample in 100 mM NaCl 25 mM NaPi. The spectra were acquired on a ^1H 1200 MHz spectrometer equipped with a TCI probe.

While ^{15}N direct detection has only recently been proposed due to the extremely low gyromagnetic ratio of this nucleus^[74–77], ^{13}C direct detection has been applied widely in biological NMR both for the study of IDPs as well as for the study of paramagnetic systems^[54,73,78–83].

Different types of carbon atoms are present in a polypeptide; C^α and C' are present on the backbone while aliphatic carbons nuclei, carboxylate carbons nuclei, aromatic carbons nuclei etc... compose the side chains of different amino acids. In IDPs, C^α s and carbon nuclei composing the side chains, always tend to cluster in narrow spectral regions depending, almost exclusively, on the localization of the atom on the side chain and the residue type. An example is provided in figure 1.7 where the C^α chemical shift dispersion of glycine residues is less than 0.5 ppm. On the contrary, carbonyl carbon possesses higher chemical shift dispersion. As shown in Figure 1.8, the chemical shift dispersion

provided by this nucleus is indeed superior to that of amide protons and its acquisition is thus preferred when sensitivity is not the limiting factor in the NMR experiment.

Additionally, carbon atoms are non-exchangeable nuclei and thus the exchange process does not affect the detectability of these nuclei. For this reason, it is possible to work in conditions of temperature and pH approaching the physiological ones^[54]. Furthermore, NMR samples of IDPs are often prepared in buffers with a high ionic strength to increase the sample quality (improved stability, reduced aggregation). High salt concentration usually hampers proton direct detection; ionic strength increases the overall sample conductivity introducing a “shielding effect” on the protein. This reduces the overall efficiency of the radio frequency pulses and introduces the necessity for increased power pulses that also increase the radiation damping effect, as well as the sample heating. Due to the lower gyromagnetic ratio of ^{13}C , the necessity to lengthen the pulses is much lower preventing all the above-mentioned problems.

One more advantage that is directly connected to the intrinsic low gyromagnetic ratio of heteronuclei, is that their transverse relaxation time generally quite favourable, provides narrow linewidths when direct detection of these nuclei is applied (Figure 1.8), resulting in a boost of resolution. The same relaxation properties enable the possibility to encode their frequency in indirect dimensions of NMR spectra for very long time without loss of signals due to transverse relaxation.

Regarding the carbonyl carbon, this atom is involved in the peptide bond which connects the carbonyl carbon of one residue (i) with the nitrogen of the following one (i+1). This connectivity is at the base of the 2D CON experiment^[84]. 2D CON, in its “exclusively heteronuclear” fashion, starts with direct perturbation of ^{13}C magnetization which is then transferred to the ^{15}N whose frequency is labelled in the indirect dimensions of the spectra and the magnetization is then transferred back to carbonyl carbon where the direct acquisition is exploited. As shown in Figure 1.9, this experiment provides superior chemical shift dispersion with respect to any 2D HN experiment. Finally, the 2D CON experiment provides also additional cross-peaks with respect to 2D HN. These peaks, which fall around 135 ppm (^{15}N dimension), arise from proline residues nuclei. These amino acids are usually very abundant in IDPs and the peculiar ^{15}N chemical shift of these residues is due to the pyrrolidine group composing the side chain that also prevents the detection of these residues with H^{N} direct detection.

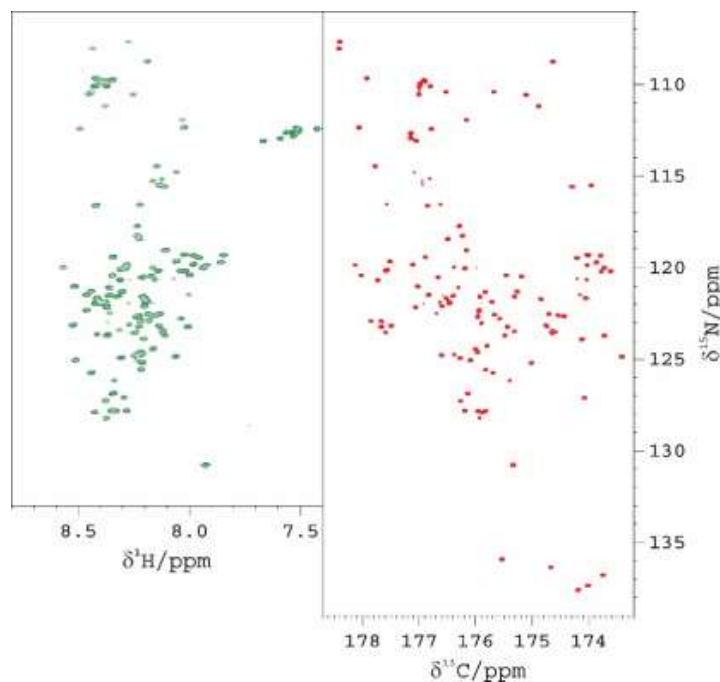


Figure 1.9: An HN-HSQC spectrum (left, green) is reported and compared with a CON (right, red). This latter experiment provides improved chemical shift dispersion in the direct dimension as already shown in Figure 1.8. The 2D HN-HSQC spectrum provides all the expected peaks for α -synuclein. However, these peaks are fairly broad due to the exchange process. In addition several resonances are found in crowded regions and are not well-resolved. On the contrary, the 2D CON spectrum provides well dispersed and narrow peaks. With the 2D CON experiment proline residues are also detected. The spectra were collected on a 600 μ M α -synuclein sample in 100 mM NaCl, 25 mM sodium phosphate. The spectra were acquired on a ^1H 700 MHz spectrometer equipped with a TXO probe.

The low intrinsic γ of ^{13}C could still represent a limitation and to improve experimental sensitivity different variants of the 2D CON experiment were developed exploiting the ^1H as a starting magnetization source. When H^{N} is used as a starting polarization source, the magnetization is then transferred to ^{15}N , its frequency is encoded in the indirect dimension, and the magnetization is transferred to the carbonyl where heteronuclear direct detection is exploited. H^{N} provides the possibility to enhance the overall recovery delay when the BEST approach is applied and to shorten the overall pulse sequence duration thanks to the shorter delays used in this variant with respect to the out-and-back approach of the exclusively heteronuclear fashion. However, exploiting H^{N} as initial magnetization pool reintroduces the dependence from the solvent exchange process and proline residues are not detected anymore. H^{α} nuclei do not suffer from solvent exchange and the possibility to exploit H^{α} -flip^[72,85], diminishes the magnetization recovery delay. This reduction becomes

particularly effective in case of slow rotational tumbling, for IDPs it can be useful when working at low temperature.

An additional feature that has to be taken into account when approaching carbon direct detection is the intrinsic strong homonuclear coupling between carbonyl carbons and alpha carbons. These two nuclear spins are connected by a strong 1J coupling of 55 Hz and this coupling evolves during the acquisition. Getting rid of this homonuclear coupling with standard composite pulse decoupling sequences is not so trivial because of giving radio frequency pulses at very similar frequencies to the one to be sampled while keeping the receiver open for acquisition. An approach is to rapidly switch between the “decoupling-mode” and the “acquisition-mode”, with a loss in sensitivity and the possible occurrence of sidebands that diminish the quality of the spectra.

Alternative strategies are the so-called “virtual decoupling approaches”^[78,86–88], one of them, summarized in Figure 1.10, consists in letting the coupling evolve, acquiring two different components of the signal, the in-phase (IP) and the anti-phase (AP) one. Once the two components are acquired, they are combined linearly and shifted to the centre of the doublet. This is the so-called IPAP approach.

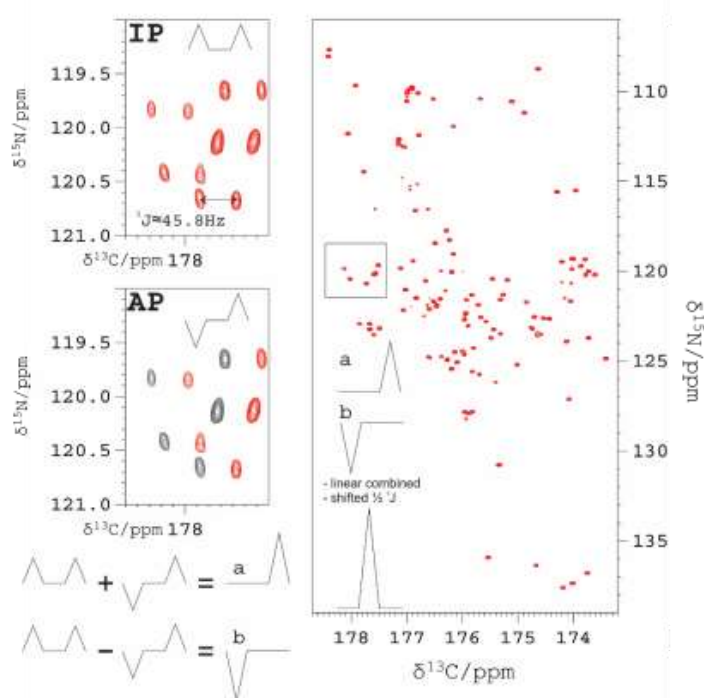


Figure 1.10. The in-phase (IP, top) and anti-phase (AP, bottom) components of the CON experiment are reported on the left for a subset of peaks. Linear combination of the IP and AP components of carbonyl signals allow us to separate the two FIDs multiplet components. These are then shifted for $\frac{1}{2} J$ CA-CO to obtain the final result, consisting in a single peak with almost twice the intensity of the single components.

Deconvolution of the coupled spectrum is also possible^[89]. In recent years, an approach based on deep neural networks has been proposed. The previously described methods rely on the necessity of recording two different components of the same doublet, with the requirement of doubling the effective phase cycle and with additional delays and pulses to create the desired coherence. This fairly new method is instead based on a deep neural network (DNN), FID-Net, that could be trained to decouple directly detected spectra using a single coupled spectrum. This method is particularly efficient when spectra with at least 2:1 signal to noise ratio can be acquired. In these conditions the acquisition can be performed in half of the time with respect to the virtual decoupling approach with a good time-saving.

1.4) Non-Uniform Sampling

Highly disordered, flexible proteins are characterized by long transverse relaxation time which enables the possibility to exploit experiments constituted by long delays and the possibility to acquire several indirect dimensions. Thanks to the nice features that 2D CON possesses, this is one of the most used experiments to obtain a fingerprint of an IDP and it has been used as a base for several 3D and multidimensional (≥ 4 D) experiments^[72,73,84,90]. Such experiments are used to fully characterise IDPs providing spectra with many indirect dimensions, useful to overcome the intrinsic low resolution usually characteristic of IDPs' spectra. One of the main problems that arise in acquiring spectra with several indirect dimensions is the NMR time required. Indeed, the time needed to conduct multidimensional experiments is directly proportional to the number of indirect evolution times sampled and according to the requirements of phase-sensitive quadrature detection and scans. Several days could be required in the case of a 4D when the full Nyquist grid is sampled, however, acquiring a single spectrum for such a long time is not always feasible. More than the cost of NMR time and disk usage needed to require and store such spectra, IDPs and proteins in general, might undergo degradation, cleavage, aggregation in a relatively short amount of time (from hours to few days). For this reason, a series of Non-Uniform Sampling (NUS) approaches have been developed^[82,91-96]. The ratio of the collected signals compared to the number of real points in the fully sampled grid is termed "sparsity".

Many NUS schemes have been proposed with different algorithms developed for reconstruction of the spectra for different applications (presence of several crowded regions, low signal to noise, several dimensions needed)^[97]. The very first idea was to uniformly sample random points in the time grid of the NMR multidimensional experiment. This approach is particularly feasible when the FIDs have none, or very little, decay in the desired time frame. Though, NMR signal decays accordingly to the transverse relaxation rates; thus, more elegant and powerful NUS schemes use schedules with exponentially weighted fashion^[96]. In this way, more points are collected where signals are stronger and the signal to noise ratio is enhanced. A different alternative is the Poisson-gap sampling method in which the gaps are at the centre of the schedule (sine-weighted sampling) or at the end (cosine-weighted), facilitating the reconstruction.

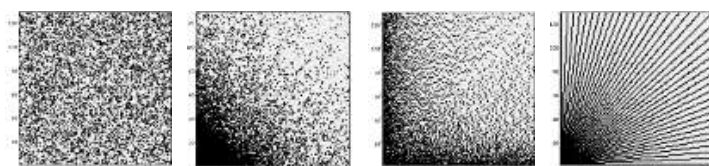


Figure 1.11: Few of the available NUS strategies are shown. From left to right: random, exponentially weighted, Poisson gap and radial sampling of a hypothetical Nyquist grid existing in a 3D experiment. Picture adapted from [97].

An alternative sample strategy is constituted by acquiring projections of multiD spectra. The APSY approach combines projection NMR spectroscopy with the automated analysis of projections achieved through the GAPRO (Geometric Analysis of PROjections) algorithm and with automated peak picking^[90,98–100]. With the APSY strategy a series of 2D NMR spectra are acquired. These spectra possess always the same direct dimension, while the second dimension is determined by a linear combination of the frequencies composing the indirect dimensions of the multidimensional experiment. The GAPRO algorithm, through iterative steps, interpolates the different peak positions, obtained in each projection, to create an N-dimensional peak list which is the final result together with the acquired 2D spectra.

An example of application of the APSY approach is reported in Figure 1.12. Exploiting the 5D-APSY-hcaCONCACON experiment, a peak list providing the correlations C'_{i-1} , N_i , C^α_i , C'_i , N_{i+1} is obtained. With this strategy, loading the peak-list on a 2D ^{13}C - ^{15}N CON spectrum. It is possible to perform the sequence specific walk through the backbone just by moving from one cross peak (C'_i , N_{i+1}) to the neighboring one (C'_{i-1} , N_i), as shown schematically in Figure 1.12 for a few residues.

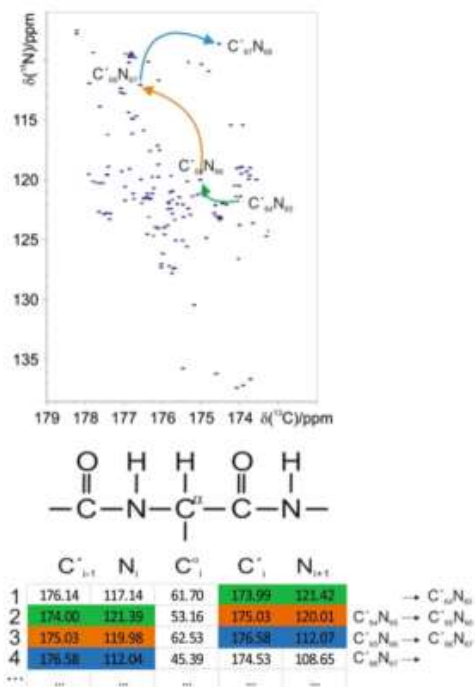


Fig. 1.12: A possible strategy used to assign α -synuclein through a 5D APSY-(HCA)CONCACON that uses the peak lists resulting from the GAPRO algorithm is here shown for residues 65–67 of α -synuclein. Four five-dimensional correlations (C'_{i-1} , N_i , C_i^α , C_i' , N_{i+1}) are reported in the table. Each five-dimensional peak contains two CON connectivity (i.e. C'_{i-1} , N_i and C_i' , N_{i+1}) that can be retrieved and then connected in the 2D-CON spectrum providing sequential assignment. The C^α chemical shift helps with the identification of the correct residue providing residue specific sequential assignment.

2. Carbon-13 direct detection as a tool for IDPs

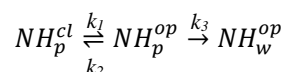
In the previous section an overview of the usefulness of ^{13}C direct detection for the study of IDPs was shown. Here I will summarize how I used ^{13}C direct detection to overcome some previously mentioned limiting factors in the study of IDPs.

Mainly, this strategy was exploited for the measurement of amide proton exchange with water proton, to increase the resolution and achieve more information from pre-existing NMR experiments, to identify new key motifs in IDPs and to monitor the interaction with small molecules. Carbon-13 direct detection was also employed in a new strategy that enables the acquisition of two experiments in the time needed for one of them and it was used for the characterization of two different challenging proteins making use of multidimensional NMR experiments.

2.1) Amide proton exchange process: DeCON to investigate the invisible

Amide proton exchange is that process in which the hydrogen atom bound to the amide nitrogen atom chemically exchanges with the hydrogens of the surrounding solvent molecules (usually water). The study of this event might help in obtaining site-resolved structural and dynamic features of a given protein.

The following chemical equation describes the overall exchange^[101,102]:



The process is described for a folded protein; the amide proton (NH_i^j) of a protein (p) could exist in two different states, a closed one (cl), where the exchange cannot occur because the amide hydrogen is buried inside the protein, and an open one (op), where the exchange process occurs with a rate k_3 . The process that unburies the proton from the closed state to the open one is characterized by the kinetic constant k_1 while the opposite process is governed by k_2 .

Despite the two processes can be clearly identified, the close to open (and vice versa) transition, and the exchange of the amide proton, NMR can study the overall effect. Through several different experiments the k_{ex} value can be measured, where

$$k_{ex} = \frac{k_1 k_3}{k_1 + k_2 + k_3}$$

This process is usually considered for globular proteins in native state where then $k_2 \gg k_1$.

Trying to disentangle as much as possible the different events, two limiting mechanisms were proposed based on the relative magnitude of k_3 and k_2 , establishing which is the fastest process once the open state has been reached: the intrinsic exchange of the amide proton ($k_3 \gg k_2$; EX1 limit) or the ratio in going back to the closed state ($k_2 \gg k_3$; EX2 limit). In the former case, $k_{ex} \approx k_1$, thus the exchange is limited by the conformational modifications that expose the amide proton to the solvent. In the latter case $k_{ex} \approx k_3/P$ where P is defined as the “protection factor”, a value that describes how much a residue is hindered in the protein.

The protection factor assumes a slightly different meaning when talking about intrinsically disordered residues. In this frame, IDPs can always be considered in an open state, the EX1 limit is then always verified. However, as widely demonstrated and discussed, an intrinsically disordered protein is very different from a completely unfolded protein as well as from a random coil chain. In these two latter cases indeed, the amide proton exchange process have been demonstrated to be more efficient than in the case of several IDPs. In 1995 Zhang and coworkers developed an algorithm (SPHERE), based on the studies by Englander, that computes the k_{ex} for a given primary sequence as if this is found in a completely random coil conformation (k_{ex}^{rc})^[103,104]. The ratio between the measured k_{ex} and the predicted k_{ex}^{rc} is the protection factor (P) in IDPs.

Several experiments were proposed to monitor and quantify this ratio. CLEANEX-PM, in its 2D version, is the most used experiment for this scope in folded proteins^[105,106]. This experiment relies on the idea of selectively perturbing the water’s spins and observing the magnetization transferred to the amide proton’s spins of the protein. The main “step forward” of this method, with respect to the previously proposed ones based on the same idea, is to suppress unwanted contributions from exchange-related NOEs/ROEs, enabling an accurate quantification. To efficiently suppress cross-relaxation and TOCSY effects the CLEANEX-PM spin-lock module ($135^\circ(x)$ $120^\circ(-x)$ $110^\circ(x)$ $110^\circ(-x)$ $120^\circ(x)$ $135^\circ(-x)$) is applied and implemented in a 2D FHSQC experiment, which is used as the read-out scheme.

The k_{ex} values can be obtained by exploiting experiments with different CLEANEX mixing times and then reporting these values against V/V_0 , where V is the peak volume at the different mixing time values and V_0 is the peak volume measured from the FHSQC spectrum (without spin-lock). The values are fitted as reported in [106] and/or through the initial slope analysis the k_{ex} values are obtained.

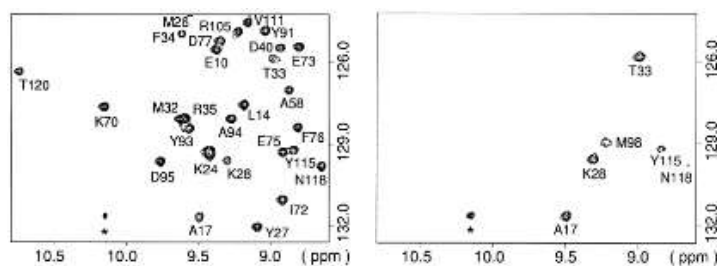


Figure 2.1 The standard HN-FHSQC spectrum is reported on the left for 1.5 mM SN protein acquired at 310 K, pH 6.8. On the right, the CLEANEX-PM spin-lock was applied with a γ B1 field of 5.1 kHz for a total duration of 40 ms. As it can be noted, several peaks disappear due to the fast chemical exchange with water. Peak marked with an asterisk is from tryptophan side chain. Picture from [106].

A second experiment, relies on the idea of observing the effect of chemical exchange on neighbouring spins that have a J -coupling to the exchanging nucleus. In this experiment, the decay of an S_x spin is observed under the application of a Carr-Purcell-Meiboom-Gill (CPMG) pulse train with and without proton decoupling^[107,108].

The CPMG train is used to suppress the long-range coupling without affecting the short-range coupling that works as a source of scalar relaxation. Thus, during the CPMG train, the S_x spin can relax due to the scalar coupling with the neighbouring I spin. If I, is an exchangeable proton, then measuring the S intensity in absence of decoupling and comparing to S in presence of decoupling will provide information about the exchange ratio of the I nucleus. This approach was used, for example, to measure k_{ex} values for $H^{\delta 1}$ and $H^{\epsilon 2}$ protons of the imidazole ring and of the NH_3^+ protons in histidine in different temperature and pH conditions^[108].

An even more elegant approach to measure chemical exchange was proposed in 1999 by Skrynnikov and Ernst, and it was named “decorrelation” spectroscopy (DeCOR)^[109]. This strategy is based on the idea of creating a double spin order operator in which one of the two involved spins undergoes chemical exchange (decorrelates). If one of the two spins, is an amide proton that decorrelates only due to the exchange process, the DeCOR scheme can be used to measure this event. In this experiment, the $2N_zH_z$ spin operator is created and then let free to evolve (decays due to chemical exchange) under variable τ_{decor} . The intensity of the peaks is measured, reported against the different τ_{decor} values and monoexponentially fitted to extract the k_{ex} rates. For precise quantification of the exchange ratio, a correction for the T_1 of $2N_zH_z$, which is active when the double spin order operator is created, has to be considered.

The main drawback of all the above-mentioned methods is that they are usually based on ^1H detection that poses several limitations to their applicability when dealing with IDPs.

For such reason we decided to implement a new method to “spy” chemical exchange of IDPs with water, based on ^{13}C direct detection, DeCON^[110]. Our proposed method follows the Skrynnikov and Ernst’s idea and the pulse sequence is reported in Figure 2.3. Starting from C'_z (a), the C'_y (b) coherence is built and then evolves in $2C'_xN_z$. Through a 90° pulse on carbonyl carbon spins along y the coherence is converted in $2C'_zN_z$ (d) and after creating $2C'_zN_y$ this is then evolved in $4C'_zN_xH_z$ (e) through the J coupling between N and H^N using a 180° band selective pulse on the amide region to avoid water perturbation. This coherence is converted in $4C'_zN_zH_z$ (f) and its decay is monitored under a variable decay τ_{decor} . At the end of this period, the remaining coherence is converted in $4C'_zN_yH_z$, back to $2C'_zN_x$ through INEPT and a pair of 90° pulses restore the $2C'_yN_z$ picked up at the end of the pulse sequence. From now on the second part of a standard CON scheme is used as a read out for the magnetization. With this strategy, a dependence from the water exchange has been reintroduced with negligible perturbation on water magnetization avoiding undesired radiation damping effects. The cross peaks can now be integrated, and their intensities can be fit to a mono-exponential decay to obtain the k_{ex} ratio, namely $k_{zzz} (I_{zzz}(\tau_{decor}) = I_0 e^{-k_{zzz}\tau_{decor}})$.

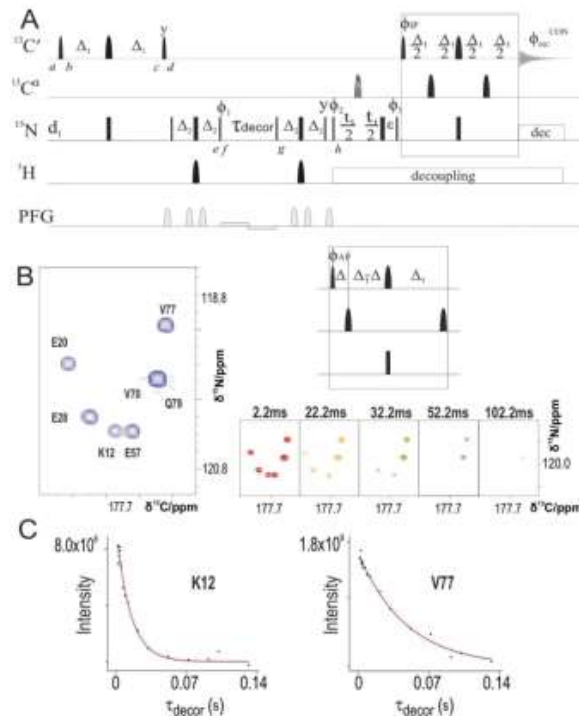


Figure 2.3: A) The proposed DeCON pulse sequence is reported in this panel. 180° selective pulses on amide region were used to selectively excite these nuclei exploiting the ^1J evolution. The selectivity also avoids water perturbation preserving from radiation damping effects. For more details refer to Section 9. A subset of peaks from the DeCON experiment is reported on the left with its assignment. Spectra recorded with different τ_{decor} are shown on the right. C) The intensities for two of the previously shown peaks are reported as a function of τ_{decor} . Picture from [110].

We compared the results obtained with our method with those deriving from the initially proposed ^1H detected variant, the HN-Décor experiment, on a sample of α -synuclein. As shown in Figure 2.4, the agreement between the exchange values obtained with the two methods is quite good for the residues detectable in both experiments.

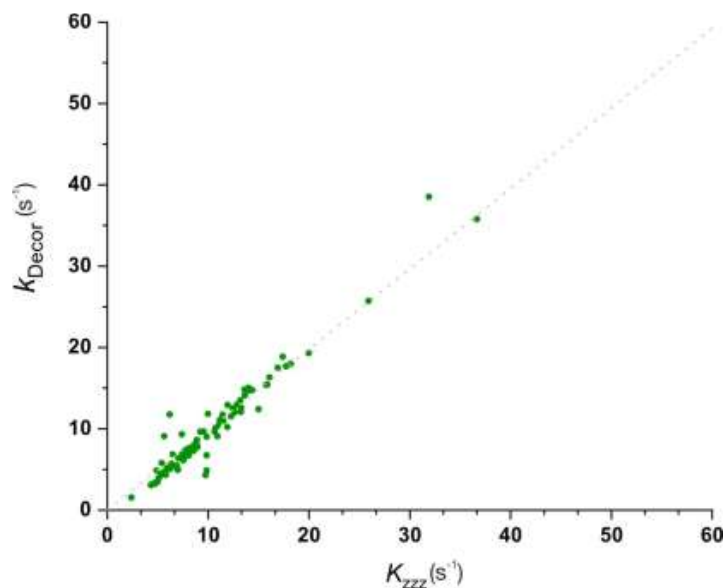


Figure 2.4: Comparison between the exchange values obtained with DeCON and DeCOR approaches. The two experiments agree for those residues that are detectable in both the experiments. These values were obtained at 298 K to enhance the number of cross-peaks visible in the HN-based experiment^[110]. Picture from [110].

Thanks to the superior chemical shift dispersion of $^{13}\text{C}'$ with respect to $^1\text{H}^{\text{N}}$, the DeCON experiment provides information for a larger number of residues. To further enhance the resolution, we have also implemented the 3D version of this experiment encoding C^{α} 's chemical shift in a third dimension (Figure 2.5).

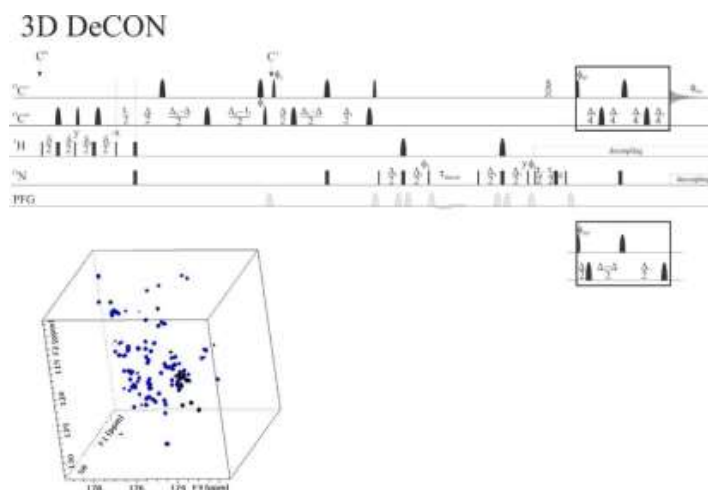


Figure 2.5: Pulse sequence used to acquire the 3D version of DeCON spectra. A screenshot of the acquired 3D spectrum is also shown. Picture from [110].

from the residue type while in the latter region the k_{ex} values seem to be influenced by the highly negatively charged residues that create a local electrostatic potential which reduce the k_{ex} values.

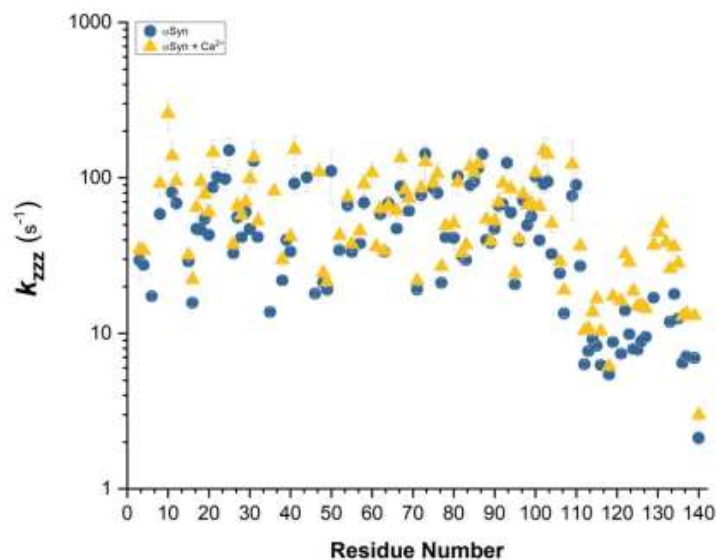


Figure 2.7: The k_{zzz} values measured with the DeCON approach are here reported on the vertical axis against the residue number. Blue circles represent the values for the protein alone in absence of calcium ions, yellow triangles represent the values upon addition of Ca^{2+} . Picture from [110].

Upon the addition of calcium ions, the overall k_{ex} values were enhanced. However, the increase was not uniform throughout the whole primary sequence. As shown in figure 2.8 residues from the N and C terminal region displayed an enhancement of their k_{ex} . In particular, Ala 18, Ala 30, Glu 115, Asp 119, Asp 121, Asn 122, Glu 123, Ala 124, Tyr 125, Ser 129, Tyr 133, Gln 134, Asp 135 and Tyr 136 experienced an over 1.9 times boost respect to the unbound form. This very localized and strong effect on these two regions could be ascribed to different reasons, a reduction in the electrostatic potential in the C-terminal region as well as in the disruption of interactions between the N-terminal and C-terminal segments.

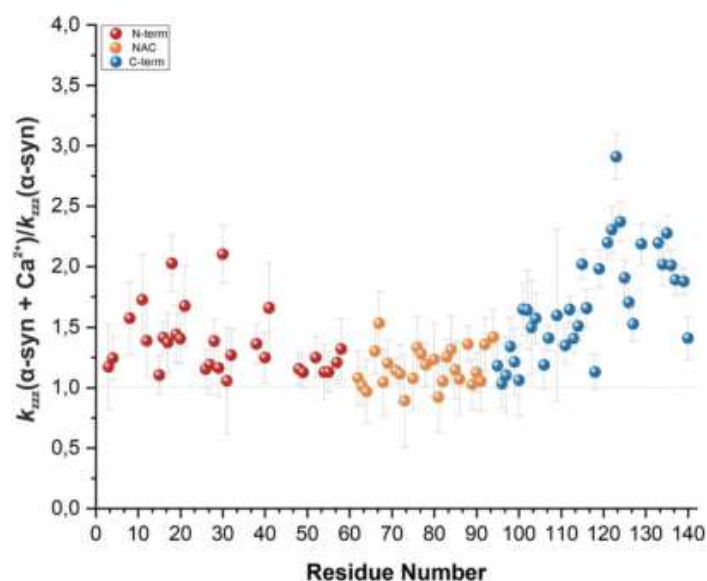


Figure 2.8: Ratio of the k_{zzz} values between the values obtained after and before the addition of calcium ions. The three colors refer to the three different regions of α -synuclein: N-terminal (1-59, red), NAC (60-95, yellow) and C-terminal (96-140, blue). The NAC region displays homogeneous enhancement while the C-terminal one shows major boost. Few scattered values are present in N-terminal domain. Picture from [110].

For a complete discussion refer to section 9 and to [110].

2.2) The CON-centric idea: acquire two experiments in the time of one

While ^{13}C nuclear spins provide a superior chemical shift dispersion and narrower line widths with respect to ^1H nuclear spins, they are also influenced by much longer longitudinal relaxation times. Thus, longer recovery delays are needed in ^{13}C excited- ^{13}C detected NMR experiments which are usually in the order of seconds. This means that the time “wasted” in waiting for the ^{13}C magnetization to be recovered is usually much longer than the experiment itself (tens/hundreds of milli-seconds). Aliphatic proton spins and carbonyl carbon spins possess very different relaxation properties with the former relaxing longitudinally faster than the latter.

^1H -start, longitudinal relaxation enhancement (LRE) experiments were developed trying to overcome this intrinsic property of carbon nuclear spins^[72]. As an example, ^1H -start ($^1\text{H}^{\text{N}}$ or $^1\text{H}^{\alpha}$ -start) experiments can be used as a valuable alternative to the ^{13}C -start exclusively heteronuclear approach

when a boost in sensitivity is required and shorter relaxation delays are needed (for example due to protein degradation). A version that provides even faster recycle delays is the ^1H -flip one in which the proton spins that are not involved in the transfer (any proton spin but $^1\text{H}^{\text{N}}$ or $^1\text{H}^{\alpha}$ in the example) can be flipped back to +z axis with a 90° or 180° pulse depending on the actual position of the spins leading to a longitudinal relaxation enhancement.

Another way to profitably use the time needed to restore the magnetization along the z axis would be to record other experiments during this delay. Thanks to technological advantages, this is now doable by exploiting the multiple receivers hardware (MR)^[111,112,121,113–120].

With this tool it is possible to acquire different FIDs on different receivers at the same moment, opening to the possibility of performing more than one experiment at the same time. Three different strategies exist for the exploitation of multiple receivers: sequential, parallel, and interleaved.

The sequential approach consists in acquiring more than one FID making use of a single initial magnetization pool. A biomolecular NMR experiment based on this approach is the “Afterglow” experiment whose spectra are reported in Figure 2.9^[122]. In the experiment, an haCACO 2D experiment is combined with a 3D haCAcoNH in a single pulse sequence. Here, the 2D CACO spectrum is recorded during the evolution of the $^1\text{J}^{\text{C}^{\text{N}}}$. The idea is that the 3D experiment shares, in one of the two indirect dimensions, the same indirect dimension of the CACO and thus C^{α} resonance can be encoded in the indirect dimensions at the early stages of the overall pulse sequence. Furthermore, the time used to directly detect C' magnetization in the 2D experiment is used as first step in the transfer of the magnetization from C' to N in the 3D experiment. Unfortunately, this latter idea also requires a compromise between resolution of the direct dimension in the 2D experiment and the overall sensitivity of the 3D experiment. The longer is the acquisition time in the bidimensional experiment, the higher is the resolution of C' dimension in the CACO spectrum. This also means a loss in magnetization for the following steps of the magnetization transfer pathway and thus a loss of sensitivity in the 3D haCAcoNH spectrum.

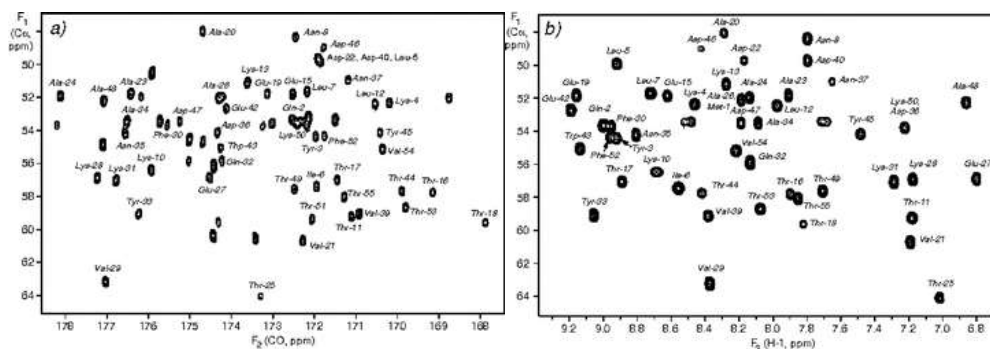


Figure 2.9: 2D haCACO spectrum is reported in panel A. This spectrum was acquired in a sequential fashion together with the 3D haCACoNH spectrum whose CACO plane is reported in panel B^[123].

In the parallel approach, instead, the FIDs of different nuclei are acquired at the same moment. This strategy found applications both in biomolecular NMR as well as in the determination of the structure of small molecules.

Regarding biomolecular applications, a possible experiment is represented by the 2D H,C'-N HSQC (Figure 2.10) where the H-N and C-N spectra are acquired in parallel^[113]. This experiment uses HN as starting magnetization pool. The magnetization is then transferred through INEPT creating the $2H_zN_z$ coherence. At this time point, nitrogen magnetization is evolved, with respect to the $^1J^{HN}$ and $^1J^{NC'}$, creating two different components proportional to N_z and $2N_zC'_z$ respectively. The two elements now follow two distinct transfer pathways with N_z that, in the end, will be converted for the detection of 2D HN spectrum. On the other hand, $2N_zC'_z$ evolves for the detection of 2D CON spectrum. The main drawback of this approach is the relative sensitivity of the two resulting spectra. Indeed, one of the two magnetizations (N_z in this case) is stored along the z axis until the other ($2N_zC'_z$) has accomplished all the needed transfer steps. In this way the desired magnetizations are created, and both the nuclear spins are ready for detection. Furthermore, splitting the initial magnetization in different sets requires adjustments for the relative distribution depending on the sensitivity of the detected nuclei. Moreover, $^1H^N$ is used as the common starting polarization source for the two experiments and results in the loss of proline resonances in the CON spectrum and of amide protons in fast exchange with the solvent.

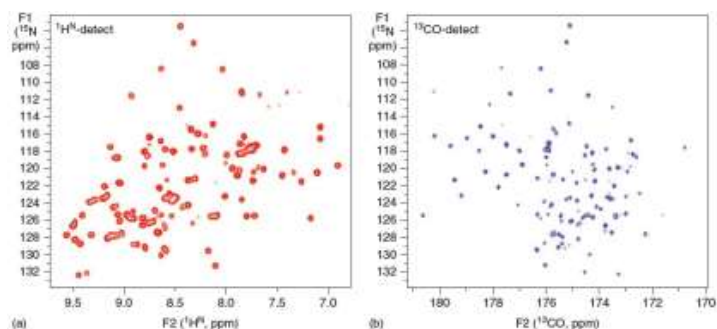


Figure 2.10: Panel A reports the 2D HN HSQC spectrum acquired in a parallel fashion together with the 2D CN HSQC whose spectrum is reported in panel B^[113].

Somewhere in the middle between the two mentioned approaches lies the “Unified Time-Optimized Interleaved Acquisition” (UTOPIA) strategy^[117]. This approach takes advantage of the detection of one (or more) FID(s) during the relaxation period of another experiment. The two experiments are thus identified as “parent” and “child”, with the child acquired during the parent’s relaxation delay. In the UTOPIA approach, the child is then acquired “for free” since the time needed for parent’s relaxation delay would be unused anyway.

As discussed previously, the 2D CON experiment represents one of the best solutions to obtain a fingerprint of an IDP. Thus, the UTOPIA strategy was exploited to obtain additional information during the $^{13}\text{C}'$ recovery delay in a sort of “Russian-Doll” approach^[120,121].

Together with the 2D CON experiment, the 2D HN experiment is usually acquired as one of the first experiments to investigate IDPs. We thus decided to combine these two experiments in the multiple receivers variant named CON//HN. The spectra obtained on α -synuclein at 298 K and pH 6.5 are reported in Figure 2.11. The main precaution taken to implement the 2D HN in the recovery delay of the 2D CON was to decouple $^{13}\text{C}'$ from ^{15}N during the evolution of this latter spin chemical shift in the 2D HN-HSQC. This was achieved with a pair of appropriately positioned 180° pulses, instead of a single one, to restore the $^{13}\text{C}'$ magnetization along the z axis and provide the starting polarization source for the subsequent CON experiment. The overall intensity of the two spectra is almost unchanged in the MR variants with respect to the standard ones. The CON experiment has only a 5% reduction in the S/N ratio while the HN one is a bit less sensitive with a reduction of 15%, not a big deal given the higher intrinsic sensitivity provided by the H^{N} nucleus.

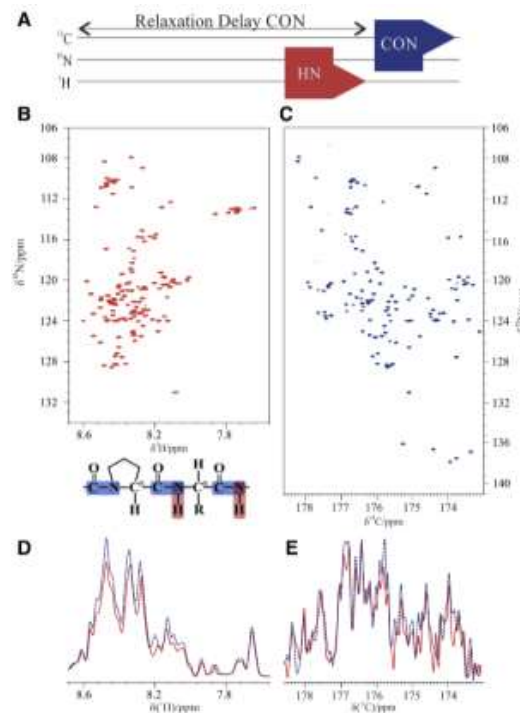


Figure 2.11: Panel A reports the schematic description of the CON//HN experiment. The relative spectra are reported in panel B (HN) and C (CON). These spectra were acquired on a ^{13}C - ^{15}N labeled α -synuclein sample at 285 K pH 6.5. Panels D and E report the projections of the HN and CON spectra respectively. Red solid traces refer to the multiple receiver experiments while blue dotted traces refer to the experiments acquired independently. Picture from ^[120].

If one wants to analyze samples with limited lifetime, the possibility to obtain two spectra in the time needed to acquire one of them becomes very useful as in the case of *in-cell* samples. Spectra acquired on a purified sample of α -synuclein at 310 K and pH 7.4 are reported in Figure 2.12.

In-cell samples are indeed affected by the presence of a high concentration of macromolecules surrounding the observed protein (leading to crowding). Thus, specific or un-specific interactions might occur between protein and other components of the cellular environment. These interactions slow down the tumbling rate of the molecule. As a consequence, the relaxation rate of the spins is increased, and an overall signal broadening is observed^[124]. Cross-peaks expected in our sample conditions are all present. When comparing the *in-cell* spectra with those acquired on the purified sample, as well with those acquired on cell lysates, it can be noted that ^1H signals are much more affected by line broadening with respect to ^{13}C ones confirming that this is a real property of the *in-cell* spectra and not due to changes in the sample.

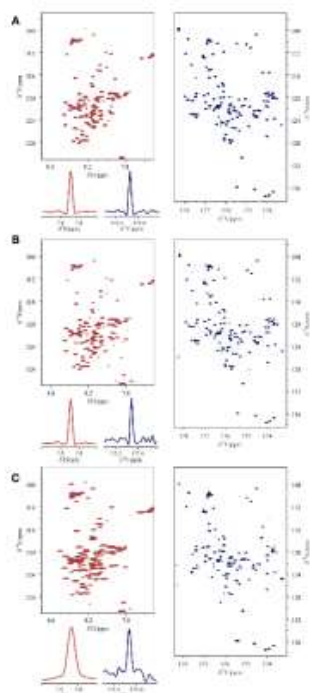


Figure 2.12: Comparison of the 2D spectra (HN left; CON right) acquired through the CON//HN experiment on ^{13}C , ^{15}N -labeled α -synuclein at 310 K. Panel A reports the spectra obtained on a purified sample. Panel B reports the spectra obtained on cells lysate resuspended in the same buffer of A. Panel C reports the spectra obtained for an in-cell sample. Picture from ^[120].

One more case that might require the acquisition of more than one spectrum at once is when a chemical reaction occurs, and it has to be monitored *via* NMR. The CON//HN experiment was used to follow the phosphorylation of α -synuclein with Fyn Kinase. Fyn Kinase is a non-receptor tyrosine kinase of the Src family which was found to highly phosphorylate tyrosine 125 in α -synuclein. Phosphorylation of Y125 represents a priming event in the efficient modification of serine 129 by CK1 kinase. This latter phosphorylation is involved in the modulation of Lewy bodies formation where S129 is selectively and widely phosphorylated, promoting fibrillation *in vitro*. In figure 2.13 the spectra for α -synuclein before and after phosphorylation are superimposed. Several changes are observed in both spectra acquired with multiple receivers. However, thanks to the superior resolution achieved with the CON experiment, additional details can be examined. In particular, peaks that in the HN spectrum fall in crowded regions, are instead well resolved in the CON spectrum such as Y125 itself.

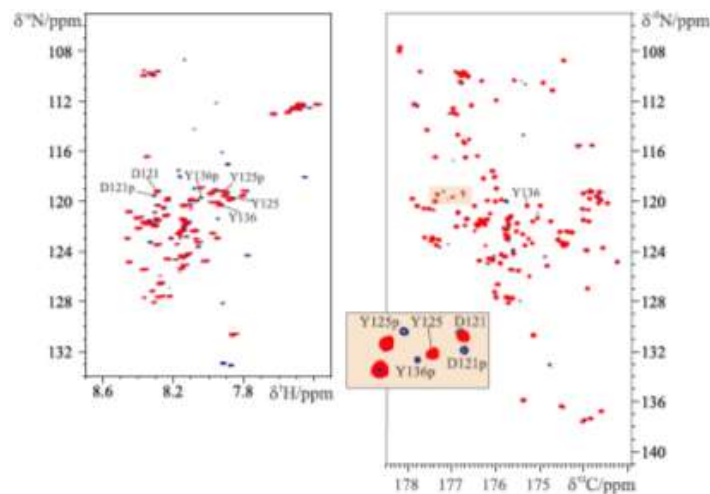


Figure 2.13: The HN (left) and CON (right) spectra of wild type a-synuclein (red) and Y125p a-synuclein (blue) acquired in the CON//HN experiment are reported and superimposed. One region of the CON is enlarged to highlight a subset of peaks that are influenced by phosphorylation. Picture from ^[120].

CON//HN is only one of the many possible combinations of experiments that can be implemented with the UTOPIA approach. Other possibilities that were implemented maintaining the CON as the parent experiment (*CON-centric* approach) are the CON//NH-BEST-TROSY^[76,125] and the CON//hCAN^[75]. Both these experiments are ¹⁵N detected. ¹⁵N detection, when not limited by sensitivity represents one of the best options for the study of IDPs thanks to the very large chemical shift dispersion and the narrow line width provided by this nucleus.

CON//NH-BEST-TROSY (Figure 2.14) exploits band selective pulses to selectively perturb amide protons and enhance longitudinal relaxation, shortening the overall time needed for the recovery delay of H^N atoms. In addition, the TROSY strategy avoids the necessity to decouple ¹H nuclear spins during ¹⁵N direct detection. Thus, the combination of the BEST approach, the TROSY one and the ¹⁵N direct detection, enables the acquisition of 4 transients of the NH experiment during the CON relaxation time which helps in compensating for the intrinsic low sensitivity of ¹⁵N. Here, the most important adjustment that was implemented in the BEST-TROSY part of the overall pulse sequence, was the introduction of the IPAP scheme. In this way the decoupling of ¹⁵N from ¹³C' and ¹³C^α during the acquisition is achieved without using standard decoupling. In this way, ¹³C' magnetization is preserved for the following CON experiment. Indeed, as it can be observed in Figure 2.12, panel D and E, the loss in S/N ratio is only minimal between the multiple receivers version of the experiment and the ones that do not make use of this approach. 2D NH-BEST-TROSY still relies on amide proton both during the initial transfer and the evolution of the indirect detection. Thus, even if detection

relies on a non-exchangeable nucleus, ^{15}N , increasing temperature, still broadens several peaks upon detection in the NH experiment while the quality of the CON is essentially maintained.

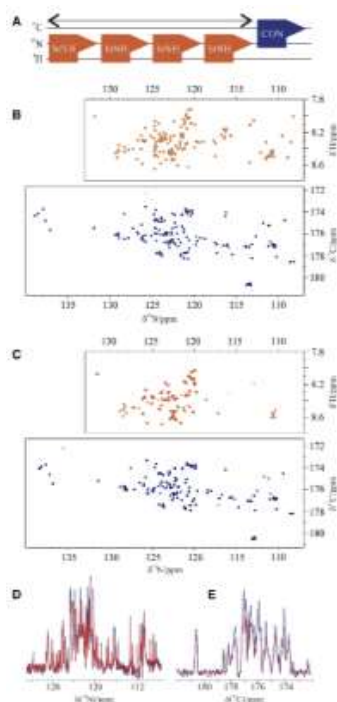


Figure 2.14: Panel A reports the schematic description of the CON//NH-BEST-TROSY experiment. The relative spectra are reported in panel B. These spectra were acquired on a ^{13}C - ^{15}N labeled α -synuclein sample at 285 K pH 6.5. Panels C report the same spectra but acquired on a ^{13}C - ^{15}N labeled α -synuclein sample at 310 K pH 7.2. Panel D and E reports the projections of the NH-BEST-TROSY and CON spectra respectively. Red solid traces refer to the multiple receiver experiments while blue dotted traces refer to the experiments acquired independently. Picture from [120].

On the contrary, the CON//hCAN experiment relies only on non-exchangeable nuclei providing spectra with a high content of information even at high temperature (Figure 2.15). Regarding the hCAN child-sequence, the magnetization starts from the $^1\text{H}^\alpha$ nuclear spin, enhancing the sensitivity of the experiment, and then is transferred to the $^{13}\text{C}^\alpha$ whose frequency is encoded in the indirect dimension. The polarization can now be transferred to two different nitrogen spins. The ^1J constant connects the C^α with the nitrogen of the same amino acid, while the ^2J constant connects it with the nitrogen of the following residue. The value of the two J constants is very similar (-11 Hz for ^1J and 7 Hz for ^2J). In our case the ^1J one was prioritized to enhance the intra-residue connectivity. Starting from H^α atoms, more than one repetition of the hCAN experiment can be achieved. In this way the overall sensitivity of this experiment was enhanced. In the end, the hCAN spectrum revealed all the

intra-residue correlations and the 81% of the inter-residue ones even if considerable overlap occurs in the C^α dimension.

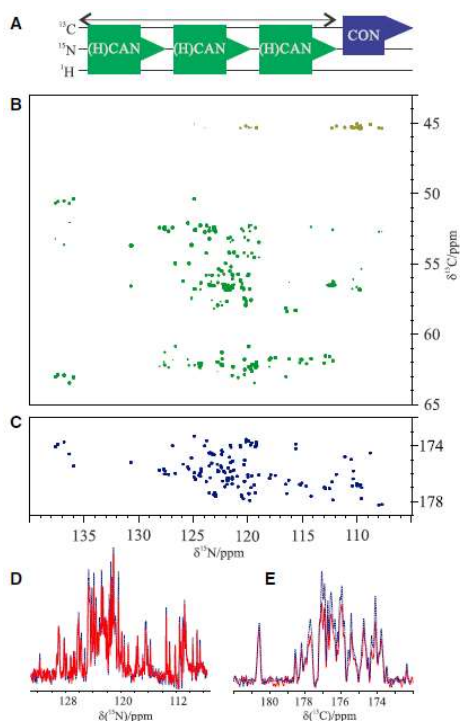


Figure 2.15: Panel A reports the schematic description of the CON//haCAN experiment. The relative spectra are reported in panel B. These spectra were acquired on a ^{13}C - ^{15}N labeled α -synuclein sample at 310 K pH 7.2. Panel D and E reports the projections of the haCAN and CON spectra respectively. Red solid traces refer to the multiple receiver experiments while blue dotted traces refer to the experiments acquired independently. Picture from ^[120].

For the complete description of the work refer to section 6 and ^[120].

2.3) First steps in IDPs' drug design: the identification of key motifs

Most of the targets for drug discovery efforts are represented by proteins and most of them are proteins with well-defined 3D structure. However, the role played by IDPs in human illnesses requires an extension in the landscape of drug discovery strategies^[4,5,42,44,45,126–128]. New challenges are thus emerging to identify novel strategies to interfere with the function (or malfunction) of IDPs (IDP's "druggability"). One of the main novelties is given by the nature of the interactions that IDPs employ for their function (fuzzy, folding-upon binding, unfolding upon binding, etc.). Indeed, IDPs' flexibility seems to provide an adaptable and specific interaction surface for binding^[45,128,129]. NMR represents a very powerful technique in this context thanks to the possibility to achieve atomic resolution information even in the case of highly flexible systems. Through this technique it is also possible to identify unexpected and new patterns, different from those of globular proteins but potentially fundamental and of general interest for IDPs.

In a work in collaboration with Prof. Ischiropoulos from the University of Pennsylvania we have investigated, through NMR, the effect of derivatives of nordihydroguaiaretic acid (NDGA) in preventing the formation of α -synuclein aggregates together with other biochemical and biophysical techniques^[130].

The role of α -synuclein in the insurgence of several neurodegenerative diseases, such as Parkinson's disease (PD) has been extensively studied^[131–136]. α -synuclein aggregation was found to be involved in the pathogenesis of these disorders. Possible drugs that might prevent α -synuclein aggregation have been investigated over the years.

The first molecules that were found to prevent aggregation were Dopamine, related catechols, different phenols, flavonoids and epigallocatechin gallate (EGCG)^[137].

Here the impact of nordihydroguaiaretic acid (NDGA) has been studied, a phenolic dibenzenediol lignan, in inhibiting α -synuclein aggregation to evaluate its potential as a chemical platform for the development of novel aggregation inhibitors.

NDGA was incubated for 24 hours with 1:1 α -synuclein and NMR was used to identify structural modifications within the protein's primary sequence. Only small changes were observed for Val 3, Phe 4, Met 5 and His 50 as reported in Figure 2.16. Besides NDGA, other similar molecules were investigated, i.e. cNDGA, the cyclized form of NDGA, and NDGA-1, a non-cyclized form containing only one pair of vicinal hydroxyl groups, whose effect were very similar to NDGA (See section 5 and [130] for a complete description).

The incubated samples were followed for several days through NMR with gradual changes of the previously reported peaks shifts. Moreover, additional changes were also found for several other peaks that might constitute a second set of α -synuclein population with a different conformation for these residues.

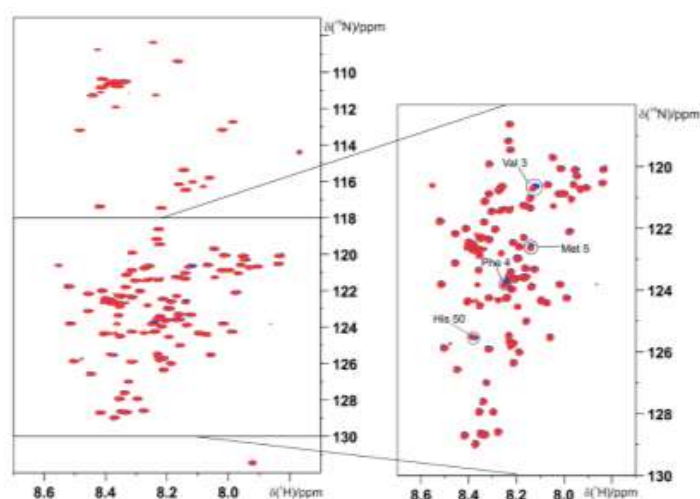


Figure 2.16: Superimposition of HN spectra of α -synuclein before (blue) and after (red) incubation 1:1 with NDGA for 24 hours. Picture from [130].

A second interaction study performed on α -synuclein regarded its interaction with calcium ions, which represent a potential trigger for the insurgence of synucleopathies, such as PD. α -synuclein is present in the presynaptic terminals and it is in close contact with microdomains, associated with neurotransmitter release, where Ca^{2+} ions are present in high concentration^[33,34,138–140]. The possibility to provide new structural and dynamic insights in the interplay between α -synuclein and calcium ions near-physiological conditions is very relevant.

For such reason, the interaction was followed by NMR, and carbon-13 detected experiments were exploited^[110]. The standard 2D-CON experiment, which provides information on the backbone's signals (i.e $\text{C}'_i\text{-N}_{i+1}$ correlations), was used together with a series of other 2D experiments, informative also on the side chains, usually seldom assigned in IDPs. In particular, the constant time version for the 2D CACO, 2D CBCACO and 2D CCCO experiments were developed and used for this purpose^[78,141]. The developed pulse sequences, together with a complete description of the work is provided in section 9 and in [110].

Thanks to this set of experiments, it was possible to detect and resolve cross-peaks from the side chains atoms. As shown in figure 2.17 for Asp2, side chains' resonances were assigned through exclusively 2D experiments. Starting from the carbonyl frequency in the CON spectrum the resonances of C^{α}_i and C^{β}_i can be identified in the CACO and CBCACO respectively. Two more cross-peaks connect these resonances to the carboxylate carbon resonance (C^{γ}). In this way the C^{β} - C^{γ} aspartate's cross-peaks were identified, and sequence-specifically assigned. Similarly, asparagine's side-chains resonances were assigned. Differently, to assign glutamate and glutamine's side chains the presence of one more cross-peak, the one involving C^{δ} , is required to achieve the completeness of information. Thus, the cross-peaks assignment of the carboxylate/carbonyl functional groups (C^{β}_i - C^{γ}_i for Asp and Asn, C^{γ}_i - C^{δ}_i for Glu and Gln) of α -synuclein amino acids was achieved.

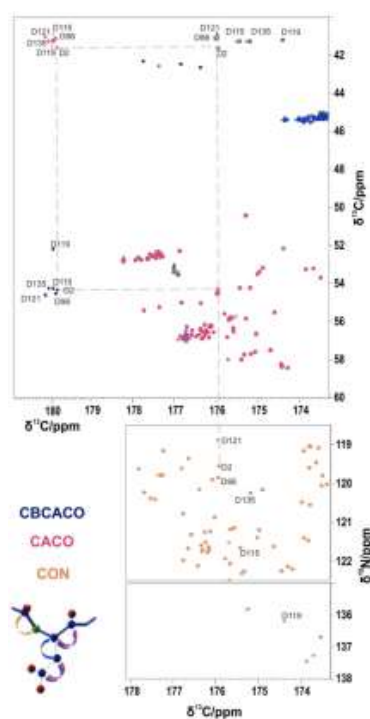


Figure 2.17: A description of the strategy used to assign in a sequence-specific manner the resonances of ^{13}C ASP2 residue through exclusively heteronuclear 2D NMR experiments. Grey dotted lines show the procedure followed to assign side-chain resonances. Starting from the carbonyl carbon resonance in the CON spectrum (orange), C^{α}_i and C^{β}_i are connected in CACO (red) and CBCACO (blue) spectra, superimposed in the figure. Finally, C^{γ}_i is correlated through C^{β}_i - C^{γ}_i and C^{α}_i - C^{γ}_i cross-peaks. Picture from [110].

Negatively charged residues are the first candidates for the binding of α -synuclein with positively charged calcium ions. Thanks to the obtained side chain's cross-peaks assignment, it is now possible to zoom-in on the putative interaction site. Following the interplay between α -synuclein and calcium

ions through CON and CACO, information were obtained both for the backbone and sidechains' atoms.

Inspecting the chemical shift perturbation effect on side chains' resonances, upon addition of Ca^{2+} ions, a clear localization on the primary sequence was found. The negatively charged residues present in the C-terminal region of the protein were much more perturbed than the ones present in the N-terminal and NAC regions (Figure 2.18).

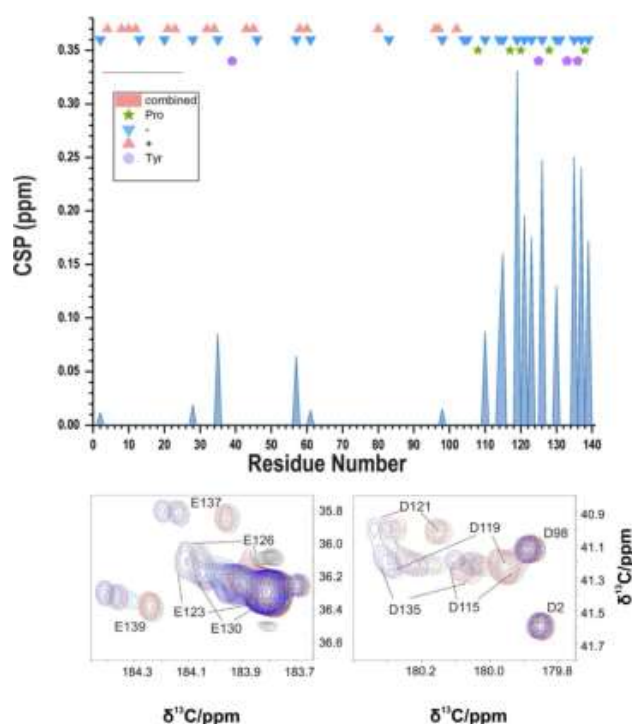


Figure 2.18: Top panel reports the chemical shift perturbation (absolute values) of aspartate and glutamate side-chain ^{13}C resonances upon interaction with Ca^{2+} . Over the graph is shown the distribution of proline (stars), tyrosine (pentagons), lysine (red triangles) and negatively charged (aspartate and glutamate, blue triangles) residues. Lower panel reports two regions of the CACO spectrum where the resonances of side chains are found. Picture from [110].

The same effect on the chemical shift values was also present on the backbone's resonances. This result still displays a localization on the C-terminal region that is rich in negatively charged residues, aromatic residues and proline residues (5 prolines in the C-terminal region). Interestingly, as shown in figure 2.19, two of these residues (P120, P138) were among the most perturbed in the sequence, which might sound unusual given the absence of binding properties for these residues. However, these amino acids are both flanked, in the primary sequence, by two negatively charged residues (D119-P120-D121 & E137-P138-E139).

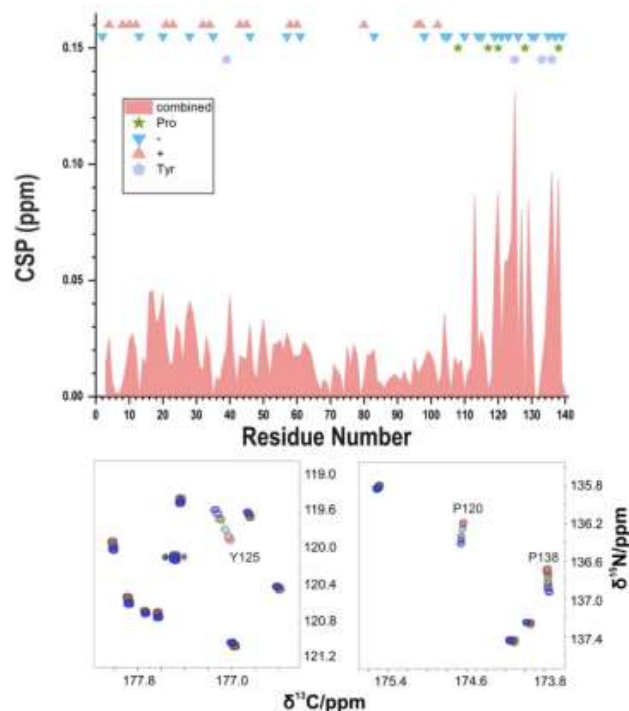


Figure 2.19: Top panel reports the chemical shift perturbation (absolute values) observed in the CON spectrum upon interaction with Ca^{2+} . The lower panels report the shift for two proline residues (P120 and P138) and for tyrosine 125, the most affected residues during the titration. The titration follows the pattern from blu ($\alpha\text{-syn}:\text{Ca}^{2+}$, 1:0) to purple ($\alpha\text{-syn}:\text{Ca}^{2+}$, 1:16). Picture from [110].

As previously mentioned, the flexible and disordered nature of α -synuclein would suggest a completely unspecific interaction with all the negatively charged groups interacting with positively charged ions in the same way. However, very different effects were detected for differently localized negatively charged residues, with the ones comprised in the C-terminal region much more perturbed than the others. In this frame, the primary sequence context has an impact in modulating the interaction of negatively charged residues with calcium ions. A closer inspection of the primary sequence revealed that two distinct tracts of the C-terminal region, quite far one from the other (D119-E126 & D135-E139), are the most perturbed and are very similar in amino acid composition. In particular, i) negatively charged residues are close but not consecutive in the primary sequence, ii) two glutamate residues are both preceded by a tyrosine residue and iii) two proline amino acids are found in the middle of negatively charged residues. Specific motifs are thus identified: tyrosine-glutamate pairs (Y125-E126 & Y136-E137) and negatively charged amino acids (E or D) separated by proline residues (D119-P120-D121 & E137-P138-E139). While the proline residues themselves do not possess any binding properties for positively charged ions, the presence of this amino acid in

the middle of two negatively charged amino acids could preform the sidechains of these latter amino acids and favour the proper orientation for calcium ion binding (Figure 2.20).

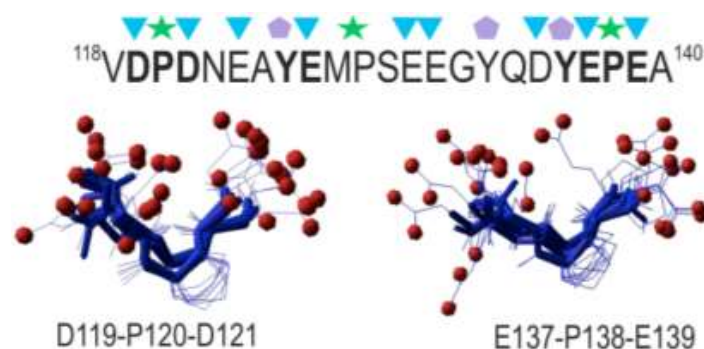


Figure 2.20: Structural models for the DPD and EPE motifs in α -synuclein. The conformers were calculated through Flexible Meccano without imposing any constraints and superimposed in MolMol. Picture from [110].

While for globular proteins the identification of binding sites is well established this is not trivial for IDPs which still represent a challenge from this point of view. To shed light on the wide versatility in which IDPs carry out their functions, the use of ^{13}C detected experiments was fundamental. The, newly identified, motifs seem thus to play a role in modulating structural and functional features of α -synuclein extending the knowledge on functional motifs enabled by disordered, flexible polypeptide tracts.

Other important motifs, and their role in modulating the compact state of an IDP, were investigated in a project in collaboration with the group of prof. R. Konrat from the university of Wien^[142].

Osteopontin (OPN), the IDP under investigation, exploits its function binding to different components of the extracellular matrix (ECM), including integrins and CD44, two membrane receptors. Osteopontin's action is regulated by several mechanisms such as post-translational modifications (phosphorylation, glycosylation, and cleavage) and splicing of OPN and CD44. Mis-regulation of OPN has been related to metastatic and aberrant cell function^[143-146].

From the structural point of view, previous studies identified the presence of regions characterized by a compact state, in the context of a fully disordered protein, fundamental for the correct exposure of OPN's interacting binding motifs like the deputed binding motif for integrin (RGD binding motif). Among the twenty natural amino acids, proline residues are very abundant in IDPs and are among those amino acids classified as "disorder promoting" residues. The presence of a pyrrolidine ring in the side chain provides peculiar structural characteristics. The steric hindrance given by the cyclic structure lowers the energy barrier present between the *cis*- and the *trans*- forms for this residue.

Indeed, while the occurrence of the *cis* conformation for any other amino acid is usually around 0.5%, for proline residues this is generally about 5-10% of the total^[147]. The specific role of proline residues in IDPs, as previously shown in the case of the interaction of α -synuclein with calcium ions, represents an interesting subject considering their important role in a number of different biological processes.

Carbon-13 detection represents one of the best tools to investigate these amino acids in IDPs. The chemical shift dispersion provided by the pyrrolidine ring ensures that these peaks are found in a very clean and precise position of the CON spectra. In IDPs, proline's nitrogen resonates between 134 ppm and 140 ppm and the carbonyl of the preceding amino acid falls between 175 ppm and 170 ppm. Moreover, due to the lack of 3D structure, the cross-peaks involving prolines' nitrogen tend to cluster in smaller regions depending on the nature of the preceding residue, resulting in "proline-fingerprint" as it can be seen in the case of one of the intrinsically disordered regions of CBP (CBP-ID4) reported in Figure 2.21^[148].

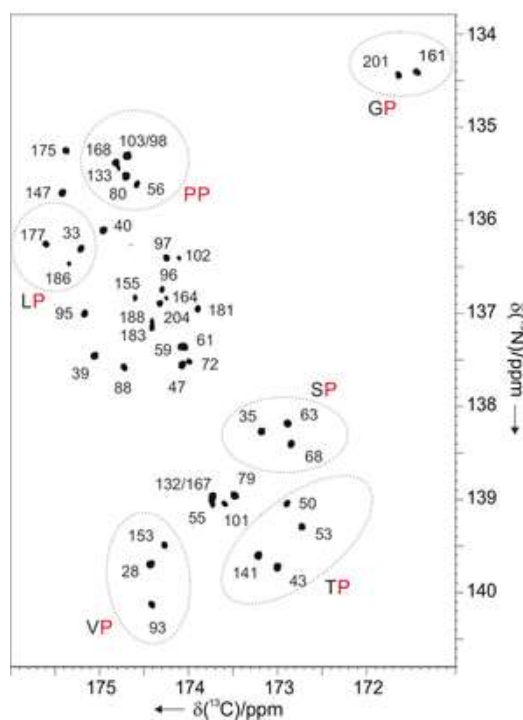


Figure 2.21: 2D CON^{Pro} spectrum showing proline residues fingerprint of ID4 is reported. C'-N connections involving the same residue-type pairs are highlighted. Picture from ^[148].

Inspecting the CON spectrum of quail Osteopontin (qOPN, the protein variant investigated), several minor forms of the major peaks were found (Figure 2.22). To better characterize these minor forms

and given the low population from which these peaks arise, a sequence-specific assignment was performed making use of ^1H -strat ^{13}C -detected 3D experiments. 3D hCBCACON, 3D hCBCANCO and 3D hCCCON experiments were indeed acquired and used for the sequence-specific assignment^[78,85,149]. These experiments were used to identify and connect the residues through ^{13}C aliphatic chemical shifts (backbone and side chains). In conjunction, a 3D hacaCOCON was also used to achieve additional information involving ^{13}C carbonyl chemical shifts^[84,150,151]. This latter experiment exploits the sensitivity improvement using proton as a starting magnetization pool and relies on isotropic mixing for the transfer of magnetization via $^3\text{J}^{\text{C}'\text{C}'}$. The ^3J value ranges between 0.3 Hz and 3 Hz depending on geometrical constraints that might limit the detection of cross-peaks. In the case of qOPN's major peaks, three residues were always connected through this experiment, from residue i to $i+2$. In the case of minor peaks, all of them presented the connection with the preceding residue while only a few of them presented all the expected connectivities. This effect is not only related to the low sensitivity given by the low population of the minor forms. Rather, it depends on specific backbone's geometrical conformations suggesting small $^3\text{J}^{\text{C}'\text{C}'}$ values.

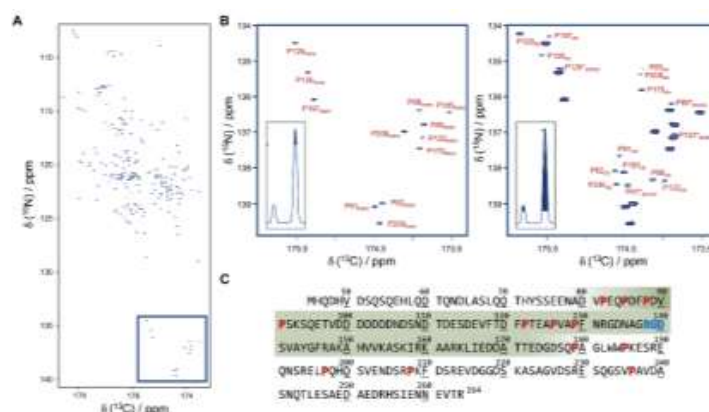


Figure 2.22: Panel A reports the 2D CON spectrum of qOPN highlighting the proline region. Panel B shows both the major (left) and the minor (right) forms of proline's peaks with their respective assignment. In the inset, the relative intensity between the major and minor forms are sketched. Panel C reports the primary sequence of qOPN construct used with proline residues highlighted in red, the integrin binding site (RGD motif) in blue. The green box highlights the residues encompassed in the compact state. The Quail OPN was preferred to the Human one due to a superior stability once the first 44 residues are removed. Picture from ^[142].

Thanks to the previously mentioned experiments all the resonances from the proline minor forms in the 2D CON experiments were successfully assigned in a sequence specific manner and are reported in Figure 2.22.

Measuring the volume ratio between the *cis*- and *trans*- forms of the cross-peaks in the 2D CON experiment, a quantification of the relative population of the two isomers of the peptide bond involving proline nitrogens has been achieved (Figure 2.23). P88, P91, P122 and P126 possess a higher *cis* population (> 10%) which correlates with the presence of an aromatic residue in position either $i\pm 1$.

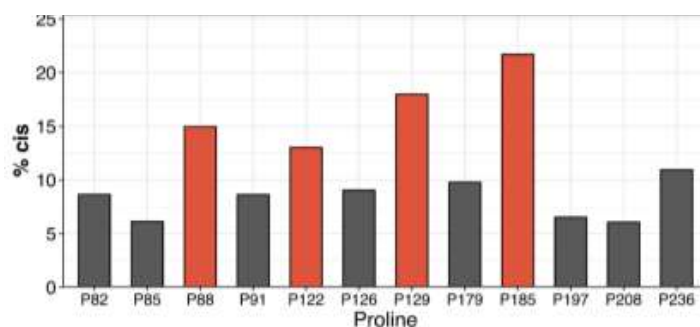


Figure 2.23: Quantification, in percentage, of *cis* population obtained by measuring the volume ratio between the *cis*- and the *trans*- forms for proline signals in the CON spectrum. The proline residues that present an aromatic amino acid in position $i\pm 1$ are reported in orange. Picture from [142].

To further understand the role of proline residues in modulating protein dynamics, $R_2(^{15}\text{N})$ measurements were conducted and revealed a reduction in the mobility for the *cis*- form with respect to the *trans*- one (Figure 2.24).

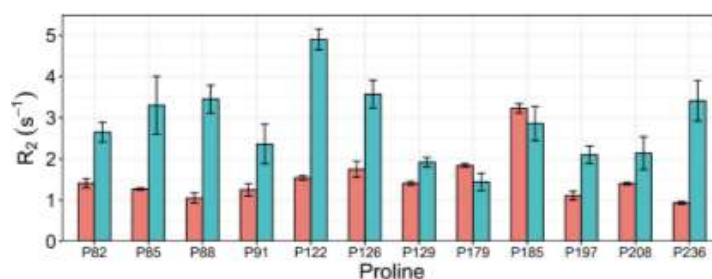


Figure 2.24: $R_2(^{15}\text{N})$ relaxation rates for *trans*-prolines are reported in orange and *cis* values are reported in cyan. Increased R_2 values are found for *cis*-prolines, especially for those residues in regions with higher differences in PPII content (Refer to Figure S5 in section 8) between the *cis* and the *trans* conformers. Picture from [142].

Using the two different sets of chemical shifts (major and minor forms), an analysis was conducted using the $\delta 2\text{D}$ software to reveal partially populated secondary structural elements^[152]. A lower propensity for the *cis*- form to sample a PPII conformation was computed. The hypothesis is that these minor forms sample higher PPI conformation providing a reduction in the flexibility of the

protein. Indeed, while PPII is an elongated left-handed helix with a helical pitch of 9.4 Å, PPI is a compacted right-handed helix with a helical pitch of 5.6 Å. These results agree with the presence of a more compact state sampled by the protein. A compact state can be defined as states characterized by reduced local mobility and with a character in between collapsed and random coil conformations. However, these states do not (or not necessary) present any secondary structure. Further studies that comprise also Paramagnetic Relaxation Enhancement (PRE) experiments revealed that only local dynamic differences are present between the *trans*- and the *cis*- form without any relevant modification in the overall compact state of the protein.

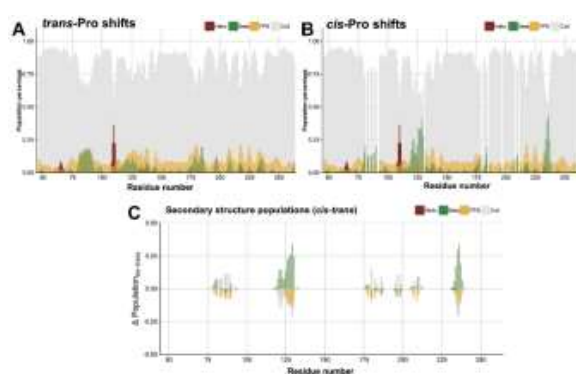


Figure 2.25: Panel A reports secondary structure population obtained by the $\delta 2D$ software using the chemical shift values of the major forms as input. Panel B reports secondary structure population obtained by the $\delta 2D$ software using the chemical shift values of the minor forms as input. Panel C reports the differences in secondary structure populations between the *cis*- and *trans*- form. Minor forms are characterized by lower PPII populations but higher β -strand conformation. Picture from ^[142].

On the contrary, to corroborate the idea that local dynamic changes are related to different side chain's conformation, $^{15}\text{N}\{-^1\text{H}\}$ heteronuclear NOE and ^{15}N 3D NOESY-HSQC were acquired. The analysis of side chain signals of tryptophan residues in the stretch W183-W184-P185 revealed differences between W183 and W184 as well as different dynamic conformation between the minor and major forms of W184. W184 minor form displays indeed a reduced mobility associated to changes in the orientation of the side chain. Steric repulsion, modulated by π -CH interaction led to reduced local mobility in the *cis*- form as illustrated in figure 2.26.

That π -CH interactions can stabilize specific conformations in Aro-Pro motifs was already described in the case of small molecules and peptides^[153–155]. The tunable effect is usually limited to prolines in position i and aromatics in positions $i\pm 1$ and $i-2$.

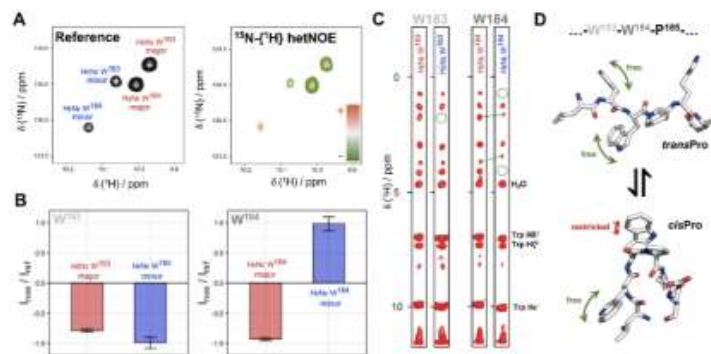


Figure 2.26: Panel A reports the heteronuclear NOE experiments (left reference, right ^{15}N -saturation) and the ratio between the saturated spectrum and the reference one ($I_{\text{sat}}/I_{\text{ref}}$) is reported in panel B for the minor and major form of W183 and W184. Panel C reports the amide $^1\text{H}^{\text{N}}$ to side-chain ^1H NOE for minor and major forms of W183 and W184, obtained in a ^{15}N 3D NOESY-HSQC experiment. Panel D reports a schematic 3D structural model of the sequence motif—WWP—in cis-Pro and trans-Pro forms. In the cis-Pro, distinct CH-aromatic interactions lead to restricted local mobility. Picture from ^[142].

The identification of patterns of amino acids in IDPs able to tune different protein characteristics might become of general interest to better understand the way they act and to test these proteins as targets for drug discovery. The examples reported here include the case of α -synuclein (DPD, EPE motifs involved in calcium binding) and that of Osteopontin (aromatic-proline pairs involved in the formation of compact states). A common feature is already emerging from these two unrelated IDPs: the key role played by prolines and aromatic residues. The possibility to characterize as many motifs as possible in IDPs could open new avenues for drug design to face the onset of diseases related to IDPs.

2.4) Viral proteins challenge carbon detection (and lose)

During my Ph.D. I also had the opportunity to work with two viral proteins with a high content of disorder. These two viral proteins were both challenging from the NMR point of view. The PNT protein from Nipah Virus (NiV-PNT) is a large protein composed by 506 residues^[156]. The Nucleocapsid protein from SARS-CoV-2 is structurally heterogeneous due to the presence of three disordered regions and two folded domains^[157].

The Nipah virus (NiV), is a zoonotic paramyxovirus responsible for severe encephalitis in humans, classified as biosecurity level 4 (BSL-4) pathogens and considered as potential bio-terrorism agent^[158]. *Exclusively heteronuclear* NMR experiments, combined with ¹H detected ones, were used to obtain sequence-specific assignment of the N-terminal domain of the Nipah virus P/V protein, a protein constituted by 506 residues whose HN and CON spectra are reported in Figure 2.27.

Three dimensional ¹³C detected experiments were acquired in their ¹H-start version. hCBCACON, hCBCANCO and hCCCON were exploited to identify the residue type and eventually connect them in a sequence-specific manner using the information provided by the carbon atoms composing the side chains^[78,85,149].

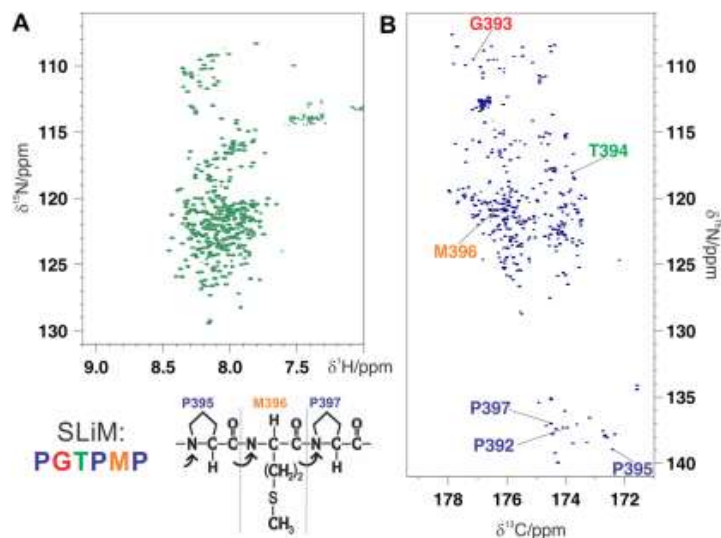


Figure 2.27: 2D $^1\text{H}^{15}\text{N}$ -BEST-TROSY spectrum is reported in panel A and ^{13}C -detected CON spectrum is reported in panel B. Both the spectra were acquired on the full-length (506 residue long) NiV-PNT. The 2D CON enables also the direct detection of the correlations involving the 26 prolines in a clean region of the spectrum. As an example of the importance of those residues, the cross peaks of the residues in a SLiM are reported. Picture from [156].

3D hCBCACON connects the C^{β}_i and C^{α}_i to C'_i , directly acquired, and N_{i+1} , providing information on the nature of the residue i using the resonance values of C^{α} and C^{β} . The coupled experiment, 3D hCBCANCO exploits the ^1J and ^2J from C^{α} to N. Thus, two sets of resonances are obtained for each C'_i nucleus, directly acquired: one comprising C^{β}_i , C^{α}_i , and N_{i+1} , the second one comprising C^{β}_{i+1} , C^{α}_{i+1} , C'_i and N_{i+1} . 3D hCCCON becomes indispensable in the case of ambiguous pairs of $\text{C}^{\beta}/\text{C}^{\alpha}$ resonances that could be referred to more than one kind of residue. This experiment provides, when existing, the resonances of C^{γ} , C^{δ} and C^{ϵ} resolving the ambiguities that could arise from the availability of only $\text{C}^{\alpha}/\text{C}^{\beta}$ chemical shift values. In conjunction, the previously described 3D haCOCON was acquired as well to obtain connections between C'_i and $\text{C}'_{i\pm 1}$ [84,150,151].

Moreover, a 5D experiment, the 5D hacaCONCACON, was exploited in its APSY version [90]. With this strategy, an automated peak peaking is performed, and a 5D peak list is also provided. Making use of the strategy reported in the introduction section for the very same experiment, connections between different 2D CON peaks can be performed and used as a starting point for the sequence-specific assignment.

Three-dimensional ^1H -detected experiments were acquired as well to complete the information provided by ^{13}C detected experiments. 3D HNCO and 3D HNcaCO were used to transfer the assignment from the ^{13}C -detected experiments to the ^1H -detected ones [9]. Moreover, 3D HNCACB

and 3D HNcoCACB were used to confirm the previously obtained C^α and C^β assignment. Finally, 3D hNcocaNNH and 3D hNcaNNH were also acquired^[159,160]. The hNcocaNNH experiment provides connectivity between H_i , N_i and N_{i+1} while the hNcaNNH experiment provide connectivity between H_i , N_i and $N_{i\pm 1}$. These experiments rely on 1H detection that do not provides the best chemical shift dispersion in the direct dimension. However, nitrogen frequencies are encoded in both the indirect dimensions providing a wide chemical shift dispersion that enable the connectivity of up to three amino acids (from N_{i-1} to N_{i+1}).

Once the 91% of the sequence-specific assignment was achieved (BMRB 50370), secondary structure propensity values were calculated making use of the online available tool ncSSP^[161]. This tool estimates the content of α -helix or β -strand for each residue based on the available chemical shift values for that residue. These values are then compared with the random coil ones^[162]. In this version of the algorithm the random coil chemical shift resonances are averaged considering neighboring amino acids in the primary sequence. To investigate backbone dynamics of NiV PNT, heteronuclear 1H - ^{15}N NOE, $R_2(^{15}N)$, and $R_1(^{15}N)$ values were measured together with CLEANEX experiment. The overall description that was obtained through these data (Figure 2.28) demonstrates the high flexibility and the disordered nature of this protein. However, looking in more detail at the data, and considered the primary sequence composition, segments with reduced dynamic and residues with possible crosstalk to each other were identified. These data, together with the ones obtained through SAXS experiment, demonstrate that this intrinsically disordered protein could be well described by an ensemble of protein with conformational heterogeneity. This heterogeneity is not unexpected given the many Short Linear Motifs (SLiMs) encompassed by this protein and predicted by the ELM website^[49]. The structural flexibility and the disordered nature of the protein represent a powerful tool to adapt the structure to possible binding partners or to perform the desired function.

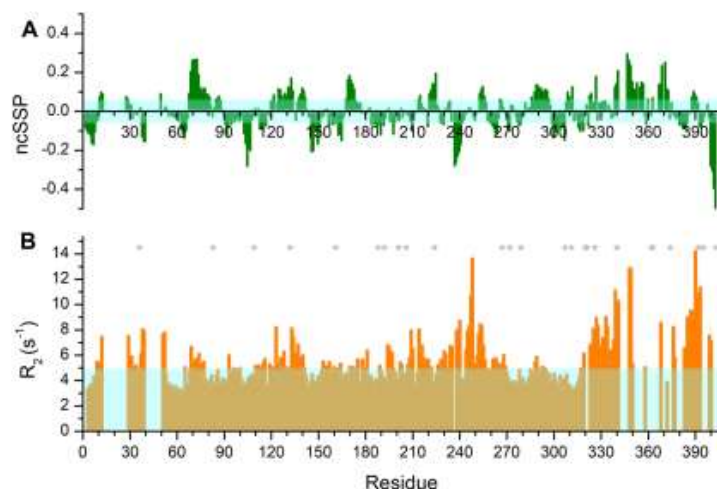


Figure 2.28: The neighbour-corrected secondary structure propensity (ncSSP) values obtained using the chemical shift values reported in the BMRB deposition 50370 are reported in panel A. The light cyan box represents the ± 0.05 value. Most of the protein is found in a disordered conformation with just a few regions displaying a moderate α -helix propensity (SSP > 0.2). $R_2(^{15}\text{N})$ values are reported as a function of residue number in panel B. The light cyan box indicates the modal value. Picture from [156].

These results, extensively discussed in section 10 and in [156] represent a starting point for obtaining atomistic information in future interaction studies involving NiV PNT. The obtained chemical shifts values on the full-length protein are indispensable to map any binding process in the context of either the isolated PNT domain or considering the complete biological context.

The SARS-CoV-2 virus is responsible for COVID-19 disease. This virus belongs to the β -class of coronavirus that can infect animals and humans. Other β -coronavirus are the severe acute respiratory syndrome virus (SARS-CoV), and the Middle East respiratory syndrome virus (MERS-CoV)^[163,164]. All these viruses attack the lower respiratory system causing viral pneumonia, while they may also affect the gastrointestinal system, heart, kidney, liver, and central nervous system leading to multiple organ diseases^[165,166].

While nowadays different vaccines have been developed and accepted by many international agencies, efficient drugs that can stop the disease are still lacking.

SARS-CoV-2 virus encodes 20 proteins, 4 of them are structural proteins: Spike (S), Envelope (E), Membrane (M) and Nucleocapsid (N).

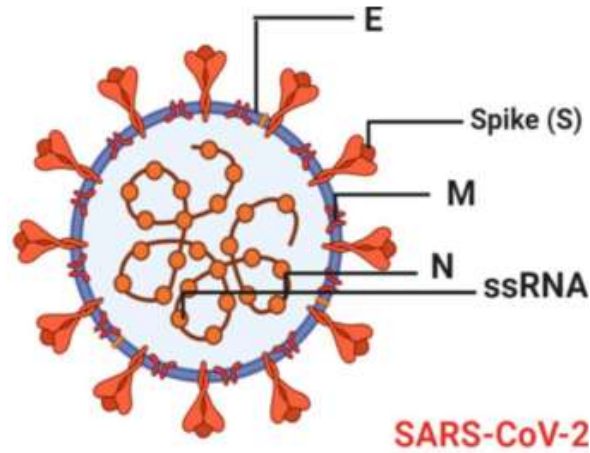


Figure 2.29: A sketch of the SARS-CoV-2 is reported highlighting the structural proteins and the single strand RNA (ssRNA)

The SARS-CoV-2 Nucleocapsid protein, is known to bind and stabilize the genomic RNA forming ribonucleoprotein complexes (RNP)^[167–170]. SARS-CoV-2 Nucleoprotein is composed by 419 residues subdivided in five domains. Bioinformatic tools were used to predict the content of disorder in this protein. Three regions were predicted to be disordered and count for about the 30% of the primary sequence^[171]. The N-terminal ordered domain (NTD) is the one deputed to the binding with RNA while the C-terminal ordered domain (CTD) is known to be the homodimerization domain and their structures were solved^[172–174].

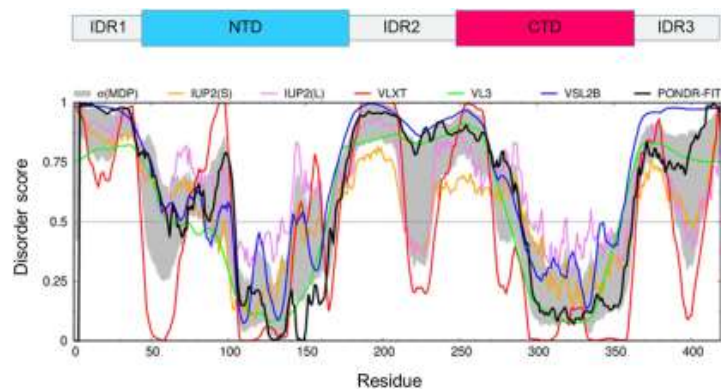


Figure 2.30: Bioinformatics analysis of the intrinsic disorder content of the SARS-CoV-2 nucleocapsid N protein obtained using IUPred short (golden line), IUPred long (purple line), PONDR® VLXT (red line), PONDR® VL3 (green line), PONDR® VSL2B (blue line), PONDR® FIT (black line). Protein segments with a disordered score much larger than 0.5 are considered disordered, while regions with disorder scores between 0.2 and 0.5 are considered as flexible. Over the plot, the domain organization used in the text is reported. Picture from ^[157].

The homology between this protein and the N protein from SARS-CoV is larger than 90%^[175]. In the protein from this latter virus, the importance of the disordered regions was proven by low resolution techniques. In these previous studies, IDRs were found to be responsible for mechanisms that regulate the formation of the RNP complex. However, structural and dynamic atomic information regarding the IDRs of SARS-CoV-2 Nucleocapsid protein still lack.

A protein construct comprising the NTD domain, and the two flanking disordered regions was thus conceived, expressed and purified (¹IDR1⁴⁷-NTD-¹⁷⁶IDR2²⁴⁸ construct, referred as NTR).

2D HN experiment demonstrated the structural heterogeneity of this protein.

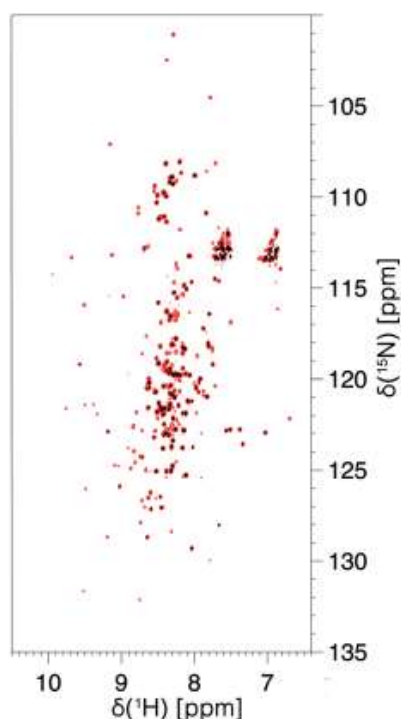


Figure 2.31: The 2D HN BEST-TROSY of NTR construct is reported. The figure displays the overlay of two different processing of the same spectrum: the black one is optimized for the resolution and the red one is optimized for the signal to noise ratio. Picture from ^[157].

In this spectrum two sets of signals can be identified, one with well-resolved peaks that can be associated to the folded NTD domain. The second set, characterized by lower dispersion but higher intensity, comprises the resonances of the intrinsically disordered residues.

Given the structural complexity of this protein a set of ¹³C direct detected experiments was used to exclusively monitor the intrinsically disordered regions. Indeed, as reported in the CON spectrum of figure 2.32, only the resonances of the disordered regions are detected.

as to resolve signals of low complexity regions such as the polyQ (238-242) or different repeats like the Asn-Arg reported in the middle panel of figure 2.32.

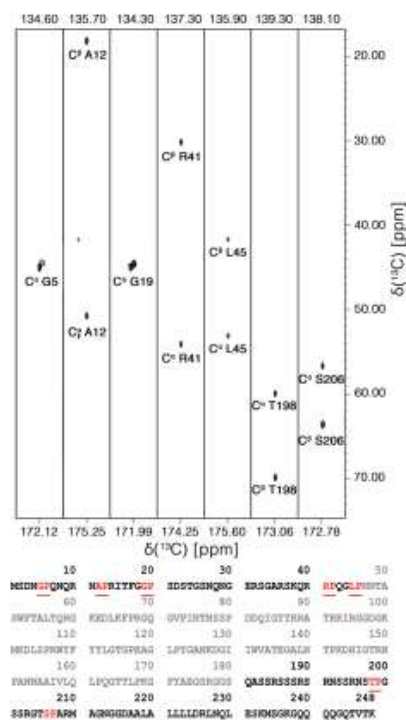


Figure 2.33: Strips derived from the 3D hCBCACON experiment extracted at the ^{15}N chemical shift of proline residues (i) are reported in the top panel. The C', C $^\alpha$ and C $^\beta$ frequencies belong to the i-1 amino acid providing the X-Pro assignment. The lower part of the figure reports the NTR primary sequence and X-Pro pairs are highlighted. Five proline residues are found in the IDR1 and two in IDR2 domain. The primary sequence of NTD domain is reported in grey. Picture from ^[157].

To complete, and verify the assignment, 3D HNCOC and 3D HNCO were acquired as well. Through this strategy the 98% of the sequence specific assignment was obtained for IDR1 (1-47, BMRB 50619) and 91% was obtained for IDR2 (176-248, BMRB 50618). To determine the structural feature of the two IDRs, H N , C $^\alpha$, C $^\beta$, C' and N chemical shift values were used to calculate secondary structure propensity. The calculation confirmed the disordered behavior of the two studied regions. A moderate propensity of α -helix conformation is sampled in one of the most hydrophobic regions of the construct, the leucine rich region (218-232) where few residues escaped detection probably due to the conformational exchange that broaden the signals. These data are in agreement with the bioinformatic analysis that predicts the presence of some structure in the region 215 – 232 while the rest of the linkers are predicted to be disordered^[171,176].

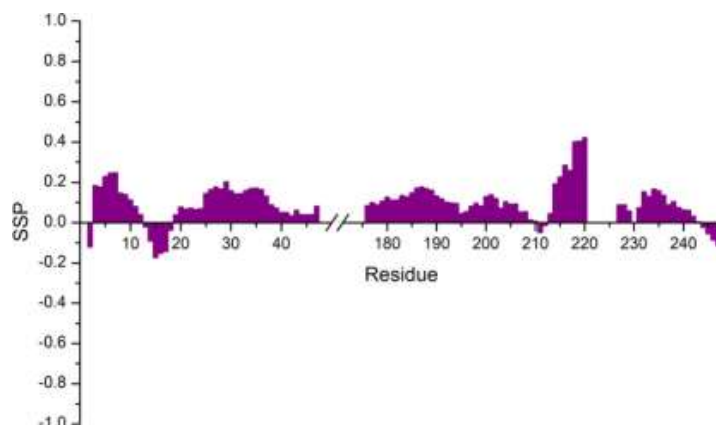


Figure 2.34: Secondary Structure Propensity (SSP) plot is reported for the assigned residues^[161]. Chemical shift values for HN, N, C', C α , and C β nuclei were used. The two regions result to be highly disordered with a slight tendency to be in an α -helix conformation for the residues 216–220. Picture from ^[157].

As for NiV PNT, the sequence specific assignment of the disordered regions of nucleoprotein represents a first step for further studies. The interactions with RNAs, with viral proteins as well as with potential drugs are just few of the possible landscapes that can now be investigated with atomic resolution, with potential consequences in unexplored approaches in drug discovery.

3. Conclusion and perspective

Solution NMR, among all the possible biophysical and biochemical techniques, has always been one of the key methods in structural biology. The possibility to obtain structural and dynamic information at atomic level represents one of the main advantages provided by NMR in protein studies.

In the last decades, IDPs emerged as an important class of proteins. These macromolecules are involved in several cellular processes, in the insurgence of many neurodegenerative diseases and exploit important functions also in different viruses.

Therefore, the study of structural disorder becomes fundamental for a complete understanding of the behaviour of these proteins.

However, the structural disorder itself poses many limitations in the study of IDPs through conventional NMR strategies. For such reasons, novel approaches that fit with the inherent properties of these proteins must be designed and applied.

My doctoral project had indeed the aim of improving the available NMR methods to study IDPs.

One of the main problems that arises when studying IDPs through NMR is the efficiency of the amide proton exchange with the solvent. This effect becomes more efficient with increasing temperature, salt concentration of the sample and pH approaching physiological conditions. The overall result is that exploiting H^N detection, peaks can be severely broadened even beyond detection. Amide proton exchange is however an important source of dynamic information and only few methods to study this process exist in the case of IDPs.

During my Ph.D. period I focused on the development and application of a method to study this event. Carbon-13 nuclei do not suffer from the exchange process. Thus, a method that relies on ^{13}C direct detection was exploited. In this experiment, named DeCON, the main idea is to observe the decay of a triple spin order operator ($4H_zC_zN_z$). The main contribution to this decay is indeed given by the exchange process. In this way it was possible to quantify the k_{ex} rate of α -synuclein in presence and absence of calcium ions at near physiological conditions revealing important dynamic changes in specific portions of the primary sequence.

The continuous increase in magnetic field strength, improvements in probe designs and renovation of the consoles allow us to think of complex NMR experiments providing different ways to collect more than one FID at the same time.

On this line the implementation of the UTOPIA approach for the collection of interleaved experiments exploiting the multiple receivers (MR) hardware was one of the main topics of my Ph.D. period. With this approach one experiment (child) is acquired during the relaxation delay of a second experiment (parent).

In all the developed MR experiments, the parent was represented by the CON experiment that relies on a quite long relaxation delay (about 2 seconds). The CON experiment was combined with HN in the CON//HN experiment. This experiment was first tested on α -synuclein and then used to monitor the same protein in *E. Coli* cells. With the same experiment the phosphorylation of Y125 was followed providing important information on the phosphorylation reaction.

CON was also combined with two different ^{15}N detected experiments. CON//NH-BEST-TROSY was used to monitor α -synuclein at two different temperatures and it was shown that a few more residues were detected at high temperature with respect to the 1H detected counterpart. Exploiting the BEST and TROSY approach, four repetitions of the NH experiment can be acquired during the CON recovery delay.

CON//hCAN experiment was tested at high temperature as well and displayed all the intra-residue peaks expected and almost the 85% of the inter-residue coherences. Exploiting H^a protons as the

starting magnetization source, three repetitions of the hCAN experiment can be acquired during the CON recovery delay.

IDPs are far from random coil conformations both from a structural and dynamic point of view. However, the driving forces that establish the physical features of IDPs are not clear at all and are not easy to be determined. Moreover, IDPs are usually characterized by unspecific binding patterns, fuzzy interactions, and low affinity for the partners. These elements hamper an easy characterization of the binding properties of these macromolecules. The identification of key motifs that modulate IDPs' features becomes thus an important topic.

Studying the interaction between α -synuclein and calcium ions, important patterns were identified. A series of exclusively heteronuclear 2D experiments were implemented to monitor the side chains of negatively charged residues and the interaction was found to be very localized in the C-terminal region of α -synuclein. Moreover, through the 2D CON, the most perturbed residues were found to be two prolines and a tyrosine. Inspecting α -synuclein's primary sequence, it was found that the two most perturbed prolines were flanked by negatively charged residues and that one more tyrosine-negatively charged amino acid pair was perturbed. In this way we identified two patterns involved in the interaction with calcium ions (DPD, EPE and YE).

Investigating Osteopontin, another key motif was identified. Indeed, the presence of an aromatic amino acid in position ± 1 respect to a proline residue seems to favour the presence of a more compact state. Indeed, measuring the population of the *cis*-form of proline residues it was found that the presence of close aromatic residues enhances the minor (*cis*-) form population. The effects were attributed to the presence of the aromatic side chains that establish π -CH interaction stabilizing the *cis*- conformation. These minor forms were found to possess a reduced mobility respect to the major (*trans*-) ones. The higher population, provided by the presence of aromatic residues, and the reduced mobility of the *cis*-Pro contribute to the presence of a compact state.

IDPs play an important role also in viruses. In the context of my doctoral project, I had the opportunity to study two different disordered viral proteins. Carbon detected NMR experiments were exploited to obtain the almost complete assignment of NiV-PNT and of the first two disordered regions from the Nucleocapsid protein of SARS-CoV-2.

In conclusion, the methods and the discoveries presented in this thesis while representing only a small step forward in the development of methods to target IDPs could be of general application and interest in this fascinating field of structural biology.

4. Bibliography

- [1] K. Wüthrich, *J. Biol. Chem.* **1990**, 265, 22059–22062.
- [2] A. Cavalli, X. Salvatella, C. M. Dobson, M. Vendruscolo, *Proc. Natl. Acad. Sci.* **2007**, 104, 9615–9620.
- [3] G. M. Clore, A. M. Gronenborn, *Crit. Rev. Biochem. Mol. Biol.* **1989**, 24, 479–564.
- [4] M. Pellecchia, I. Bertini, D. Cowburn, C. Dalvit, E. Giralt, W. Jahnke, T. L. James, S. W. Homans, H. Kessler, C. Luchinat, B. Meyer, H. Oschkinat, J. Peng, H. Schwalbe, G. Siegal, *Nat. Rev. Drug Discov.* **2008**, 7, 738–745.
- [5] A. D. Gossert, W. Jahnke, *Prog. Nucl. Magn. Reson. Spectrosc.* **2016**, 97, 82–125.
- [6] A. Zhuravleva, D. M. Korzhnev, *Prog. Nucl. Magn. Reson. Spectrosc.* **2017**, 100, 52–77.
- [7] F. Bloch, W. W. Hansen, M. Packard, *Phys. Rev.* **1946**, 69, 127–127.
- [8] M. P. Williamson, D. Marion, K. Wüthrich, *J. Mol. Biol.* **1984**, 173, 341–359.
- [9] A. Kay, L. E.; Ikura, M.; Tschudin, R.; Bax, L. E. Kay, M. Ikura, R. Tschudin, A. Bax, *J. Magn. Reson.* **1990**, 89, 496–514.
- [10] P. Symposium, *Processing* **1999**, 189, 182–189.
- [11] V. N. Uversky, C. J. Oldfield, A. K. Dunker, *Annu. Rev. Biophys.* **2008**, 37, 215–246.
- [12] W. Basile, M. Salvatore, C. Bassot, A. Elofsson, *PLoS Comput. Biol.* **2019**, 15, e1007186.
- [13] W. Basile, M. Salvatore, C. Bassot, A. Elofsson, *PLoS Comput. Biol.* **2019**, 15, e1007186.
- [14] V. N. Uversky, J. R. Gillespie, A. L. Fink, *Proteins Struct. Funct. Genet.* **2000**, 41, 415–427.
- [15] V. N. Uversky, *J. Biomol. Struct. Dyn.* **2003**, 21, 211–234.
- [16] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, A. K. Dunker, *Biophys. J.* **2007**, 92, 1439–1456.
- [17] V. Uversky, *Front. Biosci.* **2009**, 14, 5188–5238.
- [18] A. K. Dunker, M. M. Babu, E. Barbar, M. Blackledge, S. E. Bondos, Z. Dosztányi, H. J. Dyson, J. Forman-Kay, M. Fuxreiter, J. Gsponer, K.-H. Han, D. T. Jones, S. Longhi, S. J. Metallo, K. Nishikawa, R. Nussinov, Z. Obradovic, R. V. Pappu, B. Rost, P. Selenko, V.

- Subramaniam, J. L. Sussman, P. Tompa, V. N. Uversky, *Intrinsically Disord. Proteins* **2013**, *1*, e24157.
- [19] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, M. M. Babu, *Chem. Rev.* **2014**, *114*, 6589–6631.
- [20] J. Habchi, P. Tompa, S. Longhi, V. N. Uversky, *Chem. Rev.* **2014**, *114*, 6561–6588.
- [21] S. Contreras-Martos, A. Piai, S. Kosol, M. Varadi, A. Bekesi, P. Lebrun, A. N. Volkov, K. Gevaert, R. Pierattelli, I. C. Felli, P. Tompa, *Sci. Rep.* **2017**, *7*, 1–13.
- [22] R. Mukhopadhyay, J. H. Hoh, *FEBS Lett.* **2001**, *505*, 374–378.
- [23] M. E. Fealey, B. P. Binder, V. N. Uversky, A. Hinderliter, D. D. Thomas, *Biophys. J.* **2018**, *114*, 550–561.
- [24] P. Robustelli, S. Piana, D. E. Shaw, *J. Am. Chem. Soc.* **2020**, *142*, 11092–11101.
- [25] D. Kurzbach, G. Platzer, T. C. Schwarz, M. A. Henen, R. Konrat, D. Hinderberger, *Biochemistry* **2013**, *52*, 5167–5175.
- [26] D. Kurzbach, T. C. Schwarz, G. Platzer, S. Höfler, D. Hinderberger, R. Konrat, *Angew. Chemie - Int. Ed.* **2014**, *53*, 3840–3843.
- [27] M. Arbesú, G. Iruela, H. Fuentes, J. M. C. Teixeira, M. Pons, *Front. Mol. Biosci.* **2018**, *5*, DOI 10.3389/fmolb.2018.00039.
- [28] M. Arbesú, M. Pons, *Arch. Biochem. Biophys.* **2019**, *677*, 108161.
- [29] A. Sottini, A. Borgia, M. B. Borgia, K. Bugge, D. Nettels, A. Chowdhury, P. O. Heidarsson, F. Zosel, R. B. Best, B. B. Kragelund, B. Schuler, *Nat. Commun.* **2020**, *11*, 5736.
- [30] G. W. Daughdrill, M. S. Chadsey, J. E. Karlinsey, K. T. Hughes, F. W. Dahlquist, *Nat. Struct. Biol.* **1997**, *4*, 285–291.
- [31] M. S. Nielsen, H. Vorum, E. Lindersson, P. H. Jensen, *J. Biol. Chem.* **2001**, *276*, 22680–22684.
- [32] A. Binolfi, R. M. Rasia, C. W. Bertocini, M. Ceolin, M. Zweckstetter, C. Griesinger, T. M. Jovin, C. O. Fernández, *J. Am. Chem. Soc.* **2006**, *128*, 9893–9901.

- [33] A. Rcom-H'cheo-Gauthier, J. Goodwin, D. L. Pountney, *Biomolecules* **2014**, *4*, 795–811.
- [34] J. Lautenschläger, A. D. Stephens, G. Fusco, F. Ströhl, N. Curry, M. Zacharopoulou, C. H. Michel, R. Laine, N. Nespovitaya, M. Fantham, D. Pinotsi, W. Zago, P. Fraser, A. Tandon, P. St George-Hyslop, E. Rees, J. J. Phillips, A. De Simone, C. F. Kaminski, G. S. K. Schierle, *Nat. Commun.* **2018**, *9*, 712.
- [35] T. Berg, S. B. Cohen, J. Desharnais, C. Sonderegger, D. J. Maslyar, J. Goldberg, D. L. Boger, P. K. Vogt, *Proc. Natl. Acad. Sci.* **2002**, *99*, 3830–3835.
- [36] L. A. Shevde, R. S. Samant, *Matrix Biol.* **2014**, *37*, 131–141.
- [37] J. D. Mills, M. Janitz, *Neurobiol. Aging* **2012**, *33*, 1011–1024.
- [38] A. Patel, H. O. Lee, L. Jawerth, S. Maharana, M. Jahnel, M. Y. Hein, S. Stoyanov, J. Mahamid, S. Saha, T. M. Franzmann, A. Pozniakovski, I. Poser, N. Maghelli, L. A. Royer, M. Weigert, E. W. Myers, S. Grill, D. Drechsel, A. A. Hyman, S. Alberti, *Cell* **2015**, *162*, 1066–1077.
- [39] F. Theillet, L. Kalmar, P. Tompa, K. Han, P. Selenko, A. K. Dunker, G. W. Daughdrill, V. N. Uversky, *Intrinsically Disord. Proteins* **2013**, *1*, e24360.
- [40] J. Bürgi, B. Xue, V. N. Uversky, F. G. van der Goot, *PLoS One* **2016**, *11*, e0158594.
- [41] A. P. Lieberman, V. G. Shakkottai, R. L. Albin, *Annu. Rev. Pathol. Mech. Dis.* **2019**, *14*, 1–27.
- [42] R. Konrat, *J. Magn. Reson.* **2014**, *241*, 74–85.
- [43] P. Joshi, S. Chia, J. Habchi, T. P. J. Knowles, C. M. Dobson, M. Vendruscolo, *ACS Comb. Sci.* **2016**, *18*, 144–153.
- [44] G. T. Heller, F. A. Aprile, M. Vendruscolo, *Cell. Mol. Life Sci.* **2017**, *74*, 3225–3243.
- [45] G. Fuertes, L. Nevola, S. Esteban-Martín, in *Intrinsically Disord. Proteins*, Elsevier, **2019**, pp. 275–327.
- [46] B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker, V. N. Uversky, *Biochim. Biophys. Acta - Proteins Proteomics* **2010**, *1804*, 996–1010.
- [47] N. Brandes, D. Ofer, M. Linial, *Database* **2016**, *2016*, baw133.
- [48] B. Mészáros, G. Erdős, Z. Dosztányi, *Nucleic Acids Res.* **2018**, *46*, W329–W337.

- [49] M. Kumar, M. Gouw, S. Michael, H. Sámano-Sánchez, R. Pancsa, J. Glavina, A. Diakogianni, J. A. Valverde, D. Bukirova, J. Čalyševa, N. Palopoli, N. E. Davey, L. B. Chemes, T. J. Gibson, *Nucleic Acids Res.* **2019**, *48*, D296–D306.
- [50] M. Necci, D. Piovesan, S. C. E. Tosatto, *Nat. Methods* **2021**, *18*, 472–481.
- [51] T. Lazar, E. Martínez-Pérez, F. Quaglia, A. Hatos, L. B. Chemes, J. A. Iserte, N. A. Méndez, N. A. Garrone, T. E. Saldaño, J. Marchetti, A. J. V. Rueda, P. Bernadó, M. Blackledge, T. N. Cordeiro, E. Fagerberg, J. D. Forman-Kay, M. S. Fornasari, T. J. Gibson, G. N. W. Gomes, C. C. Gradinaru, T. Head-Gordon, M. R. Jensen, E. A. Lemke, S. Longhi, C. Marino-Buslje, G. Minervini, T. Mittag, A. M. Monzon, R. V. Pappu, G. Parisi, S. Ricard-Blum, K. M. Ruff, E. Salladini, M. Skepö, D. Svergun, S. D. Vallet, M. Varadi, P. Tompa, S. C. E. Tosatto, D. Piovesan, *Nucleic Acids Res.* **2021**, *49*, D404–D411.
- [52] P. Tompa, A. Fersht, *Structure and Function of Intrinsically Disordered Proteins*, Chapman And Hall/CRC, **2009**.
- [53] S. Kosol, S. Contreras-Martos, C. Cedeño, P. Tompa, *Molecules* **2013**, *18*, 10802–10828.
- [54] S. Gil, T. Hošek, Z. Solyom, R. Kümmerle, B. Brutscher, R. Pierattelli, I. C. Felli, *Angew. Chemie Int. Ed.* **2013**, *52*, 11808–11812.
- [55] M. R. Jensen, M. Zweckstetter, J. Huang, M. Blackledge, *Chem. Rev.* **2014**, *114*, 6632–6660.
- [56] I. C. Felli, R. Pierattelli, *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*, Springer International Publishing, Cham, **2015**.
- [57] H. Jane Dyson, P. E. Wright, in *Methods Enzymol.*, **2001**, pp. 258–270.
- [58] B. Brutscher, I. C. Felli, S. Gil-Caballero, T. Hošek, R. Kümmerle, A. Piai, R. Pierattelli, Z. Solyom, in *Intrinsically Disord. Proteins Stud. by NMR Spectrosc.*, **2015**, pp. 49–122.
- [59] J. B. Matthew, F. M. Richards, *J. Biol. Chem.* **1983**, *258*, 3039–44.
- [60] X. Tadeo, M. Pons, O. Millet, *Biochemistry* **2007**, *46*, 917–923.
- [61] R. L. Croke, C. O. Sallum, E. Watson, E. D. Watt, A. T. Alexandrescu, *Protein Sci.* **2008**, *17*, 1434–1445.
- [62] H. Geen, R. Freeman, *J. Magn. Reson.* **1991**, *93*, 93–141.
- [63] P. Schanda, Ě. Kupče, B. Brutscher, *J. Biomol. NMR* **2005**, *33*, 199–211.

- [64] M. Czisch, R. Boelens, *J. Magn. Reson.* **1998**, *134*, 158–160.
- [65] B. Brutscher, *Concepts Magn. Reson.* **2000**, *12*, 207–229.
- [66] P. Permi, *J. Biomol. NMR* **2002**, *23*, 201–9.
- [67] S. Mäntylahti, O. Aitio, M. Hellman, P. Permi, *J. Biomol. NMR* **2010**, *47*, 171–181.
- [68] S. Mäntylahti, M. Hellman, P. Permi, *J. Biomol. NMR* **2011**, *49*, 99–109.
- [69] M. Karjalainen, H. Tossavainen, M. Hellman, P. Permi, *J. Biomol. NMR* **2020**, *74*, 741–752.
- [70] P. Würtz, K. Fredriksson, P. Permi, *J. Biomol. NMR* **2005**, *31*, 321–330.
- [71] G. Parigi, N. Rezaei-Ghaleh, A. Giachetti, S. Becker, C. Fernandez, M. Blackledge, C. Griesinger, M. Zweckstetter, C. Luchinat, *J. Am. Chem. Soc.* **2014**, *136*, 16201–16209.
- [72] W. Bermel, I. Bertini, I. C. Felli, L. Gonnelli, W. Koźmiński, A. Piai, R. Pierattelli, J. Stanek, *J. Biomol. NMR* **2012**, *53*, 293–301.
- [73] A. Piai, T. Hošek, L. Gonnelli, A. Zawadzka-Kazimierczuk, W. Koźmiński, B. Brutscher, W. Bermel, R. Pierattelli, I. C. Felli, *J. Biomol. NMR* **2014**, *60*, 209–218.
- [74] K. Takeuchi, G. Heffron, Z. Y. J. Sun, D. P. Frueh, G. Wagner, *J. Biomol. NMR* **2010**, *47*, 271–282.
- [75] M. Gal, K. A. Edmonds, A. G. Milbradt, K. Takeuchi, G. Wagner, K. A. Kaphingst, S. Persky, C. Lachance, *J. Biomol. NMR* **2011**, *51*, 497–504.
- [76] K. Takeuchi, H. Arthanari, M. Imai, G. Wagner, I. Shimada, *J. Biomol. NMR* **2016**, *64*, 143–151.
- [77] E. B. Gibbs, R. W. Kriwacki, *Methods* **2018**, *2*, 2–9.
- [78] W. Bermel, I. Bertini, I. Felli, M. Piccioli, R. Pierattelli, *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *48*, 25–45.
- [79] W. Bermel, I. Bertini, I. C. Felli, R. Kümmerle, R. Pierattelli, *J. Magn. Reson.* **2006**, *178*, 56–64.
- [80] G. Pasat, J. S. Zintsmaster, J. W. Peng, *J. Magn. Reson.* **2008**, *193*, 226–232.
- [81] W. Bermel, I. Bertini, V. Csizmok, I. C. Felli, R. Pierattelli, P. Tompa, *J. Magn. Reson.* **2009**, *198*, 275–281.

- [82] J. Nováček, A. Zawadzka-Kazimierczuk, V. Papoušková, L. Žídek, H. Šanderová, L. Krásný, W. Koźmiński, V. Sklenář, *J. Biomol. NMR* **2011**, *50*, 1–11.
- [83] I. C. Felli, L. Gonnelli, R. Pierattelli, *Nat. Protoc.* **2014**, *9*, 2005–2016.
- [84] W. Bermel, I. Bertini, I. C. Felli, Y. M. Lee, C. Luchinat, R. Pierattelli, *J. Am. Chem. Soc.* **2006**, *128*, 3918–3919.
- [85] W. Bermel, I. C. Felli, L. Gonnelli, W. Koźmiński, A. Piai, R. Pierattelli, A. Zawadzka-Kazimierczuk, *J. Biomol. NMR* **2013**, *57*, 353–361.
- [86] M. D. Sørensen, A. Meissner, O. W. Sørensen, *J. Biomol. NMR* **1997**, *10*, 181–186.
- [87] W. Bermel, I. C. Felli, R. Kümmerle, R. Pierattelli, *Concepts Magn. Reson.* **2008**, *32*, 183–200.
- [88] I. C. Felli, R. Pierattelli, *Prog. Nucl. Magn. Reson. Spectrosc.* **2015**, *84–85*, 1–13.
- [89] G. Karunanithy, D. F. Hansen, *J. Biomol. NMR* **2021**, *75*, 179–191.
- [90] M. G. Murrall, M. Schiavina, V. Sainati, W. Bermel, R. Pierattelli, I. C. Felli, *J. Biomol. NMR* **2018**, *70*, 1–9.
- [91] E. Kupče, R. Freeman, *J. Am. Chem. Soc.* **2004**, *126*, 6429–6440.
- [92] F. Fiorito, S. Hiller, G. Wider, K. Wüthrich, *J. Biomol. NMR* **2006**, *35*, 27–37.
- [93] I. C. Felli, B. Brutscher, *ChemPhysChem* **2009**, *10*, 1356–1368.
- [94] K. Kazimierczuk, J. Stanek, A. Zawadzka-Kazimierczuk, W. Koźmiński, *Prog. Nucl. Magn. Reson. Spectrosc.* **2010**, *57*, 420–434.
- [95] K. Kazimierczuk, J. Stanek, A. Zawadzka-Kazimierczuk, W. Koźmiński, *ChemPhysChem* **2013**, *14*, 3015–3025.
- [96] K. Grudziąż, A. Zawadzka-Kazimierczuk, W. Koźmiński, *Methods* **2018**, 81–87.
- [97] M. Mobli, M. W. Maciejewski, A. D. Schuyler, A. S. Stern, J. C. Hoch, *Phys. Chem. Chem. Phys.* **2012**, *14*, 10835–10843.
- [98] S. Hiller, G. Wider, *Top. Curr. Chem.* **2012**, *316*, 21–48.
- [99] B. Krähenbühl, J. Boudet, G. Wider, *J. Biomol. NMR* **2013**, *56*, 149–154.
- [100] S. K. Dutta, P. Serrano, A. Proudfoot, M. Geralt, B. Pedrini, T. Herrmann, K. Wüthrich, *J.*

Biomol. NMR **2015**, *61*, 47–53.

- [101] S. Xu, S. Ni, M. A. Kennedy, *Biophys. J.* **2017**, *112*, 2075–2088.
- [102] A. HVIDT, *C. R. Trav. Lab. Carlsberg* **1964**, *34*, 299–317.
- [103] R. S. Molday, S. W. Englander, R. G. Kallen, *Biochemistry* **1972**, *11*, 150–158.
- [104] Y. Bai, J. S. Milne, L. Mayne, S. W. Englander, *Proteins Struct. Funct. Genet.* **1993**, *17*, 75–86.
- [105] T.-L. Hwang, S. Mori, A. J. Shaka, P. C. M. van Zijl, *J. Am. Chem. Soc.* **1997**, *119*, 6203–6204.
- [106] T.-L. Hwang, P. C. M. van Zijl, S. Mori, *J. Biomol. NMR* **1998**, *11*, 221–226.
- [107] F. Kateb, P. Pelupessy, G. Bodenhausen, *J. Magn. Reson.* **2007**, *184*, 108–113.
- [108] A. A. Sehgal, L. Duma, G. Bodenhausen, P. Pelupessy, *Chem. – A Eur. J.* **2014**, *20*, 6332–6338.
- [109] N. R. Skrynnikov, R. R. Ernst, *J. Magn. Reson.* **1999**, *137*, 276–280.
- [110] L. Pontoriero, M. Schiavina, M. G. Murralli, R. Pierattelli, I. C. Felli, *Angew. Chemie Int. Ed.* **2020**, *59*, 18537–18545.
- [111] Ě. Kupče, B. Wrackmeyer, *Appl. Organomet. Chem.* **2010**, *24*, 837–841.
- [112] E. R. Kupče, R. Freeman, *J. Magn. Reson.* **2011**, *213*, 1–13.
- [113] E. Kupče, L. E. Kay, *J. Biomol. NMR* **2012**, *54*, 1–7.
- [114] K. J. Donovan, E. Kupče, L. Frydman, *Angew. Chemie - Int. Ed.* **2013**, *52*, 4152–4155.
- [115] P. Gierth, A. Codina, F. Schumann, H. Kovacs, E. Kupče, *Magn. Reson. Chem.* **2015**, *53*, 940–944.
- [116] E. Kupce, *Mod. NMR Methodol.* **2015**, *4*, 721–731.
- [117] A. Viegas, T. Viennet, T. Yu, F. Schumann, *J. Biomol. NMR* **2016**, *64*, 9–15.
- [118] Ě. Kupče, T. D. W. Claridge, *Angew. Chemie Int. Ed.* **2017**, *56*, 11779–11783.
- [119] Ě. Kupče, K. R. Mote, P. K. Madhu, *J. Magn. Reson.* **2019**, *304*, 16–34.
- [120] M. Schiavina, M. G. Murralli, L. Pontoriero, V. Sainati, R. Kümmerle, W. Bermel, R.

- Pierattelli, I. C. Felli, *Biophys. J.* **2019**, *117*, 46–55.
- [121] M. Pons, *Biophys. J.* **2019**, *117*, 1–2.
- [122] E. Kupče, L. E. Kay, R. Freeman, *J. Am. Chem. Soc.* **2010**, *132*, 18008–18011.
- [123] L. E. Kay, R. Freeman, *J. Am. Chem. Soc.* **2010**, *2*, 3–6.
- [124] E. Luchinat, L. Banci, *IUCrJ* **2017**, *4*, 108–118.
- [125] K. Takeuchi, H. Arthanari, I. Shimada, G. Wagner, *J. Biomol. NMR* **2015**, *63*, 323–331.
- [126] G. Tóth, S. J. Gardai, W. Zago, C. W. Bertoncini, N. Cremades, S. L. Roy, M. A. Tambe, J.-C. Rochet, C. Galvagnion, G. Skibinski, S. Finkbeiner, M. Bova, K. Regnstrom, S.-S. Chiou, J. Johnston, K. Callaway, J. P. Anderson, M. F. Jobling, A. K. Buell, T. A. Yednock, T. P. J. Knowles, M. Vendruscolo, J. Christodoulou, C. M. Dobson, D. Schenk, L. McConlogue, *PLoS One* **2014**, *9*, e87133.
- [127] S. Ambadipudi, M. Zweckstetter, *Expert Opin. Drug Discov.* **2016**, *11*, 65–77.
- [128] K. Tsafou, P. B. Tiwari, J. D. Forman-Kay, S. J. Metallo, J. A. Toretzky, *J. Mol. Biol.* **2018**, *430*, 2321–2341.
- [129] G. T. Heller, P. Sormanni, M. Vendruscolo, *Trends Biochem. Sci.* **2015**, *40*, 491–496.
- [130] M. J. Daniels, J. B. Nourse, H. Kim, V. Sainati, M. Schiavina, M. G. Murralli, B. Pan, J. J. Ferrie, C. M. Haney, R. Moons, N. S. Gould, A. Natalello, R. Grandori, F. Sobott, E. J. Petersson, E. Rhoades, R. Pierattelli, I. Felli, V. N. Uversky, K. A. Caldwell, G. A. Caldwell, E. S. Krol, H. Ischiropoulos, *Sci. Rep.* **2019**, *9*, 1–17.
- [131] J. Li, M. Zhu, S. Rajamani, V. N. Uversky, A. L. Fink, *Chem. Biol.* **2004**, *11*, 1513–1521.
- [132] K. Ono, M. Yamada, *J. Neurochem.* **2006**, *97*, 105–115.
- [133] K. M. Danzer, D. Haasen, A. R. Karow, S. Moussaud, M. Habeck, A. Giese, H. Kretzschmar, B. Hengerer, M. Kostka, *J. Neurosci.* **2007**, *27*, 9220–9232.
- [134] X. Meng, L. A. Munishkina, A. L. Fink, V. N. Uversky, *Biochemistry* **2009**, *48*, 8206–8224.
- [135] J. Wagner, S. Ryazanov, A. Leonov, J. Levin, S. Shi, F. Schmidt, C. Prix, F. Pan-Montojo, U. Bertsch, G. Mitteregger-Kretzschmar, M. Geissen, M. Eiden, F. Leidel, T. Hirschberger, A. A. Deeg, J. J. Krauth, W. Zinth, P. Tavan, J. Pilger, M. Zweckstetter, T. Frank, M. Bähr, J. H. Weishaupt, M. Uhr, H. Urlaub, U. Teichmann, M. Samwer, K. Bötzel, M. Groschup, H.

- Kretzschmar, C. Griesinger, A. Giese, *Acta Neuropathol.* **2013**, *125*, 795–813.
- [136] L. Fernandes, N. Moraes, F. S. Sagrillo, A. V. Magalhães, M. C. de Moraes, L. Romão, J. W. Kelly, D. Foguel, N. P. Grimster, F. L. Palhano, *ACS Chem. Neurosci.* **2017**, *8*, 1704–1712.
- [137] K. A. Conway, *Science (80-.)*. **2001**, *294*, 1346–1349.
- [138] R. Llinás, M. Sugimori, R. B. Silver, *Science (80-.)*. **1992**, *256*, 677–679.
- [139] R. Schneggenburger, E. Neher, *Nature* **2000**, *406*, 889–993.
- [140] A. D. Stephens, M. Zacharopoulou, G. S. Kaminski Schierle, *Trends Biochem. Sci.* **2019**, *44*, 453–466.
- [141] W. Bermel, I. Bertini, L. Duma, I. C. Felli, L. Emsley, R. Pierattelli, P. R. Vasos, *Angew. Chemie - Int. Ed.* **2005**, *44*, 3089–3092.
- [142] B. Mateos, C. Conrad-Billroth, M. Schiavina, A. Beier, G. Kontaxis, R. Konrat, I. C. Felli, R. Pierattelli, *J. Mol. Biol.* **2020**, *432*, 3093–3111.
- [143] C. C. Kazanecki, D. J. Uzwiak, D. T. Denhardt, *J. Cell. Biochem.* **2007**, *102*, 912–924.
- [144] P. H. Anborgh, J. C. Mutrie, A. B. Tuck, A. F. Chambers, *J. Cell. Mol. Med.* **2010**, *14*, 2037–2044.
- [145] A. Pietras, A. M. Katz, E. J. Ekström, B. Wee, J. J. Halliday, K. L. Pitter, J. L. Werbeck, N. M. Amankulor, J. T. Huse, E. C. Holland, *Cell Stem Cell* **2014**, *14*, 357–369.
- [146] D. Kurzbach, T. C. Schwarz, G. Platzer, S. Höfler, D. Hinderberger, R. Konrat, *Angew. Chemie Int. Ed.* **2014**, *53*, 3840–3843.
- [147] T. R. Alderson, J. H. Lee, C. Charlier, J. Ying, A. Bax, *ChemBioChem* **2018**, *19*, 37–42.
- [148] M. G. Murrall, A. Piai, W. Bermel, I. C. Felli, R. Pierattelli, *ChemBioChem* **2018**, *19*, 1625–1629.
- [149] W. Bermel, I. Bertini, I. C. Felli, R. Peruzzini, R. Pierattelli, *ChemPhysChem* **2010**, *11*, 689–695.
- [150] I. C. Felli, R. Pierattelli, S. J. Glaser, B. Luy, *J. Biomol. NMR* **2009**, *43*, 187–196.
- [151] B. Mateos, R. Konrat, R. Pierattelli, I. C. Felli, *ChemBioChem* **2019**, *20*, 335–339.
- [152] C. Camilloni, A. De Simone, W. F. Vranken, M. Vendruscolo, *Biochemistry* **2012**, *51*, 2224–2231.

- [153] F. Nardi, J. Kemmink, M. Sattler, R. C. Wade, *J. Biomol. NMR* **2000**, *17*, 63–77.
- [154] K. M. Thomas, D. Naduthambi, N. J. Zondlo, *J. Am. Chem. Soc.* **2006**, *128*, 2216–2217.
- [155] N. J. Zondlo, *Acc. Chem. Res.* **2013**, *46*, 1039–1049.
- [156] M. Schiavina, E. Salladini, M. G. Murrari, G. Tria, I. C. Felli, R. Pierattelli, S. Longhi, *Sci. Rep.* **2020**, *10*, 19574.
- [157] M. Schiavina, L. Pontoriero, V. N. Uversky, I. C. Felli, R. Pierattelli, *Biomol. NMR Assign.* **2021**, *15*, 219–227.
- [158] L.-F. Wang, M. Yu, E. Hansson, L. I. Pritchard, B. Shiell, W. P. Michalski, B. T. Eaton, *J. Virol.* **2000**, *74*, 9972–9979.
- [159] C. Bracken, A. G. Palmer, J. Cavanagh, *J. Biomol. NMR* **1997**, *9*, 94–100.
- [160] Z.-Y. J. Sun, D. P. Frueh, P. Selenko, J. C. Hoch, G. Wagner, *J. Biomol. NMR* **2005**, *33*, 43–50.
- [161] K. Tamiola, F. A. A. Mulder, *Biochem. Soc. Trans.* **2012**, *40*, 1014–1020.
- [162] K. Tamiola, B. Acar, F. A. A. Mulder, *J. Am. Chem. Soc.* **2010**, *132*, 18000–18003.
- [163] P. S. Masters, *Adv. Virus Res.* **2006**, *65*, 193–292.
- [164] M. Surjit, S. K. Lal, *Infect. Genet. Evol.* **2008**, *8*, 397–405.
- [165] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, B. Cao, *Lancet* **2020**, *395*, 497–506.
- [166] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, Z. Peng, *JAMA* **2020**, *323*, 1061.
- [167] R. Rangan, I. N. Zheludev, R. J. Hagey, E. A. Pham, H. K. Wayment-Steele, J. S. Glenn, R. Das, *RNA* **2020**, *26*, 937–959.
- [168] Z. Bai, Y. Cao, W. Liu, J. Li, *Viruses* **2021**, *13*, 1115.
- [169] C. Cao, Z. Cai, X. Xiao, J. Rao, J. Chen, N. Hu, M. Yang, X. Xing, Y. Wang, M. Li, B. Zhou, X. Wang, J. Wang, Y. Xue, *Nat. Commun.* **2021**, *12*, 3917.
- [170] R. de C. A. Tavares, G. Mahadeshwar, H. Wan, N. C. Huston, A. M. Pyle, *J. Virol.* **2021**, *95*, DOI 10.1128/JVI.02190-20.

- [171] R. Giri, T. Bhardwaj, M. Shegane, B. R. Gehi, P. Kumar, K. Gadhave, C. J. Oldfield, V. N. Uversky, *Cell. Mol. Life Sci.* **2021**, *78*, 1655–1688.
- [172] S. Kang, M. Yang, Z. Hong, L. Zhang, Z. Huang, X. Chen, S. He, Z. Z. Zhou, Q. Chen, Y. Yan, C. Zhang, H. Shan, S. Chen, *Acta Pharm. Sin. B* **2020**, *10*, 1228–1238.
- [173] Y. Peng, N. Du, Y. Lei, S. Dorje, J. Qi, T. Luo, G. F. Gao, H. Song, *EMBO J.* **2020**, *39*, DOI 10.15252/embj.2020105938.
- [174] D. C. Dinesh, D. Chalupska, J. Silhan, E. Koutna, R. Nencka, V. Veverka, E. Boura, *PLOS Pathog.* **2020**, *16*, e1009100.
- [175] C. K. Chang, M. H. Hou, C. F. Chang, C. D. Hsiao, T. H. Huang, *Antiviral Res.* **2014**, *103*, 39–50.
- [176] E. Lacroix, A. R. Viguera, L. Serrano, *J. Mol. Biol.* **1998**, *284*, 173–191.

5. Chemical shift assignments

The following NMR chemical shift assignments have been deposited in the BMRB.

BMRB entry	Protein	Experimental conditions
27348	α -synuclein	20.0 mM NaPi, 200.0 mM NaCl, 0.5 mM EDTA, pH 6.5, 315.0 K
50370	NiV PNT	10.0 mM NaPi, 5 mM EDTA, 5 mM DTT, pH 6.5, 288.0 K
50618	NTR (176-248)	25.0 mM TRIS, 450.0 mM NaCl, 0.02% NaN ₃ pH 6.5, 298.0 K
50619	NTR (1-47)	25.0 mM TRIS, 450.0 mM NaCl, 0.02% NaN ₃ pH 6.5, 298.0 K

6. Cyclized NDGA modifies dynamic α -synuclein monomers preventing aggregation and toxicity

SCIENTIFIC REPORTS

OPEN

Cyclized NDGA modifies dynamic α -synuclein monomers preventing aggregation and toxicity

Received: 22 October 2018
Accepted: 4 January 2019
Published online: 27 February 2019

Malcolm J. Daniels¹, J. Brucker Nourse Jr.², Hanna Kim², Valerio Sainati³, Marco Schiavina³, Maria Grazia Murralli³, Buyan Pan⁴, John J. Ferrie⁴, Conor M. Haney⁴, Rani Moons⁵, Neal S. Gould⁶, Antonino Natalello⁷, Rita Grandori⁷, Frank Sobott^{8,9,10}, E. James Petersson⁴, Elizabeth Rhoades⁴, Roberta Pierattelli³, Isabella Felli³, Vladimir N. Uversky^{11,12}, Kim A. Caldwell², Guy A. Caldwell², Edward S. Krol¹³ & Harry Ischiropoulos^{1,6,14}

Growing evidence implicates α -synuclein aggregation as a key driver of neurodegeneration in Parkinson's disease (PD) and other neurodegenerative disorders. Herein, the molecular and structural mechanisms of inhibiting α -synuclein aggregation by novel analogs of nordihydroguaiaretic acid (NDGA), a phenolic dibenzenediol lignan, were explored using an array of biochemical and biophysical methodologies. NDGA analogs induced modest, progressive compaction of monomeric α -synuclein, preventing aggregation into amyloid-like fibrils. This conformational remodeling preserved the dynamic adoption of α -helical conformations, which are essential for physiological membrane interactions. Oxidation-dependent NDGA cyclization was required for the interaction with monomeric α -synuclein. NDGA analog-pretreated α -synuclein did not aggregate even without NDGA-analogs in the aggregation mixture. Strikingly, NDGA-pretreated α -synuclein suppressed aggregation of naïve untreated aggregation-competent monomeric α -synuclein. Further, cyclized NDGA reduced α -synuclein-driven neurodegeneration in *Caenorhabditis elegans*. The cyclized NDGA analogs may serve as a platform for the development of small molecules that stabilize aggregation-resistant α -synuclein monomers without interfering with functional conformations yielding potential therapies for PD and related disorders.

SCIENTIFIC REPORTS

OPEN

Cyclized NDGA modifies dynamic α -synuclein monomers preventing aggregation and toxicity

Received: 22 October 2018
Accepted: 4 January 2019
Published online: 27 February 2019

Malcolm J. Daniels¹, J. Brucker Nourse Jr.², Hanna Kim², Valerio Sainati³, Marco Schiavina³, Maria Grazia Murrall³, Buyan Pan⁴, John J. Ferrie⁴, Conor M. Haney⁴, Rani Moons⁵, Neal S. Gould⁶, Antonino Natalello⁷, Rita Grandori⁷, Frank Sobott^{8,9,10}, E. James Petersson⁴, Elizabeth Rhoades⁴, Roberta Pierattelli¹¹, Isabella Felli¹¹, Vladimir N. Uversky^{11,12}, Kim A. Caldwell², Guy A. Caldwell², Edward S. Krol¹³ & Harry Ischiropoulos^{1,6,14}

Growing evidence implicates α -synuclein aggregation as a key driver of neurodegeneration in Parkinson's disease (PD) and other neurodegenerative disorders. Herein, the molecular and structural mechanisms of inhibiting α -synuclein aggregation by novel analogs of nordihydroguaiaretic acid (NDGA), a phenolic dibenzenediol lignan, were explored using an array of biochemical and biophysical methodologies. NDGA analogs induced modest, progressive compaction of monomeric α -synuclein, preventing aggregation into amyloid-like fibrils. This conformational remodeling preserved the dynamic adoption of α -helical conformations, which are essential for physiological membrane interactions. Oxidation-dependent NDGA cyclization was required for the interaction with monomeric α -synuclein. NDGA analog-pretreated α -synuclein did not aggregate even without NDGA-analogs in the aggregation mixture. Strikingly, NDGA-pretreated α -synuclein suppressed aggregation of naive untreated aggregation-competent monomeric α -synuclein. Further, cyclized NDGA reduced α -synuclein-driven neurodegeneration in *Caenorhabditis elegans*. The cyclized NDGA analogs may serve as a platform for the development of small molecules that stabilize aggregation-resistant α -synuclein monomers without interfering with functional conformations yielding potential therapies for PD and related disorders.

Parkinson's disease (PD) is an age-related neurodegenerative disorder characterized by a progressive motor phenotype including tremors, rigidity, and bradykinesia. These symptoms are driven primarily by loss of dopamine-producing neurons in the *substantia nigra pars compacta*. Lewy bodies, intracellular proteinaceous inclusions, are the histopathological hallmark of PD. Immunohistochemical analysis of Lewy bodies revealed aggregated forms of α -synuclein, a 140 amino acid protein, as a major component^{1,2}. Mutations, duplications, and triplications of the gene encoding α -synuclein cause dominantly-inherited familial forms of PD³⁻⁶. Furthermore, aggregation of wild type α -synuclein is involved in the pathogenesis of a diverse group of neurodegenerative diseases^{1,7-14}.

¹Pharmacology Graduate Group, Raymond and Ruth Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA. ²Department of Biological Sciences, University of Alabama, Tuscaloosa, AL, 35487, USA. ³CERM and Department of Chemistry "Ugo Schiff", University of Florence, Sesto Fiorentino, Florence, 50019, Italy. ⁴Department of Chemistry, University of Pennsylvania, Philadelphia, PA, 19104, USA. ⁵Department of Chemistry, University of Antwerp, Antwerp, Belgium. ⁶Department of Pediatrics, Children's Hospital of Philadelphia Research Institute, Philadelphia, PA, 19104, USA. ⁷Department of Biotechnology and Biosciences, University of Milan-Bicocca, Milan, Italy. ⁸Biomolecular & Analytical Mass Spectrometry, Antwerp University, Antwerp, Belgium. ⁹Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds, United Kingdom. ¹⁰School of Molecular and Cellular Biology, University of Leeds, Leeds, United Kingdom. ¹¹Department of Molecular Medicine and Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, 33612, USA. ¹²Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, 142292, Russian Federation. ¹³College of Pharmacy & Nutrition, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. ¹⁴Children's Hospital of Philadelphia Research Institute and Systems Pharmacology and Translational Therapeutics, the Raymond and Ruth Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA. Correspondence and requests for materials should be addressed to H.I. (email: ischirop@pennmedicine.upenn.edu)

The pathological and genetic evidence implicating α -synuclein in neurodegeneration has sparked numerous studies of the normal function of α -synuclein and its role in disease pathogenesis. α -Synuclein has been implicated in synaptic function¹⁵. It preferentially binds lipid membranes with high curvature¹⁶ and may modulate neurotransmitter release and synaptic function by affecting vesicular dynamics^{17–21}.

Repeated studies across various animal models have shown that expression of mutant α -synuclein with altered aggregation kinetics causes neurodegeneration^{22,23}. Likewise, induction of α -synuclein aggregation in wildtype animals by seeding with α -synuclein aggregates—isolated from PD Lewy bodies²⁴ or generated *in vitro*^{25–28}—induces progressive neurodegeneration. Indeed, this propagation of α -synuclein aggregation is not only observed in animal models, but in dopaminergic tissue grafted into PD patients²⁹. α -Synuclein aggregation also occurs in animal models of dopamine neuron degeneration induced by oxidative chemical insult^{30,31}.

The implication of α -synuclein in PD pathogenesis stimulated several screens for small molecules that alter its aggregation. Dopamine and related catecholamines were among the first molecules found to prevent α -synuclein fibril formation³². Oxidation of vicinal hydroxyls in dopamine, and related phenols, induces formation of soluble α -synuclein oligomers that do not incorporate into fibrils^{33–35}. Recently, dopamine-induced α -synuclein oligomers were shown to contribute to neurodegeneration in mice³⁶, but the precise mechanism of oligomer toxicity remains a subject of debate³⁷. Dopamine oxidation may also contribute to neurodegeneration through modification of glucocerebrosidase and consequential lysosomal dysregulation³⁸.

Ongoing research has identified many other inhibitors of α -synuclein aggregation. Many of these inhibitors stabilize multimers or oligomers through direct interaction with α -synuclein (e.g. various phenols, catechols, and flavonoids^{35,39–45}, Anle138b^{46,47}, rifampicin⁴⁸). A well-studied member of this group is epigallocatechin gallate (EGCG), a polyphenol sharing the vicinal hydroxyls implicated in dopamine's interaction with α -synuclein. Despite their chemical similarities, dopamine and EGCG have divergent effects on the structure of α -synuclein⁴⁹. In fact, while dopamine may induce toxic oligomers, adding EGCG during cell-free α -synuclein aggregation produces species less toxic to cells³⁹. Other aggregation inhibitors alter α -synuclein-protein or -lipid interactions to prevent aggregation (e.g. NPT100-18A⁵⁰, squalamine⁵¹, PcTS^{52,53}, Hsp70⁵⁴). Recent studies have examined small molecules that directly stabilize α -synuclein monomers (e.g. BIOD303⁵⁵, nortriptiline⁵⁶, CLR01^{57,58}). However, it remains unknown whether these small molecules perturb α -synuclein's lipid interactions, which are directly implicated in its role in neurotransmitter release^{59–61}.

In this study we employed nordihydroguaiaretic acid (NDGA), a phenolic dibenzenediol lignan that inhibits α -synuclein aggregation, to examine the mechanisms of α -synuclein aggregation inhibitors and evaluate its potential utility as a chemical platform for the development of novel aggregation inhibitors^{40,62}. Recent work has exhaustively characterized the products and kinetics of NDGA oxidation as well as several novel analogs NDGA and several novel analogs⁶³. We provide evidence that cyclized NDGA, formed during oxidation, interacts with α -synuclein to produce modified monomers. We demonstrate that these aggregation-resistant monomers retain their capacity to interact with phospholipid membranes and that they inhibit aggregation of untreated α -synuclein. Further, we demonstrate for the first time that cyclized NDGA reduces α -synuclein-driven neurodegeneration in a relevant animal model. Generating α -synuclein species that retain native structural dynamics while exerting a dominant-negative effect on α -synuclein aggregation and neurotoxicity represents a new paradigm for intervening in PD and related disorders.

Results

NDGA inhibits both primary and secondary nucleation-mediated α -synuclein aggregation.

Two established assays, measurement of insoluble protein and Thioflavin-T binding, quantified the effect of NDGA on the aggregation of monomeric α -synuclein. NDGA caused concentration-dependent inhibition of insoluble α -synuclein accumulation and amyloid-like aggregate formation (Figs 1A,B and S1). Maximal inhibition was achieved at stoichiometric, equimolar concentrations of NDGA to monomeric α -synuclein. To confirm the Thioflavin-T findings, the secondary structure of aggregates was quantified by circular dichroism (CD). NDGA caused a dose-dependent decrease in the β -sheet content of the α -synuclein aggregates, as indicated by molar ellipticity decreases at 200 nm and increases at 220 nm (Fig. 1C). Finally, transmission electron microscopy was used to image the α -synuclein aggregates produced in the presence or absence of NDGA. Typical α -synuclein fibrils were observed in the absence of NDGA, but not after aggregation with equimolar NDGA (Fig. 1D). These data demonstrate that stoichiometric concentrations of NDGA inhibit the formation of amyloid-like α -synuclein fibrils.

The preceding experiments examined the effect of NDGA on a typical α -synuclein aggregation assay, in which primary nucleation is the rate-limiting step for the formation of fibrils. Aggregation of α -synuclein can also proceed by a secondary nucleation surface-catalyzed formation of fibrils. To test the effect of NDGA on secondary nucleation-dependent aggregation, α -synuclein fibrils (representing 5% of total α -synuclein) were added with monomeric α -synuclein. Equimolar NDGA (molecule to monomeric α -synuclein) inhibited the formation of insoluble α -synuclein aggregates in the presence of fibril seeds (Fig. 1E). Together, these findings demonstrate that stoichiometric NDGA prevents both primary and secondary nucleation dependent aggregation of α -synuclein.

Interaction with α -synuclein requires oxidation and cyclization of NDGA.

We used N-acetylcysteine (NAC), an electron donor, to inhibit NDGA oxidation in the presence of α -synuclein. NAC was included in a typical α -synuclein aggregation mixture at 20:1 molar excess to NDGA. After aggregation for 3 days, levels of α -synuclein aggregation were measured by the established solubility assay. mNDGA, a NDGA analog incapable of oxidation and cyclization was included as a negative control (Supplementary Fig. S2). The addition of NAC allowed α -synuclein to aggregate in the presence of NDGA, suggesting that oxidation is necessary for NDGA to inhibit aggregation (Fig. 2A,B). The inclusion of NAC also reduced levels of quinone-containing

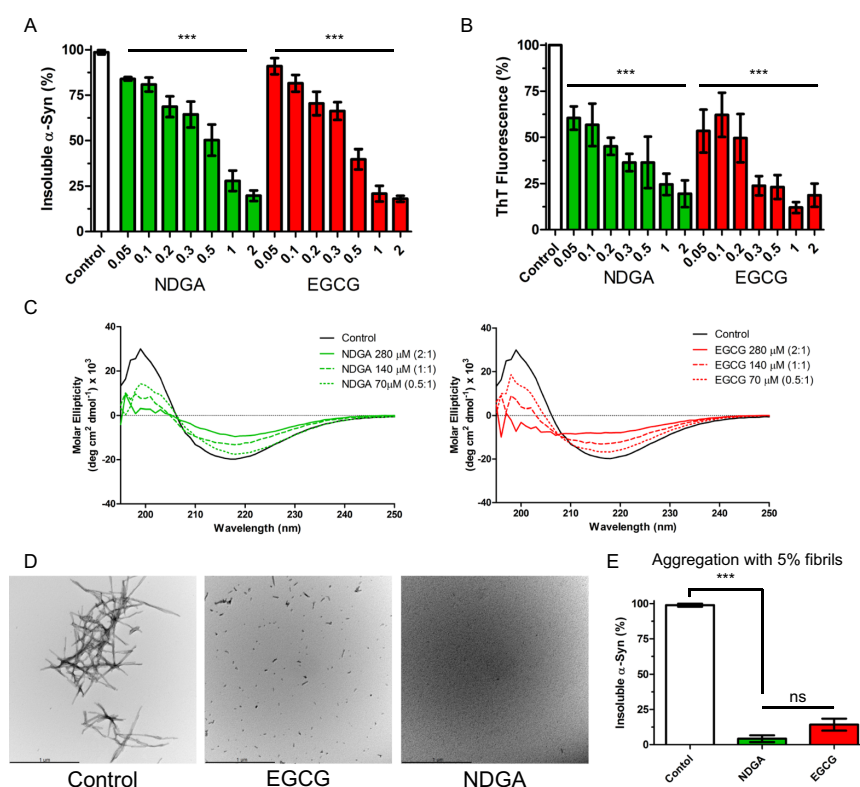


Figure 1. NDGA inhibits recombinant human α -synuclein aggregation. (A) Insoluble α -synuclein present after 7 days aggregation was reduced by NDGA and EGCG in a dose-dependent fashion as compared to solvent control. Recombinant human wildtype α -synuclein (138 μ M) was aggregated for 7 days in the presence of EGCG or NDGA at the indicated molar ratios. After aggregation, PBS-insoluble α -synuclein was separated by centrifugation (21k g for 10 min). Soluble and insoluble fractions were boiled in SDS, run by SDS-PAGE, and colloidal stained. α -Synuclein in each fraction was quantified by in-gel densitometry ($n = 3-5$, $***p < 0.001$). (B) Formation of amyloid α -synuclein fibrils quantified by Thioflavin-T was reduced by NDGA and EGCG in a dose-dependent fashion as compared to control after 7 days aggregation ($n = 3-5$, $***p < 0.001$). (C) α -Synuclein beta-sheet secondary structure was reduced by EGCG and NDGA in a dose dependent fashion. Recombinant human wildtype α -synuclein (138 μ M) was aggregated for 7 days in the presence of EGCG or NDGA at the indicated molar ratios. Secondary structure was quantified by circular dichroism. (D) Transmission electron microscopy images of α -synuclein aggregates after 3 days aggregation with small molecules at 1:1 molar ratio. (E) Insoluble α -synuclein present after 7 days aggregation in the presence of a 5% fibril seed was reduced by EGCG and NDGA. EGCG or NDGA were present at a 1:1 molar ratio. After aggregation, PBS-insoluble α -synuclein was separated by centrifugation (21k g for 10 min). Soluble and insoluble fractions were boiled in SDS, run by SDS-PAGE, and colloidal stained. α -Synuclein in each fraction was quantified by in-gel densitometry ($n = 3$, $***p < 0.001$).

NDGA oxidation products detected by near-infrared fluorescence (nIRF), a technique developed to study the oxidation-dependent interaction between α -synuclein and dopamine⁶⁴ (Fig. 2C). The correlation of increased aggregation and decreased quinone nIRF suggests a relationship between NDGA oxidation and inhibition of α -synuclein aggregation. Separately, catalase, which rapidly converts H_2O_2 to water and hydrogen gas, was included in the aggregation assay to determine whether H_2O_2 that might be produced during NDGA oxidation mediated the effects on aggregation. Inclusion of catalase had no effect on α -synuclein aggregation in the presence of NDGA nor on quinone nIRF, indicating that the effect is not H_2O_2 -dependent (Figs 2A and S3). These data indicate that inhibition of α -synuclein aggregation by NDGA and formation of quinone-modified α -synuclein is oxidation-dependent but not mediated by peroxide chemistry.

Previous work has identified the majority product of NDGA oxidation as a cyclized form⁶³. During that study, novel NDGA analogs with defined oxidation pathways were generated. Here, these analogs were used to examine the role of cyclization in NDGA's oxidation-dependent inhibition of α -synuclein aggregation. We employed NDGA and two such analogs: NDGA-1, which oxidizes with comparable kinetics but does not cyclize due to unilateral hydroxyl substitution, and NDGA-5, which cyclizes more readily than NDGA due to reduced steric

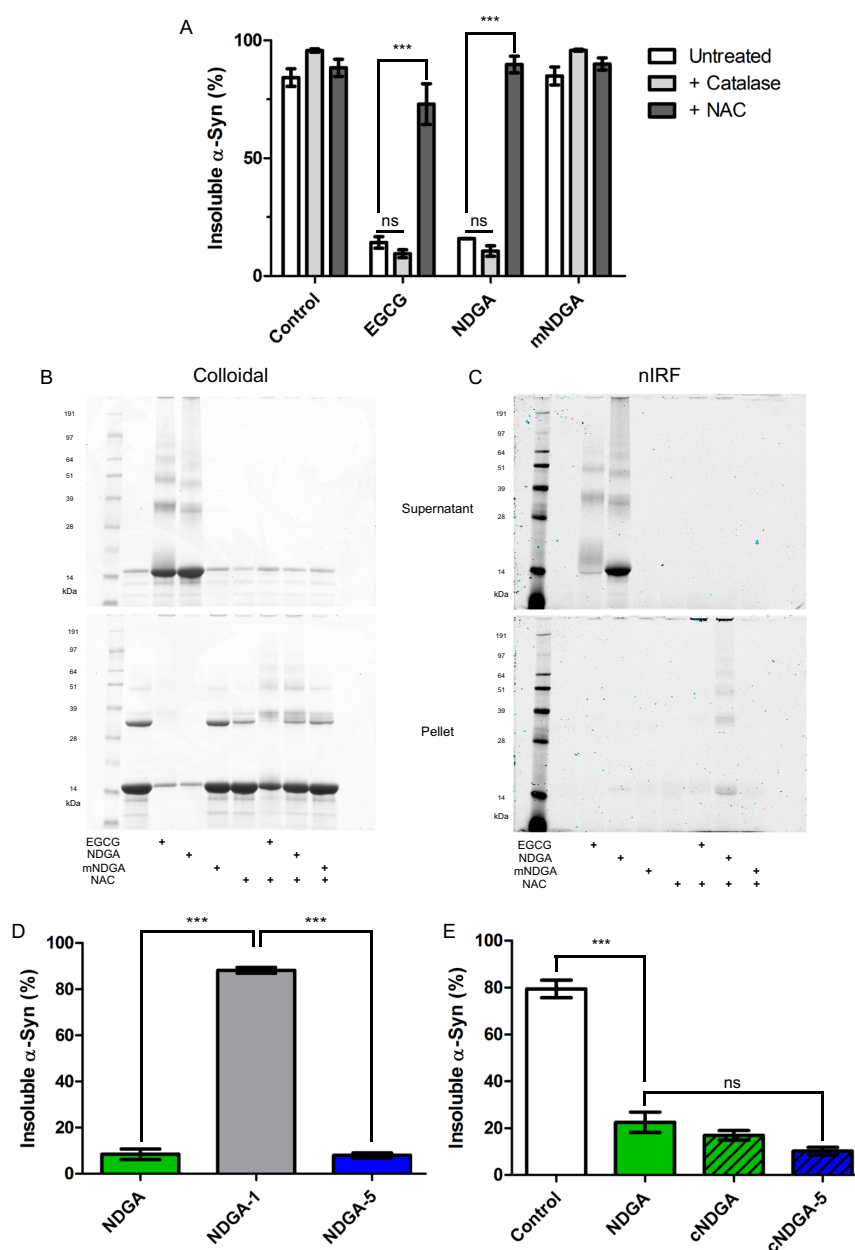


Figure 2. Interaction between NDGA and α -synuclein requires NDGA oxidation and cyclization. **(A)** NDGA treatment did not affect α -synuclein aggregation in the presence of N-acetylcysteine (NAC). α -Synuclein was aggregated for 3 days in the presence of 1:1 small molecules. After aggregation, PBS-insoluble α -synuclein was separated by centrifugation (21k g for 10 min). Soluble and insoluble fractions were boiled in SDS, run by SDS-PAGE, and colloidal stained. α -Synuclein in each fraction was quantified by in-gel densitometry. NAC was added at 20x molar excess to small molecule and catalase was added equal to 5% of protein, providing excess hydrogen peroxide decomposition capacity ($n = 3-5$, $***p < 0.001$). **(B)** Colloidal staining of representative gels showed the formation of insoluble α -synuclein aggregates in the presence of NDGA and NAC. **(C)** Near-infrared fluorescent imaging of the same gels before colloidal staining showed a reduction of quinone-dependent fluorescence in the presence of NAC. **(D)** α -Synuclein did not aggregate in the presence of cyclizable analogs, NDGA and NDGA-5, but did with non-cyclizable NDGA-1 ($n = 3$, $***p < 0.001$). **(E)** α -Synuclein did not aggregate in the presence of cyclized cNDGA and cNDGA-5 ($n = 3$). α -Synuclein was aggregated for

3 days in the presence of 1:1 small molecules. After aggregation, PBS-insoluble α -synuclein was separated by centrifugation (21k g for 10 min). Soluble and insoluble fractions were boiled in SDS, run by SDS-PAGE, and colloidal stained. α -Synuclein in each fraction was quantified by in-gel densitometry.

hindrance in the crosslinking region (Supplementary Fig. S2). α -Synuclein was aggregated for 3 days in the presence of equimolar NDGA analogs, and aggregation was quantified by the solubility assay. Both cyclizing analogs, NDGA and NDGA-5, inhibited α -synuclein aggregation, while non-cyclizing NDGA-1 did not (Fig. 2D). The products of the above experiment were also analyzed by nIRF. NDGA and NDGA-5-treated α -synuclein exhibited quinone nIRF signals, while NDGA-1-treated α -synuclein did not (Supplementary Fig. S4). These results indicate that cyclized forms of NDGA, or the process of cyclization, are required for inhibition of α -synuclein aggregation and formation of quinone-modified α -synuclein. To address this uncertainty, α -synuclein was aggregated in the presence of cNDGA and cNDGA-5, the purified cyclized forms of NDGA and NDGA-5, respectively (Supplementary Fig. S2). Using the solubility assay, both cNDGA and cNDGA-5 inhibited α -synuclein aggregation as potently as NDGA (Figs 2E and S5). This indicates that NDGA's inhibition of α -synuclein aggregation is mediated by interaction of the protein with cyclized oxidation products.

NDGA induces modest, progressive α -synuclein compaction without preventing dynamic adoption of α -helical conformation. During analysis of the interaction between NDGA and α -synuclein, equimolar incubation consistently produced nIRF-positive quinone-modified α -synuclein (Supplementary Fig. S6). Based on this observation, NDGA-treated α -synuclein was further analyzed to determine any structural effects of the interaction. Native nano-electrospray ionization mass spectra (ESI-MS), collected following 10 minute incubations of the protein-ligand mixtures, detected α -synuclein-NDGA complexes with masses matching the theoretical values of the unmodified protein and ligand⁶⁵. This result indicates that the initial interaction between α -synuclein and NDGA does not require, nor produce, covalent modifications. Complexes were observed with higher relative intensities at low protein charge states, suggesting a preferential binding to the more compact α -synuclein conformations as previously observed with EGCG. However, levels of α -synuclein binding were much lower for NDGA than previously observed with EGCG⁴⁹ (Supplementary Fig. S7).

Ion-mobility mass spectrometry (IM-MS) allows for highly sensitive detection of changes in ion compactness, mass, and charge. Of note, this allows subtle changes in protein conformation e.g. caused by ligand binding to alternative sites, to be distinguished by compactness even at the same mass and charge of the ion. Further, native IM-MS can resolve multiple different binding events by their conformational effects without averaging across the species distribution. This capability has previously been employed to show the differential effects of EGCG and dopamine on α -synuclein compaction⁴⁹. IM-MS revealed that α -synuclein compaction was unchanged by incubation with either NDGA or mNDGA (Figs 3A and S7). However, incubation with cyclized NDGA analogs cNDGA and cNDGA-5 led to α -synuclein assuming more compact conformations (Fig. 3B). α -Synuclein did non-covalently bind cyclized analogs up to 3 times, as seen by the corresponding mass shifts, and further, interaction with greater numbers of cyclized analogs correlated with increased compaction. This finding recapitulates similar results after α -synuclein incubation with EGCG⁴⁹ and indicates that interaction with cyclized NDGA causes structural compaction of α -synuclein.

NMR allowed for localization of NDGA's structural effects within the sequence of α -synuclein. After 24 hours incubation with 1:1 NDGA small variations in peak positions were observed in the N-terminus – Val 3, Phe 4, Met 5 – as well as His 50 (Fig. 3C). cNDGA and NDGA-1-treatment caused similar changes in the spectra (data not shown). Incubation with 3:1 NDGA resulted in more pronounced shifts in the same set of peaks (Supplementary Fig. S8).

These samples were followed for several days through NMR revealing a progressive variation of the peak shifts. After several days, additional changes occurred in the signals of several peaks. These peaks constitute a second set of signals, indicative of a subset of the α -synuclein population with altered conformation at these residues.

Förster resonance energy transfer (FRET) analysis of α -synuclein provided further insight into the regions altered by NDGA treatment. This examination employed several synthetic α -synuclein proteins with varying fluorescent label pairs spread throughout the amino acid sequence. Based on this methodology, FRET signal from an individual label pair can be converted to an intramolecular distance within α -synuclein and, when taken together, the multiple label pairs provided eight partially overlapped measurements covering the amino acid sequence from residue 9 to 136⁶⁶. FRET revealed progressive changes in α -synuclein conformation when treated with EGCG, while NDGA, cNDGA, and mNDGA did not induce changes, even after 24 hours (Fig. 3D,E). As with IM-MS and NMR, NDGA-treatment was found to induce minimal changes in α -synuclein conformation, while EGCG showed more marked effects.

NDGA-treated and dialyzed α -synuclein examined by CD showed no changes in secondary structure further confirming the IM-MS and NMR analyses (Fig. 3F). Despite observing limited effects of NDGA on α -synuclein structure in aqueous solution, we also examined whether NDGA treatment alters the capacity of α -synuclein to interact with hydrophobic membranes. Membrane interaction induces the N-terminus of α -synuclein to assume an α -helical secondary structure⁶⁷⁻⁶⁹. Capacity for membrane interaction is necessary for α -synuclein's putative biological functions in the synapse^{5,18,19}. Neither NDGA nor mNDGA treatment prevented α -synuclein from assuming α -helical conformations in the presence of SDS, just as occurred under control conditions (Fig. 3G). This indicates that the structural changes induced by NDGA do not restrict α -synuclein dynamic flexibility.

The ability of NDGA-treated α -synuclein to interact with membranes was further examined by fluorescence correlation spectroscopy (FCS), which has been used extensively to study the interaction of α -synuclein with lipid vesicles^{16,70}. The binding of fluorescently labelled recombinant human α -synuclein to various concentrations of

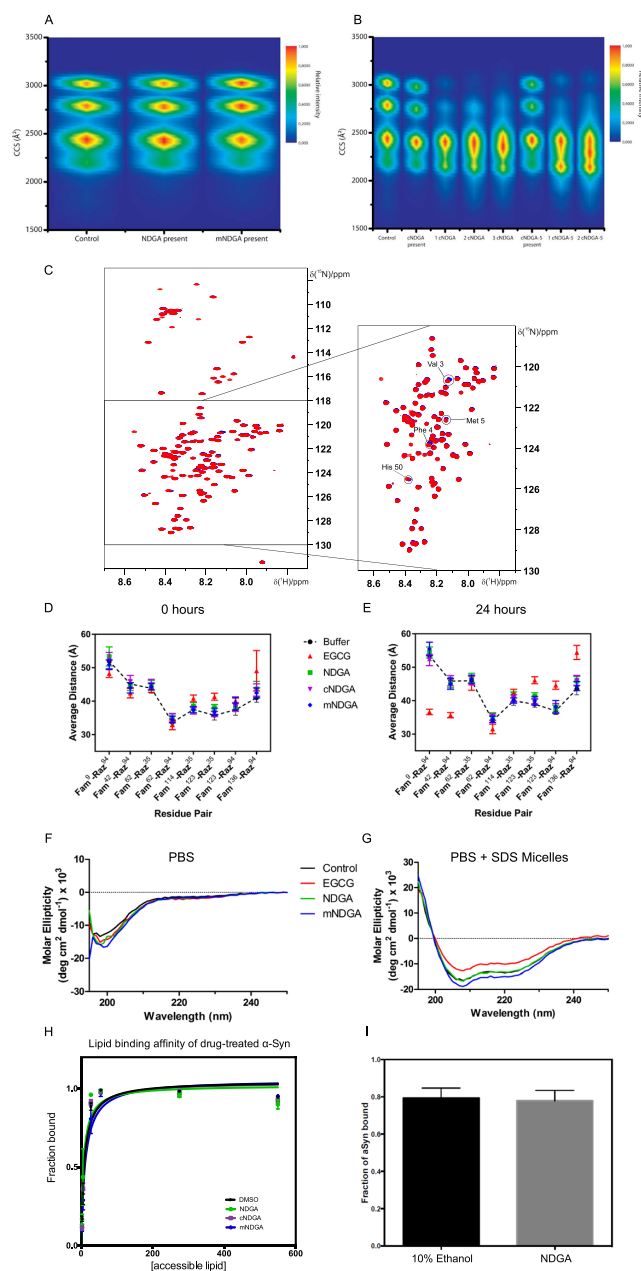


Figure 3. NDGA induces compaction of α -synuclein without preventing structural remodeling. **(A)** NDGA and mNDGA treatment did not alter α -synuclein collisional cross section measured by IM-MS. **(B)** cNDGA and cNDGA-5 induced α -synuclein compaction as measured by IM-MS. α -Synuclein was incubated with each molecule at 5:1 molar excess for 10 minutes before measurement. **(C)** NDGA treatment of α -synuclein did not induce extensive shifts in 2D NMR spectra. α -Synuclein was incubated 1:1 with NDGA for 24 hours before spectra were collected. NDGA-treated α -synuclein spectra (blue) was overlaid on solvent-treated α -synuclein (red). Fluorescein-maleimide (Fam) and tetramethylrhodamide azide (Raz) residue FRET measurement of **(D)** 0 hour and **(E)** 24 hour small molecule treatments showed progressive alteration of α -synuclein intramolecular distances by EGCG, but not NDGA, cNDGA, or mNDGA. The dashed line depicts 1 μ M α -synuclein treated with buffer (1x PBS). Treatments were 5 μ M NDGA (green square), EGCG (red triangle), cNDGA (purple inverted triangle), and mNDGA (blue diamond) ($n = 3$). **(F)** α -Synuclein secondary structure was not altered by pretreatment with NDGA. **(G)** NDGA pretreatment did not prevent α -synuclein assuming α -helical

secondary structure in the presence of SDS micelles. α -Synuclein was incubated 1:1 with NDGA, mNDGA, or solvent alone for 24 hours and then dialyzed against PBS for 24 hours. 40 mM SDS micelles were added 5 minutes before analysis. Secondary structure was quantified by circular dichroism. (H) NDGA treatment did not alter α -synuclein phospholipid binding. α -Synuclein was incubated 1:1 with NDGA analogs or solvent alone for 24 hours before fluorescence correlation spectroscopy in the presence of POPS:POPC vesicles at the indicated concentrations ($n = 3$). (I) Addition of NDGA did not displace fluorescently labeled α -synuclein from POPS:POPC vesicles ($n = 3$).

phospholipid vesicles (1:1, POPS:POPC) was compared following α -synuclein incubation for 24 hours with equimolar NDGA analogs or solvent alone. NDGA, cNDGA, and mNDGA did not alter phospholipid affinity as indicated by similar protein binding affinity at each vesicle concentration, regardless of treatment condition (Fig. 3H). FCS was also used to determine whether the addition of NDGA would displace α -synuclein already bound to phospholipid vesicles. The fraction of α -synuclein bound to vesicles was not altered by the addition of NDGA as compared to the addition of solvent alone (Fig. 3I). These data further indicate that NDGA treatment does not perturb α -synuclein interaction with hydrophobic membranes. Similarly, EGCG did not disrupt phospholipid or vesicle binding (Supplementary Fig. S9). Collectively, these results show that NDGA causes α -synuclein compaction across the sequence, without altering secondary structure or preventing structural remodeling involved in physiological function.

NDGA pretreated α -synuclein stably resists seeded aggregation. Based on the progressive remodeling of α -synuclein treated with NDGA, we analyzed the capacity of NDGA-pretreated α -synuclein to aggregate into fibrils. α -Synuclein was incubated with equimolar NDGA (as well as mNDGA and EGCG controls) for 24 hours before dialysis against excess PBS for 24 hours. This pretreated, dialyzed α -synuclein was then aggregated for 3 days before undergoing the solubility assay (as depicted in Supplementary Fig. S10). NDGA, but not mNDGA, pretreatment prevented α -synuclein aggregation (Fig. 4A). Further, pretreated, dialyzed α -synuclein was aggregated for 14 days to test the duration of the effect. Again, NDGA but not mNDGA pretreatment prevented α -synuclein aggregation (Supplementary Fig. S11). Consistent with previous data, we found that pretreatment with cyclizing (NDGA and NDGA-5) and cyclized (cNDGA and cNDGA-5) analogs prevented α -synuclein aggregation, while non-cyclizing analogs (NDGA-1 and SECO-1) did not (Supplementary Figs S12 and S13). We also examined the soluble species that remained following aggregation of pretreated, dialyzed α -synuclein using native state SEC and found that monomeric species are the majority product (Supplementary Fig. S14). Collectively, these findings demonstrate the NDGA pretreatment renders α -synuclein monomers resistant to aggregation. This effect was not observed with non-oxidizable and non-cyclizing analogs of NDGA.

Native-state size exclusion chromatographic (SEC) fractionation found that pretreatment with either NDGA or EGCG produces both monomeric and oligomeric α -synuclein species (Supplementary Fig. S6). We compared the aggregation of these two pools of α -synuclein species. α -Synuclein was incubated with equimolar concentrations of small molecules for 24 hours before oligomeric ($\geq 51 \text{ \AA}$) and monomeric ($< 51 \text{ \AA}$) species were separated using native state SEC. The resulting fractions were then aggregated for 3 days before analysis by the solubility assay. Both monomers and oligomers produced upon NDGA and EGCG pretreatment resisted aggregation, while the majority of mNDGA treatment products formed insoluble aggregates (Fig. 4b). This indicates that while aggregation-resistant α -synuclein oligomers do form after NDGA treatment, resulting monomers are also aggregation-resistant, even in the absence of oligomers.

Next, we explored whether NDGA can prevent surface-catalyzed, secondary nucleation-dependent, aggregation of α -synuclein. Pretreated, dialyzed α -synuclein was aggregated for 3 days in the presence of α -synuclein fibril seeds (equal to 5% of total α -synuclein in solution) before aggregation was quantified by the solubility assay. Again, NDGA but not mNDGA pretreatment reduced α -synuclein aggregation (Fig. 4c). This demonstrates that NDGA-pretreated, dialyzed α -synuclein does not readily undergo secondary nucleation to form species capable of incorporating into existing fibrils.

Based on the observation that NDGA treatment prevents α -synuclein aggregation, we conducted a novel analysis to determine whether aggregation of untreated α -synuclein is altered by the presence of NDGA-pretreated α -synuclein. Pretreated, dialyzed α -synuclein was generated as before, then mixed with untreated α -synuclein at varying ratios and the resulting mixtures were aggregated for 3 days before undergoing the solubility assay (as depicted in Supplementary Fig. S15). The addition of pretreated, dialyzed α -synuclein caused a super-stoichiometric reduction in insoluble α -synuclein, demonstrating that aggregation of untreated α -synuclein was reduced (Fig. 4d). This effect is demonstrated in the 20% condition wherein NDGA pretreated, dialyzed α -synuclein represents only 20% of total α -synuclein, but insoluble α -synuclein is reduced from 88% in control conditions to 37%. These data, taken in totality, indicate that pretreatment with NDGA produces α -synuclein monomers that are not only resistant to aggregation, but reduce aggregation of α -synuclein never exposed to NDGA.

cNDGA reduces α -synuclein-driven dopamine neurodegeneration in animals. We examined the effect of NDGA and cNDGA treatment on α -synuclein-driven neurodegeneration (Fig. 5A,B). We employed a widely used, *Caenorhabditis elegans* (*C. elegans*) model in which expression of human wildtype α -synuclein in dopaminergic neurons leads to progressive neurodegeneration⁷¹. Animals were treated with varying doses of NDGA, or cNDGA from hatching until the day before scoring on either day 6 or day 8. Animals were considered to be undergoing neurodegeneration if any dopamine neurons were absent or dendritic processes showed signs of dysfunction such as blebbing⁷². Representative images of animals scored during this experiment show

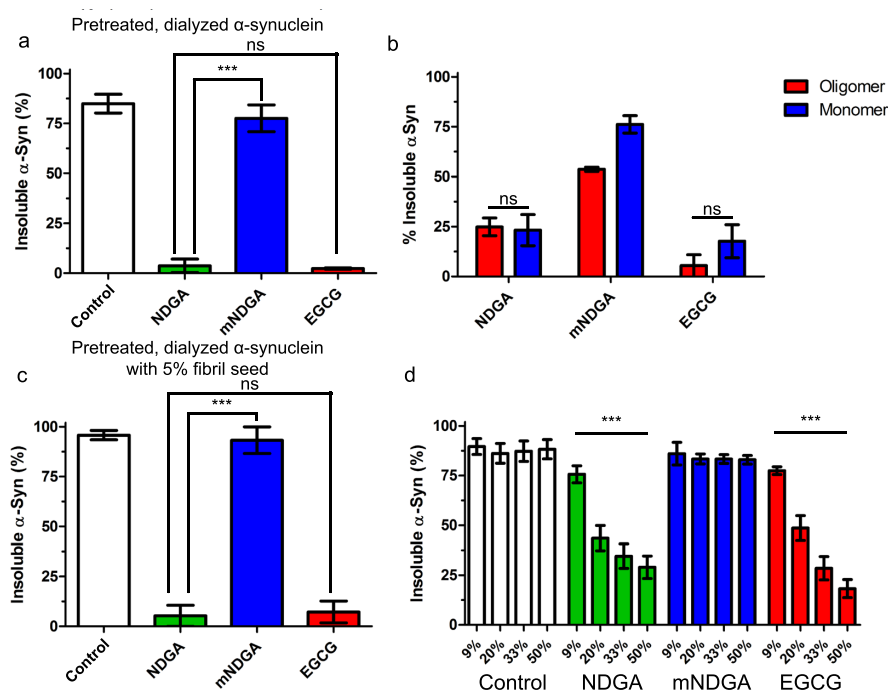


Figure 4. NDGA pretreatment prevents α -synuclein aggregation. (a) α -Synuclein did not aggregate after pretreatment with NDGA. α -Synuclein was incubated 1:1 with small molecules for 24 hours then dialyzed against PBS for 24 hours. After aggregation for 3 days, PBS-insoluble α -synuclein was separated by centrifugation (21k g for 10 min). Soluble and insoluble fractions were boiled in SDS, run by SDS-PAGE, and colloidal stained. α -Synuclein in each fraction was quantified by in-gel densitometry ($n = 3$, *** $p < 0.001$). (b) Both oligomeric and monomeric α -synuclein species induced by NDGA treatment resist aggregation. α -Synuclein was incubated 1:1 with small molecules for 24 hours then subjected to native state size exclusion chromatography. Oligomeric ($\geq 51 \text{ \AA}$) and Monomeric ($< 51 \text{ \AA}$) α -synuclein fractions were collected and aggregated separately for 3 days. Soluble and insoluble species were separated and quantified as above ($n = 3$). (c) NDGA-treated α -synuclein resisted fibrillization in the presence of 5% fibril seed. α -Synuclein was incubated 1:1 with small molecules for 24 hours then dialyzed against excess PBS for 24 hours. Untreated α -synuclein fibrils equal to 5% total protein was added immediately before mixtures were aggregated for 3 days. Soluble and insoluble species were separated and quantified as above ($n = 3$, *** $p < 0.001$). (d) NDGA pretreated, dialyzed α -synuclein inhibited aggregation of untreated α -synuclein. α -Synuclein was incubated 1:1 with small molecules for 24 hours then dialyzed against excess PBS for 24 hours. Pretreated, dialyzed α -synuclein was then mixed with untreated monomeric α -synuclein at the indicated ratios. Mixtures were aggregated for 3 days. Soluble and insoluble species were separated and quantified as above ($n = 3$, *** $p < 0.001$).

the loss of fluorescent dopamine neuron processes and bodies in solvent and NDGA-treated animals (Fig. 5B). No differences were observed in NDGA or cNDGA treated animal after day 6 (data not shown). cNDGA at all concentrations (10, 50 and 100 μM) significantly reduced the number of animals undergoing dopamine neuron degeneration at day 8 (Fig. 5A). Similarly, EGCG (50 μM) protected against neuron loss (Supplementary Fig. S16). In this model, expression of α -synuclein was limited to just six neurons with each animal. As such, total α -synuclein levels were so low as to preclude biochemical analysis of the interaction between α -synuclein and NDGA or cNDGA. Likewise, we were unable to conduct pharmacokinetic analysis of our target neurons. Despite these limited reservations, these results demonstrate, for the first time, that cNDGA can reduce neurodegeneration caused by α -synuclein.

Discussion

Because of the engagement in several neurodegenerative diseases, α -synuclein aggregation remains a promising target for therapeutic development. Using NDGA and novel analogs, we uncovered that NDGA oxidation and cyclization was required for formation of quinone-modified monomeric α -synuclein. This interaction caused slight compaction of the α -synuclein molecule but did not induce any noticeable changes in its intrinsically

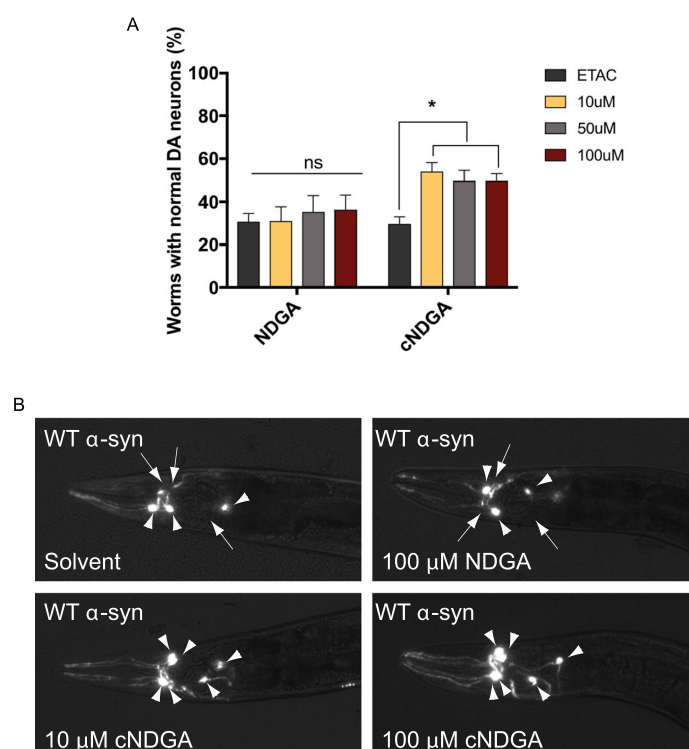


Figure 5. cNDGA reduces α -synuclein-driven neurodegeneration (A) cNDGA, but not NDGA reduces dopaminergic neurodegeneration in *C. elegans* expressing wildtype α -synuclein. Animals were exposed to each concentration of the drugs on days 0–3, 5, and 7 post-hatching. Animals were scored on day 8 post-hatching for dopaminergic neurodegeneration. The data are represented as mean \pm SEM; one-way ANOVA with Tukey's *post hoc* test for multiple comparisons ($n = 3$; 30 animals per replicate; * $p < 0.05$; ns, not significant). (B) Representative images from day 8 post-hatching. Arrowheads indicate intact neurons while arrows indicate degenerating or missing neurons. Scale bar represents 10 μ m.

disordered state. Furthermore, NDGA-treated α -synuclein retained its ability to undergo conformational changes required for interaction with the membranes, and NDGA did not dissociate protein-membrane complexes. However, interaction of α -synuclein with NDGA and its cyclizing and cyclized analogues efficiently prevented α -synuclein aggregation into fibrils. Conversely, non-cyclizing NDGA analogs had no effect on fibril formation, ensuring the aggregation resistant effects of NDGA and its analogs is not an artifact of reduced contact between α -synuclein molecules. Indeed, NDGA-treated monomers not only resisted aggregation, but prevented aggregation of naive untreated α -synuclein. Treatment with cNDGA reduced neurodegeneration in an animal model of α -synuclein neurotoxicity. Taken together, these findings suggest that NDGA and related molecules that prevent α -synuclein aggregation without perturbing its native intrinsic disorder, or associations, may be promising targets for PD and related disorders characterized by α -synuclein aggregation.

Attempts to develop therapeutic strategies targeting α -synuclein have included decreasing levels by altering expression⁷³ and clearance^{74,75}, immunological targeting^{76,77}, and altering aggregation. Many aggregation inhibitors were identified through chemical library screens before being subjected to mechanistic analysis^{32,40,41}. The neurotransmitter dopamine, and molecules that share structural and chemical similarities, were the first inhibitors identified and are among the most potent inhibitors of α -synuclein aggregation³².

Identification of dopamine and related phenols as inhibitors of α -synuclein aggregation has led to many studies of phenolic compounds as potentially neuroprotective inhibitors. The most studied of these is EGCG, which is currently undergoing clinical trials in multiple system atrophy, a synucleinopathy characterized by microglial inclusions⁷⁸. EGCG was found to be neuroprotective in animals treated with MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine), which is a prodrug to the neurotoxin MPP⁺ (1-methyl-4-phenylpyridinium) that induces α -synuclein-dependent dopaminergic neurodegeneration^{79–81}. Adding EGCG during *in vitro* α -synuclein aggregation results in generation of less toxic aggregated products³⁹. Given the structural differences between EGCG and NDGA, as well as the existing knowledge about NDGA oxidation and the availability of analogs, we hope

that studying NDGA as an α -synuclein aggregation inhibitor will provide new insight into the neuroprotective potential of inhibiting α -synuclein aggregation.

Our analysis revealed some substantial departures between the effects of NDGA and EGCG on α -synuclein. Konijnenberg *et al.* found, using IM-MS, that EGCG induces α -synuclein compaction⁴⁹, while analysis conducted by the same group for this study showed that cNDGA and cNDGA-5 induce α -synuclein compaction (Fig. 3A,B). NDGA, however, did not induce α -synuclein structural change. Interestingly, FRET found that EGCG causes substantial α -synuclein remodeling, while NDGA and cNDGA have no measurable effect on intramolecular distances (Fig. 3C). We also observed consistent differences in the patterns of SDS-stable multimers present after α -synuclein was treated with EGCG and NDGA (Figs 2B and S3, S6, S14). This trend of subtle, but measurable, differences between the effect of the two molecules leaves open the possibility of different mechanisms of their interaction with α -synuclein and differing biophysical properties of the resulting species.

Previous studies of EGCG have focused on the oligomers formed when α -synuclein is aggregated in the presence of EGCG^{39,82,83}. In this study, we found that α -synuclein aggregation in the presence of EGCG and NDGA does produce oligomers, but that monomers are the predominant product, representing more than 75% of soluble protein (Supplementary Fig. S14). Further, we found that α -synuclein monomers and oligomers induced by pretreatment with EGCG and NDGA both resisted aggregation (Fig. 4b). The corresponding aggregation experiments were conducted after applying size-exclusion chromatography, which removed from solution any EGCG or NDGA not interacting with α -synuclein, suggesting a stable interaction or modification of α -synuclein itself. Most strikingly, NDGA-pretreated species inhibited aggregation of untreated α -synuclein, suggesting a dominant-negative effect on aggregation. This effect, not previously described for EGCG or any other exogenous phenolic inhibitors, may contribute to their neuroprotective effects.

There are two proposed models of amyloid-like fibril formation and elongation, although the precise mechanism remains unclear. In the first model monomers form oligomers then stack together to form fibrils. In the second model of fibril formation, oligomers are elongated to fibrils with the addition of monomers. Under either paradigm, the data indicate that NDGA or EGCG modified oligomers do not progress to insoluble species. Possibly due to the increased compaction, monomeric α -synuclein treated with NDGA analogs, when incorporated into oligomers or fibrils may prevent further elongation. Essentially, the addition of a NDGA analog-modified α -synuclein monomer will “cap” the aggregation intermediate preventing the further addition of either monomer or oligomer. The stoichiometry would indicate that under either model the incorporation of aggregation resistant NDGA or EGCG modified monomer into oligomers or fibrils is sufficient to prevent further aggregation. The combination of a durable modification of α -synuclein and a dominant negative effect on aggregation are very desirable, given concerns over polyphenol bioavailability in the brain⁸⁴ and the hepatotoxic side effects seen with EGCG or NDGA^{78,85}.

Lipid-bound α -synuclein is resistant to aggregation⁸⁶, and appropriate membrane interactions are implicated in its physiological function⁸⁷. As such, an ideal therapeutic targeting α -synuclein aggregation would allow normal lipid interaction. Interestingly, both CD and FCS showed that the membrane interactions of α -synuclein were not perturbed by NDGA-treatment (Fig. 3G–I). Preservation of α -synuclein membrane interaction may help explain the divergent effects of NDGA and dopamine on neuron viability. Indeed, α -synuclein oligomers induce membrane destabilization and dopamine-induced oligomers may contribute to neurodegeneration in synucleinopathies^{36,37}. Further, the ability of α -synuclein to undergo native structural dynamics is necessary for formation of multimers implicated in physiological function^{59,60} and aggregation resistance^{88,89}. Future studies aimed at identifying neuroprotective α -synuclein aggregation inhibitors would be well served by incorporating methods to determine whether membrane interactions and structural dynamics are maintained, as is the case with NDGA.

NDGA pretreatment induced an NMR peak shift at His 50 (Fig. 3C). His 50 is located within the second region that assumes an α -helix during α -synuclein interaction with curved lipid membranes, falling in sequence just after the fourth lysine-rich repeat^{67,69}. His 50 also mediates an oligomer-stabilizing interaction between α -synuclein and divalent metal ions, particularly Cu(II)^{90,91}. Additionally, His 50 to Gln mutation causes autosomal dominant familial PD and accelerates α -synuclein aggregation *in vitro*^{92–94}. This suggests that interaction with His 50 could alter α -synuclein aggregation kinetics. Further investigation of the interaction between NDGA and His 50 might provide insight into NDGA's inhibition of α -synuclein aggregation. Curiously, three other residues whose positions were perturbed by NDGA binding (Val 3, Phe 4, and Met 5) are located in the close proximity to the Asp 2, which is also engaged in Cu(II) binding⁹⁵.

The chemistry underpinning dopamine and related small molecule modulators of α -synuclein aggregation includes the capacity for auto-oxidation and the presence of vicinal hydroxyl (catechol) moieties^{33,35}. In this study, we have expanded the previous observations and documented many new mechanistic insights for the biochemistry of catechol structures to inhibit the aggregation of α -synuclein. We found that NDGA-analogs require two pairs of vicinal hydroxyls to interact with α -synuclein. NDGA-1, which contains only one pair of vicinal hydroxyl groups, did not modify α -synuclein or inhibit aggregation. This indicates that NDGA cyclization, which is enabled by its two pairs of vicinal hydroxyls, is required for inhibition of α -synuclein aggregation⁶³. Indeed, cNDGA and cNDGA-5, two cyclized analogs of NDGA, were found to inhibit α -synuclein aggregation. The finding that NDGA's cyclized oxidation products are responsible for its effects on α -synuclein aggregation presents the dual opportunities to study NDGA as a prodrug, with potential for conversion by oxidation-dependent cyclization to active cNDGA, and to examine novel molecules based on the cNDGA structure in hopes of identifying inhibitors of α -synuclein aggregation.

The importance of NDGA cyclization is reinforced by the finding that cNDGA, but not NDGA, treatment reduces neurodegeneration caused by expression of human α -synuclein in dopamine neurons in *C. elegans* (Fig. 5). There are many potential explanations for the divergent effects of the two molecules, including differences in stability, uptake, metabolism, and excretion. One possibility is that NDGA does not form cNDGA under these conditions. Oxidation of NDGA likely occurs less frequently in the reducing environment of the cell⁹⁶,

and yield of cNDGA may be lower than previously observed. While the cyclization of NDGA has been precisely described under controlled conditions in buffer⁶³, it has never been examined in a complex milieu or within a living cell, where other chemistry may occur.

It could be of interest to compare the molecular mechanism of action of NDGA on α -synuclein with the effects of different small molecules on functionality of other intrinsically disordered proteins. One of the best studied examples of such systems is given by small molecule-driven inhibition of heterodimerization of the transcription factor c-Myc with its partner, Max, via a basic-helix-loop-helix-leucine zipper (bHLHZip) domain present in both proteins. Importantly, both proteins in their unbound forms are disordered, but undergo mutual coupled binding and folding when their zipper domains interact to form a helical coiled coil^{97–99}. High throughput screening uncovered several specific inhibitors that were able to bind to one of three discrete sites (residues 366–375, 375–385, and 402–409) within the 85-residue bHLHZip domain of the monomeric c-Myc^{100,101}. Solution NMR analysis revealed that interaction of said small molecules with c-Myc resulted in its local misfolding, thereby generating conformations incompatible with the heterodimerization of this protein with Max^{97–99}. Clearly, the mechanism of NDGA action, where a small molecule stabilizes specific conformations of intrinsically disordered α -synuclein, is principally different from the inhibitory activity of small molecules inducing c-Myc misfolding. Therefore, further study is warranted to determine whether the interaction between NDGA and α -synuclein represents a new mode of modulatory action of small molecules on intrinsically disordered proteins. Moreover, in-cell NMR data indicated that molecular crowding agents present in the cytosol induce compaction of α -synuclein and shielding of the aggregation prone region (hydrophobic core between residues 61–95) preventing oligomerization and aggregation¹⁰². Although ensemble-based techniques (FRET or NMR) are not able to visualize and inform on the structures of these compact states, we speculate that the compact states observed by NDGA analogs and EGCG are similar to those observed with in-cell NMR.

Collectively, these findings change our understanding of the mechanisms of phenolic α -synuclein aggregation inhibitors. We demonstrate, for the first time, that inhibition of α -synuclein aggregation by NDGA requires oxidation-dependent cyclization, reframing NDGA as a prodrug. Additionally, we show that both NDGA and EGCG stably modify α -synuclein monomers, rendering them aggregation incompetent. In the case of NDGA, this is achieved without alteration of α -synuclein secondary structure in solution, nor perturbing membrane interactions. This combination of attributes has not been previously observed in other inhibitors of α -synuclein aggregation. Further examination is certainly warranted, including a full characterization of efficacy and critically toxicity, to determine whether NDGA analogs might provide the basis for novel neuroprotective therapies for PD and related synucleinopathies.

Methods

***In vitro* α -synuclein aggregation analyses.** Recombinant human wild type α -synuclein was expressed and purified from *E. coli* as previously described¹⁰³. Purified α -synuclein was used at 2 mg/ml (138 μ M) with or without small molecules at various molar ratios in PBS with 1% DMSO. Mixtures were aggregated for 3 or 7 days at 37 °C shaking at 1400 rpm. The longer 7 day incubation was performed to ensure the treatment effects remained consistent even under prolonged aggregation. Alternatively, purified α -synuclein was incubated for 24 hours with or with equimolar small molecules in PBS with 1% DMSO. Resulting mixtures were then dialyzed against excess PBS for 24 hours at 4 °C (using mini-dialysis tubes, Thermo Fisher 69572). Mixtures were aggregated for 3 or 14 days shaking at 1400 rpm. Amyloid fibril formation was monitored by thioflavin T. Briefly, 4 μ M protein was incubated with 25 μ M ThT (Sigma Aldrich) dissolved in PBS and read at Ex of 450 and Em 482 using a fluorescent plate reader (Molecular Devices).

α -Synuclein solubility analysis and densitometric quantification. Sedimentation was performed by centrifugation at 21,000 g for 15 minutes. Supernatants were removed, and pellets were resuspended in an equal volume. Pellets and supernatants were boiled in SDS sample buffer for 3 minutes at 95 °C. α -Synuclein species were resolved by SDS-PAGE in 12% Bis/Tris gels and stained in-gel with colloidal blue (Invitrogen LC6025). Stained gels were imaged at 700 nm using LI-COR Odyssey Infrared Imaging System and densitometric quantification was conducted using LI-COR Odyssey software suite 3.0.

Near infrared fluorescence. SDS-PAGE gels were imaged as previously described⁶⁴. Briefly, gels were imaged immediately after electrophoresis at 700 nm with intensity 10 on LI-COR Odyssey Infrared Imaging System.

Size-exclusion chromatography. For size exclusion chromatography 200 μ L recombinant α -synuclein (2 mg/mL) was injected onto a Superdex 200 10/300 GL (GE Healthcare) connected to an Agilent 1100 series HPLC system and fraction collector controlled by ChemStation software version 1.04 (Agilent). Mobile phase consisted of 25 mM HEPES and 150 mM NaCl, pH 7.25 with a 0.3 mL/min flow rate. Fractions corresponding to 140–100, 100–85, 85–75, 75–68, 68–61.5, 61.5–56, 56–51, 51–47, 47–43, 43–39, 39, 39–36, 36–32, 32–29, 29–26 Å were combined and concentrated using 3 kDa NMWL ultracel microcon filters (Millipore UFC5003). SEC column elution time was calibrated to Å size using globular protein standards (GE Healthcare).

Circular dichroism. Circular dichroism spectra were obtained on a Jasco-810 spectropolarimeter maintained in the Children's Hospital of Philadelphia Protein Core facility. Protein in PBS was diluted to 20 μ M in 0.05 M KH₂PO₄ pH 7.8. Spectra were corrected for baseline measurement of an equivalent volume of PBS diluted in KH₂PO₄ buffer. Spectra were collected with a scanning width of 1 nm, at 5 nm/min with 2 accumulations per run.

Transmission electron microscopy. Transmission electron microscopy was conducted on a JEOL1010 maintained in the University of Pennsylvania Electron Microscopy Resource Laboratory. Protein solutions were diluted to 0.5 mg/mL (34.5 μ M) before mounting to carbon-coated 300 grids and negative staining with 2% uranium acetate. Images were collected with HV = 80.0 kV, Magnification 50,000x. Contrast was automatically adjusted during capture for each image, and not altered after capture. 1 μ m scale bars were included for each image presented.

Nuclear magnetic resonance spectroscopy. To provide atomic resolution information on α -synuclein, 2D ^1H - ^{15}N NMR spectra were acquired through a series of band-selective excitation short-transient transverse relaxation optimized spectroscopy heteronuclear single quantum coherence (BEST-TROSY HSQC) runs¹⁰⁴, a fast and common experiment used to evaluate if there is an effect of the different molecules on some amino acids, each of them represented by a H^{N} -N correlation (a cross-peak in the 2D spectrum).

All the NMR experiments were recorded on a Bruker Avance III spectrometer operating at 900 MHz ^1H frequency (21.14 T) equipped with a cryogenically cooled probehead for triple resonance experiments (TCL). The pulses involved in this pulse sequence were the standard ones used to investigate bio-molecules¹⁰⁵. The ^{15}N hard pulses were applied at the center of the region (117.02 ppm) and the ^1H band selective pulses were applied at the center of the amide proton region at 8.70 ppm. A spectral width of 10822.511 Hz in the direct ^1H dimension and 3647.991 Hz in the indirect ^{15}N dimension was used. All the spectra were recorded with 8 scans per increment, 4096 and 2048 points in the direct dimension and in the indirect dimension respectively and a recycle delay of 500 ms.

The experiments were recorded at 288.0 K on a 165 μ M uniformly ^{15}N labeled α -synuclein sample, prepared as previously described¹⁰⁶, in buffer (10 mM KPi, 140 mM NaCl, 0.25 mM EDTA + Roche protease inhibitors, pH 7.5). 5% D_2O was added for the lock signal. The different drugs (described above) were added as DMSO solutions at 1:1 stoichiometric ratio and left at room temperature 24 hours to incubate before NMR measurements. The same experiments were acquired on a sample containing 3 equivalents of NDGA. Care was taken to minimize the volume of the DMSO solution added and to keep it constant for all the samples investigated. All the samples were analyzed again in the following days to evaluate thermodynamic and kinetic effects.

Native nano-electrospray ionization mass spectrometry. ESI-MS experiments were conducted as previously described⁴⁹. The only methodological departure was the inclusion of DMSO to solubilize the ligands. Briefly, MS spectra were collected after 10 minute incubation of protein-ligand mixtures in pH 7.4 10 mM ammonium acetate. Spectra were collected using a hybrid quadrupole-time-of-flight mass spectrometer (QSTAR-Elite, Biosystems, Foster City, CA) equipped with a nano-ESI sample source.

Native ion mobility mass spectrometry. Native Ion mobility-mass spectrometry (IM-MS) experiments were conducted as previously described⁴⁹. The only methodological departure was the inclusion of DMSO to solubilize the ligands. Briefly, MS spectra were collected after 10 minute incubation of protein-ligand mixtures in pH 7.4 10 mM ammonium acetate. α -Synuclein was present in the mixtures at 20 μ M, ligand at 100 μ M. IM-MS was performed on a Synapt G2 HDMS (Waters, Manchester, U.K.) using nano-ESI with homemade gold-coated borosilicate capillaries.

Fluorescence correlation spectroscopy. *Expression of fluorescently-labeled α -synuclein.* α -Synuclein protein labeled at residue S9C with Alexa Fluor 488 was produced via recombinant protein production. Plasmid containing α -synuclein S9C fused to a polyhistidine-tagged GyrA intein from *Mycobacterium xenopi* (Mxe) was transformed into BL21 DE3 competent cells by heat shocking at 42 $^{\circ}\text{C}$. Single colonies grown on Ampicillin (Amp) plates were picked to inoculate primary cultures in LB supplemented with 1 μ g/mL Amp. Primary cultures were combined into secondary cultures, which were grown at 37 $^{\circ}\text{C}$ in a shaker-incubator until optical density (OD) reached \sim 0.6. Expression of the gene of interest was induced with Isopropyl β -D-1-thiogalactopyranoside (IPTG). Cells were then grown in the shaker-incubator at 18 $^{\circ}\text{C}$ overnight. After centrifugation (5000 rpm, 20 min, 4 $^{\circ}\text{C}$), cell pellets were re-suspended in re-suspension buffer (20 mM Tris, 1 mM PMSF, 1 Roche protease inhibitor tablet, pH 8.3) and sonicated in a cup in an ice bath (5 min, 1 s ON, 1 s OFF). The resulting lysate was centrifuged (14,000 rpm, 25 min, 4 $^{\circ}\text{C}$), and supernatant containing the protein of interest (POI) was purified over Ni-NTA affinity column. Intein cleavage was carried out by incubation with 200 mM β -mercaptoethanol (β ME) on a rotisserie over night at room temperature. Cleaved POI was dialyzed into 20 mM Tris pH 8 before purification over a second Ni-NTA column to remove the free intein from the sample. Flow-through containing the POI was kept, dialyzed into 20 mM Tris pH 8, and spin concentrated. Labeling with Alexa Fluor 488 (AF488) maleimide was done by adding 4 equivalents (eq.) of fluorophore dissolved in DMSO. The reaction tube was wrapped in aluminum foil and incubated at 37 $^{\circ}\text{C}$ for \sim 6 hrs. Formation of the product α -synuclein C9-AF488 was checked by matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) and polyacrylamide gel electrophoresis (SDS-PAGE). Labeled protein was purified by fast-protein liquid chromatography (FPLC) using a Hi-Trap Q 5 mL column and by reverse-phase high-performance liquid chromatography (RP-HPLC) using a C18 preparatory column. Purified protein was spin concentrated into buffer (20 mM Tris, 100 mM NaCl), and aliquots were stored at -80°C .

Treating α -synuclein with drug molecules. The concentration of α -synuclein aliquots was determined by UV-Vis at 488 nm. The drug molecules NDGA, mNDGA, and cNDGA were dissolved in DMSO and diluted 10-fold into buffer (20 mM Tris, 100 mM NaCl). A control of 10% DMSO in buffer was also made. Each aliquot of α -synuclein was respectively treated with 1 eq. of NDGA, mNDGA, and cNDGA, and DMSO-buffer. Treatment samples were incubated at room temperature for 8 hours, then dialyzed against 10 mM Tris, 100 mM NaCl overnight.

Preparation of lipid vesicles. Synthetic lipid vesicles were prepared for use in binding experiments. A mixture in 50:50 molar ratio of 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoserine (POPS) and 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) were drawn from chloroform stock and dried under nitrogen gas to form a film inside a glass vial. Films were desiccated under vacuum and re-hydrated in 3-(N-morpholino) propanesulfonic acid (MOPS) buffer. 10 Freeze-thaw cycles consisting of cooling in liquid nitrogen for 40 s and warming in a 60 °C water bath for 2 min were performed to aid the formation of uniformly sized vesicles. Using syringes fitted onto an extruder, vesicles were pushed 31 times through 50 nm pore membranes. Vesicles were determined by Dynamic Light Scattering to be monodisperse and distributed uniformly around 80 nm in diameter, consistent across different concentrations of all samples. All lipid vesicles were prepared fresh and used within 48 hours of extrusion.

Fluorescence correlation spectroscopy. Eight-well chambered coverglasses (Nunc, Rochester, NY) were prepared by plasma cleaning followed by incubation over night with polylysine-conjugated polyethylene glycol (PEG-PLL), prepared using a modified Pierce PEGylation protocol (Pierce, Rockford, IL). PEG-PLL coated Chambers were rinsed with and stored in Milli-Q water until use. FCS measurements were done on a lab-built instrument based on an Olympus IX71 microscope with a continuous emission 488 nm DPSS 50 mW laser (Spectra-Physics, Santa Clara, CA). All measurements were done at 20 °C. The laser power entering the microscope was adjusted to 4.5 μ W. Fluorescence emission collected through the objective was separated from the excitation signal through a Z488rdc long pass dichroic and an HQ600/200 m bandpass filter (Chroma, Bellows Falls, VT). Emission signal was focused through a 50 μ m optical fiber. Signal was amplified by an avalanche photodiode (Perkin Elmer, Waltham, MA) coupled to the fiber. A digital autocorrelator (Flex03Q-12, correlator.com, Bridgewater, NJ) was used to collect 30 autocorrelation curves of 30 seconds for each measurement. Fitting was done using MATLAB (The MathWorks, Natick, MA).

Binding assay of drug-treated α -synuclein to lipid vesicles. FCS was used to examine the binding affinity of drug-treated α -synuclein to lipid vesicles. Each drug-treated α -synuclein labeled with AF488 was examined in the presence of varying concentrations of lipid vesicles, and each autocorrelation curve was fit to a 2-component equation to extract the fraction of bound α -syn at each concentration. From these data, a binding curve was generated and fit to an exponential to determine the dissociation constant K_d , i.e. the concentration at which half of the protein is bound. In fitting the autocorrelation curves for α -syn in the presence of lipid vesicles, the diffusion time (τ) of bound and unbound α -syn were respectively fixed to experimentally determined values. The diffusion time of unbound protein, $\tau_{\alpha S}$, was determined by measurements of the protein in buffer without lipids. Since bound protein diffuses with the vesicles to which they are bound, the diffusion time of the vesicles, $\tau_{vesicle}$, was determined by measurements of BODIPY-labeled 50:50 POPS/POPC. The diffusion time of our unlabeled vesicles were also deduced from a calibration curve generated from the diffusion times of commercial fluorescent bead standards of 50 nm and 100 nm in diameter.

To generate a vesicle-binding curve, FCS was performed on α -syn C9-AF488 in the presence of varying concentrations (0.005 mM to 1 mM lipid) of 50:50 POPS/POPC. The accessible surface lipid concentration was calculated based on the characteristic bilayer thickness of POPS and POPC. The fraction of bound protein was extracted from the fit to each autocorrelation curve by fixing $\tau_{\alpha S}$ and globally fitting $\tau_{vesicle}$ to the same value across all concentrations.

$$G(\tau) = \frac{1}{N} \left[F_B * \frac{1}{1 + \frac{\tau}{\tau_{\alpha S}}} * \left(\frac{1}{1 + \frac{s^2 \tau}{\tau_{\alpha S}}} \right)^{1/2} + (1 - F_B) * \frac{1}{1 + \frac{\tau}{\tau_{vesicle}}} * \left(\frac{1}{1 + \frac{s^2 \tau}{\tau_{vesicle}}} \right)^{1/2} \right]$$

The fraction bound was plotted against accessible lipid concentration and fit to an exponential, from which the K_d was determined.

$$F_B = \frac{B_{max} x}{K_{d,app} + x}$$

Assay for effect of NDGA on lipid-bound α -synuclein. Untreated α -synuclein C9-AF488 (final concentration 20 μ M) was added to 250 μ L of 0.05 mM POPS/PC in a microscope chamber well. The sample was incubated for 15 minutes to ensure maximum binding interactions. NDGA stock was made by dissolving the drug in ethanol and diluting 10-fold into buffer (20 mM Tris, 100 mM NaCl). NDGA was added at 20 μ M final concentration to the vesicle-bound α -synuclein sample, and fraction of protein bound was determined using the fitting described above. A control treatment was made of 10% ethanol in buffer, and fraction of protein bound was determined in the same manner.

Förster resonance energy transfer analysis. Small molecule stock solutions containing either NDGA, EGCG, cNDGA or mNDGA were prepared in DMSO. Labeled proteins constructs, including new constructs not previously described, were prepared using methods previously described⁴⁶. Labeled positions not previously reported were confirmed by mass spectrometry of both the full-length protein and fragments from trypsin digestion using MALDI. The concentration of each protein stock was assessed by UV-Vis absorbance of the attached fluorescein-maleimide (Fam) ($\epsilon_{494} = 68,000 \text{ M}^{-1} \text{ cm}^{-1}$) and/or tetramethylrhodamide azide (Raz)

($\epsilon_{555} = 87,000 \text{ M}^{-1} \text{ cm}^{-1}$) dyes. Fluorescence assays were performed in a black, clear bottom half-diameter Greiner 96-well plate and measurements were taken on a Tecan M1000 plate reader. Each well of the plate contained a single protein sample, containing either one or both fluorescent labels, and one of the small molecules of interest. Protein samples were also measured in the absence of small molecules as a control. Each sample was prepared in 1X PBS with a final concentration of 1 μM protein. Samples containing NDGA, EGCG, cNDGA and mNDGA were prepared with a 5-fold excess of small molecules relative to the protein concentration. Small molecule stocks were prepared just prior to performing the assay. Each assay was performed by diluting the protein in buffer in each well and adding the small molecule solution, using a multichannel pipette, just prior to measuring the fluorescence. Each sample was excited at 494 nm and the emission spectrum was measured from 502–700 nm with a 1 nm step size. Each step consisted of 50 flashes at a frequency of 400 Hz with a 20 μs integration time and the gain was set to 135. The Z-position of the plate was optimized prior to the first measurement and was maintained at 21728 μm for all measurements. All measurements were performed at room temperature. After measuring the emission spectrum from each well immediately after the introduction of small molecule, the plates were sealed with parafilm and covered in aluminum foil and left at room temperature for 24 hrs. After 24 hrs the fluorescence of each sample was measured again as detailed above. The FRET between Fam and Raz probes attached to a single protein was assessed by computing the fluorescence quenching of the donor Fam fluorophore induced by the presence of the Raz acceptor as previously described⁶⁶. FRET efficiencies were converted to average interresidue distances using the Förster equation and a gaussian chain polymer model.

Analysis of dopaminergic neurodegeneration in *C. elegans*. *Generation of transgenic nematodes.* Transgenic *C. elegans* lines were generated by microinjection using previously described methods¹⁰⁷. The strain UA294 (*baEx175 a,b,c* [$P_{dat-1}::\text{WT a-syn}$, $P_{unc-54}::\text{tdTomato}$]; *vIs7* [$P_{dat-1}::\text{GFP}$]) was generated by injecting with a solution of 50 ng/ μL plasmid with either $P_{dat-1}::\text{WT a-syn}$ with a phenotypic marker ($P_{unc-54}::\text{tdTomato}$, 50 ng/ μL , for body wall muscle expression). Three distinct stable lines (*a*, *b*, *c*) were generated.

Analysis of dopaminergic neurodegeneration and NDGA analog treatments in C. elegans. NDGA and cNDGA were used at final concentrations of 10 μM , 50 μM , or 100 μM in ethyl acetate (EtAc). NDGA, cNDGA, or solvent (EtAc) was added to the surface of bacterial lawn on the nematode growth medium (NGM) Petri plates (60 mm diameter) at the corresponding final concentration and allowed to dry under the hood. For EGCG, 50 μM was added into the NGM agar plates. Three resulting independent transgenic lines (*a*, *b*, *c*) were synchronized, and exposed to corresponding concentration of NDGA, cNDGA, or solvent (EtAc) from hatching to day 3, then were transferred and maintained on NGM plates until the day of analysis. Animals on NGM plates were treated with additional NDGA, cNDGA, or solvent on days 5 and 7. For EGCG treatment, worms were transferred to fresh NGM plates daily and transferred onto fresh EGCG containing plates on days 0–3 and 5 post-hatching. For dopaminergic neurodegeneration analyses, the transgenic animals were scored as described previously⁷². Briefly, on the day of analysis, the six anterior dopaminergic neurons (four CEP (cephalic) and two ADE (anterior deirid)) were examined in 30 randomly selected nematodes with the marker transgene (tdTomato) in the body wall muscle cells. Worms were considered normal when all six anterior neurons were present without any signs of degeneration, such as broken dendritic process, cell body loss, dendritic blebbing, or a missing neuron. In total, at least 90 adult worms were analyzed for each treatment condition, at least 30 from each independent transgenic line. An average of total percentage of worms with normal neurons was reported in the study.

Reagents. Nordihydroguaiaretic acid, NDGA 74540 Aldrich; (–)- Epigallocatechin gallate, EGCG E4143 Sigma or 4524 Tocris; Terameprocol/mNDGA, T3455 Sigma; Catalase, C9322 Sigma; NDGA analogs NDGA-1, SECO-1, NDGA-5, cNDGA, and cNDGA-5 were generated, purified, and validated by Krol *et al.* at the University of Saskatchewan as previously described⁶³.

Statistical analysis. Statistical analysis was conducted using Prism 5 software (GraphPad). Unpaired students *t* test was used for comparisons between two groups. Comparisons involving multiple groups, such as molecule and dose comparisons, used two-way ANOVA with Tukey's *post hoc* test for multiple comparisons. In all cases $p < 0.05$ was considered statically significant.

Data Availability

All data generated or analyzed during this study are included in this published article (and its Supplementary Information files).

References

- Spillantini, M. G. *et al.* alpha-Synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997).
- Baba, M. *et al.* Aggregation of alpha-synuclein in Lewy bodies of sporadic Parkinson's disease and dementia with Lewy bodies. *Am. J. Pathol.* **152**, 879–84 (1998).
- Polymeropoulos, M. H. *et al.* Mutation in the α -synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045–2047 (1997).
- Krüger, R. *et al.* Ala30Pro mutation in the gene encoding alpha-synuclein in Parkinson's disease. *Nat. Genet.* **18**, 106–108 (1998).
- Chartier-Harlin, M. C. *et al.* α -Synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* **364**, 1167–1169 (2004).
- Singleton, A. B. *et al.* α -Synuclein Locus Triplication Causes Parkinson's Disease. *Science* **302**, 841 (2003).
- Wakabayashi, K., Yoshimoto, M., Tsuji, S. & Takahashi, H. Alpha-synuclein immunoreactivity in glial cytoplasmic inclusions in multiple system atrophy. *Neurosci. Lett.* **249**, 180–2 (1998).
- Spillantini, M. G., Crowther, R. A., Jakes, R., Hasegawa, M. & Goedert, M. alpha-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with lewy bodies. *Proc. Natl. Acad. Sci. USA* **95**, 6469–73 (1998).
- Gai, W. P., Power, J. H., Blumbergs, P. C. & Blessing, W. W. Multiple-system atrophy: a new alpha-synuclein disease? *Lancet (London, England)* **352**, 547–8 (1998).

10. Trojanowski, J. Q., Goedert, M., Iwatsubo, T. & Lee, V. M. Fatal attractions: abnormal protein aggregation and neuron death in Parkinson's disease and Lewy body dementia. *Cell Death Differ.* **5**, 832–7 (1998).
11. Takeda, A. *et al.* Abnormal accumulation of NACP/ α -synuclein in neurodegenerative disorders. *Am. J. Pathol.* **152**, 367–372 (1998).
12. Lücking, C. B. & Brice, A. Alpha-synuclein and Parkinson's disease. *Cell. Mol. life Sci.* **57**, 1894–908 (2000).
13. Arawaka, S., Saito, Y., Murayama, S. & Mori, H. Lewy body in neurodegeneration with brain iron accumulation type 1 is immunoreactive for alpha-synuclein. *Neurology* **51**, 887–9 (1998).
14. Wakabayashi, K., Matsumoto, K., Takayama, K., Yoshimoto, M. & Takahashi, H. NACP, a presynaptic protein, immunoreactivity in Lewy bodies in Parkinson's disease. *Neurosci. Lett.* **239**, 45–8 (1997).
15. Abeliovich, A. *et al.* Mice lacking alpha-synuclein display functional deficits in the nigrostriatal dopamine system. *Neuron* **25**, 239–52 (2000).
16. Middleton, E. R. & Rhoades, E. Effects of curvature and composition on α -synuclein binding to lipid vesicles. *Biophys. J.* **99**, 2279–2288 (2010).
17. Larsen, K. E. *et al.* Alpha-synuclein overexpression in PC12 and chromaffin cells impairs catecholamine release by interfering with a late step in exocytosis. *J. Neurosci.* **26**, 11915–22 (2006).
18. Nemani, V. M. *et al.* Increased Expression of α -Synuclein Reduces Neurotransmitter Release by Inhibiting Synaptic Vesicle Reclustering after Endocytosis. *Neuron* **65**, 66–79 (2010).
19. Murphy, D. D., Rueter, S. M., Trojanowski, J. Q. & Lee, V. M. Synucleins are developmentally expressed, and alpha-synuclein regulates the size of the presynaptic vesicular pool in primary hippocampal neurons. *J. Neurosci.* **20**, 3214–3220 (2000).
20. Greten-Harrison, B. *et al.* α -Synuclein triple knockout mice reveal age-dependent neuronal dysfunction. *Proc. Natl. Acad. Sci.* **107**, 19573–19578 (2010).
21. Scott, D. & Roy, S. α -Synuclein inhibits intersynaptic vesicle mobility and maintains recycling-pool homeostasis. *J. Neurosci.* **32**, 10129–35 (2012).
22. Giasson, B. I. *et al.* Neuronal α -synucleinopathy with severe movement disorder in mice expressing A53T human α -synuclein. *Neuron* **34**, 521–533 (2002).
23. Yang, W. *et al.* Mutant Alpha-Synuclein Causes Age-Dependent Neuropathology in Monkey Brain. *J. Neurosci.* **35**, 8345–8358 (2015).
24. Recasens, A. *et al.* Lewy body extracts from Parkinson disease brains trigger α -synuclein pathology and neurodegeneration in mice and monkeys. *Ann. Neurol.* **75**, 351–62 (2014).
25. Luk, K. C. *et al.* Pathological α -synuclein transmission initiates Parkinson-like neurodegeneration in nontransgenic mice. *Science* **338**, 949–953 (2012).
26. Sacino, A. N. *et al.* Brain Injection of α -Synuclein Induces Multiple Proteinopathies, Gliosis, and a Neuronal Injury Marker. *J. Neurosci.* **34**, 12368–12378 (2014).
27. Shimozawa, A. *et al.* Propagation of pathological α -synuclein in marmoset brain. *Acta Neuropathol. Commun.* **5**, 12 (2017).
28. Rey, N. L. *et al.* Spread of aggregates after olfactory bulb injection of α -synuclein fibrils is associated with early neuronal loss and is reduced long term. *Acta Neuropathol.* **135**, 65–83 (2018).
29. Li, J. Y. *et al.* Lewy bodies in grafted neurons in subjects with Parkinson's disease suggest host-to-graft disease propagation. *Nat. Med.* **14**, 501–503 (2008).
30. Kowall, N. W. *et al.* MPTP induces alpha-synuclein aggregation in the substantia nigra of baboons. *Neuroreport* **11**, 211–213 (2000).
31. Betarbet, R. *et al.* Chronic systemic pesticide exposure reproduces features of Parkinson's disease. *Nat. Neurosci.* **3**, 1301–1306 (2000).
32. Conway, K. A., Rochet, J. C., Bieganski, R. M. & Lansbury, J. Kinetic stabilization of the α -synuclein protofibril by a dopamine- α -synuclein adduct. *Science* **294**, 1346–1349 (2001).
33. Norris, E. H. *et al.* Reversible inhibition of α -synuclein fibrillization by dopaminochrome-mediated conformational alterations. *J. Biol. Chem.* **280**, 21212–21219 (2005).
34. Mazzulli, J. R. *et al.* Cytosolic Catechols Inhibit α -Synuclein Aggregation and Facilitate the Formation of Intracellular Soluble Oligomeric Intermediates. *J. Neurosci.* **26**, 10068–10078 (2006).
35. Meng, X., Munishkina, L. A., Fink, A. L. & Uversky, V. N. Molecular mechanisms underlying the flavonoid-induced inhibition of α -synuclein fibrillation. *Biochemistry* **48**, 8206–8224 (2009).
36. Mor, D. E. *et al.* Dopamine induces soluble α -synuclein oligomers and nigrostriatal degeneration. *Nat. Neurosci.* **20**, 1560–1568 (2017).
37. Danzer, K. M. *et al.* Different Species of α -Synuclein Oligomers Induce Calcium Influx and Seeding. *J. Neurosci.* **27**, 9220–9232 (2007).
38. Burbulla, L. F. *et al.* Dopamine oxidation mediates mitochondrial and lysosomal dysfunction in Parkinson's disease. *Science* **357**, 1255–1261 (2017).
39. Ehrnhoefer, D. E. *et al.* ECGC redirects amyloidogenic polypeptides into unstructured, off-pathway oligomers. *Nat. Struct. Mol. Biol.* **15**, 558–566 (2008).
40. Ono, K. & Yamada, M. Antioxidant compounds have potent anti-fibrillogenic and fibril-destabilizing effects for alpha-synuclein fibrils *in vitro*. *J. Neurochem.* **97**, 105–15 (2006).
41. Masuda, M. *et al.* Small molecule inhibitors of α -synuclein filament assembly. *Biochemistry* **45**, 6085–6094 (2006).
42. Meng, X., Munishkina, L. A., Fink, A. L. & Uversky, V. N. Effects of Various Flavonoids on the α -Synuclein Fibrillation Process. *Parkinsons Dis.* **2010**, 1–16 (2010).
43. Fernandes, L. *et al.* An ortho-Iminoquinone Compound Reacts with Lysine Inhibiting Aggregation while Remodeling Mature Amyloid Fibrils. *ACS Chem. Neurosci.* [acschemneuro.7b00017](https://doi.org/10.1021/acchemneuro.7b00017), <https://doi.org/10.1021/acchemneuro.7b00017> (2017).
44. Di Giovanni, S. *et al.* Entacapone and tolcapone, two catechol O-methyltransferase inhibitors, block fibril formation of α -synuclein and β -amyloid and protect against amyloid-induced toxicity. *J. Biol. Chem.* **285**, 14941–14954 (2010).
45. Lorenzen, N. *et al.* How epigallocatechin gallate can inhibit α -synuclein oligomer toxicity *in vitro*. *J. Biol. Chem.* **289**, 21299–21310 (2014).
46. Wagner, J. *et al.* Anle138b: A novel oligomer modulator for disease-modifying therapy of neurodegenerative diseases such as prion and Parkinson's disease. *Acta Neuropathol.* **125**, 795–813 (2013).
47. Deeg, A. A. *et al.* Anle138b and related compounds are aggregation specific fluorescence markers and reveal high affinity binding to α -synuclein aggregates. *Biochim. Biophys. Acta* **1850**, 1884–90 (2015).
48. Li, J., Zhu, M., Rajamani, S., Uversky, V. N. & Fink, A. L. Rifampicin inhibits alpha-synuclein fibrillation and disaggregates fibrils. *Chem. Biol.* **11**, 1513–21 (2004).
49. Konijnenberg, A. *et al.* Opposite Structural Effects of Epigallocatechin-3-gallate and Dopamine Binding to α -Synuclein. *Anal. Chem.* **88**, 8468–8475 (2016).
50. Wrasidlo, W. *et al.* A de novo compound targeting α -synuclein improves deficits in models of Parkinson's disease. *Brain* **139**, 3217–3236 (2016).
51. Perni, M. *et al.* A natural product inhibits the initiation of α -synuclein aggregation and suppresses its toxicity. *Proc. Natl. Acad. Sci.* **114**, E1009–E1017 (2017).

52. Lee, E.-N. *et al.* Phthalocyanine tetrasulfonates affect the amyloid formation and cytotoxicity of alpha-synuclein. *Biochemistry* **43**, 3704–15 (2004).
53. Fonseca-Ornelas, L. *et al.* Small molecule-mediated stabilization of vesicle-associated helical α -synuclein inhibits pathogenic misfolding and aggregation. *Nat. Commun.* **5**, 5857 (2014).
54. Danzer, K. M. *et al.* Heat-shock protein 70 modulates toxic extracellular α -synuclein oligomers and rescues trans-synaptic toxicity. *FASEB J.* **25**, 326–336 (2011).
55. Moree, B. *et al.* Small molecules detected by second-harmonic generation modulate the conformation of monomeric α -synuclein and reduce its aggregation in cells. *J. Biol. Chem.* **290**, 27582–27593 (2015).
56. Collier, T. J. *et al.* Nortriptyline inhibits aggregation and neurotoxicity of alpha-synuclein by enhancing reconfiguration of the monomeric form. *Neurobiol. Dis.* **106**, 191–204 (2017).
57. Prabhudesai, S. *et al.* A Novel 'Molecular Tweezer' Inhibitor of α -Synuclein Neurotoxicity *in Vitro* and *in Vivo*. *Neurotherapeutics* **9**, 464–476 (2012).
58. Acharya, S. *et al.* Molecular basis for preventing α -synuclein aggregation by a molecular tweezer. *J. Biol. Chem.* **289**, 10727–10737 (2014).
59. Burré, J., Sharma, M. & Südhof, T. C. α -Synuclein assembles into higher-order multimers upon membrane binding to promote SNARE complex formation. *Proc. Natl. Acad. Sci.* **111**, E4274–E4283 (2014).
60. Dettmer, U. *et al.* Loss of native α -synuclein multimerization by strategically mutating its amphipathic helix causes abnormal vesicle interactions in neuronal cells. *Hum. Mol. Genet.* **26**, 3466–3481 (2017).
61. Lou, X., Kim, J., Hawk, B. J. & Shin, Y.-K. α -Synuclein may cross-bridge v-SNARE and acidic phospholipids to facilitate SNARE-dependent vesicle docking. *Biochem. J.* **474**, 2039–2049 (2017).
62. Caruana, M. *et al.* Inhibition and disaggregation of α -synuclein oligomers by natural polyphenolic compounds. *FEBS Lett.* **585**, 1113–1120 (2011).
63. Asiamah, I., Hodgson, H. L., Maloney, K., Allen, K. J. H. & Krol, E. S. Ring substitution influences oxidative cyclisation and reactive metabolite formation of nordihydroguaiaretic acid analogues. *Bioorganic Med. Chem.* **23**, 7007–7014 (2015).
64. Mazzulli, J. R., Burbulla, L. F., Krainc, D. & Ischiropoulos, H. Detection of Free and Protein-Bound ortho-Quinones by Near-Infrared Fluorescence. *Anal. Chem.* **88**, 2399–405 (2016).
65. Natalello, A., Santambrogio, C. & Grandori, R. Are Charge-State Distributions a Reliable Tool Describing Molecular Ensembles of Intrinsically Disordered Proteins by Native MS? *J. Am. Soc. Mass Spectrom.* **28**, 21–28 (2017).
66. Ferrie, J. J. *et al.* Using a FRET Library with Multiple Probe Pairs To Drive Monte Carlo Simulations of α -Synuclein. *Biophys. J.* **114**, 53–64 (2018).
67. George, J. M., Jin, H., Woods, W. S. & Clayton, D. F. Characterization of a novel protein regulated during the critical period for song learning in the zebra finch. *Neuron* **15**, 361–372 (1995).
68. Davidson, W. S., Jonas, A., Clayton, D. F. & George, J. M. Stabilization of α -Synuclein secondary structure upon binding to synthetic membranes. *J. Biol. Chem.* **273**, 9443–9449 (1998).
69. Ulmer, T. S. & Bax, A. Comparison of structure and dynamics of micelle-bound human α -synuclein and Parkinson disease variants. *J. Biol. Chem.* **280**, 43179–43187 (2005).
70. Rhoades, E., Ramlall, T. F., Webb, W. W. & Eliezer, D. Quantification of α -synuclein binding to lipid vesicles using fluorescence correlation spectroscopy. *Biophys. J.* **90**, 4692–4700 (2006).
71. Martinez, B. A., Caldwell, K. A. & Caldwell, G. A. *C. elegans* as a model system to accelerate discovery for Parkinson disease. *Curr. Opin. Genet. Dev.* **44**, 102–109 (2017).
72. Hamamichi, S. *et al.* Hypothesis-based RNAi screening identifies neuroprotective genes in a Parkinson's disease model. *Proc. Natl. Acad. Sci. USA* **105**, 728–733 (2008).
73. Zharikov, A. D. *et al.* shRNA targeting α -synuclein prevents neurodegeneration in a Parkinson's disease model. *J. Clin. Invest.* **125**, 2721–2735 (2015).
74. Lee, H.-J. Clearance of α -Synuclein Oligomeric Intermediates via the Lysosomal Degradation Pathway. *J. Neurosci.* **24**, 1888–1896 (2004).
75. Dehay, B. *et al.* Pathogenic Lysosomal Depletion in Parkinson's Disease. *J. Neurosci.* **30**, 12535–12544 (2010).
76. Masliah, E. *et al.* Effects of alpha-synuclein immunization in a mouse model of Parkinson's disease. *Neuron* **46**, 857–68 (2005).
77. Masliah, E. *et al.* Passive immunization reduces behavioral and neuropathological deficits in an alpha-synuclein transgenic model of lewy body disease. *PLoS One* **6** (2011).
78. Levin, J. *et al.* The PROMESA-protocol: progression rate of multiple system atrophy under EGCG supplementation as anti-aggregation-approach. *J. Neural Transm.* **123**, 439–45 (2016).
79. Reznichenko, L., Kalfon, L., Amit, T., Youdim, M. B. H. & Mandel, S. A. Low dosage of rasagiline and epigallocatechin gallate synergistically restored the nigrostriatal axis in MPTP-induced parkinsonism. *Neurodegener. Dis.* **7**, 219–231 (2010).
80. Chen, M. *et al.* Tea polyphenols alleviate motor impairments, dopaminergic neuronal injury, and cerebral α -synuclein aggregation in MPTP-intoxicated parkinsonian monkeys. *Neuroscience* **286**, 383–392 (2015).
81. Thomas, B. *et al.* Resistance to MPTP-Neurotoxicity in α -synuclein knockout mice is complemented by human α -synuclein and associated with increased β -synuclein and Akt activation. *PLoS One* **6** (2011).
82. Bieschke, J. *et al.* EGCG remodels mature α -synuclein and amyloid- β fibrils and reduces cellular toxicity. *Proc. Natl. Acad. Sci.* **107**, 7710–7715 (2010).
83. Yang, J. E. *et al.* EGCG-mediated Protection of the Membrane Disruption and Cytotoxicity Caused by the 'Active Oligomer' of α -Synuclein. *Sci. Rep.* **7**, 1–10 (2017).
84. Pogačnik, L. *et al.* Potential for brain accessibility and analysis of stability of selected flavonoids in relation to neuroprotection *in vitro*. *Brain Res.* **1651**, 17–26 (2016).
85. Lambert, J. D. *et al.* Hepatotoxicity of high oral dose (–)-epigallocatechin-3-gallate in mice. *Food Chem. Toxicol.* **48**, 409–416 (2010).
86. Zhu, M. & Fink, A. L. Lipid binding inhibits α -synuclein fibril formation. *J. Biol. Chem.* **278**, 16873–16877 (2003).
87. Burré, J. *et al.* α -Synuclein promotes SNARE-complex assembly *in vivo* and *in vitro*. *Science* **329**, 1663–1667 (2010).
88. Kim, S. *et al.* GBA1 deficiency negatively affects physiological α -synuclein tetramers and related multimers. *Proc. Natl. Acad. Sci.* **115**, 201700465 (2018).
89. Dettmer, U. *et al.* Parkinson-causing α -synuclein missense mutations shift native tetramers to monomers as a mechanism for disease initiation. *Nat. Commun.* **6**, 7314 (2015).
90. Rasia, R. M. *et al.* Structural characterization of copper(II) binding to α -synuclein: Insights into the bioinorganic chemistry of Parkinson's disease. *Proc. Natl. Acad. Sci.* **102**, 4294–4299 (2005).
91. Binolfi, A. *et al.* Interaction of α -synuclein with divalent metal ions reveals key differences: A link between structure, binding specificity and fibrillation enhancement. *J. Am. Chem. Soc.* **128**, 9893–9901 (2006).
92. Appel-Cresswell, S. *et al.* Alpha-synuclein p.H50Q, a novel pathogenic mutation for Parkinson's disease. *Mov. Disord.* **28**, 811–813 (2013).
93. Proukakis, C. *et al.* A novel α -synuclein missense mutation in Parkinson disease. *Neurology* **80**, 1062–1064 (2013).
94. Ghosh, D. *et al.* The parkinson's disease-associated H50Q mutation accelerates α -synuclein aggregation *in vitro*. *Biochemistry* **52**, 6925–6927 (2013).

95. Kowalik-Jankowska, T., Rajewska, A., Jankowska, E. & Grzonka, Z. Copper(II) binding by fragments of alpha-synuclein containing M1-D2- and -H50-residues; a combined potentiometric and spectroscopic study. *Dalton Trans.* 5068–76, <https://doi.org/10.1039/b610619f> (2006).
96. Back, P. *et al.* Exploring real-time *in vivo* redox biology of developing and aging *Caenorhabditis elegans*. *Free Radic. Biol. Med.* **52**, 850–859 (2012).
97. Hammoudeh, D. I., Follis, A. V., Prochownik, E. V. & Metallo, S. J. Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-Myc. *J. Am. Chem. Soc.* **131**, 7390–401 (2009).
98. Metallo, S. J. Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol.* **14**, 481–8 (2010).
99. Dunker, A. K. & Uversky, V. N. Drugs for 'protein clouds': targeting intrinsically disordered transcription factors. *Curr. Opin. Pharmacol.* **10**, 782–8 (2010).
100. Yin, X., Giap, C., Lazo, J. S. & Prochownik, E. V. Low molecular weight inhibitors of Myc-Max interaction and function. *Oncogene* **22**, 6151–9 (2003).
101. Wang, H. *et al.* Improved low molecular weight Myc-Max inhibitors. *Mol. Cancer Ther.* **6**, 2399–408 (2007).
102. Theillet, F. X. *et al.* Structural disorder of monomeric alpha-synuclein persists in mammalian cells. *Nature* **530**, 45–50 (2016).
103. Winner, B. *et al.* *In vivo* demonstration that α -synuclein oligomers are toxic. *Proc. Natl. Acad. Sci.* **108**, 4194–4199 (2011).
104. Favier, A. & Brutscher, B. Recovering lost magnetization: Polarization enhancement in biomolecular NMR. *J. Biomol. NMR* **49**, 9–15 (2011).
105. Brutscher, B. *et al.* NMR Methods for the Study of Intrinsically Disordered Proteins Structure, Dynamics, and Interactions: General Overview and Practical Guidelines. *Adv. Exp. Med. Biol.* **870**, 49–122 (2015).
106. Huang, C., Ren, G., Zhou, H. & Wang, C. C. A new method for purification of recombinant human α -synuclein in *Escherichia coli*. *Protein Expr. Purif.* **42**, 173–177 (2005).
107. Berkowitz, L. A., Knight, A. L., Caldwell, G. A. & Caldwell, K. A. Generation of Stable Transgenic *C. elegans* Using Microinjection. *J. Vis. Exp.* 3–5, <https://doi.org/10.3791/833> (2008).

Acknowledgements

The authors thank Esteban Luna and Kelvin Luk – and the University of Pennsylvania Center for Neurodegenerative Disease Research more broadly – for generously sharing their time and reagents. The authors would like to gratefully acknowledge the support and use of resources of the CERM/CIRMMMP center of INSTRUCT-ERIC, a Landmark ESFRI project. M.D. was supported, in part, by the NIH T32 GM008076 training grant for Pharmacology. J.J.F. thanks NSF (DGE-1321851) and the Parkinson's Disease Foundation (PF-RVSA-SFW-1754) for fellowship support. C.M.H. was supported, in part, by the NIH T32 AG000255 training grant for Age Related Neurodegenerative Disease. A.N. and R.G. were supported by COST Action BM1403. E.J.P. was supported by the NIH (NS081033). E.R. was supported by the NIH NINDS (R01 NS079955). R.P. was supported, in part, by a grant of the Fondazione CR Firenze. E.K. was supported by Natural Sciences and Engineering Research Council Discovery Grant (NSERC-DG) 165915. H.I. is the Gisela and Dennis Alter research professor at the Children's Hospital of Philadelphia Research Institute. The funding sources were not involved in the study design, data interpretation, analysis or submission of this manuscript.

Author Contributions

M.D. designed, conducted, and analyzed the α -synuclein aggregation experiments. M.D. analyzed all data and wrote the manuscript. J.B.N., H.K., K.C. and G.C. designed, conducted, analyzed, and wrote description of the *C. elegans* experiments and assisted in editing the manuscript. V.S., M.S., M.G.M., R.P. and I.F. designed, conducted, analyzed, and wrote description of the N.M.R. experiments and assisted in editing the manuscript. B.P. conducted, analyzed, and wrote description of the FCS. experiments and assisted in editing the manuscript. J.J.F. conducted, analyzed, and wrote description of the FRET experiments. C.H. generated some of the labeled α -synuclein used in the FRET experiments. E.R. and E.J.P. assisted in design and analysis of the FCS and FRET experiments and assisted in editing the manuscript. R.M. and F.S. conducted, analyzed, and wrote description of the IM-MS experiments and assisted in editing the manuscript. A.N. and R.G. conducted, analyzed, and wrote description of the ESI-MS experiments and assisted in editing the manuscript. N.G. and V.U. assisted in writing and editing the manuscript. E.K. generated the cyclized NDGA analogs and assisted in experimental design, writing, and editing of the manuscript. H.I. planned, organized, and participated in the analysis and discussion of all the data and editing of the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-39480-z>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



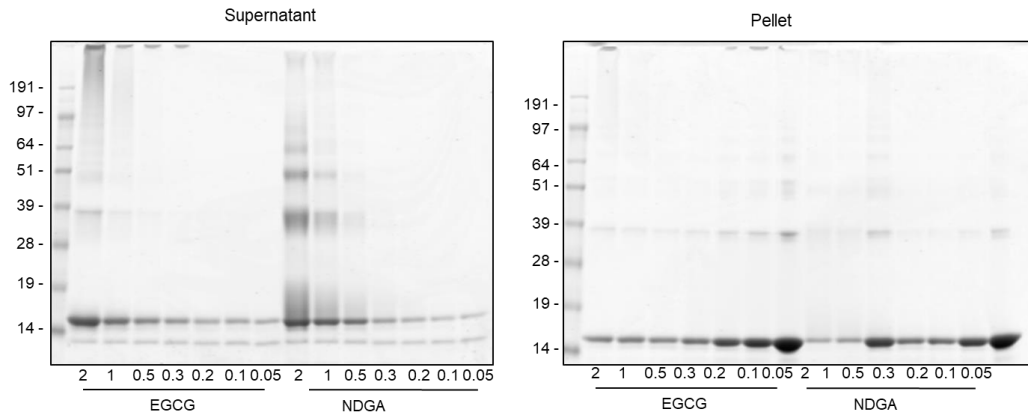
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Supplementary Figures

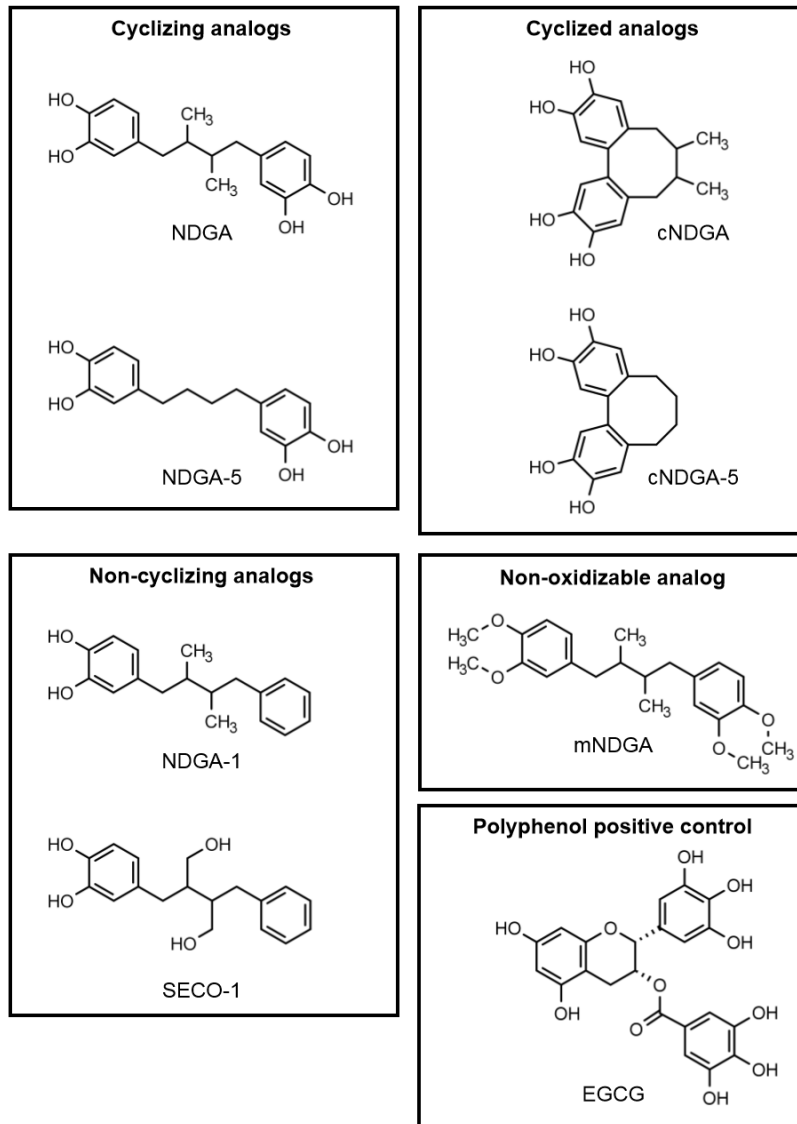
Title: Cyclized NDGA modifies dynamic α -synuclein monomers preventing aggregation and toxicity

Authors: Malcolm J. Daniels, J. Brucker Nourse Jr., Hanna Kim, Valerio Sainati, Marco Schiavina, Maria Grazia Murrari, Buyan Pan, John J. Ferrie, Conor M. Haney, Rani Moons, Neal S. Gould, Antonino Natalello, Rita Grandori, Frank Sobott, E. James Petersson, Elizabeth Rhoades, Roberta Pierattelli, Isabella Felli, Vladimir N. Uversky, Kim A. Caldwell, Guy A. Caldwell, Edward S. Krol and Harry Ischiropoulos

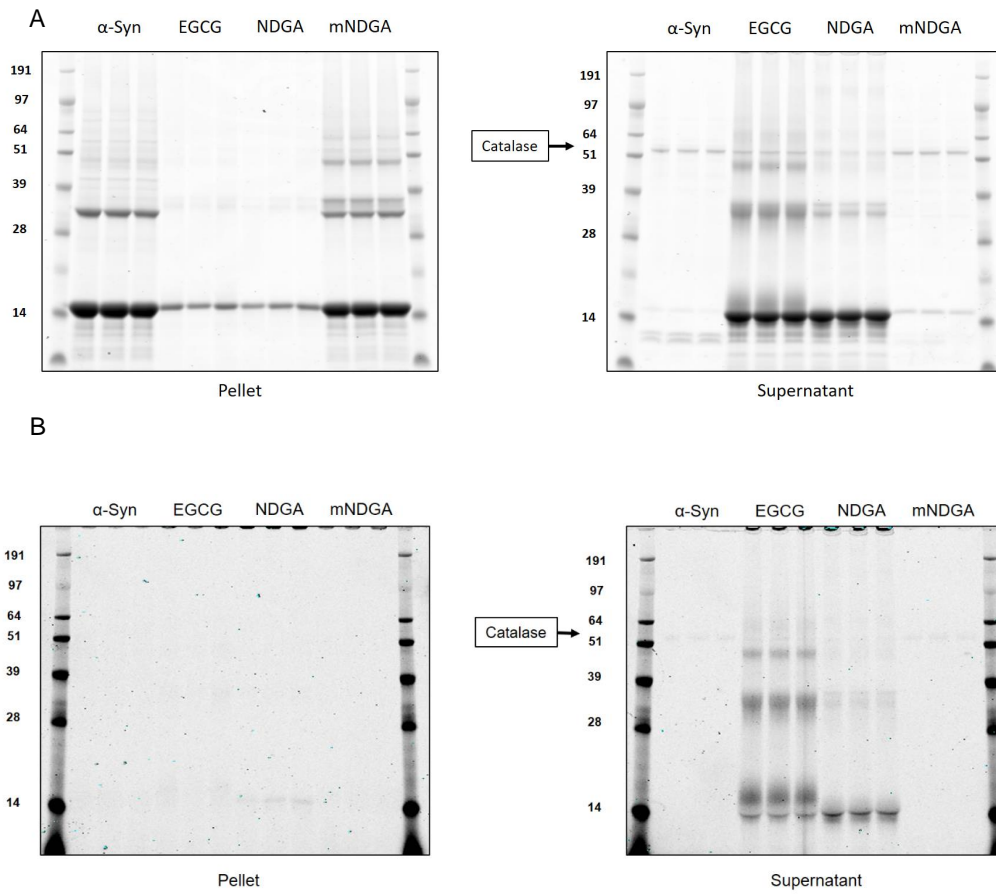


Supplementary Figure S1. Dose effect of EGCG or NDGA on α -synuclein aggregation.

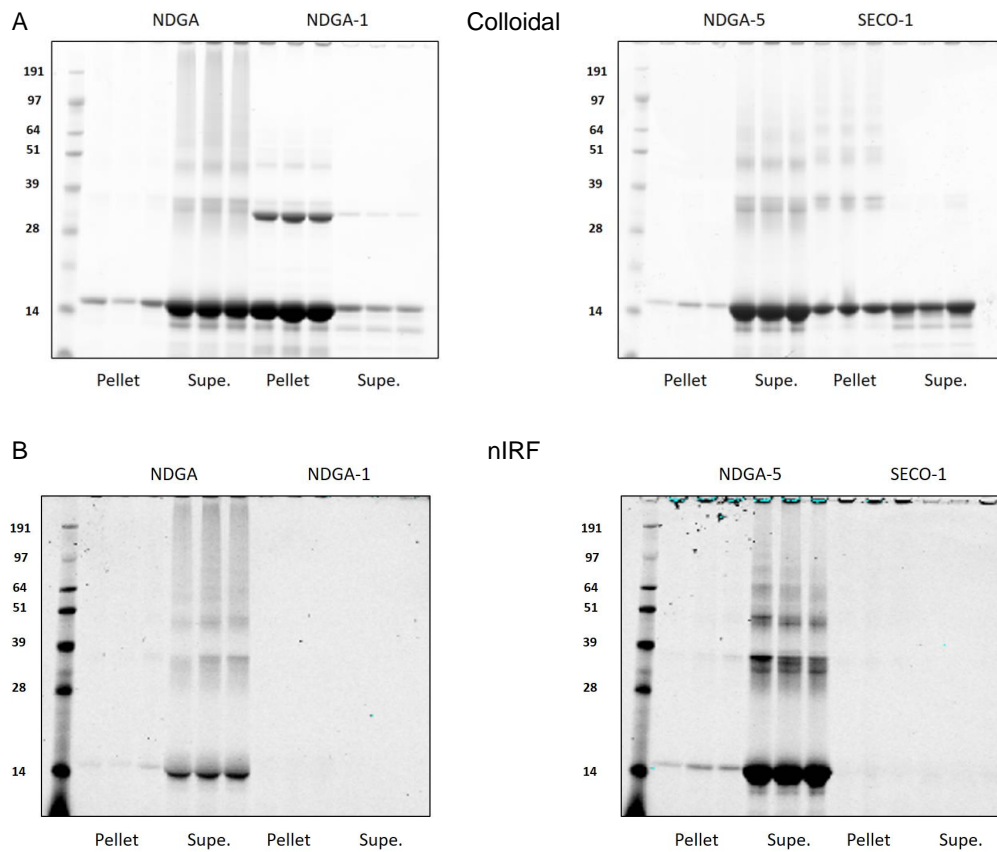
Representative images of recombinant human wildtype α -synuclein ($138 \mu\text{M}$) aggregated for 7 days in the presence of EGCG or NDGA at the indicated molar ratios. After aggregation, PBS-insoluble α -synuclein was separated by centrifugation (21k g for 10 min). Soluble (supernatant) and insoluble (pellet) fractions were boiled in SDS, run by SDS-PAGE, and colloidal stained. α -Synuclein in each fraction was quantified by in-gel densitometry. ($n = 3-5$).



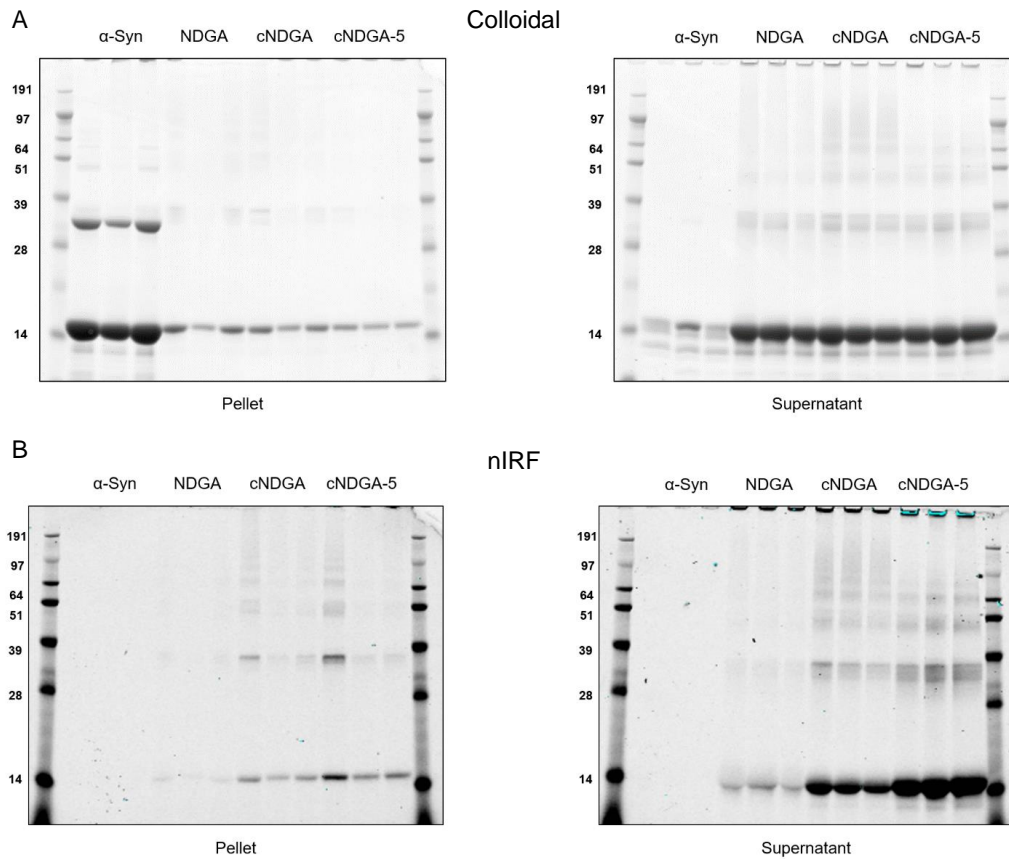
Supplementary Figure S2. Structures of EGCG, NDGA, and NDGA analogs employed in this study.



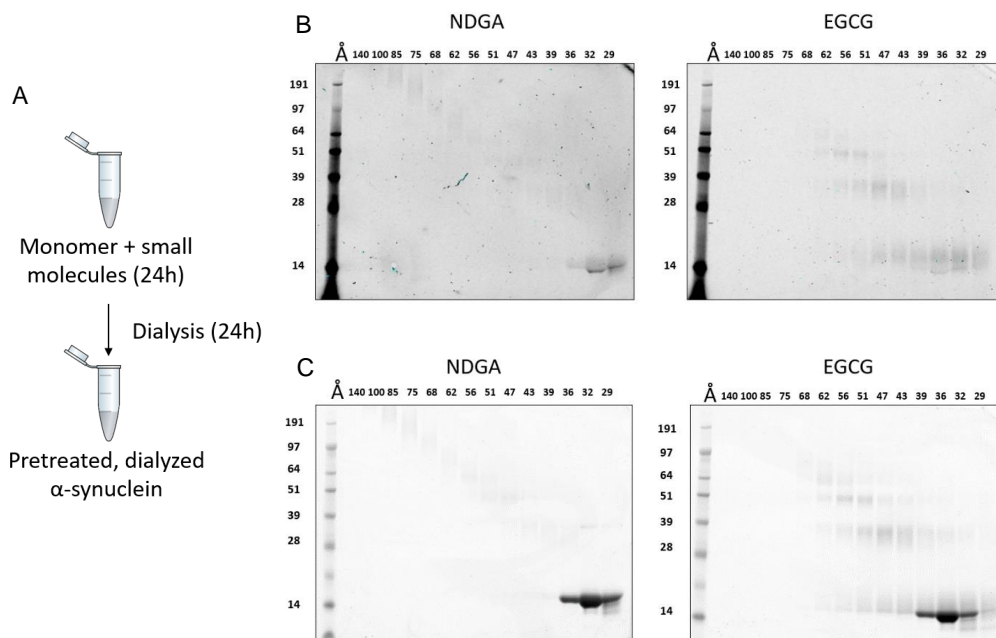
Supplementary Figure S3. Colloidal and nIRF images of α -synuclein aggregation in the presence of 5% catalase. **(a)** Colloidal stained gel **(b)** nIRF image of the same gel before colloidal staining, n=3.



Supplementary Figure S4. Colloidal and nIRF images of α -synuclein aggregation in the presence of NDGA analogs. **(a)** Colloidal images of gels after separation by centrifugation. **(b)** Near-infrared images of the same gels before colloidal staining, $n=3$.

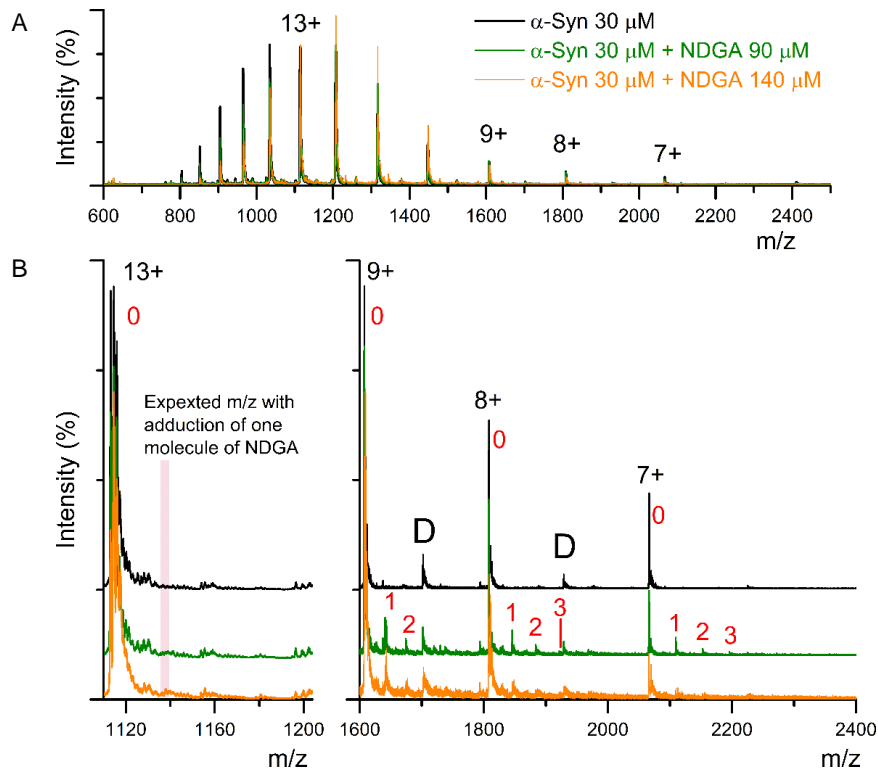


Supplementary Figure S5. Colloidal and nIRF images of α -synuclein aggregation in the presence of cyclized NDGA analogs. **(a)** Colloidal images of gels after separation by centrifugation. **(b)** Near-infrared images of the same gels before colloidal staining, $n=3$.

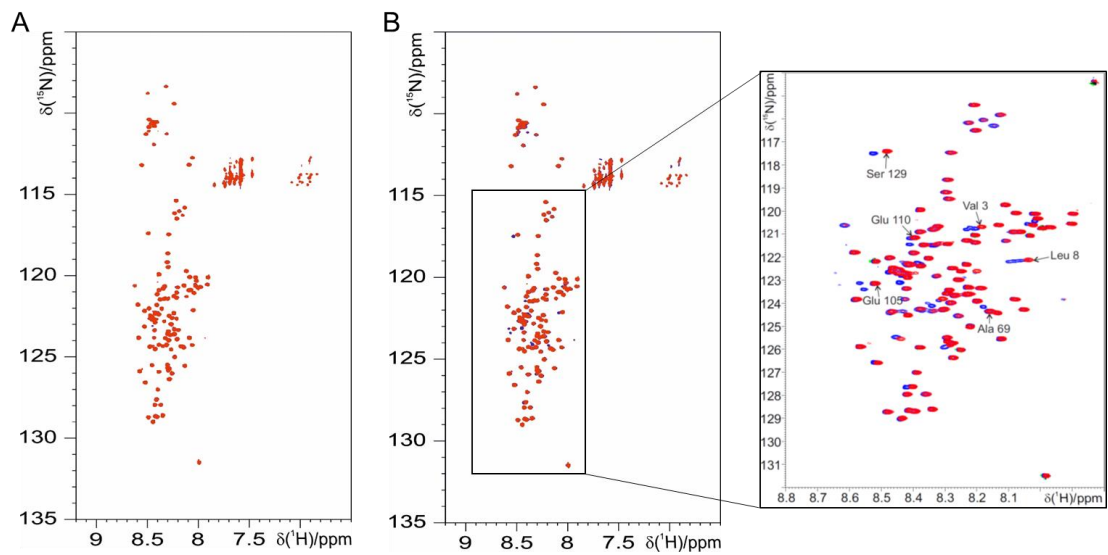


Supplementary Figure S6. NDGA pretreatment produces nIRF positive α -synuclein monomers.

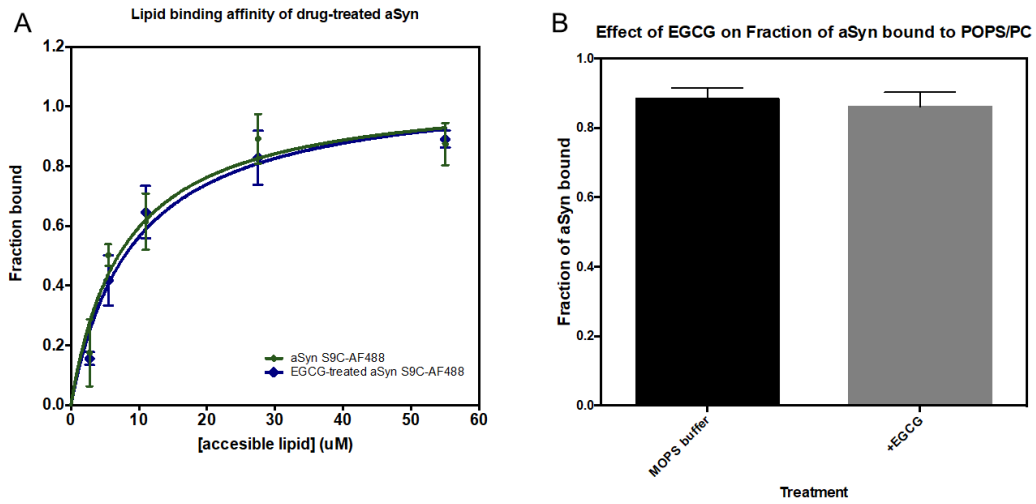
(a) Workflow for CD analysis and aggregation. (b) Near-infrared fluorescence images of NDGA and EGCG pretreated, dialyzed α -synuclein separated by native-state size-exclusion chromatography. (c) Colloidal staining of the same gels shown in panel b.



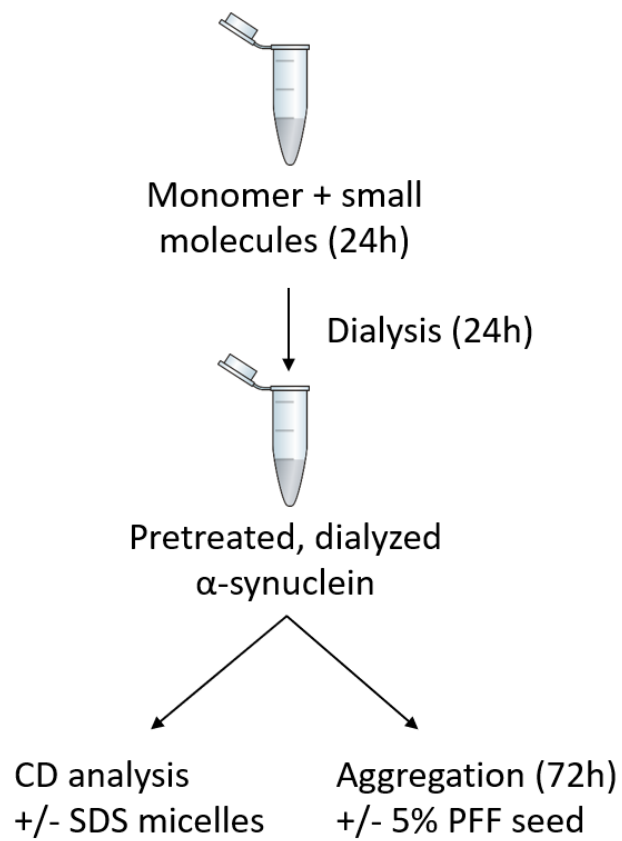
Supplementary Figure S7. Electrospray mass spectrometry characterization of α -synuclein-NDGA interaction. (a) Nano-ESI-MS spectra in positive-ion mode of 30 μ M α -synuclein in the absence and presence of NDGA at 90 μ M and 140 μ M. (b) Magnification of the spectra in the Upper panel. The number of NDGA molecules bound to α -synuclein monomers is indicated by red numbers. Dimer-specific peaks of α -synuclein are labeled as "D". Charge states are indicated by black numbers. Nano-ESI-MS spectra were collected as described in Konijnenberg et al.⁴⁹



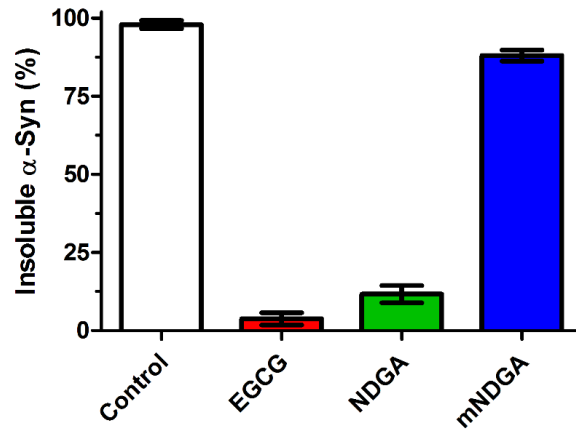
Supplementary Figure S8. 2D NMR spectra comparing α -synuclein treated with 1:1 NDGA for 24 hours (red), with (a) α -synuclein treated with 3:1 NDGA for 24 hours (blue), and (b) α -synuclein treated with 1:1 NDGA for 10 days (blue).



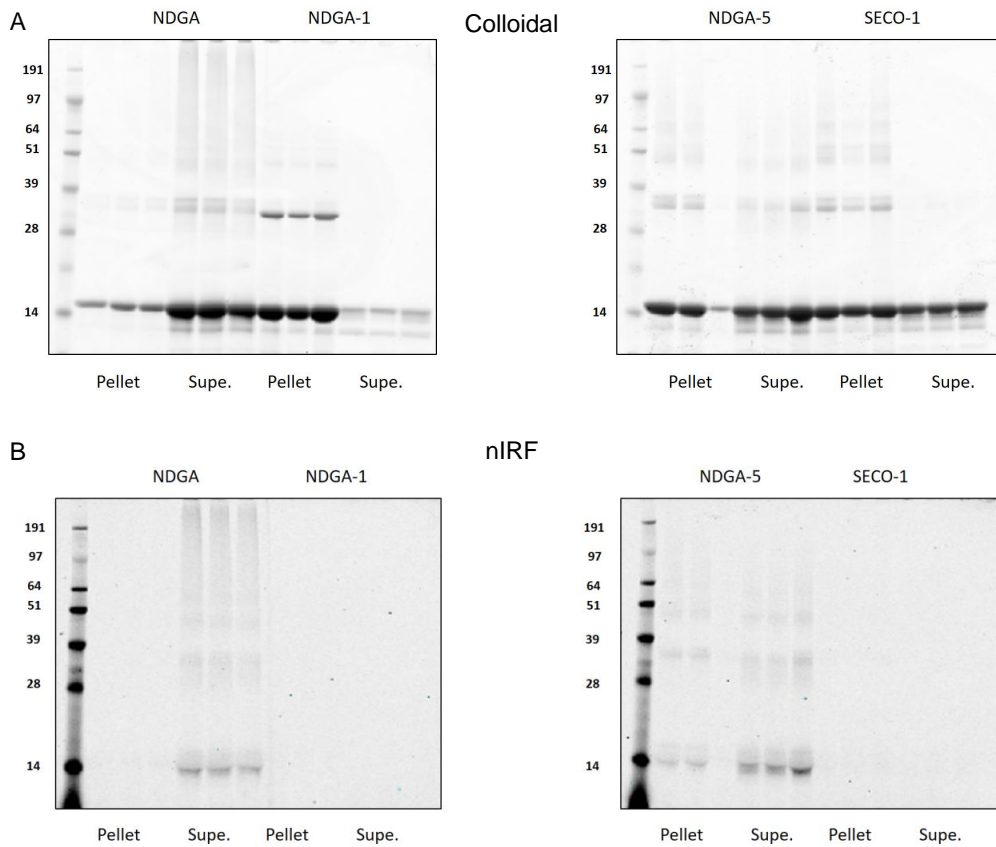
Supplementary Figure S9. EGCG does not disrupt lipid binding. (A) α -Synuclein was incubated 1:1 with EGCG or solvent alone for 24 hours before fluorescence correlation spectroscopy in the presence of POPS:POPC vesicles at the indicated concentrations. (n = 3). (B) Addition of NDGA did not displace fluorescently labeled α -synuclein from POPS:POPC vesicles. (n = 3).



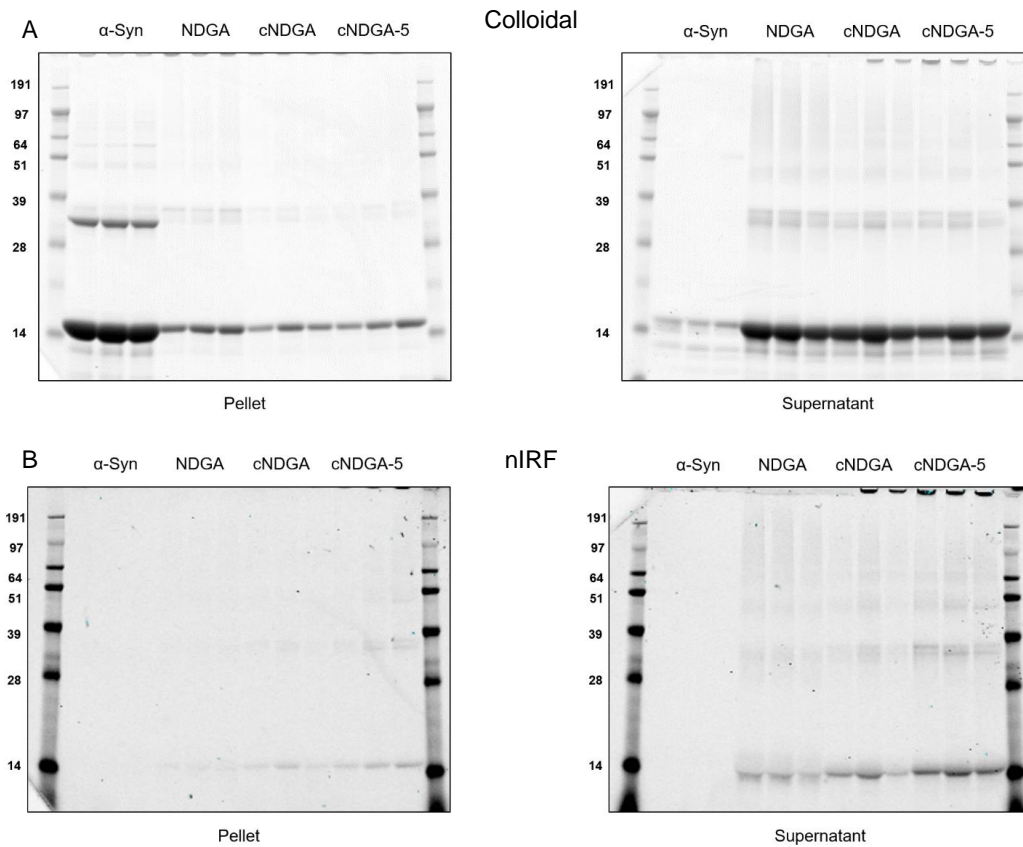
Supplementary Figure S10. Workflow for analysis of pretreated, dialyzed α -synuclein



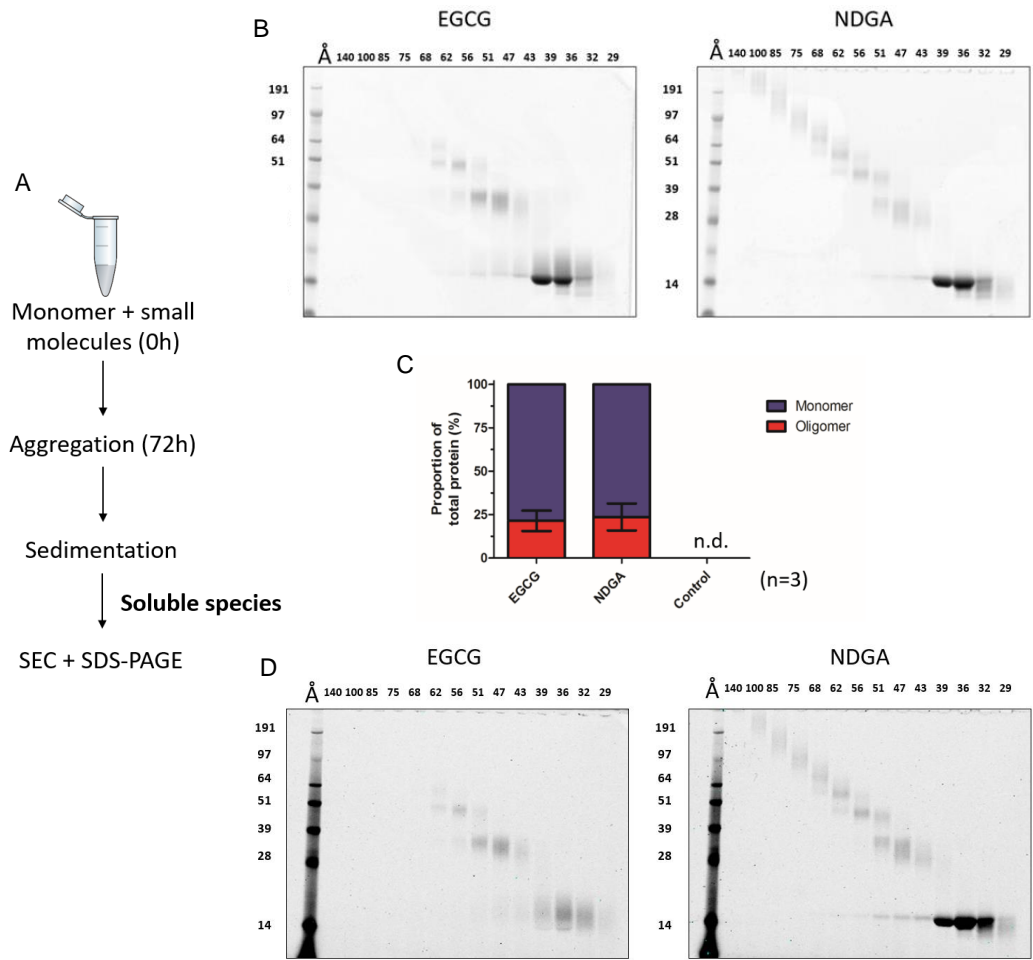
Supplementary Figure S11. NDGA pretreatment inhibits α -synuclein aggregation despite 14 days under aggregation conditions. α -Synuclein was incubated 1:1 with small molecules for 24 hours then dialyzed against PBS for 24 hours. After aggregation for 14 days, PBS-insoluble α -synuclein was separated by centrifugation (21k g for 10 min). Soluble and insoluble fractions were boiled in SDS, run by SDS-PAGE, and colloidal stained. α -Synuclein in each fraction was quantified by in-gel densitometry. (n = 3)



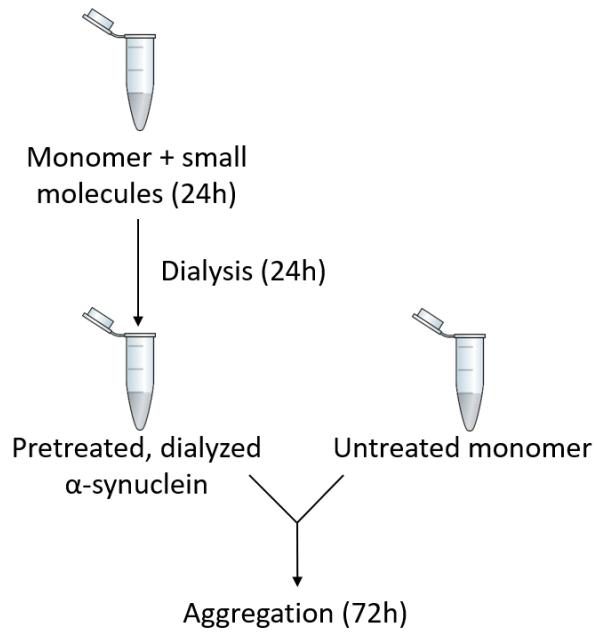
Supplementary Figure S12. Colloidal and nIRF images of NDGA analog pretreated, dialyzed α -synuclein aggregation **(a)** Colloidal images of gels after separation by centrifugation. **(b)** Near-infrared images of the same gels before colloidal staining, $n=3$.



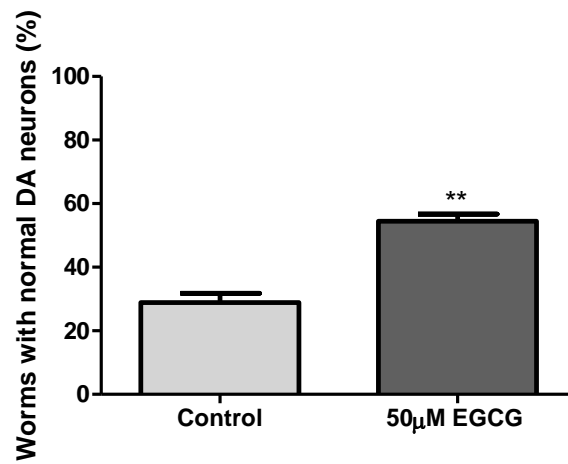
Supplementary Figure S13. Colloidal and nIRF images of cyclized NDGA analog pretreated, dialyzed α -synuclein aggregation **(a)** Colloidal images of gels after separation by centrifugation. **(b)** Near-infrared images of the same gels before colloidal staining, n=3.



Supplementary Figure S14. nIRF positive monomers are the predominant product of α -synuclein aggregation in the presence of EGCG and NDGA. (a) Workflow (b) Colloidal (c) Densitometry (d) nIRF (n=3).



Supplementary Figure S15. Workflow for aggregation of combined mixtures of pretreated, dialyzed α -synuclein and untreated α -synuclein.



Supplementary Figure S16. EGCG reduces α -synuclein-driven neurodegeneration. *C. elegans* were treated with vehicle (dH₂O) or 50µM EGCG in NGM agar on days 0-3, 5, and 7. Dopamine neurons were scored on day 8. Worms were scored as normal if no degenerative phenotypes (broken dendritic process, cell body loss, dendritic blebbing, or a missing neuron) were observed. Data represented as mean \pm SEM.; n=3 replicates with 30 worms per replicate; t-test; **p= 0.0023

7. Taking simultaneous snapshots of intrinsically disordered proteins in Action

Biophysical *Journal*
Article



Taking Simultaneous Snapshots of Intrinsically Disordered Proteins in Action

Marco Schiavina,¹ Maria Grazia Murrari,¹ Letizia Pontoriero,¹ Valerio Sainati,¹ Rainer Kümmerle,² Wolfgang Bermel,³ Roberta Pierattelli,^{1,*} and Isabella C. Felli^{1,*}

¹Magnetic Resonance Center and Department of Chemistry "Ugo Schiff," University of Florence, Sesto Fiorentino, Florence, Italy; ²Bruker BioSpin AG, Fällanden, Switzerland; and ³Bruker BioSpin GmbH, Rheinstetten, Germany

ABSTRACT Intrinsically disordered proteins (IDPs) as well as intrinsically disordered regions (IDRs) of complex protein machineries have recently been recognized as key players in many cellular functions. NMR represents a unique tool to access atomic resolution structural and dynamic information on highly flexible IDPs/IDRs. Improvements in instrumental sensitivity made heteronuclear direct detection possible for biomolecular NMR applications. The CON experiment has become one of the most useful NMR experiments to get a snapshot of an IDP/IDR in conditions approaching physiological ones. The availability of NMR spectrometers equipped with multiple receivers now enables the acquisition of several experiments simultaneously instead of one after the other. Here, we propose several variants of the CON experiment in which, during the recovery delay, a second two-dimensional experiment is acquired, either based on ¹H detection (CON/HN) or on ¹⁵N detection (CON/btNH, CON/(H)CAN). The possibility to collect simultaneous snapshots of an IDP/IDR through different two-dimensional spectra provides a novel tool to follow chemical reactions, such as the occurrence of posttranslational modifications, as well as to study samples of limited lifetime such as cell lysates or whole cells.

Taking Simultaneous Snapshots of Intrinsically Disordered Proteins in Action

Marco Schiavina,¹ Maria Grazia Murrari,¹ Letizia Pontoriero,¹ Valerio Sainati,¹ Rainer Kümmerle,² Wolfgang Bermel,³ Roberta Pierattelli,^{1,*} and Isabella C. Felli^{1,*}

¹Magnetic Resonance Center and Department of Chemistry "Ugo Schiff," University of Florence, Sesto Fiorentino, Florence, Italy; ²Bruker BioSpin AG, Fällanden, Switzerland; and ³Bruker BioSpin GmbH, Rheinstetten, Germany

ABSTRACT Intrinsically disordered proteins (IDPs) as well as intrinsically disordered regions (IDRs) of complex protein machineries have recently been recognized as key players in many cellular functions. NMR represents a unique tool to access atomic resolution structural and dynamic information on highly flexible IDPs/IDRs. Improvements in instrumental sensitivity made heteronuclear direct detection possible for biomolecular NMR applications. The CON experiment has become one of the most useful NMR experiments to get a snapshot of an IDP/IDR in conditions approaching physiological ones. The availability of NMR spectrometers equipped with multiple receivers now enables the acquisition of several experiments simultaneously instead of one after the other. Here, we propose several variants of the CON experiment in which, during the recovery delay, a second two-dimensional experiment is acquired, either based on ¹H detection (CON//HN) or on ¹⁵N detection (CON//btNH, CON//(H)CAN). The possibility to collect simultaneous snapshots of an IDP/IDR through different two-dimensional spectra provides a novel tool to follow chemical reactions, such as the occurrence of posttranslational modifications, as well as to study samples of limited lifetime such as cell lysates or whole cells.

SIGNIFICANCE The exploitation of multiple receivers available in most of the newly designed NMR spectrometers enables the simultaneous collection of several snapshots of the very same event from different atomic perspectives.

INTRODUCTION

Proteins or protein regions that lack a well-defined three-dimensional (3D) structure and that are characterized by high flexibility and disorder are widespread in living organisms and essential for protein function (1–3). Their contributions to biological processes are highly complementary to those typical of folded proteins. Indeed, intrinsically disordered proteins (IDPs) or intrinsically disordered protein regions (IDRs) are often involved in key regulatory processes for which the adaptability of the protein structure and dynamics represents a clear functional advantage (4). Their important functional role also becomes evident from the strong link that has been found between malfunction in IDPs/IDRs and many diseases such as cancer and neurodegenerative diseases. The atomic-level characterization of IDPs/IDRs has thus become a topic of central importance

also in light of the development of new drugs capable of interfering with them, a completely novel area for which the traditional approach of drug design based on the identification of well-defined binding pockets in folded proteins is obviously bound to fail (5–8).

NMR plays a central role in this area of research being the only method able to provide atomic resolution information on the structural and dynamic properties of highly flexible macromolecules (9–13). NMR observables are, however, strongly influenced by the peculiar properties of IDPs, and thus the NMR approaches should be optimized to overcome critical points (14). For example, it is well known that ¹H-detected experiments, usually the first choice on the grounds of instrumental sensitivity, do have some drawbacks for the study of IDPs, in particular when approaching physiological pH and temperature conditions. In fact, the largely solvent-exposed backbones typical of highly flexible IDPs leave amide protons accessible to the solvent and are responsible for efficient chemical exchange processes that may lead to extreme broadening of H^N resonances, reducing the sensitivity of ¹H-detected experiments. In addition, ¹H nuclear

Submitted March 29, 2019, and accepted for publication May 14, 2019.

*Correspondence: roberta.pierattelli@unifi.it or felli@cerm.unifi.it

Editor: David Eliezer.

<https://doi.org/10.1016/j.bpj.2019.05.017>

© 2019 Biophysical Society.



spins are characterized by a moderate chemical shift dispersion. These are two features that have stimulated the development of exclusively heteronuclear NMR experiments based on ^{13}C direct detection for the study of IDPs. The two-dimensional (2D) CON experiment (15,16) is now widely used to characterize highly flexible IDPs, thanks to the excellent chemical shift dispersion of the crosspeaks observed in this experiment and the possibility to reveal atomic resolution information on IDPs also in experimental conditions in which H^{N} resonances are not observable (17–28). The wide use of exclusively heteronuclear NMR experiments for protein investigations has been stimulated by improvements in instrumental sensitivity (29). Technological advances provided a jump of a factor of more than 10 in ^{13}C instrumental sensitivity, and lately, similar efforts have even enabled direct detection of ^{15}N (30–33), a nucleus characterized by a lower gyromagnetic ratio with respect to ^{13}C and thus by a very low intrinsic sensitivity.

In parallel to significant increase in instrumental sensitivity, new technologies have also become easily accessible such as the possibility to exploit NMR spectrometers with multiple receivers, enabling the design of multinuclear pulse sequences for the acquisition of more than one free induction decay (FID) for each repetition. Several different strategies exploiting multiple receivers have been proposed in the literature (34–39) to acquire more experiments in the time needed for a single one. Applications of multiple receivers for biomolecular NMR experiments were also proposed either in solution or in the solid state (36–41). One strategy consists of the acquisition of different FIDs within one main coherence transfer pathway to recover signals that would otherwise be lost or suppressed (36,37). The other strategy consists in exploiting polarization that remains unused at the end of one experiment to collect more than one spectrum in the time needed for the longest one of them (38).

The development of more sensitive instruments and the possibility to use multiple receivers stimulates the design of novel experimental strategies to investigate IDPs, and we will show some examples of the useful information that can be gathered combining the CON with other experimental schemes.

MATERIALS AND METHODS

Samples preparation

The ^{13}C , ^{15}N -labeled α -synuclein sample was prepared as previously described in the literature (42), with a final concentration of 1.0 mM in 100 mM NaCl, 50 μM ethylenediaminetetraacetic acid (EDTA), and 20 mM phosphate buffer at different pH between 6.5 and 7.4 with the addition of 2–10% D_2O for the lock. Each figure reports the details about sample conditions. The ^{13}C , ^{15}N -labeled α -synuclein samples used for in-cell experiments and for experiments on cell lysate were prepared as previously described (43) with an estimated protein concentration of 250 μM , using the same buffer as for the purified samples at pH 7.4.

For the phosphorylation reaction, 700 U of tyrosine kinase Fyn (Sigma Aldrich, St. Louis, MO) were added to a 200 μM sample of ^{13}C , ^{15}N -labeled

α -synuclein in 20 mM phosphate buffer, 100 mM NaCl, 50 μM EDTA, 2 mM dithiothreitol, 6 mM MgCl_2 , and 3 mM ATP buffer at pH 7.0 for a final volume of 300 μL , with 5% D_2O added for the lock.

NMR experiments

The NMR experiments were acquired at different temperatures on a Bruker AVANCE NEO spectrometer (Billerica, MA) operating at 700.06 MHz ^1H , 176.05 MHz ^{13}C , and 70.97 MHz ^{15}N frequencies equipped with a cryogenically cooled probehead optimized for ^{13}C -direct detection (TXO). Standard radio frequency pulses and carrier frequencies for triple resonance experiments were used and are summarized hereafter. Q5- and Q3-shaped pulses (44) of durations of 300 and 231 μs , respectively, were used for ^{13}C band-selective $\pi/2$ and π flip angle pulses except for the π pulses that should be band selective on the C^α region (Q3, 1200 μs) and for the adiabatic π pulse to invert both C' and C^α (smoothed chirp 500 μs , 20% smoothing, 80 kHz sweep width, 11.3 kHz radio frequency field strength) (45). An Eburp2 shape of duration of 1.768 ms and a Reburp shape of duration of 2.076 ms were used, respectively, for ^1H band-selective $\pi/2$ and π flip angle pulses. In the CON/HN experiment, solvent suppression was achieved through the 3:9:19 pulse scheme (46).

The ^{13}C band-selective pulses on C^α and C' were applied at the center of each region, respectively, and the ^1H -band-selective pulses were applied at the center of the H^{N} region at 8.13 ppm. Decoupling of $^{13}\text{C}^\alpha$ and ^{15}N was achieved with p5m4sp180 (Q3, 900 μs) and Waltz65 (100 μs) decoupling sequences, respectively (44,47). All gradients employed had a smoothed square shape.

All the spectra were acquired, processed, and analyzed by using Bruker TopSpin 4.0.1 software. Calibration of the spectra was achieved using 4,4-dimethyl-4-silapentane-1-sulfonic acid as a standard for ^1H and ^{13}C ; ^{15}N shifts were calibrated indirectly (48).

The CON/HN was acquired with a CON interscan delay of 1.7 s; the HN was acquired within this delay. For each increment of the CON experiment, the in-phase (IP) and antiphase (AP) components were acquired and properly combined to achieve IPAP virtual decoupling (49) as described in Fig. S1. The CON spectrum was acquired with two scans, with sweep widths of 5555 Hz (^{13}C) \times 2500 Hz (^{15}N) and 1024 \times 512 real points in the two dimensions, respectively. The HN spectrum was acquired with 4 scans (2 scans as used for the CON \times 2 additional scans because no IPAP decoupling is necessary for the HN), with sweep widths of 10,869 Hz (^1H) \times 2000 Hz (^{15}N) and 1536 \times 512 real points in the two dimensions, respectively. The total duration of the experiment was 1 h and 7 min. For comparison purposes, the two independent experiments with the same parameters were also acquired.

Similar parameters were used also for CON/HN experiments acquired for cell lysates and for in-cell samples as well as to follow the phosphorylation reaction, with the main difference being the use of four scans in the latter case because of lower sample concentration.

The CON/btNH was acquired with a CON interscan delay of 2.5 s; during this time, the btNH was repeated four times (recovery delay of 400 ms). For each increment of each experiment, the IP and AP components were acquired and properly combined to achieve IPAP virtual decoupling (49) as described in Fig. S2. The CON spectrum was acquired with 4 scans, with sweep widths of 5263 Hz (^{13}C) \times 2403 Hz (^{15}N) and 1024 \times 1024 real points in the 2 dimensions, respectively. The btNH spectrum was acquired with 128 scans (4 scans as used for the CON \times 4 repetitions of the btNH experiment \times 8 because of the different number of increments needed in the indirect dimension), with sweep widths of 5263 Hz (^{15}N) \times 2403 Hz (^1H) and 1024 \times 128 real points in the two dimensions, respectively. The total duration of the experiment was 6 h and 19 min. For comparison purposes, the two independent experiments with the same parameters were also acquired.

The CON/(H)CAN was acquired with a CON interscan delay of 2.3 s; during this time, the (H)CAN was repeated three times (recovery delay of 700 ms). For each increment of each experiment, the IP and AP

components were acquired and properly combined to achieve IPAP virtual decoupling (49) as described in Fig. S3. The CON spectrum was acquired with 8 scans, with sweep widths of 5555 ($^{13}\text{C}'$) \times 2554 Hz (^{15}N) and 1024 \times 2048 real points in the two dimensions, respectively. The (H) CAN spectrum was acquired with 192 scans (8 scans as used for the CON \times 3 repetitions of the (H)CAN experiment \times 8 because of the different number of increments needed in the indirect dimension), with sweep widths of 5555 Hz (^{15}N) \times 4000 Hz (^{13}C) and 1024 \times 256 real points in the 2 dimensions, respectively. The total duration of the experiment was 60 h and 10 min. For comparison purposes, the two independent experiments with the same parameters were also acquired.

RESULTS AND DISCUSSION

The 2D CON has by now become one of the key experiments to characterize IDPs. Many experimental variants of the basic CON pulse sequence have been developed (43,50). Important features include the nuclear spins used as starting polarization source as well as the approach employed for homonuclear decoupling. When working with sample conditions approaching the “physiological” ones, like in the presence of high ionic strength, pH above 7, and relatively high temperature, the most preferred scheme is the simplest one, which starts on ^{13}C and actively exploits heteronuclear spins only. However, it requires a relaxation delay of ~ 2 s to allow the C' magnetization to relax back to equilibrium. Can we profitably use this delay to collect additional experiments?

The CON//HN experiment

The first two spectra generally collected on an IDP are the 2D HN and 2D CON. The experimental variant designed to acquire them simultaneously is here referred to as the CON//HN, and the spectra obtained on α -synuclein are shown in Fig. 1. The pulse sequence is reported in Fig. S1 and sketched in Fig. 1 A. In brief, a very simple logic was followed in the design of the pulse sequence. The CON experiment essentially exploits only two backbone nuclear spins (C' and N). During the relaxation delay, an additional experiment can be acquired provided that different nuclear spins are perturbed and that it does not interfere with the CON itself. The ^1H - ^{15}N heteronuclear single-quantum coherence (HSQC) exploits the amide proton as a starting polarization source as well as for acquisition of the FID; ^{15}N chemical shifts are sampled in the indirect dimension, just like in the CON experiment, and this does not cause any interference. Thus, it can be profitably acquired during the 2-s recycle delay of the CON. The only compromise to combine the two experiments derives from the need to decouple ^{15}N from ^1H and ^{13}C during the chemical shift evolution period. For the CON experiment, it is very important to use composite pulse decoupling of ^1H throughout the whole pulse sequence to minimize possible influence of exchange processes of amide protons with the solvent protons, which reduce signal intensities in particular when approach-

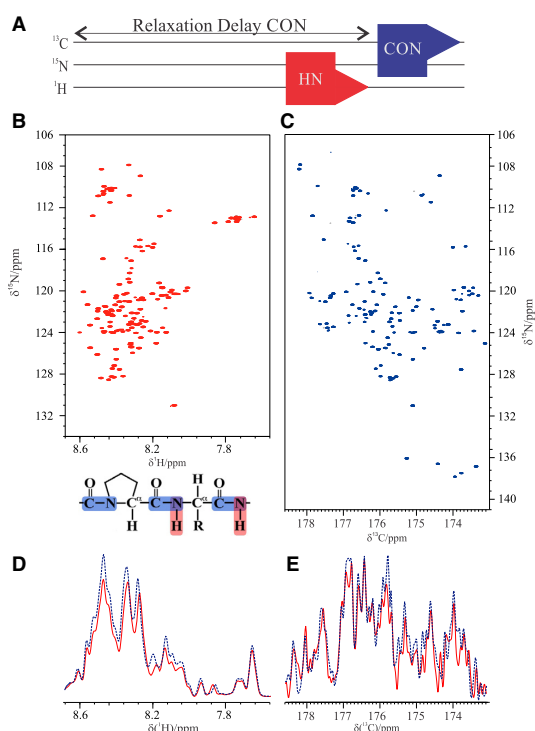


FIGURE 1 (A) Scheme of the CON//HN experiment. (B) HN and (C) CON 2D spectra acquired through the CON//HN experiment on ^{13}C , ^{15}N -labeled α -synuclein 1 mM at 285 K in 100 mM NaCl, 50 μM EDTA, and 20 mM phosphate buffer (pH 6.5) are shown. (D) A projection of the HN and the (E) CON spectra acquired with the CON//HN experiment (red solid traces) and with the corresponding experiments acquired independently (blue dotted traces) is shown. To see this figure in color, go online.

ing physiological conditions. This does not constitute a problem because the ^1H - ^{15}N HSQC pulse sequence precedes the CON one. Decoupling of ^{15}N from ^{13}C during the ^{15}N evolution period of the ^1H - ^{15}N HSQC could be achieved by a combination of two 180° pulses in appropriate positions to restore C' magnetization to equilibrium as a starting polarization source for the CON experiment. Thanks to this solution, the two spectra can be acquired simultaneously in the optimal conditions required for each of them as one can see by inspecting the quality of the spectra reported in Fig. 1. These were acquired on α -synuclein at 285 K and pH 6.5, which are conditions useful to evaluate whether all the expected resonances could be observed through the CON//HN experiment. Indeed, more than 99% of the backbone correlations could be detected in both experiments (133 out of 134 in the HN and 138 out of 139 in the CON, respectively).

Comparison of the intensities of the spectra collected through the CON//HN experiment with those acquired with exactly the same parameters but in an independent fashion

shows that reduction in signal to noise ratio (S/N) resulting from the combination of the two experiments is minimal for the CON (5%) and more significant for the HN (15%) which, however, has an intrinsically higher sensitivity and provides the same number of peaks as the independent spectrum (Fig. 1, D and E).

When going to more demanding samples of limited lifetime, the advantages of recording the CON//HN are even more striking. The spectra acquired on α -synuclein in cell are shown in Fig. 2 C and are compared with those acquired on α -synuclein on cell lysates (Fig. 2 B) as well as on purified α -synuclein (Fig. 2 A) in conditions approaching physiological ones (310 K; pH 7.4). First of all, it is interesting to observe the spectra acquired on purified α -synuclein in these conditions and compare them with the analogous one acquired at lower pH and temperature (Fig. 1). The CON results are particularly useful to access complete information (138 detected backbone correlations; 99.3% of the expected ones); the quality of the HSQC is instead reduced because of increased solvent exchange broadening at higher pH and temperature (108 detected backbone correlations, 80.6% of the expected ones). Still, the two spectra provide interesting complementary information, including qualitative information about exchange processes.

The same spectra acquired on cell lysate show that most of the signals detected on the purified sample can be observed also in cell lysates. The possibility to acquire the two spectra simultaneously is very important, in particular for cell lysates, which have limited lifetimes. It is worth noting that the linewidths of the observed crosspeaks in the two spectra are very similar to those observed for purified α -synuclein, a nontrivial observation considering the complexity of cell lysate composition. When moving to in-cell experiments, the two spectra still show that a vast majority of the expected signals can be detected, providing atomic resolution information to characterize IDPs inside whole cells. However, the increase in linewidth when going to in-cell spectra is much more pronounced for the ^1H signals with respect to ^{13}C signals. Simultaneous acquisition of the two spectra is important to confirm that it is a real property of in-cell spectra and that it does not depend on possible changes in the experimental conditions or in-cell sample quality. Therefore, the CON//HN provides at the same time information derived from the exclusively heteronuclear spectra, characterized by narrow linewidths, as well as from proton-detected spectra, characterized by higher sensitivity but more influenced by line broadening deriving from both the inhomogeneous environment as well as from exchange processes.

The CON// ^{15}N -detected experiments

^{15}N detection has interesting features for the study of IDPs because of the narrow linewidths and large dispersion of chemical shifts, provided one is in the appropriate sensitivity regime, which nowadays still requires highly

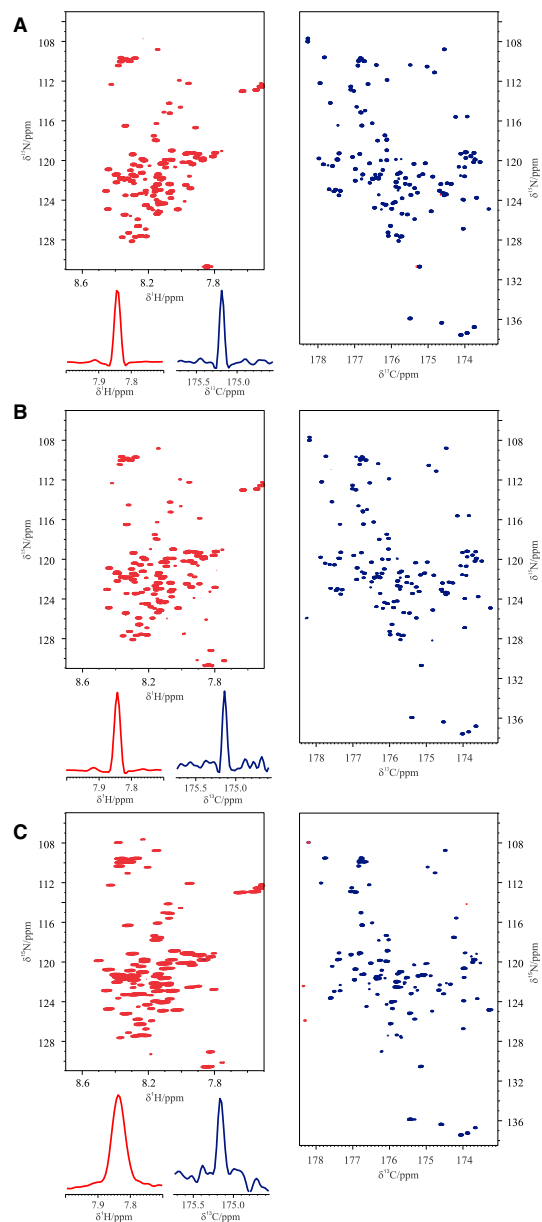


FIGURE 2 Comparison of the 2D spectra (HN *left*; CON *right*) acquired through the CON//HN experiment on ^{13}C , ^{15}N -labeled α -synuclein at 310 K in different experimental conditions: (A) purified sample in 100 mM NaCl, 50 μM EDTA, and 20 mM phosphate buffer at pH 7.4 is shown; (B) in *Escherichia coli* cells lysate resuspended in the same buffer as in (A) and (C) in-cell. The traces of a representative signal extracted from the HN (*red; left*) and the CON (*blue; right*) spectra are also reported. To see this figure in color, go online.

concentrated samples (30–33). The CON experiment was first combined with the ^{15}N -detected experiment to collect 2D HN spectra (CON//btNH) (51). The BEST-TROSY variant (52), which uses band-selective pulses for the amide protons, was selected to exploit the very efficient longitudinal relaxation enhancement of amide protons in IDPs (53–55) as well as to avoid ^1H decoupling during acquisition of the ^{15}N FID, two features that enable the acquisition of 3–4 ^{15}N FIDs during the CON relaxation delay to overcome the very low sensitivity of ^{15}N . Decoupling of ^1H during the ^{15}N evolution delay of the CON was achieved by pairs of 180° H^{N} band-selective pulses to restore the starting polarization source for the BEST-TROSY experiments. In addition, there is no proton decoupling during the acquisition of the CON. These two features reintroduce some dependence on solvent exchange processes as well as $\text{C}'\text{-H}^{\text{N}}$ couplings in the direct acquisition dimension. The major modification included in the BEST-TROSY part of the experiment was to avoid composite pulse decoupling of carbonyl carbon nuclear spins during the acquisition of the ^{15}N FID to preserve C' polarization for the following CON experiment. To this end, virtual decoupling of C' during the acquisition of the ^{15}N FID was implemented through the IPAP approach (56) in combination with band-selective decoupling of the C^α spins. The pulse sequence is reported in Fig. S2 and schematically illustrated in Fig. 3 A. The results of the CON//btNH experiments recorded on α -synuclein at different temperatures are reported in Fig. 3, B and C. The good performance of the experiment is shown by the well-resolved spectra that can be acquired; essentially all expected correlations can be detected at 285 K and pH 6.5 (138 detected backbone correlations for the CON and 133 for the btNH; 99.3% of the expected ones for each spectrum). The modifications introduced to acquire the two spectra simultaneously slightly reduce the sensitivity of the CON spectrum, which is, however, the more sensitive of the two and thus only has a minor impact on the final outcome (Fig. 3, D and E). The ^{15}N -detected BEST-TROSY HN spectrum, although less influenced by exchange broadening of amide protons than its ^1H -detected counterpart, still relies on amide protons in the initial coherence transfer step as well as in the ^1H chemical shift evolution achieved in the indirect dimension. As expected, with increasing temperature, the quality of the CON is essentially maintained, whereas the intensity of crosspeaks in the ^{15}N -detected BEST-TROSY spectrum is reduced. Still, it can be observed that even at higher temperature than physiological one (315 K), a few additional crosspeaks can be observed in the ^{15}N -detected BEST-TROSY HN spectrum that were not identified in these conditions through ^1H -detected experiments (Biological Magnetic Resonance Bank [BMRB]: 27348) (57). In particular, in the ^1H -detected experiment, 93/134 peaks were visible, whereas when performing the ^{15}N -detected variant, seven additional peaks are visible (thus 100/134), mostly due to polar resi-

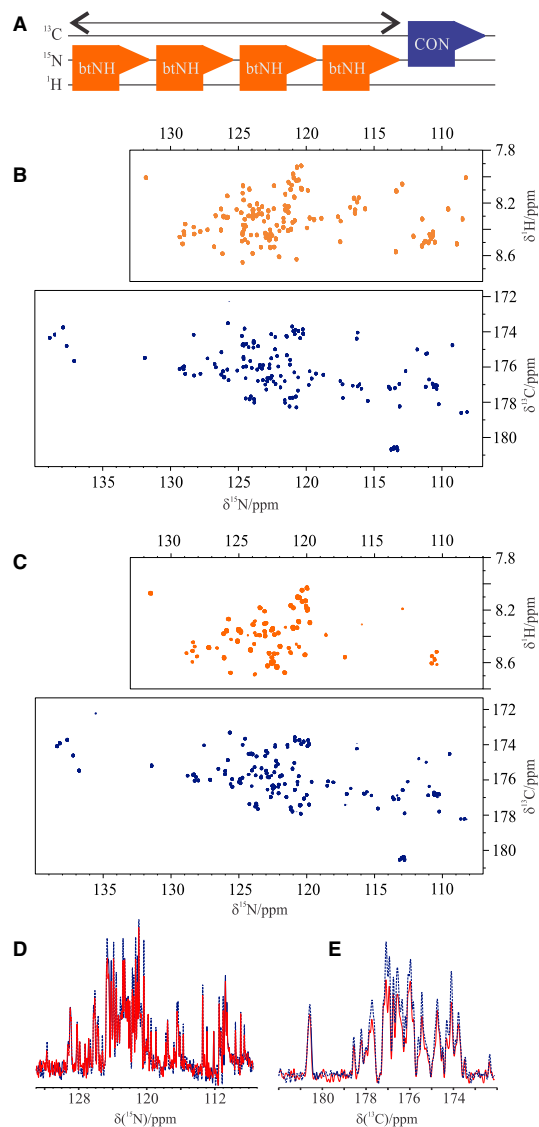


FIGURE 3 (A) Scheme of the CON/btNH experiment. The 2D spectra (btNH top; CON bottom) acquired with the CON//btNH experiment on ^{13}C , ^{15}N -labeled α -synuclein 1 mM at (B) 285 K and (C) 315 K at pH 6.5 in the same buffer reported in Fig. 1 are shown. (D) and (E) report the comparison of the S/N of the projection of the spectra acquired with the CON//btNH experiment (red solid traces) with the analogous spectra acquired independently (blue dotted traces). To see this figure in color, go online.

dues (Thr-22, -64, -44, -33, Ser-9, Lys-43, Ala-76). Even if the signal intensity is lower in the ^{15}N -detected experiment because of the lower gyromagnetic ratio of ^{15}N , it is possible to collect a few more signals because the proton magnetization is transverse for a shorter amount of time,

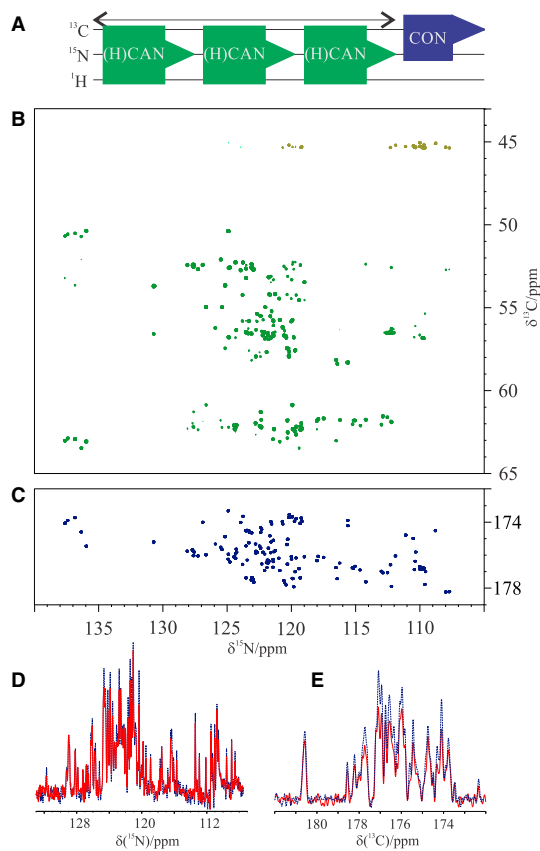


FIGURE 4 (A) Scheme of the CON//(H)CAN experiment. (B) (H)CAN and (C) CON spectra recorded on ^{13}C , ^{15}N -labeled α -synuclein at 310 K (pH 7.4) are shown. (D) and (E) report the comparison of the S/N of the projection of the spectra acquired with the CON//(H)CAN experiment (red solid traces) with the analogous spectra acquired independently (blue dotted traces). To see this figure in color, go online.

reducing losses. This shows that the method has some potential, in particular using optimized hardware (high fields and optimized probes). Another experiment based on ^{15}N detection that was combined with the CON is the (H)CAN (58). The CON//(H)CAN acquired at 310 K and pH 7.4 is shown in Fig. 4. The pulse sequence is reported in Fig. S3 and schematically illustrated in Fig. 4 A. As for the previous experiment, this magnetization transfer pathway starts from a proton, the H^α , and is then transferred to the C^α . The C^α dimension evolves as the indirect one. The magnetization can now be transferred to N of the same amino acid, through the ^1J constant, and to the previous one, through the ^2J constant. The two constants are similar and it is thus difficult to distinguish one pathway from the other. However, we decided to prioritize the intraresidue correlations, which are more sensitive and provide complementary information, with respect to the one available through the CON experi-

ment. When exploiting H^α as starting polarization source, the results are optimal at any temperature because of the nonexchangeable property of this atom. Longitudinal relaxation enhancements are expected to be modest for H^α in IDPs (55), and thus nonselective ^1H pulses can be used. More than one repetition of this experiment during the CON relaxation delay can be collected thanks to the relatively short time needed for H^α to recover to the steady state (hundreds of milliseconds). Technical compromises needed in this case are very similar to those discussed for the previous experiment. They are responsible for a small reduction in the sensitivity of the CON (Fig. 4, D and E), which, however, is the more sensitive between the two. Decoupling of ^1H and ^{13}C was achieved by pairs of 180° pulses to restore the magnetization to equilibrium for subsequent experiments. For the same reason, a virtual CO decoupling scheme is required for the acquisition of ^{15}N FIDs. To this end, we implemented the IPAP scheme, in combination with band-selective decoupling of the C^α spins.

In this spectrum, all the correlations of the intraresidue crosspeaks are observed as well as part of the interresidue ones (113/139; 81.0%). However, significant overlap still occurs, and only a fraction of them are resolved. It is worth noting however that with increasing magnetic fields and with tailored probes, the CON//(H)CAN becomes a promising tool for the study of IDPs. Finally, should long experimental times be necessary to compensate for a lower ^{15}N sensitivity, the CON//(H)CAN experiment can reveal weak crosspeaks present in the CON such as those deriving from the *cis-trans* proline isomerization, which has not been extensively studied for IDPs and has been found to be important in many regulatory processes.

Comparison with other methods

Here, we have chosen to opt for the most straightforward approach to combine two experiments: instead of simply waiting for the magnetization to recover before the next transient, we use the recovery delay of one particularly useful experiment, which would be acquired anyway, to collect an additional one providing complementary information for free. This method follows the idea of activating unexploited magnetization (UTOPIA) (34) but shifts the central interest to the ^{13}C -detected experiment, as schematically indicated in Fig. 1. The simplest variant of the CON experiment was selected because it represents the most robust one to ensure detection of all signals of an IDP near physiological conditions, including those that are invisible in HN spectra (residues experiencing fast solvent exchange and proline residues). This was combined either with ^1H -detected or with ^{15}N -detected 2D experiments.

In principle, a 3D HNCO (or a 3D HNCA), which can nowadays be collected in a very quick time thanks to projection spectroscopy (59) or nonuniform acquisition and processing strategies (60,61), could provide very similar

information. However, the results of these 3D spectra depend on the observability of amide protons. Therefore, the approach presented here is more suitable to study IDPs approaching physiological conditions (310 K; pH 7.4; ionic strength value around 0.25 M; in-cell environment) because the CON is not affected by fast solvent exchange processes.

Similar arguments hold when comparing the current approach with an elegant experimental variant of the HNCO proposed for parallel acquisition of the HN and (H^N)CON experiments (33). In this case, the two spectra (HN and (H^N)CON) are obtained by recovering magnetization from two different pathways that however both originate from H^N polarization. As a result, the two spectra are influenced by solvent exchange broadening, and the (H^N)CON does not reveal proline signals.

The relative sensitivity of the two experiments to be combined is an important issue to be considered. There is no doubt that the intrinsic sensitivity of ^{13}C is lower than that of 1H because of the lower gyromagnetic ratio. However, thanks to instrumental improvements, the sensitivity of the CON experiments acquired on highly flexible IDPs is sufficient to collect the experiment with just a few scans per increment. Therefore, it is worth combining it with the other experiment that is generally collected on an IDP, the HN correlation spectrum. The relative sensitivity of the two experiments is inverted when combining the CON with ^{15}N -detected experiments. This difference can be partly compensated by exploiting 1H as a starting polarization source, exploiting longitudinal relaxation enhancement (when possible) (49,51,52,55), or accumulating more ^{15}N FIDs. A leap in instrumental sensitivity exploiting dedicated probe heads will provide the necessary push to make this combination of experiments really amenable for every application.

From this analysis, it is clear that more than for time savings, which are generally moderate when considering combinations of simple 2D experiments, the simultaneous acquisition of two spectra is important for accessing a snapshot of an IDP in cases in which the sample either has limited stability in time or it changes because of chemical reactions as it happens when studying posttranslational modifications (PTMs). In these cases, to investigate the IDP of interest, in addition to the CON one can access complementary information through the multiple receiver CON variants proposed here.

A case study for PTMs

PTMs represent the most efficient way to modulate the biological activity of proteins after their biosynthesis through chemical reactions, catalyzed by enzymes. Indeed, it is possible to modulate the chemo-physical properties of a protein and thus its function by adding or removing specific chemical groups or through the cleavage of chemical bonds.

The diversity of PTMs found in proteins arises from the very different processes that they can modulate in a cell such as protein folding, signaling, molecular recognition, interactions, and degradation. In this way, the same protein can be handled for different aims, ranging from cell fate control to regulation of metabolism (62).

IDPs are frequently involved in regulatory and signaling functions, which can be modulated by PTMs (63–65). PTMs can trigger both local and long-range changes as well as intermolecular interactions. It has been supposed that structural flexibility could also be a biological strategy to overcome the classical problem of the “one lock/one key” model. The lack of a stable 3D structure guarantees adaptability for the enzymatic site and allows high specificity of the process, maintaining, however, low affinity (64,66).

Among a variety of possible modifications, one of the most common is phosphorylation. It was estimated that at least 2% of the human genome is constituted by kinases and more than 30% of the eukaryotic proteins are involved in phosphorylation, with 700,000 sites that can potentially act as substrates (67). Investigating this mechanism is fundamental to understand functions of these important enzymes. PTMs usually involve only a few residues, and NMR provides a unique tool to study these processes, giving access to residue-specific atomic resolution information.

It is well known that the role of some IDPs or IDRs and their link to the onset of several diseases is strictly dependent on PTMs, as phosphorylation (64–66,68). A clear example is the supposed key role of phosphorylation of specific α -synuclein residues in the modulation of Lewy bodies formation. In this widely studied process, serine 129 is selectively and extensively phosphorylated, promoting fibrillation in vitro (69–71). It has emerged that this process occurs together with phosphorylation of tyrosine 125, a priming event in the efficient modification of serine 129 by CK1 kinase, both in vitro under physiological conditions and in vivo (72–74).

To provide an example of the utility of the CON//HN experiment, we show here the phosphorylation reaction of tyrosine residues by Fyn near physiological conditions. Fyn is a nonreceptor tyrosine kinase of the Src family that has been shown to react with all the four tyrosine residues present in α -synuclein (Y39, Y125, Y133, Y136) with high-percentage levels but preferentially targeting Y125. The spectra acquired with the CON//HN experiment during the phosphorylation reaction are shown in Fig. 5. It can be observed that the changes in HN crosspeaks agree with previous observations (73) and that additional details are monitored through the CON spectrum acquired simultaneously (Fig. 5). In particular, the CON spectrum, which exploits the ^{13}C in the direct dimension, has a higher resolution with respect to the HN spectrum. Thus, we can also monitor the shift of those peaks, which, in the HN spectrum, are

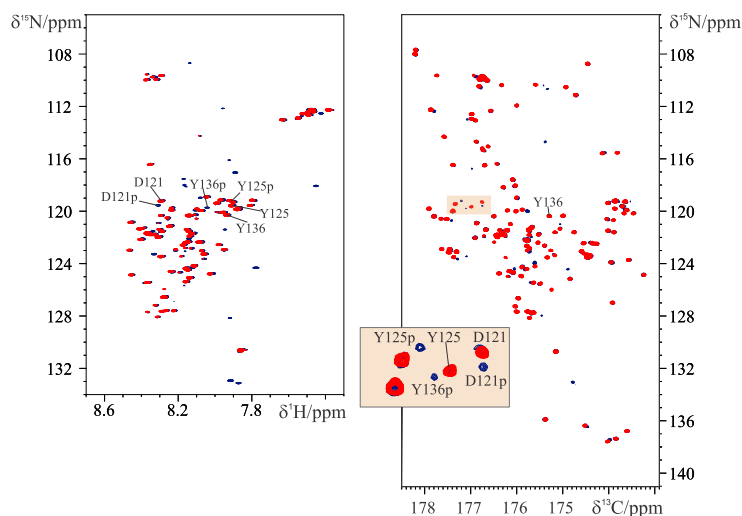


FIGURE 5 The 2D spectra acquired with the CON/HN experiment to monitor phosphorylation of ^{13}C , ^{15}N -labeled α -synuclein 200 μM with Fyn tyrosine kinase as a function of time ($t = 0$ h, red spectra; $t = 50$ h blue spectra). The HN spectra are reported on the left and the CON on the right. The highlighted region of the CON spectra is enlarged to show, as an example, a subset of α -synuclein signals that are influenced by phosphorylation. The reaction was performed at 303 K in 20 mM phosphate buffer, 100 mM NaCl, 50 μM EDTA, 2 mM dithiothreitol, 6 mM MgCl_2 , and 3 mM ATP (pH 7.0). To see this figure in color, go online.

found in crowded regions such as those of Tyr125, which is well resolved in the CON (Fig. 5).

A final comment is necessary on the specific conditions needed to follow PTMs. Most enzymes indeed fulfill their roles at physiological conditions. Therefore, the CON/HN experiment offers a particularly useful tool because in addition to the HN, often used to study PTMs, it exploits the CON, which is not influenced by exchange broadening because of the chemical properties of the nuclei participating in the magnetization transfer pathway. The fairly low substrate concentration generally required for these reactions is still sufficient to acquire good quality ^{13}C -detected spectra, demonstrating the general applicability of this approach.

CONCLUSION

Research in the area of IDPs is continuously expanding. Partly or completely disordered proteins of increasing size and complexity and with high biomedical relevance are being discovered at fast pace, adding a novel layer to structural biology. In this frame, NMR spectroscopy has a central role, enabling the investigation of systems of increasing complexity in an environment that resembles more and more the physiological one. Heteronuclear direct detection has contributed to establishing the importance of NMR in this area of research. The possibility to collect two spectra simultaneously is very attractive to study IDPs. Several experimental variants are presented here in which during the recovery delay needed to acquire the CON, an additional 2D spectrum is collected either based on ^1H detection (CON/HN) or on ^{15}N detection (CON/btNH, CON/(H)CAN). This, to our knowledge, novel approach is particularly useful to study samples of limited lifetime, such as IDPs inside whole cells or cell lysates, as well as to follow chemical reactions such as PTMs of IDPs in conditions

approaching physiological ones (pH, temperature, ionic strength, inhomogeneous in-cell environment). The availability of cryogenically cooled probeheads optimized for heteronuclear direct detection is important to alleviate the problem of the intrinsic lower sensitivity of heteronuclei. The usefulness of these experiments is bound to increase with increasing magnetic fields to take maximal advantage of the narrow linewidths of C' and N of highly flexible IDPs, enabling the study of IDPs of increasing complexity.

SUPPORTING MATERIAL

Supporting Material can be found online at <https://doi.org/10.1016/j.bpj.2019.05.017>.

AUTHOR CONTRIBUTIONS

I.C.F. and R.P. designed the research. All the authors performed research, analyzed the data, discussed the results, and wrote the manuscript.

ACKNOWLEDGMENTS

The support and the use of resources of the CERM/CIRMMMP center of Instruct-ERIC, a Landmark European Strategy Forum on Research Infrastructures project, is gratefully acknowledged. Dr. F. X. Theillet, and Dr. A. Binolfi are gratefully acknowledged for the stimulating discussions.

This work has been supported in part by a grant from the Fondazione CR Firenze and by a grant from the Italian Ministry of Foreign Affairs and International Cooperation (MAE0057283) to R.P.

REFERENCES

1. Habchi, J., P. Tompa, ..., V. N. Uversky. 2014. Introducing protein intrinsic disorder. *Chem. Rev.* 114:6561–6588.

2. Wright, P. E., and H. J. Dyson. 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16:18–29.
3. van der Lee, R., M. Buljan, ..., M. M. Babu. 2014. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* 114:6589–6631.
4. Uversky, V. N., C. J. Oldfield, and A. K. Dunker. 2008. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* 37:215–246.
5. Ambadipudi, S., and M. Zweckstetter. 2016. Targeting intrinsically disordered proteins in rational drug discovery. *Expert Opin. Drug Discov.* 11:65–77.
6. Joshi, P., S. Chia, ..., M. Vendruscolo. 2016. A fragment-based method of creating small-molecule libraries to target the aggregation of intrinsically disordered proteins. *ACS Comb. Sci.* 18:144–153.
7. Heller, G. T., F. A. Aprile, and M. Vendruscolo. 2017. Methods of probing the interactions between small molecules and disordered proteins. *Cell. Mol. Life Sci.* 74:3225–3243.
8. Tsafou, K., P. B. Tiwari, ..., J. A. Toretzky. 2018. Targeting intrinsically disordered transcription factors: changing the paradigm. *J. Mol. Biol.* 430:2321–2341.
9. Eliezer, D. 2009. Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 19:23–30.
10. I. C. Felli and R. Pierattelli, eds 2015. *Intrinsically Disordered Proteins Studied by NMR Spectroscopy* Springer, Basel, Switzerland.
11. Gil, S., T. Hošek, ..., I. C. Felli. 2013. NMR spectroscopic studies of intrinsically disordered proteins at near-physiological conditions. *Angew. Chem. Int. Ed. Engl.* 52:11808–11812.
12. Dyson, H. J., and P. E. Wright. 2001. Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Methods Enzymol.* 339:258–270.
13. Konrat, R. 2014. NMR contributions to structural dynamics studies of intrinsically disordered proteins. *J. Magn. Reson.* 241:74–85.
14. Brutscher, B., I. C. Felli, ..., Z. Solyom. 2015. NMR methods for the study of intrinsically disordered proteins structure, dynamics, and interactions: general overview and practical guidelines. *Adv. Exp. Med. Biol.* 870:49–122.
15. Bermel, W., I. Bertini, ..., R. Pierattelli. 2006. Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. *J. Am. Chem. Soc.* 128:3918–3919.
16. Bermel, W., I. Bertini, ..., R. Pierattelli. 2006. Novel ¹³C direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins. *J. Magn. Reson.* 178:56–64.
17. Pérez, Y., M. Gairí, ..., P. Bernadó. 2009. Structural characterization of the natively unfolded N-terminal domain of human c-Src kinase: insights into the role of phosphorylation of the unique domain. *J. Mol. Biol.* 391:136–148.
18. Knoblich, K., S. Whittaker, ..., U. Günther. 2009. Backbone assignment of the N-terminal polyomavirus large T antigen. *Biomol. NMR Assign.* 3:119–123.
19. Contreras-Martos, S., A. Piai, ..., P. Tompa. 2017. Linking functions: an additional role for an intrinsically disordered linker domain in the transcriptional coactivator CBP. *Sci. Rep.* 7:4676.
20. Hošek, T., E. O. Calçada, ..., R. Pierattelli. 2016. Structural and dynamic characterization of the molecular hub early region 1A (E1A) from human adenovirus. *Chemistry*. 22:13010–13013.
21. Motáčková, V., J. Nováček, ..., V. Sklenář. 2010. Strategy for complete NMR assignment of disordered proteins with highly repetitive sequences based on resolution-enhanced 5D experiments. *J. Biomol. NMR.* 48:169–177.
22. Haba, N. Y., R. Gross, ..., J. H. Chill. 2013. NMR determines transient structure and dynamics in the disordered C-terminal domain of WASP interacting protein. *Biophys. J.* 105:481–493.
23. Nováček, J., N. Y. Haba, ..., V. Sklenář. 2012. 4D non-uniformly sampled HCBCACON and ¹J(NCα)-selective HCBCANCO experiments for the sequential assignment and chemical shift analysis of intrinsically disordered proteins. *J. Biomol. NMR.* 53:139–148.
24. Lawrence, C. W., and S. A. Showalter. 2012. Carbon-detected ¹⁵N NMR spin relaxation of an intrinsically disordered protein: FCP1 dynamics unbound and in complex with RAP74. *J. Phys. Chem. Lett.* 3:1409–1413.
25. Sahu, D., M. Bastidas, and S. A. Showalter. 2014. Generating NMR chemical shift assignments of intrinsically disordered proteins using carbon-detected NMR methods. *Anal. Biochem.* 449:17–25.
26. Lopez, J., R. Schneider, ..., G. Lippens. 2016. Studying intrinsically disordered proteins under true in vivo conditions by combined cross-polarization and carbonyl-detection NMR spectroscopy. *Angew. Chem. Int. Ed. Engl.* 55:7418–7422.
27. Piai, A., E. O. Calçada, ..., R. Pierattelli. 2016. Just a flexible linker? The structural and dynamic properties of CBP-ID4 revealed by NMR spectroscopy. *Biophys. J.* 110:372–381.
28. Eftekharzadeh, B., A. Piai, ..., X. Salvatella. 2016. Sequence context influences the structure and aggregation behavior of a PolyQ tract. *Biophys. J.* 110:2361–2366.
29. Kovacs, H., D. Moskau, and M. Spraul. 2005. Cryogenically cooled probes - a leap in NMR technology. *Prog. Nucl. Mag. Res. Sp.* 46:131–155.
30. Takeuchi, K., G. Heffron, ..., G. Wagner. 2010. Nitrogen-detected CAN and CON experiments as alternative experiments for main chain NMR resonance assignments. *J. Biomol. NMR.* 47:271–282.
31. Takeuchi, K., H. Arthanari, and G. Wagner. 2016. Perspective: revisiting the field dependence of TROSY sensitivity. *J. Biomol. NMR.* 66:221–225.
32. Chhabra, S., P. Fischer, ..., H. Arthanari. 2018. ¹⁵N detection harnesses the slow relaxation property of nitrogen: delivering enhanced resolution for intrinsically disordered proteins. *Proc. Nat. Acad. Sci. USA.* 115:1710–1719.
33. Gibbs, E. B., and R. W. Kriwacki. 2018. Direct detection of carbon and nitrogen nuclei for high-resolution analysis of intrinsically disordered proteins using NMR spectroscopy. *Methods.* 138-139:39–46.
34. Kupče, E., R. Freeman, and B. K. John. 2006. Parallel acquisition of two-dimensional NMR spectra of several nuclear species. *J. Am. Chem. Soc.* 128:9606–9607.
35. Kupče, E., and R. Freeman. 2010. Molecular structure from a single NMR sequence (fast-PANACEA). *J. Magn. Reson.* 206:147–153.
36. Chakraborty, S., S. Paul, and R. V. Hosur. 2012. Simultaneous acquisition of ¹³C^α, ¹⁵N and ¹H-¹⁵N-¹⁵N sequential correlations in proteins: application of dual receivers in 3D HNN. *J. Biomol. NMR.* 52:5–10.
37. Kupče, E., and L. E. Kay. 2012. Parallel acquisition of multi-dimensional spectra in protein NMR. *J. Biomol. NMR.* 54:1–7.
38. Viegas, A., T. Viennet, ..., M. Eitzkorn. 2016. UTOPIA NMR: activating unexploited magnetization using interleaved low-gamma detection. *J. Biomol. NMR.* 64:9–15.
39. Kupče, E. 2013. NMR with multiple receivers. *Top. Curr. Chem.* 335:71–96.
40. Sharma, K., P. K. Madhu, and K. R. Mote. 2016. A suite of pulse sequences based on multiple sequential acquisitions at one and two radio-frequency channels for solid-state magic-angle spinning NMR studies of proteins. *J. Biomol. NMR.* 65:127–141.
41. Kupče, E., L. E. Kay, and R. Freeman. 2010. Detecting the “afterglow” of ¹³C NMR in proteins using multiple receivers. *J. Am. Chem. Soc.* 132:18008–18011.
42. Huang, C., G. Ren, ..., C. C. Wang. 2005. A new method for purification of recombinant human α-synuclein in *Escherichia coli*. *Protein Expr. Purif.* 42:173–177.
43. Felli, I. C., L. Gonnelli, and R. Pierattelli. 2014. In-cell ¹³C NMR spectroscopy for the study of intrinsically disordered proteins. *Nat. Protoc.* 9:2005–2016.
44. Emsley, L., and G. Bodenhausen. 1992. Optimization of shaped selective pulses for NMR using a quaternion description of their overall propagators. *J. Magn. Reson.* 97:135–148.

45. Böhlen, J. M., and G. Bodenhausen. 1993. Experimental aspects of Chirp NMR spectroscopy. *J. Magn. Reson.* 102:293–301.
46. Piotta, M., V. Saudek, and V. Sklenář. 1992. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR.* 2:661–665.
47. Freeman, R., and H. Geen. 1991. Band-selective radiofrequency pulses. *J. Magn. Reson.* 93:93–141.
48. Markley, J. L., A. Bax, ..., K. Wüthrich. 1998. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *Pure Appl. Chem.* 70:117–142.
49. Felli, I. C., and R. Pierattelli. 2015. Spin-state-selective methods in solution- and solid-state biomolecular ^{13}C NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* 84–85:1–13.
50. Felli, I. C., and R. Pierattelli. 2014. Novel methods based on ^{13}C detection to study intrinsically disordered proteins. *J. Magn. Reson.* 241:115–125.
51. Takeuchi, K., H. Arthanari, ..., G. Wagner. 2015. Nitrogen detected TROSY at high field yields high resolution and sensitivity for protein NMR. *J. Biomol. NMR.* 63:323–331.
52. Solyom, Z., M. Schwarten, ..., B. Brutscher. 2013. BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J. Biomol. NMR.* 55:311–321.
53. Schanda, P., V. Forge, and B. Brutscher. 2006. HET-SOFAST NMR for fast detection of structural compactness and heterogeneity along polypeptide chains. *Magn. Reson. Chem.* 44:S177–S184.
54. Schanda, P., H. Van Melckebeke, and B. Brutscher. 2006. Speeding up three-dimensional protein NMR experiments to a few minutes. *J. Am. Chem. Soc.* 128:9042–9043.
55. Hošek, T., S. Gil-Caballero, ..., I. C. Felli. 2015. Longitudinal relaxation properties of $^1\text{H}^{\text{N}}$ and $^1\text{H}^{\text{C}}$ determined by direct-detected ^{13}C NMR experiments to study intrinsically disordered proteins (IDPs). *J. Magn. Reson.* 254:19–26.
56. Bermel, W., I. Bertini, ..., R. Pierattelli. 2006. ^{13}C -detected protonless NMR spectroscopy of proteins in solution. *Prog. Nucl. Mag. Res. Sp.* 48:25–45.
57. Murrall, M. G., M. Schiavina, ..., I. C. Felli. 2018. ^{13}C APSY-NMR for sequential assignment of intrinsically disordered proteins. *J. Biomol. NMR.* 70:167–175.
58. Gal, M., K. A. Edmonds, ..., G. Wagner. 2011. Speeding up direct ^{15}N detection: hCaN 2D NMR experiment. *J. Biomol. NMR.* 51:497–504.
59. Hiller, S., F. Fiorito, ..., G. Wider. 2005. Automated projection spectroscopy (APSY). *Proc. Natl. Acad. Sci. USA.* 102:10876–10881.
60. Robson, S., H. Arthanari, ..., G. Wagner. 2019. Nonuniform sampling for NMR spectroscopy. *Methods Enzymol.* 614:263–291.
61. Kazimierczuk, K., J. Stanek, ..., W. Koźmiński. 2010. Random sampling in multidimensional NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* 57:420–434.
62. Cheng, H. C., R. Z. Qi, ..., H. J. Zhu. 2011. Regulation and function of protein kinases and phosphatases. *Enzyme Res.* 2011:794089.
63. Zhou, J., S. Zhao, and A. K. Dunker. 2018. Intrinsically disordered proteins link alternative splicing and post-translational modifications to complex cell signaling and regulation. *J. Mol. Biol.* 430:2342–2359.
64. Darling, A. L., and V. N. Uversky. 2018. Intrinsic disorder and post-translational modifications: the darker side of the biological dark matter. *Front. Genet.* 9:158.
65. Bah, A., and J. D. Forman-Kay. 2016. Modulation of intrinsically disordered protein function by post-translational modifications. *J. Biol. Chem.* 291:6696–6705.
66. Uversky, V. N. 2014. Wrecked regulation of intrinsically disordered proteins in diseases: pathogenicity of deregulated regulators. *Front. Mol. Biosci.* 1:6.
67. Ubersax, J. A., and J. E. Ferrell, Jr. 2007. Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* 8:530–541.
68. Fujiwara, H., M. Hasegawa, ..., T. Iwatsubo. 2002. α -synuclein is phosphorylated in synucleinopathy lesions. *Nat. Cell Biol.* 4:160–164.
69. Choi, H. S., H. Liew, ..., Y. H. Suh. 2012. Phosphorylation of α -synuclein is crucial in compensating for proteasomal dysfunction. *Biochem. Biophys. Res. Commun.* 424:597–603.
70. Mahul-Mellier, A. L., B. Fauvet, ..., H. A. Lashuel. 2014. c-Abl phosphorylates α -synuclein and regulates its degradation: implication for α -synuclein clearance and contribution to the pathogenesis of Parkinson's disease. *Hum. Mol. Genet.* 23:2858–2879.
71. Schmid, A. W., B. Fauvet, ..., H. A. Lashuel. 2013. α -synuclein post-translational modifications as potential biomarkers for Parkinson disease and other synucleinopathies. *Mol. Cell. Proteomics.* 12:3543–3558.
72. Nakamura, T., H. Yamashita, ..., S. Nakamura. 2001. Activated Fyn phosphorylates α -synuclein at tyrosine residue 125. *Biochem. Biophys. Res. Commun.* 280:1085–1092.
73. Kosten, J., A. Binolfi, ..., P. Selenko. 2014. Efficient modification of α -synuclein serine 129 by protein kinase CK1 requires phosphorylation of tyrosine 125 as a priming event. *ACS Chem. Neurosci.* 5:1203–1208.
74. Paleologou, K. E., A. W. Schmid, ..., H. A. Lashuel. 2008. Phosphorylation at Ser-129 but not the phosphomimics S129E/D inhibits the fibrillation of α -synuclein. *J. Biol. Chem.* 283:16895–16905.

Biophysical Journal, Volume 117

Supplemental Information

**Taking Simultaneous Snapshots of Intrinsically Disordered Proteins in
Action**

**Marco Schiavina, Maria Grazia Murrari, Letizia Pontoriero, Valerio Sainati, Rainer
Kümmerle, Wolfgang Bermel, Roberta Pierattelli, and Isabella C. Felli**

Supporting Material

Taking Simultaneous Snapshots of Intrinsically Disordered Proteins in Action

M. Schiavina, M. G. Murrari, L. Pontoriero, V. Sainati, R. Kümmerle, W. Bermel, R. Pierattelli, I. C. Felli

TABLE OF CONTENTS

Figure S1. CON//HN pulse sequence scheme.....	2
Figure S2. CON//btNH pulse sequence scheme.	3
Figure S3. CON//(H)CAN pulse sequence scheme.....	4
Pulse sequence CON//HN.....	5
Pulse sequence CON//btNH	10
Pulse sequence CON//(H)CAN.....	18

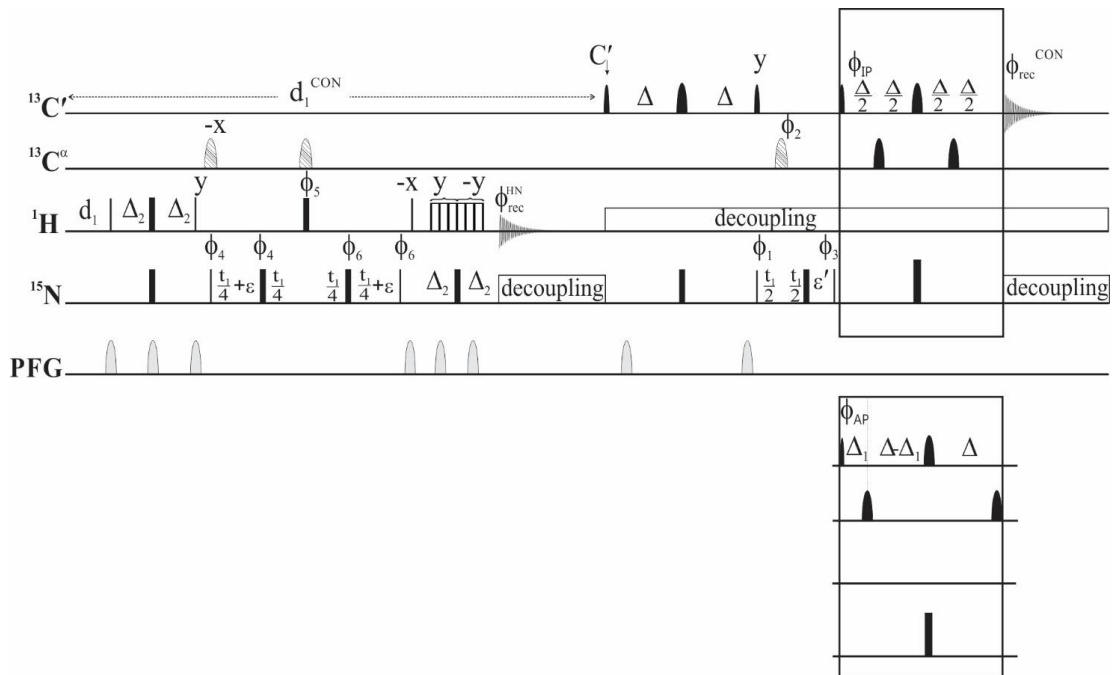


Figure S1. CON//HN pulse sequence scheme.

Narrow and wide black rectangles represent $\pi/2$ and π non-selective pulses; narrow and wide black symbols represent $\pi/2$ and π band-selective pulses (Q5 and Q3 shapes of 300 and 231 μs respectively), hatched pulses are adiabatic inversion pulse on C' and C^α (smoothed Chirp 500 μs 80 kHz sweep and 20% smooting).

The following phase cycling was employed for the CON//HN: $\phi_1 = x, -x$; $\phi_2 = 2(x), 2(-x)$; $\phi_3 = 4(x), 4(-x)$, $\phi_{\text{rec}}^{\text{IP}} = x$; $\phi^{\text{AP}} = -y$ and $\phi_{\text{rec}}^{\text{CON}} = x, -x, x, -x, -x, x, -x, x$ for the 2D-CON and $\phi_4 = x, -x$, $\phi_5 = 2(x), 2(-x)$ $\phi_6 = 4(x) 4(-x)$ and $\phi_{\text{rec}}^{\text{HN}} = x, -x, x, -x, -x, x, -x, x$ for the 2D-HN. Quadrature detection in the indirect dimension was achieved through the STATES-TPPI approach incrementing phase ϕ_1 (CON) and ϕ_4 (HN).

The length of the delays was: $d_1 = 1.7 \text{ s}$; $\Delta = 16.6 \text{ ms}$; $\Delta_1 = 4.5 \text{ ms}$; $\Delta_2 = 2.7 \text{ ms}$; $\epsilon = \frac{1}{2}$ duration of adiabatic pulse (250 μs); $\epsilon' = t_1(0) + \text{duration of adiabatic pulse (500 } \mu\text{s)}$.

Virtual decoupling of the $C'-C^\alpha$ coupling in the 2D-CON was achieved by acquiring for each increment both the IP and AP component of the signals and combining them.

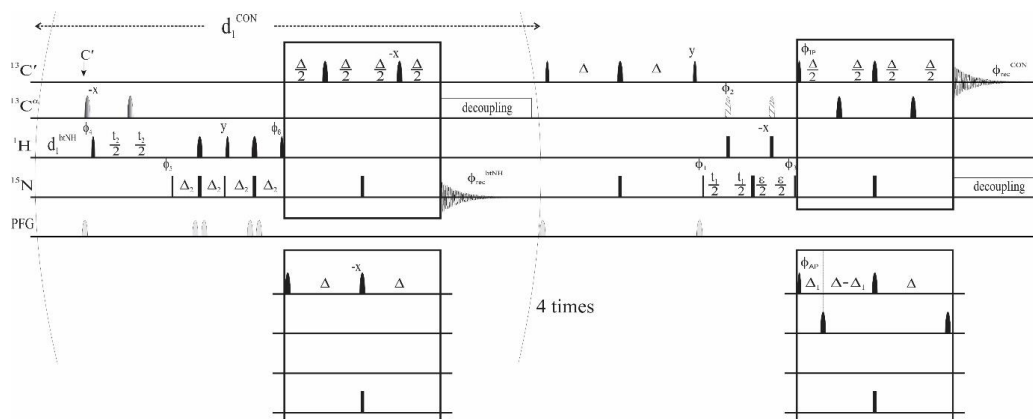


Figure S2. CON//btNH pulse sequence scheme.

Narrow and wide black rectangles represent $\pi/2$ and π non-selective pulses; narrow and wide black symbols represent $\pi/2$ and π band-selective pulses (Q5 and Q3 shapes of 300 and 231 μs respectively), hatched pulses are adiabatic inversion pulse on C' and C^α (smoothed Chirp 500 μs 80 kHz sweep and 20% smooting), the shaded shapes are Bip pulses.

The following phase cycling was employed for the CON//btNH: $\phi_1 = x, -x$; $\phi_2 = 2(x), 2(-x)$; $\phi_3 = 4(x), 4(-x)$ and $\phi_{\text{rec}}^{\text{CON}} = x, -x, x, -x, -x, x, -x, x$ for the 2D-CON and $\phi_4 = y, -y, x, -x$; $\phi_5 = y$ $\phi_6 = x$ and $\phi_{\text{rec}}^{\text{btNH}} = x, -x, -y, y$ for the 2D-btNH. Quadrature detection in the indirect dimension was achieved through the STATES-TPPI approach incrementing phase ϕ_1 for the CON and through the Echo/Antiecho approach by incrementing phase ϕ_4 and ϕ_5 .

The length of the delays was: $d_1 = 2.35$ s; $\Delta = 16.6$ ms; $\Delta_1 = 4.5$ ms; $\Delta_2 = 2.7$ ms; $\epsilon = t_1(0) +$ duration of adiabatic pulse (500 μs).

Virtual decoupling of the $C'-C^\alpha$ coupling in the 2D-CON was achieved by acquiring for each increment both the IP and AP component of the signals. The same approach was also employed to achieve heteronuclear decoupling ($C'-N$) during acquisition of the ^{15}N FIDs in order to preserve the C' polarization for the subsequent CON experiment. Band selective C^α decoupling was achieved using a Q3 pulse in a P5M4 supercycle.

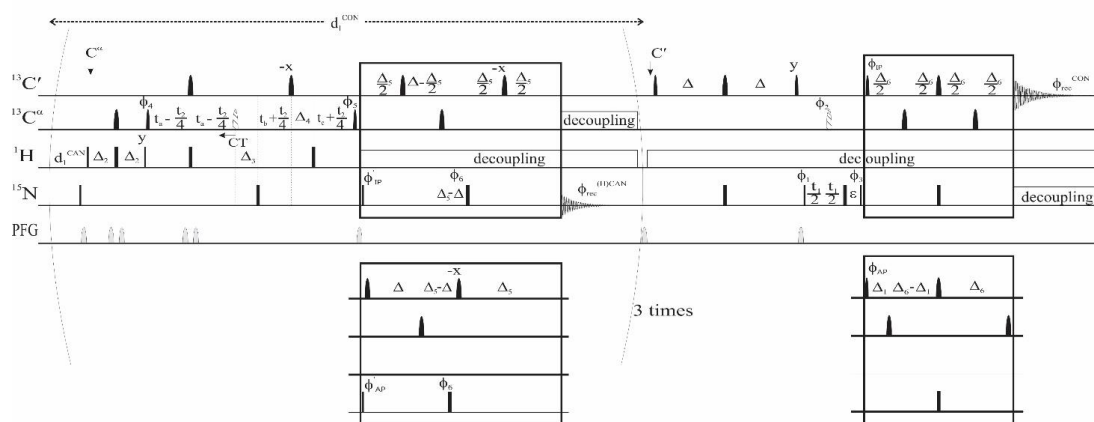


Figure S3. CON//(H)CAN pulse sequence scheme.

Narrow and wide black rectangles represent $\pi/2$ and π non-selective pulses; narrow and wide black symbols represent $\pi/2$ and π band-selective pulses (Q5 and Q3 shapes of 300 and 231 μs respectively), hatched pulses are adiabatic inversion pulse on C' and C^α (smoothed Chirp 500 μs 80 kHz sweep and 20% smooting).

The following phase cycling was employed for the CON//(H)CAN: $\phi_1 = x, -x$; $\phi_2 = 2(x), 2(-x)$; $\phi_3 = 4(x), 4(-x)$; $\phi_4 = x, -x$; $\phi_5 = 4(x), 4(-x)$; $\phi_6 = 8(x), 8(-x)$; $\phi^{IP} = x$; $\phi^{AP} = -y$ and $\phi_{\text{rec}}^{\text{CON}} = x, -x, x, -x, -x, x, -x, x$ for the 2D-CON and $\phi_4 = x, -x$, $\phi_5 = 4(x), 4(-x)$, $\phi_6 = 8(x), 8(-x)$; $\phi^{IP} = 2(x), 2(-x)$; $\phi^{AP} = 2(y), 2(-y)$ and $\phi_{\text{rec}}^{(\text{H})\text{CAN}} = x, -x, -x, x, -x, x, x, -x$ for the 2D-(H)CAN. Quadrature detection in the indirect dimension was achieved through STATES-TPPI approach incrementing phase ϕ_1 (CON) and ϕ_4 ((H)CAN).

The length of the delays was: $d_1 = 2.5$ s; $\Delta = 12.4$ ms; $\Delta_1 = 4.5$ ms; $\Delta_2 = 1.7$ ms; $\Delta_3 = 1.9$ ms; $\Delta_4 = 1.4$ ms; $\Delta_5 = 4.2$ ms; $\Delta_6 = 16.6$ ms. t_a, t_b and t_c were used to achieve the constant time mode for the $^{13}\text{C}^\alpha$ indirect dimension.

Virtual decoupling of the $C'-C^\alpha$ coupling in the CON was achieved by acquiring for each increment both the IP and AP component of the signals. The same approach was also employed to achieve heteronuclear decoupling ($C'-N$) during acquisition of the ^{15}N FIDs in order to preserve the C' polarization for the subsequent CON experiment. Band selective C^α decoupling was achieved using a Q3 pulse in a P5M4 supercycle.

Pulse sequence CON//HN

```
;ut_con_fhsqchn
;avance-version

;Dataset 1 (F1)
;fhsqcf3gpph
;avance-version (12/01/11)
;2D H-1/X correlation via double inept transfer
;
;      F1(HN) -> F3(N,t1) -> F1(HN,t2)
;
;phase sensitive
;with decoupling during acquisition
;
;(S. Mori, C. Abeygunawardana, M. O'Neil-Johnson & P.C.M. van Zijl,
; J. Magn. Reson. B 108, 94-98 (1995) )
;
;$CLASS=HighRes
;$DIM=2D
;$TYPE=
;$SUBTYPE=
;$COMMENT=

;Dataset 2 (F2)
;CON
;2D sequence with
; 13C detected correlation for triple resonance using
; inept transfer steps
;
;      F1(C=O) -> F3(N,t1) -> F1(C=O,t2)
;
;on/off resonance 13C pulses using shaped pulses
;phase sensitive (t1)
;using IPAP scheme for virtual decoupling
;(use parameterset C_CON_IASQ)
;
;W. Bermel, I. Bertini, L. Duma, I.C. Felli, L. Emsley, R. Pierattelli,
; P.R. Vasos, Angew. Chem. Int. Ed. 44, 3089-3092 (2005)
;(L. Duma, S. Hediger, A. Lesage & L. Emsley,
; J. Magn. Reson. 164, 187-195 (2003) )
;
;$CLASS=HighRes
;$DIM=2D
;$TYPE=
;$SUBTYPE=
;$COMMENT=

prosol relations=<triple>

#include <Avance.incl>
#include <Delay.incl>
#include <Grad.incl>

"p2=p1*2"
"p22=p21*2"
"d11=30m"
"d12=20u"
"d26=1s/(cnst4*4)"

"d22=4.5m"
"d23=16.6m"

"in31=in1/2"
"in32=in2/2"
```



```

"d31=3u"
"d32=3u"

"DELTA=d19-p22/2"
"DELTA1=d26-p16-d16-p27*3-d19*5-p1*2/PI"
"DELTA2=d26-p16-d16-p27*2-p0-d19*5-de-8u"
"DELTA3=d31+larger(p2,p8)/2"
"DELTA4=p21*2/PI"
"DELTA5=d32*2+p8"
"DELTA6=d23/2-p14/2"
"DELTA7=d23-d22-p14"

"TAU=d26-p16-d16-4u"

"l0=1"

"l3=td1/2"

"spoff2=0"
"spoff3=0"
"spoff5=bf2*((cnst22-cnst21)/1000000)"
"spoff8=0"
"spoff13=bf2*((cnst26-cnst21)/1000000)"

1 4u ze1
   4u ze2

2 d11 do:f1 do:f3
   3m
3 12m
4 6m
5 d1

; fhsqc

20u fq=cnst47(bf ppm):f3
d12 p11:f1 p13:f3
50u UNBLKGRAD

(p1 ph11)
4u
p16:gp1
d16
TAU p13:f3
(center (p2 ph11) (p22 ph16):f3 )
4u
TAU
p16:gp1
d16
(p1 ph12)

4u
p16:gp2
d16
(p8:sp13 ph17):f2
4u

(p21 ph13):f3
DELTA3
(p22 ph13):f3
DELTA4
d31

(center (p2 ph15) (p8:sp13 ph11):f2 )

```

```

d31
DELTA4
(p22 ph14):f3
DELTA3
(p21 ph14):f3

4u
p16:gp2
d16

(p1 ph17)
p16:gp3
d16 pl18:f1
DELTA1
p27*0.231 ph18
d19*2
p27*0.692 ph18
d19*2
p27*1.462 ph18
DELTA
(p22 ph11):f3
DELTA
p27*1.462 ph19
d19*2
p27*0.692 ph19
d19*2
p0*0.231 ph19
4u
p16:gp3
d16
4u BLKGRAD
DELTA2 pl16:f3

goscnp1 ph31 cpd3:f3

; con

4u do:f3
50u UNBLKGRAD

p16:gp5
d16

20u pl19:f1 pl3:f3
20u fq=cnst21(bf ppm):f2
20u fq=cnst57(bf ppm):f3
d12 cpds1:f1

(p13:sp2 ph1):f2
d23
(center (p14:sp3 ph1):f2 (p22 ph1):f3 )
d23
(p13:sp8 ph2):f2

p16:gp4
d16

(p21 ph3):f3
d32
(p8:sp13 ph5):f2
d32
(p22 ph1):f3
DELTA5
(p21 ph4):f3

if "10 %2 == 1"
{
(p13:sp2 ph1):f2
DELTA6
(p14:sp5 ph1):f2

```

```

        DELTA6
        (center (p14:sp3 ph1):f2 (p22 ph1):f3 )
        DELTA6
        (p14:sp5 ph1):f2
        DELTA6 pl16:f3
    }
else
    {
        (p13:sp2 ph6):f2
        d22
        (p14:sp5 ph1):f2
        DELTA7
        (center (p14:sp3 ph1):f2 (p22 ph1):f3 )
        DELTA6
        DELTA6 pl16:f3
        (p14:sp5 ph1):f2
    }

4u BLKGRAD
go2=2 ph30 cpd3:f3

d11 do:f1 do:f3 wr2 #1 if2 #1 zd2

;exp_f2 ipap
3m iu0
lo to 3 times 2

3m wr1 #0 if1 #0 zd1

;exp_f1 phase
3m ip13
3m ip16

;exp_f2 phase
3m ip3
lo to 4 times 2

;exp_f1 delay
3m id31

;exp_f2 delay
3m id32
lo to 5 times l3

exit

ph1=0
ph2=1
ph3=0 2
ph4=0 0 0 0 2 2 2 2
ph5=0 0 2 2
ph6=3

ph11=0
ph12=1
ph13=0 2
ph14=0 0 0 0 2 2 2 2
ph15=0 0 2 2
ph16=0
ph17=2
ph18=1
ph19=3

ph30=0 2 0 2 2 0 2 0
ph31=0 2 0 2 2 0 2 0

;p11 : f1 channel - power level for pulse (default)
;p13 : f3 channel - power level for pulse (default)
;pl16: f3 channel - power level for CPD/BB decoupling

```

```

;pl18: f1 channel - power level for 3-9-19-pulse (watergate)
;pl19: f1 channel - power level for CPD/BB decoupling
;sp2 : f1 channel - shaped pulse 90 degree (on resonance)
;sp3 : f1 channel - shaped pulse 180 degree (on resonance)
;sp5 : f1 channel - shaped pulse 180 degree (Ca off resonance)
;sp8 : f1 channel - shaped pulse 90 degree (on resonance)
;
;      for time reversed pulse
;sp13: f1 channel - shaped pulse 180 degree (adiabatic)
;p0 : f1 channel - 90 degree pulse at pl18 for fine adjustment
;p1 : f1 channel - 90 degree high power pulse
;p2 : f1 channel - 180 degree high power pulse
;p8 : f1 channel - 180 degree shaped pulse for inversion (adiabatic)
;p13: f1 channel - 90 degree shaped pulse
;p14: f1 channel - 180 degree shaped pulse
;p16: homospoil/gradient pulse [1 msec]
;p21: f3 channel - 90 degree high power pulse
;p22: f3 channel - 180 degree high power pulse
;p27: f1 channel - 90 degree pulse at pl18
;d1 : relaxation delay; 1-5 * T1
;d11: delay for disk I/O [30 msec]
;d12: delay for power switching [20 usec]
;d16: delay for homospoil/gradient recovery
;d19: delay for binomial water suppression
;d22: 1/(4J(COCa)) [4.5 msec]
;d23: 1/(4J(NCO)) [16.6 msec]
;d26: 1/(4J(NH))
;d31: incremented delay (2D, exp 1) [3 usec]
;d32: incremented delay (2D, exp 2) [3 usec]
;cnst4: = J(NH)
;cnst21: CO chemical shift (offset, in ppm)
;cnst22: Calpha chemical shift (offset, in ppm)
;cnst26: Call chemical shift (offset, in ppm) [101 ppm]
;cnst47: N(H) chemical shift (offset, in ppm)
;cnst57: N chemical shift (offset, in ppm)
;olp: CO chemical shift (cnst21)
;l0: flag to switch between inphase and antiphase
;l3 : loop for 2D experiment = tdl/2
;inf1: 1/SW(N) = 2 * DW(N)
;in2 : 1/SW(N) = 2 * DW(N) (inf1 of exp 2, N all)
;in31: 1/(2 * SW(N)) = DW(N)
;nd31: 2
;in32: 1/(2 * SW(N)) = DW(N)
;nd32: 2
;ns: 8 * n
;ds: >= 32
;tdl: number of experiments in F1
;      (number of experiments after IPAP processing)
;FnMODE: States-TPPI (or TPPI) in F1
;cpds1: decoupling according to sequence defined by cpdprg1
;cpd3: decoupling according to sequence defined by cpdprg3
;pcpd1: f1 channel - 90 degree pulse for decoupling sequence
;pcpd3: f3 channel - 90 degree pulse for decoupling sequence

;for z-only gradients:
;gpz1: 50%
;gpz2: 80%
;gpz3: 31%
;gpz4: 19%
;gpz5: 60%

;use gradient files:
;gpnam1: SMSQ10.100
;gpnam2: SMSQ10.100
;gpnam3: SMSQ10.100
;gpnam4: SMSQ10.100
;gpnam5: SMSQ10.100

;use AU-program splitcomb [ipap 2] to process data
;$Id: $

```

Pulse sequence CON//btNH

```
;ut_con_n_btnh
;avance-version

;Dataset 1 (F1)
;c_con_iasq
;avance-version
;CON
;2D sequence with
; 13C detected correlation for triple resonance using
;   inept transfer steps
;
;   F1(C=O) -> F3(N,t1) -> F1(C=O,t2)
;
;on/off resonance 13C pulses using shaped pulses
;phase sensitive (t1)
;using IPAP scheme for virtual decoupling
;(use parameterset C_CON_IASQ)
;
;W. Bermel, I. Bertini, L. Duma, I.C. Felli, L. Emsley, R. Pierattelli,
; P.R. Vasos, Angew. Chem. Int. Ed. 44, 3089-3092 (2005)
;(L. Duma, S. Hediger, A. Lesage & L. Emsley,
; J. Magn. Reson. 164, 187-195 (2003) )
;
;
;Dataset 2 (F2)
;dummy dataset
;
;
;Dataset 3 (F3)
;nb_hntrosy_f2ig
;avance-version (17/02/10)
;2D 15N shift correlation
;
;   F2(HN,t1) -> F3(N,t2)
;
;with refocussing of chemical shifts
;phase sensitive
;using semi-constant time in t1
;(use parameterset )
;
;(R. Schnieders, C. Richter, S. Warhaut, V. de Jesus, S. Keyhani,
; E. Duchardt-Ferner, H. Keller, J. WÄhnhert, L.T. Kuhn, A.L. Breeze,
; W. Bermel, H. Schwalbe & B. FÄrtig, ; J. Biomol. NMR 69, 31â€44 (2017) )
;(K. Takeuchi, H. Arthanari, I. Shimada & G. Wagner, J. Biomol. NMR 63, 1â€9 (2015) )
;(K. Takeuchi, H. Arthanari, M. Imai & G. Wagner, J. Biomol. NMR 64, 143â€151 (2016) )
;
;$CLASS=HighRes
;$DIM=2D
;$TYPE=
;$SUBTYPE=
;$COMMENT=

prosol relations=<triple_c>

#include <Avance.incl>
#include <Delay.incl>
#include <Grad.incl>

"p22=p21*2"
"d11=30m"
"d12=20u"
"d26=1s/(cnst4*4)"

"d22=4.5m"
```

```

"d23=16.6m"

"p42=(bwfac42/(cnst55*cnst52*bf2))*1000000"
"spw42=plw2/((p42*90.0)/(p3*totrot42))*((p42*90.0)/(p3*totrot42))*(integfac42*integ-
fac42)"
"spoal42=0.5"
"spoff42=bf2*(cnst54/1000000)-o2"

"p43=(bwfac43/(cnst55*cnst53*bf2))*1000000"
"spw43=plw2/((p43*90.0)/(p3*totrot43))*((p43*90.0)/(p3*totrot43))*(integfac43*integ-
fac43)"
"spw44=plw2/((p43*90.0)/(p3*totrot44))*((p43*90.0)/(p3*totrot44))*(integfac44*integ-
fac44)"
"spoal43=0"
"spoal44=1"
"spoff43=bf2*(cnst54/1000000)-o2"
"spoff44=bf2*(cnst54/1000000)-o2"

"in31=inf1/2"
"in32=in2/2"

"d31=3u"
"d32=in32/2-p39/2"

"d60=d1-(d61+d59)*4-d58"

"DELTA=d31"
"DELTA1=d23/2-p12/2"
"DELTA2=d23-d22-p12"
"DELTA11=d26-larger(p22,p42)/2-p16-d16"
"DELTA12=d26-larger(p22,p42)/2-p16-d16-p21*2/PI"
"DELTA13=d26-larger(p22,p42)/2-p16-d16-de"
"DELTA14=d23/2-p12/2"

"l0=0"
"l3=l22/2"
"l10=1"
"l13=td1/(8*l22)"
"l30=0"

"spoff4=bf1*((cnst26-cnst22)/1000000)"
"spoff13=bf1*((cnst26-cnst21)/1000000)"
"spoff23=0"
"spoff24=0"
"spoff25=0"
"spoff26=bf1*((cnst22-cnst21)/1000000)"
"spoff27=bf1*((cnst21-cnst22)/1000000)"

1 4u zel
4u ze3
2 d11 do:f3
3m
3 3m
4 3m
5 6m
6 9m
7 12.02m
8 3m

d60
50u UNBLKGRAD

goscnp2 ph31

```

```

p16:gp2
d16

;n_trhn

d12 p119:f1 p13:f3
20u fq=cnst22(bf ppm):f1

"cnst31=(130%2)*90"

3m ip18+cnst31

11 d61

(p39:sp4 ph17):f1

if "10 %2 == 1"
{
(p43:sp44 ph13+ph18):f2
}
else
{
(p43:sp44 ph14+ph18):f2
}
d32
(p39:sp4 ph1+ph18):f1
d32

(p21 ph15+ph18):f3
DELTA11
p16:gp3
d16
4u
(center (p42:sp42 ph1+ph18):f2 (p22 ph1+ph18):f3 )
4u
p16:gp3
d16
DELTA11
(p21 ph1+ph18):f3

(p43:sp43 ph2+ph18):f2
DELTA12
p16:gp4
d16
4u
(center (p42:sp42 ph1+ph18):f2 (p22 ph1+ph18):f3 )
4u
p16:gp4
d16
DELTA13
4u BLKGRAD
(p43:sp44 ph16+ph18):f2

if "130 %2 == 0"
{
DELTA14
(p12:sp27 ph1):f1
DELTA14
(p22 ph1):f3
DELTA14
(p12:sp27 ph17):f1
DELTA14 p119:f1
}
else
{
(p12:sp27 ph1):f1
DELTA14
DELTA14
(p22 ph1):f3
}

```

```

    (p12:sp27 ph17):f1
    DELTA14
    DELTA14 p119:f1
  }

goscnp3 ph30 cpd1:f1

4u do:f1
lo to 11 times 4

50u UNBLKGRAD

p16:gp5
dl6

;con

20u p13:f3
20u fq=cnst21(bf ppm):f1

(p11:sp23 ph1)
d23
(center (p12:sp24 ph1) (p22 ph1):f3 )
d23
(p11:sp25 ph2)

p16:gp1
dl6

(p21 ph3):f3
d31
(center (p8:sp13 ph5):f1 (p44:sp3 ph1):f2 )
d31
(p22 ph1):f3
DELTA
(center (p8:sp13 ph1):f1 (p44:sp3 ph7):f2 )
DELTA
(p21 ph4):f3

if "l10 %2 == 1"
{
  (p11:sp23 ph1)
  DELTA1
  (p12:sp26 ph1)
  DELTA1
  (center (p12:sp24 ph1) (p22 ph1):f3 )
  DELTA1
  (p12:sp26 ph1)
  DELTA1 p116:f3
}
else
{
  (p11:sp23 ph6)
  d22
  (p12:sp26 ph1)
  DELTA2
  (center (p12:sp24 ph1) (p22 ph1):f3 )
  DELTA1
  DELTA1 p116:f3
  (p12:sp26 ph1)
}

4u BLKGRAD

gol=2 ph31 cpd3:f3

dl1 do:f3 wr1 #0 if1 #0 zd1
3m wr2 #1 if2 #1 zd2

;exp_f1 ipap

```



```

3m iu10
lo to 3 times 2

;exp_f1 phase
3m ip3
lo to 4 times 2

;exp_f1 delay
3m id31
lo to 5 times 113

3m wr3 #2 if3 #2 zd3

3m iu30
lo to 6 times 2

;exp_f2 phase
3m ip15*2
3m ip16*2
3m iu0
lo to 7 times 2

;exp_f2 delay
3m id32
3m
20u
3m
3m
lo to 8 times 13

exit

ph1=0
ph2=1
ph3=0 2
ph4=0 0 0 0 2 2 2 2
ph5=0 0 2 2
ph6=3
ph7=2
ph13=1 3 0 2
ph14=3 1 0 2
ph15=1
ph16=0
ph17=2
ph18=0
ph30=0 2 3 1
ph31=0 2 0 2 2 0 2 0

;p11 : f1 channel - power level for pulse (default)
;p13 : f3 channel - power level for pulse (default)
;p116: f3 channel - power level for CPD/BB decoupling
;p119: f1 channel - power level for CPD/BB decoupling
;sp3 : f2 channel - shaped pulse 180 degree (Bip720,50,20.1)
;sp4 : f1 channel - shaped pulse 180 degree (Bip720,100,10.1)
;sp13: f1 channel - shaped pulse 180 degree (adiabatic)
;sp23: f1 channel - shaped pulse 90 degree (on resonance)
;sp24: f1 channel - shaped pulse 180 degree (on resonance)
;sp25: f1 channel - shaped pulse 90 degree (on resonance)
;
; for time reversed pulse
;sp26: f1 channel - shaped pulse 180 degree (Ca off resonance)
;sp27: f1 channel - shaped pulse 180 degree (C=O off resonance)
;sp42: f2 channel - shaped pulse 180 degree (Reburp.1000)
;sp43: f2 channel - shaped pulse 90 degree (Eburp2.1000)
;sp44: f2 channel - shaped pulse 90 degree (Eburp2tr.1000)
;p8 : f1 channel - 180 degree shaped pulse for inversion (adiabatic)
;p11: f1 channel - 90 degree shaped pulse
;p12: f1 channel - 180 degree shaped pulse

```

```

;p16: homospoil/gradient pulse [1 msec]
;p21: f3 channel - 90 degree high power pulse
;p22: f3 channel - 180 degree high power pulse
;p39: f1 channel - 180 degree shaped pulse for refocussing
;
; Bip720,100,10.1 (160us at 600.13 MHz)
;p42: f2 channel - 180 degree shaped pulse for refocussing
;
; Reburp.1000 (1.4ms at 600.13 MHz)
;p43: f2 channel - 90 degree shaped pulse for excitation
;
; Eburp2.1000/Eburp2tr.1000 (1.7ms at 600.13 MHz)
;p44: f2 channel - 180 degree shaped pulse for refocussing
;
; Bip720,50,20.1 (200us at 600.13 MHz)
;d1 : relaxation delay; 1-5 * T1
;d11: delay for disk I/O [30 msec]
;d12: delay for power switching [20 usec]
;d16: delay for homospoil/gradient recovery
;d22: 1/(4J(COCa)) [4.5 msec]
;d23: 1/(4J(NCO)) [16.6 msec]
;d26: 1/(4J)NH
;d31: incremented delay (2D, exp 1) [3 usec]
;d32: incremented delay (2D, exp 3): in32/2-p39/2
;d58: = AQ of exp 2 (1H)
;d59: = AQ of exp 3 (15N)
;d60: relaxation delay as executed
;d61: relaxation delay of exp 3
;cnst4: = J(NH)
;cnst21: CO chemical shift (offset, in ppm)
;cnst22: Calpha chemical shift (offset, in ppm)
;cnst26: Call chemical shift (offset, in ppm) [101 ppm]
;cnst43: compensation of chemical shift evolution during p43
;
; Eburp2.1000: 0.69
;cnst52: scaling factor for p42 to compensate for transition region
;
; Reburp.1000: 1.426
;cnst53: scaling factor for p43 to compensate for transition region
;
; Eburp2.1000: 1.000
;cnst54: H(N) chemical shift (offset, in ppm)
;cnst55: H(N) bandwidth (in ppm)
;o1p: CO chemical shift (cnst21)
;l0 : flag to switch between odd and even increments (exp 3)
;l3 : loop for 2D exp 1 = 122/2
;l13: loop for 2D exp 3 = td1/(8*122)
;l10: flag to switch between inphase and antiphase (exp 1)
;l22: number of experiments in F1 for exp 3 (after IPAP processing)
;l30: flag to switch between inphase and antiphase (exp 3)
;inf1: 1/SW(N) = 2 * DW(N)
;in2 : 1/SW(H) = 2 * DW(H) (inf1 of exp 3)
;in31: 1/(2 * SW(N)) = DW(N)
;nd31: 2
;in32: 1/(2 * SW(H)) = DW(H)
;nd32: 2
;ns: 8 * n
;ds: >= 32
;td1: number of experiments in F1
;
; (number of experiments after IPAP processing)
;FnMODE: States-TPPI (or TPPI) in F1
;cpd1: decoupling according to sequence defined by cpdprg1: p5m4sp180
;cpd3: decoupling according to sequence defined by cpdprg3
;pcpd1: f1 channel - 180 degree pulse for decoupling sequence
;pcpd3: f3 channel - 90 degree pulse for decoupling sequence
;cpdprg1: p5m4sp180

;for z-only gradients:
;gpz1: 50%
;gpz2: 70%
;gpz3: 31%
;gpz4: 19%
;gpz5: 45%

;use gradient files:
;gpnam1: SMSQ10.100

```

```
;gpnam2: SMSQ10.100  
;gpnam3: SMSQ10.100  
;gpnam4: SMSQ10.100  
;gpnam5: SMSQ10.100
```

```
;use AU-program splitcomb [ipap 2] to process data
```

```
;Processing (exp 3)
```

```
;PHC0(F1): 90  
;PHC1(F1): -180  
;FCOR(F1): 1
```

```
;$Id: $
```

Pulse sequence CON//**(H)**CAN

```
;ut_con_n_(h)can
;avance-version

;Dataset 1 (F1)
;n_hcan.2
;avance-version
;2D sequence with
; 15N detected correlation for triple resonance using
;   inept transfer steps
;
;   F3(Ha) -> F2(Ca,t1) -> F1(N,t2)
;
;on/off resonance 13C pulses using shaped pulses
;phase sensitive (t1)
;using constant time in t1
;(use parameterset )
;
;M. Gal, K. A. Edmonds, A. G. Milbradt, K. Takeuchi & G. Wagner,
;  J Biomol NMR 51, 497-504 (2011)
;(K. Takeuchi, G. Heffron, Z.-Y. J. Sun, D. P. Frueh & G. Wagner,
;  J Biomol NMR 47, 271-282 (2010))
;
;$CLASS=HighRes
;$DIM=2D
;$TYPE=
;$SUBTYPE=
;$COMMENT=

;Dataset 2 (F2)
;c_con_iasq
;avance-version
;CON
;2D sequence with
; 13C detected correlation for triple resonance using
;   inept transfer steps
;
;   F2(C=O) -> F1(N,t1) -> F2(C=O,t2)
;
;on/off resonance 13C pulses using shaped pulses
;phase sensitive (t1)
;using IPAP scheme for virtual decoupling
;(use parameterset )
;
;W. Bermel, I. Bertini, L. Duma, I.C. Felli, L. Emsley, R. Pierattelli,
;  P.R. Vasos, Angew. Chem. Int. Ed. 44, 3089-3092 (2005)
;(L. Duma, S. Hediger, A. Lesage & L. Emsley,
;  J. Magn. Reson. 164, 187-195 (2003) )
;
;$CLASS=HighRes
;$DIM=2D
;$TYPE=
;$SUBTYPE=
;$COMMENT=

prosol relations=<triple>

#include <Avance.incl>
#include <Delay.incl>
#include <Grad.incl>

"p2=p1*2"
"p22=p21*2"
"d3=1s/(cnst2*cnst12)"
```

```

"d4=1s/(cnst2*4)"
"d11=30m"
"d12=20u"

"d22=4.5m"
"d23=12.4m"
"d25=16.6m"
"d27=14.3m"

"d32=3u"
"d31=d27/2-d3-p22/2"
"d41=d27/2-larger(p14,p22)/2"
"d51=d23-d27/2-p14/2-p2/2"

"in31=in1/4"
"in32=in2/2"

"in41=in31"
"in51=in31"

"td1=tdmax(td1,d41*2,in41)"

"d1=(d60+d58)*3+d59+50m"

"DELTA1=d27-d23-p2/2"
"DELTA2=d3-p14/2-p22/2"
"DELTA3=d25/2"
"DELTA4=d23-d25/2-p14"
"DELTA5=d25-d23"
"DELTA6=d25/2-d12+p1*2/PI-de-4u"
"DELTA7=d32*2+p8"
"DELTA8=d23/2-p14/2"
"DELTA9=d23-d22-p14"

"spoff2=0"
"spoff3=0"
"spoff5=bf2*((cnst21-cnst22)/1000000)"
"spoff7=bf2*((cnst22-cnst21)/1000000)"
"spoff8=0"
"spoff13=bf2*((cnst26-cnst21)/1000000)"
"spoff23=bf2*((cnst23-cnst22)/1000000)"

"l0=0"
"l10=0"

"l3=td1/4"

"acqt0=0"
baseopt_echo

1 4u zel
   4u ze2
   d1

2 d11 do:f1 do:f3
   3m
3 3m
4 9m
5 3m
6 9m
7 3m

;n_hcan

```

```

8 d11 do:f2 do:f3
d60

50u UNBLKGRAD
4u p11:f1 p13:f3

(p1 ph1):f1
p16:gp1
d16 fq=cnst22(bf ppm):f2

(p21 ph1):f3
d4
(center (p14:sp3 ph1):f2 (p22 ph1):f3 )
d4
(p21 ph2):f3

(p13:sp2 ph3):f2
d41
(center (p14:sp5 ph1):f2 (p22 ph1):f3 )
d41
(p14:sp23 ph1):f2
DELTA1
(p2 ph1):f1
d51
(p14:sp5 ph10):f2
DELTA2
(p22 ph1):f3
d31
(p13:sp8 ph4):f2

4u
p16:gp2
d16 p116:f3
20u cpd3:f3

if "10 %2 == 0"
{
(p1 ph5):f1
DELTA3
(p14:sp5 ph1):f2
DELTA4
(p14:sp3 ph1):f2
DELTA5
(p2 ph6):f1
DELTA3
(p14:sp5 ph10):f2
}
else
{
(p1 ph7):f1
(p14:sp5 ph1):f2
DELTA3
DELTA4
(p14:sp3 ph1):f2
DELTA5
(p2 ph6):f1
(p14:sp5 ph10):f2
DELTA3
}

DELTA6
d12 p128:f2
4u BLKGRAD

goscnp1 ph31 cpd2:f2
lo to 8 times 3

;c_con

```

```

4u do:f2 do:f3
50m
50u UNBLKGRAD

p16:gp3
d16 fq=cnst21(bf ppm):f2

4u p11:f1 p116:f3
4u cpd3:f3

(p13:sp2 ph1):f2
d23
(center (p2 ph1):f1 (p14:sp3 ph1):f2 )
d23
(p13:sp8 ph2):f2

p16:gp4
d16

(p1 ph13):f1
d32
(p8:sp13 ph15):f2
d32
(p2 ph1):f1
DELTA7
(p1 ph14):f1

if "l10 %2 == 0"
{
(p13:sp2 ph1):f2
DELTA8
(p14:sp7 ph1):f2
DELTA8
(center (p2 ph1):f1 (p14:sp3 ph1):f2 )
DELTA8
(p14:sp7 ph1):f2
DELTA8 p119:f1
}
else
{
(p13:sp2 ph16):f2
d22
(p14:sp7 ph1):f2
DELTA9
(center (p2 ph1):f1 (p14:sp3 ph1):f2 )
DELTA8
DELTA8 p119:f1
(p14:sp7 ph1):f2
}

4u BLKGRAD

go2=2 ph30 cpd1:f1

d11 do:f1 do:f3 wr2 #1 if2 #1 zd2

;exp_f2 ipap
3m iu10
lo to 3 times 2

;exp_f2 phase
3m dp13
lo to 4 times 2

3m wr1 #0 if1 #0 zd1

;exp_f1 ipap
3m iu0

```

```

;exp_f2 delay
  3m id32
  lo to 5 times 2

;exp_f1 phase
  3m ip3
  lo to 6 times 2

;exp_f1 delay
  3m id31
  3m dd41
  3m id51

  lo to 7 times 13

exit

ph1=0
ph2=1
ph3=0 2
ph4=0 0 0 0 2 2 2 2
ph5=0 0 2 2
ph6=0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2
ph7=1 1 3 3
ph10=2

ph13=0 2
ph14=0 0 0 0 2 2 2 2
ph15=0 0 2 2
ph16=3

ph29=0
ph30=0 2 0 2 2 0 2 0
ph31=0 2 2 0 2 0 0 2

;p11 : f1 channel - power level for pulse (default)
;p12 : f2 channel - power level for pulse (default)
;p13 : f3 channel - power level for pulse (default)
;p116: f3 channel - power level for CPD/BB decoupling
;p119: f1 channel - power level for CPD/BB decoupling
;p128: f2 channel - power level for CPD/BB decoupling
;sp2: f2 channel - shaped pulse 90 degree (on resonance)
;sp3: f2 channel - shaped pulse 180 degree (on resonance)
;sp5: f2 channel - shaped pulse 180 degree (C=O off resonance)
;sp7: f2 channel - shaped pulse 180 degree (Ca off resonance)
;sp8: f2 channel - shaped pulse 90 degree (on resonance)
;
;      for time reversed pulse
;sp13: f2 channel - shaped pulse 180 degree (adiabatic)
;sp23: f2 channel - shaped pulse 180 degree (Cali off resonance)
;p1 : f1 channel - 90 degree high power pulse
;p2 : f1 channel - 180 degree high power pulse
;p8 : f2 channel - 180 degree shaped pulse for inversion (adiabatic)
;p13: f2 channel - 90 degree shaped pulse
;p14: f2 channel - 180 degree shaped pulse
;p16: homospoil/gradient pulse [1 msec]
;p21: f3 channel - 90 degree high power pulse
;p22: f3 channel - 180 degree high power pulse
;d1 : relaxation delay as executed; 1-5 * T1
;d3 : 1/(n*J(CH)), n = cnst12
;d4 : 1/(4J(CH))
;d11: delay for disk I/O [30 msec]
;d12: delay for power switching [20 usec]
;d16: delay for homospoil/gradient recovery
;d22: 1/(4J(COCa)) [4.5 msec]
;d23: 1/(4J(NCO)) [12.4 msec]
;d25: 1/(4J'(NCO)) [16.6 msec]
;d27: 1/(2J(CaCb)) [14.3 msec]
;d31 : incremented delay (2D, exp 1) = d27/2-d3-p22/2

```



```

;d32 : incremented delay (2D, exp 2) [3 usec]
;d41: decremented delay (2D, exp 1) = d27/2-larger(p14,p22)/2
;d51: incremented delay (2D) = d23-d27/2-p14/2-p2/2
;d58: = AQ of exp 1 (15N)
;d59: = AQ of exp 2 (13C)
;d60: relaxation delay for exp 1
;cnst2: = J(CH) [145 Hz]
;cnst12: for multiplicity selection = 4 for all but Gly, 5 including Gly
;cnst21: CO chemical shift (offset, in ppm)
;cnst22: Calpha chemical shift (offset, in ppm)
;cnst23: Caliphatic chemical shift (offset, in ppm)
;cnst26: Call chemical shift (offset, in ppm) [101 ppm]
;olp: CO chemical shift (cnst21)
;l0 : flag to switch between inphase and antiphase (exp 1)
;l3 : loop for 2D experiment = td1/4
;l10: flag to switch between inphase and antiphase (exp 2)
;inf1: 1/SW(Ca) = 2 * DW(Ca)
;in2 : 1/SW(N) = 2 * DW(N) (inf1 of exp 2)
;in31: 1/(4 * SW(Ca)) = (1/2) DW(Ca)
;nd31: 4
;in32: 1/(2 * SW(N)) = DW(N)
;nd32: 2
;in42: = in32
;in52: = in32
;ns: 4 * n
;ds: >= 32
;td1: number of experiments in F1
; (number of experiments after IPAP processing)
;FnMODE: States-TPPI (or TPPI) in F1
;cpd1: decoupling according to sequence defined by cpdprg1
;cpd2: decoupling according to sequence defined by cpdprg2: p5m4sp180
;cpd3: decoupling according to sequence defined by cpdprg3
;pcpd1: f1 channel - 90 degree pulse for decoupling sequence
;pcpd2: f2 channel - 180 degree pulse for decoupling sequence
;pcpd3: f3 channel - 90 degree pulse for decoupling sequence

;for z-only gradients:
;gpz1: 50%
;gpz2: 40%
;gpz3: 60%
;gpz4: 30%

;use gradient files:
;gpnam1: SMSQ10.100
;gpnam2: SMSQ10.100
;gpnam3: SMSQ10.100
;gpnam4: SMSQ10.100

;use AU-program splitcomb [ipap 2] to process data

;$Id: $

```

8. The ambivalent role of proline residues in an intrinsically disordered protein: from disorder promoters to compaction facilitators



The Ambivalent Role of Proline Residues in an Intrinsically Disordered Protein: From Disorder Promoters to Compaction Facilitators

Borja Mateos¹, Clara Conrad-Billroth¹, Marco Schiavina², Andreas Beier¹, Georg Kontaxis¹, Robert Konrat¹, Isabella C. Felli² and Roberta Pierattelli²

1 - Department of Structural and Computational Biology, University of Vienna, Max Perutz Labs, Vienna Biocenter Campus 5, 1030 Vienna, Austria

2 - CERM and Department of Chemistry "Ugo Schiff", University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Florence, Italy

Correspondence to Robert Konrat, Isabella C. Felli and Roberta Pierattelli: Robert Konrat. Fax: +43 1 4277 852202. Isabella C. Felli. Fax: +39 055 4574923. Roberta Pierattelli. Fax: +39 055 4574923. robert.konrat@univie.ac.at, felli@cerm.unifi.it, roberta.pierattelli@unifi.it
<https://doi.org/10.1016/j.jmb.2019.11.015>

Edited by Richard W. Kriwacki

Abstract

Intrinsically disordered proteins (IDPs) carry out many biological functions. They lack a stable three-dimensional structure, but rather adopt many different conformations in dynamic equilibrium. The interplay between local dynamics and global rearrangements is key for their function. In IDPs, proline residues are significantly enriched. Given their unique physicochemical and structural properties, a more detailed understanding of their potential role in stabilizing partially folded states in IDPs is highly desirable. Nuclear magnetic resonance (NMR) spectroscopy, and in particular ¹³C-detected NMR, is especially suitable to address these questions. We applied a ¹³C-detected strategy to study Osteopontin, a largely disordered IDP with a central compact region. By using the exquisite sensitivity and spectral resolution of these novel techniques, we gained unprecedented insight into *cis*-Pro populations, their local structural dynamics, and their role in mediating long-range contacts. Our findings clearly call for a reassessment of the structural and functional role of proline residues in IDPs. The emerging picture shows that proline residues have ambivalent structural roles. They are not simply *disorder promoters* but rather can, depending on the primary sequence context, act as nucleation sites for structural compaction in IDPs. These unexpected features provide a versatile mechanistic toolbox to enrich the conformational ensembles of IDPs with specific features for adapting to changing molecular and cellular environments.

© 2019 Elsevier Ltd. All rights reserved.



The Ambivalent Role of Proline Residues in an Intrinsically Disordered Protein: From Disorder Promoters to Compaction Facilitators

Borja Mateos¹, Clara Conrad-Billroth¹, Marco Schiavina², Andreas Beier¹, Georg Kontaxis¹, Robert Konrat¹, Isabella C. Felli² and Roberta Pierattelli²

1 - Department of Structural and Computational Biology, University of Vienna, Max Perutz Labs, Vienna Biocenter Campus 5, 1030 Vienna, Austria

2 - CERM and Department of Chemistry “Ugo Schiff”, University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Florence, Italy

Correspondence to Robert Konrat, Isabella C. Felli and Roberta Pierattelli: Robert Konrat. Fax: +43 1 4277 852202. Isabella C. Felli. Fax: +39 055 4574923. Roberta Pierattelli. Fax: +39 055 4574923. robert.konrat@univie.ac.at, felli@cerm.unifi.it, roberta.pierattelli@unifi.it
<https://doi.org/10.1016/j.jmb.2019.11.015>

Edited by Richard W. Kriwacki

Abstract

Intrinsically disordered proteins (IDPs) carry out many biological functions. They lack a stable three-dimensional structure, but rather adopt many different conformations in dynamic equilibrium. The interplay between local dynamics and global rearrangements is key for their function. In IDPs, proline residues are significantly enriched. Given their unique physicochemical and structural properties, a more detailed understanding of their potential role in stabilizing partially folded states in IDPs is highly desirable. Nuclear magnetic resonance (NMR) spectroscopy, and in particular ¹³C-detected NMR, is especially suitable to address these questions. We applied a ¹³C-detected strategy to study Osteopontin, a largely disordered IDP with a central compact region. By using the exquisite sensitivity and spectral resolution of these novel techniques, we gained unprecedented insight into *cis*-Pro populations, their local structural dynamics, and their role in mediating long-range contacts. Our findings clearly call for a reassessment of the structural and functional role of proline residues in IDPs. The emerging picture shows that proline residues have ambivalent structural roles. They are not simply *disorder promoters* but rather can, depending on the primary sequence context, act as nucleation sites for structural compaction in IDPs. These unexpected features provide a versatile mechanistic toolbox to enrich the conformational ensembles of IDPs with specific features for adapting to changing molecular and cellular environments.

© 2019 Elsevier Ltd. All rights reserved.

Introduction

Intrinsically disordered proteins (IDPs) exist as ensembles of rapidly interconverting conformers, therefore cannot be described by a single three-dimensional structure [1–3]. Their unique physicochemical properties derive from the polymer-like properties [4] of polypeptide chains in combination with the specific characteristics of the interaction pattern resulting from the primary sequence. Among the naturally occurring amino acids, proline has peculiar properties due to its cyclic structure, which links the backbone nitrogen with its side chain generating a pyrrolidine ring that constrains its available conformational space. These steric

restraints reduce the energy difference between *cis/trans* conformers involving peptide bonds between proline residues and the amino acids preceding them (Xaa-Pro peptide bonds). Although for noncyclic amino acids the occurrence of the *cis* conformation is less than a 0.5%, the cyclic nature of prolines slightly disfavors the *trans* conformation, leading to increased *cis* populations, typically ranging from 5 to 10% [5,6]. The energetic barrier in *cis* and *trans* interconversion is estimated to be around 84 kJ/mol⁻¹, high enough to have distinct separate species with a significant lifetime (i.e., interconversion rates of 10⁻³ s⁻¹), existing in “slow” exchange. Peptide bonds involving prolines differ substantially from those involving other amino acids and among

other factors [7,8] play an important role in protein folding processes [9].

The role of proline residues in highly flexible disordered polypeptide chains has been a long discussed topic and has involved many hypotheses [9–13]. The peculiarities of this amino acid have been a focus of study due to its pivotal role in protein–protein interactions [14], backbone rearrangements [8,15], posttranslational modifications [16,17], and secondary structure formation [18–21]. Experimental data, however, were sparse mainly because of methodological limitations.

Nuclear magnetic resonance (NMR) is a unique tool to characterize the structure and dynamics of proteins at atomic resolution. Most methods used in the study of biomolecules rely on the detection of ^1H , which offers better inherent sensitivity because of its higher gyromagnetic ratio compared with ^{13}C and ^{15}N . The general approach for a well-structured protein consists in exploiting the amide proton ($^1\text{H}^{\text{N}}$) in combination with the backbone nitrogen (^{15}N) as readout because of its ease of detection and the resulting favorable overall cross-peak dispersion that allows to observe the “fingerprint” of a protein of interest by simple 2D spectra. However, this method has limitations when dealing with IDPs, in particular when approaching physiological conditions. On one side, $^1\text{H}^{\text{N}}$ is highly sensitive to its environment, in particular to pH and temperature, which may compromise its use for studies under such conditions due to rapid exchange with the solvent. In addition, it does not provide information about proline residues, which lack the amide proton. This also complicates their sequence-specific assignment, although approaches have been developed to partially “by-pass” this limitation [22,23]. Alternative detection schemes for prolines involve aliphatic proton–carbon (^1H - $^{13}\text{C}^{\text{ali}}$) correlation spectroscopy [12,16,24]. Despite many observable correlations ($^1\text{H}^{\alpha}$ - $^{13}\text{C}^{\alpha}$ and $^1\text{H}^{\text{ali}}$ - $^{13}\text{C}^{\text{ali}}$), these experiments have very poor chemical shift dispersion characteristics and are largely insensitive to their backbone environment. Moreover, the large number of proline residues in IDPs leads to extensive cross-peak overlap that hampers their investigation at atomic resolution.

An attractive option to obtain NMR spectroscopy information on proline residues is the use of direct ^{13}C observation [25]. To our knowledge, the first reference making use of direct ^{13}C -detection for the study of prolines dates back to 1973 and was proposed by the Bovey group [26] and few years later, in 1976, by the Wüthrich group [27]. However, this approach was abandoned because of the low sensitivity of these experiments. The advances in obtaining ^{13}C -/ ^{15}N -enriched recombinant proteins and the development of cryogenically cooled NMR probeheads for the direct detection of heteronuclei [28], which have boosted sensitivity, allowed the

viability of direct ^{13}C and ^{15}N detection to study macromolecules [29–33]. More interestingly, as the chemical shifts of an NMR cross-peak depend on the two nuclei involved in the peptide bond, the 2D correlation of carbonyl carbon and amide nitrogen nuclear spins across the peptide bond ($^{13}\text{C}'_i$ - $^{15}\text{N}_{i+1}$) result in exquisite signal dispersion in the so-called CON spectrum (Fig. 1A) [34,35]. This is largely independent of tertiary structure and provides well-resolved information about all amino acids including proline, even in IDPs. In other words, CON spectra provide the most complete two-dimensional fingerprint of an IDP, and they are a key tool for their complete characterization [36,37].

Here we make use of direct ^{13}C -detection tools [38–42] to fully characterize the *cis/trans* equilibrium of proline residues in *Coturnix japonica* Osteopontin (OPN), an extracellular IDP that exerts its function by binding to numerous components of the extracellular matrix (ECM), including two types of membrane receptors: integrins and CD44. Its activity is highly regulated by posttranslational modifications (phosphorylation, glycosylation, and protein cleavage), splicing of OPN and CD44, and several binding events with different extracellular ligands. The correct interpretation of the extracellular stimuli is key for cell behavior, and a deregulation of OPN has been related to metastatic and abnormal cell function [43–47]. OPN contains regions with significant compaction that enable the correct exposure of many of the interacting motifs, like the RGD motif (Fig. 1C) [48], indicating a potential role of these compact states for its function. Thus, OPN is an ideal case to study the role of proline residues in IDPs and the relevance of *cis/trans* conformers for stabilizing local conformation and long-range contacts.

Results

Observation and signal assignment of *cis* isomers of proline residues by ^{13}C -detection

$^{13}\text{C}'_{i-1}$ - $^{15}\text{N}_i$ cross-peaks, where i is a proline, due to their peculiar nitrogen chemical shifts, appear in a well-separated region of CON spectra (Fig. 1A). A careful inspection of the spectra reveals the existence of peaks due to minor species (Fig. 1B).

Given the low population of the minor species, special emphasis was put into optimizing the sensitivity of the standard assignment experiments exploiting ^{13}C detection (3D (H)CBCACON and 3D (H)CBCANCO) [29,35,49] (see Method). Signal assignment cannot be achieved simply by comparison with the major form chemical shifts because cross-peak displacements of the minor form relative to the main peaks are substantial and do not follow a

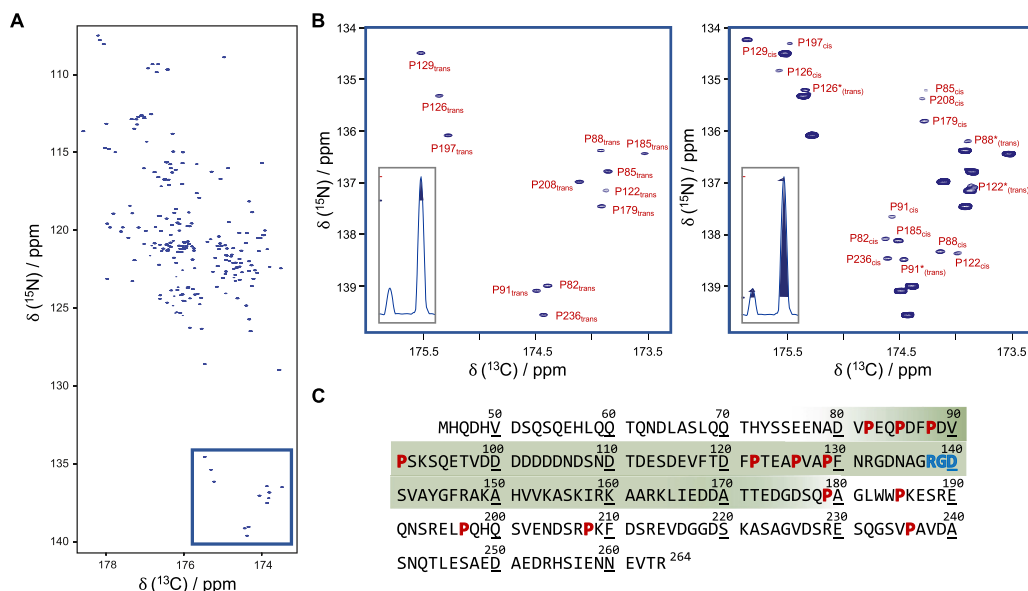


Fig. 1. Observation of prolines by ^{13}C -detection NMR. (A) ^{13}C -detected 2D CON spectrum recorded on *Coturnix japonica* Osteopontin. Note the good dispersion of cross-peaks and the presence of prolines signals around 134–140 ppm $\delta(^{15}\text{N})$. (B) Proline region in the 2D CON spectrum. The main forms of prolines are observed (left) and minor forms appear (right) when the contour levels are lowered. In the inset, the level of the 2D slice and the relative intensity of major and minor forms are reported. Tentative assignments lacking further connectivities are indicated with an asterisk. (C) Primary sequence of the *C. japonica* Osteopontin construct used in this work. Proline residues are indicated in red, the RGD motif (integrin binding site) is in blue. Underlined residues indicate amino acid spacing of ten residues (note that the sequence starts at residue M45). Residues located in the green box are part of the compact region in OPN. This compact state has extensively been characterized by paramagnetic relaxation enhancement (PRE) and interference (PRI) data, displaying cooperative unfolding under denaturing conditions [48].

simple regular pattern. Standard assignment strategies based on the correlation of C^α , C^β , and C' of consecutive residues (Fig. 2A) allow the assignment of most of the residues when spectral overlap is not pronounced. The 3D (H)CCCON [34] provides additional information about side chain shifts of aliphatic carbons (C^{all}), a key aspect to distinguish residues that have similar C^α and C^β chemical shifts. However, this strategy fails when dealing with similar sequence repeats. In particular, proline side chain carbons share very similar chemical shifts, and several ambiguities might remain during the sequence-specific assignment procedure. The 3D COCON experiment [38,39] is an excellent complementary experiment because it correlates carbonyl signals of three or more consecutive residues by transferring the magnetization across the $^3\text{J}_{\text{C}'\text{C}}$ couplings (0.5–3.0 Hz) via isotropic mixing. To detect cross-peaks involving the minor species, characterized by low signal-to-noise ratio, a sensitivity improvement was introduced into the original 3D COCON experiment. By taking advantage of the higher ^1H polarization, initial magnetization was

transferred from $^1\text{H}^\alpha$ to $^{13}\text{C}^\alpha$ and then to $^{13}\text{C}'$ before the beginning of the COCON experiment. This resulted in an approximate increase of the S/N ratio per unit time by a factor of 1.65 (Fig. S1A–B). The detection of sequential connectivities in the 3D (HCA)COCON (Fig. 2B) in combination with those observed in the 3D (H)CBCACON, 3D (H)CBCANCO and 3D (H)CCCON allowed the assignment of both major and minor forms in OPN. Under the present experimental conditions, the (HCA)COCON spectra gave the most complete information by correlating at least three (from $\delta(^{15}\text{N})$ i to $i+2$) residues in all cases and, in some cases even five residues (from $\delta(^{15}\text{N})$ $i-1$ to $i+3$). The signal intensity of the off-diagonal signals is directly dependent on the value(s) of the intervening $^3\text{J}_{\text{C}'\text{C}}$ coupling constants. In the subspectra associated with the minor forms, some of the cross-peaks connecting proline residues to their following residues show low to vanishing intensity (Fig. S1C). As the connections to the preceding residues are usually still present, this is not simply an issue of reduced spectral sensitivity reflecting the low effective concentration

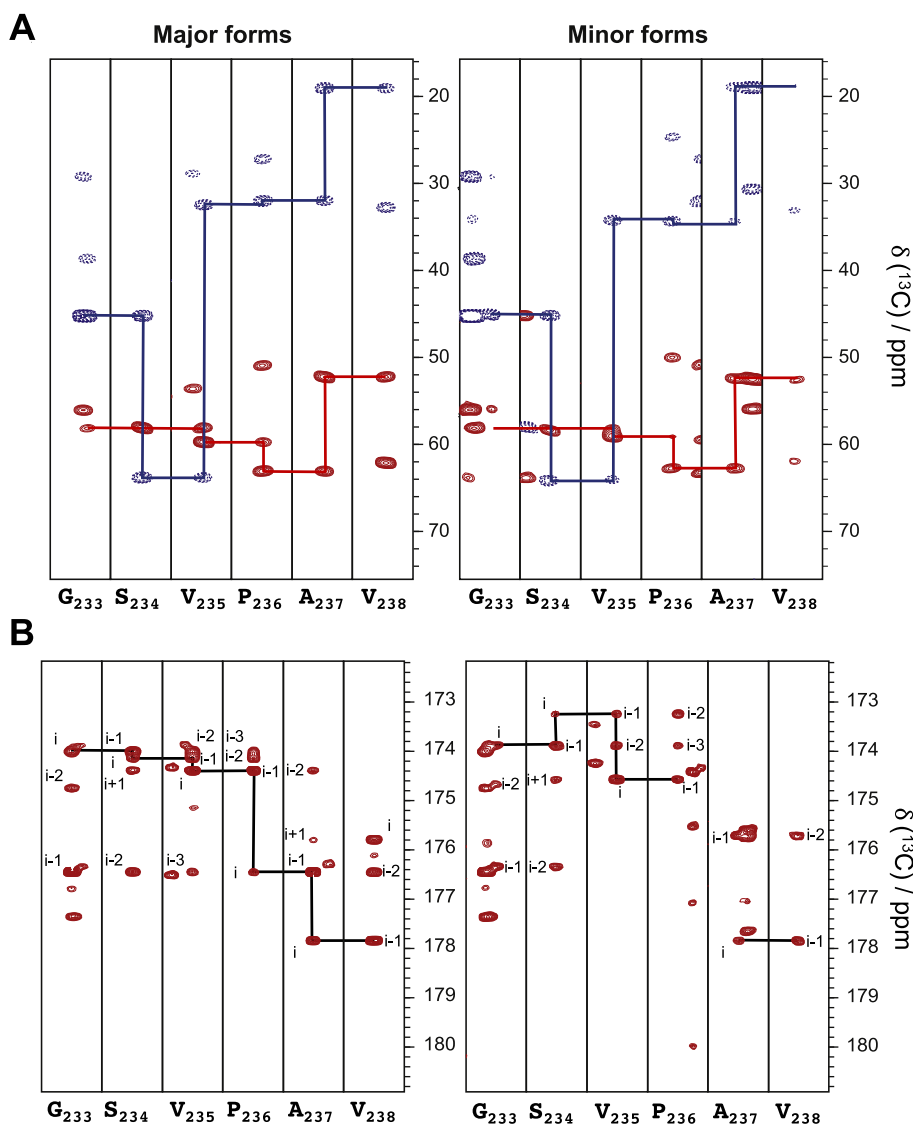


Fig. 2. Assignment of *cis* forms in prolines. Assignment strips for both *trans* and *cis* forms of *Coturnix japonica* Osteopontin OPN. (A) Strip plots of the regions 233–238 of the 3D (H)CBCACON allowing the connectivity of dipeptides by C ^{β} and C ^{α} carbons. The corresponding patterns of major and minor forms are very similar because they involve the same amino acid type in both *trans*-Pro and *cis*-Pro forms. (B) 3D (H)CACOCOC strip plots of region 233–238 showing the connectivity between at least three consecutive C' signals for both *trans* and *cis* forms. As the amplitude of $^3J_{C'C'}$ determines the intensity of the off-diagonal signals, the connectivity between a *cis*-proline (i-1) and its C-terminal neighbor (i) can have vanishingly low sensitivity because of a preferred φ angle of $[-60^\circ, -90^\circ]$ (see pair P236-A237).

of the minor species, but rather suggests a consistently small value for the $^3J_{C'iC'i+1}$ indicative of a specific backbone geometry. This is consistent with the cyclic nature of the Pro side chain limiting the $|\varphi|$

angles of the Pro residue (in position i) to values around $[-60^\circ, -90^\circ]$, thus corresponding to low values of the $^3J_{C'C'}$ coupling based on the Karplus equation [50,51] (Fig. 2B and Fig. S1D–F).

The chemical shifts of side chain carbon atoms of proline residues obtained through the analysis of the 3D (H)CCCON were very useful to confirm that the minor form observed derives from *cis/trans* isomers of Xaa-Pro peptide bonds. Indeed, as initially proposed by Schubert and others [52,53], *cis* conformations can be identified by the difference between C^β and C^γ chemical shifts, which is larger in the case of *cis*-Pro relative to *trans*-Pro. As shown in Fig. 3A, the minor forms observed in the proline region follow this pattern and are mostly in the *cis* conformation (Fig. S2).

Cis/trans isomerization at positions of proline residue also affects the neighboring residues, similarly to $^1H^N$ shift differences observed between *cis/trans*-induced forms in α -synuclein [11]. Specifically, we observed that (i) the percentage of *cis*-Pro conformers is conserved along the signals of minor forms of consecutive residues surrounding the proline (Fig. S3A) and (ii) the difference in chemical shifts between major and minor forms is large for residues $i\pm 1$ and gradually decreases until position $i\pm 3$, where it coincides with the position of the major form (Fig. S3B–C), indicating that proline isomerization has a drastic effect at the local level in a range spanning residues $i\pm 3$.

Another observation that can be easily depicted is the position of C^α and C^β shifts of residues preceding a proline (BMRB ID: 27443 and Table S4), which reflects the fact that residues preceding a Pro are biased toward an extended conformation. A typical indication of a proline in the following position is the upfield shift of C^α . This is easily recognized for all residues (see examples of V81, V90, F121, A125, A128, Q178, L196, R207 present in OPN). More interestingly, for some of these residues (V81, V90, F121, and R207), the C^β carbons experience a comparable downfield shift, thus indicating specific conformational changes at these sites.

To conclude, given the substantially improved sensitivity of the new pulse sequences, the CON cross-peaks of both *trans* and *cis* forms, including their side chains, of the 12 proline residues present in OPN could be assigned. In addition, the minor forms of several proline neighbors were assigned (11 out of 12 in the $i-1$ positions; 6 out of 12 in the $i+1$ positions; and some in $i\pm 2$ and $i\pm 3$ positions). The assignment of the minor forms can be found in Table S4. This also further extends the assignment already achieved for OPN (deposited with BMRB as ID 27443) comprising 99.1% of N, C', C^α , and C^β (major form).

Quantification of the *cis* populations

Relative populations of *trans* and *cis* species were determined by taking peak volumes from the ^{13}C -detected 2D CON spectra (Fig. 1B). It is well known

that prolines display high *cis* population, ranging from 5 to 10% in proteins, whereas even higher percentages have been observed for small peptides [54–56]. In addition, this equilibrium can be modulated by the chemical environment (for example, by the occurrence of phosphorylation [17] and/or by secondary structure propensities). In the case of OPN, experimental data clearly indicate that all proline residues sample both the *trans* and the *cis* conformation, with populations for *cis* conformers that significantly deviate from percentages reported for other IDPs (around 5–10%) like α -synuclein [11] or tau [12]. P88, P91, P122, and P126 also show an additional minor form with a *trans* signature (Figs. 1B and 3B) likely coupled with the *cis* conformation of neighboring proline residues.

It is worth noting that for P88, P122, P129, and P185, the population of the *cis* forms is greater than 10% (Fig. 3B), and this higher *cis* population correlates with the presence of aromatic residues (Phe or Trp) in either $i\pm 1$ position (Fig. 1C and Fig. S4A). Indeed, as an example, the mutation to alanine of phenylalanine 121 (preceding proline 122) induces a lower *cis* population of the peptide bond linking the two amino acids compared with the wild-type protein (from 13% to 7%, Fig. S5) confirming the importance of the direct interaction between aromatic and proline residues to stabilize the *cis* conformer.

The structural dynamics of *cis* forms

To further characterize the role of proline *cis/trans* isomerism and its impact on the structural dynamics of OPN, we measured ^{15}N relaxation parameters [57] (Figs. 4A and 5). A comparison of ^{15}N - R_2 relaxation parameters in the major and minor forms for residues other than proline showed comparable local dynamics in both states, although some specific changes occur (Fig. 4B). Interestingly, increased ^{15}N transverse relaxation rates were consistently found for residues within the region of residues 140–175. This region is known to be the most compact segment of OPN comprising most of the binding regions of this protein and housing both hydrophobic and charged residues to exploit for its stabilization [48]. In addition, residues in the regions 120–132 and 175–190 and flanking the core region of OPN also show reduced conformational dynamics resulting in higher ^{15}N transverse relaxation rates. Notably, the mutation of F121A reduces the ^{15}N - R_2 rates of these flanking regions accompanied by a moderate decompaction of the regions 140–175 (Fig. S5B). This dynamic pattern is largely conserved in the minor forms and thus suggests that there are no substantial global differences in dynamics between signals stemming from proline *cis/trans* isomerization. However, ^{15}N - R_2 and heteronuclear ^{15}N - $\{^1H\}$ NOEs (Fig. 4B) provide some

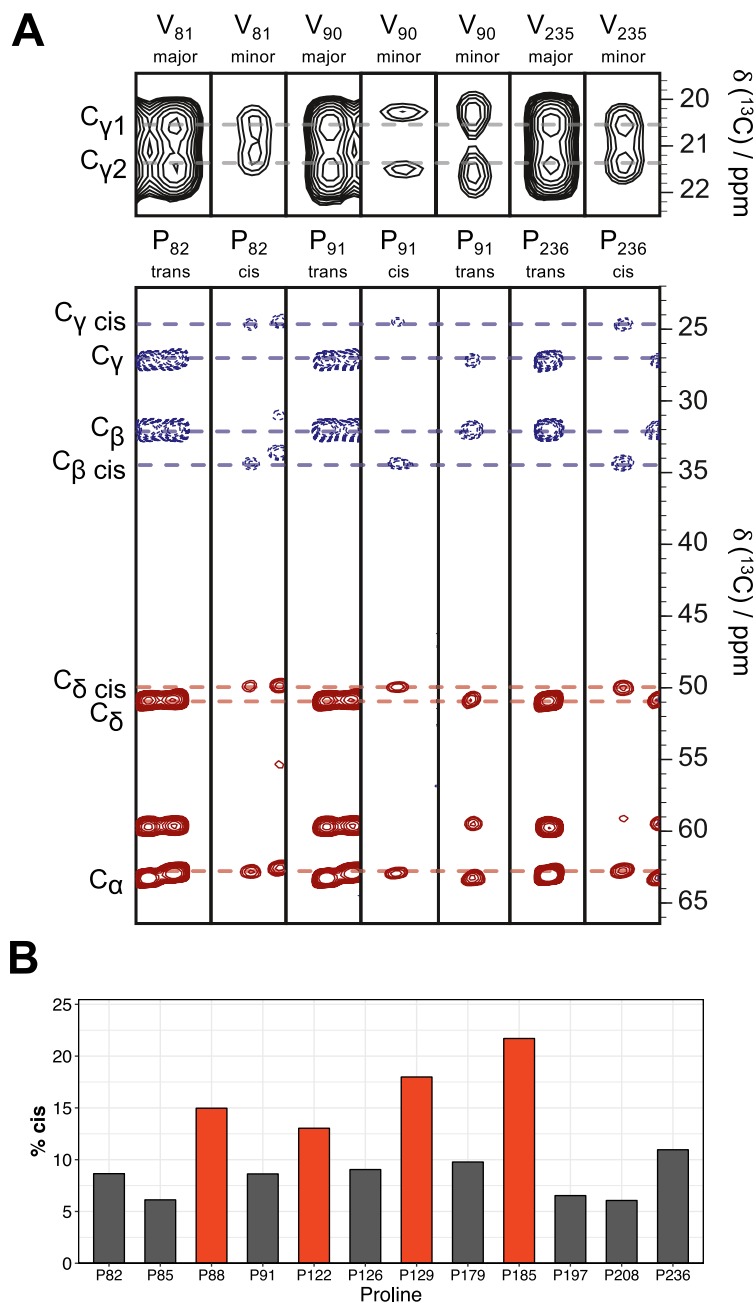


Fig. 3. Characterization of *cis*-Pro in Osteopontin. (A) Side chain C^γ shifts of valines (upper) and C^α , C^β , C^δ , and C^γ shifts of prolines (lower) of Val-Pro pairs in OPN. Characteristic C^α , C^β , and C^δ of *cis*/*trans* isoforms are indicated with dashed lines. (B) Percentage of *cis* population obtained by measuring the ratio of peak volumes between *cis* and *trans* forms for proline signals in *Coturnix japonica* Osteopontin. Orange bars correspond to proline residues i with an aromatic in position $i \pm 1$.

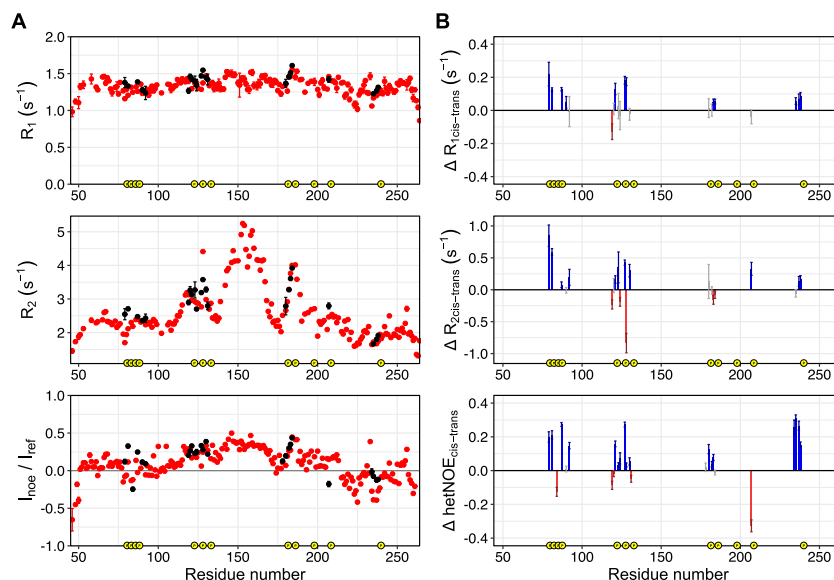


Fig. 4. ^{15}N relaxation of both major and minor forms. (A) 800 MHz ^{15}N - R_1 , ^{15}N - R_2 , and ^{15}N - $\{^1\text{H}\}$ NOE relaxation parameters of Osteopontin measured at 310 K for both major (red) and minor (black) forms of non-Pro residues. Error bars indicate the fitting errors (^{15}N - R_1 and ^{15}N - R_2) and the error propagation of intensity ratios based to the noise level (hetNOE). (B) Differences in ^{15}N relaxation parameters between minor and major forms. Statistically meaningless values are depicted in gray. Proline positions are indicated as yellow circles.

evidence for reduced local mobility in the *cis* forms presumably because of the tilt of the polypeptide chain. This is consistent with the observed increased ^{15}N - R_2 rates for the ^{15}N of *cis*-Pro (Fig. 5). Owing to the pronounced flexibility in IDPs, ^{15}N spin relaxation is typically dominated by local fluctuations on different timescales rather than the global tumbling of the protein [58–62]. In a fully extended conformation, ^{15}N transverse relaxation is dominated by fast timescale motions around the C^α - C^α direction [58,59,63–65], as described by the Gaussian axial fluctuation model [66]. Bending of the polypeptide chain due to formation of the *cis* form may restrict this dynamic mode and increase the effective local correlation time which would be accompanied by faster ^{15}N - R_2 relaxation. Several mechanisms could explain these differences in relaxation, i.e., interactions that generate local compaction. A plausible explanation could be the increase of conformations resembling those found in PPI helices in contrast to the conventional PPII form often sampled by proline-rich segments [67]. Although the changes in backbone dihedral angles are rather small, the switch from PPII to PPI results in a significant compaction of the polypeptide chain. Whereas PPII is an elongated left-handed helix, PPI is a compacted right-handed helix. The average helical pitch per residue is around 5.6 Å and thus very similar to the value found

in a typical α -helix. In contrast, the more extended PPII motif has a helical pitch of 9.4 Å. To test the hypothesis of a potential PPII-to-PPI switch, we performed a quantitative analysis of secondary structure propensities by the program $\delta 2\text{D}$ [68] separately using the shifts of major and minor forms, which shows that the *cis*-Pro leads to lower PPII populations compensated by either coil or β -strand conformations (Fig. 6). Interestingly, the largest changes in ^{15}N - R_2 rates were observed for P82, P85, P88, P122, P126, and P236. Strikingly, most of them (all except P236) belong to stretches containing consecutive P-X-X-P motifs and is thus more likely to form a continuous PPI helix. However, PPI–PPII transitions are challenging to pick up solely from chemical shift data and a cause–effect direct connection could not be obtained from these results alone.

On the role of proline isomerization in modulating global rearrangements

To probe putative changes in long-range contacts resulting from *cis/trans* isomerization of proline residues, we carried out a series of paramagnetic relaxation enhancement (PRE) experiments. We exploit ^{13}C -detected PRE by observing intensities of the paramagnetic and diamagnetic forms [69].

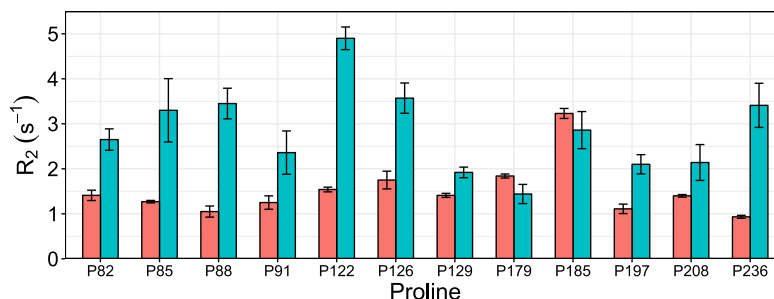


Fig. 5. ^{15}N - R_2 relaxation rates for *trans* prolines (orange) and *cis* prolines (cyan) in *Coturnix japonica* OPN measured with a ^{13}C -detection scheme. Note the increased R_2 values in *cis* prolines, especially for the residues in regions with an accentuated difference in PPII content between *cis* and *trans* conformers, as shown in Fig. S5.

$^1\text{H}^\alpha$ -start experiments are more sensitive to paramagnetic effects because protons, with their high gyromagnetic ratio compared with carbon, are reintroduced in the coherence transfer pathway to transfer coherence from protons to heteronuclei (^{13}C or ^{15}N). Thus, we used $^1\text{H}^\alpha$ -start experiments to measure ^{13}C -detected PRE to observe prolines and

get higher data set completeness. Overall, the PRE profiles of both major and minor forms obtained for the various spin label positions showed a very similar behavior (Fig. 7). Notable exceptions only occurred for residues located in the regions 120–132 (i.e., D120, F121, T123, and E124) which showed slightly larger PRE effects when the spin

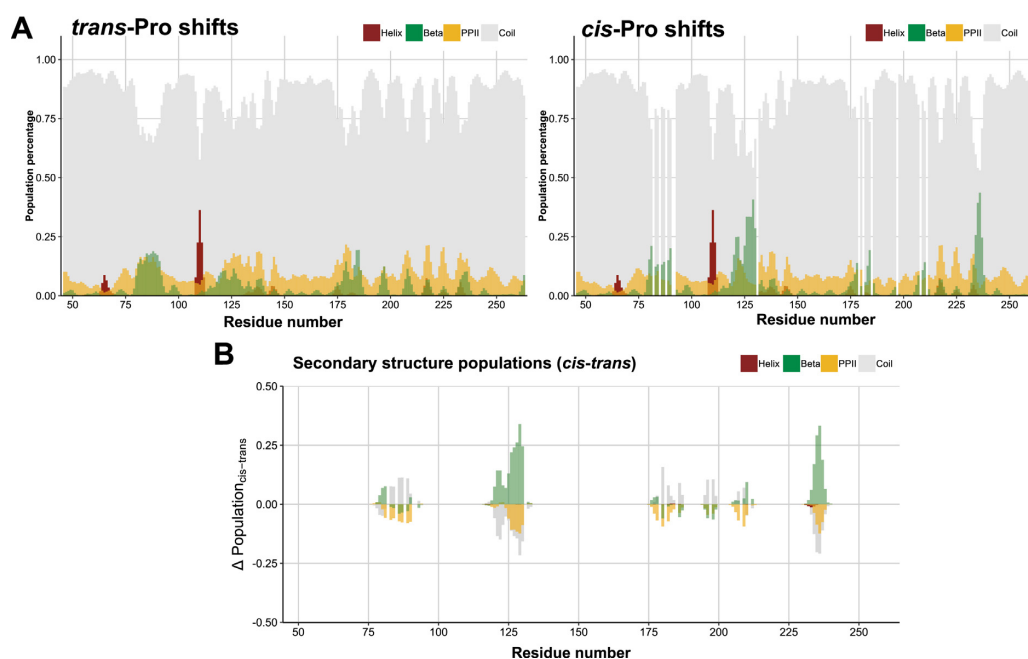


Fig. 6. Secondary structure populations from NMR shifts. (A) Secondary structure populations extracted using the $\delta^2\text{D}$ software [68] of the major forms arising from *trans*-Pro (left) or the minor forms arising from *cis*-Pro (right). (B) Secondary structure population differences. The secondary structure populations calculated with the minor forms shifts, arising from nearby *cis* prolines, were subtracted from the corresponding major forms arising from *trans* prolines. The minor forms are characterized by lower PPII populations compensated by either coil or β -strand conformations.

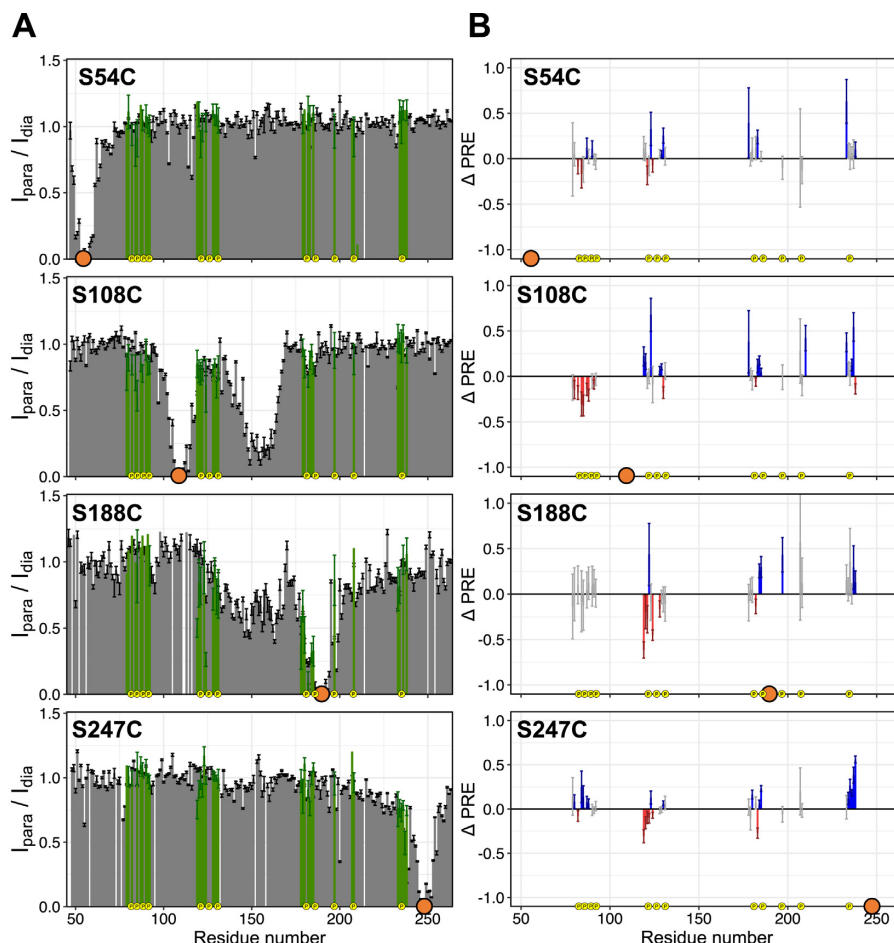


Fig. 7. ^{13}C -detected PRE experiments. (A) ^{13}C -detected PRE profiles (H^{α} -CON) of Cys mutants 54, 108, 188, and 247 (orange circles) at 310 K. Gray bars correspond to major forms and green bars to the minor ones. Error bars represent the quantification uncertainty based on the noise level. (B) Differences in ^{13}C -detected PREs between minor (stemming from *cis*-Pro) and major (stemming from *trans*-Pro) forms are plotted as a function of residue position. Negative values (red) indicate more effective PRE effects in the minor form (lower signal intensity due to larger paramagnetic effect), whereas positive differences are indicative of smaller paramagnetic relaxation. Statistically insignificant changes are depicted in gray. Proline positions are indicated as yellow circles. The position of the spin label is indicated by large orange circles.

label is located in position 188, as well as the regions 80–90 (paralleled by an increased ^{15}N R_2 rate) with the spin label at position 108.

In addition, putative changes in long-range correlations were also probed by measuring $^1\text{H}^{\text{N}}$ - R_2 rates using standard $^1\text{H}^{\text{N}}$ -detected experiments. In the interpretation of these rates, care has to be taken as $^1\text{H}^{\text{N}}$ - R_2 rates may also contain slight contributions from intermolecular exchange with bulk water. We find the largest differences between major (stemming from *trans*-Pro) and minor (stemming from *cis*-Pro) forms in the proline-rich regions located at

the termini (120–130 and 180–190) of the central region of OPN characterized by largest compaction (100–190) (Fig. S6A–B). Summing up, PRE rates were largely comparable in both Pro forms suggesting that the overall (tertiary) compaction is maintained and only local dynamics are different between *trans*-Pro and *cis*-Pro forms.

Finally, to obtain more specific information about differential side chain structural dynamics, we performed ^{15}N - $\{^1\text{H}\}$ heteronuclear NOE and ^{15}N 3D NOESY-HSQC experiments. We use the side chain signals of tryptophan residues in the stretch

(W¹⁸³-W¹⁸⁴-P¹⁸⁵) as proxies for the dynamic measurements (Fig. S9). We were particularly interested in the impact of proline *cis/trans* isomerization on the side chain dynamics of W183 and W184 as this would provide prototypical information about the consequences of Pro-aromatic side chain packing on local conformational dynamics. It can be clearly seen that minor (stemming from *cis*-Pro) and major (stemming from *trans*-Pro) forms of W183 and W184, which are directly adjacent to P185, display distinctly different conformational dynamics (Fig. S9). Whereas the two forms of W183 show nearly similar ¹⁵N-¹H} NOEs, the minor form of W184 is evidently different and shows significantly restricted motions in the (fast) ps timescale (Fig. 8A–B) and increased water exchange properties (Fig. S9C) while keeping similar ¹⁵N-T₁ and ¹⁵N-T₂ relaxation (Fig. S9A–B). In addition, different indole ¹H^N NOEs to protons of the side chains further corroborate the finding of changes in side chain orientations in the minor form of W184 (Fig. 8C). It can thus be concluded that steric repulsion due to distinct π ·CH-aromatic interactions lead to restricted local mobility in the *cis*-Pro form, as illustrated in Fig. 8D.

Discussion

Proline residues are highly abundant in IDPs, although their role is still elusive. A unique property of prolines in proteins is their significantly higher than other residues' tendency to form *cis* peptide bond isomers. The percentage of *cis*-Pro population is known to vary between 5 and 10% and has been reported for previously studied IDPs like α -synuclein [11] and tau [12]. In addition, this population can be drastically affected by phosphorylation [17] and specific protein–protein interactions [70]. In the present manuscript, we used ¹³C-detection NMR experiments to unambiguously access information about Xaa-Pro isomerization in OPN and to characterize the local and global effects of this conformational equilibrium.

The tunable stabilization of *cis* conformers by π ·CH interaction between the aromatic side chain and the proline ring has been described mainly for small molecules and peptides [55,56,71] and recently for the TAD domain of BMAL1 [72]. This stabilization is limited to prolines (at position *i*) and aromatics in positions *i*_{±1} and *i*-2. Our study provides, to our knowledge, the first experimental evidence that these weak interactions remain relevant in the context of a large IDP. We found that OPN shows particularly high *cis*-Pro populations, reaching *cis*-Pro percentages around 20% in some cases. Most importantly, the highest populations of *cis*-Pro conformers were found for prolines in close proximity to aromatic residues. The mutation of

an aromatic residue to alanine leads to a drastic reduction in *cis*-Pro population. We thus conclude that π ·CH interactions between aromatic side chains and proline rings are highly relevant for the stabilization of local structural segments in IDPs. Aromatic amino acids, leucines, and proline residues are well conserved in IDPs [5], but hydrophobic and proline residues have an opposite behavior with respect to their relative abundance when comparing ordered and disordered proteins [73]. Prolines are highly abundant in IDPs and are usually considered as disorder-promoting residues [5,74]. We thus performed a statistical analysis of the discussed motifs in globular proteins and IDPs. In our data sets (PDB for globular proteins (<https://www.rcsb.org/>) and DISPROT for disordered segments (<http://www.disprot.org/>)) (Fig. S4B–C), we found Pro residues to be more abundant in IDPs (7.3%) than in globular proteins (4.4%), whereas for aromatics the opposite was found (IDPs: 5.1%, globular proteins: 9.5%), in line with values previously reported [75]. The comparison between globular proteins vs IDPs with respect to the prevalence of aromatic proline pairs revealed that in IDPs 14% of prolines have aromatic residues in positions *i*_{±1} and *i*-2 and thus in positions where the π ·CH stacking interaction can take place, whereas, in globular proteins this number is 27%. Thus, our experimental findings that these sequence motifs are found in regions of significant structural compaction and that their mutations reduce, to a certain degree, this state are suggesting that these motifs might serve as nucleation sites for facilitating the formation of compact states in IDPs. In addition, it points out to the relevance of aromatic-proline interactions for the stabilization of underlying structure in IDPs. Although systematic studies on the biological significance of aromatic proline amino acid pairs in disordered segments are currently lacking, a first hint was provided with the Pro-Trp pair in BMAL1 [72] which is strongly conserved and key for the regulation of circadian rhythms. Another example is given by the unique domain of Src-family kinases which contains an unusual number of hydrophobic residues for an IDR and which are intercalated with prolines. Arbesú et al. [76] have shown that phenylalanine residues in this IDR promote the compaction of this protein, they hypothesize that proline residues may compensate the entropic cost of this effect by retaining correlated local flexibility, as proposed elsewhere [77]. Further evidence for the relevance of this sequence motif is provided by the fact that aromatic residues are found in prestructured molecular recognition motifs (also called MoRFs and PreSMos [78]). Our findings that π ·CH interactions significantly reduce the conformational dynamics of the minor *cis*-Pro form in W184 can thus be regarded as a prototypical case for a general mechanism to modulate backbone as well as side chain dynamics in IDPs. The compaction

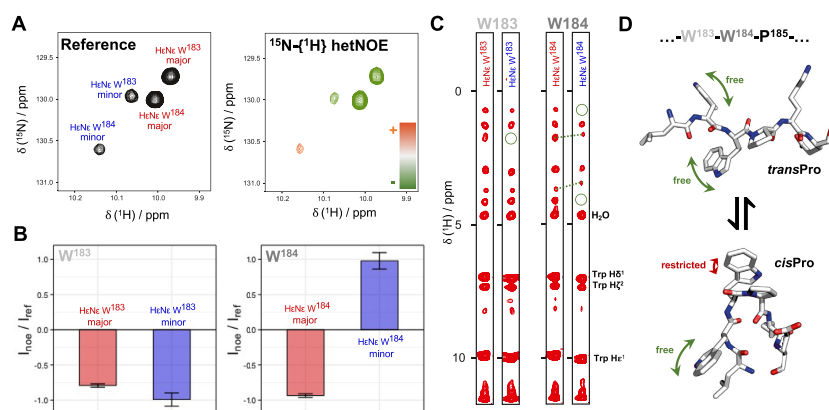
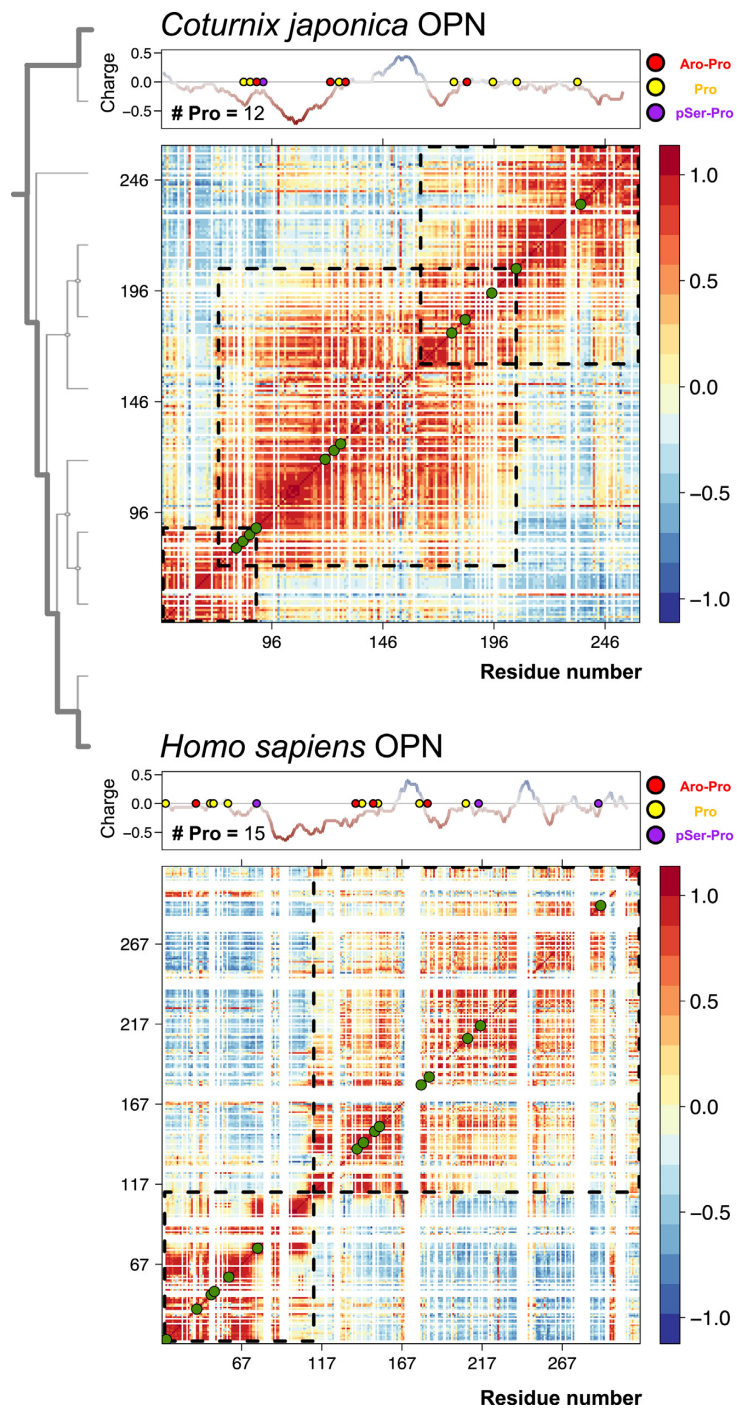


Fig. 8. Differential Trp side chain structural dynamics of minor (stemming from *cis*-Pro) and major (stemming from *trans*-Pro) forms. (A) $^{15}\text{N}\{-^1\text{H}\}$ NOE experiments (left: reference; right: ^1H -saturation), (B) $^{15}\text{N}\{-^1\text{H}\}$ intensity ratios ($I_{\text{sat}}/I_{\text{ref}}$) for minor and major forms of W183 and W184, (C) Amide $^1\text{H}^{\text{N}}$ to side-chain ^1H NOEs for minor and major forms of W183 and W184, obtained in a ^{15}N 3D NOESY-HSQC experiment. Differences in NOESY patterns are indicated by circles, and shifted resonances are marked with arrows. (D) Schematic 3D structural model of the sequence motif—WWP—in *cis*-Pro and *trans*-Pro forms. In the *cis*-Pro, distinct CH-aromatic interactions lead to restricted local mobility.



modulation provided by π -CH stacking interactions could reshape the conformational space sampled by OPN and in consequence, affects the binding events, such as the one with heparin. Previous work [48] showed an unfolding-upon-binding mechanism upon heparin interaction. This binding mechanism is possible if a certain degree of residual structure is retained in the apo-form to gain flexibility in other regions upon binding [79].

More in general, there is growing evidence that IDPs are far from being natively unfolded or disordered but rather exist as heterogeneous conformational ensembles. The distinct heterogeneities of these conformational ensembles and correlated conformational fluctuations within them can be quantitatively visualized by the Pearson correlation maps derived from PRE (and PRI) rates [80]. This is illustrated with two divergent OPNs (*Homo sapiens* vs. *C. japonica* PRE rates in Fig. S6A and C, respectively) (Fig. 9). *C. japonica* OPN contains three regions of correlated fluctuations indicating considerable structural compaction and the existence of—at least—one compact state in OPN (Fig. 9 upper). Most importantly, extremely high concentrations of urea and high-salt concentrations are needed to completely denature (unfold) the protein [48], thus indicating that both backbone structural elements and electrostatic interactions are involved in generating these compact states. Interestingly, human OPN showed a different pattern of correlated segments, where oppositely charged regions are less correlated than its evolutionary far homolog (Fig. 9 lower). To assess the potential influence of different charge distributions, we followed a strategy proposed by Pappu et al. [81], who have demonstrated the intimate relation between sequence distributions and ensemble conformations of polyampholytes with simulations and introduced the parameter κ , which takes into account not only the overall charge of the polymer but additionally also the distribution of these charges along the primary sequence. The κ value for *C. japonica* OPN (or quail OPN) is 0.262, corresponding to a typical value of a premolten globule in the interface between a weak and strong polyampholyte [79]. Interestingly, despite the low interspecies protein sequence similarity (Fig. S7), the parameter κ (Table S5) is largely similar for the individual OPN family members (Fig. S7). The observed differences in correlated fluctuations (Fig. 9) can therefore not be explained

by charge pattern but result from context-dependent interactions along OPN primary sequence. Regarding the number of prolines and their location in OPN (Fig. S8 colored circles), human OPN is significantly divergent from the quail form. In particular, the human N-terminal segment is enriched in prolines, and the quail form is depleted. On the other side, aro-Pro motifs are conserved next to the positive charged patch at the central region of the protein (Fig. S8 red circles), suggesting a relevant role of those in the central compact state.

Based on our findings, we thus conclude that in addition to the specific distribution of charges along the primary sequence (quantified by the parameter κ), the specific location of proline-aromatic sequence motifs (relative to electrostatic patches) plays a decisive role in determining the conformational space sampled by IDPs.

It is also instructive to compare our findings with a recent sequence analysis deciphering the evolutionary constraints acting on IDPs to disable a negative function (aggregation) [82]. In this study, it was shown that representative ensembles of folded proteins and IDPs respond differently to randomization (while maintaining the overall amino acid distribution) of primary sequences. Most importantly, this study unveiled that in contrast to the widespread belief that protein disorder triggers aggregation-induced diseases, IDP sequences evolved during evolution to exactly avoid amyloidogenesis and increase solubility. In other words, the finding that amyloids are sometimes associated to IDPs obscures the fact that the vast majority of IDPs emerged under the evolutionary constraint to avoid aggregation. Thus, the significance of this protein property (aggregation propensity) and its relevance for disease results from its scarcity in IDPs. Based on our finding that aro-Pro sequence motifs are significantly less abundant in IDPs, we postulate an analogous negative evolutionary constraint for IDP sequences. Because aro-Pro sequence motifs are depleted in IDPs, their occurrence points to their significance. Our study of OPN showed that the aro-Pro sequence motifs show significant local rigidification of the backbone chain and is thus in stark contrast to the conventional notion that proline residues are the origin of flexibility in IDPs. These structurally constrained sites can on the contrary act as nucleation sites for the formation of compact states in IDPs as their restricted motion reduces the entropic penalty for the formation of collapsed states. Most interestingly,

Fig. 9. Pearson correlation map calculated from the $^1\text{H}^{\text{N}}\text{-T}_2$ PRE rates of 10 cysteine mutants (*Coturnix japonica*) or 9 cysteine mutants (*Homo sapiens*). All measurements were carried out at 800 MHz and 298 K. Plots on top of the Pearson correlation map represent the charge and proline distributions. Red and purple dots correspond to context-dependent peculiar proline residues (i.e., aro-Pro and pSer-Pro sites, respectively). Green circles within the correlation map indicate the positions of proline residues along the protein sequence. Dashed squares represent regions of distinct structural compaction. Residues within one of these regions undergo correlated structural fluctuations. Coupling to other regions can be correlated (orange to red), anticorrelated (light blue to dark blue), or uncorrelated (light yellow to light blue).

the compact state(s) of OPN is (are) characterized by substantial long-range order [83]. Because there is typically a low sequence homology in related IDPs, we postulate that the conservation of (aro-Pro) sequence patterns indicates the existence of relevant and specific aro-Pro interactions and therefore suggests the existence of compact segments. It will be interesting to see whether this information can be used as a primary sequence search criterion to identify IDPs with distinct structural properties, as proposed elsewhere [84].

We therefore suggest reassessing the structural and functional roles of prolines in IDPs. It is generally accepted that prolines are highly abundant in IDPs, and on this basis are described as disorder promoting amino acids [73,74]. Based on our findings, we suggest a more subtle and ambivalent role for prolines in IDPs. Proline residues are often found in specific local secondary structural elements called PPII helices. The unique structural arrangement of residues in PPII helices does not allow for tight packing interactions, and residues in PPII helices have therefore a smaller number of direct neighbors and consequently lower structural stabilization (or disorder for that matter). However, the data obtained in our study clearly show that the structural dynamics of proline residues depends significantly on the amino acid sequence context (i.e., neighboring amino acids next to prolines in the primary sequence). Sequence patches including proline and aromatic amino acids display significant reduction of conformational flexibility, higher populations of *cis*-Pro configurations, and different π -CH stacking interactions between the proline and the aromatic ring systems.

To conclude, the observed novel proline features provide a versatile mechanism to modulate and tune the conformational ensembles of IDPs. For example, *cis/trans* isomerization will modulate the relative orientation of locally folded protein segments (i.e., regions displaying correlated structural fluctuations). The example of OPN provides a first glimpse on how the modulation of conformational ensembles in IDPs can be realized. Aro-Pro sequence motifs act as nucleation sites for the structural compaction reminiscent of diffusion-collision type of processes. The motional restriction (as evidenced by ^{15}N relaxation data) of these segments reduce the entropic penalty inherent to structure formation in IDPs. Moreover, our finding that these Pro-aromatic patches are located in hinge regions linking the three observed structural segments in OPN may point towards a general building principle. Because *cis-trans* isomerization is not only abundant in Pro-rich segments but can also be modulated by external factors (i.e., presence of prolyl-peptidyl isomerases, posttranslational modifications such as phosphorylation of neighboring Tyr or Ser/Thr, etc.), it can be anticipated that these modulation possibilities are realized by nature to endow IDPs with the stunning structural

plasticity and malleability necessary to adapt to the continuously changing molecular environment, cellular context, and thereby fulfilling their diverse biological functions.

Materials and Methods

Expression and purification of the *C. japonica* and *H. sapiens* OPN forms. *C. japonica* OPN sequence was cloned in pET11d plasmid. Several cysteine mutants and a phenylalanine-to-alanine mutation (F121A) were introduced by site-directed mutagenesis (Thermo Fisher®) and produced as the wild-type form. A total of 10 cysteine mutants of the *C. japonica* OPN (S54C, E77C, S108C, R132C, A161C, D174C, S188C, S206C, S228C, and S247C) were used. The plasmid was transformed into the *Escherichia coli* strain BL21(DE3) phage resistant via heat shock method. Protein expression was induced at OD₆₀₀ of ca. 0.8 by adding 0.4 mM IPTG (isopropyl- β -D-thiogalactopyranoside). For expression of isotopically labeled protein for NMR studies, before induction, the cells were harvested and pellets from 4 L LB resuspended in 1 L M9-Minimal Media (containing 1 g/L $^{15}\text{NH}_4\text{Cl}$ and 3 g/L ^{13}C -glucose for ^{15}N and ^{13}C labeling, respectively). The expression was carried out at 27 °C and 140 rpm overnight. The cells were harvested and resuspended in 40 mL of cold PBS (2 mM KH_2PO_4 , 8 mM Na_2HPO_4 , 2.5 mM KCl, 140 mM NaCl, 5 mM EDTA, pH 7.3); freshly prepared 1 mM DTT solution was added in the case of cysteine-mutated proteins preparations. The suspension was sonicated and warmed at 95 °C for 10 min. The lysate was centrifuged at 18000 rpm, and ammonium sulfate precipitation (saturation of 50%) carried out with the supernatant. The pellet was resuspended in PBS, diluted 1:2 with water to lower the salt concentration, and an anion-exchange chromatography (HiTrap Q, GE Healthcare) performed. The column was equilibrated with PBS and a gradient of 30% high-salt buffer (PBS containing 1 M NaCl) run in 20 min at 2 mL/min.

A pETM11 plasmid encoding for the *H. sapiens* OPN with a histidine tag was transformed in *E. coli* BL21(DE3) Rosetta phage resistant strain. The steps were the same as its homolog described above. Nine cysteine mutants of the *H. sapiens* OPN (G25C, Q50C, D90C, D130C, T185C, D210, S239C, S267C, and S311C) were produced. Regarding the purification, the harvested cell suspension was sonicated and cooked at 95 °C for 10 min. The lysate was centrifuged at 18,000 rpm, and the supernatant loaded onto a HisTrap® (GE Healthcare) affinity column. The protein was eluted with high-imidazole buffer (140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4 , 1.8 mM KH_2PO_4 , 500 mM imidazole, pH 8). The protein was concentrated in TEV cleavage buffer (140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4 , 1.8 mM KH_2PO_4 , 1 mM DTT, 1 mM EDTA, pH 8.0), TEV protease was added in a ratio of 1:50 and incubated overnight at 4 °C under agitation to get rid of the affinity tag. Further purification was carried out with ion-exchange chromatography (HiTrapQ®, GE Healthcare). The column was equilibrated with PBS and a gradient of 50% high-salt buffer (PBS containing 1 M NaCl) run in 30 min at 1 mL/min. Protein purity was confirmed via SDS-PAGE.

The protein was concentrated using a centrifugal filter and the final concentration measured by absorbance at

280 nm. The sample concentration for the assignment by ^{13}C -detection and ^{15}N relaxation rates was 2 mM. ^{13}C -detected PRE experiments were measured with protein samples about 1 mM, whereas ^1H -detected PRE experiments were measured with protein samples about 0.5 mM (*C. japonica*) and 0.3 mM (*H. sapiens*) in PBS buffer at pH 6.5.

MTSL-tagged cysteine mutants for PRE. For the MTSL-tagging, the protein was incubated with an excess of DTT (10 mM) for 15 min at room temperature. A PD-10 Desalting-Column was equilibrated with 100 mM sodium phosphate, 1 mM EDTA, and pH 8.0 buffer, the protein loaded to a total volume of 2.5 mL and eluted with 3.4 mL of buffer. The free thiol concentration was measured using a 300 μM DTNB (5,5'-dithiobis-(2-nitrobenzoic acid)) solution at 412 nm. A threefold excess of MTSL was added and the sample was incubated for 3 h at 37 °C during agitation. The free thiol concentration was measured again to confirm complete tagging of the protein. Protein purity was confirmed via SDS-PAGE electrophoresis.

^{13}C -detection assignment and ^1H -start CON PRE. NMR experiments were recorded at 310 K with a 16.4 T Bruker Avance NEO spectrometer operating at 700 MHz ^1H frequency and equipped with a cryogenically cooled probehead optimized for ^{13}C direct detection. It has to be noted that the presence of D_2O (lock solvent) in the sample gives rise to cross peaks resulting from isotopic shifts in the CON spectra due to H/D exchange. Although these signals are easy to identify because of their regular pattern of displacement, they can lead to unwanted signal overlap and complicate spectral analysis. Therefore, the lowest possible concentration of D_2O (typically 1–2%) has been used during the measurements to reduce the intensity of this set of additional cross-peaks below detection level.

In the ^{13}C -detected experiments for sequence-specific assignment of N, C', C $^\alpha$, C $^\beta$ resonances at 310 K (2D CON, 3D (H)CBCACON, 3D (H)CBCANCO, 3D (HCA) COCON, and 3D (H)CCCON)) [34,38,85], the ^{13}C carrier was placed at 174.5 ppm, the ^{15}N carrier at 123 ppm, and the ^1H carrier at 4.7 ppm (H_2O). Band-selective ^{13}C pulses were applied at 174.5 and 39 ppm to excite or invert $^{13}\text{C}'$ and $^{13}\text{C}^{\text{all}}$ spins, respectively. The following band-selective pulses were used: 300 μs with Q5 and time reversed Q5 shapes for C' and C $^{\text{all}}$ excitation, 220 μs Q3 shapes for C'/C $^{\text{all}}$ inversion/refocusing. Homonuclear virtual decoupling in the direct dimension was achieved through the IPAP approach. Detailed information about the acquisition parameters is provided (Tables S1 and S2). NMR data were processed using Bruker TopSpin 4.0. The programs CARA [86] and Sparky [87] were used to analyze and annotate the spectra. All the statistical analysis and fittings were carried out with RStudio.

^1H -T $_2$ PRE acquisition parameters. NMR experiments were recorded in on an 18.8 T Bruker Avance III HD+ spectrometer operating at 800 MHz. For the ^1H -T $_2$ PRE measurements, the measurements were carried out at 298 K, and paramagnetic samples were reduced using a threefold excess of ascorbic acid for the diamagnetic control. The ^1H -T $_2$ relaxation delays were 0.001, 0.005, 0.01, 0.02, 0.05, and 0.1 s. The paramagnetic effects were quantified as

$$\Delta\Gamma_2 = \text{H}^{\text{N}} - \text{R}_{2,\text{para}} - \text{H}^{\text{N}} - \text{R}_{2,\text{dia}} \quad (1)$$

For ^{15}N -T $_1$, ^{15}N -T $_2$, CLEANEX, and ^{15}N - $\{^1\text{H}\}$ heteronuclear NOE, the measurements were carried out at 310 K, and 10% D_2O was added as the lock solvent. See Table S3 for further details. Analysis and plotting was carried out with RStudio. For the difference plots showed in Fig. 4, Fig. 7, Figs. S5 and S6, the error associated (δQ) was calculated by propagating the fitting errors of the experimental data (δa and δb) as:

$$\delta\text{Q} = (\delta\text{a}^2 + \delta\text{b}^2)^{1/2} \quad (2)$$

^{15}N -relaxation acquisition parameters. NMR experiments were recorded in on an 18.8 T Bruker Avance III HD+ spectrometer operating at 800 MHz. The ^{15}N -T $_1$ relaxation delays were 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, and 1.28 s. The ^{15}N -T $_2$ relaxation delays were 0, 0.017, 0.034, 0.068, 0.136, 0.271, 0.407, and 0.543 s. ^{15}N - $\{^1\text{H}\}$ heteronuclear NOE was measured with an interscan delay of 10 s to avoid systematic errors in the measurement of steady-state NOE as a consequence of incomplete equilibration of the magnetization in the reference spectrum [88].

For the 2D ($\text{H}^{2\text{C}^\alpha}$)CON $^{\text{Pro}}$ experiments to determine the ^{15}N R $_2$, the delay used were 0.080, 0.160, 0.240, 0.320, 0.400, 0.480, 0.560, 0.640, and 0.800 s. The RF field strength used for the CPMG block was 3.1 kHz.

Trp side chain dynamics. We measured relaxation rates of the $^1\text{H}_{\epsilon 1}$ - $^{15}\text{N}_{\epsilon 1}$ major and minor forms (Fig. S9D). Minor forms were also observed for $^1\text{H}_{\epsilon 3}$ - $^{13}\text{C}_{\epsilon 3}$ and $^1\text{H}_{\eta}$ - $^{13}\text{C}_{\eta}$ pairs selectively labeled (Fig. S9E).

Acknowledgments

The support and the use of resources of the CERM/CIRMMP center of Instruct-ERIC is gratefully acknowledged. This work has been supported in part by iNEXT, Grant number 653706, funded by the Horizon 2020 program of the European Commission, PID: 3972 and 2676. Grants of the Fondazione CR Firenze/AIRC to RP and by the Austrian Federal Ministry of Science, Research and Economy and the National Foundation for Research, Technology and Development to RK are gratefully acknowledged. BM is grateful to the NGP-net COST action (ID: 38248) for funding a short-term mission to the Magnetic Resonance Center of the University of Florence. BM also acknowledges the DK Integrative Structural Biology funded by the Austrian Science Fund (FWF, project number AW0125821). AB is funded by the Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Austria.

Author Contributions

BM, RK, ICF, and RP conceived the project; BM, RK, ICF, and RP designed the experiments; BM and CCB produced samples; BM, MS, and CCB

performed NMR experiments with RP and ICF; BM and MS processed and analyzed the data together with RK, ICF, and RP; AB and GK contributed with new analytic tools; BM, RK, ICF, and RP wrote the manuscript, with contributions from all the other authors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jmb.2019.11.015>.

Received 20 August 2019;
Received in revised form 23 October 2019;
Accepted 14 November 2019
Available online 30 November 2019

Keywords:

Proline;
Intrinsically disordered proteins;
Osteopontin;
¹³C-detection;
NMR spectroscopy

References

- [1] A.K. Dunker, C.J. Brown, J.D. Lawson, L.M. Iakoucheva, Z. Obradović, Intrinsic disorder and protein function, *Biochemistry* 41 (2002) 6573–6582.
- [2] H.J. Dyson, P.E. Wright, Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol.* 6 (2005) 197–208, <https://doi.org/10.1038/nm1589>.
- [3] J. Habchi, P. Tompa, S. Longhi, V.N. Uversky, Introducing protein intrinsic disorder, *Chem. Rev.* 114 (2014) 6561–6588, <https://doi.org/10.1021/cr400514h>.
- [4] H. Hofmann, A. Soranno, A. Borgia, K. Gast, D. Nettels, B. Schuler, Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy, *Proc. Natl. Acad. Sci. U. S. A.* 109 (2012) 16155–16160, <https://doi.org/10.1073/pnas.1207719109>.
- [5] F. Theillet, L. Kalmar, P. Tompa, K. Han, P. Selenko, A.K. Dunker, G.W. Daughdrill, V.N. Uversky, The alphabet of intrinsic disorder, *Intrinsically Disord. Proteins* 1 (2013), e24360, <https://doi.org/10.4161/idp.24360>.
- [6] C.M. Deber, B. Brodsky, A. Rath, Proline residues in proteins, *Encycl. Life Sci.* (2001) 1–6, <https://doi.org/10.1002/9780470015902.a0003014.pub2>.
- [7] C. Dugave, L. Demange, Cis-trans isomerization of organic molecules and biomolecules: implications and applications, *Chem. Rev.* 103 (2003) 2475–2532, <https://doi.org/10.1021/cr0104375>.
- [8] U. Reimer, G. Fischer, Local structural changes caused by peptidyl-prolyl cis/trans isomerization in the native state of proteins, *Biophys. Chem.* 96 (2002) 203–212, [https://doi.org/10.1016/S0301-4622\(02\)00013-3](https://doi.org/10.1016/S0301-4622(02)00013-3).
- [9] S. Osváth, M. Gruebele, Proline can have opposite effects on fast and slow protein folding phases, *Biophys. J.* 85 (2003) 1215–1222, [https://doi.org/10.1016/S0006-3495\(03\)74557-3](https://doi.org/10.1016/S0006-3495(03)74557-3).
- [10] R.B. Perez, A. Tischer, M. Auton, S.T. Whitten, Alanine and proline content modulate global sensitivity to discrete perturbations in disordered proteins, *Proteins* (2014) 3373–3384, <https://doi.org/10.1002/prot.24692>.
- [11] T.R. Alderson, J.H. Lee, C. Charlier, J. Ying, A. Bax, Propensity for cis-Proline formation in unfolded proteins, *ChemBioChem* 19 (2018) 37–42, <https://doi.org/10.1002/cbic.201700548>.
- [12] P. Ahuja, F.X. Cantrelle, I. Huvent, X. Hanouille, J. Lopez, C. Smet, J.M. Wieruszkeski, I. Landrieu, G. Lippens, Proline conformation in a functional Tau fragment, *J. Mol. Biol.* 428 (2016) 79–91, <https://doi.org/10.1016/j.jmb.2015.11.023>.
- [13] R. Glaves, M. Baer, E. Schreiner, R. Stoll, D. Marx, Conformational dynamics of minimal elastin-like polypeptides: the role of proline revealed by molecular dynamics and nuclear magnetic resonance, *ChemPhysChem* 9 (2008) 2759–2765, <https://doi.org/10.1002/cphc.200800474>.
- [14] G. Fernandez-Ballester, C. Blanes-Mira, L. Serrano, The Tryptophan switch: changing ligand-binding specificity from type I to type II in SH3 domains, *J. Mol. Biol.* 335 (2004) 619–629, <https://doi.org/10.1016/j.jmb.2003.10.060>.
- [15] M.W. MacArthur, J.M. Thornton, Influence of proline residues on protein conformation, *J. Mol. Biol.* 218 (1991) 397–412, [https://doi.org/10.1016/0022-2836\(91\)90721-H](https://doi.org/10.1016/0022-2836(91)90721-H).
- [16] E.W. Martin, A.S. Holehouse, C.R. Grace, A. Hughes, R.V. Pappu, T. Mittag, Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation, *J. Am. Chem. Soc.* 138 (2016) 15323–15335, <https://doi.org/10.1021/jacs.6b10272>.
- [17] E.B. Gibbs, F. Lu, B. Portz, M.J. Fisher, B.P. Medellin, T.N. Laremore, Y.J. Zhang, D.S. Gilmour, S.A. Showalter, Phosphorylation induces sequence-specific conformational switches in the RNA polymerase II C-terminal domain, *Nat. Commun.* 8 (2017) 1–11, <https://doi.org/10.1038/ncomms15233>.
- [18] R.V. Pappu, G.D. Rose, A simple model for polyproline II structure in unfolded states of alanine-based peptides, *Protein Sci.* 11 (2002) 2437–2455, <https://doi.org/10.1110/Ps.0217402>.
- [19] J.A. Vila, H.A. Baldoni, D.R. Ripoll, A. Ghosh, H.A. Scheraga, Polyproline II helix conformation in a proline-rich environment: a theoretical study, *Biophys. J.* 86 (2004) 731–742, [https://doi.org/10.1016/S0006-3495\(04\)74151-X](https://doi.org/10.1016/S0006-3495(04)74151-X).
- [20] W.A. Elam, T.P. Schrank, A.J. Campagnolo, V.J. Hilser, Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites, *Protein Sci.* 22 (2013) 405–417, <https://doi.org/10.1002/pro.2217>.
- [21] M.D. Crabtree, W. Borchers, A. Poosapati, S.L. Shammass, G.W. Daughdrill, J. Clarke, Conserved helix-flanking prolines modulate IDP: target affinity by altering the lifetime of the bound complex, *Biochemistry* (2017), <https://doi.org/10.1021/acs.biochem.7b00179> acs.biochem.7b00179.
- [22] L.E. Wong, J. Maier, J. Wienands, S. Becker, C. Griesinger, Sensitivity-enhanced four-dimensional amide–amide correlation NMR experiments for sequential assignment of proline-rich disordered proteins, *J. Am. Chem. Soc.* 140 (2018) 3518–3522, <https://doi.org/10.1021/jacs.8b00215>.
- [23] Y. Yoshimura, N.V. Kulminskaya, F.A.A. Mulder, Easy and unambiguous sequential assignments of intrinsically disordered proteins by correlating the backbone ¹⁵N or ¹³C chemical shifts of multiple contiguous residues in highly

- resolved 3D spectra, *J. Biomol. NMR* 61 (2015) 109–121, <https://doi.org/10.1007/s10858-014-9890-7>.
- [24] V. Kanelis, L. Donaldson, D.R. Muhandiram, D. Rotin, J.D. Forman-Kay, L.E. Kay, Sequential assignment of proline-rich regions in proteins: application to modular binding domain complexes, *J. Biomol. NMR* 16 (2000) 253–259, <https://doi.org/10.1023/A:1008355012528>.
- [25] M.G. Murrall, A. Piai, W. Bermel, I.C. Felli, R. Pierattelli, Proline fingerprint in intrinsically disordered proteins, *ChemBioChem* 19 (2018) 1625–1629, <https://doi.org/10.1002/cbic.201800172>.
- [26] D.E. Dorman, F.A. Bovey, Carbon-13 magnetic resonance spectroscopy. Spectrum of proline in oligopeptides, *J. Org. Chem.* 38 (1973) 2379–2383, <https://doi.org/10.1021/jo00953a021>.
- [27] C. Grathwohl, K. Wüthrich, The X-Pro peptide bond as an NMR probe for conformational studies of flexible linear peptides, *Biopolymers* 15 (1976) 2025–2041, <https://doi.org/10.1002/bip.1976.360151012>.
- [28] H. Kovacs, D. Moskau, M. Spraul, Cryogenically cooled probes – a leap in NMR technology, *Prog. Nucl. Magn. Reson. Spectrosc.* 46 (2005) 131–155, <https://doi.org/10.1016/j.pnmrs.2005.03.001>.
- [29] W. Bermel, I. Bertini, I.C. Felli, M. Piccioli, R. Pierattelli, ¹³C-detected protonless NMR spectroscopy of proteins in solution, *Prog. Nucl. Magn. Reson. Spectrosc.* 48 (2006) 25–45, <https://doi.org/10.1016/j.pnmrs.2005.09.002>.
- [30] K. Takeuchi, G. Heffron, Z.Y.J. Sun, D.P. Frueh, G. Wagner, Nitrogen-detected CAN and CON experiments as alternative experiments for main chain NMR resonance assignments, *J. Biomol. NMR* 47 (2010) 271–282, <https://doi.org/10.1007/s10858-010-9430-z>.
- [31] S. Chhabra, P. Fischer, K. Takeuchi, A. Dubey, J.J. Ziarek, A. Boeszoermenyi, D. Mathieu, W. Bermel, N.E. Davey, G. Wagner, H. Arthanari, ¹⁵N detection harnesses the slow relaxation property of nitrogen: delivering enhanced resolution for intrinsically disordered proteins, *Proc. Natl. Acad. Sci. U. S. A.* 115 (2018) E1710–E1719, <https://doi.org/10.1073/pnas.1717560115>.
- [32] R.B. Pritchard, D.F. Hansen, Characterising side chains in large proteins by protonless ¹³C-detected NMR spectroscopy, *Nat. Commun.* 10 (2019) 1747, <https://doi.org/10.1038/s41467-019-09743-4>.
- [33] A. Escobedo, B. Topal, M.B.A. Kunze, J. Aranda, G. Chiesa, D. Mungianu, G. Bernardo-Seisdedos, B. Eftekharzadeh, M. Gairí, R. Pierattelli, I.C. Felli, T. Diercks, O. Millet, J. García, M. Orozco, R. Crehuet, K. Lindorff-Larsen, X. Salvatella, Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor, *Nat. Commun.* 10 (2019) 2034, <https://doi.org/10.1038/s41467-019-09923-2>.
- [34] W. Bermel, I. Bertini, I.C. Felli, R. Kümmerle, R. Pierattelli, Novel ¹³C direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins, *J. Magn. Reson.* 178 (2006) 56–64, <https://doi.org/10.1016/j.jmr.2005.08.011>.
- [35] W. Bermel, I. Bertini, I.C. Felli, R. Peruzzini, R. Pierattelli, Exclusively heteronuclear NMR experiments to obtain structural and dynamic information on proteins, *ChemPhysChem* 11 (2010) 689–695, <https://doi.org/10.1002/cphc.200900772>.
- [36] S. Gil, T. Hosek, Z. Solyom, R. Kümmerle, B. Brutscher, R. Pierattelli, I.C. Felli, NMR spectroscopic studies of intrinsically disordered proteins at near-physiological conditions, *Angew. Chem. Int. Ed.* 52 (2013) 11808–11812, <https://doi.org/10.1002/anie.201304272>.
- [37] E.B. Gibbs, R.W. Kriwacki, Direct detection of carbon and nitrogen nuclei for high-resolution analysis of intrinsically disordered proteins using NMR spectroscopy, *Methods* 138–139 (2018) 39–46, <https://doi.org/10.1016/j.jymeth.2018.01.004>.
- [38] W. Bermel, I. Bertini, I.C. Felli, Y.-M. Lee, C. Luchinat, R. Pierattelli, Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins, *J. Am. Chem. Soc.* 128 (2006) 3918–3919, <https://doi.org/10.1021/ja0582206>.
- [39] I.C. Felli, R. Pierattelli, S.J. Glaser, B. Luy, Relaxation-optimised Hartmann–Hahn transfer using a specifically Tailored MOCCA-XY16 mixing sequence for carbonyl–carbonyl correlation spectroscopy in ¹³C direct detection NMR experiments, *J. Biomol. NMR* 43 (2009) 187–196, <https://doi.org/10.1007/s10858-009-9302-6>.
- [40] S. Balayssac, B. Jiménez, M. Piccioli, ¹³C direct detected COCO-TOCSY: a tool for sequence specific assignment and structure determination in protonless NMR experiments, *J. Magn. Reson.* 182 (2006) 325–329, <https://doi.org/10.1016/j.jmr.2006.06.021>.
- [41] W. Bermel, I. Bertini, J. Chill, I.C. Felli, N. Haba, M.V.V. Kumar, R. Pierattelli, Exclusively heteronuclear ¹³C-detected amino-acid-selective NMR experiments for the study of intrinsically disordered proteins (IDPs), *ChemBioChem* 13 (2012) 2425–2432, <https://doi.org/10.1002/cbic.201200447>.
- [42] I.C. Felli, L. Gonnelli, R. Pierattelli, In-cell ¹³C NMR spectroscopy for the study of intrinsically disordered proteins, *Nat. Protoc.* 9 (2014) 2005–2016, <https://doi.org/10.1038/nprot.2014.124>.
- [43] A. Pietras, A.M. Katz, E.J. Ekström, B. Wee, J.J. Halliday, K.L. Pitter, J.L. Werbeck, N.M. Amankulor, J.T. Huse, E.C. Holland, Osteopontin-CD44 signaling in the glioma perivascular niche enhances cancer stem cell phenotypes and promotes aggressive tumor growth, *Cell Stem Cell* 14 (2014) 357–369, <https://doi.org/10.1016/j.stem.2014.01.005>.
- [44] L.A. Shevde, R.S. Samant, Role of osteopontin in the pathophysiology of cancer, *Matrix Biol.* 37 (2014) 131–141, <https://doi.org/10.1016/j.matbio.2014.03.001>.
- [45] C.C. Kazanek, D.J. Uzwiak, D.T. Denhardt, Control of osteopontin signaling and function by post-translational phosphorylation and protein folding, *J. Cell. Biochem.* 102 (2007) 912–924, <https://doi.org/10.1002/jcb.21558>.
- [46] P.H. Anborgh, J.C. Mutrie, A.B. Tuck, A.F. Chambers, Role of the metastasis-promoting protein osteopontin in the tumour microenvironment, *J. Cell Mol. Med.* 14 (2010) 2037–2044, <https://doi.org/10.1111/j.1582-4934.2010.01115.x>.
- [47] Y.U. Katagiri, J. Sleeman, H. Fujii, P. Herrlich, H. Hotta, K. Tanaka, S. Chikuma, H. Yagita, K. Okumura, M. Murakami, I. Saiki, A.F. Chambers, T. Uede, CD44 variants but not CD44s cooperate with beta1-containing integrins to permit cells to bind to osteopontin independently of arginine-glycine-aspartic acid, thereby stimulating cell motility and chemotaxis, *Cancer Res.* 59 (1999) 219–226.
- [48] D. Kurzbach, T.C. Schwarz, G. Platzer, S. Höfler, D. Hinderberger, R. Konrat, Compensatory adaptations of structural dynamics in an intrinsically disordered protein complex, *Angew. Chem. Int. Ed.* 53 (2014) 3840–3843, <https://doi.org/10.1002/anie.201308389>.

- [49] W. Bermel, I.C. Felli, L. Gonnelli, W. Kołmiński, A. Piai, R. Pierattelli, A. Zawadzka-Kazimierczuk, High-dimensionality ^{13}C direct-detected NMR experiments for the automatic assignment of intrinsically disordered proteins, *J. Biomol. NMR* 57 (2013) 353–361, <https://doi.org/10.1007/s10858-013-9793-z>.
- [50] J.-S. Hu, A. Bax, Measurement of three-bond ^{13}C – ^{13}C J-couplings between carbonyl and carbonyl/carboxyl carbons in isotopically enriched proteins, *J. Am. Chem. Soc.* 118 (1996) 8170–8171, <https://doi.org/10.1021/ja9616239>.
- [51] F. Li, J.H. Lee, A. Grishaev, J. Ying, A. Bax, High accuracy of Karplus equations for relating three-bond J couplings to protein backbone torsion angles, *ChemPhysChem* 16 (2015) 572–578, <https://doi.org/10.1002/cphc.201402704>.
- [52] M. Schubert, D. Labudde, H. Oschkinat, P. Schmieder, A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on ^{13}C chemical shift statistics, *J. Biomol. NMR* 24 (2002) 149–154, <https://doi.org/10.1023/A:1020997118364>.
- [53] Y. Shen, A. Bax, Prediction of Xaa-Pro peptide bond conformation from sequence and chemical shifts, *J. Biomol. NMR* 46 (2010) 199–204, <https://doi.org/10.1007/s10858-009-9395-y>.
- [54] H.J. Dyson, M. Rance, R.A. Houghten, R.A. Lerner, P.E. Wright, Folding of immunogenic peptide fragments of proteins in water solution, *J. Mol. Biol.* 201 (1988) 161–200, [https://doi.org/10.1016/0022-2836\(88\)90446-9](https://doi.org/10.1016/0022-2836(88)90446-9).
- [55] K.M. Thomas, D. Naduthambi, N.J. Zondlo, Electronic control of amide cis-trans isomerism via the aromatic-prolyl interaction, *J. Am. Chem. Soc.* 128 (2006) 2216–2217, <https://doi.org/10.1021/ja057901y>.
- [56] F. Nardi, J. Kemmink, M. Sattler, R.C. Wade, The cisproline(i-1)-aromatic(i) interaction: folding of the Ala-cisPro-Tyr peptide characterized by NMR and theoretical approaches, *J. Biomol. NMR* 17 (2000) 63–77.
- [57] A.G. Palmer, NMR characterization of the dynamics of biomacromolecules, *Chem. Rev.* 104 (2004) 3623–3640, <https://doi.org/10.1021/cr030413t>.
- [58] G. Parigi, N. Rezaei-Ghaleh, A. Giachetti, S. Becker, C. Fernandez, M. Blackledge, C. Griesinger, M. Zweckstetter, C. Luchinat, Long-range correlated dynamics in intrinsically disordered proteins, *J. Am. Chem. Soc.* 136 (2014) 16201–16209, <https://doi.org/10.1021/ja506820r>.
- [59] A. Abyzov, N. Salvi, R. Schneider, D. Maurin, R.W.H. Ruigrok, M.R. Jensen, M. Blackledge, Identification of dynamic modes in an intrinsically disordered protein using temperature-dependent NMR Relaxation, *J. Am. Chem. Soc.* (2016), <https://doi.org/10.1021/jacs.6b02424>.
- [60] S.N. Khan, C. Charlier, R. Augustyniak, N. Salvi, V. Déjean, G. Bodenhausen, O. Lequin, P. Pelupessy, F. Ferrage, Distribution of pico- and nanosecond motions in disordered proteins from nuclear spin relaxation, *Biophys. J.* 109 (2015) 988–999, <https://doi.org/10.1016/j.bpj.2015.06.069>.
- [61] N. Salvi, A. Abyzov, M. Blackledge, Solvent-dependent segmental dynamics in intrinsically disordered proteins, *Sci. Adv.* 5 (2019), <https://doi.org/10.1126/sciadv.aax2348>.
- [62] N. Rezaei-Ghaleh, G. Parigi, A. Soranno, A. Holla, S. Becker, B. Schuler, C. Luchinat, M. Zweckstetter, Local and global dynamics in intrinsically disordered synuclein, *Angew. Chem. Int. Ed.* 57 (2018) 15262–15266, <https://doi.org/10.1002/anie.201808172>.
- [63] G. Wagner, NMR relaxation and protein mobility, *Curr. Opin. Struct. Biol.* 3 (1993) 748–754, [https://doi.org/10.1016/0959-440X\(93\)90059-T](https://doi.org/10.1016/0959-440X(93)90059-T).
- [64] J. Wimmer, C. Schlörb, J. Klein-Seetharaman, R. Hirano, T. Ueda, T. Imoto, H. Schwalbe, Modulation of compactness and long-range interactions of unfolded lysozyme by single point mutations, *Angew. Chem. Int. Ed.* 43 (2004) 5780–5785, <https://doi.org/10.1002/anie.200460907>.
- [65] M. Xie, L. Yu, L. Bruschweiler-Li, X. Xiang, A.L. Hansen, R. Brüschweiler, Functional protein dynamics on uncharted time scales detected by nanoparticle-assisted NMR spin relaxation, *Sci. Adv.* 5 (2019) eaax5560, <https://doi.org/10.1126/sciadv.aax5560>.
- [66] R. Brueschweiler, P.E. Wright, NMR order parameters of biomolecules: a new analytical representation and application to the Gaussian Axial fluctuation model, *J. Am. Chem. Soc.* 116 (1994) 8426–8427, <https://doi.org/10.1021/ja00097a084>.
- [67] L. Shi, A.E. Holliday, H. Shi, F. Zhu, M.A. Ewing, D.H. Russell, D.E. Clemmer, Characterizing intermediates along the transition from polyproline I to polyproline II using ion mobility spectrometry-mass spectrometry, *J. Am. Chem. Soc.* 136 (2014) 12702–12711, <https://doi.org/10.1021/ja505899g>.
- [68] C. Camilloni, A. De Simone, W.F. Vranken, M. Vendruscolo, Determination of secondary structure populations in disordered states of proteins using Nuclear Magnetic Resonance chemical shifts, *Biochemistry* 51 (2012) 2224–2231, <https://doi.org/10.1021/bi3001825>.
- [69] B. Mateos, R. Konrat, R. Pierattelli, I.C. Felli, NMR characterization of long-range contacts in intrinsically disordered proteins from paramagnetic relaxation enhancement in ^{13}C direct-detection experiments, *ChemBioChem* 20 (2019) 335–339, <https://doi.org/10.1002/cbic.201800539>.
- [70] T.R. Alderson, J.L.P. Benesch, A.J. Baldwin, Proline isomerization in the C-terminal region of HSP27, *Cell Stress Chaperones* 22 (2017) 639–651, <https://doi.org/10.1007/s12192-017-0791-z>.
- [71] N.J. Zondlo, Aromatic–proline interactions: electronically tunable CH/ π interactions, *Acc. Chem. Res.* 46 (2013) 1039–1049, <https://doi.org/10.1021/ar300087y>.
- [72] C.L. Gustafson, N.C. Parsley, H. Asimgil, H.W. Lee, C. Ahlback, A.K. Michael, H. Xu, O.L. Williams, T.L. Davis, A.C. Liu, C.L. Patch, A slow conformational switch in the BMAL1 transactivation domain modulates circadian rhythms, *Mol. Cell* 66 (2017) 447–457, <https://doi.org/10.1016/j.molcel.2017.04.011>, e7.
- [73] A. Campen, R. Williams, C. Brown, J. Meng, V. Uversky, A. Dunker, TOP-IDP scale: a new amino acid scale measuring propensity for intrinsic disorder, *Protein Pept. Lett.* 15 (2008) 956–963, <https://doi.org/10.2174/092986608785849164>.
- [74] J.A. Marsh, J.D. Forman-Kay, Sequence determinants of compaction in intrinsically disordered proteins, *Biophys. J.* 98 (2010) 2383–2390, <https://doi.org/10.1016/j.bpj.2010.02.006>.
- [75] R.M. Williams, Z. Obradovi, V. Mathura, W. Braun, E.C. Garner, J. Young, S. Takayama, C.J. Brown, A.K. Dunker, The protein non-folding problem: amino acid determinants of intrinsic order and disorder, *Pac. Symp. Biocomput.* 100 (2001) 89–100.
- [76] M. Arbesú, M. Maffei, T.N. Cordeiro, J.M.C. Teixeira, Y. Pérez, P. Bernadó, S. Roche, M. Pons, The unique

- domain forms a fuzzy intramolecular complex in Src family kinases, *Structure* 25 (2017) 630–640, <https://doi.org/10.1016/j.str.2017.02.011>, e4.
- [77] M.C. Baxa, E.J. Haddadian, J.M. Jumper, K.F. Freed, T.R. Sosnick, Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations, *Proc. Natl. Acad. Sci. U. S. A.* 111 (2014) 15396–15401, <https://doi.org/10.1073/pnas.1407768111>.
- [78] C.J. Oldfield, Y. Cheng, M.S. Cortese, P. Romero, V.N. Uversky, A.K. Dunker, Coupled folding and binding with α -helix-forming molecular recognition elements, *Biochemistry* 44 (2005) 12454–12470, <https://doi.org/10.1021/bi050736e>.
- [79] D. Kurzbach, G. Platzer, T.C. Schwarz, M.A. Hemen, R. Konrat, D. Hinderberger, Cooperative unfolding of compact conformations of the intrinsically disordered protein osteopontin, *Biochemistry* 52 (2013) 5167–5175, <https://doi.org/10.1021/bi400502c>.
- [80] D. Kurzbach, A. Vanas, A.G. Flamm, N. Tarnoczi, G. Kontaxis, N. Maltar-Strmečki, K. Widder, D. Hinderberger, R. Konrat, Detection of correlated conformational fluctuations in intrinsically disordered proteins through paramagnetic relaxation interference, *Phys. Chem. Chem. Phys.* 18 (2016) 5753–5758, <https://doi.org/10.1039/C5CP04858C>.
- [81] R.K. Das, R.V. Pappu, Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues, *Proc. Natl. Acad. Sci. U. S. A.* 110 (2013) 13392–13397, <https://doi.org/10.1073/pnas.1304749110>.
- [82] Y. Naranjo, M. Pons, R. Konrat, Meta-structure correlation in protein space unveils different selection rules for folded and intrinsically disordered proteins, *Mol. Biosyst.* 8 (2012) 411–416, <https://doi.org/10.1039/c1mb05367a>.
- [83] D. Kurzbach, A. Beier, A. Vanas, A.G. Flamm, G. Platzer, T.C. Schwarz, R. Konrat, NMR probing and visualization of correlated structural fluctuations in intrinsically disordered proteins, *Phys. Chem. Chem. Phys.* 19 (2017) 10651–10656, <https://doi.org/10.1039/C7CP00430C>.
- [84] I. Pritisanac, R.M. Vernon, A.M. Moses, J.D. Forman Kay, Entropy and information within intrinsically disordered protein regions, *Entropy* 21 (2019) 662, <https://doi.org/10.3390/e21070662>.
- [85] W. Bermel, I. Bertini, V. Csizmok, I.C. Felli, R. Pierattelli, P. Tompa, H-start for exclusively heteronuclear NMR spectroscopy: the case of intrinsically disordered proteins, *J. Magn. Reson.* 198 (2009) 275–281, <https://doi.org/10.1016/j.jmr.2009.02.012>.
- [86] R.L.J. Keller, Optimizing the process of nuclear magnetic resonance spectrum analysis and computer aided resonance assignment (2005). <https://doi.org/10.3929/ETHZ-A-005068942>.
- [87] W. Lee, M. Tonelli, J.L. Markley, NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy, *Bioinformatics* 31 (2015) 1325–1327, <https://doi.org/10.1093/bioinformatics/btu830>.
- [88] F. Ferrage, A. Piserchio, D. Cowburn, R. Ghose, On the measurement of ^{15}N - ^1H nuclear Overhauser effects, *J. Magn. Reson.* 192 (2008) 302–313, <https://doi.org/10.1016/j.jmr.2008.03.011>.

Supplemental Information

The ambivalent role of proline residues in an intrinsically disordered protein: from disorder promoters to compaction facilitators

Borja Mateos, Clara Conrad-Billroth, Marco Schiavina, Andreas Beier, Georg Kontaxis, Robert Konrat, Isabella C. Felli and Roberta Pierattelli

Table S1. Acquisition parameters for ^{13}C detected experiments at 310 K.

Experiments used for OPN sequence specific assignment	Dimensions of acquired data			Spectral width (ppm)			ns ^a	d1 ^b
	t ₁	t ₂	t ₃	F ₁	F ₂	F ₃		
	<i>¹³C detected</i>							
CON	1024 (¹⁵ N)	1024 (¹³ C)		45	30		2	2
(H)CBCACON	144 (¹³ C)	128 (¹⁵ N)	1024 (¹³ C)	60	35	30	2	0.9
(H)CBCANCO	144 (¹³ C)	128 (¹⁵ N)	1024 (¹³ C)	60	35	30	4	0.9
(HCA)COCON	128 (¹³ C)	148 (¹⁵ N)	1024 (¹³ C)	12	35	30	4	1.6
(H)CCCON	196 (¹³ C)	128 (¹⁵ N)	1024 (¹³ C)	64	35	30	4	0.9

^a Number of acquired scans.
^b Relaxation delay in seconds.

Table S2. Acquisition parameters for ^{13}C detected PRE and ^{15}N -T₂ experiments at 310K.

Experiments used to measure PREs*	Dimensions of acquired data			Spectral width (ppm)			ns ^a	d1 ^b
	t ₁	t ₂	t ₃	F ₁	F ₂	F ₃		
	<i>¹³C detected</i>							
C'-CON	512 (¹⁵ N)	1024 (¹³ C)		36	31		4 ^c	6
H ^α -CON	512 (¹⁵ N)	1024 (¹³ C)		36	31		8 ^c	4
H ^N -CON	384 (¹⁵ N)	1024 (¹³ C)		31	24		4	4
(H ^α C ^α)CON ^{Pro}	220 (¹⁵ N)	1024 (¹³ C)		30	7.5		16	1.4

*Experiments were measured at 310 K. For one of the single cysteine-mutated proteins experiments, the one harbouring the paramagnetic tag at position 188, experiments were also acquired at 293 K.
^a Number of acquired scans.
^b Relaxation delay in seconds. In order to avoid unwanted effects, long inter-scan delays were used to ensure complete recovery of the magnetization to equilibrium before acquisition of the following transient.
^c Two repeats of the experiments were acquired back-to-back and added afterwards.

Table S3. Acquisition parameters for ¹H detected experiments.

Experiments used for PRE, relaxation rates and Trp dynamics of OPN	Dimensions of acquired data			Spectral width (ppm)			ns ^a	d1 ^b
	t ₁	t ₂	t ₃	F ₁	F ₂	F ₃		
	¹H detected							
PRE H ^N -T ₂ (<i>Coturnix japonica</i> OPN)	768 (¹⁵ N)	2048 (¹ H)		32	14		4	1.5
PRE H ^N -T ₂ (<i>Homo sapiens</i> OPN)	512 (¹⁵ N)	2048 (¹ H)		32	14		8	1.5
¹⁵ N-T ₁	640 (¹⁵ N)	2048 (¹ H)		32	16		8	1.4
¹⁵ N-T ₂	640 (¹⁵ N)	2048 (¹ H)		32	16		4	1.4
¹⁵ N-{ ¹ H} hetNOE	256 (¹⁵ N)	2048 (¹ H)		34	16		8	10
NOESY-HSQC	128 (¹⁵ N)	128 (¹ H)	2048 (¹ H)	36	14	14	4	0.9
¹ H- ¹³ C HSQC aromatics (o1p=120ppm)	512 (¹³ C)	2048 (¹ H)		35	16		4	1.5

^a Number of acquired scans.
^b Relaxation delay in seconds.

Table S4. Chemical shifts of *Coturnix japonica* OPN minor forms at T 310K.

	H ^N (ppm)*	C ^γ (ppm)	N (ppm)	C ^α (ppm)	C ^β (ppm)	C ^δ (ppm)	C ^γ (ppm)
N78		174.546		53.096	39.139		
A79	8.049		124.405	52.299	19.436		
D80		174.925		54.296	41.223		
V81	7.689	174.619	118.915	58.942	34.495		21.189 / 20.522
P82	-		138.067	62.732	34.495	49.837	24.714
E83		175.58		56.462	30.243		
Q84	8.01	174.249	119.888	53.096	29.844		
P85	-		135.165	62.486	34.318	49.73	24.574
D86		174.809		53.982	41.176		
F87	7.919	174.125	120.735	55.665	41.049		
P88	-		138.308	62.723	33.922	49.692	24.332
D89		175.553		54.296	40.949		
V90		174.563		58.864	34.208		21.352 / 20.031
P91	-		137.631	62.474	34.210	48.880	24.701
F118		175.747		57.191	39.579		
T119	8.06	173.668	115.784	61.777	70.015		
D120	8.115	174.482	122.529	53.888	41.102		
F121	7.935	173.972	120.534	55.703	41.338		
P122	-	176.965	138.343	62.764	32.904	49.800	24.435
T123	8.126	174.363	114.103	61.876	69.665		
E124	8.197	175.527	122.883	55.958	30.465		
A125		175.342	126.298	50.24	18.228		
P126	-	176.33	134.880	62.471	34.030	49.847	24.720
V127	7.976	174.745	118.66	60.736	33.671		
A128	8.093	175.842	126.741	50.439	17.824		
P129	-	175.831	134.232	62.395	33.772	49.754	24.229
F130	8.461	175.348	122.52	57.689	38.705		
N131	8.346		119.248	52.742			
S177		173.481		58.208	63.797		
Q178	8.033	174.271	120.842	52.908	29.799		33.503
P179	-	175.988	135.775	62.505	34.277	49.963	24.649
A180	8.407		125.12	52.619	19.264		
G181		173.876		45.154			
L182	7.867	176.783	121.1	55.217	42.08		
W183	7.881	174.596	120.184	56.444	29.515		
W184	7.781	174.499	122.377	55.163	31.082		24.563
P185	-		138.11	62.555	33.582	49.768	24.560
L196		175.460		52.400	42.101		
P197	-		134.330	62.160	34.651	50.011	24.512
S206		173.539		58.348	63.814		
R207	8.043	174.287	121.928	53.628	31.66		27.251
P208	-		135.367	62.232	34.253	50.034	24.561
K209		174.121		56.046	31.566		
F210		175.171	122.508	55.43	29.585		
D211			122.219				
Q232		176.364		55.958	29.365		
G233		173.913	109.672	45.14			
S234	8.114	173.28	115.852	58.145	63.903		
V235	7.918	174.594	120.842	58.942	34.309		21.214 / 20.38
P236	-	175.727	138.447	62.486	34.229	49.96	24.811
A237	8.523	177.862	124.75	52.388	19.002		
V238	8.102		119.172	61.954	33.255		

*Those shifts were measured at 310 K using a Bruker Avance III 800 MHz spectrometer equipped with a standard PA-TXI probehead. H^N shifts were connected to C and N shifts via the HNCO experiment.

Table S5. CIDER calculated parameters of OPN protein sequences from different species.

	N	f₋	f₊	FCR	NCPR	κ
<i>Homo sapiens</i>	314	0.23885	0.09236	0.33121	-0.14650	0.27505
<i>Cholorocebus sabaebus</i>	314	0.23567	0.09873	0.33439	-0.13694	0.27756
<i>Mus musculus</i>	294	0.23129	0.08163	0.31293	-0.14966	0.25526
<i>Rattus norvegicus</i>	317	0.23659	0.08833	0.32492	-0.14826	0.24346
<i>Oryctolagus cuniculus</i>	311	0.24437	0.08682	0.33119	-0.15756	0.24260
<i>Sus scrofa</i>	303	0.24422	0.09901	0.34323	-0.14521	0.23465
<i>Bos taurus</i>	278	0.22302	0.10432	0.32734	-0.11871	0.26739
<i>Ovis aries</i>	278	0.24101	0.10432	0.34532	-0.13669	0.26658
<i>Equus caballus</i>	313	0.24920	0.09585	0.34505	-0.15335	0.23792
<i>Gallus gallus</i>	264	0.20833	0.10227	0.31061	-0.10606	0.25812
<i>Coturnix japonica</i>	264	0.21212	0.10227	0.31439	-0.10985	0.26179

N = number of residues

f₋ = fraction of negatively charged residues

f₊ = fraction of positively charged residues

FCR = fraction of charged residues

NCPR = net charge per residue

κ = charge patterning parameter

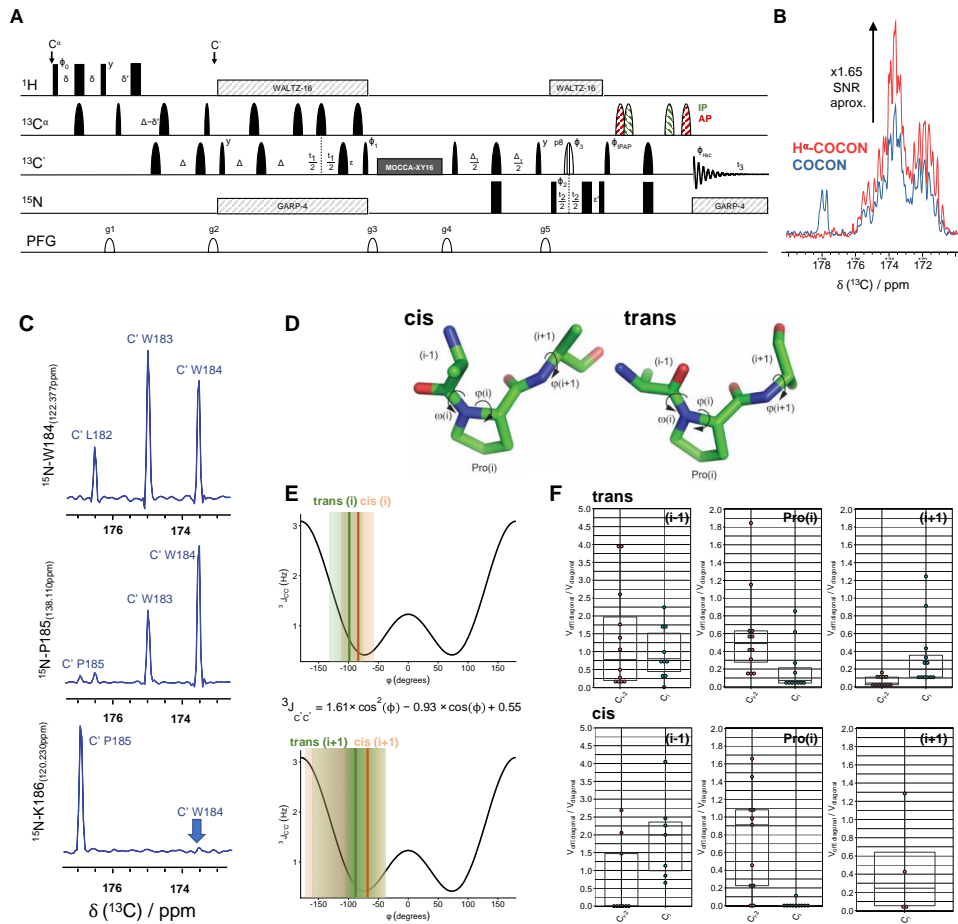


Fig. S1. (HCA)COCON experiment used to complete the unambiguous assignment of 99,5% of N, C', C^β and C^α major forms of *Coturnix japonica* OPN and 100% of *cis*-Pro signals. (A) Pulse sequence scheme. The following phase cycling was employed: $\phi_0 = x, -x$; $\phi_1 = 4(x), 4(-x)$; $\phi_2 = 2(x), 2(-x)$; $\phi_3 = 8(x), 8(-x)$; $\phi_{rec} = x, 2(-x), x, -x, 2(x), -x$; $\phi_{IP-AP} = x$; $\phi_{IP-AP} = -y$ (the IP and AP block with the C^α π pulses shifted is depicted as green and red dashed pulses, respectively). The quadrature detection was achieved through the States-TPPI method by incrementing the phase of the $\pi/2$ pulse prior to the evolution period; C^α homonuclear decoupling for C' acquisition was achieved using the IPAP scheme. The band-selective hatched pulse is an adiabatic pulse that inverts both C' and C^β nuclei. The length of the delays were set as $\delta = 1/4J_{H^{\alpha}C^{\alpha}}$, $\delta' = 1/6J_{H^{\alpha}C^{\alpha}}$, $\Delta = 1/2J_{C^{\alpha}C'}$, $\Delta_1 = 1/2J_{C^{\alpha}N}$. (B) Signal improvement of the (HCA)COCON experiment (starting from the polarization of the ¹H^α) over the conventional COCON [1]. (C)

Selected traces of a typical (HCA)COCON pattern of a preceding (upper) or following (lower) residue from a proline (middle). (D) Calculated conformations from an average of dihedral angles of a *trans* or *cis* proline. (E) Karplus curve of $^3J_{CC}$ with the most common values of the φ angle in a proline (upper) or following residue (lower). Green and orange boxes indicate the distribution of $C'_i-C'_{i+1}$ dihedral angle in PDB for *trans* or *cis* prolines (upper) and residues following *trans* or *cis* prolines (lower), respectively. Green and orange lines indicate the most common value. (F) Peak volumes of the off-diagonal signals in the (HCA)COCON spectrum normalized by the diagonal peak volume.

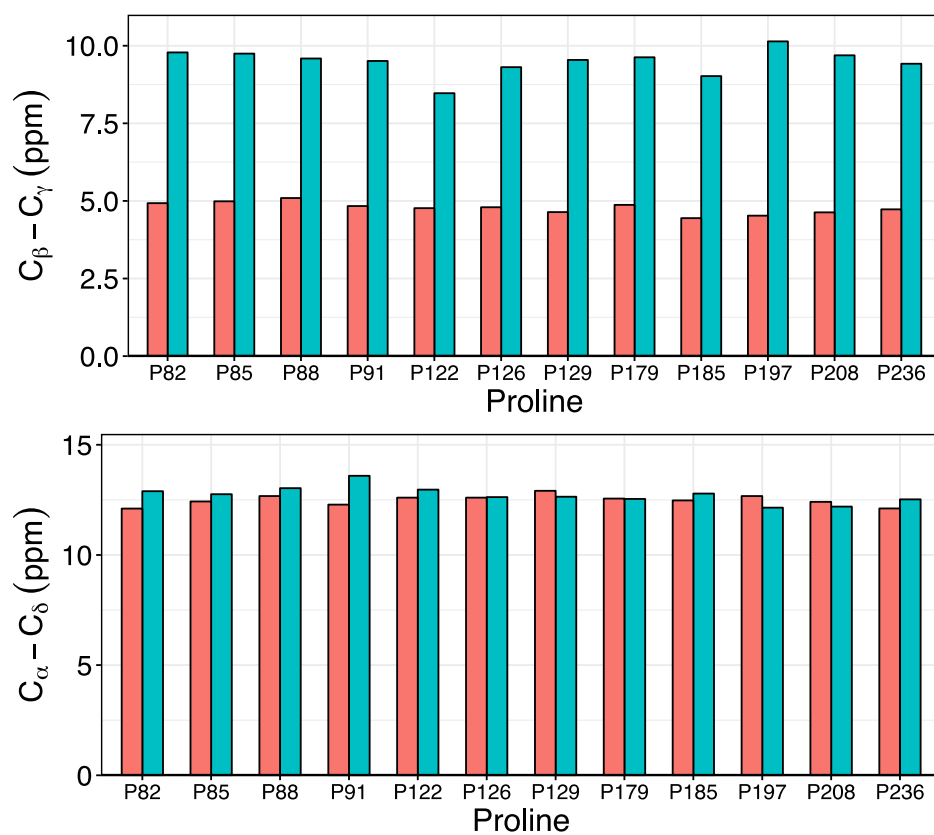


Fig. S2. Characterization of *cis* conformation by measuring the difference in proline sidechain chemical shifts of the major (salmon) and minor (cyan) forms. For some proline residues there is an additional minor form with corresponding side-chain shifts of a *trans* conformer (see **Fig. 1B**). Possible explanations are reported in the main text.

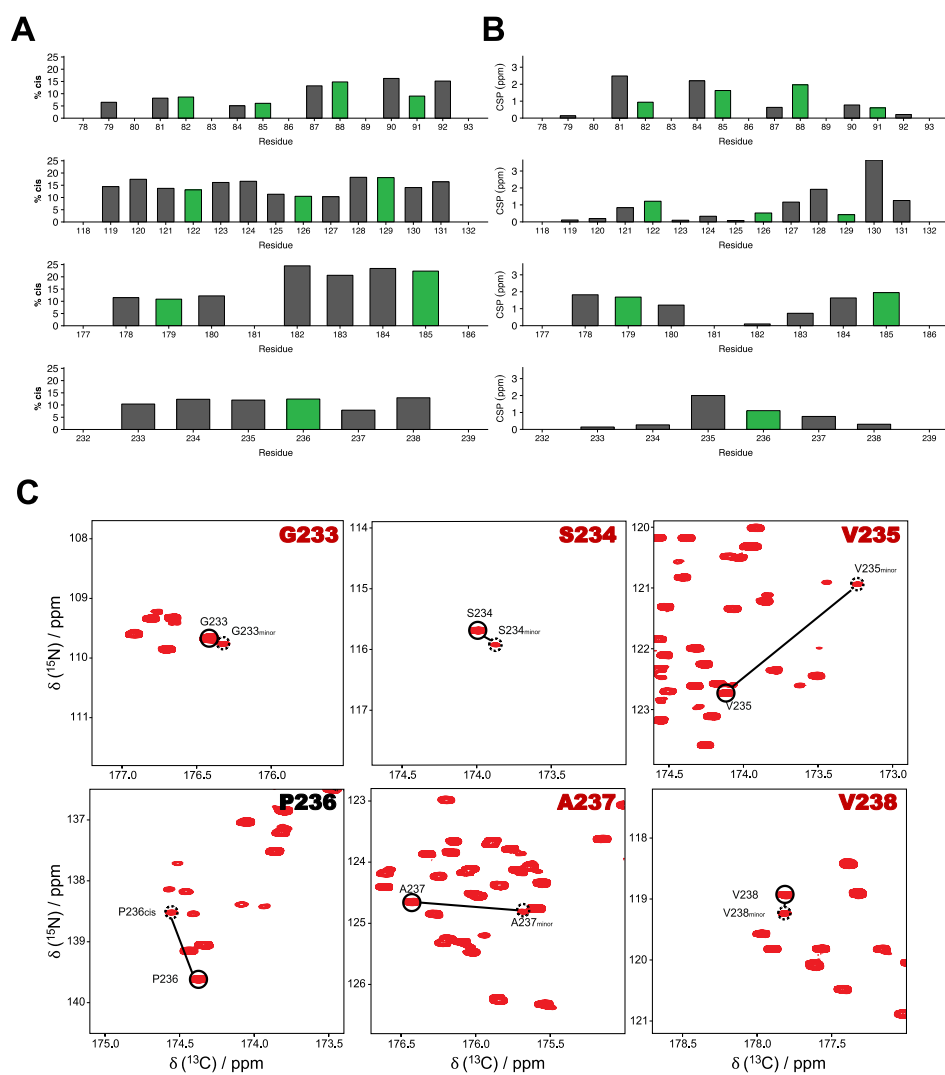


Fig. S3. (A) *Cis* population for proline residues (green) and their assigned neighboring residues (gray) as measured from relative peak volumes. The percentage of minor forms is determined by the percentage of *cis* conformer of the closest Pro. (B) Chemical shift difference (^{13}C and ^{15}N combination) between major and minor forms of proline residues (green circle) and assigned neighboring signals. (C) *cis*-Pro (dashed circle) of P236 and the minor forms generated by *cis* conformation in this proline. Note the gradual decrease of the chemical shifts difference between the major and the minor form further away from the proline.

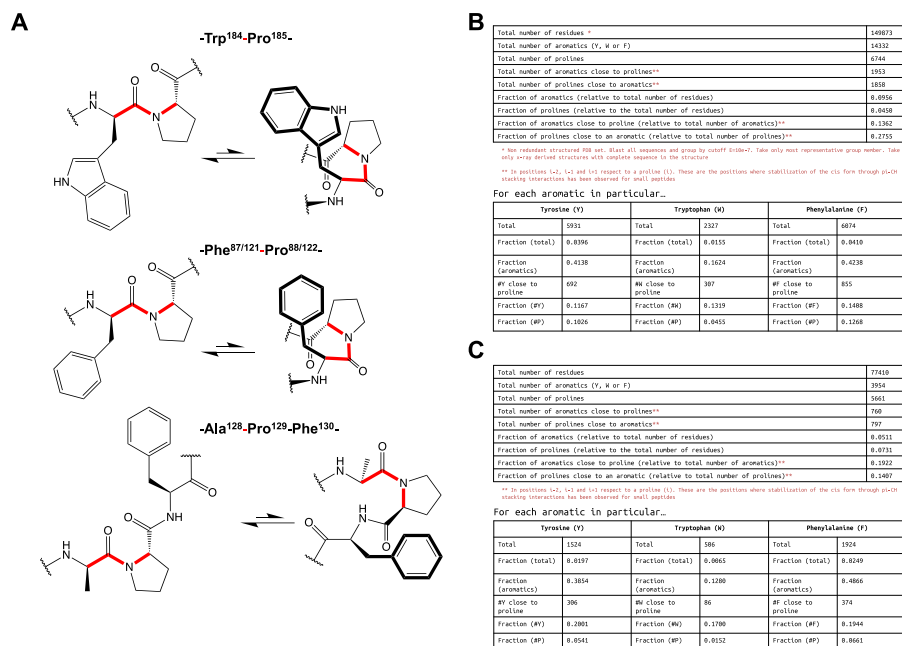


Fig. S4. (A) Schematic representation of *cis/trans* equilibrium for those proline residues in OPN with an aromatic residue in position $i \pm 1$. π -CH interactions between aromatic and proline ring can stabilize the *cis* conformation (B) Search in PDB database (folded proteins) [2] the occurrence of proline-aromatic motifs. (C) Search in DisProt database (unfolded proteins) [3] the occurrence of proline-aromatic motifs. The search of aromatic motifs were limited to position $i \pm 1$ and $i - 2$ since are the position were the π -CH *cis* stabilization was described for small peptides [4,5].

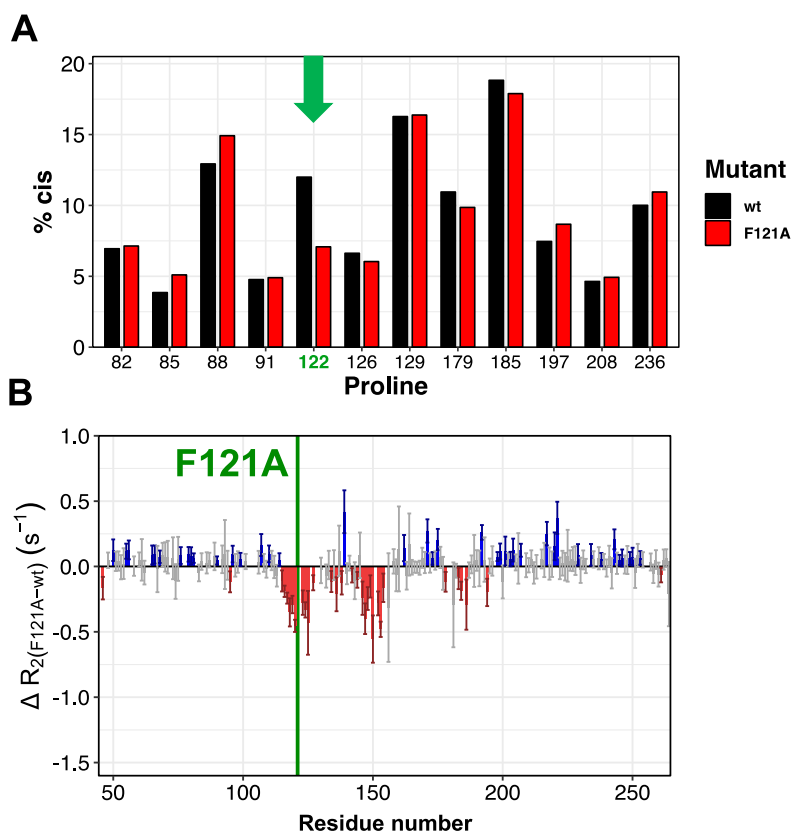


Fig. S5. Effect of mutating an aromatic residue next to a proline (Phe-Pro) with an alanine residue (Ala-Pro) in *Coturnix japonica* Osteopontin. (A) Percentage of *cis* population obtained by measuring the ratio of peak height between *cis* and *trans* forms for proline signals in CON spectra. Wt data are represented in black while data on the mutant F121A are shown in red; the green arrow indicates the neighboring proline (P122). (B) Difference (mutant – wt) in ¹⁵N-R₂ relaxation rates measured at 800 MHz, 310K. Error bars are determined as described in the Materials and Methods section. Significant negative values are colored in red while positive are in blue. Insignificant changes are colored in gray. Note the reduction in ¹⁵N-R₂ relaxation rates of the mutant in the region of the mutation as well as in a neighboring region of the compact state. The region of the mutation gains flexibility and the central region is moderately less compact.

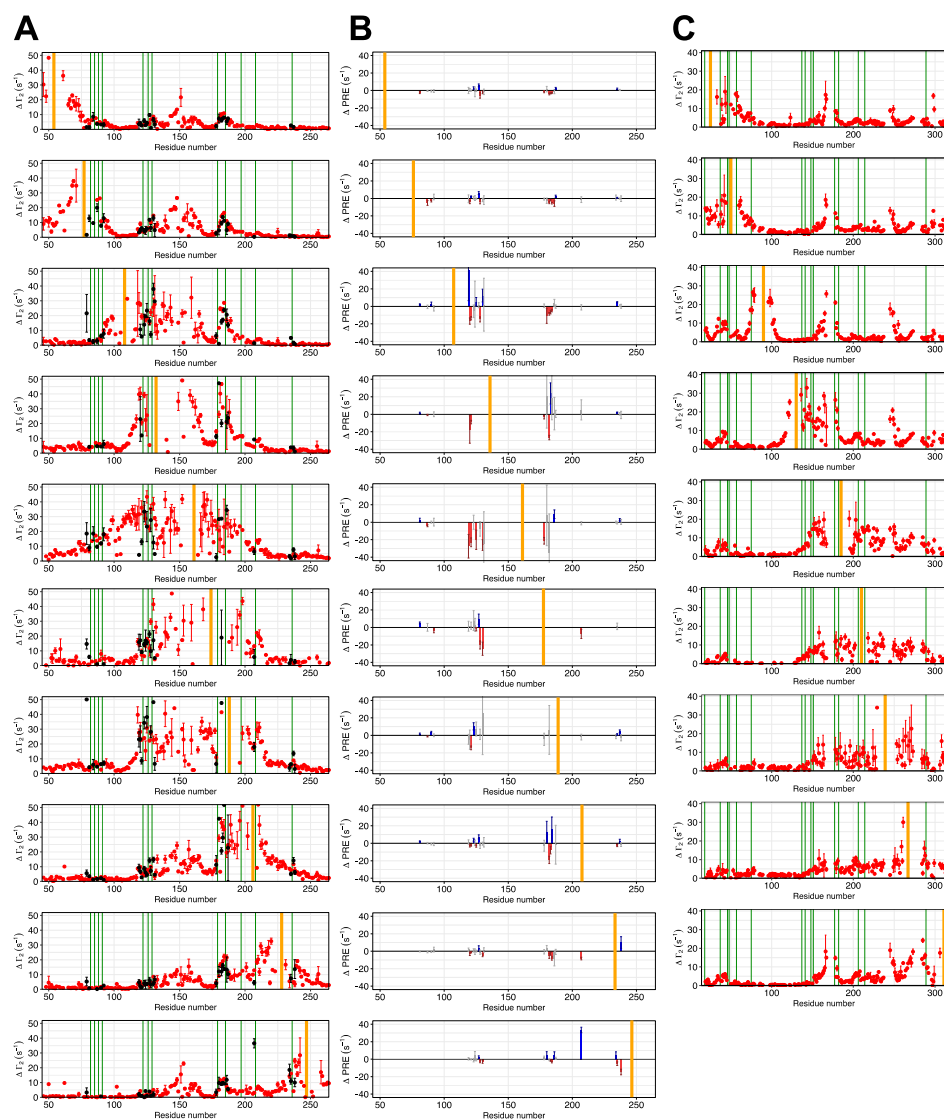


Fig. S6. 800 MHz $\Delta^1\text{H}^n\text{-}\Gamma_2$ PREs measured at T 298K for *Coturnix japonica* OPN (A) and *Homo sapiens* OPN (C). Major forms (red) and minor ones (black) with corresponding fitting error bars are reported. Vertical green lines indicate the proline residue positions. Orange vertical lines indicates the position of the spin-label. (B) Differences in $^1\text{H}^n\text{-}\Gamma_2$ PRE between minor (*cis*-Pro) and major (*trans*-Pro) form. Statistically meaningless changes are depicted in gray.

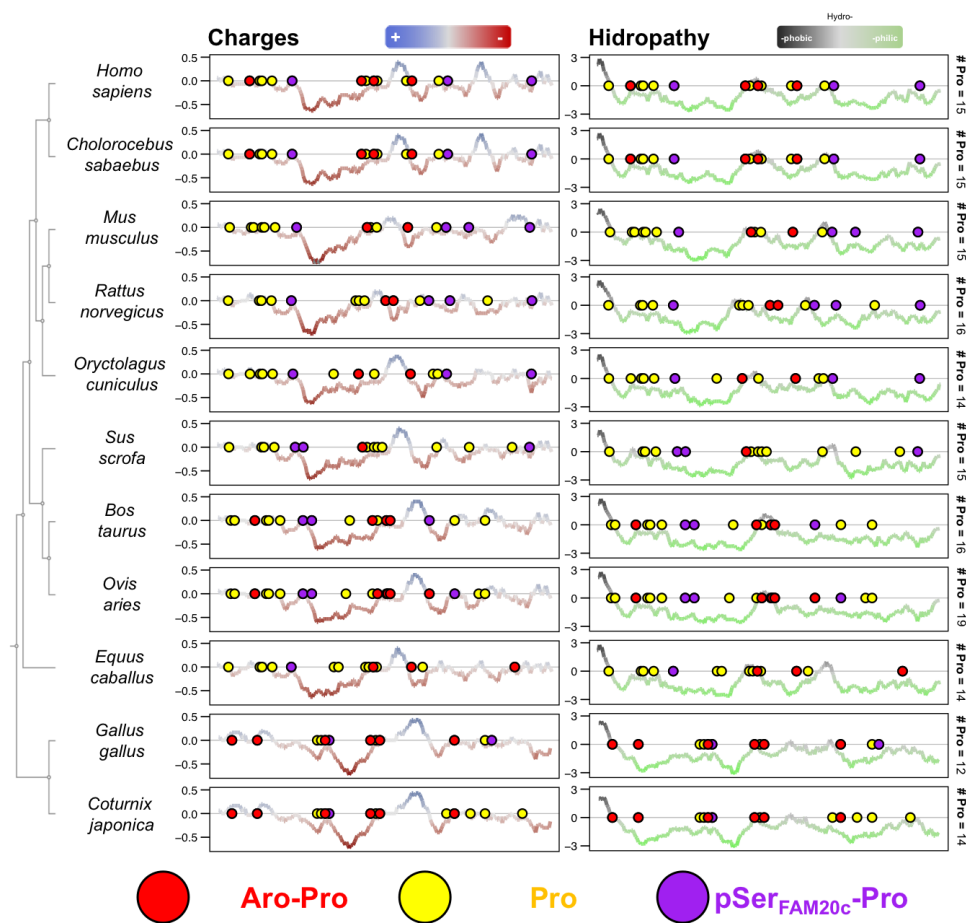


Fig. S8. Charge and hydrophobicity plots of the OPN sequences used for the MSA. The tree is based on the MSA shown in Fig. S7. The residue number is plotted on the x-axis with the corresponding hydrophobicity/charge values on the y-axis. Hydrophobicity was calculated according to the scale of Kyte & Doolittle [6] and charge by setting charged residues to +1 or -1 respectively (for histidine a value of +0.5 was assumed). The range of the hydrophobicity plot is -3.0 to 3.0, with high values corresponding to hydrophobic regions. For each residue an average value from a window of size 15 was calculated. The dotted line indicates a value of zero. Total number of prolines in the primary sequence are indicated on the right side.

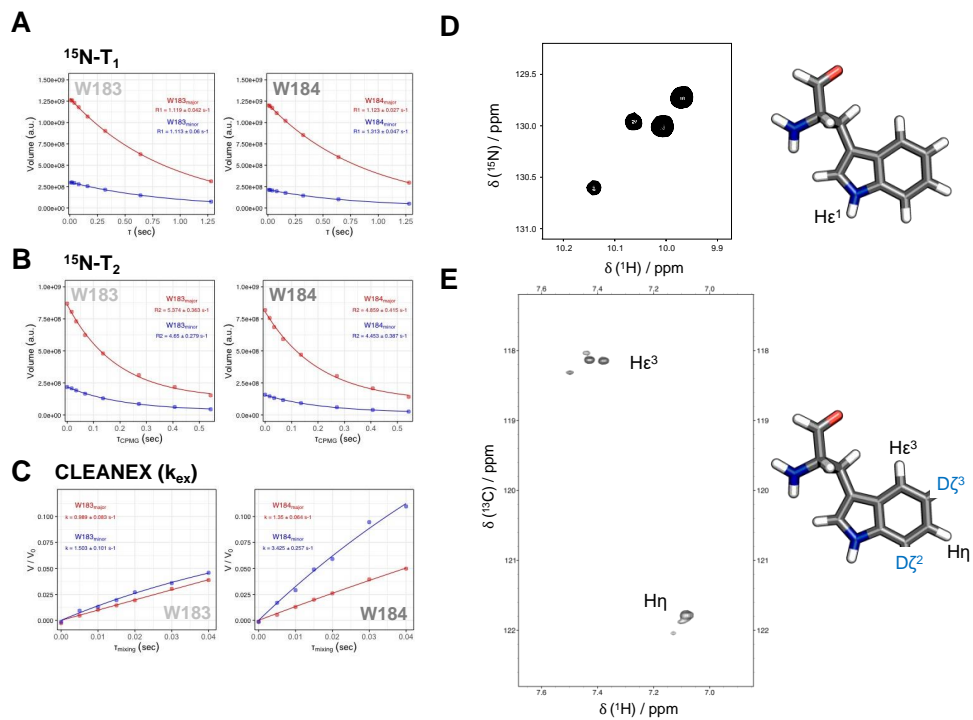


Fig. S9. Tryptophan sidechain dynamics of the Trp¹⁸³-Trp¹⁸⁴-Pro¹⁸⁵ stretch. (A) $^{15}\text{N-R}_1$ rates, (B) $^{15}\text{N-R}_2$ rates and (C) $\text{H}^{\text{N}}/\text{H}_2\text{O}$ exchange rates. Major and minor forms of W183 and W184 arising from the isomerization state of P185. (D) $\text{H}_{\epsilon 1}$ - $\text{N}_{\epsilon 1}$ major and minor forms observed in a ^1H - ^{15}N -HSQC. (E) $\text{H}_{\epsilon 3}$ - $\text{C}_{\epsilon 3}$ and H_{η} - C_{η} major and minor forms observed in a ^1H - ^{13}C -HSQC

References

- [1] W. Bermel, I. Bertini, I.C. Felli, Y.-M. Lee, C. Luchinat, R. Pierattelli, Protonless NMR Experiments for sequence-specific assignment of backbone nuclei in unfolded proteins, *J. Am. Chem. Soc.* 128 (2006) 3918–3919. doi:10.1021/ja0582206.
- [2] H.M. Berman, The Protein Data Bank, *Nucleic Acids Res.* 28 (2000) 235–242. doi:10.1093/nar/28.1.235.
- [3] D. Piovesan, F. Tabaro, I. Mičetić, M. Necci, F. Quaglia, C.J. Oldfield, M.C. Aspromonte, N.E. Davey, R. Davidović, Z. Dosztányi, A. Elofsson, A. Gasparini, A. Hatos, A. V. Kajava, L. Kalmar, E. Leonardi, T. Lazar, S. Macedo-Ribeiro, M. Macossay-Castillo, A. Meszaros, G. Minervini, N. Murvai, J. Pujols, D.B. Roche, E. Salladini, E. Schad, A. Schramm, B. Szabo, A. Tantos, F. Tonello, K.D. Tsigos, N. Veljković, S. Ventura, W. Vranken, P. Warholm, V.N. Uversky, A.K. Dunker, S. Longhi, P. Tompa, S.C.E. Tosatto, DisProt 7.0: A major update of the database of disordered proteins, *Nucleic Acids Res.* 45 (2017) D219–D227. doi:10.1093/nar/gkw1056.
- [4] K.M. Thomas, D. Naduthambi, N.J. Zondlo, Electronic control of amide cis-trans isomerism via the aromatic-prolyl interaction, *J. Am. Chem. Soc.* 128 (2006) 2216–2217. doi:10.1021/ja057901y.
- [5] N.J. Zondlo, Aromatic–Proline interactions: electronically tunable CH/ π interactions, *Acc. Chem. Res.* 46 (2013) 1039–1049. doi:10.1021/ar300087y.
- [6] J. Kyte, R.F. Doolittle, A simple method for displaying the hydrophobic character of a protein, *J. Mol. Biol.* 157 (1982) 105–132. doi:10.1016/0022-2836(82)90515-0.

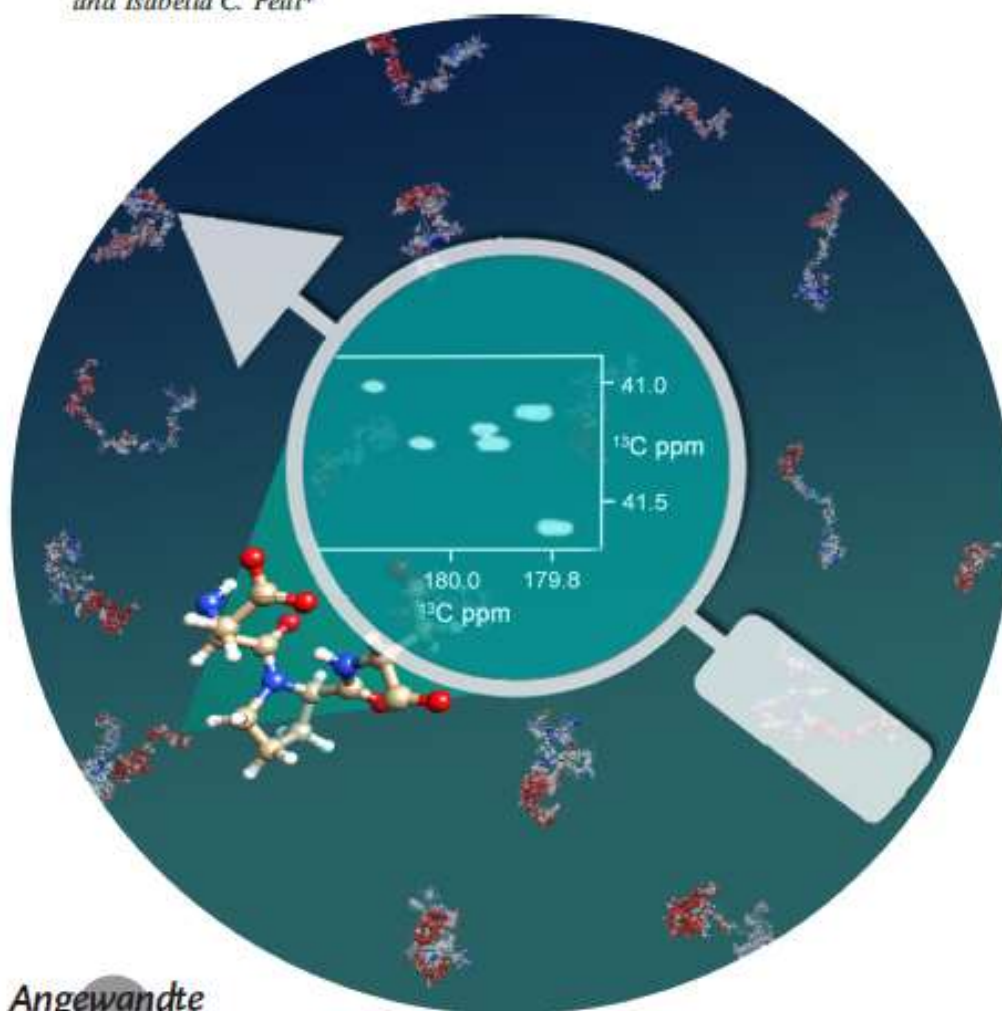
9. Monitoring the interaction of α -synuclein with calcium ions through exclusively heteronuclear Nuclear Magnetic Resonance Experiments

VIP **Protein Function** Very Important Paper

How to cite: *Angew. Chem. Int. Ed.* 2020, 59, 18537–18545
International Edition: doi.org/10.1002/anie.202008079
German Edition: doi.org/10.1002/ange.202008079

Monitoring the Interaction of α -Synuclein with Calcium Ions through Exclusively Heteronuclear Nuclear Magnetic Resonance Experiments

Letizia Pontoriero^a, Marco Schiavina^a, Maria Grazia Murrari, Roberta Pierattelli,^a and Isabella C. Felli^{a*}



Angewandte
International Edition
Chemie

Angew. Chem. Int. Ed. 2020, 59, 18537–18545

© 2020 Wiley-VCH GmbH

Wiley Online Library 18537

VIP Protein Function Very Important Paper

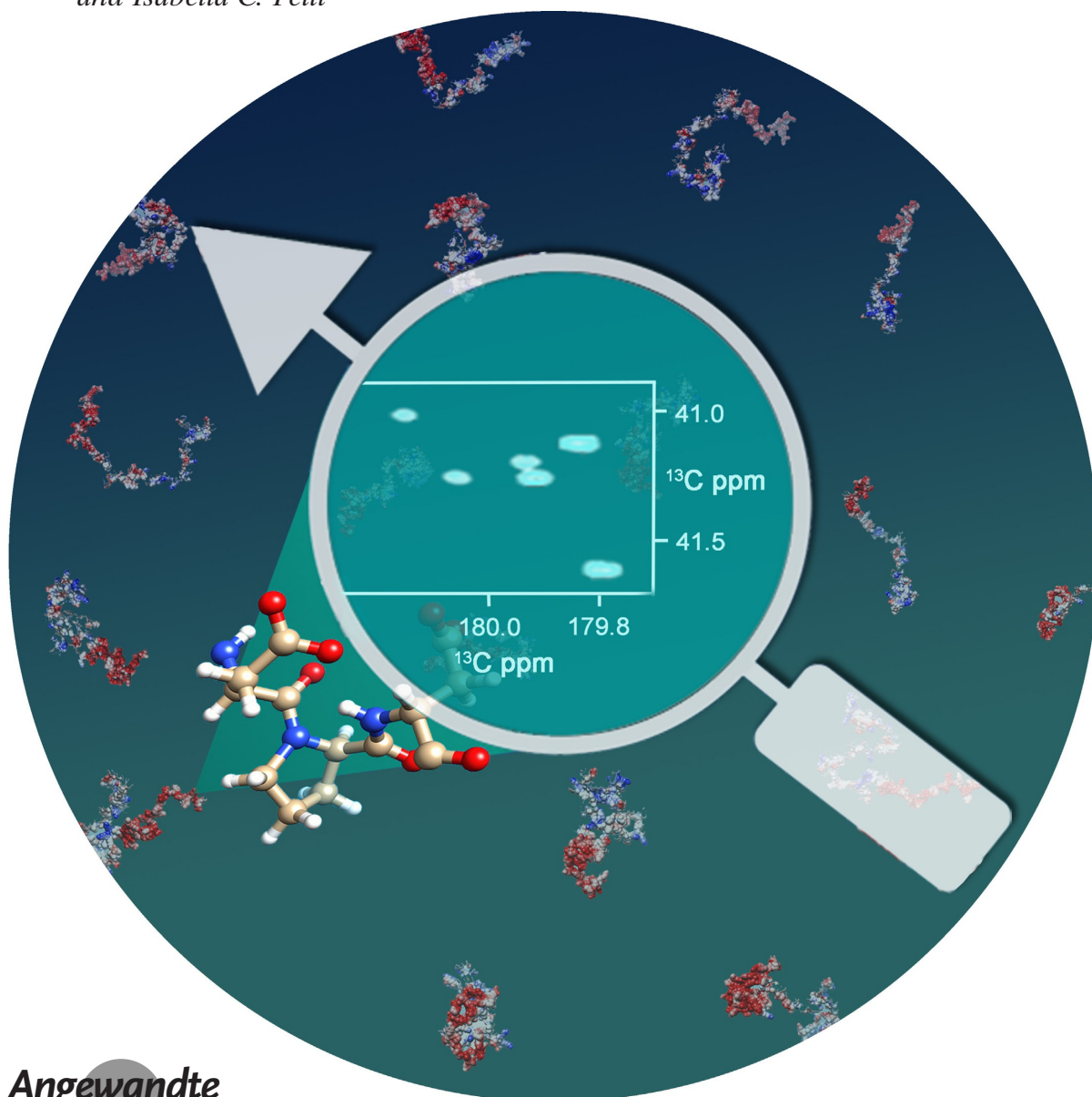
How to cite: *Angew. Chem. Int. Ed.* **2020**, *59*, 18537–18545

International Edition: doi.org/10.1002/anie.202008079

German Edition: doi.org/10.1002/ange.202008079

Monitoring the Interaction of α -Synuclein with Calcium Ions through Exclusively Heteronuclear Nuclear Magnetic Resonance Experiments

Letizia Pontoriero⁺, Marco Schiavina⁺, Maria Grazia Murrari, Roberta Pierattelli,^{*} and Isabella C. Felli^{*}



Abstract: Many properties of intrinsically disordered proteins (IDPs), or protein regions (IDRs), are modulated by the nature of amino acid side chains as well as by local solvent exposure. We propose a set of exclusively heteronuclear NMR experiments to investigate these features in different experimental conditions that are relevant for physiological function. The proposed approach is generally applicable to many IDPs/IDRs whose assignment is available in the Biological Magnetic Resonance Bank (BMRB) to investigate how their properties are modulated by different, physiologically relevant conditions. The experiments, tested on α -synuclein, are then used to investigate how α -synuclein senses Ca^{2+} concentration jumps associated with the transmission of nerve signals. Novel modules in the primary sequence of α -synuclein optimized for calcium sensing in highly flexible, disordered protein segments are identified.

Introduction

Intrinsically disordered proteins (IDPs) and protein regions (IDRs), which challenge the canonic structure-function paradigm, represent an emerging field of research in modern protein chemistry.^[1–4] Highly flexible proteins and flexible linkers of complex proteins are present in any living organism and play key roles in a variety of different cellular pathways. They lack a stable three-dimensional structure in their native conditions while retaining biological activity. Initially described using creative epithets such as “dancing proteins”, “protein clouds”, “protein chameleons”^[1] they are now widely investigated revealing novel ways through which extensive disorder and flexibility modulate protein function.

The structural and dynamic properties of IDPs/IDRs are even more influenced by the environment with respect to those of globular proteins.^[5,6] Therefore experimental tools to study them in physiologically relevant conditions, all the way to in-cell, are very useful to understand the physicochemical properties relevant for their function and misfunction. In this framework, NMR spectroscopy provides a unique investigation tool to access high resolution information.^[7,8]

Human α -synuclein is one of the most widely studied IDPs because of its involvement in several human neurodegenerative pathologies called synucleinopathies, such as Parkinson's disease (PD). Constituted by 140 amino acids, α -synuclein is intrinsically disordered in native conditions.

The primary sequence is generally subdivided into three different regions: the N-terminus (1–60), with several KTKXGV recognition motifs responsible for a net positive charge, the central, more hydrophobic, non-amyloid- β component NAC (61–94) and the C-terminal tail (95–140), a characteristic domain that is dense of negatively charged residues. Largely disordered in the monomeric state,^[9] stabilized by long range interactions between the N-terminal region and the C-terminal region,^[10,11] it adopts helical conformations when interacting with membranes through the N-terminal region^[12–15] and elongated conformations in amyloid fibrils,^[16] just to name a few snapshots on the most studied, heterogeneous structural properties of α -synuclein in different conditions.^[17] In recent years many in vitro and in vivo studies were carried out to clarify the events that lead to the insurgence of different pathologies but the structural and dynamical versatility to different local conditions encountered in neuronal cells makes it difficult to identify those factors that trigger the pathological action of the protein as the formation of toxic aggregates. Full comprehension of the pathological and physiological roles of this biomolecule is still lacking in part because of the incredible range of environment-dependent conformational plasticity, a true “chameleon protein”,^[18] that renders its investigation very challenging.

Here we would like to present a novel set of 2D NMR experiments to follow how the properties of IDPs/IDRs change in different, physiologically relevant, experimental conditions. Based on carbonyl carbon direct detection,^[19–25] these experiments provide information on backbone and side-chain chemical shifts as well as on the impact of solvent exchange at the residue level, even for those residues whose amide proton is not directly detectable. This set of 2D exclusively heteronuclear NMR experiments is tested on α -synuclein and then used to focus on its interaction with Ca^{2+} , a potential trigger for the onset of Parkinson's disease. Mainly localized in presynaptic terminals, α -synuclein is exposed to microdomains of high Ca^{2+} concentration associated with neurotransmitter release^[26,27] and could be exposed to high extracellular Ca^{2+} concentration in cell-to-cell secretion mechanisms.^[28] New insights on how a structurally and dynamically heterogeneous protein linked to the onset of Parkinson's disease senses these Ca^{2+} concentration jumps become thus very relevant to describe α -synuclein function.

Results and Discussion

Fingerprint of an IDP at physiological pH and temperature

The side chains of amino acids are seldom studied for IDPs/IDRs, even if they are expected to play important roles for their function because the extensive resonance overlap typical of their NMR spectra becomes even more pronounced when moving away from the backbone. 2D ^1H - ^{13}C correlation spectra, even if highly sensitive, only show a small fraction of resolved cross-peaks, drastically reducing their high resolution information content. Carbon-13 detected 2D NMR experiments provide a valuable source of information. A

[*] L. Pontoriero,^[†] M. Schiavina,^[†] Dr. M. G. Murralli, Prof. R. Pierattelli, Prof. I. C. Felli
CERM and Department of Chemistry “Ugo Schiff”
University of Florence
Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Florence (Italy)
E-mail: roberta.pierattelli@unifi.it
felli@cerm.unifi.it

Dr. M. G. Murralli
Present address: Department of Chemistry and Biochemistry
University of California at Los Angeles (USA)

[†] These authors contributed equally to this work.

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under:
<https://doi.org/10.1002/anie.202008079>.

fingerprint of an IDP/IDR at physiological pH and temperature can be obtained through the set of 2D exclusively heteronuclear NMR experiments based on carbonyl carbon direct detection (CON, CACO, CBCACO, and CCCO).^[29,30] However, extremely high resolution is needed to resolve resonances of side chains which cluster in very narrow spectral regions. To this end CACO, CBCACO, and CCCO pulse sequences were modified to achieve the necessary resolution to study IDPs/IDRs (Supporting Information, Figure S1); experimental variants exploiting ¹H polarization as a starting source (¹H-start) were also implemented to increase the sensitivity of the experiments (Figure S2).

Carbon-13 detected 2D NMR experiments reveal atomic resolution information for aliphatic as well as for carbonyl/carboxylate resonances of amino acid side chains (-COO⁻, -CONH₂). As an example of the quality of the spectra that can be obtained, the assignment of the resonances of the six aspartate residues present in α -synuclein is shown in Figure 1. As illustrated for Asp 2, starting from the backbone carbonyl (C^{*o*}) identified through the CON (C^{*r*}N_{*r+1*}), the resonances of C^{*o*}_{*i*} and C^{*o*}_{*i*} can be easily identified through inspection of the CACO and CBCACO. These are also correlated to the side-chain carboxylate carbon resonance (C^{*o*}_{*i*}) through two additional cross-peaks in a close but well isolated spectral region. Therefore, the C^{*o*}_{*i*}-C^{*o*}_{*i*} cross-peaks of the six aspartate residues can be easily assigned in a sequence-specific manner. Analogously, also asparagine side-chain resonances can be assigned. For glutamate and glutamine residues inspection of CCCO is also needed to unambiguously correlate the backbone carbonyl resonance to the C^{*o*}_{*i*}, C^{*o*}_{*i*}, C^{*o*}_{*i*} aliphatic side-chain resonances and finally to the C^{*o*}_{*i*} carbonyl/carboxylate resonances. The cross-peaks assignment of the carboxylate/carboxyl functional groups for α -synuclein (C^{*o*}_{*i*}-C^{*o*}_{*i*} for Asp and Asn, C^{*o*}_{*i*}-C^{*o*}_{*i*} for Glu and Gln), which fall in a very clean spectral region, is reported in Table S2 (Supporting Information).

It is worth noting that carbonyl/carboxylate side-chain resonances are seldom assigned, in particular for IDPs/IDRs. They can be detected through triple resonance experiments based on amide proton detection.^[31–33] However this approach is bound to fail in conditions in which amide protons are not detectable, such as for solvent-exposed protein backbones at physiological pH and temperature. 2D exclusively heteronuclear NMR experiments enable us to easily assign side-chain resonances, starting from the backbone assignment, adjusting chemical shifts to the conditions under investigation through inspection of a CON spectrum, followed by the analysis of the CACO/CBCACO/CCCO spectra. This constitutes a general approach to access additional key information for any IDP/IDR whose assignment is available in the Biological Magnetic Resonance Bank (BMRB, <http://www.bmrwisc.edu/>). This set of spectra thus provides a unique tool to achieve a fingerprint of an IDP near physiological conditions not only for backbone resonances but also for side chains.

Negatively charged side chains of aspartate and glutamate residues are the first candidates to establish interactions with oppositely charged polypeptide chains^[34] as well as to interact with metal ions.^[35–39] Particularly relevant for α -synuclein function is the interaction with Ca²⁺ involved in the trans-

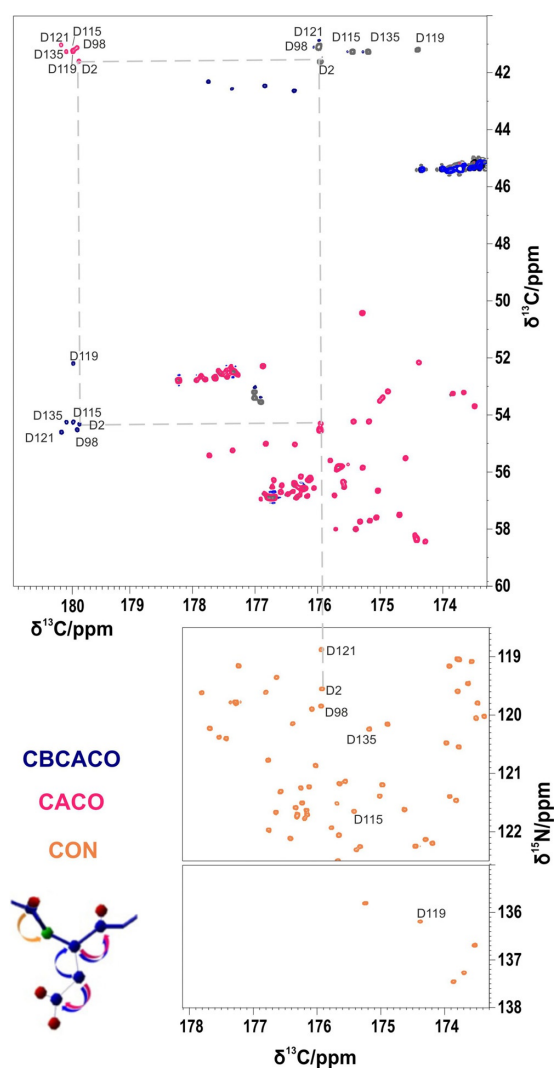


Figure 1. An illustration of the strategy used to obtain the sequence-specific assignment of the ¹³C resonances of aspartate residues through 2D exclusively heteronuclear NMR experiments. As an example, gray dotted lines indicate the steps followed to assign side-chain resonances of Asp 2. Starting from the carbonyl resonance identified in the CON spectrum (orange), C^{*o*}_{*i*} and C^{*o*}_{*i*} are identified in CACO (red) and CBCACO (blue) spectra, superimposed in the Figure, and correlated to C^{*o*}_{*i*} through the respective C^{*o*}_{*i*}-C^{*o*}_{*i*} and C^{*o*}_{*i*}-C^{*o*}_{*i*} cross-peaks in a sequence-specific manner.

mission of nervous signals.^[39–41] While intracellular Ca²⁺ concentrations are generally very low, microdomains of high Ca²⁺ concentrations are linked to the release of neurotransmitter from presynaptic terminals in all neurons.^[26] Having the assignment in hand, it is now possible to “zoom in” on the metal ion coordination sphere and access additional complementary information to that available through HN HSQC experiments.^[39]

The set of 2D exclusively heteronuclear NMR experiments, in particular the CON and the CACO, was used to monitor the changes in α -synuclein induced by the presence of Ca^{2+} . The chemical shift changes of Asp/Glu residues signals upon addition of Ca^{2+} , reported in Figure 2, show that not all of them are affected to the same extent: major changes are observed in the C-terminal region of the protein (110–140), the second part of the so-called acidic region (95–140). As expected, chemical shift changes of side-chain carboxylates are larger with respect to those observed for backbone carbonyl (C') resonances (Figure S3), reflecting a more direct effect experienced by side-chain nuclear spins upon interaction with calcium ions. No major changes in secondary structural propensity of the backbone were identified upon interaction with calcium ions (Figure S4).

To assess the general applicability of this set of ^{13}C detection experiments, the interaction studies with calcium ions were repeated using a sample with an order of magnitude lower concentration of α -synuclein (50 μM). Despite the relatively low sensitivity of ^{13}C , the experiments allowed the obtaining of a clear interaction profile (Figure S5); thanks to the inclusion of ^1H as starting polarization source the (H)CACO could be acquired in a few hours (Figures S2 and S5).

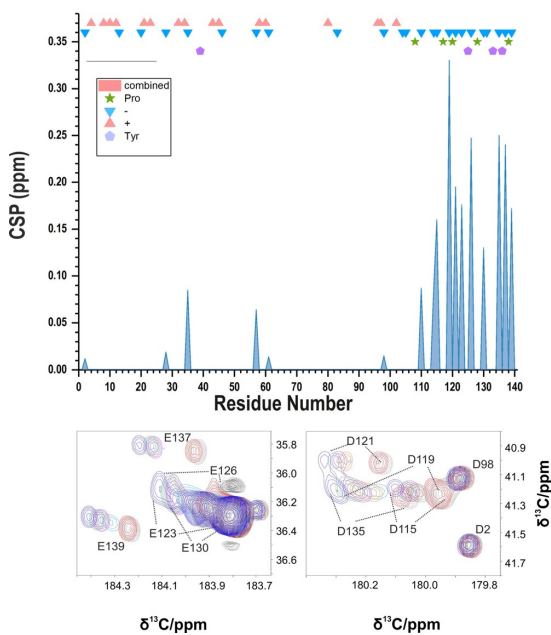


Figure 2. Chemical shift perturbation (difference in absolute value) of aspartate and glutamate side-chain ^{13}C resonances upon addition of Ca^{2+} . The symbols over the graph depict the distribution of charged, tyrosine and proline residues to evidence the particular composition along the primary sequence: Asp and Glu (blue triangles), Lys (red triangles), Pro (stars) and Tyr (pentagons). The lower panels show two regions of the CACO spectrum with cross-peaks of Asp and Glu side chains and their shifts during the titration. The different extent of the perturbation of Asp and Glu side-chain resonances is evident. Major changes are observed in the C-tail, which is rich in negative charges.

Shifting our attention to backbone nuclear spins by inspection of the combined chemical shift changes of C' and N chemical shifts,^[42] not only direct but also indirect changes derived from the interaction with Ca^{2+} can be monitored (Figure 3). Major perturbations are observed in the final part of the primary sequence (Leu 113, Pro 120, Tyr 125, Met 127, Ser 129, Tyr 136 and Pro 138).

The final part of the polypeptide chain is rich in proline residues with four of the five proline residues of the protein located between residues 117 and 138 (4 out of 22 amino acids in this region). Chemical shift changes of proline residues' signals, that could be monitored through CON spectra (Figure 3), clearly show that two of them, Pro 120 and Pro 138, are significantly perturbed upon addition of Ca^{2+} , indicating that they are involved in the interaction of α -synuclein with calcium ions. This may appear surprising since proline does not have metal binding properties. However, these two proline residues are both flanked by two negatively charged amino acids, which all experience significant chemical shift changes for carboxylate resonances. Further inspection of the most pronounced backbone chemical shift changes reveals that two tyrosine residues, which are not so common in IDPs, are also significantly perturbed by Ca^{2+} addition. These observations prompted us to inspect the

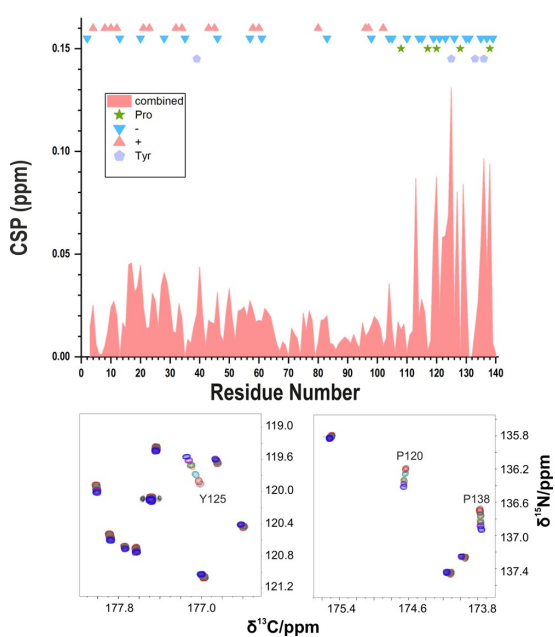


Figure 3. Chemical shift perturbation^[42] $\sqrt{0.15\delta(^{15}\text{N}_i)^2 + 0.34\delta(^{13}\text{C}_i)^2}$ of CON signals upon addition of Ca^{2+} . The lower panels show the shift for the two proline residues (Pro 120 and Pro 138 (right)) and for tyrosine 125 (left), the most affected residues during the titration. While the C-terminus is still the most perturbed part, the plot highlights a more indirect effect for the backbone resonances and also reveals an impact on many residues in the N-terminal domain. The titration follows the pattern from blue (α -syn: Ca^{2+} , 1:0) to purple (α -syn: Ca^{2+} , 1:16).

positions of proline and tyrosine residues along the primary sequence (schematically depicted in the top of Figures 2 and 3). The results show that the region of the primary sequence of α -synuclein experiencing the most pronounced changes upon Ca^{2+} interaction, that is the final part of the C-terminal region which is very rich in negatively charged amino acids while being depleted of positively charged residues, has also a peculiar abundance of proline and tyrosine residues. The NAC region instead is the one characterized by the smallest chemical shift changes while the N-terminal part of the protein experiences significant variations of backbone chemical shifts. These are less pronounced with respect to those observed for the C-terminal region and are likely to arise in part from an indirect effect of calcium binding, resulting from reduced long range electrostatic interactions between the initial and final part of the protein as observed in other studies.^[43,44] It is thus interesting to investigate whether the interaction with Ca^{2+} promotes compaction or decompaction along the primary sequence.

“Spying” chemical exchange of IDPs with water: the DeCON experiment

Exchange of amide protons with the solvent, responsible in our experimental conditions for broadening beyond detection more than half of the signals of amide protons (Figure S6),^[45] has been one of the first NMR observables used in the past to identify amide protons protected from solvent exchange by globular protein folds.^[46] On the other hand, only a little information is available so far on how solvent exchange is modulated by the properties of IDPs/IDRs. It is thus interesting to investigate this aspect in more detail. A modified variant of the CON was thus designed to reintroduce a dependence on chemical exchange processes with the solvent without perturbing the solvent resonance, still retaining the excellent resolution of CON spectra. The modified pulse sequence (Figure 4A) enables us to create three spin order operator ($4C'_2N_2H_2$) and to monitor its decorrelation due to chemical exchange, along the lines of a method initially proposed by Skrynnikov and Ernst.^[47]

The novel experiment (DeCON) can thus provide information about exchange processes of labile amide protons with the solvent also for residues that escape detection in ^1H - ^{15}N -based experiments. Starting from C'_z magnetization (a), transverse C'_y coherence is created (b) and converted into antiphase coherence $2C'_xN_z$ (c). This is then converted into $2C'_zN_y$ (d) and allowed to evolve under the effect of the $^1J_{\text{HN}}$ coupling to generate $4C'_zN_xH_z$ (e). In order to generate this latter operator a band-selective 180° pulse on the amide proton region is used to avoid perturbation of the water resonance. This operator is then converted to the three spin order $4C'_zN_zH_z$ (f) and its decay is monitored by introducing a free evolution delay (τ_{decor}). At the end of this delay the three spin order $4C'_zN_zH_z$ is converted to $4C'_zN_yH_z$ (g) which is picked up by the second part of the CON after conversion into $2C'_zN_y$ (h). It is worth noting that through this approach a dependence on solvent exchange is reintroduced with minimal perturbation of the water resonance avoiding

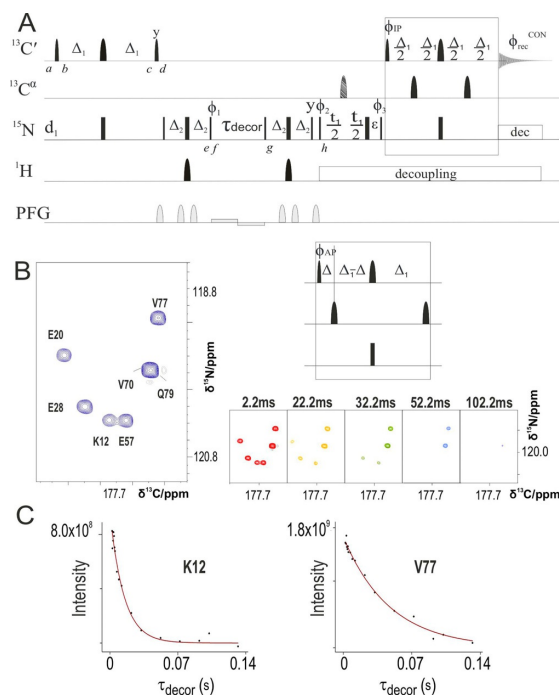


Figure 4. A) DeCON pulse sequence. The following phase cycling was employed: $\phi_1 = 2(\gamma), 2(-\gamma); \phi_2 = x, -x; \phi_3 = 4(x), 4(-x), \phi^{\text{IP}} = x; \phi^{\text{AP}} = -\gamma$ and $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x$. The length of the delays were: $\Delta = 4.5$ ms $\Delta_1 = 16.6$ ms; $\Delta_2 = 2.7$ ms; $\epsilon = t_1(0) + p180$ (500 μs). The striped pulse in the middle of the ^{15}N evolution period is an adiabatic chirp pulse that covers the whole ^{13}C spectral region. Virtual decoupling of the $C'-C''$ coupling was achieved by acquiring both the IP and AP component of the signals for each increment. The strength of the smoothed square shape gradients are: 50%, 19%, 19%, 25%, 25%, 70%; the strength of the weak bipolar gradient is 1%. Quadrature detection in the indirect dimension was obtained with the STATES-TPPI approach incrementing phase ϕ_2 . B) A portion of the 2D DeCON spectrum is shown on the left with the assignment of the cross-peaks; several spectra acquired as a function of τ_{decor} are shown on the right. C) The intensities (arbitrary units) of two of these cross-peaks are reported as a function of τ_{decor} .

radiation damping effects. As an example, a region of the spectrum obtained with this novel DeCON experiment is shown in Figure 4B as a function of τ_{decor} , the time interval in which the three spin order $4C'_zN_2H_2$ is allowed to evolve: while the signal of Val 77 is still observable with $\tau_{\text{decor}} = 52.2$ ms, one of the Lys 12 signals disappears with $\tau_{\text{decor}} = 22.2$ ms. The intensities of cross-peaks can be integrated in the series of spectra acquired with a different τ_{decor} and can be fit to a mono-exponential decay ($I_{\text{zzz}}(\tau_{\text{decor}}) = I_0 e^{-k_{\text{zzz}}\tau_{\text{decor}}}$, Figure 4C).

It is interesting to compare the results obtained through the DeCON experiment proposed here with the ones obtained through the initially proposed ^1H detected variant^[47] (HN-Decor experiment). The agreement between the data measured through the two different experiments is quite good for the residues that could be detected in both experiments

(Figure S7). The DeCON however provides information about a larger number of residues with respect to the ^1H detected variant. This is in part due to the improved resolution derived from the superior chemical shift dispersion of the $C'_{i-1}-N_i$ correlations with respect to the H_i-N_i correlations, and in part to the different magnetization transfer pathway minimizing perturbation of ^1H magnetization (in the HN coherence transfer pathway the proton magnetization is transverse both during the two INEPT steps and the acquisition, in the DeCON the proton magnetization is maintained along the z -axis). These properties of the DeCON experiment allow us to monitor a larger number of residues and to extend the range of k_{zzz} values that can be measured with respect to the ^1H detected variant. The minor contribution of longitudinal relaxation to the observed decay was evaluated investigating the decay of the $2C'_zN_z$ operator (Figure S8). Finally, a three-dimensional variant of the DeCON experiment was designed to further increase the resolution of the experiment in a third dimension exploiting C^α chemical shifts, opening the possibility of studying IDPs of increasing size (Figure S9).

The k_{zzz} values determined through the DeCON experiment are reported as a function of the residue number in Figure S10. The residues in the initial part of the polypeptide chain show significantly high values which are however quite scattered along the primary sequence, an effect largely due to the type of amino acid as indicated in Figure S10. The C-terminal region (110–140) shows significantly reduced exchange processes, in agreement with previous observations attributed to the effect of the high local negative potential.^[9] With increasing temperature or pH (or both) the k_{zzz} values increase while the trend along the protein primary sequence is maintained (data not shown).

Upon addition of Ca^{2+} to the α -synuclein sample the k_{zzz} values determined through the DeCON experiment show a global enhancement along the primary sequence, while maintaining the general trend as shown in Figure S11. The effect appears more relevant for the residues in the terminal parts; the ratio between k_{zzz} of the bound and the unbound forms is shown in Figure 5. In the central zone (residues 40–100), the average value of $k_{zzz}(\alpha\text{-synCa})/k_{zzz}(\alpha\text{-syn}) = 1.10$, while several residues in the terminal parts (1–39 and 101–140) present a higher ratio. The residues that experience the higher boost (over $\times 1.9$ times respect to the unbound form) are Ala 18, Ala 30, Glu 115, Asp 119, Asp 121, Asn 122, Glu 123, Ala 124, Tyr 125, Ser 129, Tyr 133, Gln 134, Asp 135 and Tyr 136. The global increase observed in the exchange rates could be due in part to an increase in ionic strength during the titration with CaCl_2 . In contrast, the very strong and localized effect in the terminal domains should be related to different reasons, such as reduction of the electrostatic potential in the C-terminal region and disruption of the electrostatic interactions between the N-terminal and C-terminal parts of the polypeptide chain. On the other hand the present data show no evidence of formation of a more compact state in which solvent exchange is precluded upon interaction with calcium ions.

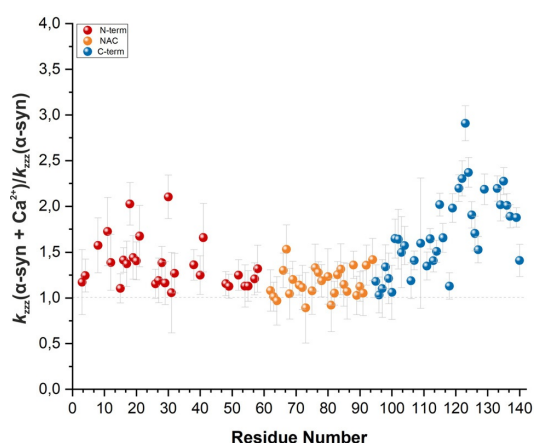


Figure 5. Ratio of the k_{zzz} values obtained from the DeCON experiments before and after addition of Ca^{2+} . The plot evidences the different effects in the increment for the three α -synuclein regions: while NAC (yellow) maintains a homogeneous trend, C-terminus (blue) shows the major boost, in line with the defined binding region. However, it is possible to see an increment also for some scattered residues in the N-terminal part (red).

Ca^{2+} sensing by α -synuclein: new insights

The high number of negatively charged amino acids in the final part of the primary sequence of α -synuclein, in combination with the disordered nature of this protein that leaves this part of the backbone largely solvent-exposed and easily accessible, provides a strong electrostatic negative potential that is likely to have an important role for its function, in particular in mediating interactions with positively charged entities (ions, small molecules, proteins, etc.). This is likely to create the initial driving force, sensed at quite long distance, for the interaction with Ca^{2+} .^[35] The negatively charged functional groups of amino acid side chains, such as carboxylate groups of aspartate and glutamate residues (COO^-), in principle all have the potential to interact with positively charged metal ions. The disordered nature of the polypeptide that leaves COO^- groups largely exposed to the solvent would suggest an unspecific effect, with all COO^- sharing similar interaction properties. Instead, we find very specific differential effects for the COO^- groups in the different parts of the polypeptide chain. The sequence context thus has an important role in mediating interactions with Ca^{2+} even in the case of IDPs. The possibility to directly observe perturbations sensed by COO^- groups through ^{13}C detection experiments allows us to zoom in on the interaction site and identify the amino acids that are the most perturbed by calcium addition. Interestingly the largest perturbations are found in the C-terminal tract for the following residues: Asp 119, Asp 121, Glu 123, Glu 126, Asp 135, Glu 137 and Glu 139. These residues belong to an extended region of the polypeptide chain (119–139, 21 amino acids long) showing that Ca^{2+} already has a strong preference for a subset of the COO^- groups in the C-terminal region in which α -synuclein is usually subdivided (95–140). Looking in more detail, two

regions, which are quite far from each other, can be identified: Asp 119–Glu 126 and Asp 135–Glu 139. These two distinct regions share very similar patterns: 1) negatively charged amino acids are close in the primary sequence but not contiguous, 2) in two cases amino acids in between negatively charged amino acids are prolines, and 3) in two cases glutamate is preceded by tyrosine. Therefore the signature in terms of amino acidic composition of these regions strongly perturbed by the addition of calcium ions is very characteristic. Specific patterns can thus be identified that are likely to play an important role in modulating calcium ion interactions: a pair of negatively charged amino acids (Asp or Glu) separated by a proline (Asp 119–Pro 120–Asp 121 and Glu 137–Pro 138–Glu 139), tyrosine–glutamate motifs (Tyr 125–Glu 126 and Tyr 136–Glu 137).

The role of a proline in between two negatively charged amino acids could thus be important to reduce local mobility and favor the proper relative orientation of negatively charged side chains for calcium binding (Figure 6). Tyrosine residues are very bulky amino acids with aromatic side chains rich in π electron density, two properties that could play a relevant role in reducing local motions and favoring interactions of highly flexible protein regions with Ca^{2+} , in particular if followed by an acidic residue providing a COO^- group. These could be key elements of specific motifs to modulate Ca^{2+} sensing in highly flexible protein tracts. The question of whether a stable complex is formed or an equilibrium between different local binding sites with similar affinities is established remains. On one hand the flexibility of the polypeptide chain provides to the system the necessary degrees of freedom to fold around a unique metal binding site, on the other hand the entropic penalty associated to folding a tract of 20 amino acids is expected to be much higher than that of multiple sites with comparable affinity in equilibrium, each of them comprising 6–8 amino acids.

It is thus interesting to zoom out and inspect chemical shift changes observed for backbone nuclear spins in this region. The major changes are observed for residues 119–129 and for residues 134–139. Interestingly the two proline residues in the new motif identified from side-chain chemical shifts (Pro 120 and Pro 138, both flanked by negatively charged amino acids) are the ones that show the largest

chemical shifts changes (Figure 3), confirming their important role in the interaction with Ca^{2+} . Tyrosine 125 also shows pronounced changes upon interaction (Figure 3) as well as Tyr 136 (data not shown). Chemical shift changes, although significant, do not indicate the formation of a defined folded state. The overall properties of this tract are still in line with a highly flexible state. Exchange properties of amide protons with the solvent, as monitored through the novel DeCON experiment, show an increase in the decorrelation upon addition of Ca^{2+} which shows that the backbone is still largely accessible to solvent exchange, definitely far from forming a protected pocket in which solvent exchange is precluded. Therefore the interaction of α -synuclein with Ca^{2+} appears more in line with a fuzzy interaction in which flexibility and disorder is maintained also upon interaction. The strong electrostatic potential of the C-terminal tail could play the initial important driving force, sensed also at long range, for the interaction with calcium ions; the identified motifs in the primary sequence could act as nucleation sites for the interaction. A number of conformations would then be easily accessible to engage other Asp/Glu residues in the interaction with calcium ions.

The interaction of α -synuclein with Ca^{2+} is likely to disrupt the interaction of the C-terminal region with the N-terminal region, rich of lysine residues. This region adopts a helical conformation when bound to membranes^[12–14] and was proposed to form long range interactions at the origin of a compact state of α -synuclein populated in solution.^[10,11,48,50] Our data confirm that chemical shift changes are observed for residues in the N-terminal region which could be explained, at least in part, by disruption of long range interactions. Indeed, most of the COO^- groups of aspartate and glutamate side chains in the N-terminal region only show modest chemical shift changes, much less than those observed in the C-terminal region. In addition, changes in solvent exchange properties are observed through DeCON for the initial part of the polypeptide chain. These could result in part by an increased solvent accessibility resulting from loss of the compact conformation. However, we cannot rule out the possibility of the occurrence of other intermolecular effects. A detailed investigation of the long range effects of Ca^{2+} addition would require additional experiments, for example exploiting paramagnetic relaxation enhancements induced by the presence of a paramagnetic tag, which might take advantage of ^{13}C NMR detection experiments.^[51]

Metal binding sites of globular, folded proteins have been extensively investigated revealing their key role in structure function relationships. The interactions of metal ions with highly flexible protein regions instead are only beginning to be investigated in detail to understand their structural and dynamic properties and their impact on protein function. The experiments proposed here provide a useful tool to investigate the interactions of flexible protein tracts with metal ions at high resolution. The example of the interaction of α -synuclein with Ca^{2+} reveals specific motifs in the protein primary sequence providing a glimpse on the wide versatility through which proteins modulate interactions with calcium ions also through high flexibility and disorder. Very few of these disordered motifs have been investigated at high

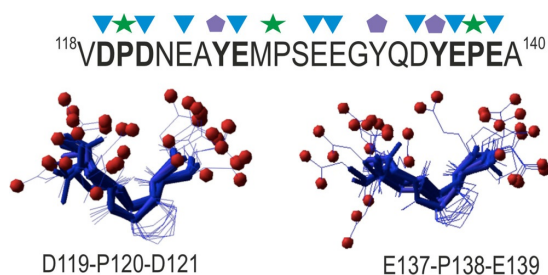


Figure 6. Structural models of the DPD and EPE motifs identified in α -synuclein as strongly perturbed as a result of Ca^{2+} concentration jumps. The structural conformers are calculated through Flexible Meccano^[48] without imposing any constraints. The Figure was obtained using MOLMOL.^[49]

resolution and many more could be studied in detail with the tools proposed here. The characteristic features identified in α -synuclein might also be useful as input for bioinformatics tools to search for similar Ca^{2+} binding patterns in disordered proteins.

α -synuclein has also been shown to interact with other metal ions.^[36–38,52] Among them Cu^{II} , Fe^{II} , Co^{II} , Ni^{II} , Mn^{II} , and several lanthanide metal ions^[38,53,54] many of which are paramagnetic. The set of experiments proposed here might be useful to provide additional insights on the mode of interaction. As an example we tested Mn^{II} , which provides very strong paramagnetic effects deriving from the high number of unpaired electrons combined with a relatively long electronic relaxation time.^[55] Previous NMR investigations revealed that Mn^{II} interacts with residues in the C-terminal region of α -synuclein, in a very similar way to what observed for Ca^{2+} .^[38] A sub-stoichiometric concentration of Mn^{II} (1/100 respect to the protein concentration) indeed shows that first carboxylate groups to be perturbed by the interaction are the same ones identified in the interaction with Ca^{2+} (Figure 7). Further additions of Mn^{II} allow to progressively zoom out and identify the first backbone resonances to be perturbed as well as the region mainly affected.

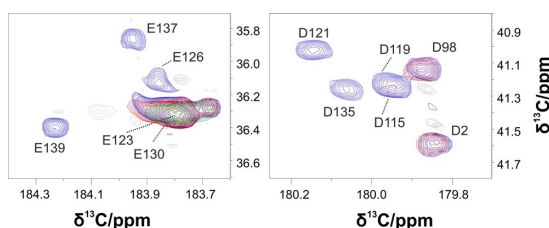


Figure 7. The panels show two regions of the CACO spectrum with cross-peaks of Asp and Glu side chains and their shifts upon Mn^{2+} addition. The Figure shows that the first carboxylate groups to be perturbed by the interaction with Mn^{2+} are the same ones identified in the interaction with Ca^{2+} .

Conclusion

The improved set of 2D exclusively heteronuclear NMR experiments based on carbonyl direct detection enabled us to resolve the signals of COO- groups of α -synuclein amino acid side chains and to monitor local solvent exposure. These experiments allowed us to zoom in on the metal ion coordination sphere, revealing novel motifs involved in the interaction with calcium ions. This represents just one example of the key role played by solvent-exposed side chains in modulating the biological function of highly flexible protein tracts. Post-translational modifications, which often involve solvent-exposed amino acid side chains, introduce another layer of complexity in modulating protein function that can be studied through the approach presented herein. The proposed experiments can thus become a tool of general interest to characterize the properties of IDPs/IDRs in physiologically relevant conditions that have not been studied

so far, thereby significantly expanding our knowledge on how protein function is modulated by disorder and flexibility.

Acknowledgements

The support and the use of resources of the CERM/CIRMMP center of Instruct-ERIC is gratefully acknowledged. This work has been supported by the Italian Ministry for University and Research (FOE funding) and by a grant of the Fondazione CR Firenze to RP.

Conflict of interest

The authors declare no conflict of interest.

Keywords: ^{13}C detection · calcium binding · IDPs · side chains · water exchange

- [1] A. K. Dunker, M. M. Babu, E. Barbar, M. Blackledge, S. E. Bondos, Z. Dosztányi, H. J. Dyson, J. Forman-Kay, M. Fuxreiter, J. Gsponer, K.-H. Han, D. T. Jones, S. Longhi, S. J. Metallo, K. Nishikawa, R. Nussinov, Z. Obradovic, R. V. Pappu, B. Rost, P. Selenko, V. Subramaniam, J. L. Sussman, P. Tompa, V. N. Uversky, *Intrinsically Disord. Proteins* **2013**, *1*, e24157.
- [2] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, M. M. Babu, *Chem. Rev.* **2014**, *114*, 6589–6631.
- [3] J. Habchi, P. Tompa, S. Longhi, V. N. Uversky, *Chem. Rev.* **2014**, *114*, 6561–6588.
- [4] M. Arbesú, M. Pons, *Arch. Biochem. Biophys.* **2019**, *677*, 108161.
- [5] L. Geist, M. A. Henen, S. Haiderer, T. C. Schwarz, D. Kurzbach, A. Zawadzka-Kazimierzczuk, S. Saxena, S. Zerko, W. Koźmiński, D. Hinderberger, R. Konrat, *Protein Sci.* **2013**, *22*, 1196–1205.
- [6] F. Theillet, A. Binolfi, T. Frembgen-Kesner, K. Hingorani, M. Sarkar, C. Kyne, C. Li, P. B. Crowley, L. Gierasch, G. J. Pielak, A. H. Elcock, A. Gershenson, P. Selenko, *Chem. Rev.* **2014**, *114*, 6661–6714.
- [7] *Intrinsically Disordered Proteins Studied by NMR Spectroscopy* (Eds.: I. C. Felli, R. Pierattelli), Springer, Heidelberg, **2015**.
- [8] J. H. Ardenkjaer-Larsen, G. S. Boebinger, A. Comment, S. Duckett, A. S. Edison, F. Engelke, C. Griesinger, R. G. Griffin, C. Hilty, H. Maeda, G. Parigi, T. Prisner, E. Ravera, J. Van Bentum, S. Vega, A. Webb, C. Luchinat, H. Schwalbe, L. Frydman, *Angew. Chem. Int. Ed.* **2015**, *54*, 9162–9185; *Angew. Chem.* **2015**, *127*, 9292–9317.
- [9] R. L. Croke, C. O. Sallum, E. Watson, E. D. Watt, A. T. Alexandrescu, *Protein Sci.* **2008**, *17*, 1434–1445.
- [10] M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, C. M. Dobson, *J. Am. Chem. Soc.* **2005**, *127*, 476–477.
- [11] C. W. Bertocini, Y. Jung, C. O. Fernández, W. Hoyer, C. Griesinger, T. M. Jovin, M. Zweckstetter, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 1430–1435.
- [12] S. Chandra, X. Chen, J. Rizo, R. Jahn, T. C. Südhof, *J. Biol. Chem.* **2003**, *278*, 15313–15318.
- [13] T. S. Ulmer, A. Bax, N. B. Cole, R. L. Nussbaum, *J. Biol. Chem.* **2005**, *280*, 9595–9603.
- [14] C. R. Bodner, C. M. Dobson, A. Bax, *J. Mol. Biol.* **2009**, *390*, 775–790.

- [15] G. Fusco, T. Pape, A. D. Stephens, P. Mahou, A. R. Costa, C. F. Kaminski, G. S. Kaminski Schierle, M. Vendruscolo, G. Veglia, C. M. Dobson, A. De Simone, *Nat. Commun.* **2016**, *7*, 12563.
- [16] M. D. Tuttle, G. Comellas, A. J. Nieuwkoop, D. J. Covell, D. A. Berthold, K. D. Klopper, J. M. Courtney, J. K. Kim, A. M. Barclay, A. Kendall, W. Wan, G. Stubbs, C. D. Schwieters, V. M. Y. Lee, J. M. George, C. M. Rienstra, *Nat. Struct. Mol. Biol.* **2016**, *23*, 409–415.
- [17] A. D. Stephens, M. Zacharopoulou, G. S. Kaminski Schierle, *Trends Biochem. Sci.* **2019**, *44*, 453–466.
- [18] V. N. Uversky, *J. Biomol. Struct. Dyn.* **2003**, *21*, 211–234.
- [19] W. Bermel, I. Bertini, I. C. Felli, R. Peruzzini, R. Pierattelli, *ChemPhysChem* **2010**, *11*, 689–695.
- [20] I. Bertini, I. C. Felli, L. Gonnelli, M. V. Vasantha Kumar, R. Pierattelli, *Angew. Chem. Int. Ed.* **2011**, *50*, 2339–2341; *Angew. Chem.* **2011**, *123*, 2387–2389.
- [21] S. Gil, T. Hošek, Z. Solyom, R. Kümmerle, B. Brutscher, R. Pierattelli, I. C. Felli, *Angew. Chem. Int. Ed.* **2013**, *52*, 11808–11812; *Angew. Chem.* **2013**, *125*, 12024–12028.
- [22] J. Lopez, P. Ahuja, M. Gerard, J. M. Wieruszkeski, G. Lippens, *J. Magn. Reson.* **2013**, *236*, 1–6.
- [23] I. C. Felli, L. Gonnelli, R. Pierattelli, *Nat. Protoc.* **2014**, *9*, 2005–2016.
- [24] E. C. Cook, G. A. Usher, S. A. Showalter, *Methods Enzymol.* **2018**, *611*, 81–100.
- [25] A. Alik, C. Boughechtouli, M. Julien, W. Bermel, R. Ghoul, S. Zinn-Justin, F. X. Theillet, *Angew. Chem. Int. Ed.* **2020**, *59*, 10411–10415; *Angew. Chem.* **2020**, *132*, 10497–10501.
- [26] R. Llinás, M. Sugimori, R. B. Silver, *Science* **1992**, *256*, 677–679.
- [27] R. Schneggenburger, E. Neher, *Nature* **2000**, *406*, 889–893.
- [28] H. J. Lee, E. J. Bae, S. J. Lee, *Nat. Rev. Neurol.* **2014**, *10*, 92–98.
- [29] W. Bermel, I. Bertini, L. Duma, I. C. Felli, L. Emsley, R. Pierattelli, P. R. Vasos, *Angew. Chem. Int. Ed.* **2005**, *44*, 3089–3092; *Angew. Chem.* **2005**, *117*, 3149–3152.
- [30] W. Bermel, I. Bertini, I. C. Felli, M. Piccioli, R. Pierattelli, *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *48*, 25–45.
- [31] L. E. Kay, M. Ikura, R. Tschudin, A. Bax, *J. Magn. Reson.* **1990**, *89*, 496–514.
- [32] M. Sattler, C. Schleucher, J. Griesinger, *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 93–158.
- [33] A. Bax, *J. Magn. Reson.* **2011**, *213*, 442–445.
- [34] C. O. Fernández, W. Hoyer, M. Zweckstetter, E. A. Jares-Erijman, V. Subramaniam, C. Griesinger, T. M. Jovin, *EMBO J.* **2004**, *23*, 2039–2046.
- [35] M. S. Nielsen, H. Vorum, E. Lindersson, P. H. Jensen, *J. Biol. Chem.* **2001**, *276*, 22680–22684.
- [36] R. M. Rasia, C. W. Bertoncini, D. Marsh, W. Hoyer, D. Cherny, M. Zweckstetter, C. Griesinger, T. M. Jovin, C. O. Fernández, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4294–4299.
- [37] Y. H. Sung, C. Rospigliosi, D. Eliezer, *Biochim. Biophys. Acta Proteins Proteomics* **2006**, *1764*, 5–12.
- [38] A. Binolfi, R. M. Rasia, C. W. Bertoncini, M. Ceolin, M. Zweckstetter, C. Griesinger, T. M. Jovin, C. O. Fernández, *J. Am. Chem. Soc.* **2006**, *128*, 9893–9901.
- [39] J. Lautenschläger, A. D. Stephens, G. Fusco, F. Ströhl, N. Curry, M. Zacharopoulou, C. H. Michel, R. Laine, N. Nespovitaya, M. Fantham, D. Pinotsi, W. Zago, P. Fraser, A. Tandon, P. St George-Hyslop, E. Rees, J. J. Phillips, A. De Simone, C. F. Kaminski, G. S. K. Schierle, *Nat. Commun.* **2018**, *9*, 712.
- [40] A. Rcom-H'cheo-Gauthier, J. Goodwin, D. L. Pountney, *Bio-molecules* **2014**, *4*, 795–811.
- [41] M. R. Post, O. J. Lieberman, E. V. Mosharov, *Front. Neurosci.* **2018**, *12*, 161.
- [42] M. P. Williamson, *Prog. Nucl. Magn. Reson. Spectrosc.* **2013**, *73*, 1–16.
- [43] Y. Yoshimura, M. A. Holmberg, P. Kukic, C. B. Andersen, A. Mata-Cabana, S. Fabio Falsone, M. Vendruscolo, E. A. A. Nollen, F. A. A. Mulder, *J. Biol. Chem.* **2017**, *292*, 8269–8278.
- [44] A. Beier, T. C. Schwarz, D. Kurzbach, G. Platzer, F. Tribuzio, R. Konrat, *J. Mol. Biol.* **2018**, *430*, 2439–2452.
- [45] M. Schiavina, M. G. Murralli, L. Pontoriero, V. Sainati, R. Kümmerle, W. Bermel, R. Pierattelli, I. C. Felli, *Biophys. J.* **2019**, *117*, 46–55.
- [46] Y. Bai, J. S. Milne, L. Mayne, S. W. Englander, T. R. Sosnick, L. Mayne, S. W. Englander, J. S. Milne, L. Mayne, S. W. Englander, *Proteins Struct. Funct. Genet.* **1993**, *17*, 75–86.
- [47] N. R. Skrynnikov, R. R. Ernst, *J. Magn. Reson.* **1999**, *137*, 276–280.
- [48] V. Ozenne, F. Bauer, L. Salmon, J. R. Huang, M. R. Jensen, S. Segard, P. Bernadó, C. Charavay, M. Blackledge, *Bioinformatics* **2012**, *28*, 1463–1470.
- [49] R. Koradi, M. Billeter, K. Wüthrich, *J. Mol. Graph.* **1996**, *14*, 51–55.
- [50] M. K. Janowska, J. Baum in *Methods in Molecular Biology* (Ed.: D. Eliezer), Springer New York, New York, **2016**, pp. 45–53.
- [51] B. Mateos, R. Konrat, R. Pierattelli, I. C. Felli, *ChemBioChem* **2019**, *20*, 335–339.
- [52] A. Santner, V. N. Uversky, *Metallomics* **2010**, *2*, 378–392.
- [53] V. N. Uversky, J. Li, A. L. Fink, *J. Biol. Chem.* **2001**, *276*, 44284–44296.
- [54] J. Bai, Z. Zhang, M. Liu, C. Li, *BMC Biophys.* **2016**, *9*, 1.
- [55] I. Bertini, C. Luchinat, G. Parigi, E. Ravera, *NMR of Paramagnetic Molecules*, Elsevier, Amsterdam, **2017**.

Manuscript received: June 6, 2020

Revised manuscript received: July 14, 2020

Accepted manuscript online: July 31, 2020

Version of record online: September 15, 2020



Supporting Information

Monitoring the Interaction of α -Synuclein with Calcium Ions through Exclusively Heteronuclear Nuclear Magnetic Resonance Experiments

Letizia Pontoriero⁺, Marco Schiavina⁺, Maria Grazia Murrari, Roberta Pierattelli,^{} and Isabella C. Felli^{*}*

anie_202008079_sm_miscellaneous_information.pdf

Table of Contents

Experimental Procedure	3
Samples preparation.....	3
NMR Spectroscopy	3
Supplementary Figures	4
Supplementary Figure S1 (Caption)	4
Supplementary Figure S1 (Figure)	5
Supplementary Figure S2 (Caption)	6
Supplementary Figure S2 (Figure)	7
Supplementary Figure S3	8
Supplementary Figure S4	9
Supplementary Figure S5	10
Supplementary Figure S6	11
Supplementary Figure S7	12
Supplementary Figure S8	13
Supplementary Figure S9	14
Supplementary Figure S10	15
Supplementary Figure S11.....	16
Supplementary Tables	17
Supplementary Table 1	17
Supplementary Table 2	18
References	19
Author contributions	19

Experimental Procedures

Samples preparation. Three ^{13}C , ^{15}N labelled α -synuclein samples were prepared as previously described in the literature^[1], lyophilized in water and stocked at -20°C . Experiments were acquired using a first sample in a 5 mm NMR tube containing 500 μL of α -synuclein with a concentration of 600 μM in 20 mM Tris-Cl buffer, $\text{pH}=7.4$. 2% D_2O was added for the lock. For the titration experiments, a batch of CaCl_2 solution was prepared with a final concentration of 100 mM in milliQ water. Seven additions of the CaCl_2 stock were performed in order to reach the following ratios, in equivalents, of α -synuclein to calcium ions: 1:0 – 1:1 – 1:2 – 1:4 – 1:6 – 1:8 – 1:10 – 1:16.

A second α -synuclein sample was used to perform the Mn^{2+} titration. Experiments were acquired in a 5 mm NMR tube containing 500 μL of α -synuclein with a concentration of 600 μM in 20 mM Tris-Cl buffer, $\text{pH}=7.4$. 2% D_2O was added for the lock. For the titration experiments, two batches of MnCl_2 solution were prepared with a final concentration of 2.5 mM and 25 mM in milliQ water. Three additions of the MnCl_2 stocks were performed in order to reach the following ratios, in equivalents, of α -synuclein to manganese ions: 1:0 – 1:0.01 – 1:0.02 – 1:0.1.

A third more diluted α -synuclein sample was prepared to perform Ca^{2+} titration. Experiments were acquired using a sample in a 5 mm NMR tube containing 500 μL of α -synuclein with a concentration of 50 μM in 20 mM Tris-Cl buffer, $\text{pH}=7.4$. 2% D_2O was added for the lock. For the titration experiments, three batches of CaCl_2 solution were prepared with a final concentration of 10 mM, 100 mM, and 1 M in buffer. Eight additions of the CaCl_2 stocks were performed in order to reach the following ratios, in equivalents, of α -synuclein to calcium ions: 1:0 – 1:1 – 1:2 – 1:4 – 1:8 – 1:16 – 1:32 – 1:64 – 1:256.

NMR Spectroscopy. The NMR experiments were acquired at 310 K on a Bruker AVANCE NEO spectrometer operating at 700.06 MHz ^1H , 176.05 MHz ^{13}C , and 70.97 MHz ^{15}N frequencies, equipped with a cryogenically cooled probehead optimized for ^{13}C -direct detection (TXO).

Experimental parameters used for the acquisition of NMR spectra are reported in Table S1 (sample concentration, type of experiment, spectral width, acquired data points, number of scans, inter-scan delays, experimental duration). Pulse lengths and carrier frequencies generally used for triple resonance experiments were used and are summarized hereafter. The ^1H carrier was placed at 4.7 ppm for non-selective hard pulses or at 8.5 ppm for band-selective pulses on the amide proton region. ^{13}C pulses were given at 176.1 ppm, 55.6 ppm and 42.6 ppm for C' , C^α and C^{all} regions. ^{15}N pulses were given at 125.0 ppm (for CON experiments) or 118.0 ppm (for HN experiments). Q5 and Q3 shapes^[2] of durations of 300 and 231 μs , respectively, were used for ^{13}C band-selective $\pi/2$ and π flip angle pulses except for the π pulses that should be band-selective on the C^α region (Q3, 900 μs), and for the adiabatic π pulse^[3] to invert both C' and C^α (smoothed Chirp 500 μs , 20 % smoothing, 80 kHz sweep width, 11.3 kHz RF field strength). Composite pulse decoupling was applied on ^1H (Waltz-16)^[4] and ^{15}N (Garp-4)^[5] with an RF field strength of 3kHz and 1kHz respectively. ^{13}C homonuclear decoupling was achieved through the IPAP virtual decoupling approach^[6]. For DeCON experiments, both for the 2D and 3D version, a Reburp shape^[7] of duration of 2076 μs was used for ^1H band-selective π flip angle pulses. These are used to generate the triple spin order $4\text{C}_2\text{N}_2\text{H}_2$ operator without perturbing the water magnetization. All gradients employed had a smoothed square shape.

All the spectra were acquired, processed and analyzed by using Bruker TopSpin 4.0.5 software. Calibration of the spectra was achieved using DSS as a standard for ^1H and ^{13}C ; ^{15}N shifts were calibrated indirectly.

Supplementary Figures

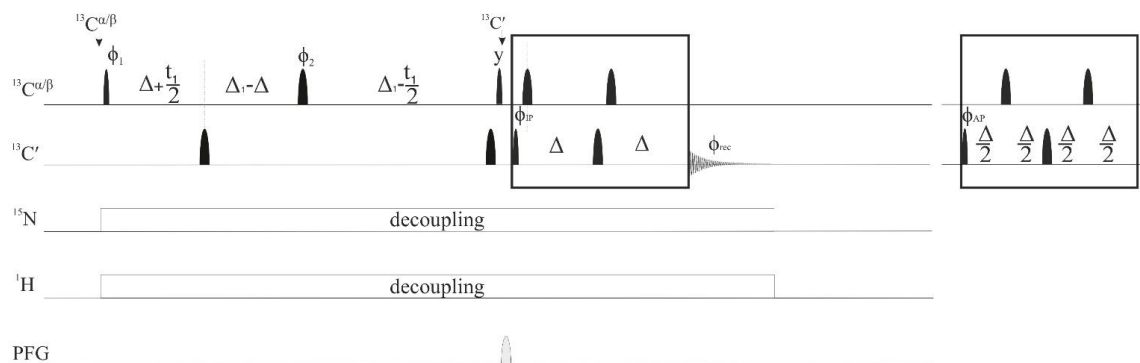
Figure S1. Pulse sequences used to acquire CACO, CBCACO, CCCO spectra. Narrow and wide black bars represent $\pi/2$ and π non-selective pulses; narrow and wide rounded black bars represent $\pi/2$ and π band-selective pulses. The pulse sequence elements reported in the boxes represent the two variants to acquire the in-phase (IP) and antiphase (AP) components of carbonyl signals needed to achieve ^{13}C homonuclear decoupling through the IPAP approach.

For CACO the following phase cycling was employed: $\phi_1 = x, -x$; $\phi_2 = 4(x), 4(y)$, $\phi^{\text{IP}} = 2(x), 2(-x)$; $\phi^{\text{AP}} = 2(-y), 2(y)$ and $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x$. The length of the delays was: $\Delta = 4.5$ ms; $\Delta_1 = 14.2$ ms. The strength of the smoothed square shape gradient was 50%. Quadrature detection in the indirect dimension was achieved through the STATES-TPPI approach incrementing phase ϕ_1 .

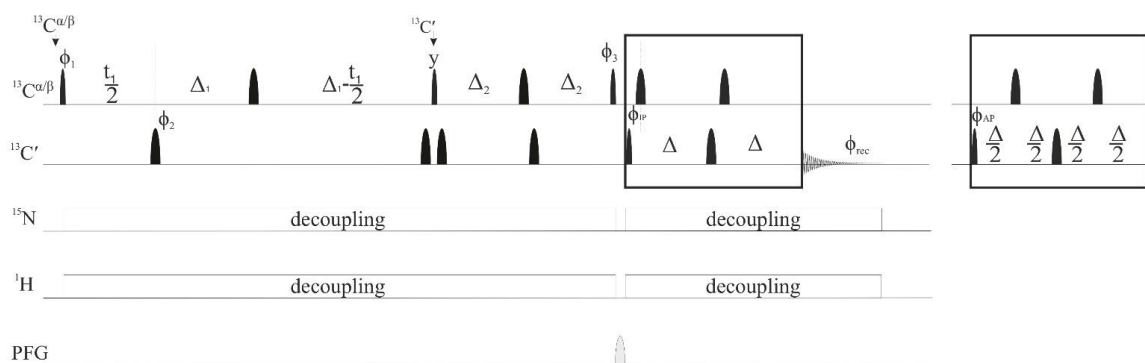
For CBCACO, the following phase cycling was employed: $\phi_1 = x, -x$; $\phi_2 = 8(x), 8(-x)$, $\phi_3 = 2(y), 2(-y)$, $\phi^{\text{IP}} = 4(x), 4(-x)$; $\phi^{\text{AP}} = 4(-y), 4(y)$ and $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x$. The length of the delays was: $\Delta = 4.5$ ms; $\Delta_1 = 11.7$ ms; $\Delta_2 = 4.0$ ms. The strength of the smoothed square shape gradient was 30%. Quadrature detection in the indirect dimension was achieved through the STATES-TPPI approach incrementing phase ϕ_1 .

For CCCO, the following phase cycling was employed: $\phi_1 = x, -x$; $\phi_2 = 2(x), 2(-x)$; $\phi^{\text{IP}} = 4(x), 4(-x)$; $\phi^{\text{AP}} = 4(-y), 4(y)$ and $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x$. The length of the delays was: $\Delta = 4.5$ ms; $\Delta_1 = 15.5$ ms. The strengths of the smoothed square shape gradients were 30%, 50% and 11%. Quadrature detection in the indirect dimension was achieved through the STATES-TPPI approach incrementing phase ϕ_1 . The ^{13}C spinlock was applied with an RF field strength of 10kHz ($1/4 \cdot p90$) with a FLOPSY sequence^[8]. The grey pulse is a band-selective pulse on the C^α region.

CACO



CBCACO



CCCO

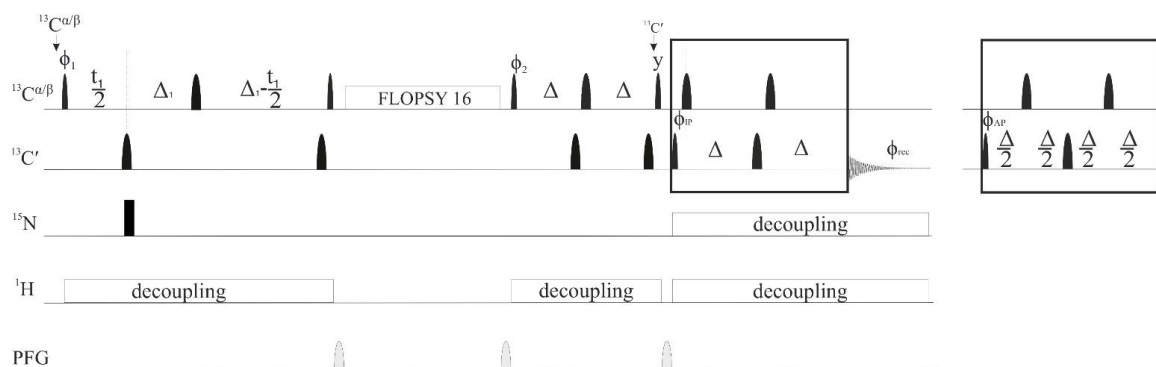


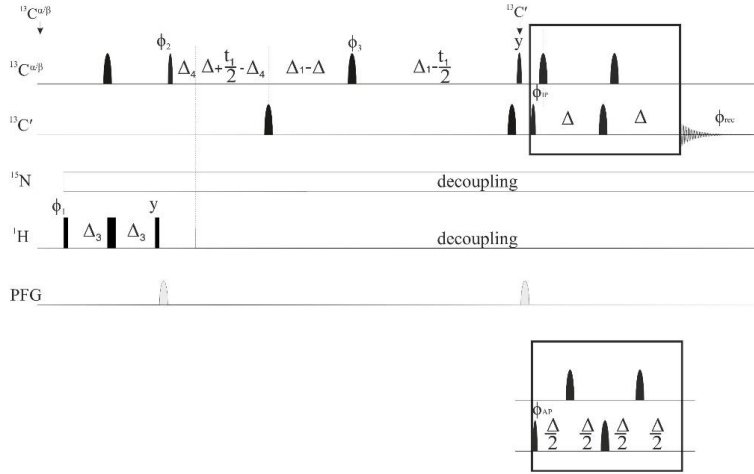
Figure S2. Pulse sequences used to acquire (H)CACO, (H)CBCACO, (H)CCCO spectra. Narrow and wide black bars represent $\pi/2$ and π non-selective pulses; narrow and wide rounded black bars represent $\pi/2$ and π band-selective pulses. The pulse sequence elements reported in the boxes represent the two variants to acquire the in-phase (IP) and antiphase (AP) components of carbonyl signals needed to achieve ^{13}C homonuclear decoupling through the IPAP approach.

For (H)CACO the following phase cycling was employed: $\phi_1 = 2(x), 2(-x)$; $\phi_2 = x, -x$; $\phi_3 = 8(x), 8(y)$; $\phi_{\text{IP}} = 4(x), 4(-x)$; $\phi_{\text{AP}} = 4(-y), 4(y)$ and $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x, -x, x, x, -x, x, -x, x, -x, x$. The length of the delays was: $\Delta = 4.5$ ms; $\Delta_1 = 14.2$ ms; $\Delta_3 = 1.8$ ms $\Delta_4 = 1.1$ ms. The strength of the smoothed square shape gradient is 30% 50%. Quadrature detection in the indirect dimension was achieved through the STATES-TPPI approach incrementing phase ϕ_2 .

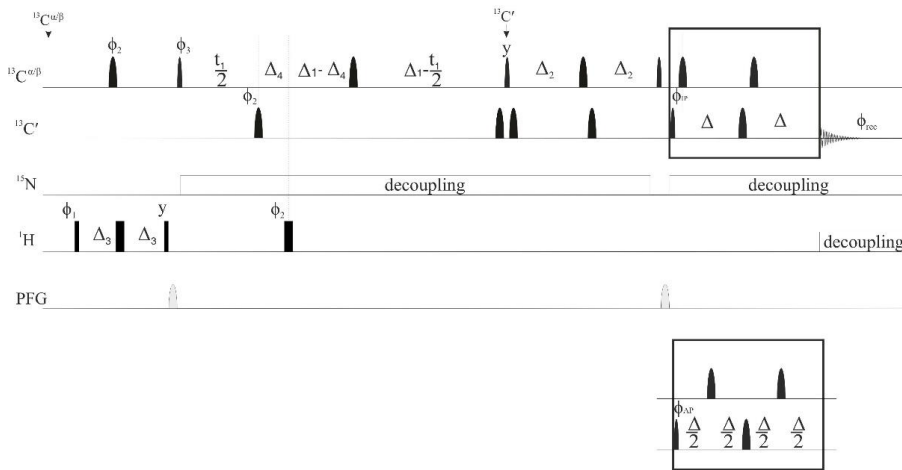
For (H)CBCACO, the following phase cycling was employed: $\phi_1 = x, -x$; $\phi_2 = 8(x), 8(-x)$, $\phi_3 = 2(x), 2(-x)$; $\phi_{\text{IP}} = 4(x), 4(-x)$; $\phi_{\text{AP}} = 4(-y), 4(y)$ and $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x$. The length of the delays was: $\Delta = 4.5$ ms; $\Delta_1 = 11.7$ ms; $\Delta_2 = 4.0$ ms $\Delta_3 = 1.8$ ms $\Delta_4 = 1.1$ ms. The strength of the smoothed square shape gradient is 30% and 50%. Quadrature detection in the indirect dimension was achieved through the STATES-TPPI approach incrementing phase ϕ_3 .

For (H)CCCO, the following phase cycling was employed: $\phi_1 = x, -x$; $\phi_2 = 2(x), 2(-x)$; $\phi_3 = 4(x), 4(-x)$; $\phi_{\text{IP}} = x$; $\phi_{\text{AP}} = -y$ and $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x$. The length of the delays was: $\Delta = 4.5$ ms; $\Delta_1 = 15.5$ ms $\Delta_3 = 1.8$ ms $\Delta_4 = 1.1$ ms. The strengths of the smoothed square shape gradients are 30%, 50% and 11%. Quadrature detection in the indirect dimension was achieved through the STATES-TPPI approach incrementing phase ϕ_2 . The ^{13}C spin lock was applied with an RF field strength of 10 kHz ($1/4^*p90$) with a FLOPSY sequence [8]. The grey pulse is a band-selective pulse on the C α region.

(H)CACO



(H)CBCACO



(H)CCCO

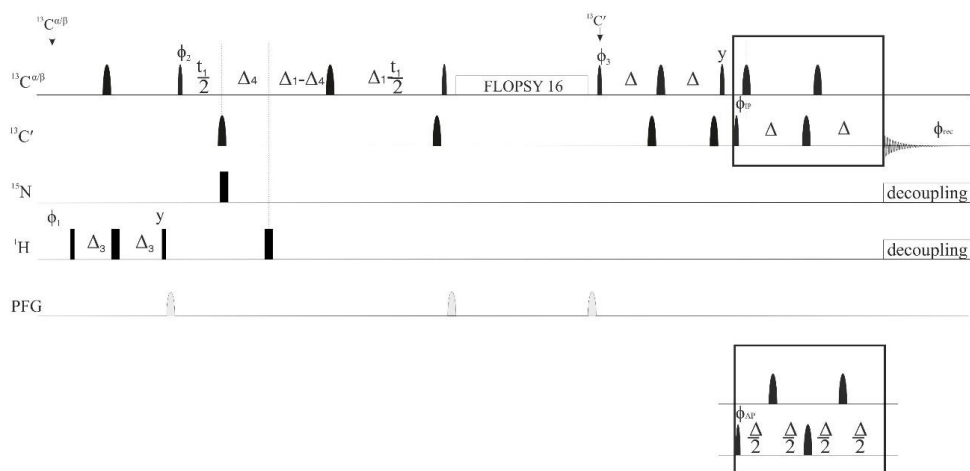


Figure S3. Comparison of chemical shift perturbations (CSP) of side chain carboxylate/carbonyl carbon chemical shifts (blue) with backbone carbonyl carbon chemical shifts (red) determined through from 2D-CACO and 2D-CON spectra ($CSP = |\Delta(\delta^{13}C)|$). Backbone CSP values are smaller in magnitude with respect to those of side chains and not necessarily maximal for Asp/Glu/Asn/Gln amino acids, reflecting a more indirect effect experienced by backbone nuclear spins upon interaction with calcium ions.

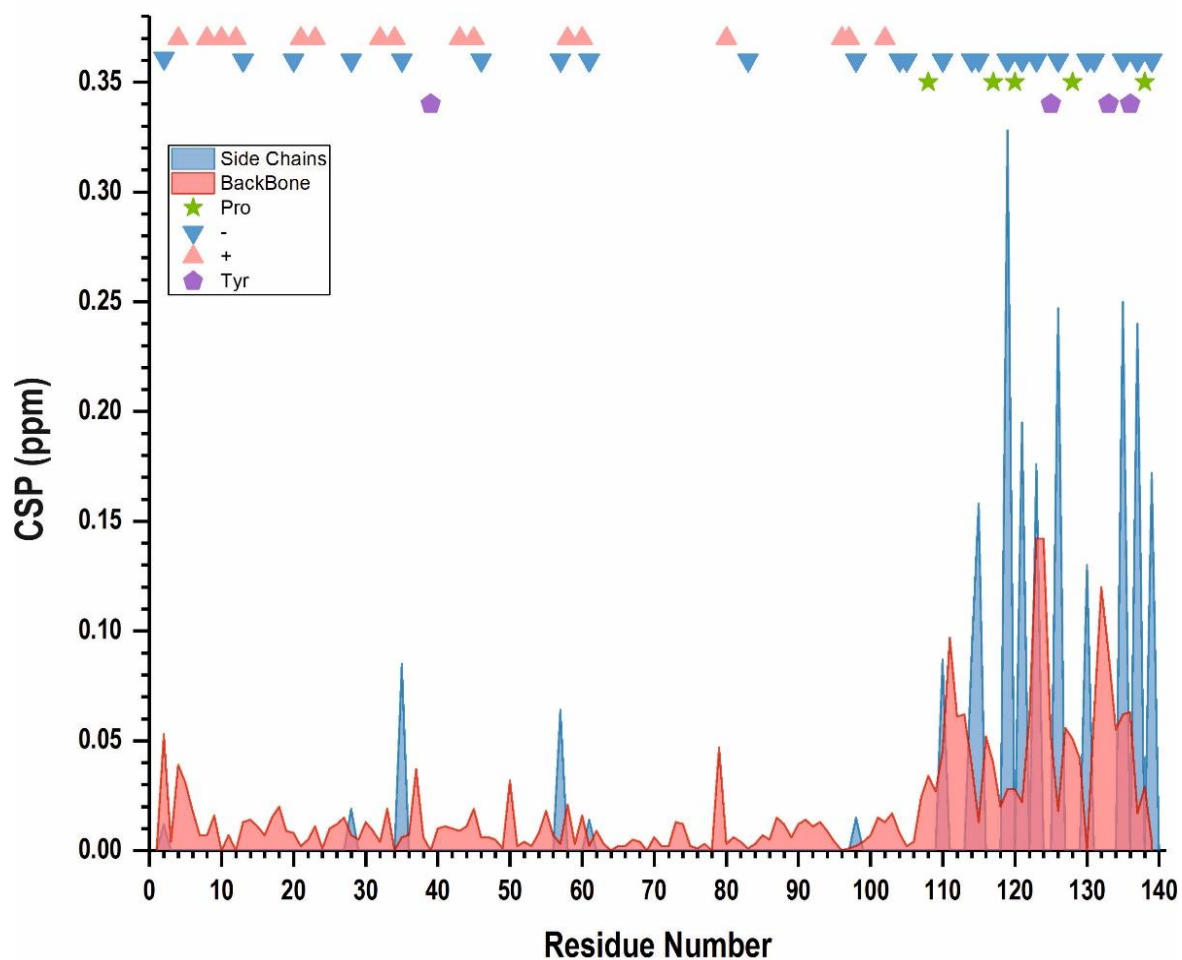


Figure S4. Secondary Structure Propensity (SSP) score obtained exploiting the Tamiola-Mulder approach^[9] (<https://st-protein02.chem.au.dk/ncIDP/>) for α -synuclein only (panel A) and α -synuclein in the presence of calcium ions (panel B). Panel C reports the differences between the SSP values obtained with and without Ca^{2+} . Chemical shifts of ^{15}N , $^{13}\text{C}^{\alpha}$, $^{13}\text{C}^{\beta}$ and $^{13}\text{C}'$ were used as input.

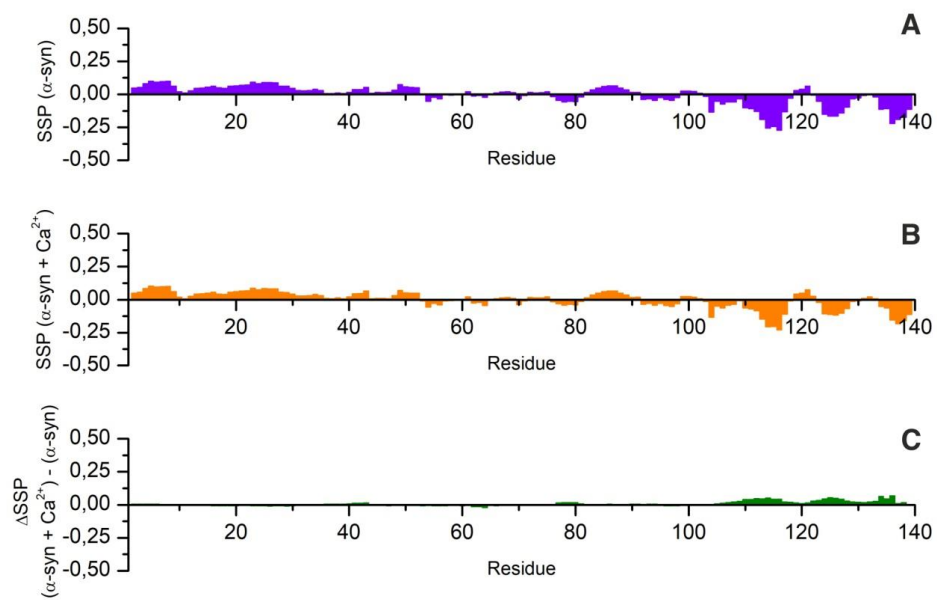


Figure S5. Superimposition of (H)CACO spectra recorded on a 50 μM α -synuclein sample in absence (black) and presence (red) of Ca^{2+} . The lower panels show two regions of the (H)CACO spectrum with cross peaks of Asp and Glu side chains and their downfield shifts upon addition of Ca^{2+} . The perturbation of Asp and Glu side chain resonances is analogous to the one observed with the more concentrated α -synuclein sample. The titration follows the pattern from black (α -syn: Ca^{2+} , 1:0) to red (α -syn: Ca^{2+} , 1:256).

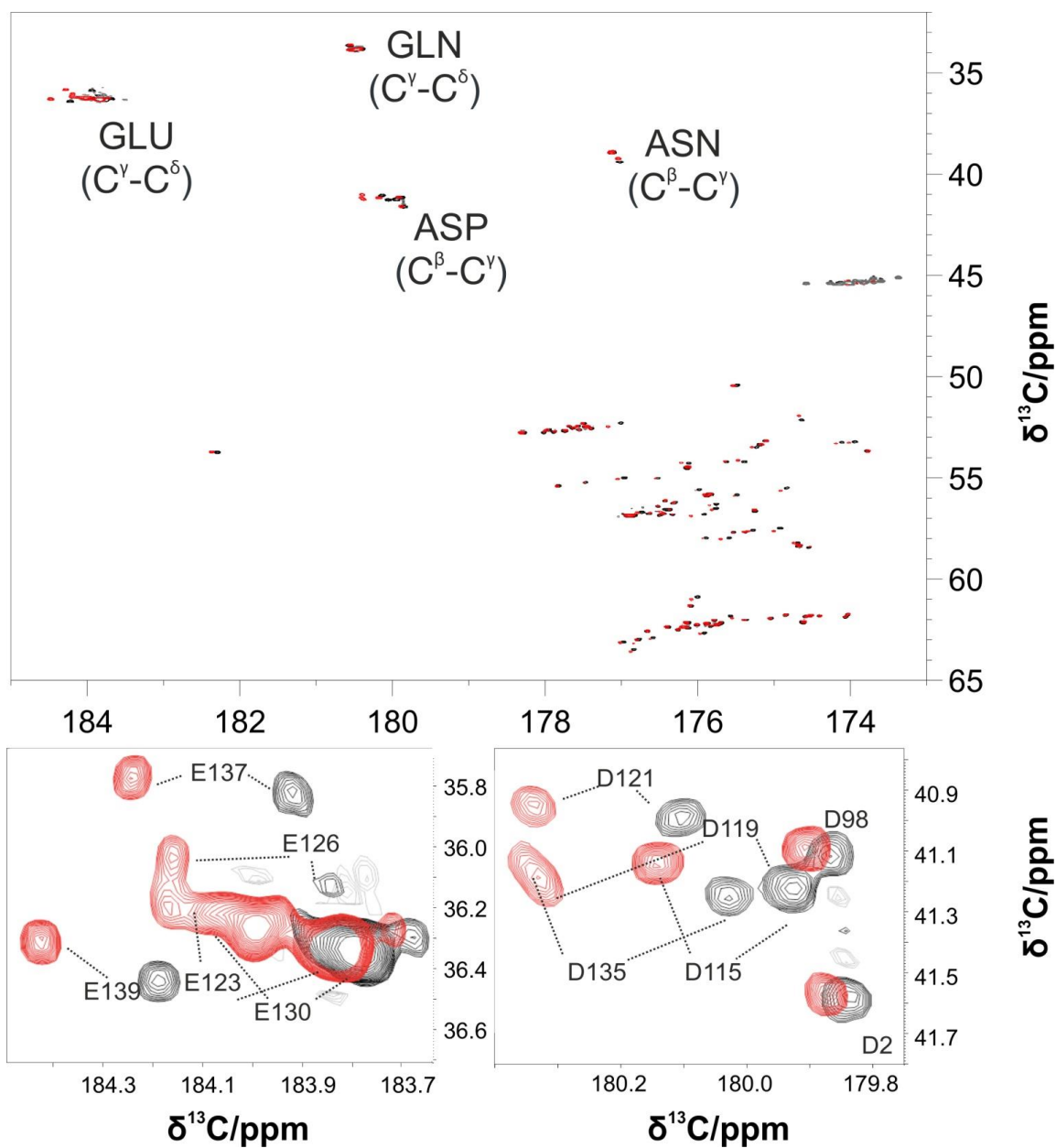


Figure S6. The figure reports in a schematic way the residues whose amide proton could be detected (0.5 indicates cross peaks close to the noise level). Backbone carbonyl carbon resonances could be detected for all peptide bonds.

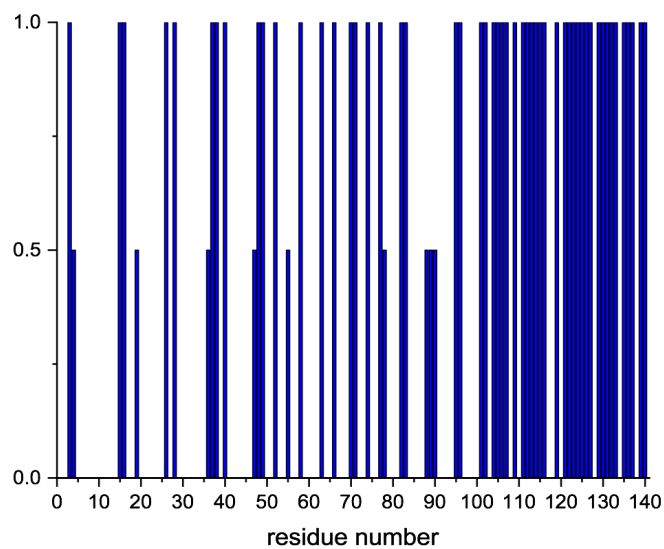


Figure S7. Comparison of DeCON and Decor performances. In both experiments the most pronounced effect leading to decorrelation of the longitudinal spin-order operators ($4C'_2N_2H_2$ and $2N_2H_2$ respectively) is the exchange of amide protons with the solvent. The agreement between the data measured through the two different experiments is good for the residues that could be detected in both experiments. In order to sample a consistent number of cross-peaks in both experiments the comparison was performed at 298 K.

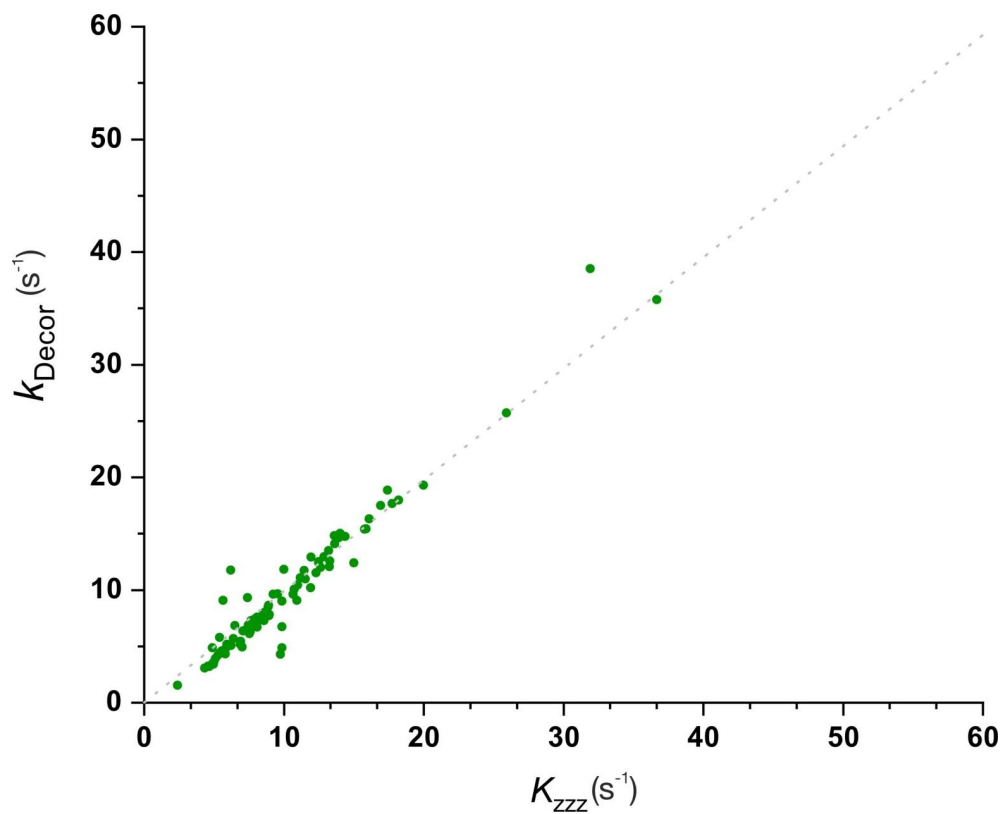


Figure S8. (A) Pulse sequence of CONzz used to determine the decay of the two spin-order operator ($2C'_zN_z$). In the CONzz experiment the following phase cycling was employed: $\phi_1 = 2(y), 2(-y)$; $\phi_2 = x, -x$; $\phi_3 = 4(x), 4(-x)$, $\phi^{IP} = x$; $\phi^{AP} = -y$ and $\phi_{rec} = x, -x, -x, x, -x, x, x, -x$. The length of the delays were: $\Delta_1 = 16.6$ ms; $\Delta = 4.5$ ms; $\varepsilon = t_1(0) + p180$ (500 μ s). The striped pulse is an adiabatic Chirp pulse to invert ^{13}C signals. Virtual decoupling of the $C'-C^{\alpha}$ coupling was achieved by acquiring for each increment both the IP and AP components of the signals. The strengths of the smoothed square shape gradient were 50% and 70%. Quadrature detection in the indirect dimension was achieved through STATES-TPPI approach incrementing phase ϕ_2 .

Intensities of cross-peaks determined in spectra acquired with the CONzz and DeCON pulse sequences as a function of τ_{decor} were fitted to a single exponential decay function. Panel (B) reports the values obtained for the decay of the two spin order ($2C'_zN_z$) measured through CONzz (k_{zz}) and for the decay of the three spin order ($4C'_zN_xH_z$) as measured through the DeCON (k_{zzz}) as a function of the residue number. Their comparison shows that the contributions of longitudinal relaxation are much smaller compared to those deriving from decorrelations due to exchange processes with the solvent. Panel (C) shows the comparison between the decay of the two-spin order measured for α -synuclein before and after addition for Ca^{2+} .

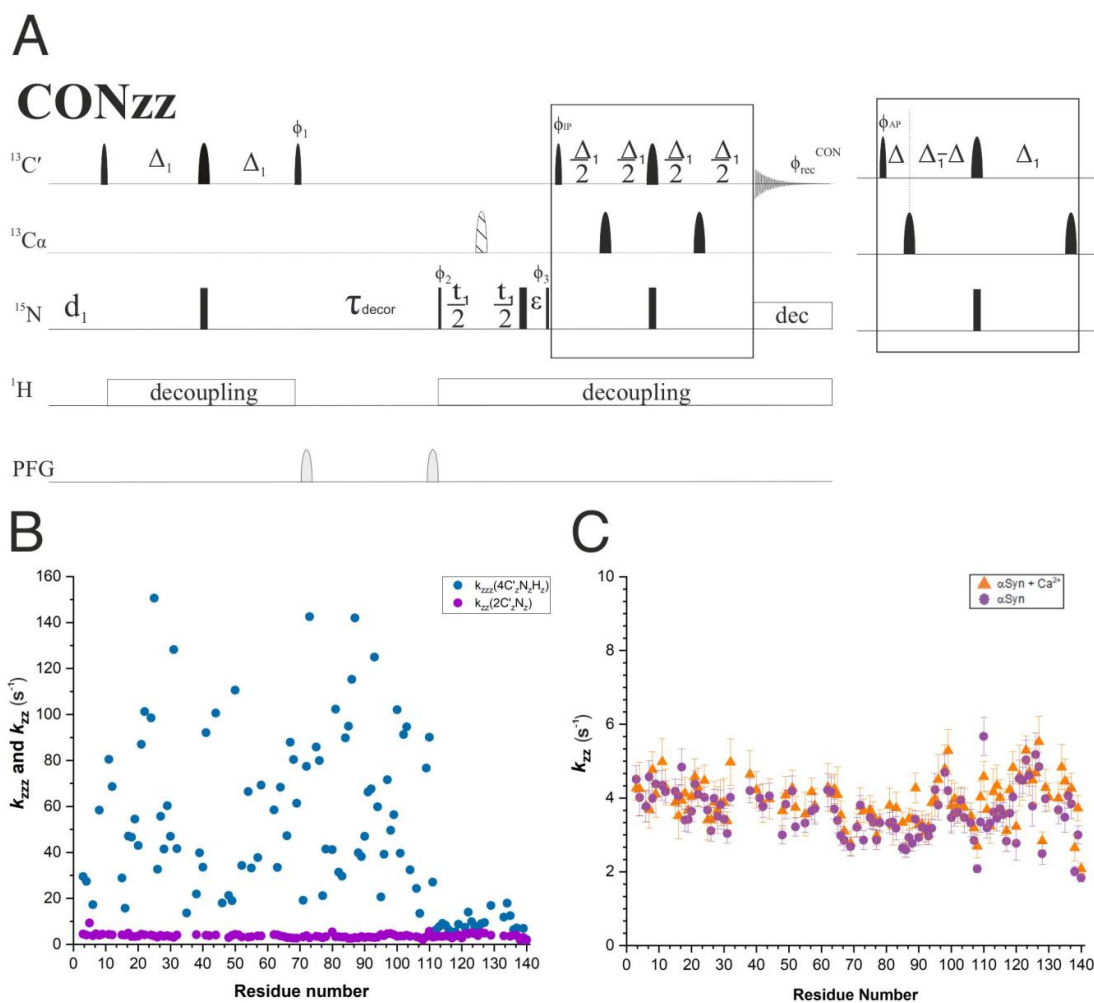


Figure S9. Pulse sequence used to acquire the 3D version of DeCON spectra. A screenshot of the acquired 3D spectrum is also shown. Narrow and wide black bars represent $\pi/2$ and π non-selective pulses; narrow and wide rounded black bars represent $\pi/2$ and π band-selective pulses. The pulse sequence elements reported in the boxes represent the two variants to acquire the in-phase (IP) and antiphase (AP) components of carbonyl signals needed to achieve ^{13}C homonuclear decoupling through the IPAP approach. The following phase cycling was employed: $\phi_1 = 8(x), 8(-x)$; $\phi_2 = x, -x$; $\phi_3 = 4(y), 4(-y)$; $\phi_4 = 2(x), 2(-x)$; $\phi_{\text{IP}} = x$ $\phi_{\text{AP}} = y$ and $\phi_{\text{rec}} = x, -x, -x, x, -x, x, -x, -x, x, x, -x, x, -x, -x, x$. The length of the delays was: $\delta = 3.6$ ms; $\delta_1 = 2.2$ ms; $\Delta = 9.0$ ms; $\Delta_1 = 33.2$ ms; $\Delta_2 = 28.4$ ms $\Delta_3 = 5.2$ ms; $\epsilon = t_1(0) + p180$ (500 μs). The strength of the smoothed square shape gradients are 30%, 50%, 19%, 19%, 80%, 25%, 25%, 70% 13%; the strength of the weak bipolar gradient is 1%. Quadrature detection in the indirect dimension was achieved through the STATES-TPPI approach incrementing phase ϕ_1 for the ^{13}C dimension and ϕ_4 for the ^{15}N dimension.

The 3D experiment was acquired on the 600 μM sample of α -synuclein using 8 scans per increment, a recovery delay of 0.9 s. 1024 points were acquired in the $^{13}\text{C}'$ direct dimension using a sweep width of 5556 Hz (31.55 ppm), 96 points were used in both the indirect dimensions using a sweep width of 1785 Hz (25.17 ppm) for the ^{15}N dimension and 2127 Hz (12.08 ppm) in the $^{13}\text{C}_\alpha$ dimension. The $^{13}\text{C}_\alpha$ dimension was folded in order to achieve a better resolution. The total duration of the experiment was 47h 29 min. Non-uniform sampling approaches can be implemented to reduce the experimental time still keeping the excellent resolution.

A series of ^{13}C - ^{15}N planes of the 3D DeCON were acquired with different τ_{decor} delays in order to monitor the decay of the signals. These experiments were performed with 8 scans per increment, a recovery delay of 0.9 s. 1024 points were acquired in the direct $^{13}\text{C}'$ dimension using a sweep width of 5556 Hz (31.55 ppm) and 64 points in the ^{15}N indirect dimension were used using a sweep width of 1785 Hz (25.17 ppm). The total duration of the experiment for the first τ_{decor} used was 20 min.

The used delays (τ_{decor}) were: 20 μs - 600 μs - 1.5 ms - 2.5 ms - 5 ms - 10 ms - 50 ms - 100 ms.

3D DeCON

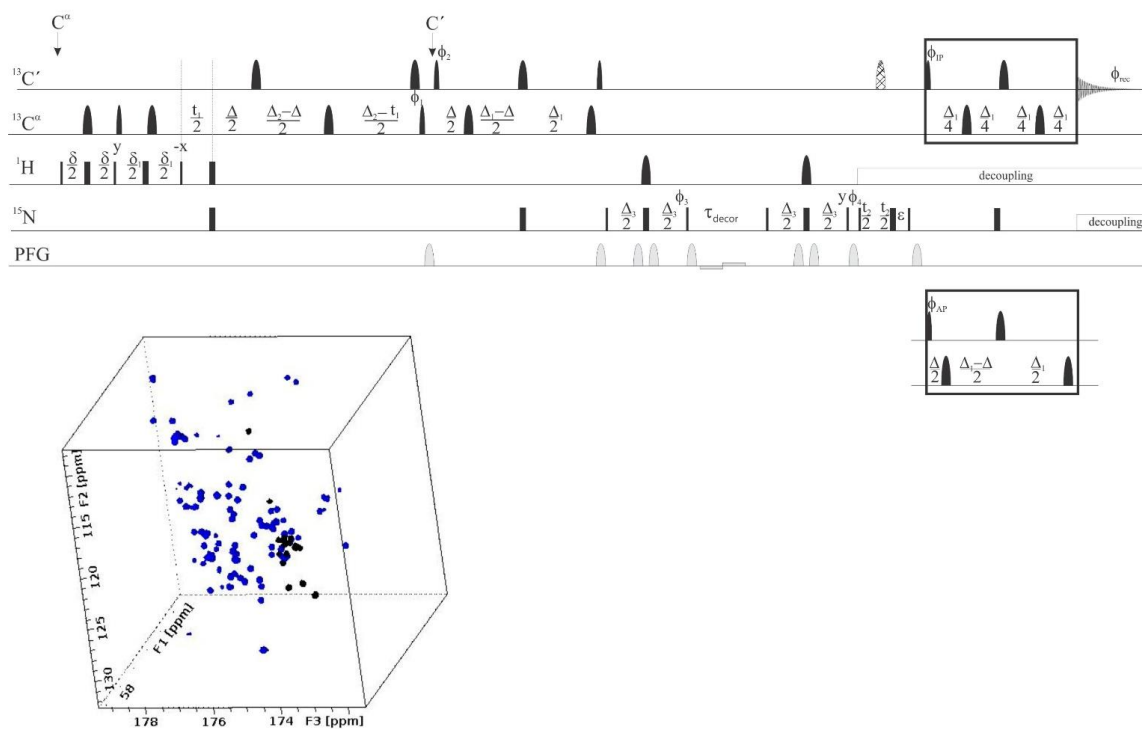


Figure S10. Values of k_{zzz} obtained from DeCON experiments on α -synuclein. The figure is color coded by residue type (upper panel). The lower panels report the data separately for four selected amino acid types (Ala, Glu, Lys, Val). The residues in the C-terminal region, however, display a different behaviour from all the others.

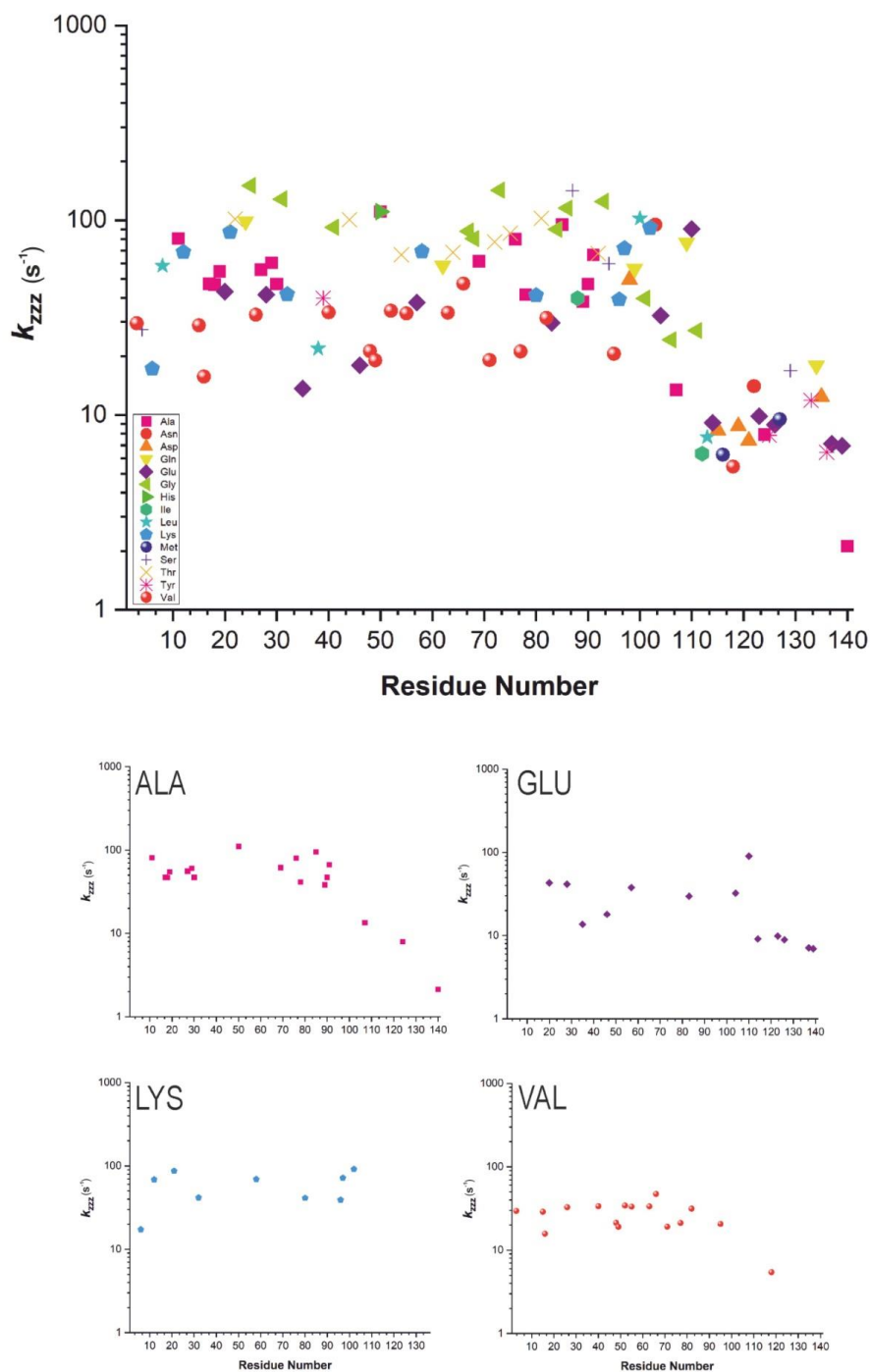
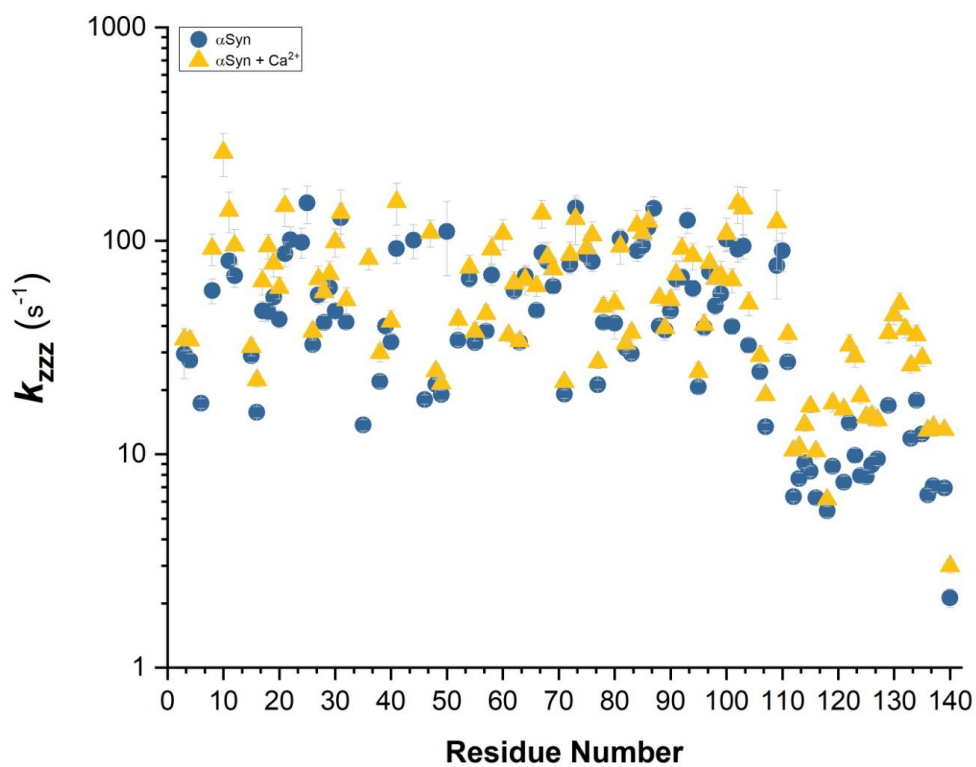


Figure S11. Values of k_{zzz} obtained from DeCON experiments on α -synuclein with (triangles) and without (circles) calcium ions.

SUPPLEMENTARY TABLES

Table S1. Experimental parameters used for the NMR experiments described in the main text.

Experiment	Dimension of acquired data (data points)		Spectral width		Number of scans	Inter scan delay (s)	Experimental duration
	F1	F2	F1	F2			
<i>α</i> -synuclein sample 600 μM							
2D-CON	800 (¹⁵ N)	1024 (¹³ C)	2273 Hz (32.03 ppm)	5556 Hz (31.55 ppm)	2	2	2 hours 2 mins 6 s
2D-CACO	330 (¹³ C)	1024 (¹³ C)	5988 Hz (34.01 ppm)	5263 Hz (34.01 ppm)	2	2	48 mins 4 s
2D-CBCACO	476 (¹³ C)	1024 (¹³ C)	10417 Hz (59.16 ppm)	5263 Hz (34.01 ppm)	2	2	1 hour 9 mins 16 s
2D-CCCO	640 (¹³ C)	1024 (¹³ C)	10417 Hz (59.16 ppm)	5263 Hz (34.01 ppm)	2	2	1 hour 24 min 9 s
2D-DeCON	200 (¹⁵ N)	1024 (¹³ C)	1786 Hz (25.17 ppm)	5555 Hz (31.55 ppm)	4	4	2 hours 5 min 11 s
2D-CONzz	200 (¹⁵ N)	1024 (¹³ C)	1786 Hz (25.17 ppm)	5555 Hz (31.55 ppm)	4	4	2 hours 5 min 11 s
2D-Dècor	192 (¹⁵ N)	2048 (¹ H)	1852 Hz (26.10 ppm)	1364 Hz (16.23 ppm)	4	3.4	46 min 14 s
<i>α</i> -synuclein sample 50 μM							
2D-CON	200 (¹⁵ N)	1024 (¹³ C)	2273 Hz (32.03 ppm)	5556 Hz (31.55 ppm)	64	1.5	12 hours 11 mins 35 s
2D-(H)CACO	330 (¹³ C)	1024 (¹³ C)	5988 Hz (34.01 ppm)	5263 Hz (34.01 ppm)	16	0.9	3 hours 9 mins 4 s
2D- (H)CBCACO	476 (¹³ C)	1024 (¹³ C)	10417 Hz (59.16 ppm)	5263 Hz (34.01 ppm)	16	0.9	4 hours 33 mins 37 s

Table S2. ¹³C chemical shifts of Asp, Asn, Glu and Gln residues of α -synuclein in 20 mM TRIS (tris-hydroxymethyl-aminomethane) buffer, 310 K, pH 7.4.

Type	Number	C'	C α	C β	C γ	C δ
ASP	2	176.09	54.26	41.61	179.83	
GLU	13	176.91	56.85	30.29	36.19	183.81
GLU	20	176.84	56.80	30.27	36.25	183.79
GLN	24	176.43	56.11	29.51	33.80	180.41
GLU	28	176.49	56.86	31.45	36.29	183.86
GLU	35	176.86	56.88	30.22	36.26	183.80
GLU	46	176.84	56.49	30.32	36.24	183.81
GLU	57	176.62	56.73	30.40	36.26	183.75
GLU	61	176.33	56.80	30.21	36.26	183.67
GLN	62	175.86	55.78	29.51	33.81	180.42
ASN	65	175.10	53.15	38.95	177.12	
GLN	79	175.90	55.86	29.38	33.82	180.48
GLU	83	176.93	56.80	30.35	36.20	183.85
ASP	98	176.15	54.46	41.10	179.87	
GLN	99	175.81	55.75	29.61	33.88	180.54
ASN	103	175.18	53.36	38.85	177.12	
GLU	104	176.37	56.53	30.43	36.24	183.90
GLU	105	176.92	56.73	30.33	36.29	183.88
GLN	109	175.88	55.84	29.61	33.78	180.44
GLU	110	176.73	56.65	30.53	36.27	183.82
GLU	114	175.76	56.48	30.64	36.30	183.80
ASP	115	175.62	54.20	41.24	179.94	
ASP	119	174.63	52.13	41.19	179.94	
ASP	121	176.13	54.51	41.00	180.14	
ASN	122	175.22	53.50	39.36	177.02	
GLU	123	175.91	56.79	30.19	36.21	183.77
GLU	126	175.48	55.82	30.75	36.05	183.85
GLU	130	176.44	56.75	30.28	36.28	183.87
GLU	131	176.82	56.85	30.34	36.28	183.78
GLN	134	174.83	55.48	29.83	33.61	180.56
ASP	135	175.39	54.21	41.25	180.06	
GLU	137	173.78	53.67	30.28	35.82	183.94
GLU	139	175.25	56.62	30.36	36.30	184.40

References

- [1] C. Huang, G. Ren, H. Zhou, C. Wang, *Protein Expr. Purif.* **2005**, *42*, 173–177.
- [2] L. Emsley, G. Bodenhausen, *Chem. Phys. Lett.* **1990**, *165*, 469–476.
- [3] J. M. Böhlen, G. Bodenhausen, *J. Magn. Reson. - Ser. A* **1993**, *102*, 293–301.
- [4] A. J. Shaka, J. Keeler, R. Freeman, *J. Magn. Reson.* **1983**, *53*, 313–340.
- [5] R. Shaka, A. J.; Barker, P. B.; Freeman, *J. Magn. Reson.* **1985**, *64*, 552.
- [6] I. C. Felli, R. Pierattelli, *Prog. Nucl. Magn. Reson. Spectrosc.* **2015**, *84–85*, 1–13.
- [7] H. Geen, R. Freeman, *J. Magn. Reson.* **1991**, *93*, 93–141.
- [8] M. Kadkhodaie, O. Rivas, M. Tan, A. Mohebbi, A. J. Shaka, *J. Magn. Reson.* **1991**, *91*, 437–443.
- [9] K. Tamiola, F. A. A. Mulder, *Biochem. Soc. Trans.* **2012**, *40*, 1014–1020.

Author Contributions

ICF and RP conceived the project. ICF and RP designed the experiments. MGM and LP produced the samples. LP and MS acquired and analyzed the data together with ICF and RP. ICF and RP wrote the manuscript with contribution from all the other authors.

10. Ensemble description of the intrinsically disordered N-terminal domain of the Nipah virus P/V protein from combined NMR and SAXS

scientific reports



OPEN Ensemble description of the intrinsically disordered N-terminal domain of the Nipah virus P/V protein from combined NMR and SAXS

Marco Schiavina^{1,2,5}, Edoardo Salladini^{4,5}, Maria Grazia Murrall^{1,2,5}, Giancarlo Tria^{2,3}, Isabella C. Felli^{1,2,5}, Roberta Pierattelli^{1,2,5} & Sonia Longhi^{4,5}

Using SAXS and NMR spectroscopy, we herein provide a high-resolution description of the intrinsically disordered N-terminal domain (PNT, aa 1–406) shared by the Nipah virus (NiV) phosphoprotein (P) and V protein, two key players in viral genome replication and in evasion of the host innate immune response, respectively. The use of multidimensional NMR spectroscopy allowed us to assign as much as 91% of the residues of this intrinsically disordered domain whose size constitutes a technical challenge for NMR studies. Chemical shifts and nuclear relaxation measurements provide the picture of a highly flexible protein. The combination of SAXS and NMR information enabled the description of the conformational ensemble of the protein in solution. The present results, beyond providing an overall description of the conformational behavior of this intrinsically disordered region, also constitute an asset for obtaining atomistic information in future interaction studies with viral and/or cellular partners. The present study can thus be regarded as the starting point towards the design of inhibitors that by targeting crucial protein–protein interactions involving PNT might be instrumental to combat this deadly virus.



OPEN

Ensemble description of the intrinsically disordered N-terminal domain of the Nipah virus P/V protein from combined NMR and SAXS

Marco Schiavina^{1,2,5}, Edoardo Salladini^{4,5}, Maria Grazia Murralli^{1,2,5}, Giancarlo Tria^{2,3}, Isabella C. Felli^{1,2}✉, Roberta Pierattelli^{1,2}✉ & Sonia Longhi⁴✉

Using SAXS and NMR spectroscopy, we herein provide a high-resolution description of the intrinsically disordered N-terminal domain (PNT, aa 1–406) shared by the Nipah virus (NiV) phosphoprotein (P) and V protein, two key players in viral genome replication and in evasion of the host innate immune response, respectively. The use of multidimensional NMR spectroscopy allowed us to assign as much as 91% of the residues of this intrinsically disordered domain whose size constitutes a technical challenge for NMR studies. Chemical shifts and nuclear relaxation measurements provide the picture of a highly flexible protein. The combination of SAXS and NMR information enabled the description of the conformational ensemble of the protein in solution. The present results, beyond providing an overall description of the conformational behavior of this intrinsically disordered region, also constitute an asset for obtaining atomistic information in future interaction studies with viral and/or cellular partners. The present study can thus be regarded as the starting point towards the design of inhibitors that by targeting crucial protein–protein interactions involving PNT might be instrumental to combat this deadly virus.

The Nipah virus (NiV), together with its close relative Hendra virus (HeV), is a zoonotic paramyxovirus responsible for severe encephalitis in humans. The NiV and HeV have been classified in the *Henipavirus* genus¹ that also comprises the later on discovered Cedar virus (CedV)². Because of their high pathogenic power, broad host range, high interspecies transmission and lack of therapeutics and vaccines, henipaviruses are classified as biosecurity level 4 (BSL-4) pathogens and are considered as potential bio-terrorism agents.

The genome of henipaviruses is made of a non-segmented, single-stranded RNA molecule of negative polarity that is encapsidated by a regular array of nucleoprotein (N) monomers to form a helical nucleocapsid. This N:RNA complex, and not naked RNA, is the substrate used by the viral polymerase for both transcription and replication. The viral polymerase is a complex consisting of the large (L) protein, which bears all the enzymatic activities, and the phosphoprotein (P). Through its interaction with both L and the nucleocapsid, the P protein acts as a tether and recruits L onto the N:RNA template. In addition, P also serves as a chaperone for both L³ and N in that it is required for proper folding/maturation of L and maintains N in a monomeric, RNA-free form⁴. Therefore P is a pivotal protein endowed with multiple functions critical for both transcription and replication.

The repertoire of P functions is further expanded by the peculiar coding capacity of the P gene. Indeed, beyond the P protein, the P gene also codes for the V and W proteins that are generated through the addition of either one (protein V) or two (protein W) non-templated guanosines at the editing site of the P messenger. The addition of these guanosines triggers a downstream frame-shift. The P, V, and W proteins therefore share a

¹Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy. ²Department of Chemistry "Ugo Schiff", University of Florence, Via della Lastruccia 3-13, 50019 Sesto Fiorentino, Italy. ³Florence Center for Electron Nanoscopy (FloCEN), University of Florence, Via della Lastruccia 3-13, 50019 Sesto Fiorentino, Italy. ⁴Lab. Architecture et Fonction des Macromolécules Biologiques (AFMB), UMR 7257, Aix-Marseille University and CNRS, 163 Avenue de Luminy, Case 932, Marseille, France. ⁵These authors contributed equally: Marco Schiavina, Edoardo Salladini and Maria Grazia Murralli. ✉email: felli@cerm.unifi.it; roberta.pierattelli@unifi.it; sonia.longhi@univ-amu.fr

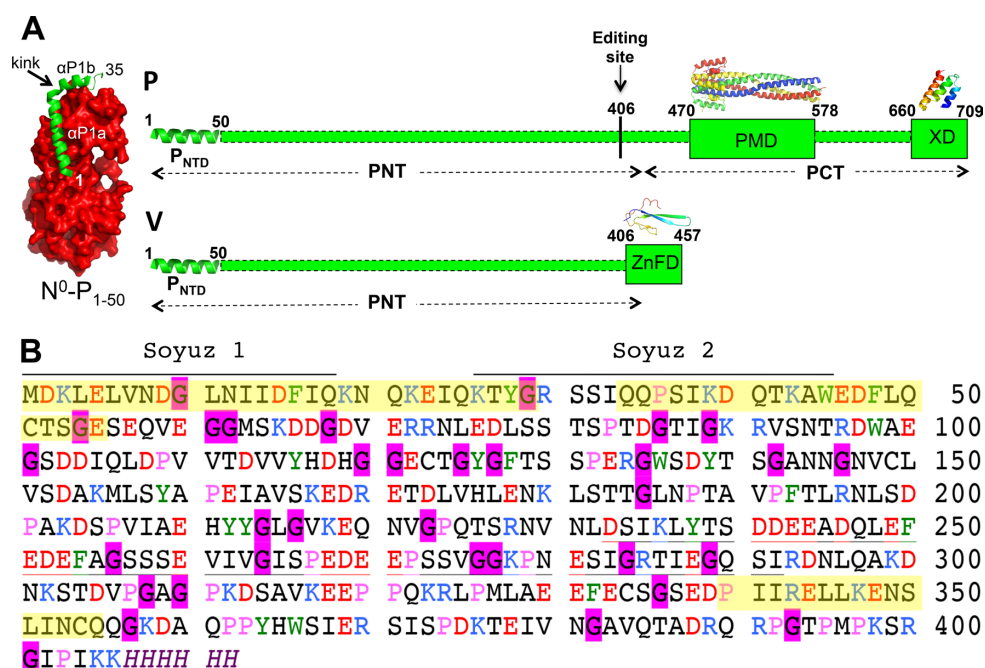


Figure 1. Modular organization of P and V, and amino acid sequence of NiV PNT. (A) Domain organization of P and V showing that P consists of two moieties, PNT and PCT, while V consists of PNT and of a zinc-finger domain (ZnFD). The P editing site is shown. Structured and disordered regions are represented as large and narrow boxes respectively. PNT: N-terminal region of P; PCT: C-terminal region of P. PMD: P multimerization domain; XD: X domain consisting of a triple α -helical bundle. The α -MoRE at the N-terminal region of P (P_{NTD}), which is partly preconfigured in solution and shown to adopt a stable α -helical conformation upon binding to the monomeric form of N^4 , is shown as a green helix. The crystal structures of NiV PMD (PDB code 4N5B)⁸ and of the NiV N^0 - P_{NTD} complex (PDB code 4CO6)⁴ are shown. The homology-derived models of NiV XD⁹ and of the ZnFD of V are also shown¹⁰. All structures were drawn using Pymol 2.0.1 (<https://pymol.org/2/>)¹¹. (B) Amino acid sequence of NiV PNT (Uniprot code Q91K91). Basic and acidic residues are shown in blue and red, respectively. Aromatic residues are shown in green, prolines in pink and glycines in black on a pink background. The soyuz1 and soyuz2 motifs conserved in *Paramyxoviridae* members are shown. The low sequence complexity region (as obtained using SEG¹² with a window size of 25 residues, trigger segment complexity of 3.0 and extension segment complexity of 3.3) is underlined. Fluctuating α -helices, as observed in NMR studies of NiV P sub-constructs¹³ are shown in yellow.

common N-terminal region (referred to as PNT) that constitutes a *bona fide* domain (i.e. a genuine functional unit) as inferred from the genetic organization of the P gene (Fig. 1A). The *Henipavirus* V and W proteins are key players in the evasion of the interferon (IFN)-mediated response via an antagonist activity of IFN signaling^{5,6}. V and W bind to STAT1, a key signal transducer in the IFN-mediated antiviral response, through their common PNT region⁷. Binding of STAT1 by V leads to inhibition of STAT1 translocation into the nucleus, whereas binding to W leads to sequestration of STAT1 in the nucleus⁷. The NiV P protein is endowed with anti-IFN function as well, indicating that the PNT domain common to P, V, and W is responsible for the IFN antagonist activity.

We previously showed that PNT from both NiV and HeV is intrinsically disordered¹⁴. The P C-terminal region (PCT) consists of an alternation of disordered regions and ordered regions (Fig. 1A). In fact, PCT comprises a disordered region (aa 407–469), referred to as « spacer » and overlapping with the reading frame of the C-terminal domain of V and W, a structured region responsible for the oligomerization of P (PMD, aa 470–578)^{8,15}, a disordered linker and a structured region (X domain, XD, aa 660–709) with a triple α -helical bundle fold (Fig. 1A)⁹. We characterized the NiV and HeV V proteins and showed that PNT conserves its disorder also within the V protein, while the zinc-finger domain (ZnFD) has a predominant β conformation¹⁰ in agreement with predictions^{16,17}. In that study, we also experimentally showed that the NiV and HeV V proteins interact with DDB1—a cellular protein whose binding to V promotes STAT1 degradation—and concluded that the ZnFD plays a crucial role in strengthening this interaction¹⁰.

Although our previous study on the V proteins from both NiV and HeV¹⁰ has contributed to illuminating the conformational behavior of these proteins and has provided a conceptual asset to design new antiviral strategies to combat the ability of these viruses to escape the innate immune response, a site-resolved description of these

proteins is still lacking. A major hurdle in this respect arises from the presence of their intrinsically disordered PNT region, which prevents crystallization and whose size constitutes a challenge for NMR studies.

To fill this gap in knowledge, and as a first step towards an atomic description of the NiV V protein, we decided to investigate NiV PNT by combining small-angle X-ray scattering (SAXS) and NMR spectroscopy. When we started this project, high-resolution data were already available for the first 35 residues of NiV PNT. Yabukarski and co-workers indeed solved the X-ray structure of a complex (herein referred to as N⁰-P₁₋₅₀) made of a truncated form of NiV N unable to self-assemble (aa 32–383) and of the first 50 residues of NiV P⁴ (Fig. 1A). Only the first 35 residues of P were defined in the electron density. This Molecular Recognition Element (MoRE) forms two α -helices (α P1a, residues 1–19; α P1b; aa 21–28) separated by a kink (Fig. 1A). In that study, the P region encompassing the first 100 residues (P₁₋₁₀₀) was also investigated by NMR spectroscopy on a small construct. Although the HN correlation NMR spectrum is typical of a disordered protein, the secondary-structure propensities (SSPs) calculated from Ca and C β secondary chemical shifts indicate the presence of five fluctuating α -helices. Upon addition of the N protein, only residues 50–100 remain detectable indicating that they remain flexible in the complex and that the N⁰-binding region of P encompasses residues 1 to 50⁴. This latter region contains two conserved motifs in *Paramyxoviridae* members (i.e. *soyuz1* ad *soyuz2*)¹⁸ (see Fig. 1B).

When we were about to complete our NMR and SAXS characterization of NiV PNT, a study was published reporting the structural description of the NiV P protein¹³. In that study, the authors combined NMR spectroscopy, SAXS, and X-ray crystallography and obtained an ensemble description of this large protein. Notably, they decided to use a “*divide et impera*” approach to disentangle the NMR spectral complexity of the large disordered N-terminal region (residues 1–474). They thus generated several overlapping constructs that altogether cover the entire P protein and then assigned the HSQC spectrum of each construct corresponding to the P disordered regions and to XD. The chemical shift-derived SSPs of the fragments revealed that the long N-terminal intrinsically disordered region and the linker connecting PMD to XD dynamically sample multiple conformations while possessing short regions of residual secondary structure. The comparison between the HSQC spectrum of the full-length NiV P protein with the HSQC spectra of the individual P sub-constructs unveiled an overall good superimposition supporting a scenario where the N-terminal moiety remains highly flexible and retains a conformational behavior in the tetramer similar to that observed in the isolated sub-constructs, with negligible inter-chain contacts¹³.

Herein, we report the results we obtained using cutting-edge, multidimensional NMR spectroscopy approaches applied to the entire PNT domain. In spite of the challenging nature of such an assignment, we succeeded in assigning as much as 91% of the peaks of this long intrinsically disordered domain. The combination of SAXS and NMR information allowed us to obtain an ensemble description of the conformational behavior of the protein in solution.

Results and discussion

NMR residue-specific structural and dynamic characterization of NiV PNT. NiV PNT is a very large intrinsically disordered protein (412 residues including the C-terminal His tag and the initial methionine), as can be inferred by inspection of the two 2D spectra, the ¹H–¹⁵N HSQC (Fig. 2A) and the ¹³C’–¹⁵N CON (Fig. 2B). The latter reveals also signals of proline residues and is characterized by improved chemical shift dispersion, two features that are very useful for the study of intrinsically disordered proteins (IDPs). An assignment strategy that combines ¹³C detected with ¹H detected NMR is thus very helpful to provide sufficient information to enable sequence-specific assignment of the resonances of the full-length protein. A series of triple-resonance 3/5D NMR experiments, specifically designed for IDPs based either on ¹H^N or ¹³C detection, were acquired^{19,20} (Supplementary Table S1). The projection reconstruction spectroscopy (APSY) approach was exploited in most of the multidimensional experiments to reduce the experimental time while preserving high spectral resolution in the indirect dimensions^{21,22}.

The ¹³C’–¹⁵N CON spectrum of NiV PNT, shown in Fig. 2B, can be considered the reference spectrum for a series of multidimensional CON-based experiments in which correlations to additional nuclear spins in the indirect dimensions of the experiments provide the needed information to achieve the sequence-specific assignment. These multidimensional spectra include the ¹³C-detected 3D (H)CBCACON²³, (H)CBCANCO²⁴, (H)COCON^{25,26}, and 5D (HCA)CONCACON²⁷ spectra. Inspection of the spectra constituted the starting point for the sequence-specific assignment^{22,27}.

This information was complemented with that available through ¹H^N detected multidimensional NMR experiments. The 3D BEST-TROSY (BT) triple-resonance experiments²⁸ HN(CO)CACB, HNCACB, HNCANNH, HNCO, HN(CO)CACB, HNCACB, and HNCOCANNH were acquired to obtain the complementary information needed to complete the assignment, augmented by a 5D BT-HN(COCAN)CONH experiment²⁹ used to resolve a few ambiguities and to confirm the chemical shift assignment obtained with the 3D spectra.

We could successfully assign the resonances of H^N, N, C’, C α , and C β nuclei of as many as 369 residues, including all proline residues (Fig. 2B), thus leading to a percentage of 91% of assigned residues for the entire protein. The chemical shifts of the previously mentioned nuclei are reported in Supplementary Table S2. The assignment is less complete in the N- and C-terminal moieties of the protein and in particular in the stretches encompassing residues 14–27, 41–49 and 342–346 either because of increased transverse relaxation or of residual signal overlap or both. The availability of heteronuclear chemical shifts enables the presence of residual secondary structure to be inferred³⁰. The neighbor-corrected structural propensities (ncSSPs) were computed by comparing the experimentally measured chemical shifts of N, C’, C α , and C β nuclei to the values expected for corresponding random coils³¹ using the tool available online at <https://st-protein02.chem.au.dk/ncSPC>³². The propensities to adopt α -helical and β -extended conformations, as obtained using the method of Mulder³², are shown in Fig. 3A

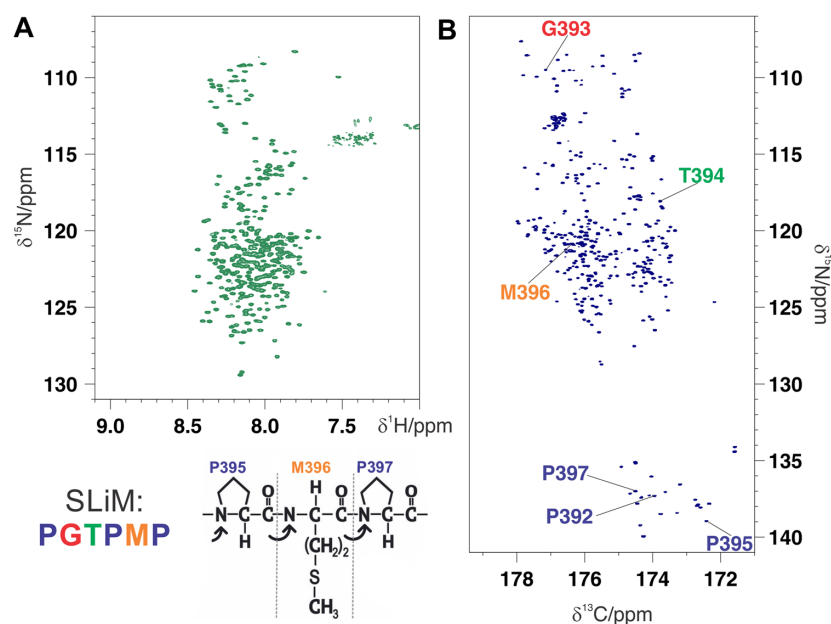


Figure 2. (A) 2D- $^1\text{H}/^{15}\text{N}$ -BEST-TROSY and (B) ^{13}C -detected CON spectra acquired on the full-length NiV PNT. The 2D CON spectrum is much more resolved and enables also the direct detection of the correlations involving the 26 prolines in a clean region of the spectrum. As an example of the importance of those residues, the cross peaks of the residues in a SLiM are reported. The complete assignment, including the ^{13}C nuclei, is reported in Supplementary Table S2.

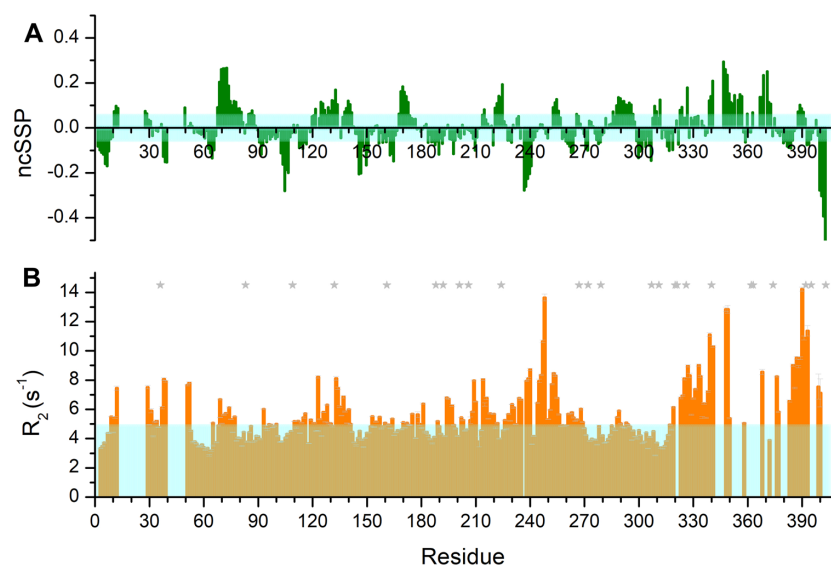


Figure 3. (A) Neighbor-corrected secondary structure propensity (ncSSP) values obtained through the online ncSSP tool (<https://st-protein02.chem.au.dk/ncSPC/>)³² using the chemical shift values reported in Supplementary Table S2. Positive and negative values correspond to α -helical and β -strand propensities, respectively. The light cyan box represents the ± 0.05 value. Few regions display a moderate α -helix propensity (SSP > 0.2) while most of the protein is largely disordered. (B) ^{15}N R_2 values reported as a function of residue number. The light cyan box indicates the modal value.

and confirm that the protein is intrinsically disordered, with SSP values in the ± 0.2 range for the majority of the primary sequence except for the regions encompassing residues 69–73, 237–240, 341–349, and 400–403.

Segments with SSP values between 0.05 and 0.2 indicate very modest deviations from random coil behavior. Moreover, the detection of local structural propensity in a polypeptide crucially depends on the choice of the reference chemical shifts for the random coil state and on possible bias induced by experimental conditions (e.g. buffer, pH, experimental temperature, etc.). Additional information is thus required to characterize the structural and dynamic properties of the protein in addition to nuclear chemical shifts.

To investigate backbone dynamics of NiV PNT, we measured heteronuclear ^1H - ^{15}N NOE, ^{15}N R_2 , and ^{15}N R_1 values. We could successfully determine relaxation rates for most of the assigned cross-peaks observed in the 2D ^1H - ^{15}N HSQC spectrum and discarded those whose overlap would have rendered the analysis ambiguous. The relaxation rates show an overall trend consistent with the considerably disordered conformation of PNT. ^1H - ^{15}N NOE values are all well below 0.5, indicating high flexibility of the backbone (Supplementary Figure S1A). Contrary to the ^{15}N R_1 values that are rather uniform along the polypeptide chain (Supplementary Figure S1B), higher than modal ^{15}N R_2 values were observed in several regions of the primary sequence, i.e. 9–52, 110–141, 207–220, 226–256, 261–269 and 308–406 (Fig. 3B). The segments encompassing residues 9–52 and the C-terminal part of the polypeptide chain (340–406) are regions that are partially unassigned in our experimental conditions. This observation, as well as the absence of correlations in the multidimensional spectra, confirms that these regions are affected by dynamical phenomena leading to high R_2 values, such as the presence of transiently populated secondary structural elements or chemical exchange broadening. The positive SSP values for the residues that can be observed in the regions 339–357 and 367–373 suggest the presence of transiently formed α -helices in these regions. The following stretch of about 30 amino acids, from residue 376 to the end of the protein primary sequence, is particularly rich in positively charged amino acids (22%), which suggests the possibility that the high nuclear relaxation rates observed in the absence of pronounced local secondary structural elements are due to the occurrence of transient electrostatic interactions (Fig. 1B and Supplementary Figure S2). Interestingly, the region encompassing residues 226–256 and 261–269, which overall shows higher than modal ^{15}N R_2 , is rich in negatively charged residues (33%) and could indeed be engaged in transient interaction with the C-terminal, positively charged stretch (376–406).

It is also interesting to note that this latter region (i.e. 376–406) is rich in positively charged amino acids (with 7 of such residues), and also contains 4 prolines (Fig. 1B and Supplementary Figure S2). This particular amino acid composition of alternating proline residues with amino acids sharing the same charge (positive in this case) could thus have a role in promoting conformations with a local net charge even in the absence of well-defined secondary structural elements. The same pattern of regions rich in amino acids sharing the same charge and proline residues is found in other regions, such as $^{267}\text{PEDEEP}^{272}$, although in this case the amino acids are negatively charged. (Fig. 1B and Supplementary Figure S2). The NMR data thus suggest a possible cross-talk between regions enriched in residues with opposite charge. The 238–256 region is negatively charged and can establish interactions with positively charged regions, such as the C-terminal one encompassing residues 376–406 (Fig. 1B and Supplementary Figure S2).

Additional amino acid stretches showing higher than modal ^{15}N R_2 values are 110–141 and 207–220. These do not show evidence of transiently populated secondary structural elements nor highly charged segments. However, looking at the amino acid composition, in particular at amino acid types that are quite rare in IDPs³³, one can observe a number of aromatic residues in these regions. These account for 22% of the residues in region 110–141 and 23% in region 207–220, indicating that these regions feature a higher amount of aromatic residues with respect to what is observed for the whole protein (6%) (Fig. 1B and Supplementary Figure S2). Aromatic residues are bulky amino acids and might be involved in local transient interactions without inducing specific secondary structural elements explaining the higher ^{15}N transverse relaxation rates observed, as also observed for other IDPs³⁴ in which aromatic residues have been shown to promote compact states.

Additional information that can be obtained by NMR derives from the investigation of exchange processes of amide protons with the solvent. Several CLEANEX experiments were acquired with different mixing times to measure the exchange rate constants³⁵. Intensities of cross peaks are measured to obtain an estimation of the k_{ex} (s^{-1}) (Supplementary Figure S1C). Thus, k_{ex} reports on solvent exchange processes of amide protons with the solvent on a per-residue basis. While this observable has been used extensively to characterize globular proteins and identify amide protons that are buried in hydrophobic cores and/or involved in hydrogen bonds through the determination of “protection factors”, in the case of IDPs it is not straightforward to interpret hydrogen exchange data in terms of structural effects as these are not easy to disentangle in a clear way from other effects such as local electrostatic potential, nature of the amino acid, and possible effects of neighboring amino acids in the primary sequence^{36,37}. On the other hand, there is no doubt that this observable is rich in information^{36–38} and that, as we make progress in understanding the different contributions and in predicting exchange values for random coil conformations³⁹, it might become a useful tool to achieve information about the structural and dynamic properties of IDPs.

Inspecting the observed k_{ex} for NiV PNT as a function of the primary sequence reveals that some regions present a k_{ex} that is higher than the modal value of 1.58 Hz, while some others are well below this value. Higher values are observed for residues 27–30, between two regions with large, positive SSP values at the edges of two transiently populated helices. Another region with high k_{ex} is the one encompassing residues 119–147, rich in aromatic residues (17%; Fig. 1B and Supplementary Figure S2), where slightly higher ^{15}N R_2 values were identified, in agreement with the presence of a solvent exposed region. Another region with significantly high rates is the one spanning residues 182–194, a region that is very rich in serine and threonine residues (38% in this segment *versus* 16% in the whole protein), two amino acids that are often solvent exposed and are characterized by high exchange rates with the solvent in general³⁹ (Fig. 1B and Supplementary Figure S2). The region presenting lower k_{ex} values (slow chemical exchange) are the regions encompassing residues 97–122 and 239–263, which are

rich in negatively charged residues (31% in the first segment, 40% in the second one) (Fig. 1B and Supplementary Figure S2), a feature that is known to cause reduced exchange processes^{37,40}.

The NMR data clearly suggest that NiV PNT is disordered and very flexible with few short and transient secondary structure elements. The N-terminus shows moderate positive secondary structure propensity for residues 68–81, while some hints of the occurrence of other transiently populated α -helices can be obtained from the combination of ^{15}N R_2 , SSP and k_{ex} values in the region encompassing residues 10–30. This region contains a segment (1–29) that adopts a helix-kink-helix conformation in the crystal structure of the N^0 -P complex⁴. Subsequent NMR studies confirmed the presence of two transiently populated α -helices at the N-terminus of NiV PNT also in solution and in the absence of the binding partner^{4,13}. The presence of these helical elements mildly populated in our experimental conditions might be responsible for the conformational exchange processes likely causing the broadening of the signals in the full construct used here. Nevertheless, a few amino acids are detectable in this region and do show SSP values that indicate α -helical propensity (Fig. 3A). The more efficient transverse relaxation in these regions, as compared to the studies performed on shorter constructs, could be due to subtle differences arising from intramolecular interactions.

The overall behavior of PNT in solution might also be influenced by the experimental conditions. Indeed, the buffer used in¹³ includes compounds that mitigate protein self-association and aggregation (150 mM NaCl, 50 mM arginine, 50 mM glutamate)⁴¹. It should be pointed out however that in other cases the same buffer was reported to induce protein compaction by minimizing the effect of intramolecular long-range electrostatic interactions⁴². The causes of these conflicting effects are not well understood⁴².

Similar considerations also hold for the 340–355 region that was shown by Jensen and co-workers to partly sample an α -helical conformation¹³. Although we could not map the core region of the transient α -helix we mapped the flanking residues that, also in this case, confirm the presence of a transient α -helix in this region. This transiently populated α -helix may correspond to a binding site for one or multiple partners whose identification will require future studies. A possible partner could be the unassembled form of the N protein, by analogy with measles virus. Indeed, the counterpart of this region in measles virus P corresponds to a transiently populated α -helix ($\alpha 4$, aa 190–198) that binds weakly N^{043} .

The transiently populated α -helices are in good agreement with predictions as obtained with both FESS (the fast secondary structure predictor implemented in FIELDS⁴⁴ (Supplementary Figure S2) and PSIPRED⁴⁵ (Supplementary Figure S3A). In terms of predicted MoREs (see Supplementary Figure S3A), a quite good agreement was found with the experimental data, as well as with already mapped binding sites, including the STAT1 (aa 110–140)¹³ and STAT2 (aa 110–140 and 230–237) binding sites⁴⁶. The 110–140 region shown to constitute a weak binding site to STAT1¹³ does not exhibit significant SSP values (all the values are below 0.2), a finding that suggests that this region might correspond to an I-MoRE (Irregular-MoRE), i.e. a region that may remain at least partially disordered after binding to the partner. This latter hypothesis is further corroborated by the fact that addition of STAT1 only triggers a decrease in the resonance intensities of the 110–140 region with no concomitant peak shift¹³.

An analysis performed on the ELM database (<https://elm.eu.org/>)⁴⁷ highlighted the presence of numerous Short Linear Motifs (SLiMs) along the sequence. SLiMs, also known as linear motifs (LMs), are short stretches of adjacent amino acids mediating protein–protein interactions and occurring within IDRs⁴⁸. SLiMs play crucial roles in cell regulation and SLiM mimicry is often used by viruses to hijack their host cellular machinery⁴⁹. A nice correlation was found between the presence of SLiMs and the dynamic behavior of NiV PNT. Indeed, 22 SLiMs were identified (Supplementary Table S3) and most of them (18) are located within the regions displaying high R_2 values (Supplementary Figure S3B). As an example, the previously mentioned C-terminal region (aa 376–406), enriched in proline residues and in positively charged residues, features several (i.e. 7) SLiMs. These SLiMs are mainly involved in interactions with kinases, phosphatases, SH3 and WW regulatory domains, consistent with a role in hijacking key cell regulatory processes. These interactions can now be investigated at atomic resolution through ^{13}C -detected experiments tailored for proline residues that allow acquiring quick snapshots in a clean spectral region^{50,51}.

SAXS ensemble characterization of NiV PNT. Synchrotron SAXS in solution measurements were performed to gain insight into overall NiV PNT conformation and motion. Linearity in the Guinier region at low angles ($sR_g < 1.0$) revealed good data quality, with no indication of protein aggregation (Fig. 4A, inset). The molecular mass determined from the forward scattering intensity at zero angle $I(0)$ was ~ 47 kDa indicating a monomer in solution and in agreement with the value (45.3 kDa) computed from the amino acid sequence using the expasy server (<https://web.expasy.org/protparam/>). Although not significantly different, the R_g extracted from $P(r)$ is slightly larger (~ 63 Å) than the one extracted from Guinier analysis (~ 61 Å). The theoretical R_g value expected for an IDP of the same length (412 amino acids) calculated using Flory's power law $R_g^{\text{IDP}} = R_0 N^{\nu}$, where R_0 is 2.54 ± 0.01 , N is number of amino acid and ν is 0.522 ± 0.001 ⁵², is ~ 59 Å. The experimentally determined R_g is therefore very close to the value expected for an IDP (for comparison the R_g expected for a random coil conformation would be ~ 72 ⁵³).

$P(r)$ yields a maximal dimension, D_{max} , of ~ 230 Å (Fig. 4B) with a long tail in the $P(r)$ function, suggesting that the protein tends to assume an overall non-compact conformation⁵⁴. The overall SAXS parameters for NiV PNT are listed in Supplementary Table S4.

The flexible nature of NiV PNT was qualitatively assessed by using the normalized Kratky plot, where the absence of a well-defined bell shape indicates a protein with intrinsically disordered regions (Fig. 4C). To achieve further insights into the conformational behavior of NiV PNT, data obtained from NMR and SAXS experiments were combined in search of an ensemble that quantitatively describes the conformational behavior of NiV PNT in solution. In a first step, a pool of 50,000 structures was generated based on the NMR data. To this end, the ncSSP

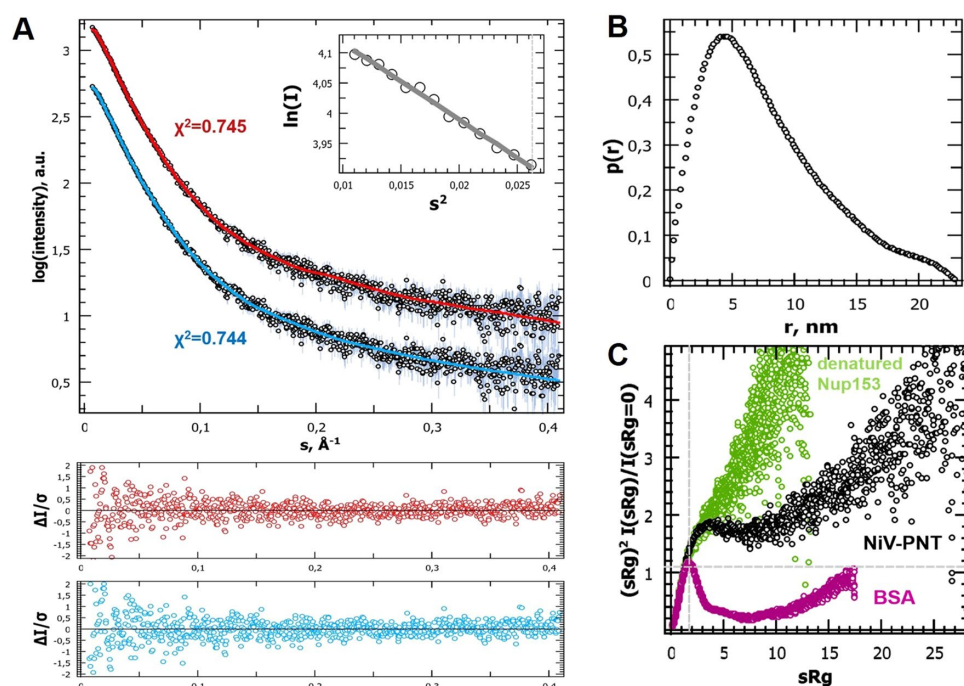


Figure 4. SAXS studies of NiV PNT. (A) SAXS scattering curve of NiV PNT and ensemble EOM 2.0 fits extracted from two distinct initial pools: (blue) generated by Flexible-Meccano considering ncSSP and (red) generated by EOM 2.0 including random Ca-only conformers. Inset: Guinier plot of the experimental scattering curve. (B) Pair distance distribution function, $P(r)$. (C) Normalized Kratky plot representation of the scattering data. The normalized scattering plots of a globular (BSA, SASDA32) and denatured (Nup153, SASDEY2) protein are shown for comparative purposes to illustrate that NiV PNT has an overall conformation in between these two extreme examples.

values larger than ± 0.05 were provided as inputs to Flexible-Meccano⁵⁵. Subsequently, from the NMR-based pool, we used Ensemble Optimization Method (EOM) 2.0 to generate a conformational sub-ensemble that best fits the experimental SAXS data. In order to minimize over-fitting, EOM attempts at minimizing the number of conformers able to fit the experimental data and usually generates ensembles consisting of 5 to 40 conformers. Based on the high flexibility expected, no repetition of conformations in the ensemble was allowed. The scattering curve back-calculated from the selected ensemble (Fig. 4A, cyan curve) fits well the experimental SAXS data as judged from both χ^2 and CorMap⁵⁶ ($\chi^2 = 0.744$, $P = 0.101$). The resulting final R_g distribution, broader than the one generated from the NMR-based pool, indicates that NiV PNT exists in solution as a randomly distributed ensemble of non-compact and highly flexible conformations. Note that successive and independent selections by EOM 2.0 consistently yielded similar R_g distributions thus attesting the reproducibility of the results (data not shown). The flexibility of the ensemble was quantified as $\text{Ensemble}_{\text{Rflex}} = \sim 91\%$ ($\text{NMR-Pool}_{\text{Rflex}} = \sim 83\%$).

The average R_g value of the ensemble (66 Å) slightly exceeds the theoretical value expected from Flory's power law (59 Å). This discrepancy may reflect sequence specificities, i.e. specific sequence attributes such as proline content and charge decoration (i.e. net charge per residue, fraction of charged residues and linear distribution of opposite charges) that were shown to be major determinants of IDP conformational properties^{57–62}. For an additional discussion of the observed discrepancy see Supplementary text.

A similar investigation was conducted by using a random pool generated by EOM 2.0 without NMR-based restraints ($\chi^2 = 0.745$, $P = 0.192$) (Fig. 4A, red curve). EOM 2.0 yielded similar results with randomness quantified as $\text{Ensemble}_{\text{Rflex}} = \sim 92\%$ ($\text{Random-Pool}_{\text{Rflex}} = \sim 83\%$) such that no significant differences was observed when using different pools (Fig. 5).

Although SAXS is poorly sensitive to secondary structure and although the presence of secondary structure elements has been reported to have a limited impact on protein compaction^{57,63}, we sought at investigating the impact of the occurrence of the N-terminal transiently populated α -helix on chain compaction. To this end, using Flexible-Meccano, we generated an initial pool in which an α -helical conformation was imposed to residues 1–30, with the experimental ncSSPs. Furthermore, we also investigated the final R_g distribution while imposing a kink centered at residue 20 (as observed in the crystal structure of the $\text{N}^0\text{-P}_{1-50}$ complex) providing NMR restraints. All the resulting sub-ensembles fit equally well the scattering data as judged from the obtained χ^2 values and even for these two cases, no differences were observed in terms of the resulting R_g distributions (Supplementary Figure S4). This probably reflects the fact that SAXS is a low-resolution technique that provides

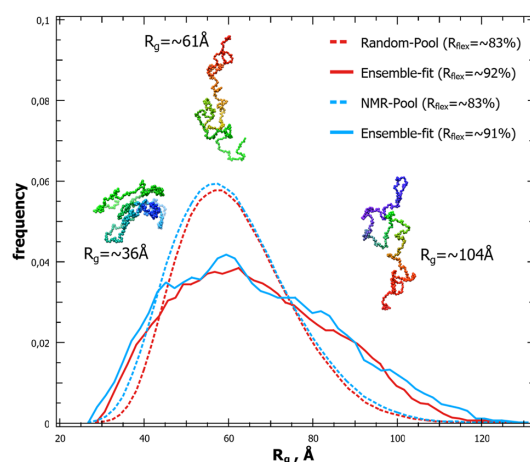


Figure 5. Modeling NiV PNT as a conformational ensemble. R_g distributions of the initial pools generated by (blue dot) Flexible-Meccano (with NMR secondary structure restraints, $\overline{R}_g = \sim 62 \text{ \AA}$) and (red dot) EOM 2.0 (without restraints, $\overline{R}_g = \sim 62 \text{ \AA}$) respectively, and of the corresponding selected ensembles (solid lines) as obtained using EOM 2.0 (both $\overline{R}_g = \sim 66 \text{ \AA}$, see also Fig. 4A). Final best ensembles contain 19 unique conformers in the case of ncSSP NMR-based ensemble and 13 unique conformers in the case of EOM 2.0 random ensemble, respectively. A cartoon representation of three conformers present in the ensembles is also displayed. The structures were drawn using Pymol 2.0.1 (<https://pymol.org/2/>)¹¹.

relevant information in terms of chain compactness but fails to capture secondary structure propensities. In addition, previous findings from others and us have highlighted that regular secondary structure does not overly contribute to protein compaction (i.e. more compact forms do not necessarily exhibit an increased content in regular secondary structure)^{57,63}. As such, it is not surprising that the presence of transiently populated α -helices in the initial ensemble does not have a significant impact on the final SAXS-derived sub-ensembles. Nevertheless, comparing SAXS-only ensembles to ensembles generated by the combined use of NMR and SAXS restraints is of interest given the growing interest towards ensemble descriptions of IDPs. In addition, combining NMR and SAXS data enables at least partly overcoming over-fitting as already discussed by Mertens and Svergun⁶⁴.

Conclusions

Albeit IDPs/IDRs are notoriously prone to undergo proteolytic degradation, we managed in generating a full-length NiV PNT sample stable enough to allow acquiring all the NMR spectra required for the assignment. Although the length of this intrinsically disordered domain is very challenging for NMR studies, the use of multidimensional NMR spectroscopy and of ¹³C-detected experiments allowed us to assign as much as 91% of the residues including prolines. The combination of SAXS and NMR data led to ensemble models of NiV PNT showing its conformational heterogeneity.

The presence of such a large disordered region shared by both the V and P proteins is likely related to their promiscuity. Structural disorder is known to serve as a determinant of protein interactivity^{65–67}. Indeed V interacts with various cellular partners, such as DDB1¹⁰, STAT proteins^{13,46,68}, CRM1^{46,69}, PLK1⁷⁰ and, possibly, nuclear factors such as IRF3. All these interactions play a crucial role in the ability of the virus to counteract the antiviral innate immune response of the host. The P protein binds to multiple partners as well, including N and the L protein (for a review see⁷¹). Therefore, intrinsic disorder represents an ergonomic solution for the virus to encode fewer proteins with more functions while keeping the genome size as small as possible. It is therefore not surprising to find large IDRs in proteins having a broad molecular partnership, such as the *Henipavirus* V and P proteins, and, more generally, in viral proteins^{72,73}.

Intrinsic disorder also represents a strategy to alleviate constraints imposed by evolution on overlapping reading frames (such in the case of PNT that overlaps with the C protein) and to buffer the deleterious effect of mutations (i.e. IDRs/IDPs are more tolerant of substitutions compared to globular proteins) (for reviews see^{71,73}). It is also well established that intrinsic disorder allows affinity and specificity to be uncoupled (i.e. it enables interactions with low affinity and decent specificity) (see^{33,74,75} and references therein cited). However, the occurrence of residual disorder in complexes involving IDPs (in the form of fuzzy appendages) and partial preconfiguration of binding motifs before binding afford a way to attenuate the entropic penalty associated to the disorder-to-order transition thereby, ultimately, modulating the binding affinity. The involvement of IDRs in protein–protein interactions that need to be finely tuned offers an exquisite means to modulate the strength of those interactions: by tuning the extent of preconfiguration of the binding motifs and/or the length of flanking fuzzy appendages, the virus can reach an optimal binding strength. It is therefore conceivable that the presence within NiV PNT of long disordered regions flanking binding motifs, including SLiMs, as well as binding sites

to N⁰ or STAT proteins, may exert a role in regulating these critical interactions. The flexibility of these flanking regions would enable regulating both the exposure of binding sites specifically recognized by multiple viral and cellular partners, and the affinity of these interactions. Ultimately, this would result in their ability to orchestrate virus replication, through hijacking of cellular pathways and evasion of the IFN response.

The present results constitute an asset for obtaining atomistic information in future interaction studies involving NiV PNT. Indeed, the availability of the chemical shifts of full-length NiV PNT will be instrumental to map the residues involved in binding to partners such as DDB1 and/or STAT1 in the context of either the isolated PNT domain or the V protein. Once detailed information is available from interaction studies, inhibitors can be rationally conceived. Therefore the present study can be regarded as a starting point towards the design of inhibitors abrogating the ability of this virus to escape the innate immune response. Given the high similarity between NiV and HeV PNT (56%), it is conceivable that the results of these studies could be extrapolated to some extent to the HeV as well.

Methods

Protein expression and purification. The NiV PNT construct, encoding residues 1–406 of the NiV P protein with a C-terminal hexahistidine tag, has been already described¹⁴. Expression of unlabeled NiV PNT was carried out as previously described¹⁴.

Isotopically labeled (either ¹⁵N–¹³C or ¹⁵N) NiV PNT samples were prepared by growing at 37 °C transformed *E. coli* T7 cells (New England Biolabs, Ipswich, MA, USA) bearing the pRARE plasmid (Novagen, Madison, WI, USA) in LB medium supplemented with 100 µg mL⁻¹ ampicillin and 34 µg mL⁻¹ chloramphenicol. When the OD₆₀₀ reached 0.6, the culture was centrifuged at 4000 rpm for 10 min and the pellet was resuspended in ¼ of the initial volume of M9 medium (6 g L⁻¹ of Na₂HPO₄, 3 g L⁻¹ of KH₂PO₄, 0.5 g L⁻¹ of NaCl, 0.246 g L⁻¹ of MgSO₄) supplemented with 1 g L⁻¹ of ¹⁵NH₄Cl and 2 g L⁻¹ of either glucose or ¹³C-glucose. After one hour at 37 °C, IPTG was added to a final concentration of 0.5 mM, and the cells were subsequently grown at 37 °C for 3.5 h.

NiV PNT was purified as described in¹⁴, except that 6 M GuHCl was added after the lysis step to both denature bacterial proteases and recover the recombinant protein also from inclusion bodies with the aim of improving protein stability and yield. After 1 h incubation at 4 °C, the sample was clarified and the supernatant was purified through immobilized metal affinity chromatography (IMAC). The fractions containing the recombinant protein were combined, and then loaded onto a Superdex 75 h 16/60 column (GE, Healthcare). The elution buffer was 10 mM sodium phosphate pH 6.5, supplemented with 5 mM EDTA and 5 mM DTT. The fractions containing the protein were collected and conserved at –20 °C.

NMR spectroscopy experiments. Immediately before NMR studies, the NiV PNT sample was thawed, centrifuged at 12,000 rpm to remove any possible protein aggregate and then concentrated using 30 kDa Amicon Ultra Centrifugal Filters (Merk Millipore, Darmstadt, Germany). All NMR spectra for assignment were acquired on ¹³C–¹⁵N-isotopically enriched NiV PNT at a concentration of about 200 µM.

The ¹H-detected spectra were acquired at 288 K with a 22.3 T Bruker Avance III 950 NMR spectrometer; the ¹³C-detected spectra were acquired at 288 K with a 16.4 T Bruker Avance NEO 700 NMR spectrometer equipped with a cryogenically cooled probehead optimized for ¹³C-direct detection. Experimental details are reported in Supplementary Table S1.

All the heteronuclear relaxation experiments (R₁, R₂ and ¹H–¹⁵N NOEs) were acquired using ¹⁵N labeled NiV PNT at about 100 µM. The spectra were recorded at 288 K on a 16.4 T Bruker Avance NEO 700 NMR spectrometer equipped with a cryogenically cooled triple resonance probehead. The ¹⁵N R₁ and R₂ experiments were acquired with 8 scans (2048 × 256 points) and a relaxation delay of 3.0 s. To determine the ¹⁵N R₁ the following delays were used: 20 ms, 60 ms, 120 ms, 180 ms, 250 ms, 400 ms, 500 ms, 600 ms, 750 ms, and 900 ms. To determine the ¹⁵N R₂ the following delays were used: 32 ms, 64 ms, 96 ms, 128 ms, 160 ms, 190 ms, 260 ms, 320 ms, 380 ms, 440 ms, and 500 ms. The ¹H–¹⁵N NOEs experiments were acquired with 64 scans (2048 × 288 points) and a relaxation delay of 6.0 s.

The CLEANEX experiments were acquired on a 100 µM sample, with 16 scans (2048 × 248 points) and a relaxation delay of 3.0 s. The following delays were used: 5 ms, 10 ms, 20 ms, and 30 ms.

NMR data processing and analysis. NMR data sets were processed using the Bruker TopSpin 4.0.6 software. CARR⁷⁶ and its tool XEASY⁷⁷ were used to analyze and annotate the spectra.

The ¹⁵N relaxation rates (R₁ and R₂) were determined by fitting the cross-peak intensity measured as a function of variable delay, to single-exponential decay using the Bruker Dynamic Center 2.4, available as a stand-alone ancillary software of TopSpin by Bruker. ¹H–¹⁵N NOE values were obtained as the ratio between peak intensity in spectra recorded with and without ¹H saturation.

The secondary structure propensity (SSP) from heteronuclear chemical shifts was determined by using the neighbor corrected structural propensity calculator (ncSPC) tool³² available online at <https://st-protein02.chem.au.dk/ncSPC>. The Mulder random coil chemical shift library³¹ was chosen for the analysis, and the average window size was left to the standard value of 5.

SAXS experiments and analysis. Samples were concentrated using 30 kDa Amicon Ultra Centrifugal Filters (Merk Millipore, Darmstadt, Germany) and loaded onto a Superdex 75 h 16/60 column using 20 mM Tris/HCl pH 8.0, 0.3 M NaCl, 5 mM DTT as elution buffer. Synchrotron X-ray scattering data were collected at ESRF BM29 beamline (Grenoble) using a PILATUS 1 M pixel detector (DECTRIS, Baden, Switzerland) at a sample-detector distance of 2.87 m and a wavelength of 0.0992 nm. This setup covers a range of momentum transfer of 0.028 < s < 4.525 nm⁻¹ ($s = 4\pi \sin(\theta)/\lambda$, where 2θ is the scattering angle). Samples were loaded using a

robotic sample changer⁷⁸ and measured at 20 °C at concentrations ranging from 0.5 to 1.2 mg/mL. 10 independent frames of 1 s each were collected for each sample and data were automatically reduced using an in-house pipeline. Further analysis was done using ATSAS 3.0.1⁷⁹. As no concentration dependence was observed (Supplementary Figure S5), the curve with less noise (at 1.2 mg/mL) was used for the SAXS analysis. The forward scattering $I(0)$ as well as the R_g were calculated using the Guinier approximation assuming that, at very small angles ($s < 1.3/R_g$), the intensity is represented as $I(s) = I(0) \cdot \exp(-sR_g)^2/3$ ^{80,81}. The forward scattering intensities were calibrated using water as reference. Linearity in the Guinier region was used to exclude sample aggregation, and the pair-distance distribution function, $P(r)$, from which the D_{max} and the R_g were estimated, was computed using GNOM⁸². Qualitative assessment of compactness *versus* structural disorder was made by transforming the scattering profiles in the so-called normalised Kratky representation $[(sR_g)^2 * I(s)/I(0) \text{ vs. } sR_g]$ ⁸³. The MM was derived by placing the scattered intensity on an absolute scale using liquid water as calibrant.

Protein flexibility was quantified using EOM 2.0⁸⁴ that assumes coexistence of a range of Ca-only conformations in solution for which an average scattering intensity fits the experimental data. A genetic algorithm (GA) is used to select ensembles with varying numbers of conformers. The GA is repeated n independent times, and the ensemble with the lowest discrepancy considered as the best solution out of n final ensembles. Furthermore, repetition of GA allows the computation of R_g distributions so that structural information about the flexibility could be extracted. The width of the distribution is used to derive the flexibility of the particles, whereby a narrow distribution indicates a rather rigid particle and broader distributions are associated with higher flexibility. Using EOM 2.0, systematic quantification of the flexibility was made by using the metric R_{flex} —which computes the Shannon information entropy of the distributions⁸⁴. Experimental error-independent goodness-of-fit was also confirmed by using the software CorMap that estimates the differences between one-dimensional spectra independently of explicit error estimates, using only data point correlations⁵⁶. All the softwares used for the SAXS data analysis are part of the ATSAS 3.0 package⁷⁹.

Accession numbers. The chemical shifts have been deposited in the Biological Magnetic Resonance Data Bank (BMRB) under the code 50370. SAXS data have been deposited in the Small Angle Scattering Biological Data Bank (SASBDB)⁸⁵ under code SASDJB5. The ensemble derived using SAXS constraints only and the one derived from the combined use of SAXS and NMR constraints have been deposited within the Protein Ensemble Database (PED-DB, <https://proteinsenemble.org/>)⁸⁶ under accession numbers PED00176 and PED00177, respectively. All-atoms pdb files, required for deposition in the PED-DB, were generated from EOM Ca-only pdb files using the PD2ca2main server (http://www.sbg.bio.ic.ac.uk/~phyre2/PD2_ca2main/)⁸⁷.

Received: 1 July 2020; Accepted: 27 October 2020

Published online: 11 November 2020

References

- Wang, L. F. *et al.* The exceptionally large genome of Hendra virus: Support for creation of a new genus within the family Paramyxoviridae. *J. Virol.* **74**, 9972–9979 (2000).
- Marsh, G. A. *et al.* Cedar virus: A novel Henipavirus isolated from Australian bats. *PLoS Pathog.* **8**, e1002836 (2012).
- Boyet, L. M. *et al.* HSP90 chaperoning in addition to phosphoprotein required for folding but not for supporting enzymatic activities of measles and Nipah virus L polymerases. *J. Virol.* **90**, 6642–6656 (2016).
- Yabukarski, F. *et al.* Structure of Nipah virus unassembled nucleoprotein in complex with its viral chaperone. *Nat. Struct. Mol. Biol.* **21**, 754–759 (2014).
- Park, M. S. *et al.* Newcastle disease virus (NDV)-based assay demonstrates interferon-antagonist activity for the NDV V protein and the Nipah virus V, W, and C proteins. *J. Virol.* **77**, 1501–1511 (2003).
- Fontana, J. M., Bankamp, B. & Rota, P. A. Inhibition of interferon induction and signaling by paramyxoviruses. *Immunol. Rev.* **225**, 46–67 (2008).
- Shaw, M. L., Garcia-Sastre, A., Palese, P. & Basler, C. F. Nipah virus V and W proteins have a common STAT1-binding domain yet inhibit STAT1 activation from the cytoplasmic and nuclear compartments, respectively. *J. Virol.* **78**, 5633–5641 (2004).
- Bruhn-Johannsen, J. F. *et al.* Crystal structure of the Nipah virus phosphoprotein tetramerization domain. *J. Virol.* **88**, 758–762 (2014).
- Habchi, J. *et al.* Characterization of the interactions between the nucleoprotein and the phosphoprotein of Henipaviruses. *J. Biol. Chem.* **286**, 13583–13602 (2011).
- Salladini, E., Delaunay, V. & Longhi, S. The Henipavirus V protein is a prevalently unfolded protein with a zinc-finger domain involved in binding to DDB1. *Mol. Biosyst.* **13**, 2254–2267 (2017).
- DeLano, W. L. The PyMOL molecular graphics system. *Proteins: Struct. Funct. Bioinf.* **30**, 442–454 (2002).
- Wootton, J. C. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.* **18**, 269–285 (1994).
- Jensen, M. R. *et al.* Structural description of the Nipah virus phosphoprotein and its interaction with STAT1. *Biophys. J.* **118**, 2470–2488 (2020).
- Habchi, J., Mamelli, L., Darbon, H. & Longhi, S. Structural disorder within henipavirus nucleoprotein and phosphoprotein: From predictions to experimental assessment. *PLoS ONE* **5**, e11684 (2010).
- Blocquel, D., Beltrandi, M., Eraldes, J., Barbier, P. & Longhi, S. Biochemical and structural studies of the oligomerization domain of the Nipah virus phosphoprotein: Evidence for an elongated coiled-coil homotrimer. *Virology* **446**, 162–172 (2013).
- Karlin, D., Ferron, F., Canard, B. & Longhi, S. Structural disorder and modular organization in Paramyxovirinae N and P. *J. Gen. Virol.* **84**, 3239–3252 (2003).
- Li, T., Chen, X., Garbutt, K. C., Zhou, P. & Zheng, N. Structure of DDB1 in complex with a paramyxovirus V protein: Viral hijack of a propeller cluster in ubiquitin ligase. *Cell* **124**, 105–117 (2006).
- Karlin, D. & Belshaw, R. Detecting remote sequence homology in disordered proteins: Discovery of conserved motifs in the N-termini of Mononegavirales phosphoproteins. *PLoS ONE* **7**, e31719 (2012).

19. Brutscher, B. *et al.* NMR methods for the study of intrinsically disordered proteins structure, dynamics, and interactions: General overview and practical guidelines. *Adv. Exp. Med. Biol.* **870**, 49–122 (2015).
20. Felli, I. C. & Pierattelli, R. Recent progress in NMR spectroscopy: Toward the study of intrinsically disordered proteins of increasing size and complexity. *IUBMB Life* **64**, 473–481 (2012).
21. Hiller, S., Fiorito, F., Wuthrich, K. & Wider, G. Automated projection spectroscopy (APSY). *Proc. Natl. Acad. Sci. USA* **102**, 10876–10881 (2005).
22. Murrall, M. G. *et al.* ¹³C APSY-NMR for sequential assignment of intrinsically disordered proteins. *J. Biomol. NMR* **70**, 167–175 (2018).
23. BermeL, W. *et al.* Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. *J. Am. Chem. Soc.* **128**, 3918–3919 (2006).
24. BermeL, W. *et al.* H-start for exclusively heteronuclear NMR spectroscopy: The case of intrinsically disordered proteins. *J. Magn. Reson.* **198**, 275–281 (2009).
25. Felli, I. C., Pierattelli, R., Glaser, S. J. & Luy, B. Relaxation-optimised Hartmann–Hahn transfer using a specifically Tailored MOCCA-XY16 mixing sequence for carbonyl–carbonyl correlation spectroscopy in ¹³C direct detection NMR experiments. *J. Biomol. NMR* **43**, 187–196 (2009).
26. Mateos, B., Konrat, R., Pierattelli, R. & Felli, I. C. NMR characterization of long-range contacts in intrinsically disordered proteins from paramagnetic relaxation enhancement in ¹³C direct-detection experiments. *ChemBioChem* **20**, 335–339 (2019).
27. BermeL, W. *et al.* High-dimensionality ¹³C direct-detected NMR experiments for the automatic assignment of intrinsically disordered proteins. *J. Biomol. NMR* **57**, 353–361 (2013).
28. Solyom, Z. *et al.* BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J. Biomol. NMR* **55**, 311–321 (2013).
29. Piai, A. *et al.* “CON–CON” assignment strategy for highly flexible intrinsically disordered proteins. *J. Biomol. NMR* **60**, 209–218 (2014).
30. Marsh, J. A., Singh, V. K., Jia, Z. & Forman-Kay, J. D. Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: Implications for fibrillation. *Protein Sci.* **15**, 2795–2804 (2006).
31. Tamiola, K., Acar, B. & Mulder, F. A. Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *J. Am. Chem. Soc.* **132**, 18000–18003 (2010).
32. Tamiola, K. & Mulder, F. A. Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochem. Soc. Trans.* **40**, 1014–1020 (2012).
33. Habchi, J., Tompa, P., Longhi, S. & Uversky, V. N. Introducing protein intrinsic disorder. *Chem. Rev.* **114**, 6561–6588 (2014).
34. Mateos, B. *et al.* The ambivalent role of proline residues in an intrinsically disordered protein: From disorder promoters to compaction facilitators. *J. Mol. Biol.* **432**, 3093–3111 (2020).
35. Hwang, T. L., van Zijl, P. C. & Mori, S. Accurate quantitation of water-amide proton exchange rates using the phase-modulated CLEAN chemical EXchange (CLEANEX-PM) approach with a Fast-HSQC (FHSQC) detection scheme. *J. Biomol. NMR* **11**, 221–226 (1998).
36. Chevelkov, V., Xue, Y., Rao, D., Forman-Kay, J. & Skrynnikov, N. ¹⁵N H/D-SOLEXSY experiment for accurate measurement of amide solvent exchange rates: Application to denatured drkN SH3. *J. Biomol. NMR* **46**, 227–244 (2010).
37. Croke, R., Sallum, C., Watson, E., Watt, E. & Alexandrescu, A. Hydrogen exchange of monomeric alpha-synuclein shows unfolded structure persists at physiological temperature and is independent of molecular crowding in *Escherichia coli*. *Protein Sci.* **17**, 1434–1445 (2008).
38. Smith, A., Zhou, L. & Pielak, G. Hydrogen exchange of disordered proteins in *Escherichia coli*. *Protein Sci.* **24**, 706–713 (2015).
39. Bai, Y., Milne, J., Mayne, L. & Englander, S. Primary structure effects on peptide group hydrogen exchange. *Proteins* **17**, 75–86 (1993).
40. Pontoriero, L., Schiavina, M., Murrall, M., Pierattelli, R. & Felli, I. Monitoring the interaction of alpha-synuclein with calcium ions through exclusively heteronuclear nuclear magnetic resonance experiments. *Angew. Chem. Int. Ed. Engl.* **59**, 18537–18545 (2020).
41. Golovanov, A., Hautbergue, G., Wilson, S. & Lian, L. A simple method for improving protein solubility and long-term stability. *J. Am. Chem. Soc.* **126**, 8933–8939 (2004).
42. Blobel, J. *et al.* Protein loop compaction and the origin of the effect of arginine and glutamic acid mixtures on solubility, stability and transient oligomerization of proteins. *Eur. Biophys. J.* **40**, 1327–1338 (2011).
43. Milles, S. *et al.* An ultraweak interaction in the intrinsically disordered replication machinery is essential for measles virus function. *Sci. Adv.* **4**, eaat7778 (2018).
44. Piovesan, D., Walsh, I., Minervini, G. & Tosatto, S. C. E. FIELDS: Fast estimator of latent local structure. *Bioinformatics* **33**, 1889–1891 (2017).
45. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
46. Rodriguez, J. J., Cruz, C. D. & Horvath, C. M. Identification of the nuclear export signal and STAT-binding domains of the Nipah virus V protein reveals mechanisms underlying interferon evasion. *J. Virol.* **78**, 5358–5367 (2004).
47. Kumar, M. *et al.* ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.* **48**, D296–D306 (2020).
48. Van Roey, K. *et al.* Short linear motifs: Ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.* **114**, 6733–6778 (2014).
49. Davey, N. E., Trave, G. & Gibson, T. J. How viruses hijack cell regulation. *Trends Biochem. Sci.* **36**, 159–169 (2011).
50. Murrall, M. G., Piai, A., BermeL, W., Felli, I. C. & Pierattelli, R. Proline fingerprint in intrinsically disordered proteins. *ChemBioChem* **19**, 1625–1629 (2018).
51. Schiavina, M. *et al.* Taking simultaneous snapshots of intrinsically disordered proteins in action. *Biophys. J.* **117**, 46–55 (2019).
52. Bernado, P. & Svergun, D. I. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst.* **8**, 151–167 (2012).
53. Fitzkee, N. & Rose, G. Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci. USA* **101**, 12497–12502 (2004).
54. Bernado, P. Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering. *Eur. Biophys. J.* **39**, 769–780 (2010).
55. Ozenne, V. *et al.* Flexible-meccano: A tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* **28**, 1463–1470 (2012).
56. Franke, D., Jeffries, C. M. & Svergun, D. Correlation map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. *Nat. Methods* **12**, 419–422 (2015).
57. Marsh, J. A. & Forman-Kay, J. D. Sequence determinants of compaction in intrinsically disordered proteins. *Biophys. J.* **98**, 2383–2390 (2010).
58. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. & Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. USA* **107**, 8183–8188 (2010).
59. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. USA* **110**, 13392–13397 (2013).
60. Das, R. K., Ruff, K. M. & Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **32**, 102–112 (2015).

61. Tedeschi, G. *et al.* Conformational response to charge clustering in synthetic intrinsically disordered proteins. *Biochim. Biophys. Acta Gen. Subj.* **1862**, 2204–2214 (2018).
62. Bianchi, G., Longhi, S., Grandori, R. & Brocca, S. Relevance of electrostatic charges in compactness, aggregation, and phase separation of intrinsically disordered proteins. *Int. J. Mol. Sci.* **21**, E6208 (2020).
63. Blocquel, D., Habchi, J., Gruet, A., Blangy, S. & Longhi, S. Compaction and binding properties of the intrinsically disordered C-terminal domain of Henipavirus nucleoprotein as unveiled by deletion studies. *Mol. Biosyst.* **8**, 392–410 (2012).
64. Mertens, H. D. T. & Svergun, D. I. Combining NMR and small angle X-ray scattering for the study of biomolecular structure and dynamics. *Arch. Biochem. Biophys.* **628**, 33–41 (2017).
65. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. *FEBS J.* **272**, 5129–5148 (2005).
66. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* **18**, 343–384 (2005).
67. Haynes, C. *et al.* Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol.* **2**, e100 (2006).
68. Rodriguez, J. J., Parisien, J. P. & Horvath, C. M. Nipah virus V protein evades alpha and gamma interferons by preventing STAT1 and STAT2 activation and nuclear accumulation. *J. Virol.* **76**, 11476–11483 (2002).
69. Rodriguez, J. J., Wang, L. F. & Horvath, C. M. Hendra virus V protein inhibits interferon signaling by preventing STAT1 and STAT2 nuclear accumulation. *J. Virol.* **77**, 11842–11845 (2003).
70. Ludlow, L. E., Lo, M. K., Rodriguez, J. J., Rota, P. A. & Horvath, C. M. Henipavirus V protein association with Polo-like kinase reveals functional overlap with STAT1 binding and interferon evasion. *J. Virol.* **82**, 6259–6271 (2008).
71. Longhi, S., Bloyet, L. M., Gianni, S. & Gerlier, D. How order and disorder within paramyxoviral nucleoproteins and phosphoproteins orchestrate the molecular interplay of transcription and replication. *Cell. Mol. Life Sci.* **74**, 3091–3118 (2017).
72. Uversky, V. N. & Longhi, S. *Flexible Viruses: Structural Disorder in Viral Proteins* (Wiley, New York, 2012).
73. Xue, B. *et al.* Structural disorder in viral proteins. *Chem. Rev.* **114**, 6880–6911 (2014).
74. Bugge, K. *et al.* Interactions by disorder—A matter of context. *Front. Mol. Biosci.* **7**, 110 (2020).
75. Olsen, J., Teilum, K. & Kragelund, B. Behaviour of intrinsically disordered proteins in protein–protein complexes with an emphasis on fuzziness. *Cell. Mol. Life Sci.* **74**, 3175–3183 (2017).
76. Keller, R. *The Computer Aided Resonance Assignment Tutorial* (Cantina Verlag, Goldau, 2004).
77. Bartels, C., Xia, T. H., Billeter, M., Guntert, P. & Wuthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **6**, 1–10 (1995).
78. Round, A. R. *et al.* Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33. *J. Appl. Crystallogr.* **41**, 913–917 (2008).
79. Franke, D. *et al.* ATSAS 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **50**, 1212–1225 (2017).
80. Guinier, A. La diffraction des rayons X aux tres petits angles; application a l'etude de phenomenes ultramicroscopiques. *Ann. Phys. (Paris)* **12**, 161–237 (1939).
81. Guinier, A. & Fournet, F. *Small Angle Scattering of X-rays* (Wiley Interscience, New York, 1955).
82. Svergun, D. Determination of the regularization parameters in indirect-trasform methods using perceptual criteria. *J. Appl. Cryst.* **25**, 495–503 (1992).
83. Durand, D. *et al.* NADPH oxidase activator p67(phox) behaves in solution as a multidomain protein with semi-flexible linkers. *J. Struct. Biol.* **169**, 45–53 (2010).
84. Tria, G., Mertens, H. D. T., Kachala, M. & Svergun, D. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCr* **2**, 202–217 (2015).
85. Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.* **43**, D357–363 (2015).
86. Lazar, T. *et al.* PED in 2021: A major update of the protein ensemble database for intrinsically disordered proteins. *Nucl. Acid Res.* (Database issue) in press, (2021).
87. Moore, B. L., Kelley, L. A., Barber, J., Murray, J. W., MacDonald, J. T. High-quality protein backbone reconstruction from alpha carbons using Gaussian mixture models. *J. Comput. Chem.* **34**, 1881–1889 (2013).

Acknowledgements

This work was carried out with the financial support of the CNRS and of the Fondazione CR Firenze to R.P. E. S. was supported by a joint doctoral fellowship from the Direction Générale de l'Armement (DGA) and Aix-Marseille University. The authors acknowledge the support of the Italian government programme “MIUR—Dipartimenti di Eccellenza 2018-2022” to the Department of Chemistry “Ugo Schiff” of the University of Florence. The authors acknowledge also the support and the use of resources of Instruct-ERIC, a Landmark ESFRI, through the R&D pilot scheme and iNEXT (H2020 Grant # 653706) for providing access to the CERM research infrastructure. The authors are thankful to Julien Perard for his help in SAXS data collection, the ESRF synchrotron for beamtime allocation and to Gerlind Sulzenbacher (AFMB lab) for efficiently managing the AFMB BAG.

Author contributions

S.L., R.P., and I.C.F. conceived and planned the experiments. E.S. purified and labeled the protein for NMR and SAXS experiments, and performed all the SAXS measurements. E.S. and M.G.M. acquired all the NMR spectra. E.S., M.G.M. and M.S. performed the sequence-specific assignment and analyzed the data under the guidance of R.P. and I.C.F. G.T. processed and analyzed SAXS data. All the authors analyzed the data. S.L. wrote the first draft of the paper, and all the authors contributed to the writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-76522-3>.

Correspondence and requests for materials should be addressed to I.C.F., R.P. or S.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

Supplementary Material

Ensemble description of the intrinsically disordered N-terminal domain of the Nipah virus P/V protein from combined NMR and SAXS

Marco Schiavina^{1,2§}, Edoardo Salladini^{4§}, Maria Grazia Murrari^{1,2§}, Giancarlo Tria^{2,3}, Isabella Felli^{1,2*}, Roberta Pierattelli^{1,2*} and Sonia Longhi^{4*}

¹Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy

²Department of Chemistry “Ugo Schiff”, University of Florence, Via della Lastruccia 3-13, 50019 Sesto Fiorentino, Italy

³Florence Center for Electron Nanoscopy (FloCEN), University of Florence, Via della Lastruccia 3-13, 50019 Sesto Fiorentino, Italy

⁴Aix Marseille Univ, CNRS, Architecture et Fonction des Macromolécules Biologiques (AFMB), UMR 7257, Marseille, France

*To whom correspondence should be sent:

felli@cerm.unifi.it

roberta.pierattelli@unifi.it

sonia.longhi@afmb.univ-mrs.fr; sonia.longhi@univ-amu.fr

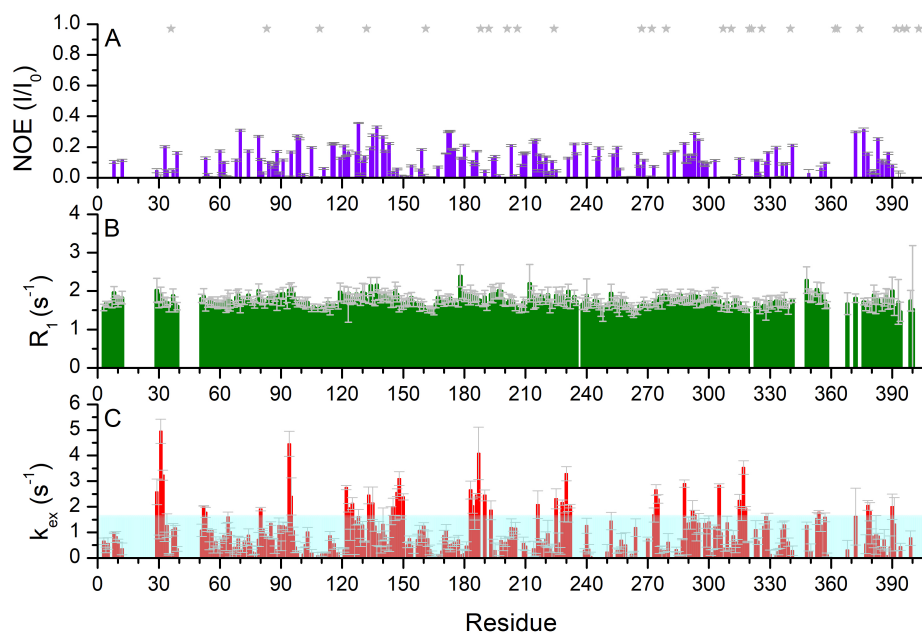
§These authors contributed equally to the work

Supplementary text

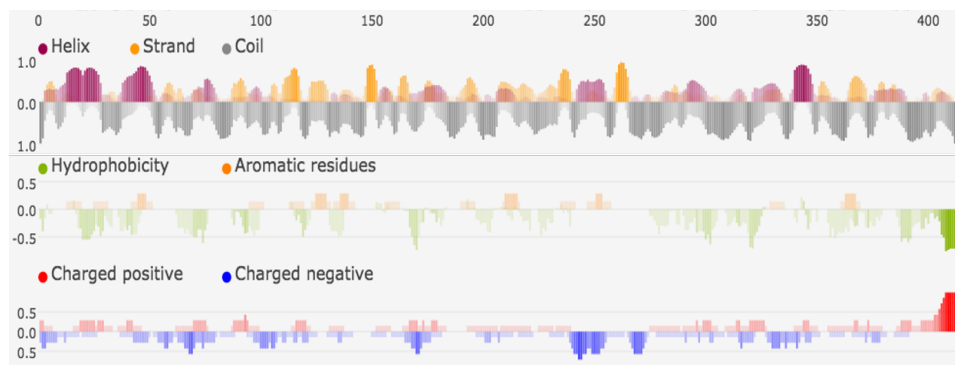
Discrepancy between experimentally observed R_g and R_g expected from Flory's power law

Flory's power law attempts at generalizing the IDP behavior while taking into account only sequence length and thus neglecting sequence specificities. Although Flory's power law remains a useful tool, the extent to which it can be generalized to all IDPs is a matter of debate, with several instances of deviations having been experimentally observed. Systematic collections of SAXS data on IDPs, such as the work of Cordeiro et al. [1], show a large scatter that indicates a lot of variation in scaling behavior. A similar study carried out Sosnick and co-workers [2] showed that several IDPs obeyed a scaling law with an exponent close to 0.6. Further, Sosnick and co-workers explained that the scaling law exponent reflects the quality of the solvent (aqueous buffer, sometimes also with denaturant) for a particular IDP and that solvent quality (and scaling exponent) can vary with the amino acid sequence. Consequently, it is not surprising to observe deviations from Flory's power law such as those we observed for NiV PNT.

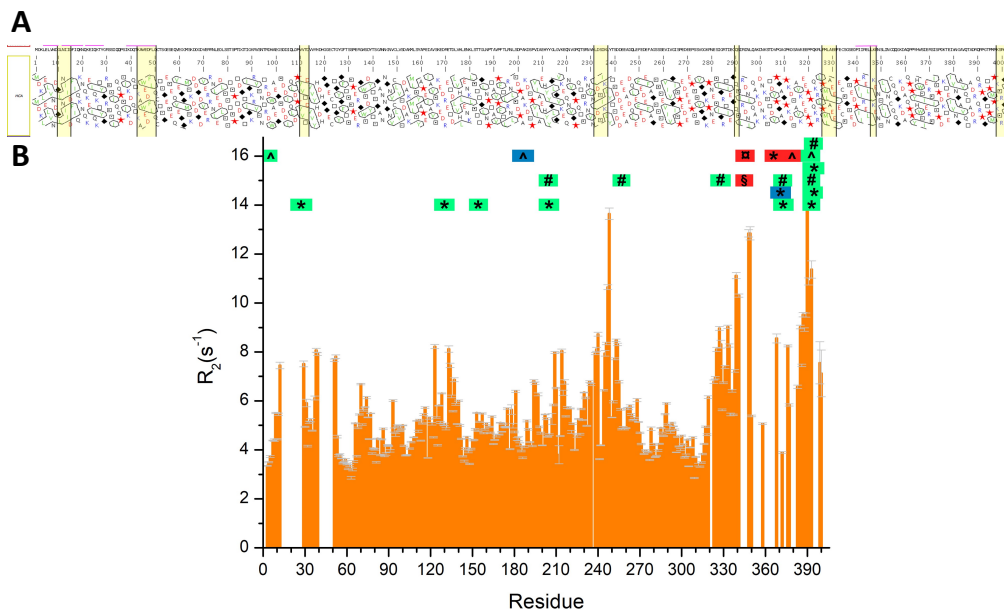
Supplementary Figures



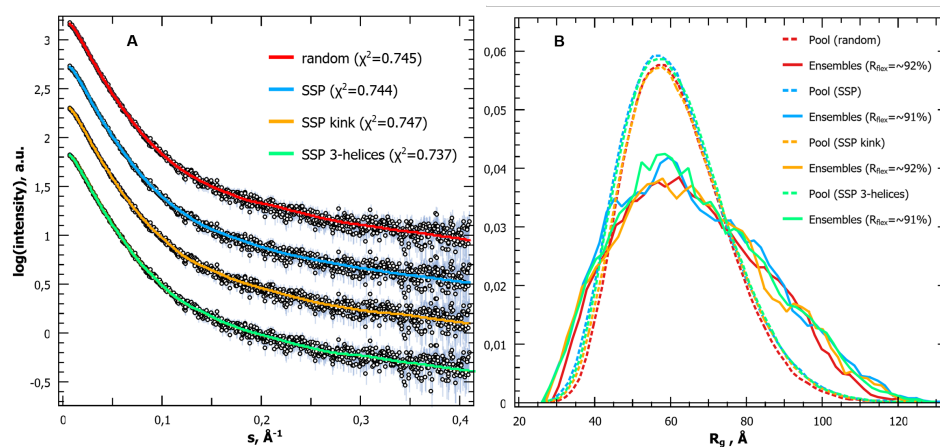
Supplementary Figure S1. (A) ^1H - ^{15}N -NOE measurements, (B) ^{15}N R_1 relaxation rates and (C) k_{ex} , as derived from CLEANEX-PM experiment using the approach reported in [3], of NiV PNT, carried out at 700 MHz and 288 K on a 100 μM ^{15}N -labeled sample. The light cyan box represents the modal value of 1.38 s^{-1} .



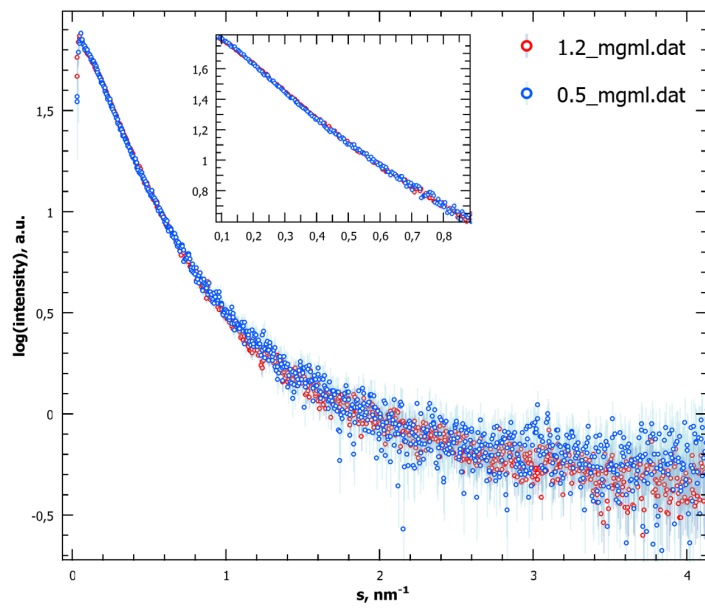
Supplementary Figure S2. Excerpt of the output generated by FELLS [4].



Supplementary Figure S3. (A) Hydrophobic cluster analysis (HCA) plot [5] of NiV PNT as obtained using MeDor [6]. Molecular Recognition Elements (MoREs), as predicted by MoRFPred [7] are shown in yellow. The α -helices, as predicted by PSIPRED [8], are shown as pink horizontal bars above the sequence. Four of them fall within the region encompassing the first 50 residues while the fifth one encompasses residues 340-348. (B) ^{15}N R_2 relaxation rates and putative Short Linear Motifs (SLiMs) as identified by the ELM database (<http://elm.eu.org/>) (see also **Supplementary Table S3**). The SLiMs are identified as follows according to the ELM classification: §, CLV (cleavage site); #, DOC (docking site); ^; LIG (ligand binding site); *, MOD (post-translational modification site), □, TRG (targeting site). The color code is representative of the conservation score (red < 0.95; blue 0.95-1; green 1).



Supplementary Figure S4. Modeling NiV PNT as a conformational ensemble. (A) Experimental scattering curve of NiV PNT and EOM 2.0 ensemble fits obtained using different initial pools as explained in **(B)**. **(B)** R_g distributions of the initial pools generated by: (red dot) EOM 2.0 (without restraints), (blue dot) Flexible-Meccano (with NMR secondary structure restraints), (orange dot) Flexible-Meccano (while imposing a kink centered at residue 20 providing NMR restraints), and (green dot) Flexible-Meccano (with α -helices imposed to residues 1-30). R_g distributions of the corresponding selected ensembles obtained using EOM 2.0 are shown as solid lines.



Supplementary Figure S5. Superimposition of the experimental scattering curves of NiV PNT collected at 0.5 mg/mL (*light blue*) and 1.2 mg/mL (*orange*). Inset: zoom on the superimposition at very low angle.

Supplementary Table S1. Acquisition parameters.

Experiments	Dimension of acquired data			Spectral width (ppm)			NS ^a	d1 (s) ^b
	t ₁	t ₂	t ₃	F ₁	F ₂	F ₃		
¹H detected								
¹ H- ¹⁵ N BEST-TROSY	1024 (¹⁵ N)	8192 (¹ H)		35	15		2	0.05
BT HNCO	224 (¹³ C)	112 (¹⁵ N)	4096(¹ H)	6	24	12	4	0.13
BT HN(CA)CO	224 (¹³ C)	112 (¹⁵ N)	4096(¹ H)	6	24	12	8	0.13
BT HNCACB	148 (¹³ C)	128 (¹⁵ N)	4096(¹ H)	34	25	12	4	0.20
BT HN(CO)CACB	220 (¹³ C)	128 (¹⁵ N)	4096(¹ H)	65	25	12	16	0.20
BT (H)N(COCA)NNH	134 (¹⁵ N)	134 (¹⁵ N)	4096(¹ H)	25	25	12	8	0.20
BT (H)N(CA)NNH	148 (¹⁵ N)	128 (¹⁵ N)	4096(¹ H)	25	25	12	16	0.20
¹³C detected								
CON	1264 (¹⁵ N)	1024 (¹³ C)		40	30		8	2.00
(H)CBCACON	164 (¹³ C)	128 (¹⁵ N)	1024 (¹³ C)	69	34	30	4	0.80
(H)CBCANCO	164 (¹³ C)	128(¹⁵ N)	1024 (¹³ C)	69	34	30	8	0.80
(HCA)COCON	108 (¹³ C)	128 (¹⁵ N)	1024 (¹³ C)	12	35	30	8	1.6
^a number of acquired scans								
^b relaxation delay in seconds								

Supplementary Table S2. Chemical shift (ppm) values for assigned residues of NiV PNT (aa 1-406) in 10 mM sodium phosphate buffer, pH 6.5, supplemented with 5 mM EDTA and 5 mM DTT at 288 K. ¹H resonances were calibrated with respect to the signal of 2,2-dimethylsilapentane-5-sulfonic acid (DSS). ¹³C chemical shifts were referred to external DSS. ¹⁵N chemical shifts were referred indirectly to the ¹³C standard using the conversion factor derived from the ratio of NMR frequencies [9].

Number	Type	H ^N	N	C ^α	C ^β	C ^γ
2	ASP			54.28	41.361	175.708
3	LYS	8.442	122.574	56.413	32.901	176.375
4	LEU	8.364	124.439	55.136	42.051	177.295
5	GLU	8.389	122.708	56.187	30.119	176.175
6	LEU	8.389	123.935	54.96	42.217	177.281
7	VAL	8.147	121.906	62.253	32.653	175.97
8	ASN	8.566	123.424	52.81	38.905	174.984
9	ASP	8.354	121.881	54.531	40.954	176.829
10	GLY	8.351	109.211	45.541	//	174.413
11	LEU	8.024	121.695	55.304	42.398	177.189
12	ASN	8.523	120.187	53.115	38.681	175.206
13	ILE	8.116	122.14	61.464	37.802	176.434
28	TYR			57.809	38.751	176.383
29	GLY	8.33	110.865	45.188	//	174.11
30	ARG	8.189	121.072	55.965	30.725	176.561
31	SER	8.447	117.458	58.295	63.755	174.65
32	SER	8.365	118.503	58.114	63.675	174.439
33	ILE	8.075	122.448	61.065	38.431	176.188
34	GLN	8.381	124.958	55.45	29.448	175.584
35	GLN	8.441	123.814	53.63	28.727	173.924
36	PRO	//	137.762	63.004	31.969	176.799
37	SER	8.512	116.99	58.12	63.879	174.891
38	ILE	8.25	123.527	61.297	38.488	176.108
39	LYS	8.233	125.953	55.974	32.744	175.928
40	ASP	8.19	124.36	53.578	41.887	174.951
50	GLN			55.797	28.988	176.216
51	CYS	8.251	120.433	58.71	28.988	175.145
52	THR	8.243	116.528	61.724	69.66	174.744
53	SER	8.322	118.652	58.433	63.755	175.115
54	GLY	8.434	111.549	45.136	//	174.328
55	GLU	8.301	121.182	56.599	30.209	176.802
56	SER	8.388	117.207	58.48	63.82	174.581
57	GLU	8.442	123.256	56.548	30.092	176.422
58	GLN	8.396	122.26	55.494	29.182	175.944
59	VAL	8.244	122.538	62.016	32.775	176.366
60	GLU	8.612	125.912	56.769	29.974	177.126
61	GLY	8.559	111.586	45.352	//	174.837
62	GLY	8.304	109.203	45.208	//	174.204

63	MET	8.262	120.25	55.294	32.929	176.391
64	SER	8.467	118.256	58.236	63.705	174.638
65	LYS	8.489	124.246	56.245	33.084	182.359
66	ASP		122.234	54.163	41.216	176.104
67	ASP	8.299	122.051	54.539	40.991	177.092
68	GLY	8.433	109.548	45.52	//	174.322
69	ASP	8.233	121.339	54.441	40.969	176.87
70	VAL	8.012	120.993	63.215	32.26	176.789
71	GLU	8.385	123.271	57.19	29.627	177.111
72	ARG	8.184	122.315	56.583	30.127	182.756
73	ARG		122.199	56.473	30.554	176.481
74	ASN	8.468	119.99	53.402	38.337	175.743
75	LEU	8.209	122.931	55.915	41.952	177.977
76	GLU	8.298	121.069	56.82	30.046	176.498
77	ASP	8.241	121.499	54.525	41.028	176.729
78	LEU	8.254	123.868	55.469	41.971	178.059
79	SER	8.317	116.578	59.127	63.656	174.955
80	SER	8.211	117.849	58.34	63.375	174.85
81	THR	8.111	115.844	61.47	69.597	174.487
82	SER	8.332	120.306	56.297	63.398	172.887
83	PRO	//	138.777	63.239	32.057	177.483
84	THR	8.273	113.976	61.61	69.423	174.658
85	ASP	8.226	123.041	54.233	41.078	176.925
86	GLY	8.425	110.271	45.575	//	174.758
87	THR	8.162	114.355	62.251	69.569	175.008
88	ILE	8.102	123.563	61.341	38.479	176.862
89	GLY	8.465	113.381	45.229	//	173.91
90	LYS	8.093	121.528	55.885	33.148	176.506
91	ARG	8.413	123.779	55.903	30.882	176.346
92	VAL	8.312	122.73	62.031	32.444	176.147
93	SER	8.425	120.081	57.842	63.784	174.248
94	ASN	8.562	121.952	53.026	38.694	175.427
95	THR	8.141	114.825	61.859	69.497	174.541
96	ARG	8.297	123.397	55.901	30.52	175.879
97	ASP	8.276	122.047	54.167	40.873	176.008
98	TRP	7.982	122.191	57.283	29.406	175.815
99	ALA	7.981	126.156	52.146	19.395	177.258
100	GLU	8.132	120.87	56.836	30.019	177.123
101	GLY	8.441	111.21	45.364	//	174.334
102	SER	8.133	115.975	58.149	63.729	174.402
103	ASP		122.637	54.237	41.132	175.987
104	ASP		121.228	54.21	40.965	176.102
105	ILE	7.986	121.751	61.046	38.75	176.063
106	GLN	8.446	125.596	55.073	29.221	181.626
107	LEU		125.209	54.615	42.472	176.888
108	ASP	8.358	123.461	52.441	40.517	174.337

109	PRO	//	136.753	62.929	31.973	176.774
110	VAL	8.235	121.564	62.47	32.349	176.515
111	VAL	8.291	125.514	62.194	32.596	176.31
112	THR	8.219	118.719	61.523	69.76	174.079
113	ASP	8.291	123.727	54.184	41.058	175.988
114	VAL	8.063	121.424	62.286	32.505	175.825
115	VAL	8.093	124.952	61.953	32.665	175.534
116	TYR	8.332	125.489	57.716	38.686	175.382
117	HIS	8.18	122.235	54.912	29.933	173.766
118	ASP		122.463	54.197	40.977	176.189
119	HIS		120.293	55.715	29.092	175.43
120	GLY		110.422	45.644	//	174.805
121	GLY	8.337	109.628	45.128	//	174.298
122	GLU	8.378	121.141	56.562	30.035	176.717
123	CYS	8.495	120.86	58.416	27.696	174.899
124	THR	8.195	116.622	61.906	69.593	174.995
125	GLY	8.323	111.484	45.006	//	173.84
126	TYR	8.107	120.694	58.131	38.632	176.408
127	GLY	8.298	111.183	45.086	//	173.776
128	PHE	8.004	120.424	57.757	39.569	176.018
129	THR	8.134	116.324	61.441	69.839	174.074
130	SER	8.271	118.775	57.984	63.801	174.043
131	SER	8.321	119.125	57.208	63.365	172.969
132	PRO	//	138.58	63.552	31.971	177.151
133	GLU		121.107	56.692	29.715	176.739
134	ARG		121.976	56.188	30.518	176.75
135	GLY	8.345	110.23	45.062	//	173.987
136	TRP	8.044	121.536	57.453	29.54	176.496
137	SER	8.047	117.617	58.185	63.917	173.765
138	ASP	8.092	122.307	54.197	40.889	176.225
139	TYR	8.118	121.266	58.319	38.378	176.408
140	THR	8.029	115.609	61.747	69.603	174.747
141	SER	8.205	118.459	58.505	63.51	175.212
142	GLY	8.32	111.424	45.369	//	174.075
143	ALA	8.091	124.024	52.638	19.115	177.698
144	ASN	8.421	117.972	53.21	38.677	175.158
145	ASN	8.329	119.571	53.112	38.572	175.775
146	GLY	8.371	109.267	45.449	//	174.003
147	ASN	8.237	119.241	53.204	38.857	175.255
148	VAL	8.081	120.933	62.161	32.721	175.884
149	CYS	8.468	124.257	58.321	27.867	174.393
150	LEU	8.423	126.314	55.016	42.172	177.312
151	VAL	8.158	121.967	62.078	32.816	176.156
152	SER	8.361	120.016	58.232	63.87	174.325
153	ASP	8.281	123.517	54.108	41.092	176.307
154	ALA	8.244	125.15	53.019	18.777	178.278

155	LYS	8.258	120.081	56.431	32.591	176.994
156	MET		120.976	55.322	32.37	176.359
157	LEU	8.138	123.547	55.142	42.129	177.225
158	SER	8.1	116.395	58.113	63.737	173.799
159	TYR	8.076	122.731	57.409	38.85	174.841
160	ALA	8.129	128.213	49.979	18.424	174.8
161	PRO	//	135.873	62.732	32.018	176.89
162	GLU	8.544	121.5	56.592	30.003	176.534
163	ILE	8.147	123.078	60.825	38.609	175.797
164	ALA	8.368	129.433	52.284	19.142	177.495
165	VAL	8.157	120.831	61.982	32.787	176.286
166	SER	8.52	121.029	57.819	63.919	174.625
167	LYS	8.541	124.676	56.553	32.953	182.756
168	GLU		122.028	56.938	29.824	176.417
169	ASP	8.333	122.327	54.431	40.982	176.362
170	ARG	8.246	121.621	56.839	30.41	176.738
171	GLU	8.484	122.042	56.968	29.889	177.057
172	THR	8.073	114.959	62.256	69.552	174.429
173	ASP	8.304	123.177	54.472	40.945	176.351
174	LEU	8.116	122.536	55.469	42.151	177.748
175	VAL	7.939	120.558	62.556	32.457	182.359
176	HIS		122.481	56.173	30.088	176.182
177	LEU	8.227	123.961	55.406	42.204	177.409
178	GLU	8.499	134.621	56.554	29.971	176.395
179	ASN	8.405	120.443	53.228	38.705	175.348
180	LYS	8.294	122.531	56.413	32.76	176.739
181	LEU	8.253	123.128	55.14	42.102	177.676
182	SER	8.3	116.992	58.095	63.522	175.206
183	THR	8.252	116.02	61.556	69.29	175.067
184	THR	8.101	116.057	62.007	69.558	175.195
185	GLY	8.38	111.739	45.067	//	173.865
186	LEU	8.057	121.71	54.623	42.355	176.968
187	ASN	8.527	121.172	51.204	38.689	173.492
188	PRO	//	137.262	63.39	32.134	177.22
189	THR	8.162	114.097	61.619	69.427	174.239
190	ALA	8.145	127.195	52.154	19.266	177.361
191	VAL	8.139	122.043	59.75	32.395	174.666
192	PRO	//	139.927	62.868	32.052	
193	PHE			58.154	39.197	175.924
194	THR	7.947	116.593	61.534	69.813	173.907
195	LEU	8.123	124.993	55	42.132	177.167
196	ARG		122.315	55.824	30.569	175.788
197	ASN	8.483	120.498	52.924	38.577	175.346
198	LEU	8.306	123.645	55.305	41.896	177.464
199	SER	8.203	116.281	58.297	63.768	173.83
200	ASP	8.227	123.903	52.066	41.146	174.637

201	PRO	//	138.145	63.541	32.057	177.106
202	ALA	8.36	123.379	52.362	18.723	178.252
203	LYS	8.041	120.228	56.301	32.941	176.411
204	ASP	8.206	121.126	54.148	41.014	175.918
205	SER	8.128	117.61	56.399	63.239	172.63
206	PRO	//	138.517	62.91	31.969	176.778
207	VAL		121.35	62.453	32.429	176.308
208	ILE	8.242	126.511	60.554	38.278	175.849
209	ALA	8.36	129.238	52.332	19.222	177.484
210	GLU	8.321	120.645	56.528	30.149	176.248
211	HIS		119.796	55.277	29.83	174.076
212	TYR	8.12	122.285	57.706	38.72	175.408
213	TYR	8.227	123.021	57.803	38.661	176.068
214	GLY	7.733	109.975	45.072	//	173.918
215	LEU	8.105	121.831	55.214	42.176	178.114
216	GLY	8.497	110.534	45.13	//	174.137
217	VAL	7.925	120.037	62.28	32.502	176.389
218	LYS	8.42	125.601	56.192	32.93	176.609
219	GLU	8.42	122.815	56.832	30.151	176.342
220	GLN	8.384	121.653	55.771	29.441	175.633
221	ASN	8.539	120.656	53.27	38.805	174.975
222	VAL	8.118	120.234	62.007	32.77	176.279
223	GLY	8.337	112.992	44.404	//	171.87
224	PRO	//	134.827	63.163	31.958	177.368
225	GLN	8.643	121.26	55.782	29.262	176.419
226	THR	8.19	115.827	61.866	69.78	174.662
227	SER	8.355	118.693	58.325	63.645	174.438
228	ARG	8.366	123.369	55.907	30.727	175.904
229	ASN	8.481	120.856	53.055	38.566	175.292
230	VAL	8.11	120.821	62.247	32.704	175.725
231	ASN	8.485	122.351	53.13	38.61	175.54
232	LEU	8.284	123.655	55.533	42.087	177.467
233	ASP		120.594	54.73	40.903	176.556
234	SER	8.039	115.772	58.502	63.843	174.554
235	ILE	7.957	122.749	61.256	38.502	176.443
236	LYS		125.929	56.171	32.783	175.947
237	LEU		124.132	54.907	42.494	176.799
238	TYR	8.189	121.293	57.449	38.949	175.831
239	THR	8.178	116.371	61.064	70.108	174.297
240	SER	8.372	118.267	58.283	63.864	174.407
241	ASP	8.446	122.59	54.318	41.053	175.912
242	ASP	8.216	120.929	54.193	40.898	176.093
243	GLU		115.527	56.736	30.163	176.859
244	GLU	8.362	121.914	56.55	29.963	176.555
245	ALA	8.194	125.165	52.662	19.224	177.8
246	ASP	8.295	119.971	54.435	40.834	176.377

247	GLN	8.237	120.773	55.638	29.159	175.973
248	LEU	8.219	123.532	55.381	42.183	175.98
249	GLU		121.615	56.592	30.067	176.245
250	PHE	8.12	120.831	57.649	39.701	176.345
251	GLU		123.087	56.241	30.433	175.879
252	ASP	8.36	122.361	54.18	41.256	176.558
253	GLU	8.494	122.68	57.046	29.876	176.421
254	PHE	8.299	120.735	57.481	38.972	175.876
255	ALA	8.049	126.075	52.907	19.009	178.168
256	GLY	8.013	108.329	45.283	//	174.31
257	SER	8.199	116.132	58.444	64.075	174.836
258	SER	8.469	118.473	58.391	63.839	174.702
259	SER	8.359	118.399	58.316	63.747	174.362
260	GLU	8.345	123.4	56.413	30.456	176.191
261	VAL	8.19	122.998	62.454	32.674	175.871
262	ILE	8.361	127.308	60.648	38.301	176.184
263	VAL	8.348	126.627	62.079	32.587	176.428
264	GLY	8.466	113.609	45.007	//	173.625
265	ILE	8.038	120.671	60.661	39.021	176.301
266	SER	8.644	122.987	56.287	63.387	172.987
267	PRO	//	138.655	63.565	31.962	177.398
268	GLU	8.536	120.422	56.898	29.787	176.473
269	ASP	8.185	121.131	54.48	41.265	176.1
270	GLU	8.103	121.375	56.026	30.788	176.267
271	GLU	8.46	124.449	54.247	29.68	174.751
272	PRO	//	138.513	63.119	31.871	177.237
273	SER	8.505	116.413	58.255	63.576	174.978
274	SER	8.367	118.604	58.161	63.639	174.779
275	VAL	8.135	121.976	62.438	32.359	176.93
276	GLY	8.5	113.043	45.147	//	174.69
277	GLY	8.219	109.11	45.031	//	173.825
278	LYS	8.305	122.451	53.918	32.54	174.397
279	PRO	//	137.988	63.078	31.978	176.878
280	ASN	8.6	119.42	53.381	38.541	175.537
281	GLU	8.404	122.217	56.565	30.215	176.547
282	SER	8.383	117.274	58.352	63.567	174.74
283	ILE	8.062	123.031	61.51	38.573	176.892
284	GLY	8.49	113.129	45.253	//	174.012
285	ARG	8.123	121.187	55.895	31.165	176.542
286	THR	8.325	117.044	61.922	69.972	174.632
287	ILE	8.352	124.31	61.107	38.227	176.442
288	GLU	8.567	125.934	56.889	29.936	177.205
289	GLY	8.497	110.783	45.439	//	174.398
290	GLN	8.173	120.418	55.934	29.487	176.185
291	SER	8.456	117.814	58.628	63.564	174.811
292	ILE	8.183	123.425	61.611	38.451	176.599

293	ARG	8.296	124.815	56.477	30.804	176.301
294	ASP	8.266	121.597	54.505	40.947	176.235
295	ASN	8.324	119.432	53.37	38.471	175.657
296	LEU	8.222	122.403	55.598	41.725	177.737
297	GLN	8.221	120.662	55.717	29.142	175.956
298	ALA	8.209	125.683	52.578	19.025	177.973
299	LYS	8.268	120.942	56.287	32.899	176.389
300	ASP	8.279	121.454	53.914	41.073	175.989
301	ASN	8.332	120.182	53.089	38.627	175.386
302	LYS	8.347	122.053	56.266	32.864	176.951
303	SER	8.374	117.573	58.363	63.514	174.965
304	THR	8.181	115.721	61.521	69.42	174.268
305	ASP	8.275	123.777	54.133	40.935	175.82
306	VAL	8.145	122.636	59.954	32.581	174.563
307	PRO	//	140.658	63.311	31.913	177.724
308	GLY	8.545	110.626	45.027	//	173.927
309	ALA	8.13	124.251	52.218	19.635	178.011
310	GLY	8.341	109.241	44.396	//	171.878
311	PRO	//	135.131	62.888	32.075	177.451
312	LYS	8.53	122.165	56.075	32.816	176.624
313	ASP	8.329	122.028	54.255	41.265	176.374
314	SER	8.249	117.02	58.483	63.601	174.31
315	ALA	8.307	126.509	52.637	19.096	177.747
316	VAL	8.048	120.444	62.144	32.318	176.096
317	LYS	8.424	126.83	55.866	33.026	176.178
318	GLU		123.828	55.971	30.518	176.086
319	GLU		125.158	54.116	29.75	174.056
320	PRO	//	139.197	61.246	30.654	174.812
321	PRO	//	135.805	62.852	32.126	176.885
322	GLN		121.633	55.777	29.431	176.061
323	LYS		123.688	56.06	33.064	176.194
324	ARG		123.635	55.638	30.928	175.959
325	LEU		126.166	55.409	41.284	174.872
326	PRO	//	138.636	63.001	31.817	176.668
327	MET	8.395	120.954	55.478	33.111	176.129
328	LEU	8.299	124.244	54.885	42.33	177.107
329	ALA	8.338	125.304	52.773	19.083	177.961
330	GLU	8.427	120.055	56.837	30.035	176.641
331	GLU	8.293	121.607	57.258	29.999	176.252
332	PHE	8.293	120.807	57.511	39.718	175.732
333	GLU	8.318	122.971	56.416	30.171	176.216
334	CYS	8.418	121.018	58.283	27.958	174.819
335	SER	8.575	119.309	58.65	63.71	175.203
336	GLY	8.526	111.966	45.252	//	174.29
337	SER	8.183	115.84	58.538	63.919	174.546
338	GLU	8.51	123.099	56.333	30.239	175.924

339	ASP	8.329	124.294	52.309	41.455	174.957
340	PRO	//	137.86	63.873	32.145	177.714
341	ILE		120.437	62.092	32.785	
347	LYS			57.066	32.344	177.426
348	GLU		121.531	57.156	29.793	176.922
349	ASN	8.381	119.576	53.578	38.587	175.692
350	SER	8.201	116.632	59.077	63.517	174.87
351	LEU	8.15	123.758	55.57	41.94	177.8
352	ILE	8.15	120.115	61.369	38.281	176.266
353	ASN	8.375	122.042	53.274	38.504	175.416
354	CYS	8.247	119.98	58.757	27.778	174.831
355	GLN	8.463	122.809	56.061	29.038	176.025
356	GLN		121.704	56.073	29.413	176.554
357	GLY		110.906	45.12	//	176.553
358	LYS		121.369	56.494	32.831	176.598
359	ASP		121.099	54.189	40.889	175.785
360	ALA		124.36	52.205	19.143	177.426
361	GLN		121.178	53.258	28.868	173.577
362	PRO	//	139.123	61.159	30.659	174.715
363	PRO	//	135.238	61.311	30.542	
367	SER			58.121	63.607	174.359
368	ILE	8.16	122.844	60.766	38.684	176.047
369	GLU	8.394	121.674	56.045	29.129	
371	SER			61.897	70.026	174.335
372	ILE		122.611	60.71	38.706	176.069
373	SER		121.749	55.827	62.993	173.044
374	PRO	//	138.261	63.266	31.913	176.633
375	ASP		120.308	54.029	40.826	176.37
376	LYS	8.26	122.211	56.16	32.668	176.704
377	THR	8.18	116.072	62.234	69.633	174.497
378	GLU	8.417	124.137	56.251	30.283	176.112
379	ILE	8.244	123.608	60.887	38.202	176.38
380	VAL	8.32	126.034	62.061	32.675	175.995
381	ASN	8.664	123.963	53.346	38.59	175.767
382	GLY	8.402	109.942	45.222	//	173.745
383	ALA	8.064	124.1	52.191	19.337	177.781
384	VAL		120.325	62.272	32.571	176.375
385	GLN	8.578	125.355	55.493	29.494	176.174
386	THR	8.238	116.61	61.775	69.659	174.576
387	ALA	8.407	126.586	52.64	19.048	177.603
388	ASP	8.304	119.877	54.336	40.89	176.435
389	ARG	8.171	121.551	56.063	30.367	176.272
390	GLN	8.344	121.252	55.626	29.136	175.817
391	ARG	8.359	124.111	53.946	29.925	174.248
392	PRO	//	138.018	63.215	32.101	177.458
393	GLY	8.561	110.186	44.986	//	174.042

394	THR	8.033	117.341	59.909	69.96	172.713
395	PRO	//	139.659	62.975	32.109	176.659
396	MET	8.315	121.560	54.025	32.318	174.806
397	PRO	//	137.716	63.258	32.026	
398	LYS			56.289	32.818	176.829
399	SER	8.355	117.483	58.222	63.698	174.563
400	ARG	8.487	123.745	56.05	30.658	176.627
401	GLY	8.393	110.215	44.887	//	
402	ILE			52.868	41.367	175.221
403	PRO	//	136.107	62.932	31.897	

Supplementary Table S3. Predicted Short Linear Motifs (SLiMs) with a conservation score ≥ 0.88 as provided by the ELM database (see also **Supplementary Figure S3B**). The full report is available at:

http://elm.eu.org/cgimodel.py?fun=smartResult&userId=QiKRrpLIDz&EXPECT_CUTOFF=100&r=1&bg=on

SLiM Code	Type	Residues	Conservation Score
LIG_WD40_WDRS_VDV_2	DKEL	2-6	1.00
MOD_GSK3_1	IQKTYGRS	24-31	1.00
MOD_CK2_1	GFTSSPE	127-133	1.00
MOD_SUMO_rev_2	SDAKML	152-157	1.00
LIG_WD40_WDRS_VDV_2	TTGLNPTA	183-190	0.97
DOC_WW_Pin1_4	AKDSPY	202-207	1.00
MOD_Pro_Dkin_1	AKDSPVI	202-208	1.00
DOC_USP7_MATH_1	AGDSSS	255-259	1.00
DOC_PP2A_B56_1	LPMLAEE	325-331	1.00
CLV_PCSK_SKI1_1	RELLK	343-347	0.90
TRG_Pf_PMV_PEXEL_1	RELLKE	343-348	0.90
MOD_GSK3_1	YHWSIERS	364-371	0.88
MOD_PIK_1	IERSISP	368-374	0.97
DOC_WW_Pin1_4	RSISPD	370-375	1.00
MOD_Pro_Dkin_1	RSISPDK	370-376	1.00
LIG_FHA_1	DKTEIVN	375-381	0.88
LIG_SH3_3	RQRPGTP	389-395	1.00
DOC_WW_Pin1_4	RPGTPM	391-395	1.00
MOD_DYRK1A_RPXSP_1	RPGTP	391-395	1.00
MOD_Pro_Dkin_1	RPGTPMP	391-397	1.00
MOD_CDK_CPxxK_3	RPGTPMPK	391-398	1.00
DOC_CKS1_1	PGTPMP	392-397	1.00

Supplementary Table S4. SAXS data-collection and scattering-derived parameters.

Data-collection parameters	
Data source and instrument (beamline)	ESRF (BM29)
Instrument (detector)	PILATUS 1M
Beam geometry (mm ²)	0.7 x 0.7
Wavelength (Å)	0.992
Detector distance (m)	2.847
<i>s</i> range (nm ⁻¹)	0.028-4.525
Exposure time (sec)	10 (10 frames x 1.00sec)
Concentration (mg ml ⁻¹)	1.2
Temperature (K)	293
Structural parameters	
<i>I</i> (0) (cm ⁻¹) [from <i>P</i> (<i>r</i>)]	69.53±0.1
<i>R</i> _g (nm) [from <i>P</i> (<i>r</i>)]	6.36±0.2
<i>I</i> (0) (cm ⁻¹) [from Guinier approximation]	69.62±0.7
<i>R</i> _g (nm) [from Guinier approximation]	6.17±0.12
<i>D</i> _{max} (nm)	23±1
Molecular-mass determination	
Molecular mass <i>M</i> _r (kDa) [from <i>water calibration</i>]	~47
Calculated monomeric <i>M</i> _r (kDa) [from <i>sequence</i>]*	~45
Software employed	
Primary data reduction	PIPELINE
Data processing	ATSAS 3.0.1
Flexibility assessment	EOM 2.0

* as obtained using the ProtParam tool of the expasy server (<https://web.expasy.org/protparam/>)

References

- 1 Cordeiro, T.N., Herranz-Trillo, F., Urbanek, A., Estaña, A., Cortés, J., Sibille, N. & Bernadó, P. Structural characterization of highly flexible proteins by small-angle scattering. *Adv. Exp. Med. Biol.* 1009,107-129, (2017).
- 2 Riback, J.A., Bowman, M.A., Zmyslowski, A.M., Knoverek, C.R., Jumper, J.M., Hinshaw, J.R., Kaye, E.B., Freed, K.F., Clark, P.L. & Sosnick, T.R. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water *Science* 358, 238-241, (2017).
- 3 Hwang, T. L., van Zijl, P. C. & Mori, S. Accurate quantitation of water-amide proton exchange rates using the phase-modulated CLEAN chemical EXchange (CLEANEX-PM) approach with a Fast-HSQC (FHSQC) detection scheme. *J. Biomol. NMR* 11, 221-226, (1998).
- 4 Piovesan, D., Walsh, I., Minervini, G. & Tosatto, S. C. E. FIELDS: fast estimator of latent local structure. *Bioinformatics* 33, 1889-1891, (2017).
- 5 Callebaut, I. *et al.* Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell. Mol. Life Sci.* 53, 621-645., (1997).
- 6 Lieutaud, P., Canard, B. & Longhi, S. MeDor: a metasever for predicting protein disorder. *BMC Genomics* 9, S25, (2008).
- 7 Disfani, F. M. *et al.* MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* 28, i75-83, (2012).
- 8 Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292, 195-202, (1999).
- 9 Markley, J. L. *et al.* Recommendations for the presentation of NMR structures of proteins and nucleic acids--IUPAC-IUBMB-IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. *Eur. J. Biochem.* 256, 1-15, (1998).

11. The Highly flexible disordered regions of the SARS-CoV-2 nucleocapsid N protein within the 1-248 residue construct: sequence-specific resonance assignments through NMR

Biomolecular NMR Assignments (2021) 15:219–227
<https://doi.org/10.1007/s12104-021-10009-8>

ARTICLE



The highly flexible disordered regions of the SARS-CoV-2 nucleocapsid N protein within the 1–248 residue construct: sequence-specific resonance assignments through NMR

Marco Schiavina^{1,2} · Letizia Pontoriero^{1,2} · Vladimir N. Uversky³ · Isabella C. Felli^{1,2} · Roberta Pierattelli^{1,2}

Received: 10 December 2020 / Accepted: 15 February 2021 / Published online: 3 March 2021
© The Author(s) 2021

Abstract

The nucleocapsid protein N from SARS-CoV-2 is one of the most highly expressed proteins by the virus and plays a number of important roles in the transcription and assembly of the virion within the infected host cell. It is expected to be characterized by a highly dynamic and heterogeneous structure as can be inferred by bioinformatics analyses as well as from the data available for the homologous protein from SARS-CoV. The two globular domains of the protein (NTD and CTD) have been investigated while no high-resolution information is available yet for the flexible regions of the protein. We focus here on the 1–248 construct which comprises two disordered fragments (IDR1 and IDR2) in addition to the N-terminal globular domain (NTD) and report the sequence-specific assignment of the two disordered regions, a step forward towards the complete characterization of the whole protein.

Keywords SARS-CoV-2 · Covid-19 · Nucleocapsid protein · NMR spectroscopy · ¹³C detection · IDPs



The highly flexible disordered regions of the SARS-CoV-2 nucleocapsid N protein within the 1–248 residue construct: sequence-specific resonance assignments through NMR

Marco Schiavina^{1,2} · Letizia Pontoriero^{1,2} · Vladimir N. Uversky³ · Isabella C. Felli^{1,2} · Roberta Pierattelli^{1,2}

Received: 10 December 2020 / Accepted: 15 February 2021 / Published online: 3 March 2021
© The Author(s) 2021

Abstract

The nucleocapsid protein N from SARS-CoV-2 is one of the most highly expressed proteins by the virus and plays a number of important roles in the transcription and assembly of the virion within the infected host cell. It is expected to be characterized by a highly dynamic and heterogeneous structure as can be inferred by bioinformatics analyses as well as from the data available for the homologous protein from SARS-CoV. The two globular domains of the protein (NTD and CTD) have been investigated while no high-resolution information is available yet for the flexible regions of the protein. We focus here on the 1–248 construct which comprises two disordered fragments (IDR1 and IDR2) in addition to the N-terminal globular domain (NTD) and report the sequence-specific assignment of the two disordered regions, a step forward towards the complete characterization of the whole protein.

Keywords SARS-CoV-2 · Covid-19 · Nucleocapsid protein · NMR spectroscopy · ¹³C detection · IDPs

Biological context

Coronaviruses (CoVs) are relatively large viruses containing a single-stranded positive-sense RNA genome encapsulated within a membrane envelope (Cui et al. 2019). There are four classes of CoVs, called α , β , γ , and δ , with the class β -coronavirus including CoVs that can infect humans, such as the severe acute respiratory syndrome virus (SARS-CoV),

the Middle East respiratory syndrome virus (MERS-CoV), and the COVID-19 causative agent SARS-CoV-2 (Masters 2006; Surjit and Lal 2008). Similar to SARS-CoV and MERS-CoV, SARS-CoV-2 attacks the lower respiratory system causing viral pneumonia, but it may also affect the gastrointestinal system, heart, kidney, liver, and central nervous system leading to multiple organ failure (Huang et al. 2020; Wang et al. 2020). The severe rate of this virus spread, based on its unexpectedly high infectivity, demands rapid action towards both the development of a vaccine and potent viral inhibitors to weaken or eliminate major life-threatening symptoms.

The SARS-CoV-2 nucleocapsid protein N is a structurally heterogeneous, 419 amino-acid-long, multidomain RNA-binding protein that is found inside the viral envelope (Fig. 1). This protein, as already established for its SARS-CoV homologue, stabilizes viral RNA by forming a ribonucleoprotein complex (RNP) and plays a fundamental role in the transcription and assembly of the virion once the host cell is infected (Chang et al. 2009, 2014). The self-association of the N protein is also responsible for the formation of a shell, the capsid, which protects the genetic material from external agents. The N protein includes two functional domains known as N- and C-terminal domains, or NTD and CTD respectively, that are

Marco Schiavina and Letizia Pontoriero have contributed equally.

✉ Isabella C. Felli
felli@cerm.unifi.it

✉ Roberta Pierattelli
roberta.pierattelli@unifi.it

¹ Magnetic Resonance Center – CERM, University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, FI, Italy

² Department of Chemistry “Ugo Schiff”, University of Florence, Via della Lastruccia 3-13, 50019 Sesto Fiorentino, FI, Italy

³ Department of Molecular Medicine and USF Health Byrd Alzheimer’s Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA

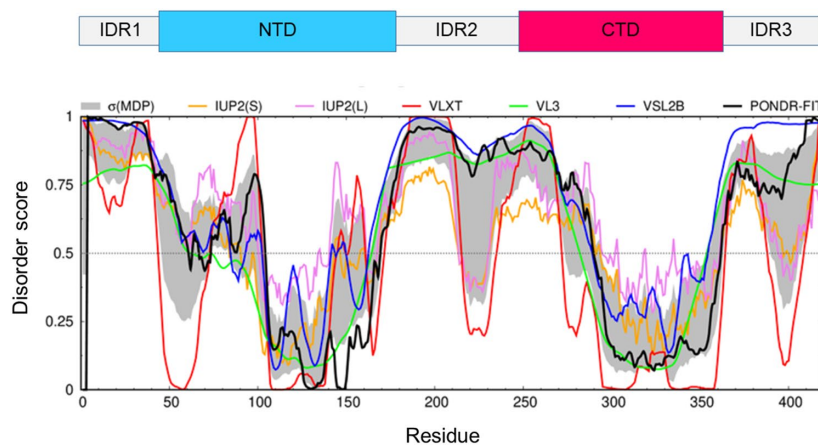


Fig. 1 Bioinformatics analysis of the intrinsic disorder predisposition of the SARS-CoV-2 nucleocapsid N protein obtained using IUPred short (golden line), IUPred long (purple line), PONDNR@ VLXT (red line), PONDNR@ VL3 (green line), PONDNR@ VSL2B (blue line), PONDNR@ FIT (black line). The gray shadow region signifies the error distribution $\sigma(\text{MDP})$ around the mean disorder profile

responsible for RNA binding (NTD) and homo-dimerization (CTD) (Chang et al. 2006). Bioinformatics analysis predicts the presence of three long intrinsically disordered regions in the polypeptide chain as reported in Fig. 1 (Giri et al. 2020). These regions are believed to be responsible for an intricate mechanism that leads to the regulation of the formation of the RNP complex. They are also engaged in many interactions with other viral proteins or host proteins, as was already demonstrated for the homologous nucleocapsid protein of the CoV that causes SARS (Chang et al. 2014; Giri et al. 2020). To date there is no structural and dynamic information with atomic resolution for the entire N protein due to its highly disordered nature. The structures of the globular NTD and CTD domains have been determined (Kang et al. 2020; Peng et al. 2020; Dinesh et al. 2020). However, there is no atomic resolution information on the disordered parts of this protein. On the other hand, the role of disorder is not accidental and is very relevant for the modulation of the mechanisms leading to the infection (Goh et al. 2012, 2013). In addition, the N proteins of the different variants of CoVs seem to be genetically stable (Giri et al. 2020), which makes them excellent candidates for developing antiviral therapies that have not been explored to date.

In this frame, we provide here the backbone assignment of the two disordered regions flanking the NTD, the N-terminal IDR1 and the serine-rich disordered region IDR2, in the 1–248 residue construct (IDR1-NTD-IDR2). These data will contribute to the efforts of the research consortium covid19-nmr (www.covid19-nmr.de) enabling follow-up

applications, such as residue-resolved drug screening and interaction mapping.

Methods and experiments

Construct design

This study uses the SARS-CoV-2 NCBI reference genome entry NC_045512.2, identical to GenBank entry MN908947.3. The definition of domain boundaries for the IDR1-NTD-IDR2 fragment (1–248) was guided by the SARS-CoV homologue (Chang et al. 2014).

A codon-optimized expression construct of SARS-CoV-2 IDR1-NTD-IDR2 inserted into the pET29b(+) plasmid was obtained from Twist Bioscience.

Sample preparation

Uniformly ^{13}C , ^{15}N -labelled IDR1-NTD-IDR2 protein was expressed in *E. coli* strain BL21 (DE3). The culture was grown in 1 L LB medium at 37 °C until OD_{600} reached 0.8, then transferred in 250 mL of labelled minimal medium (4x) containing 0.25 g/L $^{15}\text{NH}_4\text{Cl}$ (Cambridge Isotope Laboratories), 0.75 g/L $[\text{U}]^{13}\text{C}_6\text{-D-glucose}$ (Eurisotop). After 1 h of metabolite clearance, the culture was induced with 0.2 mM isopropyl-beta-thiogalactopyranoside (IPTG) at 18 °C for 16/18 h.

The cell pellet was resuspended in 25 mM 2-Amino-2-(hydroxymethyl)-1,3-propanediol (TRIS), 1.0 M sodium

chloride, 5% glycerol, DNase, RNase and 500 μ L of 100 \times stock of protease inhibitor cocktail (SIGMA) at pH 8.

Cells were disrupted by sonication. The supernatant was cleared by centrifugation (50', 30,000 \times g, 4 $^{\circ}$ C), then the cleared supernatant was dialyzed overnight at 4 $^{\circ}$ C into 25 mM TRIS pH 7.2 (binding buffer).

The protein was purified with ion-exchange chromatography using an HiTrap SP FF 5 mL column and a 70% gradient of 25 mM TRIS, 1 M NaCl pH 7.2. Fractions containing pure protein were pooled and concentrated using 15 mL and 0.5 mL Centricon centrifugal concentrators (MW cut-off 10 kDa).

Final NMR samples were 280 μ M IDR1-NTD-IDR2, 25 mM TRIS pH 6.5, 450 mM sodium chloride, 0.02% NaN₃, 5% (v/v) D₂O in water.

NMR experiments

All the NMR experiments were acquired at 298 K. Carbon-13 direct detected NMR experiments were acquired on a 16.4 T Bruker AVANCE NEO spectrometer operating at 700.06 MHz ¹H, 176.05 MHz ¹³C, and 70.97 MHz ¹⁵N frequencies, equipped with a 5 mm cryogenically cooled probehead optimized for ¹³C direct detection (TXO). Proton direct detected NMR experiments were acquired on a 28.3 T Bruker AVANCE NEO spectrometer operating at 1200.85 MHz ¹H, 301.97 MHz ¹³C, and 121.70 MHz ¹⁵N equipped with a 3 mm cryogenically cooled triple-resonance probehead (TCI).

Backbone assignment was performed by analyzing 2D and 3D ¹H and ¹³C direct detected experiments. In particular, 2D-[¹H, ¹⁵N]-HSQC, 2D-[¹H, ¹⁵N]-BEST-TROSY (BT), 2D-CON, 2D-(H)CACO and 2D-(H)CBCACO experiments were performed. Moreover, a series of 3D experiments were acquired: 3D-(H)CBCACON, 3D-(H)CBCANCO, 3D-BT-HNCACB, and 3D-BT-HN(CO)CACB. To compare the resonance values obtained through the carbon detected spectra with the ones obtained with the proton detected ones, 3D-HNCO and 3D-HN(CA)CO were also collected.

All the 2D-¹³C detected experiments were acquired in a version optimized for the detection of the highly flexible regions of the protein (Felli and Pierattelli 2012). Carbon-13 homonuclear decoupling was achieved through the IPAP virtual decoupling approach (Bermel et al. 2006a). 2D-(H)CACO and 2D-(H)CBCACO exploit constant-time evolution in the indirect dimension (Pontoriero et al. 2020). The 2D-CON was acquired both with the ¹³C start variant (Bermel et al. 2006b) as well as with the 2D-(HCA)CON variant (Bermel et al. 2009) to ensure direct detection of proline ¹⁵N resonances. 3D-(H)CBCACON and 3D-(H)CBCANCO (Bermel et al. 2009) were acquired with high resolution in all detected dimensions. Most relevant acquisition parameters are reported in Table 1.

Pulse lengths and carrier frequencies generally used for triple resonance experiments were used for the ¹³C detected experiments and are summarized hereafter. The ¹H carrier was placed at 4.7 ppm for non-selective hard pulses. ¹³C pulses were given at 176.7 ppm, 55.9 ppm, and 45.7 ppm for C', C $^{\alpha}$ and C $^{\text{ali}}$ regions, respectively.

Table 1 Experimental parameters used to collect the NMR experiments

Experiments	Dimension of acquired data			Spectral width (ppm)			NS ^a	d1 + aq (s) ^b	Spectrometer frequency (¹ H) (MHz)
	t1	t2	t3	F1	F2	F3			
¹ H detected									
¹ H- ¹⁵ N BEST-TROSY	512 (¹⁵ N)	9676 (¹ H)		41	15		16	0.5	1200
BT-HNCACB	96 (¹³ C)	90 (¹⁵ N)	6144 (¹ H)	75	41	14	96	0.2	1200
BT-HN(CO)CACB	96 (¹³ C)	80 (¹⁵ N)	6144 (¹ H)	75	41	14	96	0.2	1200
HN(CA)CO	128 (¹³ C)	128 (¹⁵ N)	4096 (¹ H)	7	28	18	8	1.0	1200
HNCO	128 (¹³ C)	220 (¹⁵ N)	4096 (¹ H)	7	28	18	4	1.0	1200
¹³ C detected									
CON	512 (¹⁵ N)	1024 (¹³ C)		34	31		32	1.6	700
(HCA)CON	220 (¹⁵ N)	1024 (¹³ C)		40	31		16	0.9	700
(H)CACO	330 (¹³ C)	1024 (¹³ C)		34	30		32	1.0	700
(H)CBCACO	476 (¹³ C)	1024 (¹³ C)		59	30		32	1.0	700
(H)CBCACON	128 (¹³ C)	96 (¹⁵ N)	1024 (¹³ C)	58	34	30	4	1.0	700
(H)CBCANCO	96 (¹³ C)	96 (¹⁵ N)	1024 (¹³ C)	58	34	30	16	1.0	700

^aNumber of acquired scans

^bRelaxation delay (acquisition time plus recovery delay d1)

^{15}N pulses were given at 124.0 ppm. Q5 and Q3 shapes (Emsley and Bodenhausen 1990) of durations of 300 and 231 μs , respectively, were used for ^{13}C band-selective $\pi/2$ and π flip angle pulses except for the π pulses that should be band-selective on the C^α region (Q3, 900 μs), and for the adiabatic π pulse (Böhlen and Bodenhausen 1993) to invert both C' and C^α (smoothed Chirp 500 μs , 20% smoothing, 80 kHz sweep width, 11.3 kHz RF field strength). Composite pulse decoupling was applied on ^1H (Waltz-16) (Shaka et al. 1983) and ^{15}N (Garp-4) (Shaka et al. 1985) with an RF field strength of 3 kHz and 1 kHz respectively.

^1H detected experiments, acquired at 1.2 GHz, exploited the BEST-TROSY approach (3D-BT-HNCACB and 3D-BT-HN(CO)CACB) or the sensitivity enhanced approach (3D-HNCO and 3D-HN(CA)CO) for the 3D experiments. The 2D- $[^1\text{H}, ^{15}\text{N}]$ -BEST-TROSY used sensitivity-enhanced gradient echo/antiecho coherence selection (Czisch and Boelens 1998; Schulte-Herbrüggen and Sørensen 2000) and Band-Selective Excitation Short-Transient (BEST) (Schanda et al. 2006; Lescop et al. 2007; Solyom et al. 2013) approach using exclusively shaped proton pulses. The inter-scan delay was set to 0.2 s. A 2D- $[^1\text{H}, ^{15}\text{N}]$ -HSQC was also acquired in its fast version which exploits Watergate 3-9-19 pulses for water suppression (Mori et al. 1995). 3D-BT-HNCACB, and 3D-BT-HN(CO)CACB used echo/antiecho gradient selection and semi-constant time in the ^{15}N dimension (Schulte-Herbrüggen and Sørensen 2000; Solyom et al. 2013). 3D-HNCO and 3D-HN(CA)CO used sensitivity enhanced approach and selective pulse on the solvent for the water suppression (Kay et al. 1994). C' and $\text{C}^\alpha/\text{C}^\beta$ selective excitation was exploited through band selective pulses.

Carrier frequencies used for triple resonance experiments in ^1H detected experiments were the same as for ^{13}C detected experiments except for the ^{15}N carrier placed at 118.0 ppm. Pulse shapes and lengths for ^{13}C band-selective pulses were G4 (Emsley and Bodenhausen 1992) and Q3 (Emsley and Bodenhausen 1990) shapes of durations of 205 and 128 μs , respectively, used for ^{13}C band-selective $\pi/2$ and π flip angle pulses except for the π pulses that should be band-selective on the C^α region (Q3, 525 μs). The ^1H band-selective pulses on the amide region were Pc9 (Kupce and Freeman 1994) or Eburp2 (Geen and Freeman 1991) for the $\pi/2$ and Reburp (Geen and Freeman 1991) or Bip (Smith et al. 2001) for π pulses.

All the spectra were acquired, processed, and analysed by using Bruker TopSpin 4.0.8 software. Chemical shifts were referenced using the ^1H and ^{13}C shifts of DSS. Nitrogen chemical shifts were referenced indirectly using the conversion factor derived from the ratio of NMR frequencies (Markley et al. 1998).

The sequence-specific assignment was performed with the aid of CARA (Keller 2004) and its tool NEASY (Bartels et al. 1995).

Bioinformatics tools

Several commonly utilized bioinformatics tools were used to predict or evaluate some of the protein features. Peculiarities of the distribution of intrinsic disorder predisposition along the amino acid sequence of the SARS-CoV-2 nucleocapsid protein N were evaluated by several members of the PONDR family (PONDR® VLXT (Romero et al. 2001), PONDR® VL3 (Obradovic et al. 2003), PONDR® VSL2 (Obradovic et al. 2005), and PONDR® FIT (Xue et al. 2010)), together with the two versions of IUPred2A designed to predict short and long disordered regions (Mészáros et al. 2018).

The online tool ncSPC available at <https://st-protein02.chem.au.dk/ncSPC/> was used to calculate the secondary structure propensity with the obtained assignment (Tamiola and Mulder 2012).

Assignments and data deposition

The 2D HN spectrum recorded on the IDR1-NTD-IDR2 (1–248) construct of the SARS-CoV-2 nucleocapsid protein N is shown in Fig. 2. The 2D HN spectrum clearly shows a set of well-resolved NMR signals deriving from the globular NTD domain, as one can verify by superimposing the available sequence-specific assignment (BMRB 34511, Dinesh et al. 2020). In addition, a set of signals, with smaller dispersion and higher intensity, are observed. These are expected to originate from the flexible and disordered fragments of the protein (black contours in Fig. 2).

The 2D CON spectrum (Fig. 3) provides information regarding the highly flexible and disordered protein regions. Due to the very different structural and dynamic properties of the globular NTD domain, with the chosen set-up the NMR signals of this region are very weak or absent in the 2D CON. This is exploited to selectively detect the resonances deriving from the two disordered protein regions. Proline residues can be directly monitored through the observation of the $\text{C}'_{i-1}\text{-N}_i$ correlations that fall in a very clean region of the CON spectrum ($132 < \delta(^{15}\text{N}) < 140$ ppm). The observation of only 7 well-resolved cross-peaks in this region (out of 17 expected for this construct) indeed confirms that C' direct detection selectively picks up the signals of the disordered regions (5 proline residues present in the IDR1 region and 2 in the IDR2 one, Fig. 3 bottom squared region).

Sequence-specific assignment of the resonances can be performed by combining the information available in the 2D ^{13}C -detected spectra with that provided by two 3D experiments, the (H)CBCACON and the (H)CBCANCO (Bermel et al. 2009).

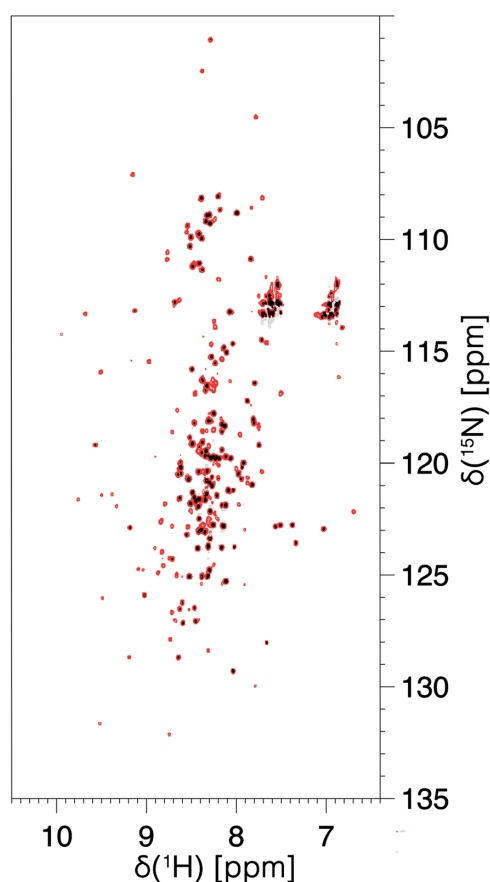


Fig. 2 The 2D HN BEST-TROSY of IDR1-NTD-IDR2 construct of the SARS-CoV-2 nucleocapsid protein. The figure shows the superimposition of two different processing of the same spectrum: the black one is optimized for the resolution and the red one is optimized for the signal to noise ratio. The spectrum was collected on a 28.3 T Bruker AVANCE NEO spectrometer operating at 1200.85 MHz ^1H , 301.97 MHz ^{13}C , and 121.70 MHz ^{15}N equipped with a 3 mm cryogenically cooled triple-resonance probehead (TCI)

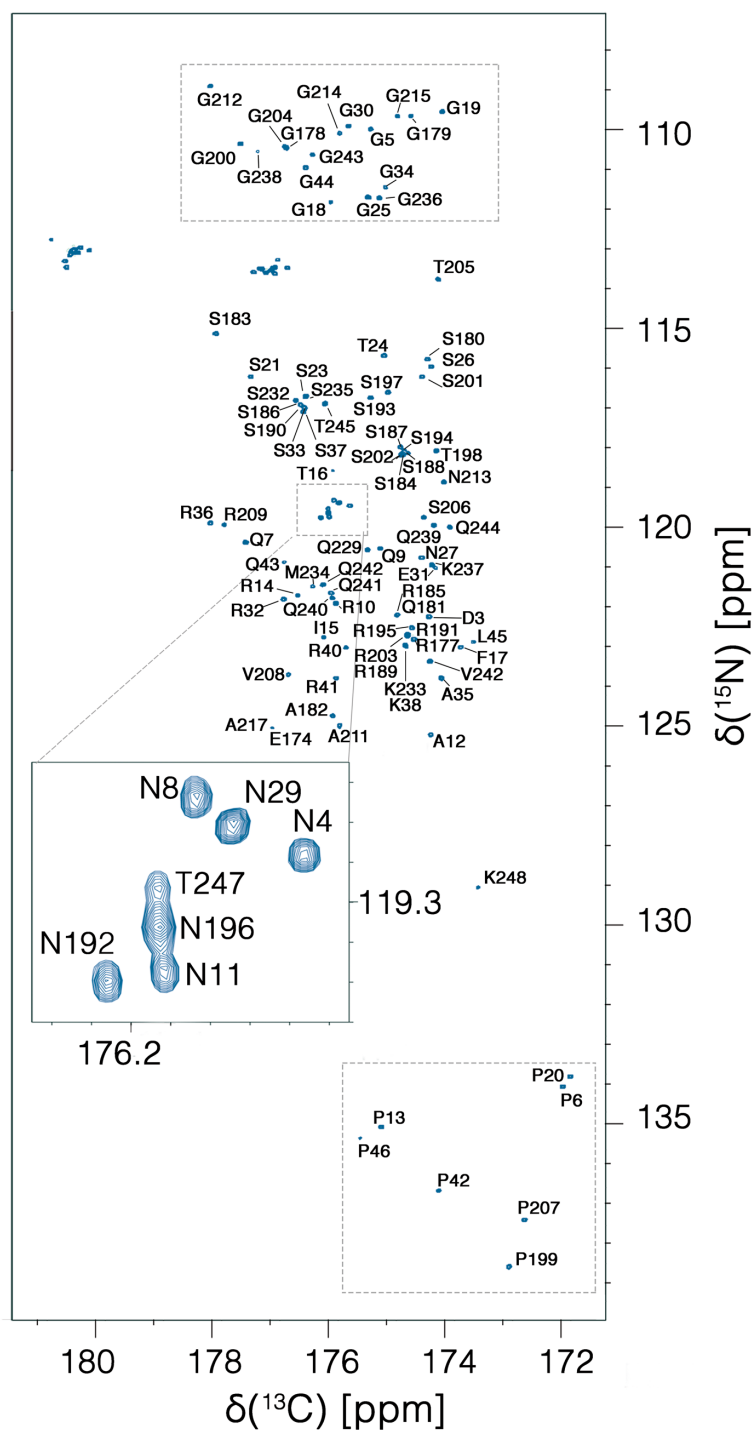
It is worth noting that proline resonances provide a useful starting point for sequence-specific assignment. The particular ^{15}N chemical shift range expected for proline nitrogen signals (N_i) and the fact that this is correlated to resonances of the preceding amino acid (C'_{i-1} , C^α_{i-1} , C^β_{i-1}) through the 2D CON and 3D (H)CBCACON spectra constitute two features that allow us to unambiguously identify the type of dipeptide (X_{i-1} -Pro $_i$ pair) that gives rise to specific signals as highlighted in Fig. 4. Indeed, the characteristic chemical shifts of C^α and C^β resonances enable us to recognize glycine, alanine, serine, and threonine residues; the remaining X-Pro pairs can then be easily identified as deriving from leucine and arginine residues by comparison with the

primary sequence of the protein. Therefore, already at this very early stage of the sequence-specific assignment process, most of the observed resonances in this region could be assigned to specific amino acids uniquely considering the type of X-Pro pairs present in the intrinsically disordered regions (all resonances could be unambiguously assigned except for the two Gly-Pro pairs). Similarly, inspecting the opposite region of the CON spectrum at low ^{15}N chemical shifts (Fig. 3, top squared region) allows us to identify correlations involving ^{15}N nuclear spins of glycine residues; correlation to the carbonyl carbon of the previous amino acid (C'_{i-1} - N_i) contributes to an excellent resolution allowing us to count 16 resolved cross peaks in this region in the simple 2D mode. This is in line with the number of glycine residues present in the flexible disordered fragments. The classification of these resonances in X_{i-1} -Gly $_i$ pairs achieved through inspection of the (H)CBCACON provides further input for their identification, as described above for the case of X_{i-1} -Pro $_i$ pairs. Complete comparative analysis of the 3D (H)CBCACON and 3D (H)CBCANCO spectra enables the identification of the vast majority of the expected resonances of disordered regions. The excellent resolution obtained in the 2D reference spectra, the CON as well as the (H)CACO and (H)CBCACO, provides valuable support for the analysis of crowded regions of the spectra and to the discrimination between different residue types (Pontoriero et al. 2020).

The information retrieved for the intrinsically disordered regions of the spectra can then be used as a starting point to identify the spin systems also in $^1\text{H}^{\text{N}}$ detected 3D spectra. The latter are much more crowded due to more extensive cross-peak overlap, as well as because the signals of the globular region are also observed. In addition, cross peak intensities are highly heterogeneous due to the different structural and dynamic properties of the globular and disordered domains as well as due to the effects of solvent exchange processes. Therefore, the combined analysis of the two datasets greatly simplifies the identification of the signals deriving from the intrinsically disordered regions. As a further aid to discriminate the different sets of signals, spectra can be processed to enhance resolution, at the expense of signal-to-noise, taking advantage of the long-lived ^{15}N coherences of highly flexible regions of the protein as well as exploiting the long FID acquisition times that are possible through the BEST-TROSY approach (Schanda et al. 2006; Lescop et al. 2007; Solyom et al. 2013).

As a result, 98% of the disordered fragment IDR1 (only the first methionine is missing) (BMRB 50619) and 91% of the fragment IDR2 (BMRB 50618) could be assigned in a sequence-specific manner (C' , C^α , C^β , N , H^{N}) (vide infra). It is interesting to note how the combined use of these complementary datasets ($^{13}\text{C}'$ - and $^1\text{H}^{\text{N}}$ -detected 3D experiments) provides information that is particularly useful to achieve sequence-specific assignment of intrinsically disordered

Fig. 3 The 2D-CON of IDR1-NTD-IDR2 construct of the SARS-CoV-2 nucleocapsid protein. The high resolution provided by this experiment allows us to easily resolve resonances in the usually very crowded Gly-region (upper squared region) and to directly observe correlations involving proline residues (lower squared region). In the expansion shown in the center of the map the resolution of several repeating fragments comprising asparagine residues can be appreciated (the assignment reported is referred to the amide nitrogen of the mentioned amino acid). The spectrum was acquired on a 16.4 T Bruker AVANCE NEO spectrometer operating at 700.06 MHz ^1H , 176.05 MHz ^{13}C , and 70.97 MHz ^{15}N frequencies, equipped with a 5 mm cryogenically cooled probehead optimized for ^{13}C direct detection (TXO)



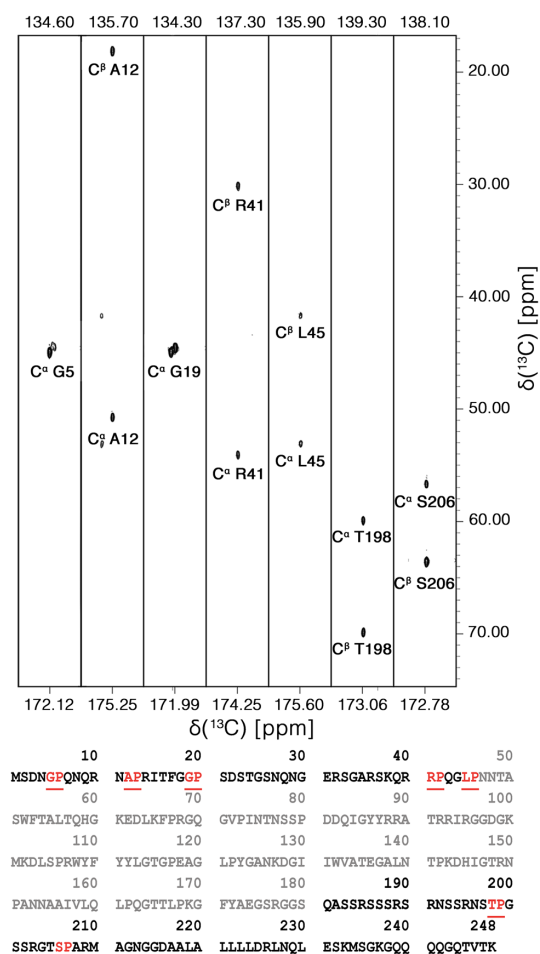


Fig. 4 Seven strips derived from the 3D-(H)CBCACON experiment extracted at the ^{15}N chemical shift of proline residues. The C^{α} , C^{β} and C^{γ} frequencies belong to the preceding amino acid leading to the X-Pro assignment. The lower part of the figure reports the IDR1-NTD-IDR2 primary sequence in which X-Pro pairs are highlighted. Five proline residues are found in the IDR1 and two in IDR2 domain. The primary sequence of NTD domain is reported in grey. The 3D spectrum was acquired on a 16.4 T Bruker AVANCE NEO spectrometer operating at 700.06 MHz ^1H , 176.05 MHz ^{13}C , and 70.97 MHz ^{15}N frequencies, equipped with a 5 mm cryogenically cooled probehead optimized for ^{13}C direct detection (TXO)

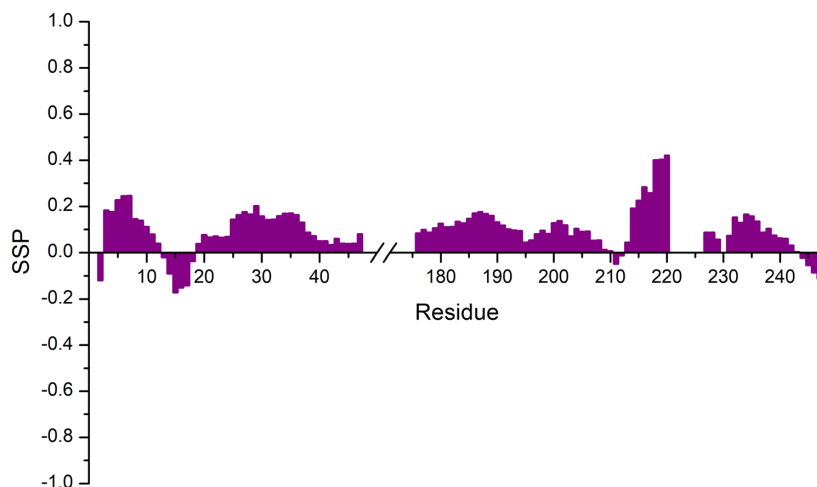
regions also within highly heterogeneous proteins. The set of 2D spectra (HN, CON, (H)CACO, (H)CBCACO), provided they are acquired with high resolution, then becomes a very useful tool to achieve atomic resolution for the vast majority of the amino acids in the highly flexible disordered regions of complex, heterogeneous proteins.

The first two disordered regions of the N protein from SARS-CoV-2 (IDR1 and IDR2) can now be investigated at atomic resolution providing experimental information regarding the many interaction sites that can be predicted through different approaches (Kumar et al. 2008; Giri et al. 2020). The resonances of characteristic amino acids involved in interactions with RNA, such as arginine, serine, glutamine, and glycine residues, which are very abundant in the IDR1 and IDR2 disordered domains, can be detected and most of them can be resolved already in the 2D mode also at physiological pH and temperature conditions. Several signals in low complexity regions, such as the polyQ (238–242) or some repeats located in different positions in the primary sequence (for example the Asn-Arg region reported in the expanded panel in the middle of Fig. 3) can be resolved allowing their high-resolution investigation.

Chemical shifts were then used to determine secondary structural propensities as shown in Fig. 5. The data confirm the disordered nature of these fragments, with a moderate propensity to sample a helical conformation in the leucine-rich region (218–232), where few residues (Leu 221, Leu 222, Leu 223, Leu 224, Asp 225, Arg 226, and Leu 230) escaped detection likely because of the signal broadening due to conformational exchange. These experimental results are in agreement with the bioinformatics analysis reported in Fig. 1, which predicts a high extent of disorder for the two IDR regions as well as the presence of some structure in the region 215–232.

The NMR resonance assignments of the IDR1 and IDR2 domains of the N protein from SARS-CoV-2 open the way to understanding the role of these flexible parts of the nucleocapsid protein in modulating its function. The suite of ^{13}C detected 2D experiments (CON, (H)CACO, (H)CBCACO) in conjunction with 2D HN correlation experiments provide an excellent tool to monitor at atomic resolution their role in the interactions with RNA, with viral proteins or with proteins of the host, as well as with small molecules as potential drugs, opening the way to radically novel, unexplored approaches in drug discovery.

Fig. 5 Secondary Structure Propensity (SSP) plot obtained with the assignment reported on the BMRB (50619 and 50618) for the two assigned regions 1–47 and 176–248. Chemical shift values for ^1H , ^{15}N , C' , C^α , and C^β nuclei were used. The two regions result to be highly disordered with a slight tendency to be in an α -helix conformation for the residues 216–220



Acknowledgements The support and the use of resources of the CERM/CIRMMMP centre of Instruct-ERIC is gratefully acknowledged. This work was supported in part by the Fondazione CR Firenze and by the Italian Ministry for University and Research (FOE funding). Fabio Almeida, Andreas Schlundt and Leonardo Gonnelli are acknowledged for the stimulating discussions.

Author contributions ICF and RP conceived the project and planned the experiments. LP purified and labelled the protein for NMR experiments. MS and LP acquired and analysed the data under the supervision of ICF and RP. VNU performed the bioinformatics analysis. All the authors contributed to writing the paper.

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement. Fondazione CR Firenze; Italian Ministry for University and Research (FOE funding).

Availability of data and materials The chemical shift values for the ^1H , ^{13}C and ^{15}N resonances of the first two flexible linkers of the SARS-CoV-2 nucleoprotein have been deposited in the BioMagResBank (<https://www.bmrb.wisc.edu>) under accession number 50619 (IDR1, residues 1–47) and 50618 (IDR2, residues 176–248). Spectral raw data (upon request) and assignments are also accessible through <https://covid19-nmr.de>.

Declarations

Conflict of interest The authors declare no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bartels C, Xia TH, Billeter M, Güntert P, Wütrich K (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J Biomol NMR* 6:1–10. <https://doi.org/10.1007/BF00417486>
- Bermel W, Bertini I, Felli IC, Kümmerle R, Pierattelli R (2006a) Novel ^{13}C direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins. *J Magn Reson* 178:56–64. <https://doi.org/10.1016/j.jmr.2005.08.011>
- Bermel W, Bertini I, Felli IC, Lee YM, Luchinat C, Pierattelli R (2006b) Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins. *J Am Chem Soc* 128:3918–3919. <https://doi.org/10.1021/ja0582206>
- Bermel W, Bertini I, Csizmek V, Felli IC, Pierattelli R, Tompa P (2009) H-start for exclusively heteronuclear NMR spectroscopy: the case of intrinsically disordered proteins. *J Magn Reson* 198:275–281. <https://doi.org/10.1016/j.jmr.2009.02.012>
- Böhlen JM, Bodenhausen G (1993) Experimental aspects of chirp NMR spectroscopy. *J Magn Reson Ser A* 102:293–301. <https://doi.org/10.1006/jmra.1993.1107>
- Chang CK, Sue SC, Yu TH, Hsien CM, Tsai CK, Chiang YC, Lee SJ, Hsiao HH, Wu WJ, Chang WL, Lin CH, Huang TH (2006) Modular organization of SARS coronavirus nucleocapsid protein. *J Biomed Sci* 13:59–72. <https://doi.org/10.1007/s11373-005-9035-9>
- Chang CK, Hsu YL, Chang YH, Chao FA, Wu MC, Huang YS, Hu CK, Huang TH (2009) Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging. *J Virol* 83:2255–2264. <https://doi.org/10.1128/jvi.02001-08>
- Chang CK, Hou MH, Chang CF, Hsiao CD, Huang TH (2014) The SARS coronavirus nucleocapsid protein—forms and functions. *Antiviral Res* 103:39–50. <https://doi.org/10.1016/j.antiviral.2013.12.009>
- Cui J, Li F, Shi ZL (2019) Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* 17:181–192. <https://doi.org/10.1038/s41579-018-0118-9>
- Czisch M, Boelens R (1998) Sensitivity enhancement in the TROSY Experiment. *J Magn Reson* 134:158–160. <https://doi.org/10.1006/jmre.1998.1483>

- Dinesh DC, Chalupska D, Silhan J, Veverka V, Boura E (2020) Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLoS Pathog* 16:e1009100. <https://doi.org/10.1371/journal.ppat.1009100>
- Emsley L, Bodenhausen G (1990) Gaussian pulse cascades: new analytical functions for rectangular selective inversion and in-phase excitation in NMR. *Chem Phys Lett* 165:469–476. [https://doi.org/10.1016/0009-2614\(90\)87025-M](https://doi.org/10.1016/0009-2614(90)87025-M)
- Emsley L, Bodenhausen G (1992) Optimization of shaped selective pulses for NMR using a quaternion description of their overall propagators. *J Magn Reson* 97:135–148. [https://doi.org/10.1016/0022-2364\(92\)90242-Y](https://doi.org/10.1016/0022-2364(92)90242-Y)
- Felli IC, Pierattelli R (2012) Recent progress in NMR spectroscopy: toward the study of intrinsically disordered proteins of increasing size and complexity. *IUBMB Life* 64:473–481. <https://doi.org/10.1002/iub.1045>
- Geen H, Freeman R (1991) Band-selective radiofrequency pulses. *J Magn Reson* 93:93–141. [https://doi.org/10.1016/0022-2364\(91\)90034-Q](https://doi.org/10.1016/0022-2364(91)90034-Q)
- Giri R, Bhardwaj T, Shegane M, Gehi BR, Kumar P, Godhava K, Oldfield CJ, Uversky VN (2020) Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses. *Cell Mol Life Sci* 25:1–34. <https://doi.org/10.1007/s00018-020-03603-x>
- Goh GKM, Dunker AK, Uversky VN (2012) Understanding viral transmission behavior via protein intrinsic disorder prediction: coronaviruses. *J Pathog* 2012:738590. <https://doi.org/10.1155/2012/738590>
- Goh GKM, Dunker AK, Uversky V (2013) Prediction of intrinsic disorder in MERS-CoV/HCoV-EMC supports a high oral-fecal transmission. *PLoS Curr* 5:1–25. <https://doi.org/10.1371/currents.outbreaks.22254b58675cdebc256dbe3c5aa6498b>
- Huang C, Wang Y, Li X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395:497–506. [https://doi.org/10.1016/S0140-6736\(20\)30183-5](https://doi.org/10.1016/S0140-6736(20)30183-5)
- Kang S, Yang M, Hong Z, Zhang L, Huang Z, Chen X, He S, Zhou ZZ, Chen Q, Yan Y, Zhang C, Shan H, Chen S (2020) Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm Sin B* 10:1228–1238. <https://doi.org/10.1016/j.apsb.2020.04.009>
- Kay LE, Xu GY, Yamazaki T (1994) Enhanced-Sensitivity triple-resonance spectroscopy with minimal H₂O saturation. *J Magn Reson Ser A* 109:129–133. <https://doi.org/10.1006/jmra.1994.1145>
- Keller R (2004) The computer aided resonance assignment tutorial. Goldau, Switz. Cantina Verlag 1–81
- Kumar M, Gromiha MM, Raghava GPS (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins Struct Funct Genet* 71:189–194. <https://doi.org/10.1002/prot.21677>
- Kucec E, Freeman R (1994) Wideband excitation with polychromatic pulses. *J Magn Reson Ser A* 108:268–273. <https://doi.org/10.1006/jmra.1994.1123>
- Lescop E, Schanda P, Brutscher B (2007) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. *J Magn Reson* 187:163–169. <https://doi.org/10.1016/j.jmr.2007.04.002>
- Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sukes BD, Wright PE, Wütrich K (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids. *Pure Appl Chem* 70:117–142. <https://doi.org/10.1023/A:1008290618449>
- Masters PS (2006) The molecular biology of coronaviruses. *Adv Virus Res* 65:193–292. [https://doi.org/10.1016/S0065-3527\(06\)66005-3](https://doi.org/10.1016/S0065-3527(06)66005-3)
- Mészáros B, Erdős G, Dosztányi Z (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* 46:W329–W337. <https://doi.org/10.1093/nar/gky384>
- Mori S, Abeygunawardana C, Johnson MO, Vanzijl PCM (1995) Improved sensitivity of HSQC spectra of exchanging protons at short interscan delays using a new Fast HSQC (FHSQC) detection scheme that avoids water saturation. *J Magn Reson Ser B* 108:94–98. <https://doi.org/10.1006/jmrb.1995.1109>
- Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins Struct Funct Genet* 53:566–572. <https://doi.org/10.1002/prot.10532>
- Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins Struct Funct Genet* 61:176–182. <https://doi.org/10.1002/prot.20735>
- Peng Y, Du N, Lei Y, Dorje S, Qi J, Luo T, Gao GF, Song H (2020) Structures of the SARS-CoV-2 nucleocapsid and their perspectives for drug design. *EMBO J* 39:1–12. <https://doi.org/10.15252/embj.2020105938>
- Pontoriero L, Schiavina M, Murrall MG, Pierattelli R, Felli IC (2020) Monitoring the interaction of α -synuclein with calcium ions through exclusively heteronuclear nuclear magnetic resonance experiments. *Angew Chem Int Ed* 59:18537–18545. <https://doi.org/10.1002/anie.202008079>
- Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Ak D (2001) Sequence complexity of disordered protein. *Proteins* 42:38–48. [https://doi.org/10.1002/1097-0134\(20010101\)42:1<3c38::aid-prot50>3e3.0.co;2-3](https://doi.org/10.1002/1097-0134(20010101)42:1<3c38::aid-prot50>3e3.0.co;2-3)
- Schanda P, Van Melckebeke H, Brutscher B (2006) Speeding up three-dimensional protein NMR experiments to a few minutes. *J Am Chem Soc* 128:9042–9043. <https://doi.org/10.1021/ja062025p>
- Schulte-Herbrüggen T, Sørensen OW (2000) Clean TROSY: compensation for relaxation-induced artifacts. *J Magn Reson* 144:123–128. <https://doi.org/10.1006/jmre.2000.2020>
- Shaka AJ, Keeler J, Freeman R (1983) Evaluation of a new broadband decoupling sequence: WALTZ-16. *J Magn Reson* 53:313–340. [https://doi.org/10.1016/0022-2364\(83\)90035-5](https://doi.org/10.1016/0022-2364(83)90035-5)
- Shaka AJ, Barker PB, Freeman R (1985) Computer-optimized decoupling scheme for wideband applications and low-level operation. *J Magn Reson* 64:547–552. [https://doi.org/10.1016/0022-2364\(85\)90122-2](https://doi.org/10.1016/0022-2364(85)90122-2)
- Smith MA, Hu H, Shaka AJ (2001) Improved broadband inversion performance for NMR in liquids. *J Magn Reson* 151:269–283. <https://doi.org/10.1006/jmre.2001.2364>
- Solyom Z, Schwarten M, Geist L, Konrat R, Willbold D, Brutscher B (2013) BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J Biomol NMR* 55:311–321. <https://doi.org/10.1007/s10858-013-9715-0>
- Surjit M, Lal SK (2008) The SARS-CoV nucleocapsid protein: a protein with multifarious activities. *Infect Genet Evol* 8:397–405. <https://doi.org/10.1016/j.meegid.2007.07.004>
- Tamiola K, Mulder FAA (2012) Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochem Soc Trans* 40:1014–1020. <https://doi.org/10.1042/BST20120171>
- Wang D, Hu B, Hu C et al (2020) Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* 323:1061. <https://doi.org/10.1001/jama.2020.1585>
- Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* 1804:996–1010. <https://doi.org/10.1016/j.bbapap.2010.01.011>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

12. Large-scale recombinant production of the SARS-CoV-2 proteome for high-throughput and structural biology applications

OPEN ACCESS



Edited by:

Qian Fan,
Hainan University, China

Reviewed by:

David Douglas Eicher,
Hemphill State University, United States

Luis G. Gilboa,
National Polytechnic Institute of Mexico

(INMEXAT), Mexico

*Correspondence:

Anja Böckmann

a.boeckmann@bioinf.uni-frankfurt.de

Harald Schwalbe

schwalbe@mr.uni-frankfurt.de

Martin Hengsbach

hengsbach@mr.uni-frankfurt.de

Andreas Schlund

schlund@bio.uni-frankfurt.de

¹ These authors have contributed equally to this work

and shared first authorship

² These authors share last authorship

³ Specialty section:

This article was submitted to

Structural Biology,

a section of the journal

Frontiers in Molecular Biosciences

Received: 03 January 2021

Accepted: 04 February 2021

Published: 10 May 2021

Citation:

Altmann H, Kim SB, Gurell NG, Gurell M,

Nikolichos M, Albrecht R, Albrecht M, Albrecht C,

Albrecht T, Albrecht L, de Amorim GC,

Anderson TH, Anderson CG, Anoma G, Bae JK,

Bae A, Baderger M, Becker J, Böhmann A,

Böckmann A, Böckmann M, Camacho-Zarco AR,

Chakraborty T, Chao P, Cheloni D, Chikara A,

Chu J, Cole L, Crosby ME, de Jesus V,

Demchenko K, Fell K, Fiebig J, Fiedler M,

Fischer M, Fölscher M, Gao G, Fritsch G, Gallo A,

Gandhi S, Gao J, Gharib N, Gharib M, Gharib R,

Gharib S, Gharib T, Gharib U, Gharib V, Gharib W,

Gharib X, Gharib Y, Gharib Z, Gharib AA, Gharib AB,

Gharib AC, Gharib AD, Gharib AE, Gharib AF,

Gharib AG, Gharib AH, Gharib AI, Gharib AJ,

Gharib AK, Gharib AL, Gharib AM, Gharib AN,

Gharib AO, Gharib AP, Gharib AQ, Gharib AR,

Gharib AS, Gharib AT, Gharib AU, Gharib AV,

Gharib AW, Gharib AX, Gharib AY, Gharib AZ,

Gharib BA, Gharib BB, Gharib BC, Gharib BD,

Gharib BE, Gharib BF, Gharib BG, Gharib BH,

Gharib BI, Gharib BJ, Gharib BK, Gharib BL,

Gharib BM, Gharib BN, Gharib BO, Gharib BP,

Gharib BQ, Gharib BR, Gharib BS, Gharib BT,

Gharib BU, Gharib BV, Gharib BW, Gharib BX,

Gharib BY, Gharib BZ, Gharib CA, Gharib CB,

Gharib CC, Gharib CD, Gharib CE, Gharib CF,

Gharib CG, Gharib CH, Gharib CI, Gharib CJ,

Gharib CK, Gharib CL, Gharib CM, Gharib CN,

Gharib CO, Gharib CP, Gharib CQ, Gharib CR,

Gharib CS, Gharib CT, Gharib CU, Gharib CV,

Gharib CW, Gharib CX, Gharib CY, Gharib CZ,

Gharib DA, Gharib DB, Gharib DC, Gharib DD,

Gharib DE, Gharib DF, Gharib DG, Gharib DH,

Gharib DI, Gharib DJ, Gharib DK, Gharib DL,

Gharib DM, Gharib DN, Gharib DO, Gharib DP,

Gharib DQ, Gharib DR, Gharib DS, Gharib DT,

Gharib DU, Gharib DV, Gharib DW, Gharib DX,

Gharib DY, Gharib DZ, Gharib EA, Gharib EB,

Gharib EC, Gharib ED, Gharib EE, Gharib EF,

Gharib EG, Gharib EH, Gharib EI, Gharib EJ,

Gharib EK, Gharib EL, Gharib EM, Gharib EN,

Gharib EO, Gharib EP, Gharib EQ, Gharib ER,

Gharib ES, Gharib ET, Gharib EU, Gharib EV,

Gharib EW, Gharib EX, Gharib EY, Gharib EZ,

Gharib FA, Gharib FB, Gharib FC, Gharib FD,

Gharib FE, Gharib FF, Gharib FG, Gharib FH,

Gharib FI, Gharib FJ, Gharib FK, Gharib FL,

Gharib FM, Gharib FN, Gharib FO, Gharib FP,

Gharib FQ, Gharib FR, Gharib FS, Gharib FT,

Gharib FU, Gharib FV, Gharib FW, Gharib FX,

Gharib FY, Gharib FZ, Gharib GA, Gharib GB,

Gharib GC, Gharib GD, Gharib GE, Gharib GF,

Gharib GG, Gharib GH, Gharib GI, Gharib GJ,

Gharib GK, Gharib GL, Gharib GM, Gharib GN,

Gharib GO, Gharib GP, Gharib GQ, Gharib GR,

Gharib GS, Gharib GT, Gharib GU, Gharib GV,

Gharib GW, Gharib GX, Gharib GY, Gharib GZ,

Gharib HA, Gharib HB, Gharib HC, Gharib HD,

Gharib HE, Gharib HF, Gharib HG, Gharib HH,

Gharib HI, Gharib HJ, Gharib HK, Gharib HL,

Gharib HM, Gharib HN, Gharib HO, Gharib HP,

Gharib HQ, Gharib HR, Gharib HS, Gharib HT,

Gharib HU, Gharib HV, Gharib HW, Gharib HX,

Gharib HY, Gharib HZ, Gharib IA, Gharib IB,

Gharib IC, Gharib ID, Gharib IE, Gharib IF,

Gharib IG, Gharib IH, Gharib II, Gharib IJ,

Gharib IK, Gharib IL, Gharib IM, Gharib IN,

Gharib IO, Gharib IP, Gharib IQ, Gharib IR,

Gharib IS, Gharib IT, Gharib IU, Gharib IV,

Gharib IW, Gharib IX, Gharib IY, Gharib IZ,

Gharib JA, Gharib JB, Gharib JC, Gharib JD,

Gharib JE, Gharib JF, Gharib JG, Gharib JH,

Gharib JI, Gharib JJ, Gharib JK, Gharib JL,

Gharib JM, Gharib JN, Gharib JO, Gharib JP,

Gharib JQ, Gharib JR, Gharib JS, Gharib JT,

Gharib JU, Gharib JV, Gharib JW, Gharib JX,

Gharib JY, Gharib JZ, Gharib KA, Gharib KB,

Gharib KC, Gharib KD, Gharib KE, Gharib KF,

Gharib KG, Gharib KH, Gharib KI, Gharib KJ,

Gharib KK, Gharib KL, Gharib KM, Gharib KN,

Gharib KO, Gharib KP, Gharib KQ, Gharib KR,

Gharib KS, Gharib KT, Gharib KU, Gharib KV,

Gharib KW, Gharib KX, Gharib KY, Gharib KZ,

Gharib LA, Gharib LB, Gharib LC, Gharib LD,

Gharib LE, Gharib LF, Gharib LG, Gharib LH,

Gharib LI, Gharib LJ, Gharib LK, Gharib LL,

Gharib LM, Gharib LN, Gharib LO, Gharib LP,

Gharib LQ, Gharib LR, Gharib LS, Gharib LT,

Gharib LU, Gharib LV, Gharib LW, Gharib LX,

Gharib LY, Gharib LZ, Gharib MA, Gharib MB,

Gharib MC, Gharib MD, Gharib ME, Gharib MF,

Gharib MG, Gharib MH, Gharib MI, Gharib MJ,

Gharib MK, Gharib ML, Gharib MM, Gharib MN,

Gharib MO, Gharib MP, Gharib MQ, Gharib MR,

Gharib MS, Gharib MT, Gharib MU, Gharib MV,

Gharib MW, Gharib MX, Gharib MY, Gharib MZ,

Gharib NA, Gharib NB, Gharib NC, Gharib ND,

Gharib NE, Gharib NF, Gharib NG, Gharib NH,

Gharib NI, Gharib NJ, Gharib NK, Gharib NL,

Gharib NM, Gharib NN, Gharib NO, Gharib NP,

Gharib NQ, Gharib NR, Gharib NS, Gharib NT,

Gharib NU, Gharib NV, Gharib NW, Gharib NX,

Gharib NY, Gharib NZ, Gharib OA, Gharib OB,

Gharib OC, Gharib OD, Gharib OE, Gharib OF,

Gharib OG, Gharib OH, Gharib OI, Gharib OJ,

Gharib OK, Gharib OL, Gharib OM, Gharib ON,

Gharib OO, Gharib OP, Gharib OQ, Gharib OR,

Gharib OS, Gharib OT, Gharib OU, Gharib OV,

Gharib OW, Gharib OX, Gharib OY, Gharib OZ,

Gharib PA, Gharib PB, Gharib PC, Gharib PD,

Gharib PE, Gharib PF, Gharib PG, Gharib PH,

Gharib PI, Gharib PJ, Gharib PK, Gharib PL,

Gharib PM, Gharib PN, Gharib PO, Gharib PP,

Gharib PQ, Gharib PR, Gharib PS, Gharib PT,

Gharib PU, Gharib PV, Gharib PW, Gharib PX,

Gharib PY, Gharib PZ, Gharib QA, Gharib QB,

Gharib QC, Gharib QD, Gharib QE, Gharib QF,

Gharib QG, Gharib QH, Gharib QI, Gharib QJ,

Gharib QK, Gharib QL, Gharib QM, Gharib QN,

Gharib QO, Gharib QP, Gharib QQ, Gharib QR,

Gharib QS, Gharib QT, Gharib QU, Gharib QV,

Gharib QW, Gharib QX, Gharib QY, Gharib QZ,

Gharib RA, Gharib RB, Gharib RC, Gharib RD,

Gharib RE, Gharib RF, Gharib RG, Gharib RH,

Gharib RI, Gharib RJ, Gharib RK, Gharib RL,

Gharib RM, Gharib RN, Gharib RO, Gharib RP,

Gharib RQ, Gharib RR, Gharib RS, Gharib RT,

Gharib RU, Gharib RV, Gharib RW, Gharib RX,

Gharib RY, Gharib RZ, Gharib SA, Gharib SB,

Gharib SC, Gharib SD, Gharib SE, Gharib SF,

Gharib SG, Gharib SH, Gharib SI, Gharib SJ,

Gharib SK, Gharib SL, Gharib SM, Gharib SN,

Gharib SO, Gharib SP, Gharib SQ, Gharib SR,

Gharib SS, Gharib ST, Gharib SU, Gharib SV,

Gharib SW, Gharib SX, Gharib SY, Gharib SZ,

Gharib TA, Gharib TB, Gharib TC, Gharib TD,

Gharib TE, Gharib TF, Gharib TG, Gharib TH,

Gharib TI, Gharib TJ, Gharib TK, Gharib TL,

Gharib TM, Gharib TN, Gharib TO, Gharib TP,

Gharib TQ, Gharib TR, Gharib TS, Gharib TT,

Gharib TU, Gharib TV, Gharib TW, Gharib TX,

Gharib TY, Gharib TZ, Gharib UA, Gharib UB,

Gharib UC, Gharib UD, Gharib UE, Gharib UF,

Gharib UG, Gharib UH, Gharib UI, Gharib UJ,

Gharib UK, Gharib UL, Gharib UM, Gharib UN,

Gharib UO, Gharib UP, Gharib UQ, Gharib UR,

Gharib US, Gharib UT, Gharib UU, Gharib UV,

Gharib UW, Gharib UX, Gharib UY, Gharib UZ,

Gharib VA, Gharib VB, Gharib VC, Gharib VD,

Gharib VE, Gharib VF, Gharib VG, Gharib VH,

Gharib VI, Gharib VJ, Gharib VK, Gharib VL,

Gharib VM, Gharib VN, Gharib VO, Gharib VP,

Gharib VQ, Gharib VR, Gharib VS, Gharib VT,

Gharib VU, Gharib VV, Gharib VW, Gharib VX,

Gharib VY, Gharib VZ, Gharib WA, Gharib WB,

Gharib WC, Gharib WD, Gharib WE, Gharib WF,

Gharib WG, Gharib WH, Gharib WI, Gharib WJ,

Gharib WK, Gharib WL, Gharib WM, Gharib WN,

Gharib WO, Gharib WP, Gharib WQ, Gharib WR,

Gharib WS, Gharib WT, Gharib WU, Gharib WV,

Gharib WW, Gharib WX, Gharib WY, Gharib WZ,

Gharib XA, Gharib XB, Gharib XC, Gharib XD,

Gharib XE, Gharib XF, Gharib XG, Gharib XH,

Gharib XI, Gharib XJ, Gharib XK, Gharib XL,

Gharib XM, Gharib XN, Gharib XO, Gharib XP,

Gharib XQ, Gharib XR, Gharib XS, Gharib XT,

Gharib XU, Gharib XV, Gharib XW, Gharib XX,

Gharib XY, Gharib XZ, Gharib YA, Gharib YB,

Gharib YC, Gharib YD, Gharib YE, Gharib YF,

Gharib YG, Gharib YH, Gharib YI, Gharib YJ,

Gharib YK, Gharib YL, Gharib YM, Gharib YN,

Gharib YO, Gharib YP, Gharib YQ, Gharib YR,

Gharib YS, Gharib YT, Gharib YU, Gharib YV,

Gharib YW, Gharib YX, Gharib YY, Gharib YZ,

Gharib ZA, Gharib ZB, Gharib ZC, Gharib ZD,

Gharib ZE, Gharib ZF, Gharib ZG, Gharib ZH,

Gharib ZI, Gharib ZJ, Gharib ZK, Gharib ZL,

Gharib ZM, Gharib ZN, Gharib ZO, Gharib ZP,

Gharib ZQ, Gharib ZR, Gharib ZS, Gharib ZT,

Gharib ZU, Gharib ZV, Gharib ZW, Gharib ZX,

Gharib ZY, Gharib ZZ, Gharib AA, Gharib AB,

Gharib AC, Gharib AD, Gharib AE, Gharib AF,

Gharib AG, Gharib AH, Gharib AI, Gharib AJ,

Gharib AK, Gharib AL, Gharib AM, Gharib AN,

Gharib AO, Gharib AP, Gharib AQ, Gharib AR,

Gharib AS, Gharib AT, Gharib AU, Gharib AV,

Gharib AW, Gharib AX, Gharib AY, Gharib AZ,

Gharib BA, Gharib BB, Gharib BC, Gharib BD,

Gharib BE, Gharib BF, Gharib BG, Gharib BH,

Gharib BI, Gharib BJ, Gharib BK, Gharib BL,

Gharib BM, Gharib BN, Gharib BO, Gharib BP,

Gharib BQ, Gharib BR, Gharib BS, Gharib BT,

Gharib BU, Gharib BV, Gharib BW, Gharib BX,

Gharib BY, Gharib BZ, Gharib CA, Gharib CB,

Gharib CC, Gharib CD, Gharib CE, Gharib CF,

Gharib CG, Gharib CH, Gharib CI, Gharib CJ,

Gharib CK, Gharib CL, Gharib CM, Gharib CN,

Gharib CO, Gharib CP, Gharib CQ, Gharib CR,

Gharib CS, Gharib CT, Gharib CU, Gharib CV,

Gharib CW, Gharib CX, Gharib CY, Gharib CZ,

Gharib DA, Gharib DB, Gharib DC, Gharib DD,

Gharib DE, Gharib DF, Gharib DG, Gharib DH,

Gharib DI, Gharib DJ, Gharib DK, Gharib DL,

Gharib DM, Gharib DN, Gharib DO, Gharib DP,

Gharib DQ, Gharib DR, Gharib DS, Gharib DT,

Gharib DU, Gharib DV, Gharib DW, Gharib DX,

Gharib DY, Gharib DZ, Gharib EA, Gharib EB,

Gharib EC, Gharib ED, Gharib EE, Gharib EF,

Gharib EG, Gharib EH, Gharib EI, Gharib EJ,

OPEN ACCESS



Edited by:

Qian Han,
Hainan University, China

Reviewed by:

David Douglas Boehr,
Pennsylvania State University, United States
Luis G. Bribea,
National Polytechnic Institute of Mexico
(QINVESTAV), Mexico

*Correspondence:

Anja Böckmann
a.boeckmann@ibcp.fr
Harald Schwalbe
schwalbe@nmr.uni-frankfurt.de
Martin Hengesbach
hengesbach@nmr.uni-frankfurt.de
Andreas Schlundt
schlundt@bio.uni-frankfurt.de

[†]These authors have contributed equally to this work
and share first authorship

[‡]These authors share last authorship

Specialty section:

This article was submitted to
Structural Biology,
a section of the journal
Frontiers in Molecular Biosciences

Received: 13 January 2021

Accepted: 04 February 2021

Published: 10 May 2021

Citation:

Altinokic N, Kom SM, Qureshi NS, Dujardin M, Ninot-Pedrosa M, Abele R, Abi Saad MJ, Alfano C, Almeida FCL, Alshamleh I, de Amorim GC, Anderson TK, Anobom CD, Anorma C, Bains JK, Bax A, Blackledge M, Blechar J, Böckmann A, Brigandat L, Bula A, Büttiker M, Camacho-Zarco AR, Carlomagno T, Caruso IP, Ceylan B, Chaikwad A, Chu F, Cole L, Crosby MG, de Jesus V, Dhamotharan K, Felli IC, Ferner J, Fleischmann Y, Fogeron M-L, Fourkotiis NK, Fuks C, Fürtig B, Gallo A, Gande SL, Gerez JA, Ghosh D, Gomes-Neto F, Gorbatyuk O, Guseva S, Hacker C, Häfner S, Hao B, Hargittay B, Hender-Wildman K, Hoch JC, Hohmann KF, Hutchison MT, Jaudzems K, Jovic K, Kaderli J, Kalniņš G, Kanape I, Kirchdoerfer FN, Kirkpatrick J, Knapp S, Krishnathas R, Kutz F, zur Lage S, Lambert R, Lang A, Laurents D, Lecoq L, Linhard V, Lohr F, Malki A, Bessa LM, Martin RW, Matzel T, Maurin D, McNutt SW, Mebus-Antunes NC, Meier BH, Meiser N, Mompeán M, Monaca E, Montserret R, Mariño Perez L, Moser C, Muhle-Goll C, Neves-Martins TC, Ni X, Norton-Baker B, Pierattelli R, Pontoriero L, Pustovalova Y, Ohlenschläger O, Orts J, Da Poian AT, Pyper DJ, Richter C, Riek R, Rienstra CM, Robertson A, Pinheiro AS, Sabbatella R, Salvi N, Saxena K, Schulte L, Schiavina M, Schwalbe H, Silber M, Almeida MDS, Sprague-Piercy MA, Spyroulias GA, Sreeramulu S, Tants J-N, Tars K, Torres F, Töws S, Treviño MA, Trucks S, Tsika AC, Varga K, Wang Y, Weber ME, Weigand JE, Wiedemann C, Wimer-Bartoschek J, Wirtz Martin MA, Zahnder J, Hengesbach M and Schlundt A (2021) Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications. *Front. Mol. Biosci.* 6:653148. doi: 10.3389/fmolb.2021.653148

Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications

Nadide Altinokic^{1,2†}, Sophie Marianne Korn^{2,3†}, Nusrat Shahin Qureshi^{1,2†}, Marie Dujardin^{4†}, Martí Ninot-Pedrosa^{4†}, Rupert Abele⁵, Marie Jose Abi Saad⁶, Caterina Alfano⁷, Fabio C. L. Almeida^{8,9}, Islam Alshamleh^{1,2}, Gisele Cardoso de Amorim^{8,10}, Thomas K. Anderson¹¹, Cristiane D. Anobom^{8,12}, Chelsea Anorma¹³, Jasleen Kaur Bains^{1,2}, Adriaan Bax¹⁴, Martin Blackledge¹⁵, Julius Blechar^{1,2}, Anja Böckmann^{4*†}, Louis Brigandat⁴, Anna Bula¹⁶, Matthias Büttiker⁶, Aldo R. Camacho-Zarco¹⁵, Teresa Carlomagno^{17,18}, Icaro Putinhon Caruso^{8,9,19}, Betül Ceylan^{1,2}, Apirat Chaikwad^{20,21}, Feixia Chu²², Laura Cole⁴, Marquise G. Crosby²³, Vanessa de Jesus^{1,2}, Karthikeyan Dhamotharan^{2,3}, Isabella C. Felli^{24,25}, Jan Ferner^{1,2}, Yanick Fleischmann⁶, Marie-Laure Fogeron⁴, Nikolaos K. Fourkotiis²⁶, Christin Fuks¹, Boris Fürtig^{1,2}, Angelo Gallo²⁶, Santosh L. Gande^{1,2}, Juan Atilio Gerez⁶, Dhiman Ghosh⁶, Francisco Gomes-Neto^{8,27}, Oksana Gorbatyuk²⁸, Serafima Guseva¹⁵, Carolin Hacker²⁹, Sabine Häfner³⁰, Bing Hao²⁸, Bruno Hargittay^{1,2}, K. Hender-Wildman¹¹, Jeffrey C. Hoch²⁸, Katharina F. Hohmann^{1,2}, Marie T. Hutchison^{1,2}, Kristaps Jaudzems¹⁶, Katarina Jović²², Janina Kaderli⁶, Gints Kalniņš³¹, Iveta Kanape¹⁶, Robert N. Kirchdoerfer¹¹, John Kirkpatrick^{17,18}, Stefan Knapp^{20,21}, Robin Krishnathas^{1,2}, Felicitas Kutz^{1,2}, Susanne zur Lage¹⁸, Roderick Lambert³, Andras Lang³⁰, Douglas Laurents³², Lauriane Lecoq⁴, Verena Linhard^{1,2}, Frank Lohr^{2,33}, Anas Malki¹⁵, Luiza Mamigonian Bessa¹⁵, Rachel W. Martin^{13,23}, Tobias Matzel^{1,2}, Damien Maurin¹⁵, Seth W. McNutt²², Nathane Cunha Mebus-Antunes^{8,9}, Beat H. Meier⁶, Nathalie Meiser¹, Miguel Mompeán³², Elisa Monaca⁷, Roland Montserret⁴, Laura Mariño Perez¹⁵, Celine Moser³⁴, Claudia Muhle-Goll³⁴, Thais Cristtina Neves-Martins^{8,9}, Xiamonin Ni^{20,21}, Brenna Norton-Baker¹³, Roberta Pierattelli^{24,25}, Letizia Pontoriero^{24,25}, Yulia Pustovalova²⁸, Oliver Ohlenschläger³⁰, Julien Orts⁶, Andrea T. Da Poian⁹, Dennis J. Pyper^{1,2}, Christian Richter^{1,2}, Roland Riek⁶, Chad M. Rienstra³⁵, Angus Robertson¹⁴, Anderson S. Pinheiro^{8,12}, Raffaele Sabbatella⁷, Nicola Salvi¹⁵, Krishna Saxena^{1,2}, Linda Schulte^{1,2}, Marco Schiavina^{24,25}, Harald Schwalbe^{1,2*†}, Mara Silber³⁴, Marcius da Silva Almeida^{8,9}, Marc A. Sprague-Piercy²³, Georgios A. Spyroulias²⁶, Sridhar Sreeramulu^{1,2}, Jan-Niklas Tants^{2,3}, Kaspars Tars³¹, Felix Torres⁶, Sabrina Töws³, Miguel Á. Treviño³², Sven Trucks¹, Aikaterini C. Tsika²⁶, Krisztina Varga²², Ying Wang¹⁷, Marco E. Weber⁶, Julia E. Weigand³⁶, Christoph Wiedemann³⁷, Julia Wimer-Bartoschek^{1,2}, Maria Alexandra Wirtz Martin^{1,2}, Johannes Zahnder⁶, Martin Hengesbach^{1*†} and Andreas Schlundt^{2,3*†}

¹Institute for Organic Chemistry and Chemical Biology, Goethe University Frankfurt, Frankfurt am Main, Germany, ²Center of Biomolecular Magnetic Resonance (BMRZ), Goethe University Frankfurt, Frankfurt am Main, Germany, ³Institute for Molecular Biosciences, Goethe University Frankfurt, Frankfurt am Main, Germany, ⁴Molecular Microbiology and Structural Biochemistry, UMR 5086, CNRS/Lyon University, Lyon, France, ⁵Institute for Biochemistry, Goethe University Frankfurt, Frankfurt am Main, Germany, ⁶Swiss Federal Institute of Technology, Laboratory of Physical Chemistry, ETH Zurich, Zurich, Switzerland, ⁷Structural Biology and Biophysics Unit, Fondazione Ri.MED, Palermo, Italy, ⁸National Center of Nuclear Magnetic Resonance (CNRMN, CENABIO), Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, ⁹Institute of Medical Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, ¹⁰Multidisciplinary Center for Research in Biology (NUMPEX), Campus Duque de Caxias Federal University of Rio de Janeiro, Duque de Caxias, Brazil, ¹¹Institute for Molecular Virology, University of Wisconsin-Madison, Madison, WI, United States, ¹²Institute of Chemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, ¹³Department of Chemistry, University of California, Irvine, CA, United States, ¹⁴LCP, NIDDK, NIH, Bethesda, MD, United States, ¹⁵Univ. Grenoble Alpes, CNRS, CEA, IBS, Grenoble, France, ¹⁶Latvian Institute of Organic Synthesis, Riga, Latvia, ¹⁷BMWZ and Institute of Organic

Chemistry, Leibniz University Hannover, Hannover, Germany, ¹⁸Group of NMR-Based Structural Chemistry, Helmholtz Centre for Infection Research, Braunschweig, Germany, ¹⁹Multuser Center for Biomolecular Innovation (CMIB), Department of Physics, São Paulo State University (UNESP), São José do Rio Preto, Brazil, ²⁰Institute of Pharmaceutical Chemistry, Goethe University Frankfurt, Frankfurt am Main, Germany, ²¹Structural Genomics Consortium, Buchmann Institute for Molecular Life Sciences, Frankfurt am Main, Germany, ²²Department of Molecular, Cellular, and Biomedical Sciences, University of New Hampshire, Durham, NH, United States, ²³Department of Molecular Biology and Biochemistry, University of California, Irvine, CA, United States, ²⁴Magnetic Resonance Centre (CERM), University of Florence, Sesto Fiorentino, Italy, ²⁵Department of Chemistry "Ugo Schiff", University of Florence, Sesto Fiorentino, Italy, ²⁶Department of Pharmacy, University of Patras, Patras, Greece, ²⁷Laboratory of Toxinology, Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, ²⁸Department of Molecular Biology and Biophysics, UConn Health, Farmington, CT, United States, ²⁹Signals GmbH & Co. KG, Frankfurt am Main, Germany, ³⁰Leibniz Institute on Aging—Fritz Lipmann Institute (FLI), Jena, Germany, ³¹Latvian Biomedical Research and Study Centre, Riga, Latvia, ³²"Rocasolano" Institute for Physical Chemistry (IQFR), Spanish National Research Council (CSIC), Madrid, Spain, ³³Institute of Biophysical Chemistry, Goethe University Frankfurt, Frankfurt am Main, Germany, ³⁴BG-4, Karlsruhe Institute of Technology, Karlsruhe, Germany, ³⁵Department of Biochemistry and National Magnetic Resonance Facility at Madison, University of Wisconsin-Madison, Madison, WI, United States, ³⁶Department of Biology, Technical University of Darmstadt, Darmstadt, Germany, ³⁷Institute of Biochemistry and Biotechnology, Charles Tanford Protein Centre, Martin Luther University Halle-Wittenberg, Halle/Saale, Germany

The highly infectious disease COVID-19 caused by the *Betacoronavirus* SARS-CoV-2 poses a severe threat to humanity and demands the redirection of scientific efforts and criteria to organized research projects. The international *COVID19-NMR* consortium seeks to provide such new approaches by gathering scientific expertise worldwide. In particular, making available viral proteins and RNAs will pave the way to understanding the SARS-CoV-2 molecular components in detail. The research in *COVID19-NMR* and the resources provided through the consortium are fully disclosed to accelerate access and exploitation. NMR investigations of the viral molecular components are designated to provide the essential basis for further work, including macromolecular interaction studies and high-throughput drug screening. Here, we present the extensive catalog of a holistic SARS-CoV-2 protein preparation approach based on the consortium's collective efforts. We provide protocols for the large-scale production of more than 80% of all SARS-CoV-2 proteins or essential parts of them. Several of the proteins were produced in more than one laboratory, demonstrating the high interoperability between NMR groups worldwide. For the majority of proteins, we can produce isotope-labeled samples of HSQC-grade. Together with several NMR chemical shift assignments made publicly available on covid19-nmr.com, we here provide highly valuable resources for the production of SARS-CoV-2 proteins in isotope-labeled form.

Keywords: COVID-19, SARS-CoV-2, nonstructural proteins, structural proteins, accessory proteins, intrinsically disordered region, cell-free protein synthesis, NMR spectroscopy

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, SCoV2) is the cause of the early 2020 pandemic coronavirus lung disease 2019 (COVID-19) and belongs to *Betacoronaviruses*, a genus of the *Coronaviridae* family covering the α - δ genera (Leao et al., 2020). The large RNA genome of SCoV2 has an intricate, highly condensed arrangement of coding sequences (Wu et al., 2020). Sequences starting with the main start codon contain an open reading frame 1 (ORF1), which codes for two distinct, large polypeptides (pp), whose relative abundance is governed by the action of an RNA pseudoknot structure element. Upon RNA folding, this element causes a -1 frameshift to allow the continuation of translation, resulting in the generation of a 7,096-amino acid 794 kDa polypeptide. If the pseudoknot is not formed, expression of the first ORF generates a 4,405-amino acid 490 kDa polypeptide. Both the short and long polypeptides translated from this ORF (pp1a and pp1ab, respectively) are posttranslationally cleaved by virus-encoded

proteases into functional, nonstructural proteins (nsps). ORF1a encodes eleven nsps, and ORF1ab additionally encodes the nsps 12–16. The downstream ORFs encode structural proteins (S, E, M, and N) that are essential components for the synthesis of new virus particles. In between those, additional proteins (accessory/auxiliary factors) are encoded, for which sequences partially overlap (Finkel et al., 2020) and whose identification and classification are a matter of ongoing research (Nelson et al., 2020; Pavesi, 2020). In total, the number of identified peptides or proteins generated from the viral genome is at least 28 on the evidence level, with an additional set of smaller proteins or peptides being predicted with high likelihood.

High-resolution studies of SCoV and SCoV2 proteins have been conducted using all canonical structural biology approaches, such as X-ray crystallography on proteases (Zhang et al., 2020) and methyltransferases (MTase) (Krafcikova et al., 2020), cryo-EM of the RNA polymerase (Gao et al., 2020; Yin et al., 2020), and liquid-state (Almeida et al., 2007; Serrano et al., 2009; Cantini et al., 2020; Gallo et al., 2020; Korn et al., 2020a; Korn et al., 2020b;

TABLE 1 | SCoV2 protein constructs expressed and purified, given with the genomic position and corresponding PDBs for construct design.

Protein genome position (nt) ^a	Trivial name construct expressed	Size (aa)	Boundaries	MW (kDa)	Homol. SCoV (%) ^b	Template PDB ^c	SCoV2 PDB ^d
nsp1 266–805	Leader	180		19.8	84		
	Full-length	180	1–180	19.8	83		
nsp2 806–2,719	Full-length	116	13–127	12.7	85	2GDT	7K7P
	638	70.5	68				
nsp3 2,720–8,554	C-terminal IDR (CtDR)	45	557–601	4.9	55		
	1,945	217.3	76				
a a b c c c c d e Y	Ub-like (Ubl) domain	111	1–111	12.4	79	2IDY	7KAG
	Ub-like (Ubl) domain + IDR	206	1–206	23.2	58		
	Macrodomain	170	207–376	18.3	74	6VXS	6VXS
	SUD-N	140	409–548	15.5	69	2W2G	
	SUD-NM	267	409–675	29.6	74	2W2G	
	SUD-M	125	551–675	14.2	82	2W2G	
	SUD-MC	195	551–743	21.9	79	2KQV	
	SUD-C	64	680–743	7.4	73	2KAF	
	Papain-like protease PL ^{pro}	318	743–1,060	36	83	6W9C	6W9C
	NAB	116	1,088–1,203	13.4	87	2K87	
nsp5 10,055–10,972	CoV-Y	308	1,638–1,945	34	89		
	Main protease (M^{pro})	306		33.7	96		
nsp7 11,843–12,091	Full-length ^e	306	1–306	33.7	96	6Y84	6Y84
	83	9.2	99				
nsp8 12,092–12,685	Full-length	83	1–83	9.2	99	6W1Q	6W1Q
	198	21.9	98				
nsp9 12,686–13,024	Full-length	198	1–198	21.9	97	6W1Q	6W1Q
	113	12.4	97				
nsp10 13,025–13,441	Full-length	113	1–113	12.4	97	6W4B	6W4B
	139	14.8	97				
nsp13 16,237–18,039	Full-length	139	1–139	14.8	97	6W4H	6W4H
	Helicase	601	66.9	100			
nsp14 18,040–19,620	Full-length	601	1–601	66.9	100	6ZSL	6ZSL
	Exonuclease/ methyltransferase	527	59.8	95			
	Full-length	527	1–527	59.8	95	5NFY	
nsp15 19,621–20,658	MTase domain	240	288–527	27.5	95		
	Endonuclease	346	38.8	89			
nsp16 20,659–21,552	Full-length	346	1–346	38.8	89	6W01	6W01
	Methyltransferase	298	33.3	93			
ORF3a 25,393–26,220	Full-length	298	1–298	33.3	93	6W4H	6W4H
	275	31.3	72				
ORF4 26,245–26,472	Full-length	275	1–275	31.3	72	6XDC	6XDC
	Envelope (E) protein	75	8.4	95			
ORF5 26,523–27,387	Full-length	75	1–75	8.4	95	5X29	7K3G
	Membrane glycoprotein (M)	222	25.1	91			
ORF6 27,202–27,387	Full-length	222	1–222	25.1	91		
	61	7.3	69				
	Full-length	61	1–61	7.3	69		

(Continued on following page)

TABLE 1 | (Continued) SCoV2 protein constructs expressed and purified, given with the genomic position and corresponding PDBs for construct design.

Protein genome position (nt) ^a	Trivial name construct expressed	Size (aa)	Boundaries	MW (kDa)	Homol. SCoV (%) ^b	Template PDB ^c	SCoV2 PDB ^d
ORF7a 27,394–27,759		121		13.7	85		
	Ectodomain (ED)	66	16–81	7.4	85	1XAK	6W37
ORF7b 27,756–27,887		43		5.2	85		
	Full-length	43	1–43	5.2	85		
ORF8 27,894–28,259		121		13.8	32		
	Full-length	121	1–121	13.8	32		
	ΔORF8 w/o signal peptide	106	16–121	12	41	7JTL	7JTL
ORF9a 28,274–29,533	Nucleocapsid (N)	419		45.6	91		
	IDR1-NTD-IDR2	248	1–248	26.5	90		
	NTD-SR	169	44–212	18.1	92		
	NTD	136	44–180	14.9	93	6YI3	6YI3
	CTD	118	247–364	13.3	96	2JW8	7C22
ORF9b 28,284–28,574		97		10.8	72		
	Full-length	97	1–97	10.8	72	6Z4U	6Z4U
ORF14 28,734–28,952		73		8	n.a		
	Full-length	73	1–73	8	n.a		
ORF10 29,558–29,674		38		4.4	29		
	Full-length	38	1–38	4.4	29		

^aGenome position in nt corresponding to SCoV2 NCBI reference genome entry NC_045512.2, identical to GenBank entry MN908947.3.

^bSequence identities to SCoV are calculated from an alignment with corresponding protein sequences based on the genome sequence of NCBI Reference NC_004718.3.

^cRepresentative PDB that was available at the beginning of construct design, either SCoV or SCoV2.

^dRepresentative PDB available for SCoV2 (as of December 2020).

^eAdditional point mutations in fl-construct have been expressed.

n.a.: not applicable.

Kubatova et al., 2020; Tonelli et al., 2020) and solid-state NMR spectroscopy of transmembrane (TM) proteins (Mandala et al., 2020). These studies have significantly improved our understanding on the functions of molecular components, and they all rely on the recombinant production of viral proteins in high amount and purity.

Apart from structures, purified SCoV2 proteins are required for experimental and preclinical approaches designed to understand the basic principles of the viral life cycle and processes underlying viral infection and transmission. Approaches range from studies on immune responses (Esposito et al., 2020), antibody identification (Jiang et al., 2020), and interactions with other proteins or components of the host cell (Bojkova et al., 2020; Gordon et al., 2020). These examples highlight the importance of broad approaches for the recombinant production of viral proteins.

The research consortium *COVID19-NMR* founded in 2020 seeks to support the search for antiviral drugs using an NMR-based screening approach. This requires the large-scale production of all druggable proteins and RNAs and their NMR resonance assignments. The latter will enable solution structure determination of viral proteins and RNAs for rational drug design and the fast mapping of compound binding sites. We have recently produced and determined secondary structures of SCoV2 RNA *cis*-regulatory elements in near completeness by NMR spectroscopy, validated by DMS-

MaPseq (Wacker et al., 2020), to provide a basis for RNA-oriented fragment screens with NMR.

We here compile a compendium of more than 50 protocols (see **Supplementary Tables S11–S123**) for the production and purification of 23 of the 30 SCoV2 proteins or fragments thereof (summarized in **Tables 1, 2**). We defined those 30 proteins as existing or putative ones to our current knowledge (see later discussion). This compendium has been generated in a coordinated and concerted effort between >30 labs worldwide (**Supplementary Table S1**), with the aim of providing pure mg amounts of SCoV2 proteins. Our protocols include the rational strategy for construct design (if applicable, guided by available homolog structures), optimization of expression, solubility, yield, purity, and suitability for follow-up work, with a focus on uniform stable isotope-labeling.

We also present protocols for a number of accessory and structural E and M proteins that could only be produced using wheat-germ cell-free protein synthesis (WG-CFPS). In SCoV2, accessory proteins represent a class of mostly small and relatively poorly characterized proteins, mainly due to their difficult behavior in classical expression systems. They are often found in inclusion bodies and difficult to purify in quantities adequate for structural studies. We thus here exploit cell-free synthesis, mainly based on previous reports on production and purification of viral membrane proteins in general (Fogeron et al., 2015b; Fogeron et al., 2017; Jirasko

TABLE 2 | Summary of SCoV2 protein production results in *Covid19-NMR*.

Construct expressed	Yields (mg/L) ^a or (mg/ml) ^b	Results	Comments	BMRB	Supplementary Material
nsp1 fl	5	NMR assigned	Expression only at >20°C; after 7 days at 25°C partial proteolysis	50620 ^d	SI1
GD	>0.5	HSQC	High expression; mainly insoluble; higher salt increases stability (>250 mM)		
nsp2 CIDR	0.7–1.5	NMR assigned	Assignment with His-tag shown in (Mompéan et al., 2020)	50687 ^c	SI2
nsp3 Ubl	0.7	HSQC	Highly stable over weeks; spectrum overlays with Ubl + IDR		SI3
Ubl + IDR	2–3	NMR assigned	Highly stable for >2 weeks at 25°C	50446 ^d	
Macrodomain	9	NMR assigned	Highly stable for >1 week at 25°C and > 2 weeks at 4°C	50387 ^d 50388 ^d	
SUD-N	14	NMR assigned	Highly stable for >10 days at 25°C	50448 ^d	
SUD-NM	17	HSQC	Stable for >1 week at 25°C		
SUD-M	8.5	NMR assigned	Significant precipitation during measurement; tendency to dimerize	50516 ^d	
SUD-MC	12	HSQC	Stable for >1 week at 25°C		
SUD-C	4.7	NMR assigned	Stable for >10 days at 25°C	50517 ^d	
PL ^{pro}	12	HSQC	Solubility-tag essential for expression; tendency to aggregate		
NAB	3.5	NMR assigned	Highly stable for >1 week at 25°C; stable for >5 weeks at 4°C	50334 ^d	
CoV-Y	12	HSQC	Low temperature (<25°C) and low concentrations (<0.2 mM) favor stability; gradual degradation at 25°C; lithium bromide in final buffer supports solubility		
nsp5 fl	55	HSQC	Impaired dimerization induced by artificial N-terminal residues		SI4
nsp7 fl	17	NMR assigned	Stable for several days at 35°C; stable for >1 month at 4°C	50337 ^d	SI5
nsp8 fl	17	HSQC	Concentration dependent aggregation; low concentrations favor stability		SI6
nsp9 fl	4.5	NMR assigned	Stable dimer for >4 months at 4°C and >2 weeks at 25°C	50621 ^d 50622 ^d 50513	SI7
nsp10 fl	15	NMR assigned	Zn ²⁺ addition during expression and purification increases protein stability; stable for >1 week at 25°C	50392	SI8
nsp13 fl	0.5	HSQC	Low expression; protein unstable; concentration above 20 µM not possible		SI9
nsp14 fl	6	Pure protein	Not above 50 µM; best storage: with 50% (v/v) glycerol; addition of reducing agents		SI10
MTase	10	Pure protein	As fl nsp14; high salt (>0.4 M) for increased stability; addition of reducing agents		
nsp15 fl	5	HSQC	Tendency to aggregate at 25°C		SI11
nsp16 fl	10	Pure protein	Addition of reducing agents; 5% (v/v) glycerol favorable; highly unstable		SI12
ORF3a fl	0.6	Pure protein	Addition of detergent during expression (0.05% Brij-58); stable protein		SI13
E protein fl	0.45	Pure protein	Addition of detergent during expression (0.05% Brij-58); stable protein		SI14

(Continued on following page)

TABLE 2 | (Continued) Summary of SCoV2 protein production results in *Covid19-NMR*.

Construct expressed	Yields (mg/L) ^a or (mg/ml) ^b	Results	Comments	BMRB	Supplementary Material
M Protein					
fl	0.33	Pure protein	Addition of detergent during expression (0.05% Brij-58); stable protein		SI15
ORF6					
fl	0.27	HSQC	Soluble expression without detergent; stable protein; no expression with STREP-tag at N-terminus		SI16
ORF7a					
ED	0.4	HSQC	Unpurified protein tends to precipitate during refolding, purified protein stable for 4 days at 25°C		SI17
ORF7b					
fl	0.6	HSQC	Tendency to oligomerize; solubilizing agents needed		SI18
fl	0.27	HSQC	Addition of detergent during expression (0.1% MNG-3); stable protein		
ORF8					
fl	0.62	HSQC	Tendency to oligomerize		SI19
ΔORF8	0.5	Pure protein			
N protein					
IDR1-NTD- IDR2	12	NMR assigned	High salt (>0.4 M) for increased stability	50618, 50619, 50558, 50557 ^d	SI20
NTD-SR	3	HSQC			
NTD	3	HSQC		34511	
CTD	2	NMR assigned	Stable dimer for >4 months at 4°C and >3 weeks at 30°C	50518 ^d	
ORF9b					
fl	0.64	HSQC	Expression without detergent, protein is stable		SI21
ORF14					
fl	0.43	HSQC	Addition of detergent during expression (0.05% Brij-58); stable in detergent but unstable on lipid reconstitution		SI22
ORF10					
fl	2	HSQC	Tendency to oligomerize; unstable upon tag cleavage		SI23

^aYields from bacterial expression represent the minimal protein amount in mg/L independent of the cultivation medium. *Italic values indicate yields from CFPS.*

^bYields from CFPS represent the minimal protein amount in mg/ml of wheat-germ extract.

^cCOVID19-nmr BMRB depositions yet to be released.

^dCOVID19-nmr BMRB depositions.

et al., 2020b). Besides yields compatible with structural studies, ribosomes in WG extracts further possess an increased folding capacity (Netzer and Hartl, 1997), favorable for those more complicated proteins.

We exemplify in more detail the optimization of protein production, isotope-labeling, and purification for proteins with different individual challenges: the nucleic acid-binding (NAB) domain of nsp3e, the main protease nsp5, and several auxiliary proteins. For the majority of produced and purified proteins, we achieve >95% purity and provide ¹⁵N-HSQC spectra as the ultimate quality measure. We also provide additional suggestions for challenging proteins, where our protocols represent a unique resource and starting point exploitable by other labs.

MATERIALS AND METHODS

Strains, Plasmids, and Cloning

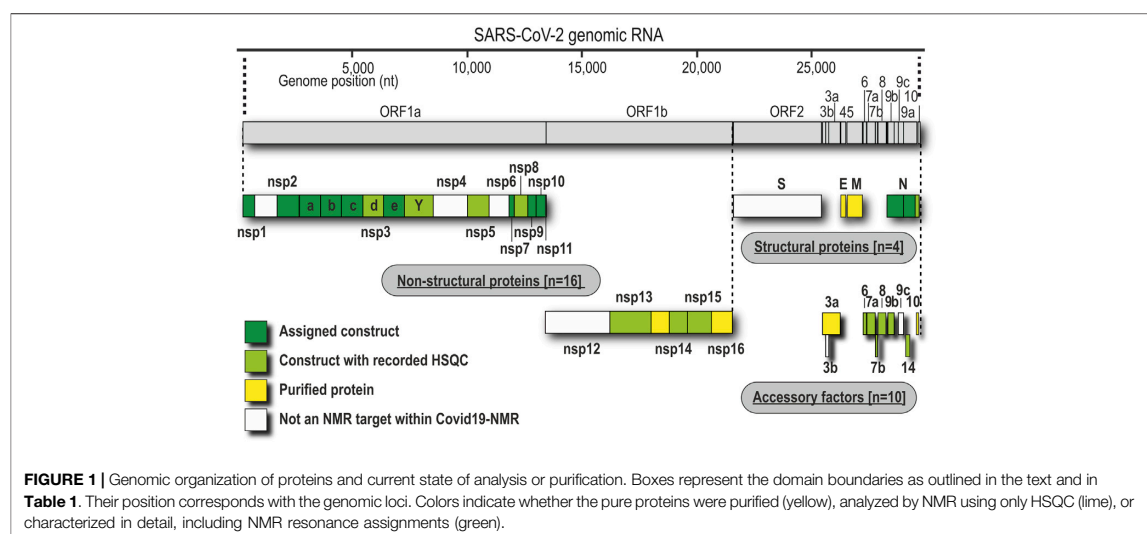
The rationale of construct design for all proteins can be found within the respective protocols in **Supplementary Tables S11–S123**. For bacterial production, *E. coli* strains and expression plasmids are given; for WG-CFPS, template

vectors are listed. Protein coding sequences of interest have been obtained as either commercial, codon-optimized genes or, for shorter ORFs and additional sequences, annealed from oligonucleotides prior to insertion into the relevant vector. Subcloning of inserts, adjustment of boundaries, and mutations of genes have been carried out by standard molecular biology techniques. All expression plasmids can be obtained upon request from the *COVID19-NMR* consortium (<https://covid19-nmr.com/>), including information about coding sequences, restriction sites, fusion tags, and vector backbones.

Protein Production and Purification

For SCoV2 proteins, we primarily used heterologous production in *E. coli*. Detailed protocols of individual full-length (fl) proteins, separate domains, combinations, or particular expression constructs as listed in **Table 1** can be found in the (**Supplementary Tables S11–S123**).

The ORF3a, ORF6, ORF7b, ORF8, ORF9b, and ORF14 accessory proteins and the structural proteins M and E were produced by WG-CFPS as described in the **Supplementary Material**. In brief, transcription and translation steps have



been performed separately, and detergent has been added for the synthesis of membrane proteins as described previously (Takai et al., 2010; Fogeron et al., 2017).

NMR Spectroscopy

All amide correlation spectra, either HSQC- or TROSY-based, are representative examples. Details on their acquisition parameters and the raw data are freely accessible through <https://covid19-nmr.de> or upon request.

RESULTS

In the following, we provide protocols for the purification of SCoV2 proteins sorted into 1) nonstructural proteins and 2) structural proteins together with accessory ORFs. **Table 1** shows an overview of expression constructs. We use a consequent terminology of those constructs, which is guided by domains, intrinsically disordered regions (IDRs) or other particularly relevant sequence features within them. This study uses the SCoV2 NCBI reference genome entry NC_045512.2, identical to GenBank entry MN908947.3 (Wu et al., 2020), unless denoted differently in the respective protocols. Any relevant definition of boundaries can also be found in the SI protocols.

As applicable for a major part of our proteins, we further define a standard procedure for the purification of soluble His-tagged proteins that are obtained through the sequence of IMAC, TEV/Ulp1 Protease cleavage, Reverse IMAC, and Size-exclusion chromatography, eventually with individual alterations, modifications, or additional steps. For convenient reading, we will thus use the abbreviation IPRS to avoid redundant protocol description. Details for every protein,

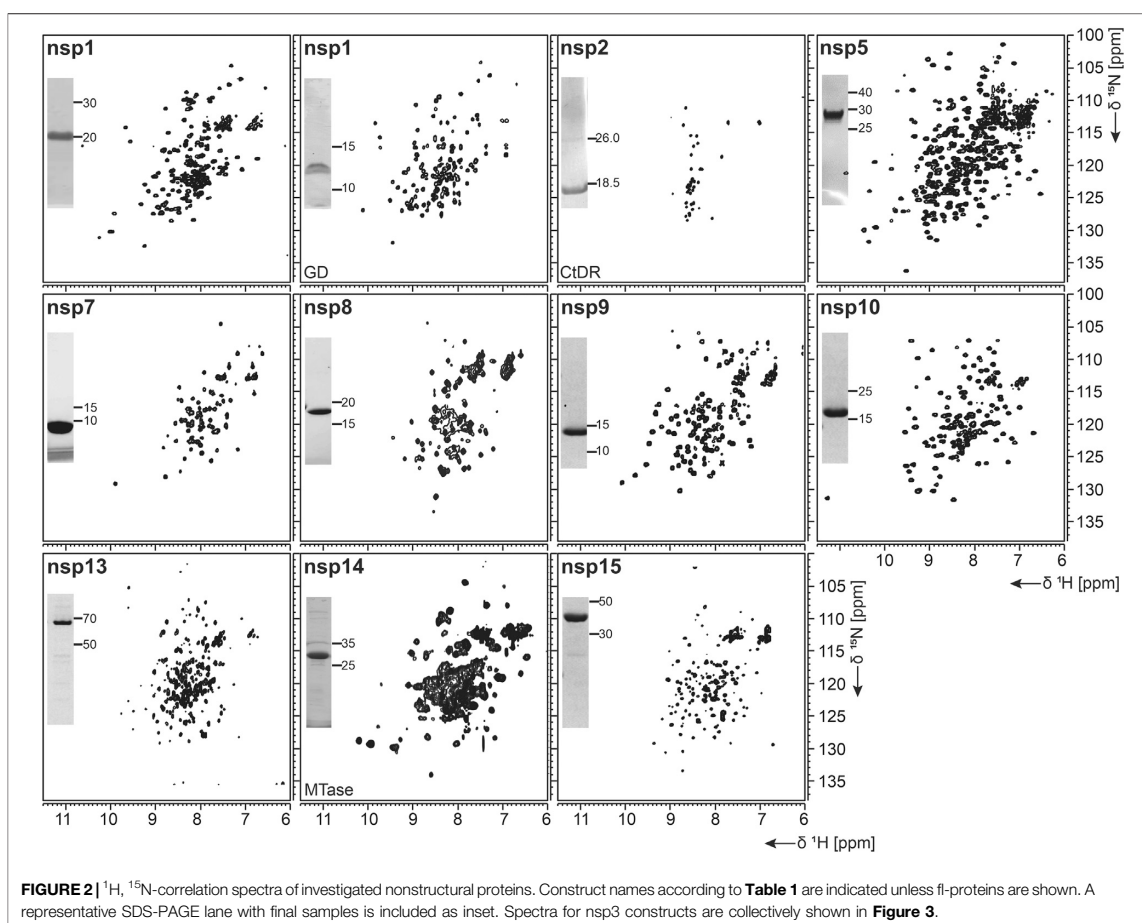
including detailed expression conditions, buffers, incubation times, supplements, storage conditions, yields, and stability, can be found in the respective **Supplementary Tables S11–S123** (see also **Supplementary Tables S1, S2**) and **Tables 1, 2**.

Nonstructural Proteins

We have approached and challenged the recombinant production of a large part of the SCoV2 nsps (**Figure 1**), with great success (**Table 2**). We excluded nsp4 and nsp6 (TM proteins), which are little characterized and do not reveal soluble, folded domains by prediction (Oostra et al., 2007; Oostra et al., 2008). The function of the very short (13 aa) nsp11 is unknown, and it seems to be a mere copy of the nsp12 amino-terminal residues, remaining as a protease cleavage product of ORF1a. Further, we left out the RNA-dependent RNA polymerase nsp12 in our initial approach because of its size (>100 kDa) and known unsuitability for heterologous recombinant production in bacteria. Work on NMR-suitable nsp12 bacterial production is ongoing, while other expert labs have succeeded in purifying nsp12 for cryo-EM applications in different systems (Gao et al., 2020; Hillen et al., 2020). For the remainder of nsps, we here provide protocols for fl-proteins or relevant fragments of them.

nsp1

nsp1 is the very N-terminus of the polyproteins pp1a and pp1ab and one of the most enigmatic viral proteins, expressed only in α - and β -CoVs (Narayanan et al., 2015). Interestingly, nsp1 displays the highest divergence in sequence and size among different CoVs, justifying it as a genus-specific marker (Snijder et al., 2003). It functions as a host shutoff factor by suppressing innate immune functions and host gene expression (Kamitani et al.,



2006; Narayanan et al., 2008; Schubert et al., 2020). This suppression is achieved by an interaction of the nsp1 C-terminus with the mRNA entry tunnel within the 40 S subunit of the ribosome (Schubert et al., 2020; Thoms et al., 2020).

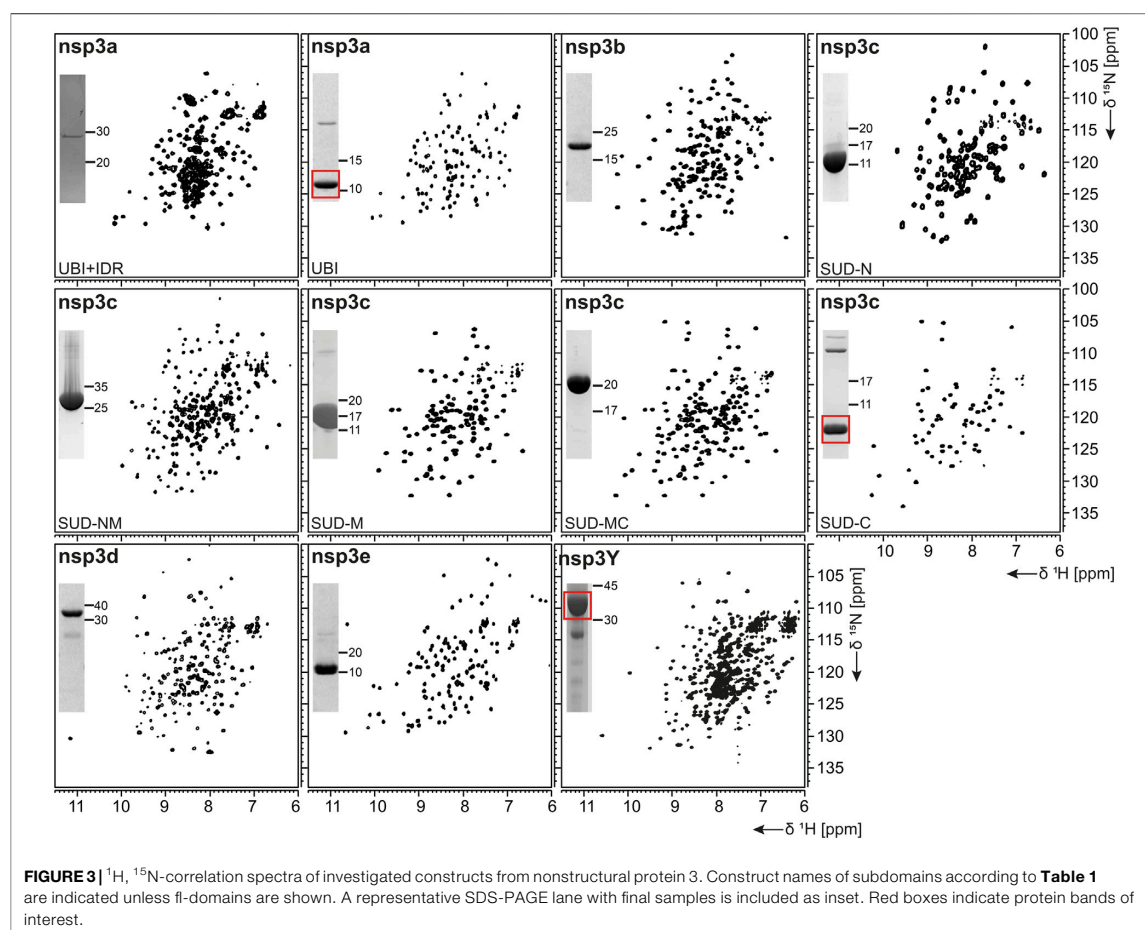
As summarized in **Table 1**, fl-domain boundaries of nsp1 were chosen to contain the first 180 amino acids, in analogy to its closest homolog from SCoV (Snijder et al., 2003). In addition, a shorter construct was designed, encoding only the globular core domain (GD, aa 13–127) suggested by the published SCoV nsp1 NMR structure (Almeida et al., 2007). His-tagged fl nsp1 was purified using the IPRS approach. Protein quality was confirmed by the available HSQC spectrum (**Figure 2**). Despite the flexible C-terminus, we were able to accomplish a near-complete backbone assignment (Wang et al., 2021).

Interestingly, the nsp1 GD was found to be problematic in our hands despite good expression. We observed insolubility, although buffers were used according to the

homolog SCoV nsp1 GD (Almeida et al., 2007). Nevertheless, using a protocol comparable to the one for fl nsp1, we were able to record an HSQC spectrum proving a folded protein (**Figure 2**).

nsp2

nsp2 has been suggested to interact with host factors involved in intracellular signaling (Cornillez-Ty et al., 2009; Davies et al., 2020). The precise function, however, is insufficiently understood. Despite its potential dispensability for viral replication in general, it might be a valuable model to gain insights into virulence due to its possible involvement in the regulation of global RNA synthesis (Graham et al., 2005). We provide here a protocol for the purification of the C-terminal IDR (CtDR) of nsp2 from residues 557 to 601, based on disorder predictions [PrDOS (Ishida and Kinoshita, 2007)]. The His-Trx-tagged peptide was purified by IPRS. Upon dialysis, two IEC steps were performed: first anionic and then cationic, with good final yields (**Table 1**). Stability and



purity were confirmed by an HSQC spectrum (**Figure 2**) and a complete backbone assignment (Mompean et al., 2020; **Table 2**).

nsp3

nsp3, the largest nsp (Snijder et al., 2003), is composed of a plethora of functionally related, yet independent, subunits. After cleavage of nsp3 from the fl ORF1-encoded polypeptide chain, it displays a 1945-residue multidomain protein, with individual functional entities that are subclassified from nsp3a to nsp3e followed by the ectodomain embedded in two TM regions and the very C-terminal CoV-Y domain. The soluble nsp3a-3e domains are linked by various types of linkers with crucial roles in the viral life cycle and are located in the so-called viral cytoplasm, which is separated from the host cell after budding off the endoplasmic reticulum and contains the viral RNA (Wolff et al., 2020). Remarkably, the nsp3c substructure comprises three subdomains, making nsp3

the most complex SCoV2 protein. The precise function and eventual RNA-binding specificities of nsp3 domains are not yet understood. We here focus on the nsp3 domains a-e and provide elaborated protocols for additional constructs carrying relevant linkers or combinations of domains (**Table 1**). Moreover, we additionally present a convenient protocol for the purification of the C-terminal CoV-Y domain.

nsp3a

The N-terminal portion of nsp3 is comprised of a ubiquitin-like (Ubl) structured domain and a subsequent acidic IDR. Besides its ability to bind ssRNA (Serrano et al., 2007), nsp3a has been reported to interact with the nucleocapsid (Hurst et al., 2013; Khan et al., 2020), playing a potential role in virus replication. We here provide protocols for the purification of both the Ubl (aa 1-111) and fl nsp3a (aa 1-206), including the acidic IDR (Ubl + IDR **Table 1**). Domain boundaries were defined similar to the published NMR structure of SCoV nsp3a (Serrano et al., 2007). His-

tagged nsp3a Ubl + IDR and GST-tagged nsp3a Ubl were each purified via the IPRS approach. nsp3a Ubl yielded mM sample concentrations and displayed a well-dispersed HSQC spectrum (**Figure 3**). Notably, the herein described protocol also enables purification of fl nsp3a (Ubl + IDR) (**Tables 1, 2**). Despite the unstructured IDR overhang, the excellent protein quality and stability allowed for near-complete backbone assignment [**Figure 3**, (Salvi et al., 2021)].

nsp3b

nsp3b is an ADP-ribose phosphatase macrodomain and potentially plays a key role in viral replication. Moreover, the de-ADP-ribosylation function of nsp3b protects SCoV2 from antiviral host immune response, making nsp3b a promising drug target (Frick et al., 2020). As summarized in **Table 1**, the domain boundaries of the herein investigated nsp3b are residues 207–376 of the nsp3 primary sequence and were identical to available crystal structures with PDB entries 6YWM and 6YWL (unpublished). For purification, we used the IPRS approach, which yielded pure fl nsp3b (**Table 2**). Fl nsp3b displays well-dispersed HSQC spectra, making this protein an amenable target for NMR structural studies. In fact, we recently reported near-to-complete backbone assignments for nsp3b in its apo and ADP-ribose-bound form (Cantini et al., 2020).

nsp3c

The SARS unique domain (SUD) of nsp3c has been described as a distinguishing feature of SCoVs (Snijder et al., 2003). However, similar domains in more distant CoVs, such as MHV or MERS, have been reported recently (Chen et al., 2015; Kusov et al., 2015). nsp3c comprises three distinct globular domains, termed SUD-N, SUD-M, and SUD-C, according to their sequential arrangement: N-terminal (N), middle (M), and C-terminal (C). SUD-N and SUD-M develop a macrodomain fold similar to nsp3b and are described to bind G-quadruplexes (Tan et al., 2009), while SUD-C preferentially binds to purine-containing RNA (Johnson et al., 2010). Domain boundaries for SUD-N and SUD-M and for the tandem-domain SUD-NM were defined in analogy to the SCoV homolog crystal structure (Tan et al., 2009). Those for SUD-C and the tandem SUD-MC were based on NMR solution structures of corresponding SCoV homologs (**Table 1**) (Johnson et al., 2010). SUD-N, SUD-C, and SUD-NM were purified using GST affinity chromatography, whereas SUD-M and SUD-MC were purified using His affinity chromatography. Removal of the tag was achieved by thrombin cleavage and final samples of all domains were prepared subsequent to size-exclusion chromatography (SEC). Except for SUD-M, all constructs were highly stable (**Table 2**). Overall protein quality allowed for the assignment of backbone chemical shifts for the three single domains (Gallo et al., 2020) and good resolved HSQC spectra also for the tandem domains (**Figure 3**).

nsp3d

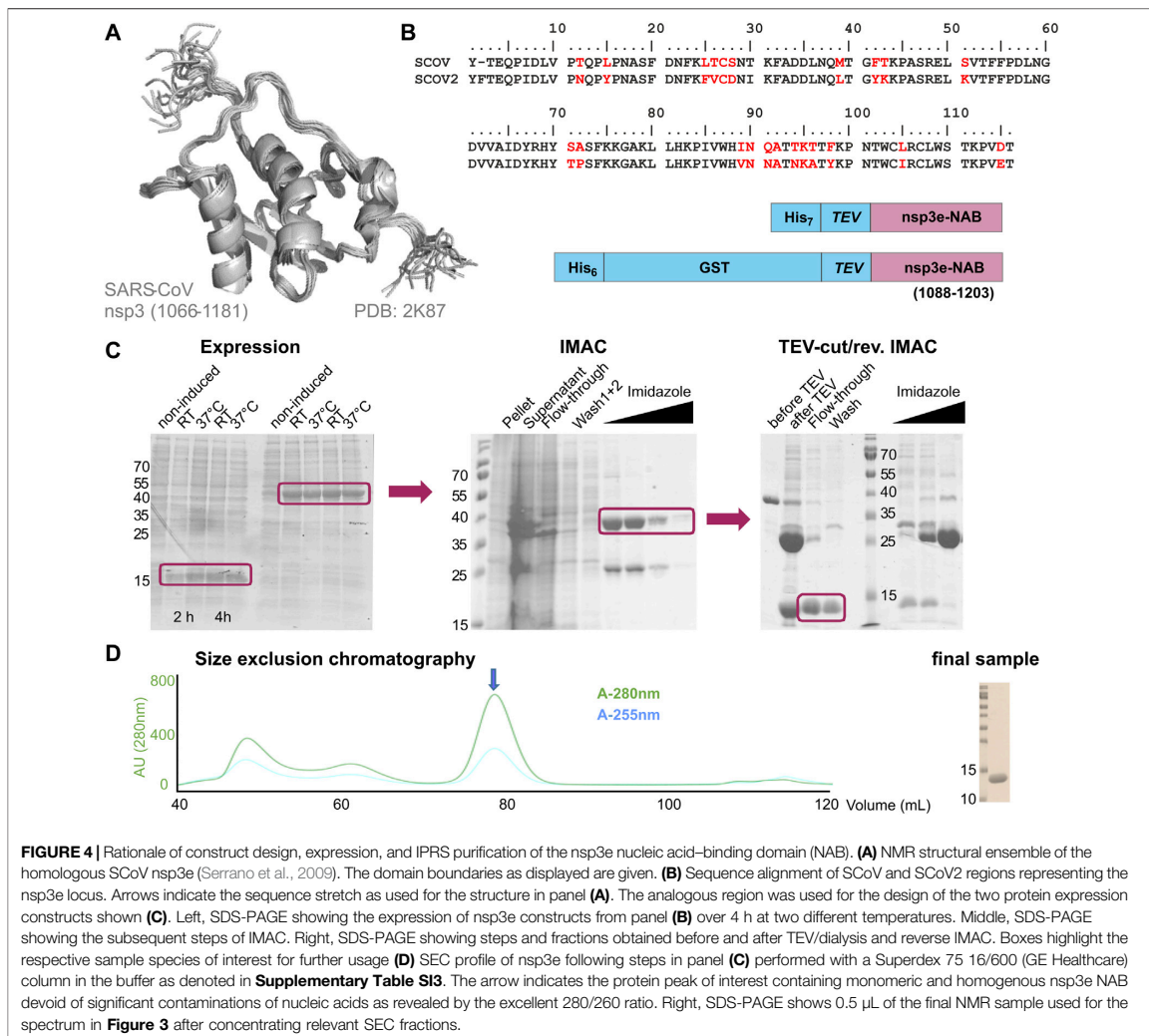
nsp3d comprises the papain-like protease (PL^{Pro}) domain of nsp3 and, hence, is one of the two SCoV2 proteases that are responsible for processing the viral polypeptide chain and generating functional proteins (Shin et al., 2020). The domain boundaries of PL^{Pro} within nsp3 are set by residues 743 and 1,060 (**Table 1**). The protein is particularly challenging, as it is prone to misfolding and rapid precipitation. We prepared His-tagged and His-SUMO-tagged PL^{Pro}. The His-tagged version mainly remained in the insoluble fraction. Still, mg quantities could be purified from the soluble fraction, however, greatly misfolded. Fusion to SUMO significantly enhanced protein yield of soluble PL^{Pro}. The His-SUMO-tag allowed simple IMAC purification, followed by cleavage with Ulp1 and isolation of cleaved PL^{Pro} via a second IMAC. A final purification step using gel filtration led to pure PL^{Pro} of both unlabeled and 15N-labeled species (**Table 2**). The latter has allowed for the acquisition of a promising amide correlation spectrum (**Figure 3**).

nsp3e

nsp3e is unique to *Betacoronaviruses* and consists of a nucleic acid-binding domain (NAB) and the so-called group 2-specific marker (G2M) (Neuman et al., 2008). Structural information is rare; while the G2M is predicted to be intrinsically disordered (Lei et al., 2018); the only available experimental structure of the nsp3e NAB was solved from SCoV by the Wüthrich lab using solution NMR (Serrano et al., 2009). We here used this structure for a sequence-based alignment to derive reasonable domain boundaries for the SCoV2 nsp3e NAB (**Figures 4A,B**). The high sequence similarity suggested using nsp3 residues 1,088–1,203 (**Table 1**). This polypeptide chain was encoded in expression vectors comprising His- and His-GST tags, both cleavable by TEV protease. Both constructs showed excellent expression, suitable for the IPRS protocol (**Figure 4C**). Finally, a homogenous NAB species, as supported by the final gel of pooled samples (**Figure 4D**), was obtained. The excellent protein quality and stability are supported by the available HSQC (**Figure 3**) and a published backbone assignment (Korn et al., 2020a).

nsp3Y

nsp3Y is the most C-terminal domain of nsp3 and exists in all coronaviruses (Neuman et al., 2008; Neuman, 2016). Together, though, with its preceding regions G2M, TM 1, the ectodomain, TM2, and the Y1-domain, it has evaded structural investigations so far. The precise function of the CoV-Y domain remains unclear, but, together with the Y1-domain, it might affect binding to nsp4 (Hagemeijer et al., 2014). We were able to produce and purify nsp3Y (CoV-Y) comprising amino acids 1,638–1,945 (**Table 1**), yielding 12 mg/L with an optimized protocol that keeps the protein in a final NMR buffer containing HEPES and lithium bromide. Although the protein still shows some tendency to aggregate and degrade (**Table 2**), and despite its relatively large size, the spectral quality is excellent (**Figure 3**). nsp3 CoV-Y appears suitable for an NMR backbone



assignment carried out at lower concentrations in a deuterated background (ongoing).

nsp5

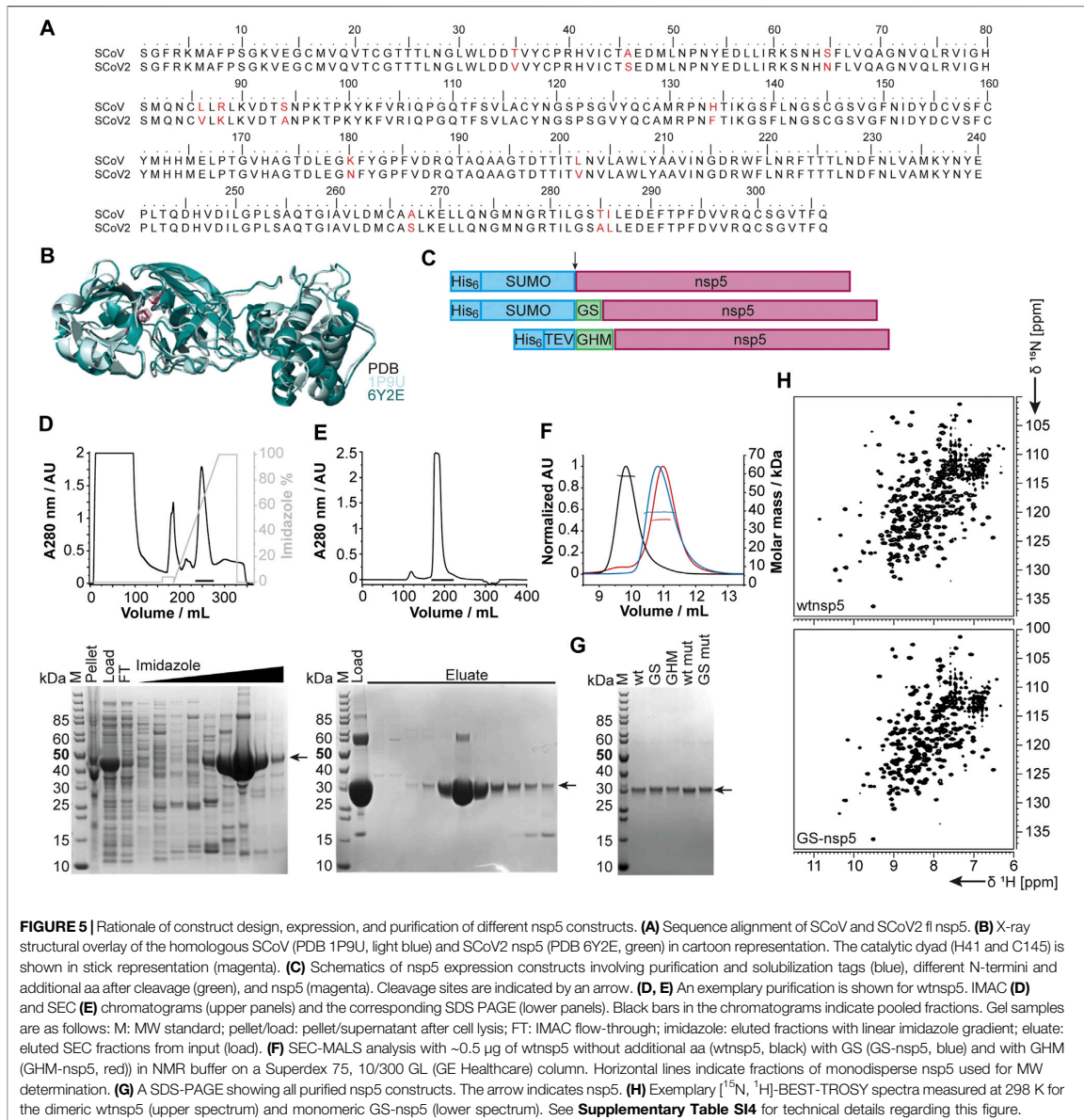
The functional main protease nsp5 (M^{Pro}) is a dimeric cysteine protease (Ullrich and Nitsche, 2020). Amino acid sequence and 3D structure of SCoV [PDB 1P9U (Anand et al., 2003)] and SCoV2 (PDB 6Y2E [Zhang et al., 2020]) homologs are highly conserved (**Figures 5A,B**). The dimer interface involves the N-termini of both monomers, which puts considerable constraints on the choice of protein sequence for construct design regarding the N-terminus.

We thus designed different constructs differing in the N-terminus: the native N-terminus (wt), a GS mutant with the additional N-terminal residues glycine and serine as His-SUMO

fusion, and a GHM mutant with the amino acids glycine, histidine, and methionine located at the N-terminus with His-tag and TEV cleavage site (**Figure 5C**). Purification of all proteins via the IPRS approach (**Figures 5D,E**) yielded homogenous and highly pure protein, analyzed by PAGE (**Figure 5G**), mass spectrometry, and 2D [^{15}N , 1H]-BEST TROSY spectra (**Figure 5H**). Final yields are summarized in **Table 2**.

nsp7 and nsp8

Both nsp7 and nsp8 are auxiliary factors of the polymerase complex together with the RNA-dependent RNA polymerase nsp12 and have high sequence homology with SCoV (100% and 99%, respectively) (Gordon et al., 2020). For nsp7 in complex with nsp8 or for nsp8 alone, additional functions in RNA synthesis priming have been proposed (Tvarogova et al., 2019;



Konkolova et al., 2020). In a recent study including an RNA-substrate-bound structure (Hillen et al., 2020), both proteins (with two molecules of nsp8 and one molecule of nsp7 for each nsp12 RNA polymerase) were found to be essential for polymerase activity in SCoV2. For both fl-proteins, a previously established expression and IPRS purification strategy for the SCoV proteins (Kirchdoerfer and Ward, 2019) was successfully transferred, which resulted in decent yields of

reasonably stable proteins (Table 2). Driven by its intrinsically oligomeric state, nsp8 showed some tendency toward aggregation, limiting the available sample concentration. The higher apparent molecular weight and limited solubility are also reflected in the success of NMR experiments. While we succeeded in a complete NMR backbone assignment of nsp7 (Tonelli et al., 2020), the quality of the spectra obtained for nsp8 is currently limited to the HSQC presented in Figure 2.

nsp9

The 12.4 kDa ssRNA-binding nsp9 is highly conserved among *Betacoronaviruses*. It is a crucial part of the viral replication machinery (Miknis et al., 2009), possibly targeting the 3'-end stem-loop II (s2m) of the genome (Robertson et al., 2005). nsp9 adopts a fold similar to oligonucleotide/oligosaccharide-binding proteins (Egloff et al., 2004), and structural data consistently uncovered nsp9 to be dimeric in solution (Egloff et al., 2004; Sutton et al., 2004; Miknis et al., 2009; Littler et al., 2020). Dimer formation seems to be a prerequisite for viral replication (Miknis et al., 2009) and influences RNA-binding (Sutton et al., 2004), despite a moderate affinity for RNA *in vitro* (Littler et al., 2020).

Based on the early available crystal structure of SCoV2 nsp9 (PDB 6W4B, unpublished), we used the 113 aa fl sequence of nsp9 for our expression construct (Table 1). Production of either His- or His-GST-tagged fl nsp9 yielded high amounts of soluble protein in both natural abundance and ¹³C- and ¹⁵N-labeled form. Purification *via* the IPRS approach enabled us to separate fl nsp9 in different oligomer states. The earliest eluted fraction represented higher oligomers, was contaminated with nucleic acids and was not possible to concentrate above 2 mg/ml. This was different for the subsequently eluting dimeric fl nsp9 fraction, which had a A260/280 ratio of below 0.7 and could be concentrated to >5 mg/ml (Table 2). The excellent protein quality and stability are supported by the available HSQC (Figure 2), and a near-complete backbone assignment (Dudas et al., 2021).

nsp10

The last functional protein encoded by ORF1a, nsp10, is an auxiliary factor for both the methyltransferase/exonuclease nsp14 and the 2'-O-methyltransferase (MTase) nsp16. However, it is required for the MTase activity of nsp16 (Krafcikova et al., 2020), it confers exonuclease activity to nsp14 in the RNA polymerase complex in SCoV (Ma et al., 2015). It contains two unusual zinc finger motifs (Joseph et al., 2006) and was initially proposed to comprise RNA-binding properties. We generated a construct (Table 1) containing an expression and affinity purification tag on the N-terminus as reported for the SCoV variant (Joseph et al., 2006). Importantly, additional Zn²⁺ ions present during expression and purification stabilize the protein significantly (Kubatova et al., 2020). The yield during isotope-labeling was high (Table 2), and tests in unlabeled rich medium showed the potential for yields exceeding 100 mg/L. These characteristics facilitated in-depth NMR analysis and a backbone assignment (Kubatova et al., 2020).

nsp13

nsp13 is a conserved ATP-dependent helicase that has been characterized as part of the RNA synthesis machinery by binding to nsp12 (Chen et al., 2020b). It represents an interesting drug target, for which the available structure (PDB 6ZSL) serves as an excellent basis (Table 1). The precise molecular function, however, has remained enigmatic since it is not clear whether the RNA unwinding function is required for making ssRNA accessible for RNA synthesis (Jia et al., 2019) or whether it is required for proofreading and backtracking (Chen

et al., 2020b). We obtained pure protein using a standard expression vector, generating a His-SUMO-tagged protein. Following Ulp1 cleavage, the protein showed limited protein stability in the solution (Table 2).

nsp14

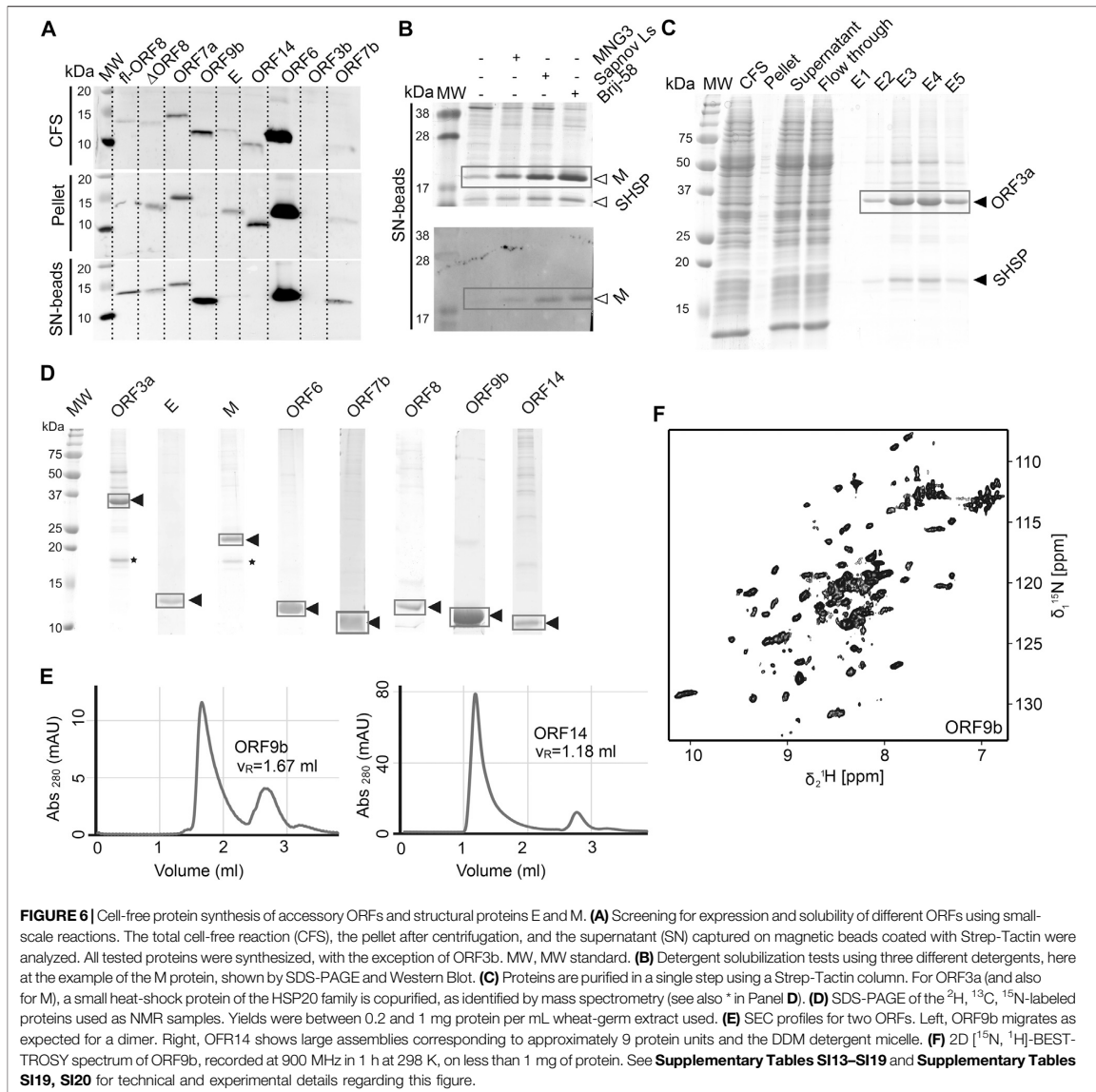
nsp14 contains two domains: an N-terminal exonuclease domain and a C-terminal MTase domain (Ma et al., 2015). The exonuclease domain interacts with nsp10 and provides part of the proofreading function that supports the high fidelity of the RNA polymerase complex (Robson et al., 2020). Several unusual features, such as the unusual zinc finger motifs, set it apart from other DEDD-type exonucleases (Chen et al., 2007), which are related to both nsp10 binding and catalytic activity. The MTase domain modifies the N7 of the guanosine cap of genomic and subgenomic viral RNAs, which is essential for the translation of viral proteins (Thoms et al., 2020). The location of this enzymatic activity within the RNA synthesis machinery ensures that newly synthesized RNA is rapidly capped and thus stabilized. As a strategy, we used constructs, which allow coexpression of both nsp14 and nsp10 (pRSFDuet and pETDuet, respectively). Production of isolated fl nsp14 was successful, however, with limited yield and stability (Table 2). Expression of the isolated MTase domain resulted in soluble protein with 27.5 kDa mass that was amenable to NMR characterization (Figure 2), although only under reducing conditions and in the presence of high (0.4 M) salt concentration.

nsp15

The poly-U-specific endoribonuclease nsp15 was one of the very first SCoV2 structures deposited in the PDB [6VWW, (Kim et al., 2020)]. Its function has been suggested to be related to the removal of U-rich RNA elements, preventing recognition by the innate immune system (Deng et al., 2017), even though the precise mechanism remains to be established. The exact role of the three domains (N-terminal, middle, and C-terminal catalytic domain) also remains to be characterized in more detail (Kim et al., 2020). Here, the sufficient yield of fl nsp15 during expression supported purification of pure protein, which, however, showed limited stability in solution (Table 2).

nsp16

The MTase reaction catalyzed by nsp16 is dependent on nsp10 as a cofactor (Krafcikova et al., 2020). In this reaction, the 2'-OH group of nucleotide +1 in genomic and subgenomic viral RNA is methylated, preventing recognition by the innate immune system. Since both nsp14 and nsp16 are in principle susceptible to inhibition by MTase inhibitors, a drug targeting both enzymes would be highly desirable (Bouvet et al., 2010). nsp16 is the last protein being encoded by ORF1ab, and only its N-terminus is formed by cleavage by the M^{Pro} nsp5. Employing a similar strategy to that for nsp14, nsp16 constructs were designed with the possibility of nsp10 coexpression. Expression of fl nsp16 resulted in good yields, when expressed both isolated and together with nsp10. The protein, however, is in either case



unstable in solution and highly dependent on reducing buffer conditions (Table 2). The purification procedures of nsp16 were adapted with minor modifications from a previous X-ray crystallography study (Rosas-Lemus et al., 2020).

Structural Proteins and Accessory ORFs

Besides establishing expression and purification protocols for the nsps, we also developed protocols and obtained pure mg quantities of the SCoV2 structural proteins E, M, and N, as well as literally all accessory proteins. With the exception of the relatively well-behaved nucleocapsid (N) protein, SCoV2 E, M,

and the remaining accessory proteins represent a class of mostly small and relatively poorly characterized proteins, mainly due to their difficult behavior in classical expression systems.

We used wheat-germ cell-free protein synthesis (WG-CFPS) for the successful production, solubilization, purification, and, in part, initial NMR spectroscopic investigation of ORF3a, ORF6, ORF7b, ORF8, ORF9b, and ORF14 accessory proteins, as well as E and M in mg quantities using the highly efficient translation machinery extracted from wheat-germs (Figures 6A–D).

ORF3a

The protein from ORF3a in SCoV2 corresponds to the accessory protein 3a in SCoV, with homology of more than 70% (Table 1). It has 275 amino acids, and its structure has recently been determined (Kern et al., 2020). The structure of SCoV2 3a displays a dimer, but it can also form higher oligomers. Each monomer has three TM helices and a cytosolic β -strand rich domain. SCoV2 ORF3a is a cation channel, and its structure has been solved by electron microscopy in nanodiscs. In SCoV, 3a is a structural component and was found in recombinant virus-like particles (Liu et al., 2014), but is not explicitly needed for their formation. The major challenge for NMR studies of this largest accessory protein is its size, independent of its employment in solid state or solution NMR spectroscopy.

As most other accessory proteins described in the following, ORF3a has been produced using WG-CFPS and was expressed in soluble form in the presence of Brij-58 (Figure 6C). It is copurified with a small heat-shock protein of the HSP20 family from the wheat-germ extract. The protocol described here is highly similar to that of the other cell-free synthesized accessory proteins. Where NMR spectra have been reported, the protein has been produced in a ^2H , ^{13}C , ^{15}N uniformly labeled form; otherwise, natural abundance amino acids were added to the reaction. The proteins were further affinity-purified in one step using Strep-Tactin resin, through the Strep-tag II fused to their N- or C-terminus. For membrane proteins, protein synthesis and also purification were done in the presence of detergent.

About half a milligram of pure protein was generally obtained per mL of extract, and up to 3 ml wheat-germ extract have been used to prepare NMR samples.

ORF3b

The ORF3b protein is a putative protein stemming from a short ORF (57 aa) with no homology to existing SCoV proteins (Chan et al., 2020). Indeed, ORF3b gene products of SCoV2 and SCoV are considerably different, with one of the distinguishing features being the presence of premature stop codons, resulting in the expression of a drastically shortened ORF3b protein (Konno et al., 2020). However, the SCoV2 nucleotide sequence after the stop codon shows a high similarity to the SCoV ORF3b. Different C-terminal truncations seem to play a role in the interferon-antagonistic activity of ORF3b (Konno et al., 2020). ORF3b is the only protein that, using WG-CFPS, was not synthesized at all; i.e., it was neither observed in the total cell-free reaction nor in supernatant or pellet. This might be due to the premature stop codon, which was not considered. Constructs of ORF3b thus need to be redesigned.

ORF4 (Envelope Protein, E)

The SCoV2 envelope (E) protein is a small (75 amino acids), integral membrane protein involved in several aspects of the virus' life cycle, such as assembly, budding, envelope formation, and pathogenicity, as recently reviewed in (Schoeman and Fielding, 2020). Structural models for SCoV (Surya et al.,

2018) and the TM helix of SCoV2 (Mandala et al., 2020) E have been established. The structural models show a pentamer with a TM helix. The C-terminal part is polar, with charged residues interleaved, and is positioned on the membrane surface in SCoV. E was produced in a similar manner to ORF3a, using the addition of detergent to the cell-free reaction.

ORF5 (Membrane Glycoprotein, M)

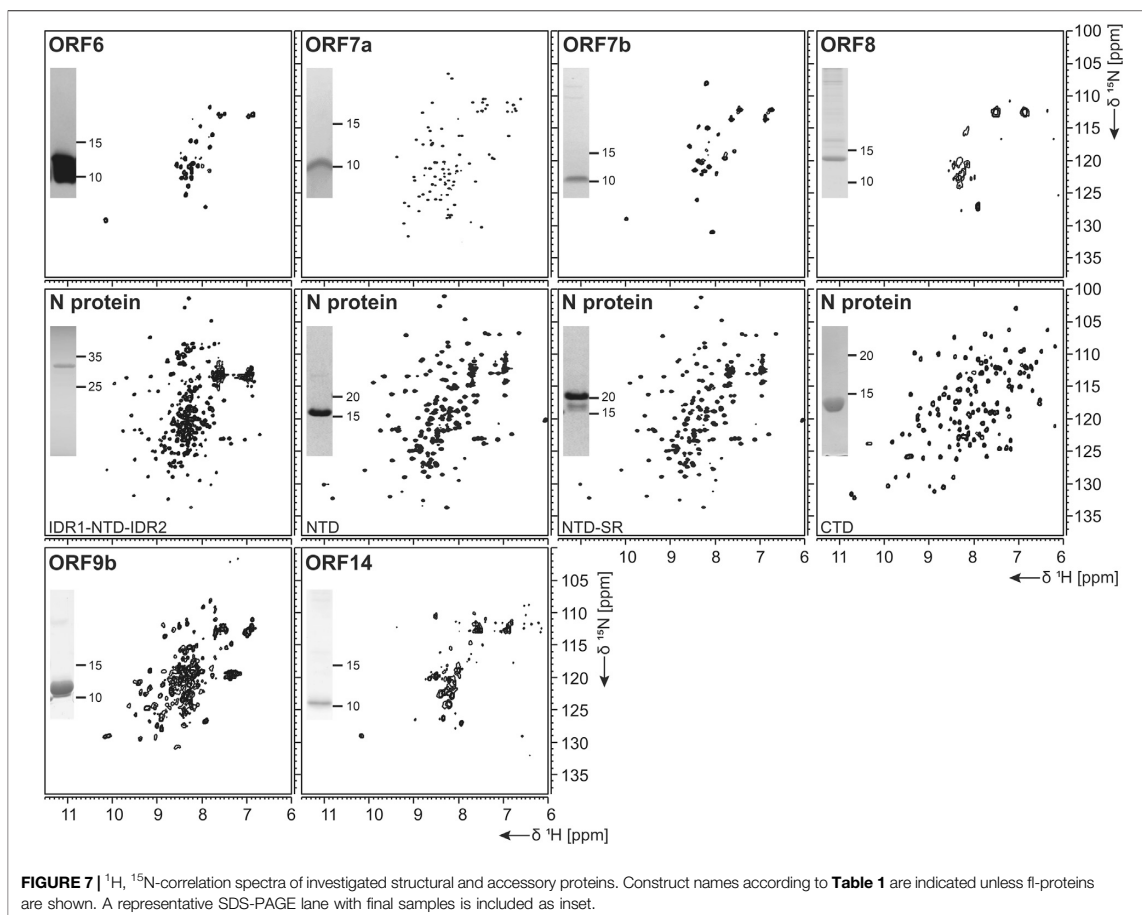
The M protein is the most abundant protein in the viral envelope and is believed to be responsible for maintaining the virion in its characteristic shape (Huang et al., 2004). M is a glycoprotein and sequence analyses predict three domains: A C-terminal endodomain, a TM domain with three predicted helices, and a short N-terminal ectodomain. M is essential for viral particle assembly. Intermolecular interactions with the other structural proteins, N and S to a lesser extent, but most importantly E (Vennema et al., 1996), seem to be central for virion envelope formation in coronaviruses, as M alone is not sufficient. Evidence has been presented that M could adopt two conformations, elongated and compact, and that the two forms fulfill different functions (Neuman et al., 2011). The lack of more detailed structural information is in part due to its small size, close association with the viral envelope, and a tendency to form insoluble aggregates when perturbed (Neuman et al., 2011). The M protein is readily produced using cell-free synthesis in the presence of detergent; as ORF3a, it is copurified with a small heat-shock protein of the HSP20 family (Figure 6B). Membrane-reconstitution will likely be necessary to study this protein.

ORF6

The ORF6 protein is incorporated into viral particles and is also released from cells (Huang et al., 2004). It is a small protein (61 aa), which has been found to concentrate at the endoplasmic reticulum and Golgi apparatus. In a murine coronavirus model, it was shown that expressing ORF6 increased virulence in mice (Zhao et al., 2009), and results indicate that ORF6 may serve an important role in the pathogenesis during SCoV infection (Liu et al., 2014). Also, it showed to inhibit the expression of certain STAT1-genes critical for the host immune response and could contribute to the immune evasion. ORF6 is expressed very well in WG-CFPS; the protein was fully soluble with detergents and partially soluble without them and was easily purified in the presence of detergent, but less efficiently in the absence thereof. Solution NMR spectra in the presence of detergent display narrow but few resonances, which correspond, in addition to the C-terminal STREP-tag, to the very C-terminal ORF6 protein residues.

ORF7a

SCoV2 protein 7a (121 aa) shows over 85% homology with the SCoV protein 7a. While the SCoV2 7a protein is produced and retained intracellularly, SCoV protein 7a has also been shown to be a structural protein incorporated into mature virions (Liu et al., 2014). 7a is one of the accessory proteins, of which a (partial) structure has been determined at high



resolution for SCoV2 (PDB 6W37). However, the very N-terminal signal peptide and the C-terminal membrane anchor, both highly hydrophobic, have not been determined experimentally yet.

Expression of the ORF7a ectodomain (ED) with a GB1 tag (Bogomolovas et al., 2009) was expected to produce reasonable yields. The IPFS purification resulted in a highly stable protein, as evidenced by the NMR data obtained (**Figure 7**).

ORF7b

Protein ORF7b is associated with viral particles in a SARS context (Liu et al., 2014). Protein 7b is one of the shortest ORFs with 43 residues. It shows a long hydrophobic stretch, which might correspond to a TM segment. It shows over 93% sequence homology with a bat coronavirus 7b protein (Liu et al., 2014). There, the cysteine residue in the C-terminal part is not conserved, which might facilitate structural studies. ORF7b has been synthesized successfully both from bacteria and by WG-CFPS in the presence of detergent and could be purified using a STREP-tag (**Table 2**). Due to the necessity of solubilizing agent

and its obvious tendency to oligomerize, structure determination, fragment screening, and interaction studies are challenging. However, we were able to record the first promising HSQC, as shown in **Figure 7**.

ORF8

ORF 8 is believed to be responsible for the evolution of *Betacoronaviruses* and their species jumps (Wu et al., 2016) and to have a role in repressing the host response (Tan et al., 2020). ORF 8 (121 aa) from SCoV2 does not apparently exist in SCoV on the protein level, despite the existence of a putative ORF. The sequences of the two homologs only show limited identity, with the exception of a small 7 aa segment, where, in SCoV, the glutamate is replaced with an aspartate. It, however, aligns very well with several coronaviruses endemic to animals, including Paguma and Bat (Chan et al., 2020). The protein comprises a hydrophobic peptide at its very N-terminus, likely corresponding to a signal peptide; the remaining part does not show any specific sequence features. Its

structure has been determined (PDB 7JTL) and shows a similar fold to ORF7a (Flower et al., 2020). In this study, ORF8 has been used both with (fl) and without signal peptide (Δ ORF8). We first tested the production of ORF8 in *E. coli*, but yields were low because of insolubility. Both ORF8 versions have then been synthesized in the cell-free system and were soluble in the presence of detergent. Solution NMR spectra, however, indicate that the protein is forming either oligomers or aggregates.

ORF9a (Nucleocapsid Protein, N)

The nucleocapsid protein (N) is important for viral genome packaging (Luo et al., 2006). The multifunctional RNA-binding protein plays a crucial role in the viral life cycle (Chang et al., 2014) and its domain architecture is highly conserved among coronaviruses. It comprises the N-terminal intrinsically disordered region (IDR1), the N-terminal RNA-binding globular domain (NTD), a central serine/arginine-(SR-) rich intrinsically disordered linker region (IDR2), the C-terminal dimerization domain (CTD), and a C-terminal intrinsically disordered region (IDR3) (Kang et al., 2020).

N represents a highly promising drug target. We thus focused our efforts not exclusively on the NTD and CTD alone, but, in addition, also provide protocols for IDR-containing constructs within the N-terminal part.

N-Terminal Domain

The NTD is the RNA-binding domain of the nucleocapsid (Kang et al., 2020). It is embedded within IDRs, functions of which have not yet been deciphered. Recent experimental and bioinformatic data indicate involvement in liquid-liquid phase separation (Chen et al., 2020a).

For the NTD, several constructs were designed, also considering the flanking IDRs (Table 1). In analogy to the available NMR [PDB 6YI3, (Dinesh et al., 2020)] and crystal [PDB 6M3M, (Kang et al., 2020)] structures of the SCoV2 NTD, boundaries for the NTD and the NTD-SR domains were designed to span residues 44–180 and 44–212, respectively. In addition, an extended IDR1-NTD-IDR2 (residues 1–248) construct was designed, including the N-terminal disordered region (IDR1), the NTD domain, and the central disordered linker (IDR2) that comprises the SR region. His-tagged NTD and NTD-SR were purified using IPRS and yielded approx. 3 mg/L in ^{15}N -labeled minimal medium. High protein quality and stability are supported by the available HSQC spectra (Figure 7).

The untagged IDR1-NTD-IDR2 was purified by IEC and yielded high amounts of ^{13}C , ^{15}N -labeled samples of 12 mg/L for further NMR investigations. The quality of our purification is confirmed by the available HSQC (Figure 7), and a near-complete backbone assignment of the two IDRs was achieved (Guseva et al., 2021; Schiavina et al., 2021). Notably, despite the structurally and dynamically heterogeneous nature of the N protein, the mentioned N constructs revealed a very good long-term stability, as shown in Table 2.

C-Terminal Domain

Multiple studies on the SCoV2 CTD, including recent crystal structures (Ye et al., 2020; Zhou et al., 2020), confirm the domain as dimeric. Its ability to self-associate seems to be necessary for viral replication and transcription (Luo et al., 2006). In addition, the CTD was shown to, presumably nonspecifically, bind ssRNA (Zhou et al., 2020).

Domain boundaries for the CTD were defined to comprise amino acids 247–364 (Table 1), in analogy to the NMR structure of the CTD from SCoV (PDB 2JW8, [Takeda et al., 2008]). Gene expression of His- or His-GST-tagged CTD yielded high amounts of soluble protein. Purification was achieved via IPRS. The CTD eluted as a dimer judged by its retention volume on the size-exclusion column and yielded good amounts (Table 2). The excellent protein quality and stability are supported by the available HSQC spectrum (Figure 7) and a near-complete backbone assignment (Korn et al., 2020b).

ORF9b

Protein 9b (97 aa) shows 73% sequence homology to the SCoV and also to bat virus (bat-SL-CoVZXC21) 9b protein (Chan et al., 2020). The structure of SCoV2 ORF9b has been determined at high resolution (PDB 6Z4U). Still, a significant portion of the structure was not found to be well ordered. The protein shows a β -sheet-rich structure and a hydrophobic tunnel, in which bound lipid was identified. How this might relate to membrane binding is not fully understood at this point. The differences in sequence between SCoV and SCoV2 are mainly located in the very N-terminus, which was not resolved in the structure (PDB 6Z4U). Another spot of deviating sequence not resolved in the structure is a solvent-exposed loop, which presents a potential interacting segment. ORF9b has been synthesized as a dimer (Figure 6E) using WG-CFPS in its soluble form. Spectra show a well-folded protein, and assignments are underway (Figure 6F).

ORF14 (ORF9c)

ORF14 (73 aa) remains, at this point in time, hypothetical. It shows 89% homology with a bat virus protein (bat-SL-CoVZXC21). It shows a highly hydrophobic part in its C-terminal region, comprising two negatively charged residues and a charged/polar N-terminus. The C-terminus is likely mediating membrane interaction. While ORF14 has been synthesized in the wheat-germ cell-free system in the presence of detergent and solution NMR spectra have been recorded, they hint at an aggregated protein (Figure 6E). Membrane-reconstitution of ORF14 revealed an unstable protein, which had been degraded during detergent removal.

ORF10

The ORF10 protein is comprised of 38 aa and is a hypothetical protein with unknown function (Yoshimoto, 2020). SCoV2 ORF10 displays 52.4% homology to SCoV ORF9b. The protein sequence is rich in hydrophobic residues, rendering expression and purification challenging. Expression of ORF10 as His-Trx-tagged or His-SUMO tagged fusion protein was possible; however, the ORF10 protein is poorly soluble and shows

partial unfolding, even as an uncleaved fusion protein. Analytical SEC hints at oligomerization under the current conditions.

DISCUSSION

The ongoing SCoV2 pandemic and its manifestation as the COVID-19 disease call for an urgent provision of therapeutics that will specifically target viral proteins and their interactions with each other and RNAs, which are crucial for viral propagation. Two “classical” viral targets have been addressed in comprehensive approaches soon after the outbreak in December 2019: the viral protease nsp5 and the RNA-dependent RNA polymerase (RdRp) nsp12. While the latter turned out to be a suitable target using the repurposed compound Remdesivir (Hillen et al., 2020), nsp5 is undergoing a broad structure-based screen against a battery of inhibitors in multiple places (Jin et al., 2020; Zhang et al., 2020), but with, as of yet, the limited outcome for effective medication. Hence, a comprehensive, reliable treatment of COVID-19 at any stage after the infection has remained unsuccessful.

Further viral protein targets will have to be taken into account in order to provide inhibitors with increased specificity and efficacy and preparative starting points for following potential generations of (SARS-)CoVs. Availability of those proteins in a recombinant, pure, homogenous, and stable form in milligrams is, therefore, a prerequisite for follow-up applications like vaccination, high-throughput screening campaigns, structure determination, and mapping of viral protein interaction networks. We here present, for the first time, a near-complete compendium of SCoV2 protein purification protocols that enable the production of large amounts of pure proteins.

The COVID19-NMR consortium was launched with the motivation of providing NMR assignments of all SCoV2 proteins and RNA elements, and enormous progress has been made since the outbreak of COVID-19 for both components [see Table 2 and (Wacker et al., 2020)]. Consequently, we have put our focus on producing proteins in stable isotope-labeled forms for NMR-based applications, e.g., the site-resolved mapping of interactions with compounds (Li and Kang, 2020). Relevant to a broad scientific community, we here report our protocols to suite perfectly any downstream biochemical or biomedical application.

Overall Success and Protein Coverage

As summarized in Table 2, we have successfully purified 80% of the SCoV2 proteins either in full or providing relevant fragments of the parent protein. Those include most of the nsp5, where all of the known/predicted soluble domains have been addressed (Figure 1). For a very large part, we were able to obtain protein samples of high purity, homogeneity, and fold for NMR-based applications. We would like to point out a number of CoV proteins that, evidenced by their HSQCs, for the first time, provide access to structural information, e.g., the PL^{Pro} nsp3d and nsp3Y. Particularly for the nsp3 multidomain protein, we here present soluble samples of

almost the complete cytosolic region with more than 120 kDa in the form of excellent 2D NMR spectra (Figure 3), a major part of which fully backbone-assigned. We thus enable the exploitation of the largest and most enigmatic multifunctional SCoV2 protein through individual domains in solution, allowing us to study their concerted behavior with single residue resolution. Similarly, for nsp2, we provide a promising starting point for studying the so far neglected, often uncharacterized, and apparently unstructured proteins.

Driven by the fast-spreading COVID-19, we initially left out proteins that require advanced purification procedures (e.g., nsp12 and S) or where *a priori* information was limited (nsp4 and nsp6). This procedure seems justified with the time-saving approach of our effort in favor of the less attended proteins. However, we are in the process of collecting protocols for the missing proteins.

Different Complexities and Challenges

The compilation of protein production protocols, initially guided by information from CoV homologs (Table 1), has confronted us with very different levels of complexity. With some prior expectation toward this, we have shared forces to quickly “work off” the highly conserved soluble and small proteins and soon put focus into the processing of the challenging ones. The difficulties in studying this second class of proteins are due to their limited sequence conservation, no prior information, large molecular weights, insolubility, and so forth.

The nsp3e NAB represents one example where the available NMR structure of the SCoV homolog provided a *bona fide* template for selecting initial domain boundaries (Figure 4). The transfer of information derived from SCoV was straightforward; the transferability included the available protocol for the production of comparable protein amounts and quality, given the high sequence identity. In such cases, we found ourselves merely to adapt protocols and optimize yields based on slightly different expression vectors and *E. coli* strains.

However, in some cases, such transfer was unexpectedly not successful, e.g., for the short nsp1 GD. Despite intuitive domain boundaries with complete local sequence identity seen from the SCoV nsp1 NMR structure, it took considerable efforts to purify an analogous nsp1 construct, which is likely related to the impaired stability and solubility caused by a number of impacting amino acid exchanges within the domain’s flexible loops. In line with that, currently available structures of SCoV2 nsp1 have been obtained by crystallography or cryo-EM and include different buffers. As such, our initial design was insufficient in terms of taking into account the parameters mentioned above. However, one needs to consider those particular differences between the nsp1 homologs as one of the most promising target sites for potential drugs as they appear to be hotspots in the CoV evolution and will have essential effects for the molecular networks, both in the virus and with the host (Zust et al., 2007; Narayanan et al., 2015; Shen et al., 2019; Thoms et al., 2020).

A special focus was put on the production of the SCoV2 main protease nsp5, for which NMR-based screenings are ongoing. The

main protease is critical in terms of inhibitor design as it appears under constant selection, and novel mutants remarkably influence the structure and biochemistry of the protein (Cross et al., 2020). In the present study, the expression of the different constructs allowed us to characterize the protein in both its monomeric and dimeric forms. Comparison of NMR spectra reveals that the constructs with additional amino acids (GS and GHM mutant) display marked structural differences to the wild-type protein while being structurally similar among themselves (Figure 5H). The addition of two residues (GS) interferes with the dimerization interface, despite being similar to its native N-terminal amino acids (SGFR). We also introduced an active site mutation that replaces cysteine 145 with alanine (Hsu et al., 2005). Intriguingly, this active site mutation C145A, known to stabilize the dimerization of the main protease (Chang et al., 2007), supports dimer formation of the GS added construct (GS-nsp5 C145A) shown by its 2D NMR spectrum overlaying with the one of wild-type nsp5 (Supplementary Table S14). The NMR results are in line with SEC-MALS analyses (Figure 5F). Indeed, the additional amino acids at the N-terminus shift the dimerization equilibrium toward the monomer, whereas the mutation shifts it toward the dimer despite the N-terminal aa additions. This example underlines the need for a thorough and precise construct design and the detailed biochemical and NMR-based characterization of the final sample state. The presence of monomers vs. dimers will play an essential role in the inhibitor search against SCoV2 proteins, as exemplified by the particularly attractive nsp5 main protease target.

Exploiting Nonbacterial Expression

As a particular effort within this consortium, we included the so far neglected accessory proteins using a structural genomics procedure supported by wheat-germ cell-free protein synthesis. This approach allowed us previously to express a variety of difficult viral proteins in our hands (Fogeron et al., 2015a; Fogeron et al., 2015b; Fogeron et al., 2016; Fogeron et al., 2017; Wang et al., 2019; Jirasko et al., 2020a). Within the workflow, we especially highlight the straightforward solubilization of the membrane proteins through the addition of detergent to the cell-free reaction, which allowed the production of soluble protein in milligram amounts compatible with NMR studies. While home-made extracts were used here, very similar extracts are available commercially (Cell-Free Sciences, Japan) and can thus be implemented by any lab without prior experience. Also, a major benefit of the WG-CFPS system for NMR studies lies in the high efficiency and selectivity of isotopic labeling. In contrast to cell-based expression systems, only the protein of interest is produced (Morita et al., 2003), which allows bypassing extensive purification steps. In fact, one-step affinity purification is in most cases sufficient, as shown for the different ORFs in this study. Samples could be produced for virtually all proteins, with the exception of the ORF3b construct used. With new recent insight into the stop codons present in this ORF, constructs will be adapted, which shall overcome the problems of ORF3b production (Konno et al., 2020).

For two ORFs, 7b and 8, we exploited a paralleled production strategy, i.e., both in bacteria and via cell-free synthesis. For those challenging proteins, we were, in principle, able to obtain pure samples from either expression system. However, for ORF7b, we found a strict dependency on detergents for follow-up work from both approaches. ORF8 showed significantly better solubility when produced in WG extracts compared to bacteria. This shows the necessity of parallel routes to take, in particular, for the understudied, biochemically nontrivial ORFs that might represent yet unexplored but highly specific targets to consider in the treatment of COVID-19.

Downstream structural analysis of ORFs produced with CFPS remains challenging but promising progress is being made in the light of SCoV2. Some solution NMR spectra show the expected number of signals with good resolution (e.g., ORF9b). As expected, however, most proteins cannot be straightforwardly analyzed by solution NMR in their current form, as they exhibit too large objects after insertion into micelles and/or by inherent oligomerization. Cell-free synthesized proteins can be inserted into membranes through reconstitution (Fogeron et al., 2015a; Fogeron et al., 2015b; Fogeron et al., 2016; Jirasko et al., 2020a; Jirasko et al., 2020b). Reconstitution will thus be the next step for many accessory proteins, but also for M and E, which were well produced by WG-CFPS. We will also exploit the straightforward deuteration in WG-CFPS (David et al., 2018; Wang et al., 2019; Jirasko et al., 2020a) that circumvents proton back-exchange, rendering denaturation and refolding steps obsolete (Tonelli et al., 2011). Nevertheless, the herein presented protocols for the production of non-nsp5 by WG-CFPS instantly enable their employment in binding studies and screening campaigns and thus provide a significant contribution to soon-to-come studies on SCoV2 proteins beyond the classical and convenient drug targets.

Altogether and judged by the ultimate need of exploiting recombinant SCoV2 proteins in vaccination and highly paralleled screening campaigns, we optimized sample amount, homogeneity, and long-term stability of samples. Our freely accessible protocols and accompanying NMR spectra now offer a great resource to be exploited for the unambiguous and reproducible production of SCoV2 proteins for the intended applications.

DATA AVAILABILITY STATEMENT

Assignments of backbone chemical shifts have been deposited at BMRB for proteins, as shown in Table 2, indicated by their respective BMRB IDs. All expression constructs are available as plasmids from <https://covid19-nmr.de/>.

AUTHOR CONTRIBUTIONS

NA, SK, NQ, MD, MN, ABö, HS, MH, and AS designed the study, compiled the protocols and NMR data, and wrote the manuscript. All authors contributed coordinative or practical work to the study. All authors contributed to the creation and collection of protein protocols and NMR spectra.

FUNDING

This work was supported by Goethe University (Corona funds), the DFG-funded CRC: “Molecular Principles of RNA-Based Regulation,” DFG infrastructure funds (project numbers: 277478796, 277479031, 392682309, 452632086, 70653611), the state of Hesse (BMRZ), the Fondazione CR Firenze (CERM), and the IWB-EFRE-program 20007375. This project has received funding from the European Union’s Horizon 2020 research and innovation program under Grant Agreement No. 871037. AS is supported by DFG Grant SCHL 2062/2-1 and by the JQYA at Goethe through project number 2019/AS01. Work in the lab of KV was supported by a CoRE grant from the University of New Hampshire. The FLI is a member of the Leibniz Association (WGL) and financially supported by the Federal Government of Germany and the State of Thuringia. Work in the lab of RM was supported by NIH (2R01EY021514) and NSF (DMR-2002837). BN-B was supported by the NSF GRFP. MC was supported by NIH (R25 GM055246 MBRS IMSD), and MS-P was supported by the HHMI Gilliam Fellowship. Work in the labs of KJ and KT was supported by Latvian Council of Science Grant No. VPP-COVID 2020/1-0014. Work in the UPAT’s lab was supported by the INSPIRED (MIS 5002550) project, which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure,” funded by the Operational Program “Competitiveness, Entrepreneurship and Innovation” (NSRF 2014–2020) and cofinanced by Greece and the EU (European Regional Development Fund) and the FP7 REGPOT CT-2011-285950–“SEE-DRUG” project (purchase of UPAT’s 700 MHz NMR equipment). Work in the CM-G lab was supported by the Helmholtz society. Work in the lab of ABö was supported by the CNRS, the French National Research Agency (ANR, NMR-SCoV2- ORF8), the Fondation de la Recherche Médicale (FRM, NMR-SCoV2-ORF8), and the IR-RMN-THC Fr3050 CNRS. Work in the lab of BM was supported by the Swiss National Science Foundation (Grant number 200020_188711), the Günthard Stiftung für Physikalische Chemie, and the ETH Zurich. Work in the labs of ABö and BM was supported by a common grant from SNF (grant 31CA30_196256). This work was supported by the ETH Zurich, the grant ETH 40 18 1, and the grant Krebsliga KFS 4903 08 2019. Work in the lab of the IBS Grenoble was supported by the Agence Nationale de Recherche (France)

REFERENCES

- Almeida, M. S., Johnson, M. A., Herrmann, T., Geralt, M., and Wüthrich, K. (2007). Novel beta-barrel fold in the nuclear magnetic resonance structure of the replicase nonstructural protein 1 from the severe acute respiratory syndrome coronavirus. *J. Virol.* 81 (7), 3151–3161. doi:10.1128/JVI.01939-06
- Anand, K., Ziebuhr, J., Wadhvani, P., Mesters, J. R., and Hilgenfeld, R. (2003). Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* 300 (5626), 1763–1767. doi:10.1126/science.1085658
- Bogomolovas, J., Simon, B., Sattler, M., and Stier, G. (2009). Screening of fusion partners for high yield expression and purification of bioactive viscotoxins. *Protein Expr. Purif.* 64 (1), 16–23. doi:10.1016/j.pep.2008.10.003
- Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., et al. (2020). Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* 583 (7816), 469–472. doi:10.1038/s41586-020-2332-7
- Bouvet, M., Debarnot, C., Imbert, I., Selisko, B., Snijder, E. J., Canard, B., et al. (2010). *In vitro* reconstitution of SARS-coronavirus mRNA cap methylation. *PLoS Pathog.* 6 (4), e1000863. doi:10.1371/journal.ppat.1000863
- Cantini, F., Banci, L., Altincekic, N., Bains, J. K., Dhamotharan, K., Fuks, C., et al. (2020). (1)H, (13)C, and (15)N backbone chemical shift assignments of the apo and the ADP-ribose bound forms of the macromolecule of SARS-CoV-2 non-structural protein 3b. *Biomol. NMR Assign.* 14 (2), 339–346. doi:10.1007/s12104-020-09973-4
- Chan, J. F., Kok, K. H., Zhu, Z., Chu, H., To, K. K., Yuan, S., et al. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* 9 (1), 221–236. doi:10.1080/22221751.2020.1719902

RA-COVID SARS2NUCLEOPROTEIN and European Research Council Advanced Grant DynamicAssemblies. Work in the CA lab was supported by Patto per il Sud della Regione Siciliana–CheMIS grant (CUP G77B17000110001). Part of this work used the platforms of the Grenoble Instruct-ERIC center (ISBG; UMS 3518 CNRS-CEA-UGA-EMBL) within the Grenoble Partnership for Structural Biology (PSB), supported by FRISBI (ANR-10-INBS-05-02) and GRAL, financed within the University Grenoble Alpes graduate school (Ecoles Universitaires de Recherche) CBH-EUR-GS (ANR-17-EURE-0003). Work at the UW-Madison was supported by grant numbers NSF MCB2031269 and NIH/NIAID AI123498. MM is a Ramón y Cajal Fellow of the Spanish AEI-Ministry of Science and Innovation (RYC2019-026574-I), and a “La Caixa” Foundation (ID 100010434) Junior Leader Fellow (LCR/BQ/PR19/11700003). Funded by project COV20/00764 from the Carlos III Institute of Health and the Spanish Ministry of Science and Innovation to MM and DVL. VDJ was supported by the Boehringer Ingelheim Fonds. Part of this work used the resources of the Italian Center of Instruct-ERIC at the CERM/ CIRMMMP infrastructure, supported by the Italian Ministry for University and Research (FOE funding). CF was supported by the Stiftung Polytechnische Gesellschaft. Work in the lab of JH was supported by NSF (RAPID 2030601) and NIH (R01GM123249).

ACKNOWLEDGMENTS

The authors thank Leonardo Gonnelli and Katharina Targaczewski for the valuable technical assistance. IBS acknowledges integration into the Interdisciplinary Research Institute of Grenoble (IRIG CEA). They acknowledge the Advanced Technologies Network Center of the University of Palermo to support infrastructures.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2021.653148/full#supplementary-material>

- Chang, C. K., Hou, M. H., Chang, C. F., Hsiao, C. D., and Huang, T. H. (2014). The SARS coronavirus nucleocapsid protein—forms and functions. *Antivir. Res.* 103, 39–50. doi:10.1016/j.antiviral.2013.12.009
- Chang, H. P., Chou, C. Y., and Chang, G. G. (2007). Reversible unfolding of the severe acute respiratory syndrome coronavirus main protease in guanidinium chloride. *Biophys. J.* 92 (4), 1374–1383. doi:10.1529/biophysj.106.091736
- Chen, H., Cui, Y., Han, X., Hu, W., Sun, M., Zhang, Y., et al. (2020a). Liquid-liquid phase separation by SARS-CoV-2 nucleocapsid protein and RNA. *Cell Res.* 30, 1143. doi:10.1038/s41422-020-00408-2
- Chen, J., Malone, B., Llewellyn, E., Grasso, M., Shelton, P. M. M., Olinares, P. D. B., et al. (2020b). Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex. *Cell* 182 (6), 1560–1573. doi:10.1016/j.cell.2020.07.033
- Chen, P., Jiang, M., Hu, T., Liu, Q., Chen, X. S., and Guo, D. (2007). Biochemical characterization of exoribonuclease encoded by SARS coronavirus. *J. Biochem. Mol. Biol.* 40 (5), 649–655. doi:10.5483/bmbrep.2007.40.5.649
- Chen, Y., Savinov, S. N., Mielech, A. M., Cao, T., Baker, S. C., and Mesecar, A. D. (2015). X-ray structural and functional studies of the three tandemly linked domains of non-structural protein 3 (nsp3) from murine hepatitis virus reveal conserved functions. *J. Biol. Chem.* 290 (42), 25293–25306. doi:10.1074/jbc.M115.662130
- Cornillez-Ty, C. T., Liao, L., Yates, J. R., 3rd, Kuhn, P., and Buchmeier, M. J. (2009). Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. *J. Virol.* 83 (19), 10314–10318. doi:10.1128/JVI.00842-09
- Cross, T. J., Takahashi, G. R., Diessner, E. M., Crosby, M. G., Farahmand, V., Zhuang, S., et al. (2020). Sequence characterization and molecular modeling of clinically relevant variants of the SARS-CoV-2 main protease. *Biochemistry* 59 (39), 3741–3756. doi:10.1021/acs.biochem.0c00462
- David, G., Fogeron, M. L., Schledorn, M., Montserret, R., Haselmann, U., Penzel, S., et al. (2018). Structural studies of self-assembled subviral particles: combining cell-free expression with 110 kHz MAS NMR spectroscopy. *Angew. Chem. Int. Ed. Engl.* 57 (17), 4787–4791. doi:10.1002/anie.201712091
- Davies, J. P., Almsay, K. M., McDonald, E. F., and Plate, L. (2020). Comparative multiplexed interactomics of SARS-CoV-2 and homologous coronavirus non-structural proteins identifies unique and shared host-cell dependencies. *bioRxiv* [Epub ahead of print]. doi:10.1101/2020.07.13.201517
- Deng, X., Hackbart, M., Mettelman, R. C., O'Brien, A., Mielech, A. M., Yi, G., et al. (2017). Coronavirus nonstructural protein 15 mediates evasion of dsRNA sensors and limits apoptosis in macrophages. *Proc. Natl. Acad. Sci. U.S.A.* 114 (21), E4251–E4260. doi:10.1073/pnas.1618310114
- Dinesh, D. C., Chalupska, D., Silhan, J., Koutna, E., Nencka, R., Veverka, V., et al. (2020). Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLoS Pathog.* 16 (12), e1009100. doi:10.1371/journal.ppat.1009100
- Dudas, F. D., Puglisi, R., Korn, S. M., Alfano, C., Kelly, G., Monaca, E., et al. (2021). Backbone chemical shift spectral assignments of coronavirus-2 non-structural protein nsp9. *Biomol. NMR Assign.* 2021, 1–10. doi:10.1007/s12104-020-09992-1
- Egloff, M. P., Ferron, F., Campanacci, V., Longhi, S., Rancurel, C., Dutartre, H., et al. (2004). The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world. *Proc. Natl. Acad. Sci. U.S.A.* 101 (11), 3792–3796. doi:10.1073/pnas.0307877101
- Esposito, D., Mehalko, J., Drew, M., Snead, K., Wall, V., Taylor, T., et al. (2020). Optimizing high-yield production of SARS-CoV-2 soluble spike trimers for serology assays. *Protein Expr. Purif.* 174, 105686. doi:10.1016/j.pep.2020.105686
- Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., et al. (2020). The coding capacity of SARS-CoV-2. *Nature* 589, 125. doi:10.1038/s41586-020-2739-1
- Flower, T. G., Buffalo, C. Z., Hooy, R. M., Allaire, M., Ren, X., and Hurley, J. H. (2020). Structure of SARS-CoV-2 ORF8, a rapidly evolving coronavirus protein implicated in immune evasion. *bioRxiv* [Epub ahead of print]. doi:10.1101/2020.08.27.270637
- Fogeron, M. L., Badillo, A., Jirasko, V., Gouttenoire, J., Paul, D., Lancien, L., et al. (2015a). Wheat germ cell-free expression: two detergents with a low critical micelle concentration allow for production of soluble HCV membrane proteins. *Protein Expr. Purif.* 105, 39–46. doi:10.1016/j.pep.2014.10.003
- Fogeron, M. L., Badillo, A., Penin, F., and Böckmann, A. (2017). Wheat germ cell-free overexpression for the production of membrane proteins. *Methods Mol. Biol.* 1635, 91–108. doi:10.1007/978-1-4939-7151-0_5
- Fogeron, M. L., Jirasko, V., Penzel, S., Paul, D., Montserret, R., Danis, C., et al. (2016). Cell-free expression, purification, and membrane reconstitution for NMR studies of the nonstructural protein 4B from hepatitis C virus. *J. Biomol. NMR* 65 (2), 87–98. doi:10.1007/s10858-016-0040-2
- Fogeron, M. L., Paul, D., Jirasko, V., Montserret, R., Lacabanne, D., Molle, J., et al. (2015b). Functional expression, purification, characterization, and membrane reconstitution of non-structural protein 2 from hepatitis C virus. *Protein Expr. Purif.* 116, 1–6. doi:10.1016/j.pep.2015.08.027
- Frick, D. N., Viridi, R. S., Vuksanovic, N., Dahal, N., and Silvaggi, N. R. (2020). Molecular basis for ADP-ribose binding to the Mac1 domain of SARS-CoV-2 nsp3. *Biochemistry* 59 (28), 2608–2615. doi:10.1021/acs.biochem.0c0309
- Gallo, A., Tsika, A. C., Fourkiotis, N. K., Cantini, F., Banci, L., Sreeramulu, S., et al. (2020). 1H, 13C and 15N chemical shift assignments of the SUD domains of SARS-CoV-2 non-structural protein 3c: “the N-terminal domain-SUD-N”. *Biomol. NMR Assign.* 2020, 1–5. doi:10.1007/s12104-020-09987-y
- Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., et al. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* 368 (6492), 779–782. doi:10.1126/science.abb7498
- Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583 (7816), 459–468. doi:10.1038/s41586-020-2286-9
- Graham, R. L., Sims, A. C., Brockway, S. M., Baric, R. S., and Denison, M. R. (2005). The nsp2 replicase proteins of murine hepatitis virus and severe acute respiratory syndrome coronavirus are dispensable for viral replication. *J. Virol.* 79 (21), 13399–13411. doi:10.1128/JVI.79.21.13399-13411.2005
- Guseva, S., Perez, L. M., Camacho-Zarco, A., Bessa, L. M., Salvi, N., Malki, A., et al. (2021). (1)H, (13)C and (15)N Backbone chemical shift assignments of the n-terminal and central intrinsically disordered domains of SARS-CoV-2 nucleoprotein. *Biomol. NMR Assign.* doi:10.1007/s12104-021-10014-x
- Hagemeyer, M. C., Monastyrska, I., Griffith, J., van der Sluijs, P., Voortman, J., van Bergen en Henegouwen, P. M., et al. (2014). Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. *Virology* 458–459, 125–135. doi:10.1016/j.virol.2014.04.027
- Hillen, H. S., Kocic, G., Farnung, L., Dienemann, C., Tegunov, D., and Cramer, P. (2020). Structure of replicating SARS-CoV-2 polymerase. *Nature* 584 (7819), 154–156. doi:10.1038/s41586-020-2368-8
- Hsu, M. F., Kuo, C. J., Chang, K. T., Chang, H. C., Chou, C. C., Ko, T. P., et al. (2005). Mechanism of the maturation process of SARS-CoV 3CL protease. *J. Biol. Chem.* 280 (35), 31257–31266. doi:10.1074/jbc.M502577200
- Huang, Y., Yang, Z. Y., Kong, W. P., and Nabel, G. J. (2004). Generation of synthetic severe acute respiratory syndrome coronavirus pseudoparticles: implications for assembly and vaccine production. *J. Virol.* 78 (22), 12557–12565. doi:10.1128/JVI.78.22.12557-12565.2004
- Hurst, K. R., Koetzner, C. A., and Masters, P. S. (2013). Characterization of a critical interaction between the coronavirus nucleocapsid protein and nonstructural protein 3 of the viral replicase-transcriptase complex. *J. Virol.* 87 (16), 9159–9172. doi:10.1128/JVI.01275-13
- Ishida, T., and Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* 35, W460–W464. doi:10.1093/nar/gkm363
- Jia, Z., Yan, L., Ren, Z., Wu, L., Wang, J., Guo, J., et al. (2019). Delicate structural coordination of the severe acute respiratory syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res.* 47 (12), 6538–6550. doi:10.1093/nar/gkz409
- Jiang, H. W., Li, Y., Zhang, H. N., Wang, W., Yang, X., Qi, H., et al. (2020). SARS-CoV-2 proteome microarray for global profiling of COVID-19 specific IgG and IgM responses. *Nat. Commun.* 11 (1), 3581. doi:10.1038/s41467-020-17488-8
- Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., et al. (2020). Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582 (7811), 289–293. doi:10.1038/s41586-020-2223-y
- Jirasko, V., Lakomek, N. A., Penzel, S., Fogeron, M. L., Bartenschlager, R., Meier, B. H., et al. (2020a). Proton-detected solid-state NMR of the cell-free synthesized

- α -helical transmembrane protein NS4B from hepatitis C virus. *ChemBiochem*. 21 (10), 1453–1460. doi:10.1002/cbic.201900765
- Jirasko, V., Lends, A., Lakomek, N. A., Fogeron, M. L., Weber, M. E., Malär, A. A., et al. (2020b). Dimer organization of membrane-associated NSSA of hepatitis C virus as determined by highly sensitive 1 H-detected solid-state NMR. *Angew. Chem. Int. Ed.* 60 (10), 5339–5347. doi:10.1002/anie.202013296
- Johnson, M. A., Chatterjee, A., Neuman, B. W., and Wüthrich, K. (2010). SARS coronavirus unique domain: three-domain molecular architecture in solution and RNA binding. *J. Mol. Biol.* 400 (4), 724–742. doi:10.1016/j.jmb.2010.05.027
- Joseph, J. S., Saikatendu, K. S., Subramanian, V., Neuman, B. W., Brooun, A., Griffith, M., et al. (2006). Crystal structure of nonstructural protein 10 from the severe acute respiratory syndrome coronavirus reveals a novel fold with two zinc-binding motifs. *J. Virol.* 80 (16), 7894–7901. doi:10.1128/JVI.00467-06
- Kamitani, W., Narayanan, K., Huang, C., Lokugamage, K., Ikegami, T., Ito, N., et al. (2006). Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. *Proc. Natl. Acad. Sci. U.S.A.* 103 (34), 12885–12890. doi:10.1073/pnas.0603144103
- Kang, S., Yang, M., Hong, Z., Zhang, L., Huang, Z., Chen, X., et al. (2020). Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm. Sin. B* 10 (7), 1228–1238. doi:10.1016/j.apsb.2020.04.009
- Kern, D. M., Sorum, B., Mali, S. S., Hoel, C. M., Sridharan, S., Remis, J. P., et al. (2020). Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs. *bioRxiv* 17, 156554. doi:10.1101/2020.06.17.156554
- Khan, M. T., Zeb, M. T., Ahsan, H., Ahmed, A., Ali, A., Akhtar, K., et al. (2020). SARS-CoV-2 nucleocapsid and Nsp3 binding: an in silico study. *Arch. Microbiol.* 203, 59. doi:10.1007/s00203-020-01998-6
- Kim, Y., Jedrzejczak, R., Maltseva, N. I., Wilamowski, M., Endres, M., Godzik, A., et al. (2020). Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Protein Sci.* 29 (7), 1596–1605. doi:10.1002/pro.3873
- Kirchdoerfer, R. N., and Ward, A. B. (2019). Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat. Commun.* 10 (1), 2342. doi:10.1038/s41467-019-10280-3
- Konkolova, E., Klima, M., Nencka, R., and Boura, E. (2020). Structural analysis of the putative SARS-CoV-2 primase complex. *J. Struct. Biol.* 211 (2), 107548. doi:10.1016/j.jsb.2020.107548
- Konno, Y., Kimura, I., Uriu, K., Fukushi, M., Irie, T., Koyanagi, Y., et al. (2020). SARS-CoV-2 ORF3b is a potent interferon antagonist whose activity is increased by a naturally occurring elongation variant. *Cell Rep.* 32 (12), 108185. doi:10.1016/j.celrep.2020.108185
- Korn, S. M., Dharmotharan, K., Fürtig, B., Hengesbach, M., Löhr, F., Qureshi, N. S., et al. (2020a). 1H, 13C, and 15N backbone chemical shift assignments of the nucleic acid-binding domain of SARS-CoV-2 non-structural protein 3e. *Biomol. NMR Assign.* 14 (2), 329–333. doi:10.1007/s12104-020-09971-6
- Korn, S. M., Lambert, R., Fürtig, B., Hengesbach, M., Löhr, F., Richter, C., et al. (2020b). 1H, 13C, and 15N backbone chemical shift assignments of the C-terminal dimerization domain of SARS-CoV-2 nucleocapsid protein. *Biomol. NMR Assign.* 2020, 1–7. doi:10.1007/s12104-020-09995-y
- Krafcikova, P., Silhan, J., Nencka, R., and Boura, E. (2020). Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. *Nat. Commun.* 11 (1), 3717. doi:10.1038/s41467-020-17495-9
- Kubatova, N., Qureshi, N. S., Altincekic, N., Abele, R., Bains, J. K., Ceylan, B., et al. (2020). 1H, 13C, and 15N backbone chemical shift assignments of coronavirus-2 non-structural protein Nsp10. *Biomol. NMR Assign.* 2020, 1–7. doi:10.1007/s12104-020-09984-1
- Kusov, Y., Tan, J., Alvarez, E., Enjuanes, L., and Hilgenfeld, R. (2015). A G-quadruplex-binding macromolecule within the “SARS-unique domain” is essential for the activity of the SARS-coronavirus replication-transcription complex. *Virology* 484, 313–322. doi:10.1016/j.virol.2015.06.016
- Leao, J. C., Gusmao, T. P. L., Zazar, A. M., Leao Filho, J. C., Barkokebas Santos de Faria, A., Morais Silva, I. H., et al. (2020). Coronaviridae-old friends, new enemy! *Oral Dis.* 2020, 13447. doi:10.1111/odi.13447
- Lei, J., Kusov, Y., and Hilgenfeld, R. (2018). Nsp3 of coronaviruses: structures and functions of a large multi-domain protein. *Antivir. Res.* 149, 58–74. doi:10.1016/j.antiviral.2017.11.001
- Li, Q., and Kang, C. (2020). A practical perspective on the roles of solution NMR spectroscopy in drug discovery. *Molecules* 25 (13), 2974. doi:10.3390/molecules25132974
- Littler, D. R., Gully, B. S., Colson, R. N., and Rossjohn, J. (2020). Crystal structure of the SARS-CoV-2 non-structural protein 9, Nsp9. *iScience* 23 (7), 101258. doi:10.1016/j.isci.2020.101258
- Liu, D. X., Fung, T. S., Chong, K. K., Shukla, A., and Hilgenfeld, R. (2014). Accessory proteins of SARS-CoV and other coronaviruses. *Antivir. Res.* 109, 97–109. doi:10.1016/j.antiviral.2014.06.013
- Luo, H., Chen, J., Chen, K., Shen, X., and Jiang, H. (2006). Carboxyl terminus of severe acute respiratory syndrome coronavirus nucleocapsid protein: self-association analysis and nucleic acid binding characterization. *Biochemistry* 45 (39), 11827–11835. doi:10.1021/bi0609319
- Ma, Y., Wu, L., Shaw, N., Gao, Y., Wang, J., Sun, Y., et al. (2015). Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proc. Natl. Acad. Sci. U.S.A.* 112 (30), 9436–9441. doi:10.1073/pnas.1508686112
- Mandala, V. S., McKay, M. J., Shcherbakov, A. A., Dregni, A. J., Kolocouris, A., and Hong, M. (2020). Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers. *Nat. Struct. Mol. Biol.* 27, 1202. doi:10.1038/s41594-020-00536-8
- Miknis, Z. J., Donaldson, E. F., Umland, T. C., Rimmer, R. A., Baric, R. S., and Schultz, L. W. (2009). Severe acute respiratory syndrome coronavirus nsp9 dimerization is essential for efficient viral growth. *J. Virol.* 83 (7), 3007–3018. doi:10.1128/JVI.01505-08
- Mompean, M., Trevino, M. A., and Laurents, D. V. (2020). Towards targeting the disordered SARS-CoV-2 nsp2 C-terminal region: partial structure and dampened mobility revealed by NMR spectroscopy. *bioRxiv* [Epub ahead of print]. doi:10.1101/2020.11.09.374173
- Morita, E. H., Sawasaki, T., Tanaka, R., Endo, Y., and Kohno, T. (2003). A wheat germ cell-free system is a novel way to screen protein folding and function. *Protein Sci.* 12 (6), 1216–1221. doi:10.1110/ps.0241203
- Narayanan, K., Huang, C., Lokugamage, K., Kamitani, W., Ikegami, T., Tseng, C. T., et al. (2008). Severe acute respiratory syndrome coronavirus nsp1 suppresses host gene expression, including that of type I interferon, in infected cells. *J. Virol.* 82 (9), 4471–4479. doi:10.1128/JVI.02472-07
- Narayanan, K., Ramirez, S. I., Lokugamage, K. G., and Makino, S. (2015). Coronavirus nonstructural protein 1: common and distinct functions in the regulation of host and viral gene expression. *Virus Res.* 202, 89–100. doi:10.1016/j.virusres.2014.11.019
- Nelson, C. W., Arden, Z., Goldberg, T. L., Meng, C., Kuo, C. H., Ludwig, C., et al. (2020). Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic. *Elife* 9, 59633. doi:10.7554/eLife.59633
- Netzer, W. J., and Hartl, F. U. (1997). Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* 388 (6640), 343–349. doi:10.1038/41024
- Neuman, B. W., Joseph, J. S., Saikatendu, K. S., Serrano, P., Chatterjee, A., Johnson, M. A., et al. (2008). Proteomics analysis unravels the functional repertoire of coronavirus nonstructural protein 3. *J. Virol.* 82 (11), 5279–5294. doi:10.1128/JVI.02631-07
- Neuman, B. W., Kiss, G., Kunding, A. H., Bhella, D., Baksh, M. F., Connelly, S., et al. (2011). A structural analysis of M protein in coronavirus assembly and morphology. *J. Struct. Biol.* 174 (1), 11–22. doi:10.1016/j.jsb.2010.11.021
- Neuman, B. W. (2016). Bioinformatics and functional analyses of coronavirus nonstructural proteins involved in the formation of replicative organelles. *Antivir. Res.* 135, 97–107. doi:10.1016/j.antiviral.2016.10.005
- Oostra, M., Hagemeyer, M. C., van Gent, M., Bekker, C. P., te Lintelo, E. G., Rottier, P. J., et al. (2008). Topology and membrane anchoring of the coronavirus replication complex: not all hydrophobic domains of nsp3 and nsp6 are membrane spanning. *J. Virol.* 82 (24), 12392–12405. doi:10.1128/JVI.01219-08
- Oostra, M., te Lintelo, E. G., Deijs, M., Verheije, M. H., Rottier, P. J., and de Haan, C. A. (2007). Localization and membrane topology of coronavirus nonstructural protein 4: involvement of the early secretory pathway in replication. *J. Virol.* 81 (22), 12323–12336. doi:10.1128/JVI.01506-07
- Pavesi, A. (2020). New insights into the evolutionary features of viral overlapping genes by discriminant analysis. *Virology* 546, 51–66. doi:10.1016/j.virol.2020.03.007

- Robertson, M. P., Igel, H., Baertsch, R., Haussler, D., Ares, M., Jr., and Scott, W. G. (2005). The structure of a rigorously conserved RNA element within the SARS virus genome. *PLoS Biol.* 3 (1), e5. doi:10.1371/journal.pbio.0030005
- Robson, F., Khan, K. S., Le, T. K., Paris, C., Demirbag, S., Barfuss, P., et al. (2020). Coronavirus RNA proofreading: molecular basis and therapeutic targeting. *Mol. Cell* 79 (5), 710–727. doi:10.1016/j.molcel.2020.07.027
- Rosas-Lemus, M., Minasov, G., Shuvalova, L., Inniss, N. L., Kiryukhina, O., Wiersum, G., et al. (2020). The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. *bioRxiv* [Epub ahead of print]. doi:10.1101/2020.04.17.047498
- Salvi, N., Bessa, L. M., Guseva, S., Camacho-Zarco, A., Maurin, D., Perez, L. M., et al. (2021). ¹H, ¹³C and ¹⁵N backbone chemical shift assignments of SARS-CoV-2 nsp3a. *Biomol. NMR Assign.* 2021, 1–4. doi:10.1007/s12104-020-10001-8
- Schoeman, D., and Fielding, B. C. (2020). Is there a link between the pathogenic human coronavirus envelope protein and immunopathology? A review of the literature. *Front. Microbiol.* 11, 2086. doi:10.3389/fmicb.2020.02086
- Schubert, K., Karousis, E. D., Jomaa, A., Scaiola, A., Echeverria, B., Gurzeler, L. A., et al. (2020). SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nat. Struct. Mol. Biol.* 27 (10), 959–966. doi:10.1038/s41594-020-0511-8
- Serrano, P., Johnson, M. A., Almeida, M. S., Horst, R., Herrmann, T., Joseph, J. S., et al. (2007). Nuclear magnetic resonance structure of the N-terminal domain of nonstructural protein 3 from the severe acute respiratory syndrome coronavirus. *J. Virol.* 81 (21), 12049–12060. doi:10.1128/JVI.00969-07
- Serrano, P., Johnson, M. A., Chatterjee, A., Neuman, B. W., Joseph, J. S., Buchmeier, M. J., et al. (2009). Nuclear magnetic resonance structure of the nucleic acid-binding domain of severe acute respiratory syndrome coronavirus nonstructural protein 3. *J. Virol.* 83 (24), 12998–13008. doi:10.1128/JVI.01253-09
- Schiavina, M., Pontoriero, L., Uversky, V. N., Felli, I. C., and Pierattelli, R. (2021). The highly flexible disordered regions of the SARS-CoV-2 nucleocapsid N protein within the 1-248 residue construct: Sequence-specific resonance assignments through NMR. *Biomol NMR Assign.* [in press].
- Shen, Z., Wang, G., Yang, Y., Shi, J., Fang, L., Li, F., et al. (2019). A conserved region of nonstructural protein 1 from alphacoronaviruses inhibits host gene expression and is critical for viral virulence. *J. Biol. Chem.* 294 (37), 13606–13618. doi:10.1074/jbc.RA119.009713
- Shin, D., Mukherjee, R., Grewe, D., Bojkova, D., Baek, K., Bhattacharya, A., et al. (2020). Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. *Nature* 587 (7835), 657–662. doi:10.1038/s41586-020-2601-5
- Snijder, E. J., Bredendiek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L., et al. (2003). Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol.* 331 (5), 991–1004. doi:10.1016/s0022-2836(03)00865-9
- Surya, W., Li, Y., and Torres, J. (2018). Structural model of the SARS coronavirus E channel in LMPG micelles. *Biochim. Biophys. Acta Biomembr.* 1860 (6), 1309–1317. doi:10.1016/j.bbmem.2018.02.017
- Sutton, G., Fry, E., Carter, L., Sainsbury, S., Walter, T., Nettleship, J., et al. (2004). The nsp9 replicase protein of SARS-coronavirus, structure and functional insights. *Structure* 12 (2), 341–353. doi:10.1016/j.str.2004.01.016
- Takai, K., Sawasaki, T., and Endo, Y. (2010). Practical cell-free protein synthesis system using purified wheat embryos. *Nat. Protoc.* 5 (2), 227–238. doi:10.1038/nprot.2009.207
- Takeda, M., Chang, C. K., Ikeya, T., Güntert, P., Chang, Y. H., Hsu, Y. L., et al. (2008). Solution structure of the c-terminal dimerization domain of SARS coronavirus nucleocapsid protein solved by the SAIL-NMR method. *J. Mol. Biol.* 380 (4), 608–622. doi:10.1016/j.jmb.2007.11.093
- Tan, J., Vonrhein, C., Smart, O. S., Bricogne, G., Bollati, M., Kusov, Y., et al. (2009). The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. *Plos Pathog.* 5 (5), e1000428. doi:10.1371/journal.ppat.1000428
- Tan, Y., Schneider, T., Leong, M., Aravind, L., and Zhang, D. (2020). Novel immunoglobulin domain proteins provide insights into evolution and pathogenesis of SARS-CoV-2-related viruses. *mBio* 11 (3). doi:10.1128/mBio.00760-20
- Thoms, M., Buschauer, R., Ameismeier, M., Koepke, L., Denk, T., Hirschenberger, M., et al. (2020). Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. *Science* 369 (6508), 1249–1255. doi:10.1126/science.abc8665
- Tonelli, M., Singarapu, K. K., Makino, S., Sahu, S. C., Matsubara, Y., Endo, Y., et al. (2011). Hydrogen exchange during cell-free incorporation of deuterated amino acids and an approach to its inhibition. *J. Biomol. NMR* 51 (4), 467–476. doi:10.1007/s10858-011-9575-4
- Tonelli, M., Rienstra, C., Anderson, T. K., Kirchdoerfer, R., and Henzler-Wildman, K. (2020). ¹H, ¹³C, and ¹⁵N backbone and side chain chemical shift assignments of the SARS-CoV-2 non-structural protein 7. *Biomol. NMR Assign.* 2020, 1–5. doi:10.1007/s12104-020-09985-0
- Tvarogová, J., Madhugiri, R., Bylapudi, G., Ferguson, L. J., Karl, N., and Ziebuhr, J. (2019). Identification and characterization of a human coronavirus 229E nonstructural protein 8-associated RNA 3'-terminal adenylyltransferase activity. *J. Virol.* 93 (12), e00291–e00319. doi:10.1128/JVI.00291-19
- Ullrich, S., and Nitsche, C. (2020). The SARS-CoV-2 main protease as drug target. *Bioorg. Med. Chem. Lett.* 30 (17), 127377. doi:10.1016/j.bmcl.2020.127377
- Venema, H., Godeke, G. J., Rossen, J. W., Voorhout, W. F., Horzinek, M. C., Opstelten, D. J., et al. (1996). Nucleocapsid-independent assembly of coronavirus-like particles by co-expression of viral envelope protein genes. *EMBO J.* 15 (8), 2020–2028. doi:10.1002/j.1460-2075.1996.tb00553.x
- Wacker, A., Weigand, J. E., Akabayov, S. R., Altincekic, N., Bains, J. K., Banijamali, E., et al. (2020). Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. *Nucleic Acids Res.* 48, 12415. doi:10.1093/nar/gkaa1013
- Wang, Y., Kirkpatrick, J., Zur Lage, S., Korn, S. M., Neissner, K., Schwalbe, H., et al. (2021). ¹H, ¹³C, and ¹⁵N backbone chemical-shift assignments of SARS-CoV-2 non-structural protein 1 (leader protein). *Biomol NMR Assign.* doi:10.1007/s12104-021-10019-6
- Wang, S., Fogeron, M. L., Schledorn, M., Dujardin, M., Penzel, S., Burdette, D., et al. (2019). Combining cell-free protein synthesis and NMR into a tool to study capsid assembly modulation. *Front. Mol. Biosci.* 6, 67. doi:10.3389/fmolb.2019.00067
- Wolff, G., Limpens, R. W. A. L., Zevenhoven-Dobbe, J. C., Laugks, U., Zheng, S., de Jong, A. W. M., et al. (2020). A molecular pore spans the double membrane of the coronavirus replication organelle. *Science* 369 (6509), 1395–1398. doi:10.1126/science.abd3629
- Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579 (7798), 265–269. doi:10.1038/s41586-020-2008-3
- Wu, Z., Yang, L., Ren, X., Zhang, J., Yang, F., Zhang, S., et al. (2016). ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. *J. Infect. Dis.* 213 (4), 579–583. doi:10.1093/infdis/jiv476
- Ye, Q., West, A. M. V., Silletti, S., and Corbett, K. D. (2020). Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. *Protein Sci.* 29, 1890. doi:10.1002/pro.3909
- Yin, W., Mao, C., Luan, X., Shen, D. D., Shen, Q., Su, H., et al. (2020). Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science* 368 (6498), 1499–1504. doi:10.1126/science.abc1560
- Yoshimoto, F. K. (2020). The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. *Protein J.* 39 (3), 198–216. doi:10.1007/s10930-020-09901-4
- Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., et al. (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors. *Science* 368 (6489), 409–412. doi:10.1126/science.abb3405
- Zhao, J., Falcón, A., Zhou, H., Netland, J., Enjuanes, L., Pérez Breña, P., et al. (2009). Severe acute respiratory syndrome coronavirus protein 6 is required for optimal replication. *J. Virol.* 83 (5), 2368–2373. doi:10.1128/JVI.02371-08
- Zhou, R., Zeng, R., Von Brunn, A., and Lei, J. (2020). Structural characterization of the C-terminal domain of SARS-CoV-2 nucleocapsid protein. *Mol. Biomed.* 1 (2), 1–11. doi:10.1186/s43556-020-00001-4
- Züst, R., Cervantes-Barragán, L., Kuri, T., Blakqori, G., Weber, F., Ludewig, B., et al. (2007). Coronavirus non-structural protein 1 is a major pathogenicity factor: implications for the rational design of coronavirus vaccines. *PLoS Pathog.* 3 (8), e109. doi:10.1371/journal.ppat.0030109

Conflict of Interest: CH was employed by Signals GmbH & Co. KG.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Altincekic, Korn, Qureshi, Dujardin, Ninot-Pedrosa, Abele, Abi Saad, Alfano, Almeida, Alshamleh, de Amorim, Anderson, Anobom, Anorma, Bains, Bax, Blackledge, Blechar, Böckmann, Brigandat, Bula, Bütikofer, Camacho-Zarco, Carlomagno, Caruso, Ceylan, Chaikuad, Chu, Cole, Crosby, de Jesus, Dharmotharan, Felli, Ferner, Fleischmann, Fogeron, Fourkiotis, Fuks, Fürtig, Gallo, Gande, Gerez, Ghosh, Gomes-Neto, Gorbatyuk, Guseva, Hacker, Häfner, Hao, Hargittay, Henzler-Wildman, Hoch, Hohmann, Hutchison, Jaudzems, Jović, Kaderli, Kalniņš, Kaņepe, Kirchoerfer, Kirkpatrick, Knapp, Krishnathas, Kutz, zur

Lage, Lambertz, Lang, Laurents, Lecoq, Linhard, Löhr, Malki, Bessa, Martin, Matzel, Maurin, McNutt, Mebus-Antunes, Meier, Meiser, Mompeán, Monaca, Montserret, Mariño Perez, Moser, Muhle-Goll, Neves-Martins, Ni, Norton-Baker, Pierattelli, Pontoriero, Pustovalova, Ohlenschläger, Orts, Da Poian, Pyper, Richter, Riek, Rienstra, Robertson, Pinheiro, Sabbatella, Salvi, Saxena, Schulte, Schiavina, Schwalbe, Silber, Almeida, Sprague-Piercy, Spyroulias, Sreeramulu, Tants, Täs, Torres, Töws, Treviño, Trucks, Tsika, Varga, Wang, Weber, Weigand, Wiedemann, Wirmer-Bartoschek, Wirtz Martin, Zehnder, Hengesbach and Schlundt. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

GLOSSARY

aa Amino acid	IPRS IMAC-protease cleavage-reverse IMAC-SEC;
BEST Band-selective excitation short-transient	M Membrane protein
BMRB Biomagnetic resonance databank	MERS Middle East Respiratory Syndrome
CFPS Cell-free protein synthesis	MHV Murine hepatitis virus
CoV Coronavirus	M^{pro} Main protease
CTD C-terminal domain	MTase Methyltransferase
DEDD Asp-Glu-Glu-Asp	N Nucleocapsid protein
DMS Dimethylsulfate	NAB Nucleic acid-binding domain
E Envelope protein	nsp Nonstructural protein
ED Ectodomain	NTD N-terminal domain
fl Full-length	PL^{pro} Papain-like protease
GB1 Protein G B1 domain	RdRP RNA-dependent RNA polymerase
GD Globular domain	S Spike protein
GF Gel filtration	SARS Severe Acute Respiratory Syndrome
GST Glutathione-S-transferase	SEC Size-exclusion chromatography
His Hisx-tag	SUD SARS unique domain
HSP Heat-shock protein	SUMO Small ubiquitin-related modifier
HSQC Heteronuclear single quantum coherence	TEV Tobacco etch virus
IDP Intrinsically disordered protein	TM Transmembrane
IDR Intrinsically disordered region	TROSY Transverse relaxation-optimized spectroscopy
IEC Ion exchange chromatography	Trx Thioredoxin
IMAC Immobilized metal ion affinity chromatography	Ubl Ubiquitin-like domain
	Ulp1 Ubiquitin-like specific protease 1
	WG Wheat-germ.

Table S1: Overview of labs as assigned to protocols

Protein	System	Main protocol from group(s)	Protocol as „additional information“ from group(s)
nsp1	Bacterial	Carlomagno (fl), Schlundt (GD)	Schlundt (fl)
nsp2	Bacterial	Laurents (CtDR)	-
nsp3a	Bacterial	Blackledge (UBI+IDR), Schlundt (UBI)	-
nsp3b	Bacterial	Schwalbe (Macrodomain)	Alfano (Macrodomain)
nsp3c	Bacterial	Spyroulias (SUD-N, SUD-NM, SUD-M, SUD-MC, SUD-C)	-
nsp3d	Bacterial	Schwalbe (PL ^{pro})	Schwalbe (PL ^{pro})
nsp3e	Bacterial	Schlundt (NAB)	Schlundt (NAB)
nsp3Y	Bacterial	Hoch (CoV-Y)	-
nsp5	Bacterial	Schwalbe (fl)	Schwalbe (fl, A-D), Orts (fl, E), Varga (fl, F), Bax ((fl, G-H), Martin ((fl, I)
nsp7	Bacterial	Henzler-Wildman/Kirchdoerfer (fl)	-
nsp8	Bacterial	Henzler-Wildman/Kirchdoerfer (fl)	-
nsp9	Bacterial	Schlundt (fl)	Schlundt (fl, A), Alfano (fl, B)
nsp10	Bacterial	Schwalbe (fl)	Jaudzems (fl)
nsp13	Bacterial	Schwalbe (fl)	Schwalbe (fl)
nsp14	Bacterial	Jaudzems (fl, MTase)	Schwalbe (fl)
nsp15	Bacterial	Schwalbe (fl)	-
nsp16	Bacterial	Jaudzems (fl)	Jaudzems (fl)
ORF3a	Cell-free	Böckmann (fl)	-
Envelope (ORF4)	Cell-free	Böckmann/Meier (fl)	-
Membrane (ORF5)	Cell-free	Böckmann/Meier (fl)	Böckmann/Meier (fl)
ORF6	Cell-free	Böckmann (fl)	Böckmann (fl)
ORF7a	Bacterial	Muhle-Goll (ED)	-
ORF7b	Bacterial	Schwalbe (fl)	Schwalbe (fl, A-E)
	Cell-free	Böckmann (fl)	-

ORF8	Bacterial	Wiedemann/Ohlenschläger (fl-L84S) Alfano (w/o signal peptide (Δ))	Wiedemann/Ohlenschläger (fl)
	Cell-free	Böckmann (fl, Δ)	-
Nucleo-capsid (ORF9a)	Bacterial	Pierattelli/Felli (IDR1-NTD-IDR2), Almeida (NTD-SR, NTD), Schlundt (CTD)	-
ORF9b	Cell-free	Böckmann (fl)	Böckmann (fl, A-B)
ORF14	Cell-free	Böckmann/Meier (fl)	-
ORF10	Bacterial	Schwalbe (fl)	Schwalbe (fl, A-D)

Table S2: Abbreviations used throughout the SI

Abbreviation	Full name
aa	Amino acid
AC	Affinity chromatography
BEST	Band-selective Excitation Short-Transient
BisTris	2,2-Bis(hydroxymethyl)-2,2',2''-nitrilotriethanol
bME	2-mercaptoethanol
BMRB	Biomagnetic Resonance Databank
Brij 58	Polyethylene glycol hexadecyl ether, Polyoxyethylene (20) cetyl ether
CFPS	Cell-free protein synthesis
CFS	Cell-free sample
CoV	Coronavirus
CTD	C-terminal domain
DDM	n-dodecyl β -D-maltoside
<i>E. coli</i>	<i>Escherichia coli</i> cells
ED	Ectodomain
fl	Full-length
GB1	Protein G B1 domain
GD	Globular domain
GST	Glutathione-S-transferase
His ₆ (analog His ₇)	Hexahistidine tag
HSQC	Heteronuclear single quantum coherence
IDR	Intrinsically disordered region
IEC	Ion exchange chromatography
IMAC	Immobilized metal ion affinity chromatography
Inv.	Inverse
IPTG	Isopropyl- β -d-thiogalactopyranoside
LB medium	Lysogeny broth medium
M9 medium	M9 minimal medium
MOPS	3-(N-morpholino)propanesulfonic acid, 4-morpholinepropanesulfonic acid
M ^{pro}	Main protease

MTase	Methyltransferase
MWCO	Molecular weight cut-off
NAB	Nucleic acid-binding domain
NaPi/KPi	Sodium/potassium phosphate
NA	Not available
n.d.	Not defined/no information available
nsp	Non-structural protein
NTA	Nitrilotriacetic acid
NTD	N-terminal domain
o.n.	Overnight
OD ₆₀₀	Optical density at 600 nm
ORF	Open reading frame
PDB	Protein Data Bank
PL ^{pro}	Papain-like protease
rt	Room temperature
S, SARS	Severe acute respiratory syndrome
SD	Superdex
SEC	Size exclusion chromatography
SN	Soluble fraction, supernatant
SUD	SARS unique domain
SUMO	Small ubiquitin-like modifier
TCEP	Tris-(2-carboxyethyl)-phosphin
TEV	Tobacco etch virus
Triton X-100	4-(1,1,3,3-Tetramethylbutyl)-phenyl-polyethylenglykol
TROSY	Transverse relaxation-optimized spectroscopy
Trx	Thioredoxin
Ubl	ubiquitin-like domain
Ulp1	ubiquitin-like specific protease 1
WB	Western blot
WG(E)	Wheat germ (extract)
YT medium	Yeast extract-tryptone medium

SI1: nsp1

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp1
2	Region/Name/Further Specification
	nsp1 / Leader protein
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEKGVLPQ LEQPYVFIKRS DARTAPHGHVMVELVAELEGIQYGRSGETLGVLPVPHVGEIPVAYRKVLLRKNG NKGAGGHSYGADLKSFDLGDDELGTDPEYEDFQENWNTKHSSGV TRELMRELNGG
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
fl	aa 1-180 (fl nsp1)
GD	aa 13-127 of fl nsp1
5	Ratio for construct design
fl	fl sequence according to NCBI Reference Sequence YP_009725297.1
GD	In analogy to the available NMR structure (PDB 2GDT) of nsp1 SCoV 13-127
6	Sequence homology (to SCoV)
fl	Identity: 83%; similarity: 89%
GD	Identity: 85%; similarity: 90%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PBD 2GDT, 2HSX SCoV2: PBD 7K3N, 7K7P, 6ZN5, 7JQC, 7K5I
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: BMRB 7014 SCoV2: BMRB 50620

Table 2: Protein Expression

1	Expression vector
fl	pETM11 (Gunter Stier, EMBL Heidelberg)
GD	pKM263 (GenScript)
2	Purification-/Solubility-Tag
fl	N-terminal His ₆
GD	N-terminal His ₆
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein

fl	19.90 kDa / 12,950 M ⁻¹ cm ⁻¹ / 5.37
GD	12.93 kDa / 4,470 M ⁻¹ cm ⁻¹ / 6.22
5	Comments on sequence of expressed construct
fl	N-terminal „GA" two artificial residues due to TEV-cleavage and construct design
GD	N-terminal „GAMA" four artificial residues due to TEV-cleavage and construct design
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
fl	0.6 mM IPTG at OD ₆₀₀ 0.7
GD	1 mM IPTG
10	Cultivation temperature and time
fl	16°C for 18-20 h
GD	16°C for 18-20 h

Table 3a: Protein Purification (fl nsp1)

1	Buffer List
A	50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 100 mM Na ₂ SO ₄ , 5% (v/v) glycerol, 5 mM imidazole, 1 mM TCEP-HCl (cell disruption / immobilized metal affinity chromatography (IMAC) / TEV-cleavage).
B	50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 100 mM Na ₂ SO ₄ , 1 mM EDTA, 1 mM TCEP-HCl (SEC).
C	50 mM NaPi (pH 6.5), 200 mM NaCl, 2 mM DTT, 2 mM EDTA (final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus one tablet of EDTA free protease inhibitor cocktail (Roche), 100 µg of lysozyme (Carl Roth), and 50 µg of deoxyribonuclease (DNase) (New England Biolabs)) by sonication.
B	IMAC (gravity flow Ni ²⁺ -NTA (Cytiva)), washed first with buffer 1A and then with buffer 1A containing additional 2 M LiCl, before eluting with 300 mM imidazole in buffer 1A .
C	Desalting and TEV-cleavage (0.5 mg TEV protease per 1 L culture) o.n. in buffer 1A .
D	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer 1B .
E	NMR sample preparation in buffer 1C .

Table 3b: Protein Purification (GD nsp1)

1	Buffer List
A	50 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 mM imidazole, 4 mM DTT (cell disruption / IMAC/ dialysis after IMAC / TEV-cleavage).
B	25 mM NaPi (pH 7.0), 250 mM NaCl, 2 mM TCEP-HCl, 0.02% (w/v) NaN ₃ (SEC / final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus 100 μ L protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni ²⁺ -NTA), Elution with 150-500 mM imidazole in buffer 1A .
C	Dialysis o.n. in in buffer 1A .
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer 1A .
E	SEC on HiLoad SD 75 16/600 (GE Healthcare) in buffer 1B .
F	NMR sample preparation in buffer 1B .

Table 4: Final samples

1	Yield
fl	5 mg/L ¹³ C, ¹⁵ N-M9 medium
GD	< 0.5 mg/L ¹⁵ N-M9 medium
2	Stability
fl	No significant precipitation or degradation observed after storage at 4°C for 3 weeks. Relatively stable during NMR measurements at 25°C for ~7 days, despite some proteolysis of disordered C-terminal tail.
GD	Stable during several weeks storage at 4°C.
3	Comment on applicability
fl	Suitable for NMR structure determination, fragment screening, interaction studies.
GD	purification needs optimization to obtain more soluble protein

Additional information

Constructs	Conditions	Comments
aa 1-180 (fl nsp1); His ₇ (pET-TEV-Nco (GenScript)), TEV-cleavage site, N-terminal 2 artificial residues "GA".	As above for GD nsp1.	Yields 2.4 mg/L ¹⁵ N, ¹³ C-M9 medium. Obvious degradation during measurement. Storage at 4°C not advisable. Higher salt concentration seems to slightly improve stability.

SI2: nsp2

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp2
2	Region/Name/Further Specification
	C-terminal IDR (CtDR)
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTE RSEKSYELQTPFEIKLAKKFDTFNGECPNFVFPPLNSIIKTIQPRVEKKKLDGFMGRIRSVYPV NECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGATTCGYLPQNAVVKIYCP ACHNSEVGPESHSLAEYHNESGLKTILRKGGRITAFGGCVFSYVGCHNKCAWVPRASANIGCN HTGVVGESEGLNDNLEILQKEKVNINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFK QIVESCGNFKVTKGKAKKGAWNIGEQQSILSPLYAFASEAARVVRSIFSRTLETAQNSVRVLQK AAITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTFNIFGTVEKLPVL DWLEEKFKEGVEFLRDGWEIVKFISTCACEIVGGQIVTCAKEIKESVQTFKLVNKFALCADSII IGGAKLKLALNLGETFVTHSKGLYRKCYSREETGLLMPLKAPKEIFLEGETLPTEVLTEEVVLK TGDLPQLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEKYCALAPNMMVNTNTFTLKG
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 557-601 of complete nsp2 (Ct-DR)
5	Ratio for construct design
	Based on disorder predictions (PrDOS (Ishida and Kinoshita, 2007))
6	Sequence homology (to SCoV)
	Identity: 55%; similarity: 68%
7	Published structures (SCoV2 or homologue variants)
	-
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: 50687

Table 2: Protein Expression

1	Expression vector
	Home made plasmid derived from pET28b(+) (EMD Biosciences) containing the codifying sequence for thioredoxin A from <i>E. coli</i> and TEV protease cleavage site instead of thrombin.
2	Purification-/Solubility-Tag
	N-terminal His ₆ -Trx
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein

	4.92 kDa / - / 3.9
5	Comments on sequence of expressed construct
	N-terminal „G“, one artificial residue due to TEV-cleavage.
6	Used expression strain
	<i>E. coli</i> BL21 star (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.5 mM IPTG at OD ₆₀₀ 0.6
10	Cultivation temperature and time
	37°C until induction. Following induction, incubation at 25°C for 17 h

Table 3: Protein Purification

1	Buffer List
A	50 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 mM imidazole (cell lysis, IMAC1 and 2).
B	5 mM Tris-HCl (pH 8.0), 20 mM NaCl (dialysis after IMAC1/TEV cleavage).
C	5 mM histidine (pH 5.4), 5 mM NaCl (dialysis after IMAC2 and anionic IEC).
D	10 mM acetic acid (pH 4.3), 5 mM NaCl (dialysis after cationic IEC).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell lysis in 1A (plus 5 µL Halt protease inhibitor (Thermo) and lysozyme 20 µg/mL).
B	IMAC1 (HisTrap crude 5 mL, Cytiva). Elution 10-500 mM imidazole in buffer 1A .
C	Dialysis in buffer 1B and TEV cleavage (4°C, 17 h).
D	IMAC2 (after TEV cleavage) (HisTrap crude 5 mL, Cytiva). Elution 10-500 mM imidazole in buffer 1A (protein expected in flow-through).
E	Dialysis in buffer 1C (4°C, 17 h).
F	Anionic IEC, elution 10-1,000 mM NaCl in buffer 1C .
G	Dialysis in buffer 1C (4°C, 48 h).
H	Cationic IEC. Elution 10-1,000 mM NaCl in buffer 1D (protein expected in flow-through).

Table 4: Final sample

1	Yield
	1.5 mg/L LB medium, 0.7-1.5 mg/L ¹³ C, ¹⁵ N-M9 medium
2	Stability
	No visible precipitation after two weeks at 4°C.
3	Comment on applicability
	Suitable for NMR structure determination, fragment screening, interaction studies.

SI3: nsp3a

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp3
2	Region/Name/Further Specification
	nsp3a Ubiquitin-like domain (Ubl) + IDR
3	Sequence of “fl” protein (aa 1-206 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)
	APTKVTFGDDTVIEVQGYKSVNITFELDERIDKVLNEKCSAYTVELGTEVNEFACVVADAVIKT LQPVSELLTPLGIDLDEWSMATYYLFDSEGEFKLASHMYCSFYPPDEDEEEGDCEEEEFEPSTQY EYGTEDDYQGKPLEFGATSAAALQPEEEQEEDWLDDDSQQTVGQQDGSSEDNQTTTIQTIVEVQP QLEMELTPVVQTIE
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
Ubl+ IDR	aa 1-206 of complete nsp3
Ubl	aa 1-111 of complete nsp3
5	Ratio for construct design
Ubl+ IDR	Based on homologous structure from SCoV.
Ubl	Based on disorder prediction, folded domain and SCoV Ubl1.
6	Sequence homology (to SCoV)
Ubl+ IDR	Identity: 58%; Similarity: 75%
Ubl	Identity: 79%; Similarity: 89%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 2GRI; 2IDY
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: BMRB 7019 SCoV2: BMRB 50446

Table 2: Protein Expression

1	Expression vector
Ubl+ IDR	pET-TEV-Nco (GenScript)
Ubl	pKM263 (GenScript)
2	Purification-/Solubility-Tag
Ubl+ IDR	N-terminal His ₆

Ubl	N-terminal His ₆ -GST
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
Ubl+ IDR	23.50 kDa / 24,410 M ⁻¹ cm ⁻¹ / 3.62
Ubl	12.72 kDa / 14,440 M ⁻¹ cm ⁻¹ / 4.08
5	Comments on sequence of expressed construct
Ubl+ IDR	N-terminal "GAM" three artificial residues due to TEV-cleavage and construct design.
Ubl	N-terminal "GAMG" four artificial residues due to TEV-cleavage and construct design.
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	1 mM IPTG at OD ₆₀₀ 0.6-0.8
10	Cultivation temperature and time
Ubl+ IDR	37°C for 5 h
Ubl	18°C for 18 h

Table 3a: Protein Purification (Ubl + IDR)

1	Buffer List
A	50 mM Tris-HCl (pH 8.0), 150 mM NaCl and complete EDTA-free tablet (cell disruption).
B	50 mM Tris-HCl (pH 8.0) and 150 mM NaCl (wash buffer).
C	50 mM Tris-HCl (pH 8.0), 150 mM NaCl and 500 mM imidazole (elution buffer).
D	50 mM Tris-HCl (pH 8.0), 150 mM NaCl and 5 mM bME (TEV cleavage).
E	50 mM NaPi (pH 6.5), 250 mM NaCl (final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Resuspension of cell pellet in 50 mL per liter of culture of 1A at 4°C.
B	Cell disruption by sonication on ice.

C	Clarification of lysate by centrifugation at 16,000 g for 30 min at 4°C.
D	Loading of lysate on Ni ²⁺ -loaded IMAC resin (ThermoFisher scientific) pre-equilibrated with 1B at 22°C.
E	Wash IMAC resin with 50 bed volumes of 1B .
F	Elute protein from IMAC resin with 5 bed volumes of 1C .
G	TEV cleavage with 1 mg TEV per 50 mg protein by dialysis against 1D for 18 h at 4°C.
H	Removal of uncleaved protein and tag by elution through Ni ²⁺ -loaded IMAC resin pre-equilibrated with 1B at 22°C.
I	Wash with 5 bed volumes of 1B .
J	SEC with HiLoad SD 75 pg column (GE Healthcare) pre-equilibrated with 1E at 4°C.

Table 3b: Protein Purification (Ubl)

1	Buffer List
A	50 mM NaPi (pH 6.5), 300 mM NaCl, 10 mM imidazole, 2 mM TCEP-HCl (Cell disruption / IMAC)
B	25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM DTT, 0.02% NaN ₃ (dialysis after IMAC / TEV-cleavage)
C	25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl, 0.02% NaN ₃ , pH7 (SEC / final NMR buffer)
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus 100 µL protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni ²⁺ -NTA), Elution with 150-500 mM imidazole in buffer 1A
C	Dialysis o.n. in in buffer 1B
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer 1B
E	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer 1C
F	NMR sample preparation in buffer 1C

Table 4: Final sample

1	Yield
Ubl+ IDR	0.7 mg/L ¹⁵ N-M9 medium
Ubl	2-3 mg/L ¹⁵ N-M9 medium
1b	A260/280 ratio
Ubl+ IDR	0.57
Ubl	0.6
2	Stability
Ubl+	2 weeks at 25°C.

IDR	
Ubl	Very stable over weeks.
3	Comment on applicability
Ubl+ IDR	Stable for NMR assignments and screening
Ubl	Stable for NMR assignments and screening (spectra overlay with folded part of nsp3a Ubl + IDR above.)

SI3: nsp3b

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp3
2	Region/Name/Further Specification
	nsp3b / Macrodomain
3	Sequence of “fl” protein (aa 207-376 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)
	VNSFSGYLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATNNAMQV ESDDYIATNGPLKVGGSCLVLSGHNLAHLHVVGPVNVNKGEDIQLLSAYENFNQHEVLLAPL LSAGIFGADPIHSLRVCVDTVRTNVYLAVF DNLYDKLVSSFLEMK
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 207-376 of complete nsp3
5	Ratio for construct design
	Based on homologous structure from SCoV (PDB 6VXS).
6	Sequence homology (to SCoV)
	Identity: 74%; similarity: 84%
7	Published structures (SCoV2 or homologue variants)
	SCoV2: PDB 6W6Y, 6YWM, 6YWL, 6YWK, 6WEY, 7KG3, 6W02, 6WOJ, 6WEN, 6WCF, 6VXS, 7JME
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV2: BMRB 50387 (apo), 50388 (holo)

Table 2: Protein Expression

1	Expression vector
	pET28a(+) (GenScript)
2	Purification-/Solubility-Tag
	N-terminal His ₆
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	18.65 kDa / 10,430 M ⁻¹ cm ⁻¹ / 7.20
5	Comments on sequence of expressed construct
	N-terminal “GHM” three artificial residues due to TEV-cleavage and construct design.

6	Used expression strain
	<i>E. coli</i> T7 Express
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.2 mM IPTG at OD ₆₀₀ 0.6-0.7
10	Cultivation temperature and time
	18-20°C for 16-18 h

Table 3: Protein Purification

1	Buffer List
A	25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME (cell disruption / IMAC).
B	25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 mM bME (dialysis after IMAC / TEV-cleavage).
C	25 mM BisTris (pH 6.5), 150 mM NaCl, 3 mM TCEP-HCl (SEC / final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus one tablet of EDTA free protease inhibitor cocktail (Merck)) by microfluidization.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer 1A .
C	TEV-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer 1B .
D	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with 500 mM imidazole in buffer 1A .
E	SEC (HiLoad 26/600 SD 200 pg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer 1C (elution volume 245-290 mL).
F	NMR sample preparation in buffer 1C .

Table 4: Final sample

1	Yield
	94 mg/L ¹⁵ N-M9 medium, 9 mg/L ¹³ C, ¹⁵ N-M9 medium
2	Stability
	Stable throughout measurement (7 days, 298 K). No significant precipitation or degradation observed after storage at 4°C for 2 weeks.
3	Comment on applicability

Suitable for NMR structure determination, fragment screening, interaction studies.

Additional information

Constructs	Conditions	Comments
aa 206-374 of complete nsp3; His ₆ -GST (mod pET9d), TEV-cleavage site, N-terminal "GAM" three artificial residues. Based on boundaries from crystal structure (PDB 6W6Y).	IMAC buffer: 50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 5% (v/v) glycerol, 50 mM imidazole, 1 mM DTT. Cleavage buffer: 50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 1 mM DTT. SEC/final buffer: 20 mM NaPi (pH 7.4), 150 mM NaCl, 3 mM TCEP-HCl.	Yields 30 mg/L LB medium. No significant precipitation or degradation observed after storage at 4°C for 10 days. Suitable for NMR studies, fragment-based screening, interaction studies.

SI3: nsp3c

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp3
2	Region/Name/Further Specification
SUD-N	nsp3c / SARS Unique Domain (SUD) -N
SUD-NM	nsp3c / SUD-NM
SUD-M	nsp3c / SUD-M
SUD-MC	nsp3c / SUD-MC
SUD-C	nsp3c / SUD-C
3	Sequence of “fl” protein (aa 409-743 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)
	QDDKKIKACVEEVTTTLEETKFLTENLLYIDINGNLHPDSATLVSDIDITFLKKDAPYIVGDVV QEGVLTAVVIPTKKAGGTTEMLAKALRKVPTDNYITTYPGQLNGYTVEEAKTVLKKCKSAFY ILPSIISNEKQEILGTVSWNLREMLAHAEETRKLMPVCVETKAIVSTIQRKYKGIKIQEGVVDYG ARFYFYTSKTTVASLINTLNDLNETLVTMPLGYVTHGLNLEEAAARYMRSCLKVPATVSVSPDA VTAYNGYLTSSSKTPEEHFIETISLAGSYKDWYSYSGQSTQLGIEFLKRGDKSVYYTSPNPTTFHLD GEVITFDNLKTLLS
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
SUD-N	aa 409-548 of complete nsp3
SUD-NM	aa 409-675 of complete nsp3
SUD-M	aa 551-675 of complete nsp3
SUD-MC	aa 551-743 of complete nsp3
SUD-C	aa 680-743 of complete nsp3
5	Ratio for construct design
SUD-N	Based on X-ray structure of homologue nsp3c from SCoV (PDB 2W2G).
SUD-NM	Based on X-ray structure of homologue nsp3c from SCoV (PDB 2W2G).
SUD-M	Based on X-ray structure of homologue nsp3c from SCoV (PDB 2W2G).
SUD-MC	Based on NMR structure of homologue nsp3c from SCoV (PDB 2KQV, 2KQW).
SUD-C	Based on NMR structure of homologue nsp3c from SCoV (PDB 2KAF).
6	Sequence homology (to SCoV)

SUD-N	Identity: 69%, similarity: 81.6%
SUD-NM	Identity: 74%, similarity: 85.4%
SUD-M	Identity: 82%, similarity: 89.6%
SUD-MC	Identity: 79%, similarity: 88.7%
SUD-C	Identity: 73%, similarity: 87.7%
7	Published structures (SCoV2 or homologue variants)
	-
8	(Published) assignment (SCoV2 or homologue variants)
SUD-N	SCoV2: BMRB 50448
SUD-NM	Ongoing
SUD-M	SCoV2: BMRB 50516 SUD-M
SUD-MC	Ongoing
SUD-C	SCoV2: BMRB 50517 SUD-C

Table 2: Protein Expression

1	Expression vector
SUD-N	pGEX4T1 (Addgene)
SUD-NM	pGEX4T1 (Addgene)
SUD-M	pET28a(+) (Addgene)
SUD-MC	pET28a(+) (Addgene)
SUD-C	pGEX4T1 (Addgene)
2	Purification-/Solubility-Tag
SUD-N	N-terminal GST
SUD-NM	N-terminal GST
SUD-M	N-terminal His ₆
SUD-MC	N-terminal His ₆
SUD-C	N-terminal GST

3	Cleavage Site
	Thrombin
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
SUD-N	15.54 kDa / 8,940 M ⁻¹ cm ⁻¹ / 5.04
SUD-NM	29.60 kDa / 26,360 M ⁻¹ cm ⁻¹ / 6.03
SUD-M	14.27 kDa / 17,420 M ⁻¹ cm ⁻¹ / 8.71
SUD-MC	21.94 kDa / 28,880 M ⁻¹ cm ⁻¹ / 6.58
SUD-C	7.42 kDa / 11,460 M ⁻¹ cm ⁻¹ / 4.82
5	Comments on sequence of expressed construct
SUD-N	N-terminal „GS" two artificial residues due to thrombin-cleavage
SUD-NM	N-terminal „GS" two artificial residues due to thrombin-cleavage
SUD-M	N-terminal „GSHM" four artificial residues due to thrombin-cleavage and cloning
SUD-MC	N-terminal „GSHM" four artificial residues due to thrombin-cleavage and cloning
SUD-C	N-terminal „GS" two artificial residues due to thrombin-cleavage
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	1 mM IPTG at OD ₆₀₀ 0.6-0.8
10	Cultivation temperature and time
	18°C for 18-20 h

Table 3a: Protein Purification (SUD-N and SUD-NM)

1	Buffer List
A	50 mM Tris-HCl (pH 8.0), 300 mM NaCl (cell disruption / affinity chromatography (AC)).
B	50 mM NaPi (pH 7.2), 50 mM NaCl, 2 mM EDTA, 2 mM DTT (SEC / NMR buffer).

C	50 mM Tris-HCl (pH 8.0), 10 mM reduced glutathione (elution buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus 25 μ L protease inhibitor cocktail (Sigma Aldrich P8849) and 2 mM DTT) by sonication, after sonication incubation with 25 μ L DNase (1 mg/mL) for 10 min on ice.
B	AC - GSTrap (GE Healthcare) (wash buffer 1A).
C	Cleavage on column (100 μ L thrombin (10 mg/mL) per 0.5 L culture) at 4°C for 16 h.
D	Elution of SUD-N, SUD-NM after cleavage with buffer 1A , elution of GST with buffer 1C and buffer exchange with Amicon Ultra 15 mL centrifugal filter membrane (10,000 MWCO) (Merck Millipore) to buffer 1B .
E	SEC - SD Increase 75 10/300 GL (GE Healthcare) in buffer 1B .
F	NMR sample preparation in buffer 1B .

Table 3b: Protein Purification (SUD-M and SUD-MC)

1	Buffer List
A	50 mM Tris-HCl (pH 8.0), 500 mM NaCl (Cell disruption / IMAC).
B SUD-M	50 mM NaPi (pH 7.2), 50 mM NaCl, 2 mM EDTA, 2 mM DTT (SEC / NMR buffer).
B SUD-MC	50 mM NaPi (pH 7.6), 50 mM NaCl, 2 mM EDTA, 2 mM DTT (SEC / NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus 10 mM imidazole and 25 μ L protease inhibitor cocktail (Sigma Aldrich P8849) and 2 mM DTT) by sonication, before and after sonication incubation with 50 μ L DNase (1 mg/mL) for 15 min on ice.
B	IMAC - HisTrap (Ni ²⁺) (GE Healthcare), a step gradient elution of imidazole in buffer 1A (10, 20, 40, 100, 200, 400 mM). SUD-M eluted mostly in 100 mM imidazole in buffer 1A and a small amount in fraction 200 mM imidazole in buffer 1A . SUD-MC eluted mostly in 100 mM imidazole in buffer 1A and a small amount in 40 mM imidazole in buffer 1A .
C	Buffer exchange with Amicon Ultra 15 mL centrifugal filter membrane (10,000 MWCO) (Merck Millipore) in buffer 1B SUD-M and SUD-MC respectively.
D	Cleavage in solution (100 μ L thrombin (10 mg/mL) per 0.5 L culture) for SUD-M : 1 h at 4°C and then 1 h at rt; SUD-MC : 16 h at 4°C.
E	SEC - Superdex Increase 75 10/300 GL (GE Healthcare) in buffer 1C-SUD-M , 1C-SUD-MC .
F	NMR sample preparation in buffer 1C-SUD-M , 1C-SUD-MC .

Table 3c: Protein Purification (SUD-C)

1	Buffer List
A	50 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10% (v/v) glycerol (cell disruption / AC).
B	50 mM Tris-HCl (pH 8.0), 300 mM NaCl (AC).

C	50 mM Tris-HCl (pH 8.0), 10 mM reduced glutathione (elution buffer).
D	50 mM NaPi (pH 7.2), 50 mM NaCl, 2 mM EDTA, 2 mM DTT (SEC / NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus 25 μ L protease inhibitor cocktail (Sigma Aldrich P8849) and 2 mM DTT) by sonication, after sonication incubation with 25 μ L DNase (1 mg/mL) for 10 min on ice.
B	AC with GSTrap (GE Healthcare) (wash buffer 1A and then wash with buffer 1B).
C	Elution with buffer 1C , buffer exchange with Amicon Ultra 15 mL centrifugal filter membrane (10,000 MWCO) (Merck Millipore) to buffer 1D .
D	Cleavage in solution (350 μ L thrombin (10 mg/mL) per 0.5 L culture) at 37°C for 5 h.
E	SEC on SD Increase 75 10/300 GL (GE Healthcare) in buffer 1D .
F	NMR sample preparation in buffer 1D .

Table 4: Final sample

1	Yield
SUD-N	13.92 mg/L ^{15}N or ^{13}C , ^{15}N -M9 medium
SUD-NM	17.25 mg/L ^{15}N or ^{13}C , ^{15}N -M9 medium
SUD-M	8.47 mg/L ^{15}N or ^{13}C , ^{15}N -M9 medium
SUD-MC	12.06 mg/L ^{15}N or ^{13}C , ^{15}N -M9 medium
SUD-C	4.70 mg/L ^{15}N or ^{13}C , ^{15}N -M9 medium
1b	A260/280 ratio
SUD-N	0.55
SUD-NM	0.50
SUD-M	0.81
SUD-MC	0.62
SUD-C	0.71
2	Stability
SUD-N	Stable throughout NMR spectra acquisition (10 days, 298 K). No significant precipitation or degradation observed after thawing from -80°C. Very stable construct.
SUD-NM	Stable throughout measurement (7 days, 298 K). No significant precipitation or degradation observed after defrosting from -80°C.
SUD-M	Not very stable throughout spectra acquisition, 10 days 298 K. Significant precipitation observed after thawing from storage at -80°C. Forms dimers without reducing agent observable even by SDS-page.

SUD-MC	Stable throughout measurement (7 days, 298 K). No significant precipitation or degradation observed after thawing from -80°C.
SUD-C	Stable throughout measurement (10 days, 298 K). No significant precipitation or degradation observed after thawing from -80°C. Stable construct.
3	Comment on applicability
	Suitable for NMR structure determination, fragment screening, interaction studies.

SI3: nsp3d

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp3
2	Region/Name/Further Specification
	nsp3d / papain-like protease / PL ^{pro}
3	Sequence of “fl” protein (aa 743-1060 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)
	SLREVRTIKVFTTVDNINLHTQVVDMSMTYGQQFGPTYLDGADVTKIKPHNSHEGKTFYVLPN DDTLRVEAFEYYHTTDPSPFLGRYMSALNHTKKWKYPQVNGLTSLIKWADNNCYLATALLTQQ IELKFNPPALQDAYRARAGEAANFCALILAYCNKTVGELGDVRETMSYLFQHANLDSCRVL NVVCKTCGQQTTLKGVAVMYMGTLSEYQFKKGVQIPCTCGKQATKYLQQESPFVMMMSA PPAQYELKHGTFTCASEYTGNYQCGHYKHITSKETLYCIDGALLTKSSEYKGPITDVFYKENSY TTTIK
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 743-1060 of complete nsp3
5	Ratio for construct design
	Based on homologous structure from SCoV (PDB 4M0W)
6	Sequence homology (to SCoV)
	Identity: 83%; similarity: 91%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 4M0W, 2FE8 SCoV2: PDB 6W9C
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Protein Expression

1	Expression vector
	pE-SUMO (LifeSensors)
2	Purification-/Solubility-Tag
	N-terminal His ₆ -SUMO
3	Cleavage Site
	Ulp1
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	35.99 kDa / 45,270 M ⁻¹ cm ⁻¹ / 8.17
5	Comments on sequence of expressed construct

	No artificial residues due to Ulp1-cleavage and construct design.
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.2 mM IPTG at OD ₆₀₀ 0.6-0.7 (addition of 50 μM ZnCl ₂)
10	Cultivation temperature and time
	18-20°C for 16-18 h

Table 3: Protein Purification

1	Buffer List
A	20 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 mM imidazole, 50 μM ZnCl ₂ , 10 mM bME (cell disruption / IMAC).
B	10 mM HEPES (pH 7.4), 100 mM NaCl, 50 μM ZnCl ₂ , 10 mM bME (dialysis after IMAC / TEV-cleavage).
C	10 mM HEPES (pH 7.4), 100 mM NaCl, 50 μM ZnCl ₂ , 5 mM DTT (SEC).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (addition of 50 μM ZnCl ₂) by microfluidization.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer 1A .
C	Ulp1-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer 1B .
D	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with 500 mM imidazole in buffer 1A .
E	SEC (HiLoad 26/600 SD 75 μg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer 1C (elution volume 180-220 mL).

Table 4: Final sample

1	Yield
	12 mg/L ¹⁵ N-M9 medium
2	Stability
	Tendency to aggregate.
3	Comment on applicability

Suitable for fragment screening, interaction studies.

Additional information

Constructs	Conditions	Comments
aa 743-1060 of complete nsp3; His ₆ (pET28a(+)) (GenScript), TEV-cleavage site, N-terminal "GHM" three artificial residues.	Native (as above)	Weak expression, less protein.

SI3: nsp3e

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp3
2	Region/Name/Further Specification
	nsp3e / NAB globular domain
3	Sequence of “fl” protein (aa 1080-1203 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)
	YFTEQPIDLVPNQYPNASFDNFKFVCDNIKFADDLNQLTGYKKPASRELKVTFFPDLNGDVVA IDYKHYTPSFKKGAKLLHKPIVWHVNNATNKATYKPNTWCIRCLWSTKPVET
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1088-1203 of complete nsp3
5	Ratio for construct design
	Based on boundaries from NMR structure of homologue nsp3e from SARS-CoV (2K87).
6	Sequence homology (to SCoV)
	Identity: 82%; similarity: 89%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 2K87
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: BMRB 15723; SCoV2: BMRB 50334

Table 2: Protein Expression

1	Expression vector
	pKM263 (GenScript)
2	Purification-/Solubility-Tag
	N-terminal His ₆ -GST
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	13.75 kDa / 25,565 M ⁻¹ cm ⁻¹ / 8.9
5	Comments on sequence of expressed construct
	N-terminal „GAMG” four artificial residues due to TEV-cleavage and construct design.
6	Used expression strain

	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	1 mM IPTG at OD ₆₀₀ 0.7
10	Cultivation temperature and time
	20-22°C for 18-20 h

Table 3: Protein Purification

1	Buffer List
A	50 mM NaPi (pH 6.5), 300mM NaCl, 10 mM imidazole, 2 mM TCEP-HCl (cell disruption / IMAC).
B	25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM DTT, 0.02% (w/v) NaN ₃ (dialysis after IMAC / TEV-cleavage).
C	25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl, 0.02% (w/v) NaN ₃ (SEC / final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus 100 µL protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni ²⁺ -NTA) (Carl Roth, Germany), elution with 150-500 mM imidazole in buffer 1A .
C	Dialysis o.n. in buffer 1B .
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer 1B .
E	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer 1C .
F	NMR sample preparation in buffer 1C .

Table 4: Final sample

1	Yield
	3.5 mg/L ¹³ C, ¹⁵ N-M9 medium
2	A260/280 ratio
	0.74
3	Stability
	Stable throughout measurement (7 days, 298 K). No significant precipitation or degradation observed after storage at 4°C for 5 weeks.
4	Comment on applicability

Suitable for NMR structure determination, fragment screening, interaction studies.

Additional information

Constructs	Conditions	Comments
NAB (aa 1088-1203) of complete nsp3; His ₇ (pET-TEV-Nco (GenScript)), TEV-cleavage site, N-terminal "GAMG" four artificial residues.	As above.	Works as well, but slightly less expression and yield.

SI3: nsp3Y

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp3
2	Region/Name/Further Specification
	nsp3-Y / Cov-Y
3	Sequence of “fl” protein (aa 1638-1945 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)
	DTFCAGSTFISDEVARDLSLQFKRPINPTDQSSYIVDSVTVKNGSIHLYFDKAGQKTYERHLSLSHF VNLDNLRANNTKGSLPINVIVFDGKSKCEESSAKSASVYYSQLMCQPILLDDQALVSDVGDSAE VAVKMFDAYVNTFSSTFNVPMEKLTAVATAEAEAKNVSLDNVLSSTFISAARQGFVDSVET KDVVECLKLSHQSDIEVTGDCNNYMLTYNKVENMTPRDLGACIDCSARHINAQVAKSHNIAL IWNVKDFMSLSEQLRKQIRSAAKKNNLPFKLTATTRQVVNVVTTKIALKGG
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1638-1945 of complete nsp3
5	Ratio for construct design (detailed and comprehensible)
	We took the C-terminal part of nsp3 after predicted transmembrane region and Y1 domain that consists of two sequential zinc finger motifs.
6	Sequence homology (to SCoV)
	Identity: 89%; similarity: 96%
7	Published structures (SCoV2 or homologue variants)
	-
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Protein Expression

1	Expression vector
	pET28b(+) (GenScript)
2	Purification-/Solubility-Tag
	N-terminal His ₆
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	34 kDa / 17,420 M ⁻¹ cm ⁻¹ / 6.66
5	Comments on sequence of expressed construct

	N-terminal „G" one artificial residue due to TEV-cleavage.
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labeling)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.5 mM IPTG at OD ₆₀₀ 0.7
10	Cultivation temperature and time
	18°C for 15-16 h

Table 3: Protein Purification

1	Buffer List
A	20 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 mM imidazole, 0.1 mM PMSF, 5 mM bME, 0.1 mg/mL lysozyme, cOmplete EDTA-free inhibitor (Cell disruption).
B	20 mM Tris-HCl (pH 8.0), 300 mM NaCl, 20 mM imidazole (IMAC).
C	50 mM Tris-HCl (pH 8.0), 200 mM NaCl, 2 mM DTT (TEV-cleavage).
D	50 mM HEPES (pH 6.9), 200 mM LiBr, 5 mM DTT.
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A by sonication.
B	IMAC (gravity flow Ni ²⁺ -NTA) (Thermo Scientific), wash with buffer 1B and elution with 250 mM imidazole in buffer 1B .
C	TEV-cleavage (5% (w/w) TEV protease per approximate amount of the protein) in buffer 1C o.n. at rt.
D	Inv. IMAC (gravity flow Ni ²⁺ -NTA) in buffer 1C .
E	SEC on 10/300 GL SD 200 (GE Healthcare) in buffer 1D .

Table 4: Final sample

1	Yield
	12 mg/L ¹³ C, ¹⁵ N-M9 medium
2	Stability
	Stable at 25°C at protein concentration below 0.4 mM for 3 to 5 days or at 30°C o.n.. The protein gradually degrades at rt. After one week, we observe an additional band on SDS gel at ~27 kDa.
3	Comment on applicability

The protein is suitable for NMR assignment and protein interaction studies at low temperature (20-25°C) and reasonably low concentration (< 0.2 mM).

SI4: nsp5

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp5
2	Region/Name/Further Specification
	3C-like protease (3CL ^{pro}) / main protease (M ^{pro})
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDVVVYCPRHVICTSEDMLNPNYEDLLIRKSN HNFLVQAGNVQLRVIGHSMQNCVLKLVDTANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVY QCAMRPNFTIKGSFLNGSCGSGVGFNIDYDCVSFCYMHMELPTGVHAGTDLEGNFYGPVDRQ TAQAAGDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILG PLSAQTGIAVLDMCASLKELLQNGMNGRTILGSALLEDEFTPFDDVVRQCSGVTFQ
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-306 (fl nsp5)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 96%; similarity: 99.7%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 1P9U, 6LU7 SCoV2: PDB 6Y2E, 5R7Y, 6Y84, 7K3T
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: BMRB 17251

Table 2: Protein Expression

1	Expression vector
	pE-SUMO (LifeSensors)
2	Purification-/Solubility-Tag
	N-terminal His ₆ -SUMO
3	Cleavage Site
	Ulp1
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	33.80 kDa / 32,890 M ⁻¹ cm ⁻¹ / 5.95
5	Comments on sequence of expressed construct
	No artificial residues due to TEV-cleavage and construct design.

6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.2 mM IPTG at OD ₆₀₀ 0.6-0.7
10	Cultivation temperature and time
	18-20°C for 16-18 h

Table 3: Protein Purification

1	Buffer List
A	50 mM NaPi (pH 7.5), 300 mM NaCl, 5 mM imidazole, 5% (v/v) glycerol, 10 mM bME (cell disruption / IMAC).
B	50 mM NaPi (pH 7.0), 300 mM NaCl, 10 mM bME, 5% (v/v) glycerol (dialysis after IMAC / Ulp1-cleavage).
C	25 mM NaPi (pH 7.5), 150 mM NaCl, 2 mM TCEP-HCl (SEC buffer).
D	10 mM NaPi (pH 7.0), 0.5 mM TCEP-HCl (final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus one tablet of EDTA free protease inhibitor cocktail (Merck)) by microfluidization.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer 1A .
C	Ulp1-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer 1B .
D	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with 500 mM imidazole in buffer 1A .
E	SEC (HiLoad 26/600 SD 75 µg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer 1C (elution volume 170-210 mL).
F	NMR sample preparation in buffer 1D .

Table 4: Final sample

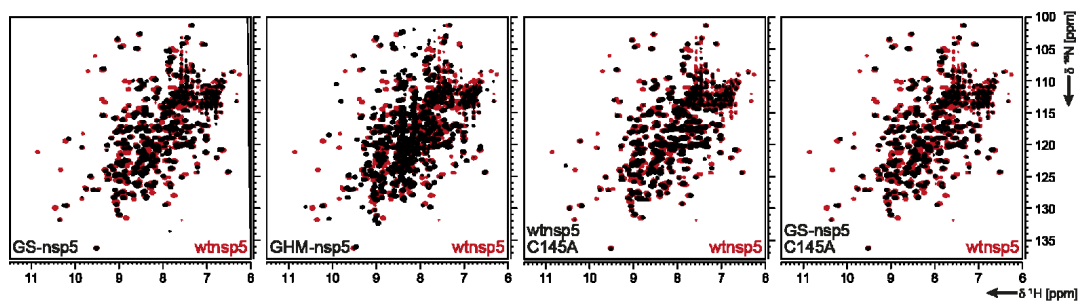
1	Yield
	55 mg/L ¹⁵ N-M9 medium
2	Stability
	No significant precipitation or degradation observed after storage at -80°C for a month.

3	Comment on applicability
	Suitable for NMR structure determination, fragment screening, interaction studies.

Additional information

	Constructs	Conditions	Comments
A	aa 1-306 (fl nsp5) C145A mutation; His ₆ -SUMO (pE-SUMO (LifeSensors)), Ulp1-cleavage site, no N-terminal artificial residues.	Native (as above)	Comparable to fl nsp5 expression and purification, similar yield (80 mg/L ¹⁵ N-M9 medium).
B	aa 1-306 (fl nsp5); His ₆ -SUMO (pE-SUMO (LifeSensors)), Ulp1-cleavage site, N-terminal "GS" two artificial residues.	Native (as above)	Comparable to fl nsp5 expression and purification, similar yield (55 mg/L ¹⁵ N-M9 medium, 36 mg/L ¹³ C, ¹⁵ N-M9 medium, 20 mg/L ² H, ¹³ C, ¹⁵ N E. coli-OD2 CDN medium (Silantes)).
C	aa 1-306 (fl nsp5) C145A mutation; His ₆ -SUMO (pE-SUMO (LifeSensors)), Ulp1-cleavage site, N-terminal "GS" two artificial residues.	Native (as above)	Comparable to fl nsp5 expression and purification, similar yield (55 mg/L ¹⁵ N-M9 medium).
D	aa 1-306 (fl nsp5); His ₆ (pet28a(+)) (GenScript), TEV-cleavage site; N-terminal "GHM" three artificial residues.	Native (as above) IMAC buffer (1A): 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 5% (v/v) glycerol, 10 mM bME	Comparable to fl nsp5 purification, however, less expression/yield (35 mg/L ¹⁵ N-M9 medium, 10 mg/L ¹³ C, ¹⁵ N-M9 medium).
E	aa 1-306 (fl nsp5); GST and His ₆ -tag (pET-28a(+)) (GenScript), TEV and auto cleavage site for M ^{pro} , N-terminal „GS“ and C-terminal "GPHHHHHH" ten artificial residues.	IMAC buffer: 50 mM Tris-HCl (pH 8.0), 200 mM NaCl, 20 mM imidazole. SEC-buffer: 50 mM NaPi (pH 7.6), 50 mM NaCl, 0.02% (w/v) NaN ₃ . NMR-buffer: 50 mM NaPi (pH 7.6), 50 mM NaCl, 0.02% (w/v) NaN ₃ , 5 mM bME.	Yields 20 mg/L ¹⁵ N-M9 medium. The protein is stable up to 350 μM in NMR buffer at 25°C for at least 7 days. At 50 μM and at 4°C, the protein is stable for ~15 days. The protein is not suitable for freeze/thaw and results in precipitation.
F	aa 1-306 (fl nsp5); C-terminal His ₆ -tag (pET21b(+)) (GenScript), human rhinovirus 3-C protease cleavage site, N-terminal "M" additional aa, however our mass spectrum results suggest that M1 was removed by <i>E. coli</i> methionine aminopeptidase.	IMAC buffer: 20 mM Tris-HCl (pH 7.33), 150 mM NaCl, 20 mM imidazole. Storage buffer: 20 mM Tris-HCl (pH 7.33), 150 mM NaCl.	Yields 5 mg/L ¹⁵ N-M9 medium. Stable for 2-3 weeks at 4°C at low micromolar concentration.
G	aa 1-306 (fl nsp5) C145A mutation; His ₆ -GB1 (pET24a(+)) (GenScript), TEV-cleavage site, no artificial residues.	IMAC buffer: 20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 20 mM imidazole, 0.5 mM TCEP-HCl. SEC/NMR buffer: 10 mM NaPi (pH 7.0), 0.5 mM TCEP-HCl.	Yields ≥ 70 mg/L ¹⁵ N, ² H, ¹⁵ N-M9, and ² H, ¹³ C, ¹⁵ N-M9 medium. 1-2 mM sample stable for several weeks at 25°C. Negligible precipitation on freeze-thaw. Samples stable for ≥ 3 months at 80°C. Sample precipitation in buffer: 10 mM NaPi (pH 7.0), 0.4 M GdnHCl.

H	aa 1-306 (fl nsp5); His ₆ -GB1 (pET24a+) (GenScript), TEV-cleavage site, no artificial residues.	As above (G).	Negligible expression when induced in ¹⁵ N-M9 medium at 25°C, 30°C, and 37°C, with 0.5-1 mM IPTG.
I	aa 1-306 (fl nsp5); His ₆ -GST (pGEX-6p-1 (Genewiz)), autolytic and HRV 3C cleavage site, no artificial residues.	IMAC buffer: 25 mM Tris-HCl (pH 7.8), 150 mM NaCl, 5 mM imidazole, 1 mM bME. Cleavage buffer: 25 mM Tris-HCl (pH 7.8), 150 mM NaCl, 1 mM DTT. SEC buffer: 25 mM Tris-HCl (pH 7.8), 150 mM NaCl, 1 mM DTT, 1 mM EDTA.	40-60 mg/mL autoinduction Media ZYM-5052. Stored at 1 mg/mL at -20°C with 30% v/v glycerol in SEC buffer. Also stored at 25 mg/mL at -80°C in SEC buffer. Flash frozen. Neither show loss of activity compared to non-frozen samples.



Overlays of [¹⁵N, ¹H]-BEST TROSY spectra of wt nsp5 (red) with the other constructs (black). From left to right: N-terminally GS added nsp5 (GS-nsp5), GHM added (GHM-nsp5), the active site mutants C145A with native N-terminus (wt nsp5 C145A), and GS added mutant (GS-nsp5 C145A).

SI5: nsp7

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp7
2	Region/Name/Further Specification
	nsp7
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	SKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCVQLHNDILLAKDTTEAFEKMSVLLSVLLSMQ GAVDINKLCEEMLDNRATLQ
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-83 (fl nsp7)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 98.8%; similarity: 100%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 2KYS, 1YSY, 6NUS, 6NUR, 2AHM, SCoV2: PDB 7BV2, 7BV1, 6YYT, 7BTF, 6WQD, 6WTC, 6WIQ, 6M71, 6YHU, 6XEZ, 6M5I, 7CTT, 7C2K, 7BW4, 7BZF, 7JLT, 7AAP, 6XIP, 6XQB
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: PDB 1YSY, BMRB 6513, PDB 2KYS, BMRB 16981 SCoV2: BMRB 50337

Table 2: Protein Expression

1	Expression vector
	pET46
2	Purification-/Solubility-Tag
	N-terminal His ₆ , enterokinase
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	9.24 kDa / 5500 cm ⁻¹ M ⁻¹ / 5.2
5	Comments on sequence of expressed construct
	N-terminal "G" an artificial residue due to TEV-cleavage.

6	Used expression strain
	<i>E. coli</i> Rosetta2 pLysS
7	Cultivation medium
	M9 (uniformly ¹⁵ N, ¹³ C-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.5 mM IPTG at OD ₆₀₀ 0.8
10	Cultivation temperature and time
	16°C for 14-16 h

Table 3: Protein Purification

1	Buffer List
A	10 mM HEPES (pH 7.4), 300 mM NaCl, 30 mM imidazole, 2 mM DTT.
B	10 mM HEPES (pH 7.4), 300 mM NaCl, 300 mM imidazole, 2 mM DTT.
C	10 mM MOPS (pH 7.0), 150 mM NaCl, 2 mM DTT.
2	Purification steps (with corresponding buffer(s) and incubation times)
	Cell lysis in buffer 1A by microfluidizer operating at 20,000 psi. Lysates were cleared by centrifugation at 25,000 g for 30 min and then filtered through a 0.45 µm filter.
A	Ni-NTA Agarose beads (Qiagen) were added to cleared lysates and incubated for 30 min. Beads were collected by centrifugation and then loaded onto a gravity column. Beads were washed twice with 10 column volumes of buffer 1A . Protein was eluted with 5 column volumes of buffer 1B .
B	Eluted protein was cleaved with 1% (w/w) TEV protease o.n. at rt while dialyzing the protein into 1 L buffer 1C . Uncleaved protein was removed by inv. Ni-NTA binding.
C	Protein was concentrated using a 10 kDa MWCO (Amicon) concentrator and purified on an SD 200 Increase 10/300 (GE Life Sciences) size exclusion column, AKTApure (GE Life Sciences) using buffer 1C .

Table 4: Final sample

1	Yield
	17 mg/L ¹³ C, ¹⁵ N-M9 medium
1b	A260/280 ratio
	0.5
2	Stability
	NMR sample stable at 4°C for a month, at 35°C for several days before degradation occurs.
3	Comment on applicability

Suitable for NMR-based screening applications.

SI6: nsp8

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp8
2	Region/Name/Further Specification
	nsp8
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	AIASEFSSLPSYAAFATAQEAYEQAVANGDSEVVLKCLKKSLNVAKSEFDRDAAMQRKLEKM ADQAMTQMYKQARSEDKRAKVTSAMQTMFLTMLRKLNDALNNIINNARDGCVPLNIIPLTT AAKLMVVIPDYNTYKNTCDGTTFTYASALWEIQVVDADSKIVQLSEISMDNSPNLAWPLIVT ALRANSAVKLQ
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-198 (fl nsp8)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 97%; similarity: 98%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 6NUS, 6NUR, 2AHM, SCoV2: PDB 7C2K, 7BV2, 7BV1, 7CTT, 6M5I, 7BW4, 6XEZ, 7BZF, 6XQB, 6M7I, 6YYT, 7BTF, 7JLT, 7AAP, 6WIQ, 6XIP, 6WQD, 6WTC, 6YHU
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Protein Expression

1	Expression vector
	pET46
2	Purification-/Solubility-Tag
	N-terminal His ₆ , enterokinase
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	21.94 kDa / 19,940 cm ² M ⁻¹ / 6.5
5	Comments on sequence of expressed construct
	N-terminal "G" an artificial residue due to TEV-cleavage.

6	Used expression strain
	<i>E. coli</i> Rosetta2 pLysS
7	Cultivation medium
	M9 (uniformly ¹⁵ N-, ¹³ C-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.5 mM IPTG at OD ₆₀₀ 0.8
10	Cultivation temperature and time
	16°C for 16-18 h

Table 3: Protein Purification

1	Buffer List
A	10 mM HEPES (pH 7.4), 300 mM NaCl, 30 mM imidazole, 2 mM DTT.
B	10 mM HEPES (pH 7.4), 300 mM NaCl, 300 mM imidazole, 2 mM DTT.
C	10 mM MOPS (pH 7.0), 300 mM NaCl, 2 mM DTT.
D	10 mM MOPS (pH 7.0), 150 mM NaCl, 2 mM DTT.
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell lysis in buffer 1A by microfluidizer operating at 20,000 psi. Lysates were cleared by centrifugation at 25,000 g for 30 min and then filtered through a 0.45 µm filter. Ni-NTA Agarose beads (Qiagen) were added to cleared lysates and incubated for 30 min. Beads were collected by centrifugation and then loaded onto a gravity column. Beads were washed twice with 10 column volumes of buffer 1A . Protein was eluted with 5 column volumes of buffer 1B .
B	Eluted protein was cleaved with 1% (w/w) TEV protease o.n. at rt while dialyzing the protein into 1 L buffer 1C . Uncleaved protein was removed by inverse Ni-NTA binding.
C	Protein was concentrated using a 10 kDa MWCO (Amicon) concentrator and purified on an SD 200 Increase 10/300 (GE Life Sciences) size exclusion column, AKTApure (GE Life Sciences) using buffer 1D .

Table 4: Final sample

1	Yield
	17 mg/L ¹³ C, ¹⁵ N-M9 medium
1b	A260/280 ratio
	0.5
2	Stability
	Concentration dependent aggregation of nsp8 observed in the range of 0.1-1.1 mM by NMR.

3	Comment on applicability
	Suitable for NMR-based screening approach.

SI7: nsp9

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp9
2	Region/Name/Further Specification
	nsp9
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	NNELSPVALRQMSCAAGTTQTACTDDNALAYNTTKGGRFVLALLSDLQDLKWARFPKSDGT GTIYTELEPPCRFVTDTPKGPKVKYLYFIKGLNNLNRGMVGLSLAATVRLQ
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-113 (fl nsp9)
5	Ratio for construct design (detailed and comprehensible)
	In analogy to the available crystal structure (PDB 1QZ8) of nsp9 SCoV, fl sequence.
6	Sequence homology (to SCoV)
	Identity: 97%; similarity: 97%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 3EE7 (G104E), 1UW7, 1QZ8 SCoV2: PDB 6WXd, 6W4B, 6W9Q
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: BMRB 6501 SCoV2: BMRB 50621, 50622

Table 2: Protein Expression

1	Expression vector
	pKM263 (GenScript)
2	Purification-/Solubility-Tag
	N-terminal His ₆ -GST
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	12,7 kDa / 13,075 M ⁻¹ cm ⁻¹ / 9.1
5	Comments on sequence of expressed construct
	N-terminal „GAMG" four artificial residues due to TEV-cleavage and construct design
6	Used expression strain

	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	1 mM IPTG at OD ₆₀₀ 0.7
10	Cultivation temperature and time
	20-22°C for 18-20 h

Table 3: Protein Purification

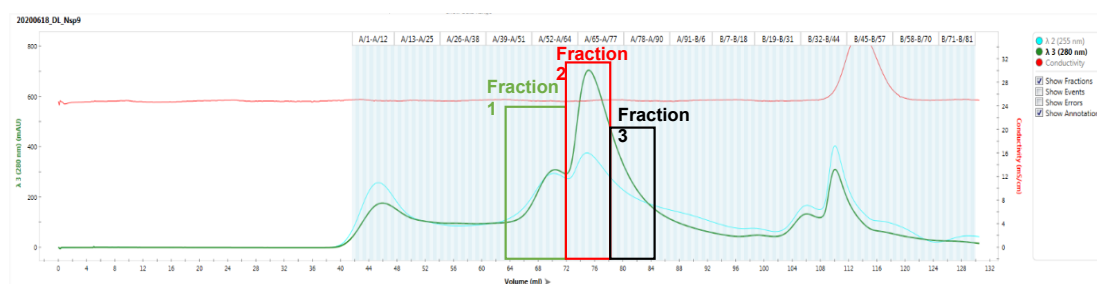
1	Buffer List
A	50 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 mM imidazole, 4 mM DTT (cell disruption / IMAC/ dialysis after IMAC / TEV-cleavage).
B	25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl, 0.02% (w/v) NaN ₃ (SEC / final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus 100 µL protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni ²⁺ -NTA (Carl Roth)), Elution with 150-500 mM imidazole in buffer 1A .
C	Dialysis o.n. in in buffer 1A .
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer 1A .
E	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer 1B . See relevant peak in attached SEC profile.
F	NMR sample preparation in buffer 1B .

Table 4: Final sample

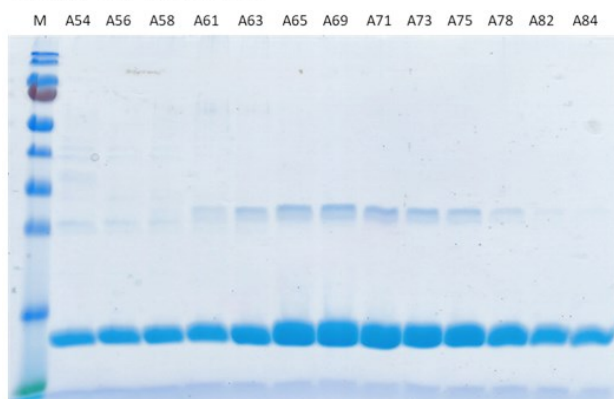
1	Yield
	4.5 mg/L ¹³ C, ¹⁵ N-M9 medium
1b	A260/280 ratio
	0.7
2	Stability
	Stable dimer. Storage at 4°C possible.
3	Comment on applicability
	Conditions for NMR structure determination may need to be optimized (concerning line width due to dimeric state). Backbone assignment and screening successful.

Additional information

	Constructs	Conditions	Comments
A	aa 1-113 (fl nsp9); His ₇ (pET-TEV-Nco (GenScript)), TEV-cleavage site, N-terminal "GAMG" four artificial residues.	As above.	Expression and purification as for GST-tagged fl nsp9, but lower expression and yield.
B		<p>IMAC buffer: 25 mM NaPi (pH 7.4), 300 mM NaCl, 20 mM imidazole, 1 mM DTT.</p> <p>Cleavage buffer: 25 mM NaPi (pH 7.4), 150 mM NaCl, 1 mM DTT.</p> <p>SEC/NMR buffer A: 25 mM NaPi (pH 7.0), 150 mM NaCl, 1 mM DTT, 150 mM NaCl, 2 mM TCEP-HCl.</p> <p>SEC/NMR buffer B: 25 mM NaAc (pH 5.0), 150 mM NaCl, 2 mM TCEP-HCl.</p>	<p>3 mg/L ¹³C, ¹⁵N-M9 medium. Sample in Buffer A looked degraded (from the ¹⁵N HSQC) after 5 days of ¹³C 3D NMR experiments at 298 K. Less degradation was observed for sample in Buffer B after same period. Suitable for NMR studies, fragment-based screening, interaction studies.</p>



H6-GST-TEV-Nsp9 (BL21)



SEC profile of TEV-cleaved His₆-GST-fl nsp9 (HiLoad 16/600 SD 75, GE Healthcare) and SDS gel of corresponding fractions. (Ladder: PageRuler™ prestained, Thermo Fischer)
 Main peak (fraction 2 - corresponding to SEC fractions A 61 to A73) was subsequently used for NMR.

SI8: nsp10

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1a and ORF1ab; nsp10
2	Region/Name/Further Specification
	nsp10
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	AGNATEVPANSTVLSFCFAVDAAKAYKDYLASGGQPITNCVKMLCTHTGTGQAITVTPEAN MDQESFGGASCCLYCRCHIDHPNPKGFCDLKGKYVQIPTTCANDPVGFTLKNTVCTVCGMWK GYGCSCDQLREPLQ
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-139 (fl nsp10)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 97%; similarity: 99%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 5C8S, 5NFY, 2FYG, 2XYQ, 2XYV, 2XYW SCoV2: PDB 6W4H, 6W61, 7JYY, 7C2I, 7BQ7, 2G9T
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV2: BMRB 50392

Table 2: Protein Expression

1	Expression vector
	pET21b(+) (GenScript)
2	Purification-/Solubility-Tag
	N-terminal His ₆
3	Cleavage Site
	-
4	Molecular weight / Extinction coefficient / pI - of protein
	16.24 kDa / 12,950 M ⁻¹ cm ⁻¹ / 6.72
5	Comments on sequence of expressed construct
	N-terminal "MGSDKIHSHHHH" twelve artificial residues due to construct design
6	Used expression strain

	<i>E. coli</i> T7 Express
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.5 mM IPTG at OD ₆₀₀ 0.6-0.7 (addition of 50 μM ZnCl ₂)
10	Cultivation temperature and time
	18-20°C for 16-18 h

Table 3: Protein Purification

1	Buffer List
A	25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME (cell disruption / IMAC)
B	50 mM NaPi (pH 7.5), 50 mM NaCl, 5 mM DTT (SEC / final NMR buffer)
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus one tablet of EDTA free protease inhibitor cocktail (Merck) and addition of 50 μM ZnCl ₂) by microfluidization.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer 1A .
C	SEC (HiLoad 26/600 SD 75 μg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer 1B (elution volume 175-225 mL).
D	NMR sample preparation in buffer 1B .

Table 4: Final sample

1	Yield
	25 mg/L ¹⁵ N-M9 medium, 15 mg/L ¹³ C, ¹⁵ N-M9 medium
2	Stability
	Stable throughout measurement (6 days, 298 K). No significant precipitation or degradation observed after storage at -80°C for 2 months.
3	Comment on applicability
	Suitable for NMR structure determination, fragment screening, interaction studies.

Additional information

Constructs	Conditions	Comments
aa 1-139 (fl nsp10); His ₆ (pMCSG53 (BEI Resources, cat.	IMAC-buffer: 50 mM Tris-HCl (pH 9.0), 0.5 M NaCl, 10 mM bME,	Yields 30-40 mg/L 2xTY medium. Can be flash-frozen in liquid

NR-52425)), TEV cleavage site, N-terminal "SNM" three artificial residues.	2 mM MgCl ₂ , 0.1% (v/v) Triton X-100, 5-10% (v/v) glycerol, 50 mM imidazole. SEC-buffer: 20 mM HEPES (pH 8.5), 0.5 M NaCl, 10 mM bME, 2 mM MgCl ₂ , 5% (v/v) glycerol, 20 mM imidazole.	nitrogen and stored at 20°C, used for nsp14 and nsp16 stabilization at 1:1 molar ratios.
--	--	--

SI9: nsp13

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1ab; nsp13
2	Region/Name/Further Specification
	NTPase / helicase domain / RNA 5'-triphosphatase
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	AVGACVLCNSQTSLRCGACIRRPFLCCKCCYDHVISTSHKLVLSVNPYVCNAPGCDVTDVTQL YLGGMSSYYCKSHKPPISFPLCANGQVFLYKNTCVGSDNVTDFNAIATCDWTNAGDYILANTC TERLKLFAAETLKATEETFKLSYGIATVREVLSRELHLSWEVKGKPRPPLNRNYVFTGYRVTKN SKVQIGEYTFEKGDYGDVYRGTTTYKLVNGDYFVLTSHVTMPLSAPTLVPQEHYVRITGLY PTLNISDEFSSNVANYQKVGMMQKYSTLQGGPGTGKSHFAIGLALYYPSARIVYTACSHAARDAL CEKALKYLPIDKCSRIIPARARVECFDKFKVNSTLEQYVFTVNALPETTADIVVFDEISMATNY DLSVFNARLRAKHVYIGDPAQLPAPRTLLTKGTLEPEYFNSVCRMKTIGPDMFLGTCRRCPA EIVDTVSALVYDNKLAHKDKSAQCFKMFYKGVITHDVSSAINRPQIGVVREFLTRNPAWRKA VFISPYNSQNAVASKILGLPTQTVDSSQSGSEYDYVIFTQTTETAHSCNVNRFNVAITRAKVGILCI MSDRDLYDKLQFTSLEIPRRNVATLQ
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	1-601 aa (fl nsp13)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 99.8%; similarity: 100%
7	Published structures (SCoV2 or homologue variants)
	SCoV2: PDB 6ZSL, 6JYT, 6XEZ
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Protein Expression

1	Expression vector
	pE-SUMO (LifeSensors)
2	Purification-/Solubility-Tag
	N-terminal His ₆ -SUMO
3	Cleavage Site
	Ulp1
4	Molecular weight / Extinction coefficient / pI - of cleaved protein

	66.85 kDa / 67,160 M ⁻¹ cm ⁻¹ / 8.66
5	Comments on sequence of expressed construct
	No artificial residues due to Ulp1-cleavage and construct design.
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.2 mM IPTG at OD ₆₀₀ 0.6-0.7 (addition of 50 μM ZnCl ₂)
10	Cultivation temperature and time
	18-20°C for 16-18 h

Table 3: Protein Purification

1	Buffer List
A	25 mM Tris (pH 8.0), 300 mM NaCl, 5 mM imidazole, 5% (v/v) glycerol, 10 mM bME (cell disruption / IMAC).
B	20 mM BisTris (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl (SEC/ final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus one tablet of EDTA free protease inhibitor cocktail (Merck) and addition of 50 μM ZnCl ₂) by microfluidization.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer 1A .
C	SEC (HiLoad 26/600 SD 200 pg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer 1B (elution volume 210-240 mL).
D	NMR sample preparation in buffer 1B .

Table 4: Final sample

1	Yield
	0.5 mg/L ¹⁵ N-M9 medium
2	Stability
	Aggregation at > 20 μM under these conditions.
3	Comment on applicability
	Not suitable for NMR experiments.

Additional information

Constructs	Conditions	Comments
aa 1-601 (fl nsp13); His ₆ (pET28a+) (GenScript), TEV-cleavage site, N-terminal "GHM" three artificial residues.	Native (as above)	Weak expression, instable protein.

SI10: nsp14

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1ab; nsp14
2	Region/Name/Further Specification
	nsp14 / 3'-to-5' exonuclease / guanine N7-methyltransferase (MTase)
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	AENVTLGLFKDCSKVITGLHPTQAPTHLSVDTKFKTEGLCVDIPGIPKDMTYRRLISMMGFKMNY QVNGYPNMFITREEAIRHVRAWIGFDVEGCHATREAVGTNLPLQLGFSTGVNLVAVPTGYVDT PNNTDFSRVSAKPPPG
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
fl	aa 1-527 (fl nsp14)
MTase	aa 288-527 (MTase domain)
5	Ratio for construct design
fl	fl protein
MTase	In analogy to SCoV structure (PDB 5C8U)
6	Sequence homology (to SCoV)
fl	Identity: 95%; similarity: 99%
MTase	Identity: 95%, similarity: 97%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 5C8U, 5C8S, 5C8T, 5NFY
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Protein Expression

1	Expression vector
fl	pRSF-Duet1 (Novagen)
MTase	pET28a (Novagen)
2	Purification-/Solubility-Tag
fl	N-terminal His ₆
MTase	N-terminal His ₆
3	Cleavage Site
fl	TEV

MTase	Thrombin
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
fl	60.01 kDa / 91,660 M ⁻¹ cm ⁻¹ / 7.79
MTase	27.82 kDa / 48,970 M ⁻¹ cm ⁻¹ / 7.19
5	Comments on sequence of expressed construct
fl	N-terminal "GSM" three artificial residues due to construct design.
MTase	N-terminal "GSHM" four artificial residues due to construct design.
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	2xTY for protein production, LB for transformation and maintenance
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	1 mM IPTG at OD ₅₄₀ 0.5-0.6
10	Cultivation temperature and time
	20°C for 18-20 h

Table 3: Protein Purification (fl nsp14 and nsp14 MTase)

1	Buffer List
A	50 mM Tris-HCl (pH 9.0), 0.5 M NaCl, 10 mM bME, 2 mM MgCl ₂ , 0.1% (v/v) Triton X-100, 10% (v/v) glycerol, 50 mM imidazole (cell disruption).
B	50 mM Tris-HCl (pH 9.0), 0.5 M NaCl, 10 mM bME, 2 mM MgCl ₂ , 5% (v/v) glycerol, 50 mM imidazole (IMAC).
C	50 mM Tris-HCl (pH 9.0), 0.5 M NaCl, 10 mM bME, 2 mM MgCl ₂ , 5% (v/v) glycerol, 1 M imidazole (IMAC).
D	20 mM HEPES (pH 8.5), 0.5 M NaCl, 10 mM bME, 2 mM MgCl ₂ , 5% (v/v) glycerol, 20 mM imidazole (SEC).
E	20 mM potassium phosphate (pH 8.0), 0.25 M KCl (Screening).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A by sonication in pulse mode (0.5 s on /0.5 s off) for 10 min.
B	IMAC (gravity flow or batch Ni ²⁺ -NTA) (GE Healthcare), washing with buffer 1B , elution with 1C .
C-fl	[Optional] Overnight incubation with TEV protease at 4°C. The ratio was 1 mg of TEV protease per 20-40 mg of nsp14 protein.

C- MTase	[Optional] Overnight incubation with thrombin protease at 4°C. The ratio was 1-2 U of thrombin protease per 3-4 mg of MTase nsp14 protein.
D	SEC on SD 200 16/600 column (GE Healthcare) in buffer 1D (elution volume 75-95 mL).
E-fl	[Optional] Separation of TEV protease and uncleaved nsp14 material with IMAC, collection of flow through in buffer 1D .
E- MTase	[Optional] Separation of thrombin protease and uncleaved MTase nsp14 material with IMAC, collection of flow through in buffer 1D .
F	For fragment screening the buffer is exchanged to 1E .
G	[Optional] If higher concentrations or increased stability of fl nsp14 is desired, nsp10 should be added at 1:1 molar ratio.

Table 4: Final sample

1	Yield
fl	6 mg/L 2xTY medium
MTase	~ 10 mg/L 2xTY medium
1b	A260/280 ratio
fl	0.6
MTase	0.6
2	Stability
fl	The fl nsp14 construct tends to be unstable at concentrations above 3 mg/mL without reducing agent (TCEP-HCl or bME). Unstable at 4°C longer than one week. Freezing is not advisable; storage in 50% (v/v) glycerol at -20°C is preferable.
MTase	The MTase construct is even more unstable, and requires the presence of reducing agent (TCEP-HCl or bME) and NaCl at least in 400 mM concentration.
3	Comment on applicability
	Suitable for fragment screening and enzymatic activity assays.

Additional information

Constructs	Conditions	Comments
Fl nsp14; His ₆ (pETDuet (GenScript)), no cleavage site, N-terminal "MGSSHHHHHSQDP" 14 artificial residues.	IMAC-buffer: 25 mM Tris/HCl (pH 8.5), 300 mM NaCl, 5 mM imidazole, 10 mM bME, 5% (v/v) glycerol. SEC-buffer: 25 mM Tris/HCl (pH 8.5), 300 mM NaCl, 5 mM DTT, 5% (v/v) glycerol	Yields 14 mg/L ¹⁵ N-M9 medium. Tendency to aggregate.

SI11: nsp15

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1ab; nsp15
2	Region/Name/Further Specification
	nsp15 / NendoU / Endonuclease
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	SLENVAFNVVNKGHFDGQQGEVPSIINNTVYTKVDGVDVELFENKTTLPVNVAFELWAKRNI KPVPEVKILNNLGVDIAANTVIWDYKRDAPAHISTIGVCSMTDIAKKPTETICAPLTVFFDGRVD GQVDLFRNARNGVLITEGSVKGLQPSVGPQASLNGVTLIGEAVKTQFNYYKKVDGVVQQLPE TYFTQSRNLQEFKPRSQMEIDFLELAMDEFIERYKLEGYAFEHIVYGDFSHSQLGGLHLLIGLAK RFKESPFLEDFIPMDSTVKNYFITDAQTGSSKCVCSVIDLLDDDFVEIISQDLSVVSQVVKVTI DYTEISFMLWCKDGHVETFPKQLQ
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-346 (fl nsp15)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 89%; similarity: 98%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 2H85 SCoV2: PDB 6W01
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Protein Expression

1	Expression vector
	pET28a(+) (GenScript)
2	Purification-/Solubility-Tag
	N-terminal His ₆
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	39.14 kDa / 32,890 M ⁻¹ cm ⁻¹ / 5.12
5	Comments on sequence of expressed construct

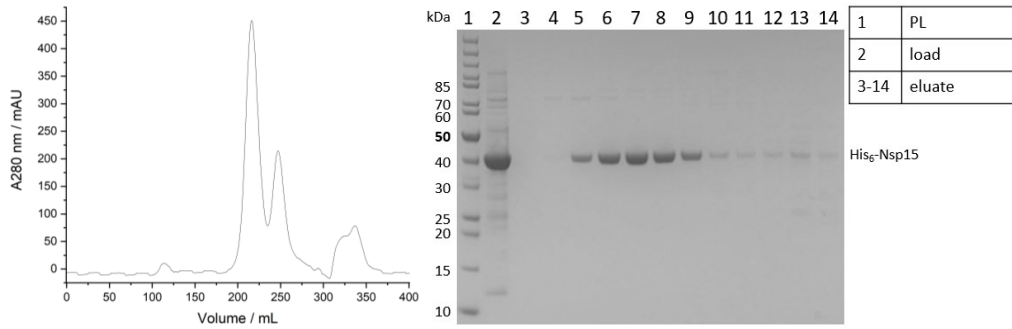
	N-terminal “GHM” three artificial residues due to TEV-cleavage and construct design
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.2 mM IPTG at OD ₆₀₀ 0.6-0.7
10	Cultivation temperature and time
	18-20°C for 16-18 h

Table 3: Protein Purification

1	Buffer List
A	25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 5% (v/v) glycerol, 10 mM bME (cell disruption / IMAC).
B	25 mM NaPi (pH 7.5), 300 mM NaCl, 2 mM TCEP-HCl (SEC/ final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus one tablet of EDTA free protease inhibitor cocktail (Merck)) by microfluidization.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer 1A .
C	SEC (HiLoad 26/600 SD 200 pg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer 1B (elution volume 200-260 mL).
D	NMR sample preparation in buffer 1B .

Table 4: Final sample

1	Yield
	5 mg/L ¹⁵ N-M9 medium
2	Stability
	Tendency to aggregate at rt.
3	Comment on applicability
	Suitable for fragment screening and interaction studies.



Analytical SEC of nsp15. Protein was eluted from 200-260 mL (left panel) with corresponding SDS-PAGE of SEC with fractions analyzed from 190-260 mL (right panel).

SI12: nsp16

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF1ab; nsp16
2	Region/Name/Further Specification
	nsp16 / 2'-O-ribose methyltransferase (2'-O-MTase)
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	SSQAWQPGVAMPNLYKMQRMLLEKCDLQNYGDSATLPKGIMMNVAKYTQLCQYLNTLTLAV PYNMRVIHFGAGSDKGVAPGTAVLRQWLPTGTLLVDSLNDVSDADSTLIGDCATVHTANK WDLIISDMYDPKTKNVTKENDSKEGFFTYICGFIQQKLALGGSVAIKITEHSWNADLYKLMGHF AWWTAFVTNVNASSEAFILGCNYLKGPREQIDGYVMHANYIFWRNTNPIQLSSYSLFDMMSKFP LKLRGTAVMSLKEGQINDMILSLLSKGRLLIRENNRVVISSDVLVNN
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-298 (fl nsp16)
5	Ratio for construct design
	Based on fl annotation boundaries of YP_009725311.1 protein entry in NC_045512.2.
6	Sequence homology (to SCoV)
	Identity: 93%; similarity: 99%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 3R24, 2XYR, 2XYQ SCoV2: PDB 7JYY, 6W4H, 6YZ1, 7BQ7, 7C2I, 6W61
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Protein Expression

1	Expression vector
	pRSF-Duet1 (Novagen)
2	Purification-/Solubility-Tag
	N-terminal His ₆
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	33.67 kDa / 55,790 M ⁻¹ cm ⁻¹ / 7.76
5	Comments on sequence of expressed construct
	N-terminal „GSMA" - four artificial residues due to TEV-cleavage and construct design.

6	Used expression strain
	<i>E. coli</i> BL21(DE3)
7	Cultivation medium
	2xTY
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	1 mM IPTG at OD ₅₄₀ 0.5-0.6
10	Cultivation temperature and time
	20°C for 18-20 h

Table 3: Protein Purification

1	Buffer List
A	50 mM Tris-HCl (pH 9.0), 500 mM NaCl, 10 mM bME, 2 mM MgCl ₂ , 0.1% (v/v) Triton X-100, 10% (v/v) glycerol, 50 mM imidazole (cell disruption).
B	50 mM Tris-HCl (pH 9.0), 500 mM NaCl, 10 mM bME, 2 mM MgCl ₂ , 5% (v/v) glycerol, 50 mM imidazole (IMAC).
C	50 mM Tris-HCl (pH 9.0), 500 mM NaCl, 10 mM bME, 2 mM MgCl ₂ , 5% (v/v) glycerol, 1 M imidazole (IMAC).
D	20 mM HEPES (pH 8.5), 500 mM NaCl, 10 mM bME, 2 mM MgCl ₂ , 5% (v/v) glycerol, 20 mM imidazole (SEC).
E	20 mM KPi (pH 8.0), 200 mM KCl, 1 mM MgCl ₂ , 2 mM DTT (Screening).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A by sonication in pulse mode (0.5 s on /0.5 s off) for 10 min.
B	IMAC (gravity flow or batch Ni ²⁺ -NTA) (GE Healthcare), washing with buffer 1B , elution with 1C .
C	[Optional] Overnight incubation with TEV protease at 4°C. The ratio was 1 mg of TEV protease per 20-40 mg of nsp16 protein.
D	SEC on SD 200 16/600 column (GE Healthcare) in buffer 1D (elution volume 90-100 mL).
E	[Optional] Separation of TEV protease and uncleaved nsp16 material with IMAC, collection of flow through in buffer 1D .
F	nsp10 is added at 1:1 molar ratio – necessary for stability and activity.
G	For fragment screening the buffer is exchanged to 1E .

Table 4: Final sample

1	Yield
	~ 10-15 mg/L 2xTY medium.

1b	A260/280 ratio
	0.55
2	Stability
	Extremely unstable in non-reducing conditions, presence of reducing agents is essential. Presence of 5% (v/v) glycerol is also desirable for increased stability. Can be flash-frozen in liquid nitrogen and stored at -20°C.
3	Comment on applicability
	Suitable for fragment screening.

Additional information

Constructs	Conditions	Comments
F1 nsp16; His ₆ (pMCSG53 (BEI Resources, cat. NR-52427)), TEV-cleavage site, N-terminal „SNM” three artificial residues.	As above	~ 5 mg/L 2xTY medium). Purity and stability is comparable to the “GSMA” construct above.

SI13: ORF3a

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF3a
2	Region/Name/Further Specification
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFQSASKIITLK KRWQLALS KGVHFCNLLLLFVTVYSHLLLVAAGLEAPFLYLYALVYFLQSIN FVRIIMRLWLC WKCRSKNPLLYDANYFLCWHTNCYDYCIPYNSVTSSIVITSGDGTTSPISEHDYQIGGYTEKWE SGVKDCVVLHSYFTSDYYQLYSTQLSTDTGVEHVTFEYFNKIVDEPEEHVQIHTIDGSSGVVNPV MEPIYDEPTTTTSVPL
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-275 (fl ORF3a)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 72.4%; similarity: 90.2%
7	Published structures (SCoV2 or homologue variants)
	SCoV2: PDB 6XDC
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Cell-free Protein Synthesis

1	Expression vector
	pEU-E01-MCS (Cell-Free Sciences)
2	Purification-/Solubility-Tag
	C-terminal Strep tag II (WSHPQFEK)
3	Cleavage Site
	-
4	Molecular weight / Extinction coefficient / pI - of protein
	32.32 kDa / 64,205 M ⁻¹ cm ⁻¹ / 5.67
5	Comments on sequence of expressed construct
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.

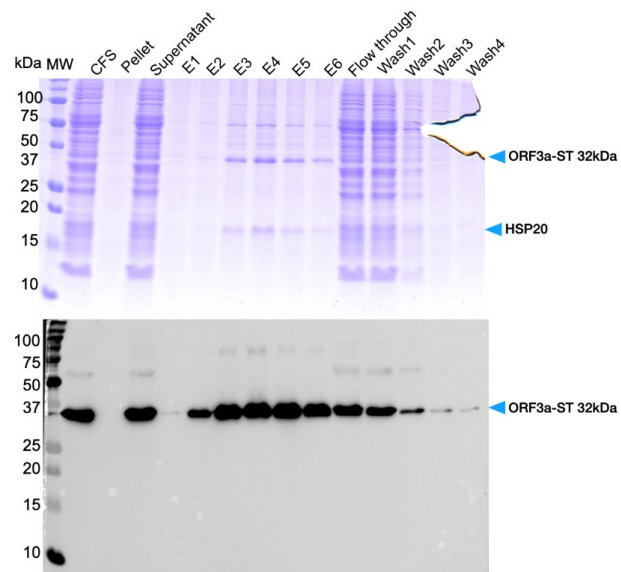
6	Feeding buffer
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix and 0.05% (w/v) Brij-58.
7	Translation mix
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix 0.05% (w/v) Brij-58.
8	Protein synthesis temperature and time
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

1	Buffer List
A	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 0.1% (w/v) DDM (wash buffer).
B	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 2.5 mM desthiobiotin, and 0.1% (w/v) DDM (elution buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Harvest total CFS.
B	Incubate with benzonase for 30 min on a wheel, at rt.
C	Centrifuge for 30 min at 20,000 g, 4°C.
D	Harvest the soluble fraction (SN).
E	Equilibrate the Strep-Tactin column (IBA Lifesciences) with 2 CV of 1A (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of 1A .
H	Elute the protein of interest with 1B .

Table 4: Final sample

1	Yield
	0.6 mg/mL WGE
1b	A260/280 ratio
	1.08
2	Stability
	Stable at 4°C for at least 2 weeks.
3	Comment on applicability
	ORF3a-ST is eluted with small heat shock protein (SHSP, 18 kDa) from wheat.



WG-CFPS in the presence of detergent, and Strep-tag purification of ORF3a. SDS-PAGE (upper panel) and WB (lower panel).

SI14: ORF4 (Envelope (E) protein)

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF 4; Envelope (E) protein
2	Region/Name/Further Specification
	E protein
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MYSFVSEETGTLIVNSVLLFLAFVVLLVTLAILTALRLCAYCCNIVNVS LVKPSFYVYSRVKNL NSSRVPDLLV
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-75 (fl ORF4)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 94.7%; similarity: 97.4%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 5X29
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: BMRB 36049

Table 2: Cell-free Protein Synthesis

1	Expression vector
	pEU-E01-MCS (Cell-free Sciences)
2	Purification-/Solubility-Tag
	C-terminal Strep tag II (WSHPQFEK)
3	Cleavage Site
	-
4	Molecular weight / Extinction coefficient / pI - of protein
	9.56 kDa / 11,460 M ⁻¹ cm ⁻¹ / 8.55
5	Comments on sequence of expressed construct
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.
6	Feeding buffer

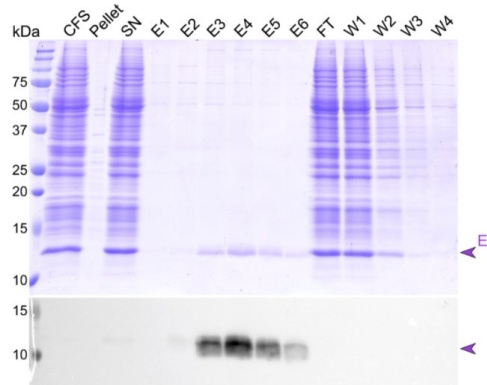
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix and 0.05% (w/v) Brij-58.
7	Translation mix
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix 0.05% (w/v) Brij-58.
8	Protein synthesis temperature and time
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

1	Buffer List
A	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 0.1% (w/v) DDM (wash buffer).
B	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 2.5 mM desthiobiotin, and 0.1% (w/v) DDM (elution buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Harvest total CFS.
B	Incubate with benzonase for 30 min on a wheel, at rt.
C	Centrifuge for 30 min at 20,000 g, 4°C.
D	Harvest the soluble fraction (SN).
E	Equilibrate the Strep-Tactin column (IBA Lifesciences) with 2 CV of 1A (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of 1A .
H	Elute the protein of interest with 1B .

Table 4: Final sample

1	Yield
	0.45 mg/mL WGE
1b	A260/280 ratio
	1.52
2	Stability
	Stable at least a few days at rt.
3	Comment on applicability
	E protein cannot be sedimented and is thus not directly available for solid-state NMR. Lipid reconstitution will be needed.



WG-CFPS in the presence of detergent, and Strep-tag purification of E (ORF4). SDS-PAGE (upper panel) and WB (lower panel).

SI15: ORF5 (M protein)

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF5; Membrane glycoprotein (M)
2	Region/Name/Further Specification
	M protein
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MADSNGTITVEELKKLLEQWNLVIGFLFTWICLLQFAYANRNRFLYIIKLIFLWLLWPVTLACF VLAAYRINWITGGIAIAMAACLVGLMWLSYFIASFRLFARTRSMWSFNPETNILLNVPLHGTILT RPLLESELVIGAVILRGHLRIAGHHLGRCDIKDLPKEITVATSRTLSYYKLGASQRVAGDSGFAA YSRYRIGNYKLNTDHSSSDNIALLVQ
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-222 (fl ORF5)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 90.5%; similarity: 98.2%
7	Published structures (SCoV2 or homologue variants)
	-
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Cell-free Protein Synthesis

1	Expression vector
	pEU-E01-MCS (Cell-Free Sciences)
2	Purification-/Solubility-Tag
	C-terminal Strep tag II (WSHPQFEK)
3	Cleavage Site
	-
4	Molecular weight / Extinction coefficient / pI - of protein
	26.35 kDa / 57,660 M ⁻¹ cm ⁻¹ / 9.48
5	Comments on sequence of expressed construct
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.

6	Feeding buffer
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT, 6 mM (average concentration) amino acid mix, and 0.05% (w/v) Brij-58.
7	Translation mix
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, 6 mM (average concentration), and amino acid mix 0.05% (w/v) Brij-58.
8	Protein synthesis temperature and time
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

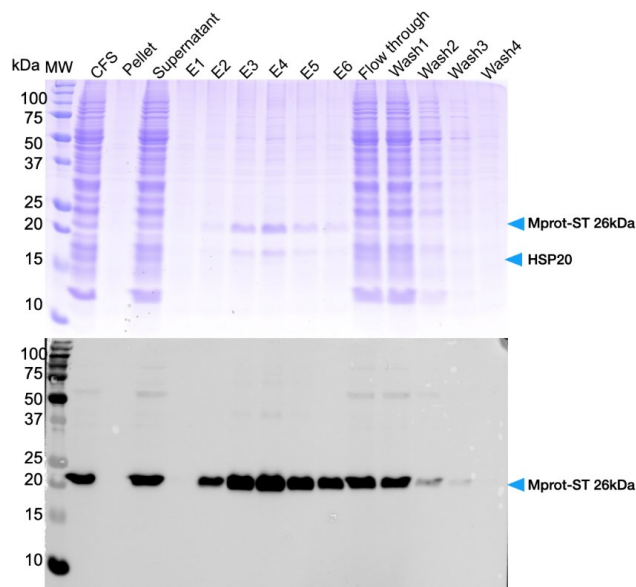
1	Buffer List
A	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 0.1% (w/v) DDM (wash buffer).
B	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 2.5 mM desthiobiotin, and 0.1% (w/v) DDM (elution buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Harvest total CFS.
B	Incubate with benzonase for 30 min on a wheel, at rt.
C	Centrifuge for 30 min at 20,000 g, 4°C.
D	Harvest the soluble fraction (SN).
E	Equilibrate the Strep-Tactin column (IBA Lifesciences) with 2 CV of 1A (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of 1A .
H	Elute the protein of interest with 1B .

Table 4: Final sample

1	Yield
	0.33 mg/mL WGE
1b	A260/280 ratio
	1.16
2	Stability
	Stable at 4°C for at least 2 weeks.
3	Comment on applicability
	Mprotein-ST (and ST-Mprot) is eluted with small heat shock protein (SHSP 18 kDa) from wheat.

Additional information

Constructs	Conditions	Comments
F1 ORF5; Strep tag II (pEU-E01-MCS (Cell-Free Sciences)); no cleavage site; N-terminal "WSHPQFEK" eight artificial residues.	As above, but: - Purification: 1 mM DTT was added in purification buffers 1A and 1B. - Tab. 3.2B: 0.25% (w/v) DDM is added and incubated on the wheel for 1 h. - Tab. 3.2C: 40,000 g for 40 min. - Tab. 3.2E: added Strep beads for batch purification (200 μ L 50% (w/v) suspension per well) and incubated on the wheel for 1.5 h.	Works as well with similar yield (0.39 mg/mL) and purity.



WG-CFPS in the presence of detergent, and Strep-tag purification of M (ORF5). SDS-PAGE (upper panel) and WB (lower panel).

SI16: ORF6

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF6
2	Region/Name/Further Specification
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLIKNLSKSLTENKYSQLDEEQPMEID
4	Protein boundaries - amino acid numbering (according to NCBI Reference Sequence NC_045512.2):
	aa 1-61 (fl ORF6)
5	Ratio for construct design (detailed and comprehensible)
	fl protein
6	Sequence homology (to SCoV)
	Identity: 68.9%; similarity: 93.4%
7	Published structures (SCoV2 or homologue variants)
	-
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Cell-free Protein Synthesis

1	Expression vector
	pEU-E01-MCS (Cell-Free Sciences)
2	Purification-/Solubility-Tag
	C-terminal Strep tag II (WSHPQFEK)
3	Cleavage Site
	-
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	8470.85 kDa / 13,980 M ⁻¹ cm ⁻¹ / 4.89
5	Comments on sequence of expressed construct
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.
6	Feeding buffer

	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix
7	Translation mix
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix
8	Protein synthesis temperature and time
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

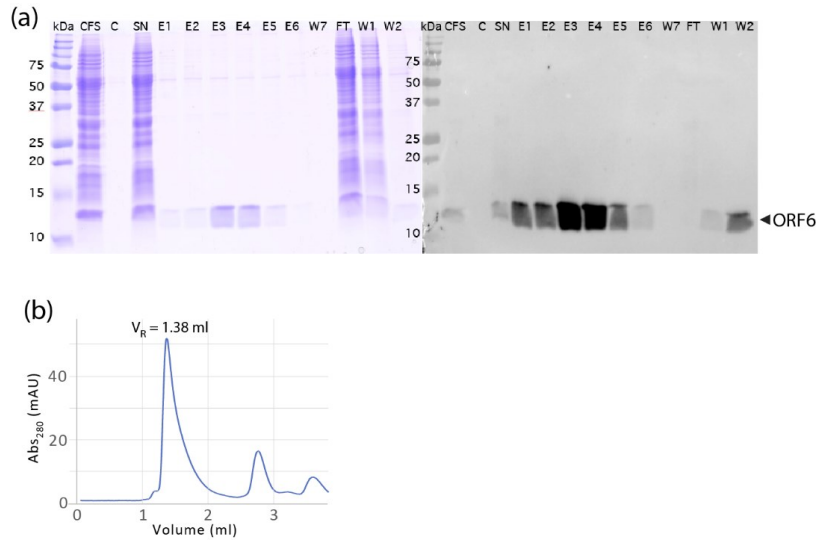
1	Buffer List
A	20 mM NaPi (pH 6.5), 50 mM NaCl (wash buffer).
B	20 mM NaPi (pH 6.5), 50 mM NaCl, 2.5 mM desthiobiotin (elution buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Harvest total CFS.
B	Incubate with benzonase for 30 min on a wheel, at rt.
C	Centrifuge for 30 min at 20,000 g, 4°C.
D	Harvest the soluble fraction (SN).
E	Equilibrate the Strep-Tactin column (IBA Lifesciences) with 2 CV of 1A (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of 1A .
H	Elute the protein of interest with 1B .

Table 4: Final sample

1	Yield
	0.27 mg/mL of WGE and total production of 875 µg for NMR samples
1b	A260/280 ratio
	1.36
2	Stability
	stable
3	Comment on applicability
	Positioning the Strep tag at the N-terminus abolished synthesis.

Additional information

Constructs	Conditions	Comments
F1 ORF6; Strep tag II (pEU-E01-MCS (Cell-Free Sciences)), no cleavage site, N-terminal "WSHPQFEK" eight artificial residues.		No expression observed.



(a) WG-CFPS and Strep-tag purification of ORF6. SDS-PAGE (left panel) and WB (right panel). **(b) SEC profile of ORF6.**

SI17: ORF7a

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF7a
2	Region/Name/Further Specification
	Ectodomain (ED)
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MKIIILFLALITLATCELYHYQECVRGTTVLLKEPCSSGTYESGNSPFHPLADNKFALTCFSTQFAFA CPDGVKHHVYQLRARSVSPKLFIRQEEVQELYSPIFLIVAAIVFITLCFTLKRKTE
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 16-81 (ectodomain of ORF7a)
5	Ratio for construct design (detailed and comprehensible)
	Only the ectodomain without signaling peptide. Transmembrane helix is also not included in the construct.
6	Sequence homology (to SCoV)
	Identity: 85.3%; similarity: 95.9%
7	Published structures (SCoV2 or homologue variants)
	SCoV: PDB 1XAK, 1YO4
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: BMRB 6824

Table 2: Protein Expression

1	Expression vector
	pET24d-GB1 (Novagen, modified by G. Stier (Bogomolovas et al., 2009))
2	Purification-/Solubility-Tag
	N-terminal His ₆ -GB1
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	7.49 kDa / 6,210 M ⁻¹ cm ⁻¹ / 6.99
5	Comments on sequence of expressed construct
	N-terminal „G" one artificial residue due to TEV-cleavage and construct design.
6	Used expression strain

	<i>E.coli</i> (DE3) BL21
7	Cultivation medium
	M9 (uniformly ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.2 mM IPTG at OD ₆₀₀ 0.7
10	Cultivation temperature and time
	25°C for 18-20 h

Table 3: Protein Purification

1	Buffer List
A	20 mM Tris-HCl (pH 8.0), 6 M GdnHCl, 500 mM NaCl, 5 mM imidazole, 2 mM bME (Cell disruption / solubilization of pellet).
B	20 mM Tris (pH 8.0), 6 M GdnHCl, 500 mM NaCl, 10 mM imidazole, 2 mM bME (IMAC1).
C	50 mM NaPi (pH 8.0), 300 mM NaCl, 10 mM imidazole, 2 mM bME (IMAC2).
D	1 mM acetate-D4 (pH 5.0) (final NMR-buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption and solubilization of pellet in buffer 1A .
B	IMAC, gravity flow Ni ²⁺ -NTA (Qiagen), elution with 200 mM imidazole in buffer 1B .
C	Dialysis against buffer 1C .
D	TEV-cleavage (1 mg TEV protease per 10 mL protein solution) o.n. in buffer 1C .
E	Inv. IMAC, elution with 200 mM imidazole in buffer 1C .
F	Dialysis of flow-through of inv. IMAC against 1D and concentrate (NMR-sample).

Table 4: Final sample

1	Yield
	0.4 mg/L ¹⁵ N-M9 medium
1b	A260/280 ratio
	0.7
2	Stability
	Stable throughout measurement (1 day, 298/315 K). No precipitation or degradation observed after four days at rt.

3	Comment on applicability
	Suitable for NMR structure determination, fragment screening, interaction studies.

SI18: ORF7b

Tabel 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF7b
2	Region/Name/Further Specification
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MIELSLIDFY LCFLAFLFL VLIMLIIFWF SLELQDHNET CHA
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-43 (fl ORF7b)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: 85.4%; similarity: 97.2%
7	Published structures (SCoV2 or homologue variants)
	-
8	(Published) assignment (SCoV2 or homologue variants)
	-

Bacterial

Table 2: Protein Expression

1	Expression vector
	pThiorex (GenScript)
2	Purification-/Solubility-Tag
	N-terminal His ₆ -Trx
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	5.37 kDa / 6,990 M ⁻¹ cm ⁻¹ / 4.17
5	Comments on sequence of expressed construct
	N-terminal "GA(M)G" three artificial residues due to TEV-cleavage and construct design.
6	Used expression strain

	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.2 mM IPTG at OD ₆₀₀ 0.7
10	Cultivation temperature and time
	18-20°C for 16-18 h

Table 3: Protein Purification with detergent

1	Buffer List
A	25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME (cell disruption).
B	25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME, 1.5% (w/v) DDM (Solubilization of pellet).
C	25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 mM imidazole, 10 mM bME, 0.02% (w/v) DDM (IMAC).
D	25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl, 0.02% (w/v) DDM (SEC/final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus one tablet of EDTA free protease inhibitor cocktail (Merck)) by microfluidization.
B	Solubilization of pellet after lysis 1B (plus one tablet of EDTA free protease inhibitor cocktail (Merck)).
C	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer 1C .
D	TEV-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer 1C
E	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with 500 mM imidazole in buffer 1C .
F	Rebuffer flow-through of inv. IMAC in buffer 1D (NMR sample).
G	Analytical SEC (SD 75 Increase 10/300 GL (GE Healthcare), ÄKTA start (GE Healthcare)) in buffer 1D .

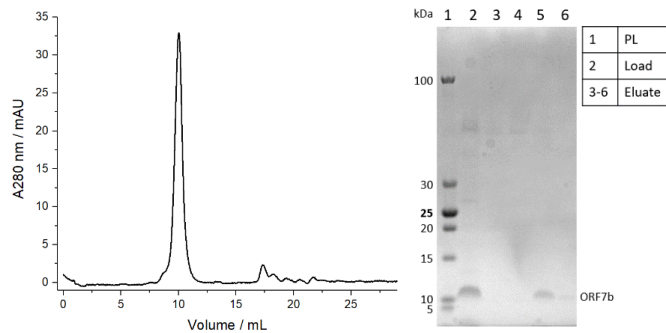
Table 4: Final sample

1	Yield
	0.6 mg/L ¹⁵ N-M9 medium
2	Stability
	Stable throughout measurement (2 days, 283/298 K). No significant precipitation or degradation observed after storage at 4°C for 3 months.

3	Comment on applicability
	Due to necessity of solubilizing agent and tendency to oligomerize structure determination, fragment screening, and interaction studies are hindered.

Additional information

	Constructs	Conditions	Comments
A	As above	Native IMAC buffer: 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME. SEC buffer: 25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl	Nearly no protein was extracted in soluble fraction.
B		Denaturing Solubilizing buffer: 25 mM Tris-HCl (pH 8.0), 6 M GdnHCl, 300 mM NaCl, 5 mM imidazole. IMAC wash buffer: 25 mM Tris-HCl (pH 8.0), 8 M urea, 300 mM NaCl, 5 mM imidazole. Renaturing buffer: 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME. IMAC elution buffer: 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 500 mM imidazole, 10 mM bME.	After refolding and cleavage degradation of protein.
C	Fl ORF7b; His ₆ -SUMO (pE-SUMO (GenScript)), Ulp1-cleavage site, no artificial residues.	Native IMAC buffer: as above SEC buffer: 25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl.	Protein is soluble with fusion, runs in exclusion volume of SD 200 columns, degrades after cleavage. NMR shows SUMO is mostly unfolded.
D		Detergent IMAC buffer: 50 mM NaPi (pH 7.0), 200 mM NaCl, 0.1% (v/v) Triton X-100, 5 mM imidazole, 10 mM bME. SEC buffer: 25 mM NaPi (pH 6.0), 50 mM NaCl, 0.01% (v/v) Triton X-100, 2 mM TCEP-HCl.	Copurification of impurities, runs in exclusion volume of SD 200 columns. NMR shows severely broadened and poorly dispersed resonances hinting to oligomerization.
E		Semi-denaturing IMAC buffer: 50 mM Tris-HCl (pH 8.0), 2 M urea, 300 mM NaCl, 10 mM imidazole, 10 mM bME. SEC buffer: 25 mM NaPi (pH 6.5), 50 mM NaCl, 2 M urea, 5 mM DTT.	Degrades after cleavage.



Analytical SEC of ORF7b. Protein was in exclusion volume (9-11 mL, left panel) with corresponding SDS-PAGE of SEC with fractions analyzed from 7-11 mL elution volume (right panel).

Cell-free

Table 2: Cell-free Protein Synthesis

1	Expression vector
	pEU-E01-MCS (Cell-Free Sciences)
2	Purification-/Solubility-Tag
	C-terminal Strep tag II (WSHPQFEK)
3	Cleavage Site
	-
4	Molecular weight / Extinction coefficient / pI - of protein
	5.37 kDa / 6,990 M ⁻¹ cm ⁻¹ / 4.17
5	Comments on sequence of expressed construct
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.
6	Feeding buffer
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix and 0.1% (w/v) MNG-3.
7	Translation mix
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix, 0.1% (w/v) MNG-3.
8	Protein synthesis temperature and time
	22°C for 16 h without agitation (bilayer method).

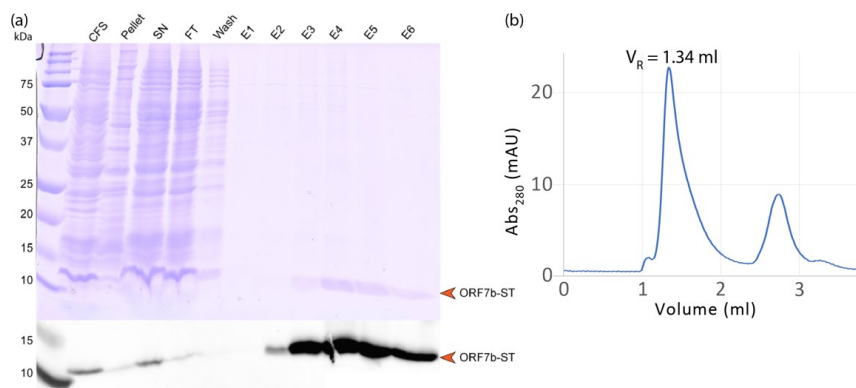
Table 3: Protein Purification

1	Buffer List
A	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 0.1% (w/v) DDM (wash buffer).

B	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 2.5 mM desthiobiotin, and 0.1% (w/v) DDM (elution buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Harvest total CFS.
B	Incubate with benzonase for 30 min on a wheel, at rt.
C	Centrifuge for 30 min at 20,000 g, 4°C.
D	Harvest the soluble fraction (SN).
E	Equilibrate the Strep-Tactin column (IBA Lifesciences) with 2 CV of 1A (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of 1A .
H	Elute the protein of interest with 1B .

Table 4: Final sample

1	Yield
	0.27 mg/mL of WGE and total production of 880 µg for NMR samples
1b	A260/280 ratio
	1.36
2	Stability
	Stable in detergent over several days.
3	Comment on applicability
	Needs reconstitution into membranes for further structural analysis.



(a) WG-CFPS in presence of detergent and Strep-tag purification of ORF7b. SDS-PAGE (upper panel) and WB (lower panel). (b) SEC profile of ORF7b.

SI19: ORF8

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF8
2	Region/Name/Further Specification
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MKFLVFLGIITVAAFHQECSLQSC TQHQP YVVD DPCPIHFY SKWYIRV GARKSAPLIELCVDEA GSKSPIQYIDIGNYTVSCSPFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDFI
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
ORF8	aa 1-121 (fl ORF8 = ORF8)
ORF8_m	aa 1-121 (fl ORF8) with L84S mutation (~ isolate 2019-nCoV_HKU-SZ-002a_2020).
ΔORF8	aa 16-121 (without signal peptide = ΔORF8)
5	Ratio for construct design (detailed and comprehensible)
ORF8	fl protein
ΔORF8	Protein after the hypothetical cleavage of the N-terminal Signal Peptide
6	Sequence homology (to SCoV)
ORF8	Identity: 31.7%; similarity: 70.7%
ΔORF8	Identity: 40.5%; similarity: 66.7%
7	Published structures (SCoV2 or homologue variants)
	SCoV2: 7JTL, 7JX6
8	(Published) assignment (SCoV2 or homologue variants)
	-

Bacterial

Table 2: Protein Expression

1	Expression vector
ORF8_m	pPK1154 (GenScript)
ΔORF8	pET22b (+) (Merck/Novagen)
2	Purification-/Solubility-Tag

ORF8_m	N-terminal His ₆ -SUMO
ΔORF₈	N-terminal His ₆ -GST
3	Cleavage Site
ORF8_m	Ulp1
ΔORF₈	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
ORF8_m	13.80 kDa / 15,930 M ⁻¹ cm ⁻¹ / 5.42
ΔORF₈	12.54 kDa / 15,930 M ⁻¹ cm ⁻¹ / 5.15
5	Comments on sequence of expressed construct
ORF8_m	No artificial residues due to Ulp1-cleavage and construct design.
ΔORF₈	N-terminal "GAMG" three artificial residues due to TEV-cleavage and construct design.
6	Used expression strain
ORF8_m	<i>E. coli</i> BL21 (DE3)
ΔORF₈	<i>E. coli</i> BL21 (DE3) pLysS
7	Cultivation medium
ORF8_m	LB / M9 (uniformly ¹⁵ N-labelled)
ΔORF₈	LB
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
ORF8_m	0.5 mM IPTG at OD ₆₀₀ 0.6
ΔORF₈	0.5 mM IPTG at OD ₆₀₀ 0.6-0.7
10	Cultivation temperature and time
ORF8_m	16-20°C for 16-18 h
ΔORF₈	18°C for 16-18 h

Table 3a: Protein Purification (ORF8)

1	Buffer List
A	10 mM NaPi (pH 8.0), 300 mM NaCl, 10 mM imidazole, 0.5 mM DTT (Cell disruption).
B	10 mM NaPi (pH 8.0), 300 mM NaCl, 10 mM imidazole, 0.5 mM DTT (Solubilization of pellet).
C	10 mM NaPi (pH 8.0), 300 mM NaCl, 0.5 mM DTT (IMAC).
D	50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM DTT, 0.2% (w/v) NP40.
E	50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM DTT.
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (plus one tablet of EDTA free protease inhibitor cocktail (Merck)) by French-press.
B	Solubilization of pellet after lysis 1B (plus one tablet of EDTA free protease inhibitor cocktail (Merck)).
C	IMAC (Nickel-NTA-Agarose, QIAGEN) by hand, elution with 250 mM imidazole in buffer 1C .
D	Ulp1-cleavage (Protein/Ulp1 ratio 10:1) o.n. at 21°C in buffer 1D .
E	Rebuffer in buffer 1E .

Table 3b: Protein Purification (Δ ORF8)

1	Buffer List
A	50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 5% (v/v) glycerol, 50 mM imidazole (cell disruption/IMAC).
B	50 mM Tris-HCl (pH 8.0), 150 mM NaCl (TEV-cleavage).
C	20 mM NaPi (pH 7.4), 150 mM NaCl, 1 mM EDTA (SEC final buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A (supplemented with 0.5 mg/mL lysozyme, 10 μ g/mL DNaseI, 5 mM MgCl ₂ , cOmplete™ EDTA-free protease inhibitors) by incubation for 30 min at RT followed by sonication at 43% amplitude for 2 minutes (1 s on, 1 s off). Extraction of the periplasmic fraction: added 0.1% (v/v) Triton to the total sample after sonication, and incubated 15 min at 4°C. Centrifugation at 24.700 g for 40 min at 4°C. Recovering of the soluble fraction and filtration using 0.45 μ m syringe filters.
B	IMAC (HisTrap FF Crude (GE Healthcare), ÄKTA Pure 25 M1 (GE Healthcare)), binding with buffer 1A supplemented with 50 mM imidazole, elution with imidazole gradient up to 500 mM in buffer 1A .
C	TEV-cleavage (Protein/TEV ratio 1:10) at 4°C, o.n. in buffer 1B .
D	Inv. IMAC (HisTrap FF Crude (GE Healthcare), ÄKTA Pure 25 M1 (GE Healthcare)), binding with buffer 1A supplemented with 50 mM imidazole, elution with imidazole gradient up to 500 mM in buffer 1A .
E	SEC on Increase 10/300 S75 (GE Healthcare) at 4°C in buffer 1C .

Table 4: Final sample

1	Yield
ORF8 m	<0.5 mg/L LB mg/mL ¹⁵ N-M9 medium

Δ ORF8	0.5 mg/L LB medium
2	Stability
ORF8m	Not determined.
Δ ORF8	No significant precipitation or degradation observed after storage at 4°C for 1 week.
3	Comment on applicability
ORF8m	Weak expression into soluble fraction, 30%/70% soluble/inclusion bodies. After purification extremely low yield for NMR studies.
Δ ORF8	Very low yield. It would be very expensive to prepare a labelled sample for NMR studies.

Additional information (bacterial expression)

Constructs	Conditions	Comments
ORF8 with L84S mutation; His ₆ (pPK1151 (Genscript)), TEV-cleavage site, N-terminal "GS" two artificial residues.	As above for ORF8m, only LB medium.	No expression.

Cell-free

Table 2: Cell-free Protein Synthesis

1	Expression vector
	pEU-E01-MCS (Cell-Free Sciences)
2	Purification-/Solubility-Tag
ORF8	C-terminal Strep tag II (WSHPQFEK)
Δ ORF8	N-terminal Strep tag II (WSHPQFEK)
3	Cleavage Site
	-
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
ORF8	15.00 kDa / 21,805 M ⁻¹ cm ⁻¹ / 5.64
Δ ORF8	13.53 Da / 21,805 M ⁻¹ cm ⁻¹ / 5.39
5	Comments on sequence of expressed construct
ORF8	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.
Δ ORF8	N-terminal "M" and C-terminal "SAWSHPQFEK" eleven artificial residues due to construct design.

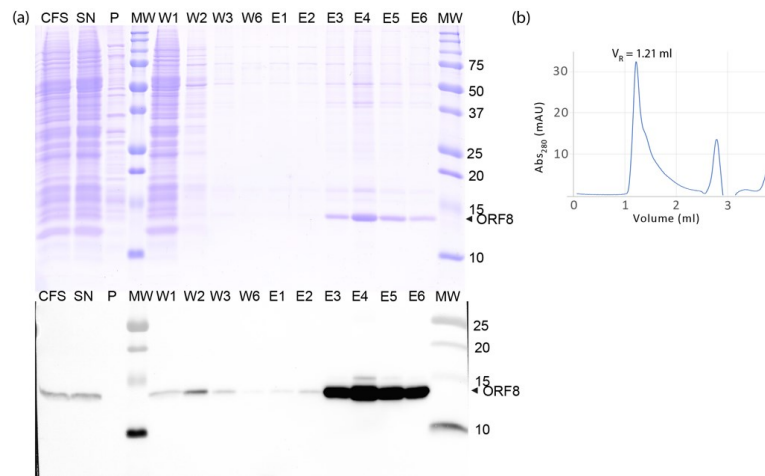
6	Feeding buffer
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix and 0.05% (w/v) Brij-58.
7	Translation mix
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix 0.05% (w/v) Brij-58.
8	Protein synthesis temperature and time
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification (ORF8a and ORF8b)

1	Buffer List
A	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 0.1% (w/v) DDM (wash buffer).
B	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 2.5 mM desthiobiotin, and 0.1% (w/v) DDM (elution buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Harvest total CFS.
B	Incubate with benzonase for 30 min on a wheel, at rt.
C	Centrifuge for 30 min at 20,000 g, 4°C.
D	Harvest the soluble fraction (SN).
E	Equilibrate the Strep-Tactin column (IBA Lifesciences) with 2 CV of 1A (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of 1A .
H	Elute the protein of interest with 1B .

Table 4: Final sample

1	Yield
	0.62 mg/mL WGE after purification. Total of 683 µg for the NMR samples
1b	A260/280 ratio
	0.7
2	Stability
	Stable at 4°C for weeks.
3	Comment on applicability
	Protein very sensitive to dilution-concentration steps. Purity is sufficient for NMR as other cell-free proteins are not labelled.



(a) WG-CFPS in presence of detergent and Strep-tag purification of ORF8. SDS-PAGE (upper panel) and WB (lower panel). (b) SEC profile of ORF8.

SI20: ORF9a (Nucleocapsid (N) protein)

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF9a; Nucleocapsid (N) phosphoprotein
2	Region/Name/Further Specification
	N-terminal disordered region (aa 1-43, IDR1) / N-terminal RNA binding domain (aa 44-180, NTD) / serine-arginine (SR) rich motif (aa 181-212, SR) / central disordered linker (aa 181-248, IDR2) / C-terminal dimerization domain (247-364) / C-terminal disordered region (aa 365-419, IDR3)
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MSDNGPQNQRNAPRITFGGSPDSTGSNQNNGERSGARSKQRRPQGLPNNTASWFTALTQHGKED LKFPFGQGVPIINTSSPDDQIGYYRRATRRIRGGDGKMKDLSPRWYFYLLGTGPEAGLPYGAN KDGIIWVATEGALNTPKDHIGTRNPANNAIIVLQLPQGTTLPKGFYAEGRGGSQASSRSSRSR NSSRNSTPGSSRGTSPARMAGNGGDAALALLLDRLNQLESKMSGKGQQQQGQTVTKKSAAE ASKKPRQKRTATKAYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKHWPOIAQFAPSASAFFG MSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKKADET QALPQRQKKQQTVLLPAADLDDFSKQLQSQSMSSADSTQA
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
IDR1- NTD- IDR2	aa 1-248 (of fl ORF9a)
NTD- SR	aa 44-212 (of fl ORF9a)
NTD	aa 44-180 (of fl ORF9a)
CTD	aa 247-364 (of fl ORF9a)
5	Ratio for construct design (detailed and comprehensible)
IDR1- NTD- IDR2	Based on boundaries from SCoV homolog.
NTD- SR	In analogy to the available NMR (PDB 6YI3) and crystal (6M3M) structures of N-NTD SCoV2.
NTD	In analogy to the available NMR (PDB 6YI3) and crystal (6M3M) structures of N-NTD SCoV2.
CTD	In analogy to the available NMR structure (PDB 2JW8) of N-CTD from SCoV.
6	Sequence homology (to SCoV)
IDR1- NTD- IDR2	Identity: 90%; similarity: 94%
NTD- SR	Identity: 92%; similarity: 96%
NTD	Identity: 93%; similarity: 97%
CTD	Identity: 96%; similarity: 98%
7	Published structures (SCoV2 or homologue variants)

	SCoV: PDB 2JW8, 2CJR SCoV2: PDB 6YI3, 6M3M, 6VYO, 6WKP, 6WZO, 6WJI, 6YUN, 6ZCO, 7CE0, 7C22
8	(Published) assignment (SCoV2 or homologue variants)
	SCoV: BMRB 15511 (CoV) SCoV2: PDB 6YI3, BMRB 34511 (NTD), BMRB 50518 (CTD), BRMB 50619 (IDR1), BMRB 50618 (IDR2), BMRB 50557 (IDR1), BMRB 50558 (IDR2).

Table 2: Protein Expression

1	Expression vector
IDR1-NTD-IDR2	pET29a(+) (Twistbioscience)
NTD-SR	pET-28a(+) (GenScript)
NTD	pET-28a(+) (GenScript)
CTD	pKM263 (GenScript)
2	Purification-/Solubility-Tag
IDR1-NTD-IDR2	-
NTD-SR	N-terminal His ₆
NTD	N-terminal His ₆
CTD	N-terminal His ₆ -GST
3	Cleavage Site
IDR1-NTD-IDR2	-
NTD-SR	TEV
NTD	TEV
CTD	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
IDR1-NTD-IDR2	26.52 kDa / 26,930 M ⁻¹ cm ⁻¹ / 10.57
NTD-SR	18.10 kDa / 26,930 M ⁻¹ cm ⁻¹ / 10.35
NTD	14.85 kDa / 26,930 M ⁻¹ cm ⁻¹ / 9.60
CTD	13.56 kDa / 16,960 M ⁻¹ cm ⁻¹ / 9.77

5a	Comments on sequence of expressed construct
IDR1-NTD-IDR2	No artificial residues due to construct design.
NTD-SR	No artificial residues due to TEV-cleavage and construct design.
NTD	No artificial residues due to TEV-cleavage and construct design.
CTD	N-terminal „GAMG" four artificial residues due to TEV-cleavage and construct design.
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)
7	Cultivation medium
IDR1-NTD-IDR2	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
NTD-SR	LB / M9 (uniformly ¹⁵ N-labelled)
NTD	LB / M9 (uniformly ¹⁵ N-labelled)
CTD	LB / M9 (uniformly ¹⁵ N or ¹³ C, ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
IDR1-NTD-IDR2	0.2 mM IPTG at OD ₆₀₀ 0.8
NTD-SR	0.2 mM IPTG at OD ₆₀₀ 0.7
NTD	0.2 mM IPTG at OD ₆₀₀ 0.7
CTD	1 mM IPTG at OD ₆₀₀ 0.7
10	Cultivation temperature and time
IDR1-NTD-IDR2	Cells are grown at 37°C in 1 L LB until OD ₆₀₀ 0.8, then transferred in 250 mL labelled minimal medium (4x). After 1 h of metabolite clearance, the culture is induced at 18°C for 16-18 h. For unlabelled protein, culture is induced at OD ₆₀₀ 0.9.
NTD-SR	16-18°C for 16-18 h
NTD	16-18°C for 16-18 h
CTD	20-22°C for 18-20 h

Table 3a: Protein Purification (IDR1-NTD-IDR2)

1	Buffer List
A	25 mM Tris-HCl (pH 8.0), 1 M NaCl, 5% (v/v) glycerol, RNase, DNase, proteases inhibitor cocktail (SIGMAFAST™ tablet, 500 µL of 100x stock) (lysis buffer).
B	25 mM Tris-HCl (pH 7.2) (dialysis after lysis and binding buffer).
C	25 mM Tris-HCl (pH 7.2), 1 M NaCl (elution buffer).
D	25 mM Tris-HCl (pH 7.2), 450 mM NaCl, 0.02% (w/v) NaN ₃ (NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell lysis in Buffer 1A by sonication (30 min with pulse 1 s on, 10 s off). It is crucial to add a cocktail of proteases inhibitors in lysis buffer; this step is crucial to preserve construct integrity.
B	Dialysis O/N at 4°C in Buffer 1B for buffer exchange.
C	Ion Exchange chromatography with HiTrap SP FF 5 mL column (GE Healthcare), gradient elution with buffer 1C . The protein eluted at 45-50% gradient.

Table 3b: Protein Purification (NTD and NTD-SR)

1	Buffer List
A	50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 20 mM imidazole, 10% (v/v) glycerol, 0.01 mg/mL DNase, 5 mM MgCl ₂ and protease inhibitor cocktail (Sigma) (cell disruption).
B	50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 20 mM imidazole, 10% (v/v) glycerol (IMAC).
C	50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 500 mM imidazole, 10% (v/v) glycerol (IMAC).
D	50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 1 mM DTT (dialysis after IMAC / TEV-cleavage).
E	20 mM Na ₂ HPO ₄ (pH 6.5), 50 mM NaCl, 500 µM PMSF, 3 mM NaN ₃ , 3 mM EDTA (final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Cell disruption in buffer 1A by sonication.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer 1B and 1C .
C	TEV-cleavage (1:10 (v/v) TEV:protein solution) during dialysis o.n. in buffer 1D .
D	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer 1B and 1C .
E	NMR sample preparation in buffer 1E .

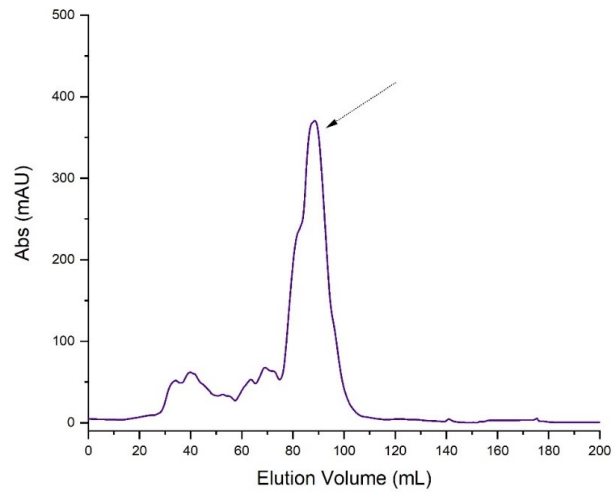
Table 3c: Protein Purification (CTD)

1	Buffer List
A	50 mM NaPi (pH 7.4), 150 mM NaCl, 10 mM imidazole (cell disruption / IMAC/ dialysis after IMAC / TEV-cleavage).
B	25 mM NaPi (pH 6.0), 50 mM NaCl, 0.5 mM EDTA, 0.02% (w/v) NaN ₃ (SEC / final NMR buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)

A	Cell disruption in buffer 1A (plus 100 μ L protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni ²⁺ -NTA), Elution with 150-500 mM imidazole in buffer 1A .
C	Dialysis o.n. in in buffer 1A .
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer 1A .
E	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer 1B .
F	NMR sample preparation in buffer 1B .

Table 4: Final sample

1	Yield
IDR1-NTD-IDR2	12 mg/L ¹³ C, ¹⁵ N M9 medium
NTD-SR	3 mg/L ¹⁵ N M9 medium
NTD	3 mg/L ¹⁵ N M9 medium
CTD	2 mg/L ¹³ C, ¹⁵ N-M9 medium
1b	A260/280 ratio
IDR1-NTD-IDR2	0.63
NTD-SR	0.7
NTD	0.7
CTD	0.55
2	Stability
IDR1-NTD-IDR2	Protein is stable for at least one1 week at working conditions (298 K).
NTD-SR	Stable throughout measurement (15 days, 298 K). No significant precipitation or degradation observed after storage at 4°C for 5 weeks.
NTD	Stable throughout measurement (15 days, 298 K). No significant precipitation or degradation observed after storage at 4°C for 5 weeks.
CTD	Stable throughout measurement (7 days, 303 K). No significant precipitation or degradation observed after storage at 4°C for 8 weeks. Tolerates temperature up to 315 K.
3	Comment on applicability
	All suitable for NMR structure determination, fragment screening, interaction studies.



Chromatogram of IEC of aa 1-248 construct. Protein is eluted at 45% gradient of Buffer 1B, fractions from 85-100 mL were collected.

SI21: ORF9b

Table 1: General Information

1	Protein Name
	ORF9b
2	Region/Name/Further Specification
3	Sequence of fl protein
	MDPKISEMHP ALRLVDPQIQ LAVTRMENAV GRDQNNVGPK VYPIILRLGS PLSLNMARKT LNSLEDKAFQ LTPIAVQMTK LATTEELPDE FVVVTVK
4	Protein boundaries of expressed construct
	aa 1-97 (fl ORF9b)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity : 72.4%; similarity: 95.0%
7	Published structures (SCoV2 or homologue variants)
	SCoV2: PDB 6Z4U
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Cell-free Protein Synthesis

1	Expression vector
	pEU-E01-MCS (Cell-Free Sciences)
2	Purification-/Solubility-Tag
	C-terminal Strep tag II (WSHPQFEK)
3	Cleavage Site
	-
4	Molecular weight / Extinction coefficient / pI - of protein
	11.99 kDa / 6,990 M ⁻¹ cm ⁻¹ / 6.73
5	Comments on sequence of expressed construct
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.
6	Feeding buffer

	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix
7	Translation mix
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix
8	Protein synthesis temperature and time
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

1	Buffer List
A	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA.
B	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 2.5 mM desthiobiotin.
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Harvest total CFS.
B	Incubate with benzonase for 30 min on a wheel, at rt.
C	Centrifuge for 30 min at 20,000 g, 4°C.
D	Harvest the soluble fraction (SN).
E	Equilibrate the Strep-Tactin column (IBA Lifesciences) with 2 CV of 1A (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of 1A .
H	Elute the protein of interest with 1B .

Table 4: Final sample

1	Yield
	0.64 mg/mL WGE and total production of 1338 µg for NMR samples.
1b	A260/280 ratio
	0.76
2	Stability
	Stable at 4°C for a week.
3	Comment on applicability
	Protein studied at pH 6, 7.5 and pH 8. Methionine gets oxidized without DTT in the buffer.

Additional information

	Constructs	Conditions	Comments
A	F1 ORF9b; Strep tag II (pEU-E01-MCS (Cell-Free Sciences)); no cleavage site; C-terminal "WSHPQFEK" eight artificial residues.	As above with 0.1% (w/v) DDM	NMR shows severely broadened resonances due to oligomerization or protein micelles.
B		As above without DTT	Methionines get oxidated.

SI22: ORF14

Table 1: General Information

1	Protein Name
	ORF14
2	Region/Name/Further Specification
3	Sequence of fl protein
	MLQSCYNFLKEQHCQKASTQKGAEEAAVKPLLVP HHVVATVQEIQQA AVGELLLLLEWLAMA VM LLLLCCCLTD
4	Protein boundaries of expressed construct
	aa 1-73 (fl ORF14)
5	Ratio for construct design
	fl protein
6	Sequence homology (to SCoV)
	Identity: NA; similarity: NA
7	Published structures (SCoV2 or homologue variants)
	-
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Cell-free Protein Synthesis

1	Expression vector
	pEU-E01-MCS (Cell-Free Sciences)
2	Purification-/Solubility-Tag
	N-terminal Strep tag II (WSHPQFEK)
3	Cleavage Site
	-
4	Molecular weight / Extinction coefficient / pI - of protein
	9.26 kDa / 12,490 M ⁻¹ cm ⁻¹ / 6.01
5	Comments on sequence of expressed construct
	N-terminal "WSHPQFEKGGG" eleven artificial residues due to construct design.
6	Feeding buffer

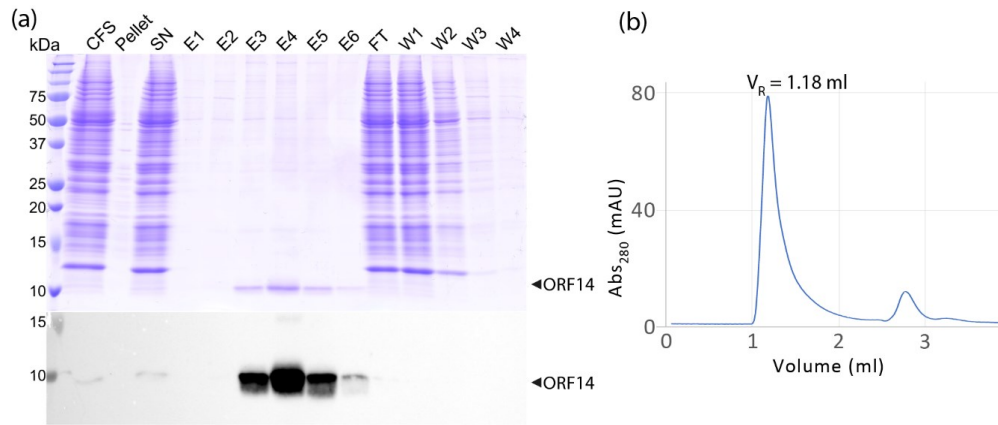
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix and 0.05% (w/v) Brij-58.
7	Translation mix
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix 0.05% (w/v) Brij-58.
8	Protein synthesis temperature and time
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

1	Buffer List
A	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 0.1% (w/v) DDM (wash buffer).
B	100 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA, 2.5 mM desthiobiotin, and 0.1% (w/v) DDM (elution buffer).
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Harvest total CFS.
B	Incubate with benzonase for 30 min on a wheel, at rt.
C	Centrifuge for 30 min at 20,000 g, 4°C.
D	Harvest the soluble fraction (SN).
E	Equilibrate the Strep-Tactin column (IBA Lifesciences) with 2 CV of 1A (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of 1A .
H	Elute the protein of interest with 1B .

Table 4: Final sample

1	Yield
	0.43 mg/mL WGE
1b	A260/280 ratio
	1.06
2	Stability
	protein has proved unstable during lipid insertion using cyclodextrin for detergent removal
3	Comment on applicability
	Solution NMR shows severely broadened resonances hinting to oligomerization or too big protein micelles. Lipid reconstitution is ongoing.



(a) WG-CFPS in presence of detergent and Strep-tag purification of ORF14. SDS-PAGE (upper panel) and WB (lower panel). (b) SEC profile of ORF14.

SI23: ORF10

Table 1: General Information

1	Protein Name (according to NCBI Reference Sequence NC_045512.2)
	ORF10
2	Region/Name/Further Specification
3	Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)
	MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT
4	Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)
	aa 1-38 (fl ORF10)
5	Ratio for construct design
	Hypothetical fl protein.
6	Sequence homology (to SCoV)
	Identity: 29%; similarity: 52% with ORF9b
7	Published structures (SCoV2 or homologue variants)
	-
8	(Published) assignment (SCoV2 or homologue variants)
	-

Table 2: Protein Expression

1	Expression vector
	pThiorex (GenScript)
2	Purification-/Solubility-Tag
	N-terminal His ₆ -Trx
3	Cleavage Site
	TEV
4	Molecular weight / Extinction coefficient / pI - of cleaved protein
	4.45 kDa / 4,470 M ⁻¹ cm ⁻¹ / 7.93
5	Comments on sequence of expressed construct
	N-terminal "GA" two artificial residues due to TEV-cleavage and construct design
6	Used expression strain
	<i>E. coli</i> BL21 (DE3)

7	Cultivation medium
	LB / M9 (uniformly ¹⁵ N-labelled)
8	Induction system
	IPTG inducible T7 promoter
9	Induction of protein expression
	0.2 mM IPTG at OD ₆₀₀ 0.6-0.7
10	Cultivation temperature and time
	18-20°C for 16-18 h

Table 3: Protein Purification

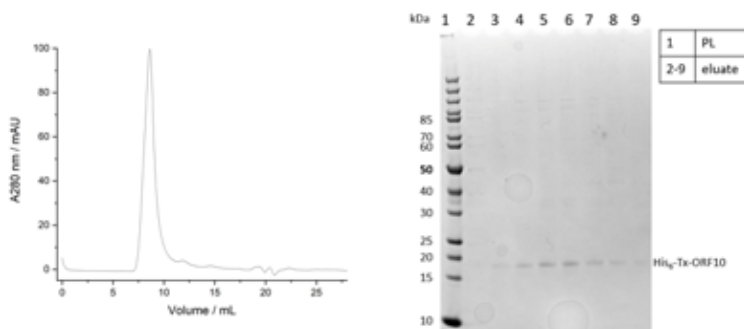
1	Buffer List
A	25 mM Tris (pH 8.0), 6 M GdnHCl, 300 mM NaCl, 5 mM imidazole (Solubilization)
B	25 mM Tris (pH 8.0), 8 M urea, 300 mM NaCl, 5 mM imidazole (IMAC - wash)
C	25 mM Tris (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME (IMAC - elution)
D	25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl.
2	Purification steps (with corresponding buffer(s) and incubation times)
A	Solubilization of cell pellet and inclusion bodies in 1A (plus one tablet of EDTA free protease inhibitor cocktail (Merck)).
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), washed with buffer 1B , refolded on column in buffer 1C , elution with imidazole gradient up to 500 mM in buffer 1C .
C	Analytic TEV-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer 1C .
D	Analytical SEC (SD 75 Increase 10/300 GL (GE Healthcare), ÄKTA start (GE Healthcare)) in buffer 1D .

Table 4: Final sample

1	Yield
	2 mg/L (¹⁵ N-M9) His ₆ -SUMO-fused
2	Stability
	Degrades after cleavage
3	Comment on applicability
	Tendency to oligomerize (exclusion volume of SD 75 column).

Additional information

	Constructs	Conditions	Comments
A	As above	Native IMAC buffer: 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME. SEC buffer: 25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl.	Nearly no protein was extracted in soluble fraction (in inclusion bodies)
B	Fl ORF10; His ₆ -SUMO (pE-SUMO (GenScript)), Ulp1-cleavage site, no artificial residues.	Native IMAC buffer: as above SEC buffer: 25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl.	Protein is mostly soluble with fusion, partial degradation (copurification of His ₆ -SUMO), runs in exclusion volume of SD 200 columns, degrades after cleavage. NMR shows SUMO is mostly unfolded.
C		Detergent IMAC buffer: 50 mM NaPi (pH 7.0), 200 mM NaCl, 0.1% (v/v) Triton X-100, 5 mM imidazole, 10 mM bME. SEC buffer: 25 mM NaPi (pH 6.0), 50 mM NaCl, 0.01% (v/v) Triton X-100, 2 mM TCEP-HCl.	Copurification of impurities, runs in exclusion volume of SD 75 columns hinting to oligomerization. Degrades after cleavage.
D		Semi-denaturing IMAC buffer: 50 mM Tris-HCl (pH 8.0), 2 M urea, 300 mM NaCl, 10 mM imidazole, 10 mM bME. SEC buffer: 25 mM NaPi (pH 6.5), 50 mM NaCl, 2 M urea, 5 mM DTT.	Degrades after cleavage.



Analytical SEC of His₆-Trx-ORF10. Protein was in exclusion volume (8.5-12 mL, left panel) with corresponding SDS-PAGE of SEC with fractions analyzed from 8-12 mL elution volume (right panel).