



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE

DOTTORATO DI RICERCA IN  
SCIENZE CHIMICHE

CICLO XXXV

COORDINATORE Prof. Annamaria Papini

**Development of new strategies to  
investigate intrinsically disordered  
proteins and their interactions**

Settore Scientifico Disciplinare CHIM/03

**Dottorando**

Dott. Pontoriero Letizia

Letizia Pontoriero

**Tutore**

Prof. Pierattelli Roberta

Roberta Pierattelli

**Co-Tutore**

Prof. Felli Isabella Caterina

Isabella Caterina Felli

**Coordinatore**

Prof. Papini Annamaria

---

Anni 2019/2022





*To anyone who believed in me*

## **ACKNOWLEDGMENTS**

After these three intense years of research and the writing of the thesis, this section proves to be the most difficult to flesh out. This doesn't mean I'm confused about who to thank but it is not easy to find the correct words to do so.

For me, these PhD experiences have been a gamble, it has brought me in front of challenges that I was not sure I would be able to face. In particular, at the beginning, I did not feel appropriate because I started this journey at an age when, at least in Italy, one should already have a good postdoctoral curriculum. But now, close to be 35 years old, I found "my place" and I am certain that I have lived three of the most important and formative years of my life, not only from the scientific field but especially from the human one.

For this reason, I can only begin by thanking Prof. Roberta Pierattelli and Prof. Isabella Felli who welcomed me when I was still an undergraduate student for my master's thesis. I thank them for being advisors even before being supervisors, for believing in me and in my enthusiasm from the very first moment. I especially thank them for being by my side and helping me through some very difficult personal moments like the last two months.

I then thank all the undergraduate students I had the pleasure to train and to work with: Francesca, Eleonora, Naomi, Peppe, Maria, Tessa and Angela because, above all, they taught me the beauty of confrontation, reminding me that there is always something to learn!

A special thanks goes to Lorenzo Bracaglia, one of the sweetest people I know, who went from being a thesis student to a colleague as well as

a friend (Lore, sorry for invading your desk!!).

Thanks to all the people with whom I had the honor and pleasure to confront and collaborate: Prof. Vladimir Uversky, Prof. Annamaria Papini, Prof. Alexandre Bonvin, Prof. Robert Konrat, Prof. Alejandro Vila, Prof. Luigi Messori, Prof. Harald Schwalbe and all the Covid-19 NMR consortium, Dr. Andreas Binolfi, Dr. Andreas Schlundt, Dr. Sophie Korn and Dr. Lara Massai. I reserve a special mention to Dr. Lisandro Gonzalez, also for introducing me to the art of yerba mate.

Another special mention goes to Prof. Marco Fragai for his patience given all the times he had to explain to me how to run ITC measurements.

Thanks to the CERM, my second home, and to all the "past and present" CERMIANS who have hosted and "put up with" me over the years, from professors to technicians and researchers, not forgetting the PhD students.

An unique and honorable mention goes to Marco Schiavina, my "third supervisor" and my partner in crime. I have scientifically and humanely grown up with Marco and I believe that part of the credits and recognitions I have gained during these 4 years of working together goes to him. My PhD period would not have been the same without him and, probably, I have to thank my PhD for giving me one of the best friends I could desire.

A special thanks, which might seem obvious but is not, goes to my family members Claudia, Rosanna and Pino, my pillars, for giving me roots and wings, for supporting me and always believing in me despite everything. Without them I would not be here. To them, I owe the credit

for the woman I am becoming.

Last but not least, there is Stefano, the love of my life. I must first apologize to him for some of the difficult moments that have accompanied these three years. We are building our own family and Stefano has been patient to wait and help me, most importantly, he has never obstructed the path of my personal fulfillment. Stefano, I love you and will love you forever. Thank you.

“Even as a youngster, though, I could not bring myself to believe that if knowledge presented danger, the solution was ignorance. To me, it always seemed that the solution had to be wisdom. You did not refuse to look at danger, rather you learned how to handle it safely, After all, that is the point of the challenge posed to man since a group of primates evolved into our species. Any technological innovation can be dangerous: fire has been from the beginning, and language even more so; both can be said to still be dangerous nowadays, but no man could be said to be so without fire and speech.”

Isaac Asimov, from *The Caves of Steel*



# TABLE OF CONTENTS

List of Abbreviations	i
ABSTRACT	iii
AIM	iv
INTRODUCTION	1
The flavours of Disorder	1
NMR: the golden technique	4
CHAPTER 1	8
DEALING WITH PHYSIOLOGICAL CONDITIONS	8
1.1 Fingerprints of IDPs at physiological conditions	8
1.2 Spying chemical exchange with water	17
1.3 <sup>13</sup> C toolbox highlights novel insights on $\alpha$ -synuclein	22
CHAPTER 2	25
VIRAL MODULAR PROTEINS: THE CASE OF THE N PROTEIN FROM SARS-COV-2	25
2.1 The Nucleocapsid protein	25
2.2 “Divide and conquer”: the NTR construct	27
2.3 The promiscuous interactions of NTR	36
2.4 Increasing the complexity: the full-length N protein	57
CONCLUSION	72
Bibliography	74

## List of Abbreviations

**AF** AlphaFold

**BMRB** Biological Magnetic Resonance Bank

**CPMG** Carr-Purcell-Meiboom-Gill

**ECM** ExtraCellular Matrix

**EMSA** Electrophoretic Mobility Shift Assay

**GAG** GlycosAmineGlycan

**gRNA** genomic RNA

**H<sup>N</sup>** amide proton

**IDP** Intrinsically Disordered Protein

**IDR** Intrinsically Disordered Region

**IEC** Ions Exchange Chromatography

**LCR** Low Complexity Region

**LLPS** Liquid-Liquid Phase Separation

**MR** Multiple Receivers

**MTSL** S-(1-oxyl-2,2,5,5,-tetramethyl-2,5,-dihydro-1H- pyrrol-3-yl)  
methylmethane-sulfonothiolate

**NAC** Non-Amyloid Component

**NMR** Nuclear Magnetic Resonance

**NTD** N-Terminal Domain

**NTR** N-Terminal Region

**PDB** Protein Data Bank

**PTM** Post Translational Modification

**PPi** Protein-Protein interactions

**PRE** Paramagnetic Relaxation Enhancements



**RBP** RNA-Binding Protein  
**RNP** RiboNucleoProtein  
**SEC** Size Exclusion Chromatography  
**SLIM** Short LInear Motif  
**SUV** Small Unilamellar Vesicle  
**TF** Trascrption factor  
**TRIS** Tris(hydroxymethyl)aminomethan

## **ABSTRACT**

The interest in Intrinsically Disordered Proteins (IDPs) and Intrinsically Disordered Regions (IDRs) of complex proteins arises from the very different processes they can modulate in cells and viruses. These highly flexible molecules are biological tools that can be handled for different aims ranging from cell fate control to regulation of metabolism. For these reasons, their study covers different fields of research, from chemistry to medicine. Nuclear Magnetic Resonance (NMR) spectroscopy plays a central role in the characterization of IDPs/IDRs being a unique method able to provide atomic information on their structural and dynamic features. The peculiar properties of IDPs strongly influence the NMR observables thus improved experimental methods are needed in order to study these proteins and their interactions, especially when approaching physiological conditions. It is thus important to develop novel experiments optimized for the properties of IDPs and overcome the experimental limitations linked to their biochemical and biophysical properties. In this framework,  $^{13}\text{C}$ -detected experiments and high-field instrumentation provide a unique strategy to investigate IDPs/IDRs at atomic resolution in different experimental conditions and to disentangle the information on the heterogeneous nature of complex multi-domain proteins while mimicking the physiological milieu.

## **KEYWORDS**

Intrinsically disordered proteins, intrinsically disordered regions, modular protein, NMR, heteronuclear detection,  $^{13}\text{C}$  detection, high-field NMR, SARS-CoV 2

## AIM

The central aim of this PhD project consists of the development and application of novel NMR approaches to overcome critical points arising when investigating complex IDPs of high biomedical relevance. Modular proteins, such as Transcription Factors (TF) and RNA Binding Proteins (RBP), are in general very rich in IDRs but high-resolution information on these portions is often lacking. These disordered fragments are referred to as “flexible linkers” with the implicit assumption that their only role consists in connecting the functional globular domains, which is seldom true. Thus, the investigation of the flexible linkers of complex protein machineries represents an important objective to complete the description of modular ensembles at high resolution. Modularity is often exploited by viruses and linked to diseases, such as the onset of cancer and neurodegeneration. Therefore, shifting the attention to the IDRs of complex proteins can reveal novel modules that are functional but not yet described in the Protein Data Bank <sup>1</sup> (PDB, [www.rcsb.org](http://www.rcsb.org)). Deep knowledge of the features of IDPs and IDRs is expected to open novel opportunities to design specific molecules able to interact with flexible stretches.

In this context, direct detection of heteronuclei such as <sup>13</sup>C represents an attractive alternative to the canonic <sup>1</sup>H detected NMR because it involves only non-exchangeable nuclear spins and is characterized by a large chemical shift dispersion. Several variants of exclusively heteronuclear NMR experiments can be designed to take advantage of the properties of the low-γ and provide additional information through different kinds of NMR observables. The <sup>13</sup>C-direct detected

experiments offer the opportunity to determine exchange rates even in cases in which amide protons are not directly observable. Investigation of how the exchange rates are modulated by the properties of IDPs in their native state and upon interactions with different targets has not been addressed in detail and would provide novel information.

The structure-function paradigm and current drug discovery approaches, based on seeking the optimal fit between complementary well-defined surfaces, are pillars of modern protein chemistry. IDPs/IDRs, characterized by many conformers accessible at room temperature, do not fit into this picture. The potential impact of shifting the focus to IDPs/IDRs as targets for drug development is extraordinary. To fully realize this objective a thorough understanding of their structural and dynamic properties is required, together with radically novel approaches to interfere with their function/misfunction and design novel drugs. The development of novel NMR experiments for the high-resolution characterization of complex modular proteins represents an important contribution towards these ambitious objectives.

In this frame, I focused my PhD project on the design, implementation, and application of  $^{13}\text{C}$  methodologies by studying two IDPs of biomedical relevance:  $\alpha$ -synuclein and the modular Nucleocapsid protein (N) of SARS-CoV 2. The achieved results are here presented in two chapters. An introduction to describe the main characters of this research work is also provided. The first chapter regards the design of  $^{13}\text{C}$ -NMR experiments to challenge physiological-like experimental conditions, in order to obtain atomic details on side chain nuclei and

information on amide proton ( $H^N$ )-water exchange at model physiological conditions values. The second chapter discusses the application of the proposed methodologies to study the multi-domain protein N of SARS-CoV 2 through the use of a 248 residues protein construct and the wild-type full-length protein. The recent pandemic strongly impacted biomedical research and highlighted the urgency of finding novel strategies to investigate challenging systems such as the N protein. The assignment of the 1-248 residues is provided together with structural characterization studies performed with high field NMR instrumentation, including the first commercially available spectrometer operating at 1200 MHz (28.2 T magnet). The interactions with different biological relevant targets, including the viral genomic RNA, are investigated.

The scientific publications resulting from my doctoral research are also enclosed.

# INTRODUCTION

## The flavours of Disorder

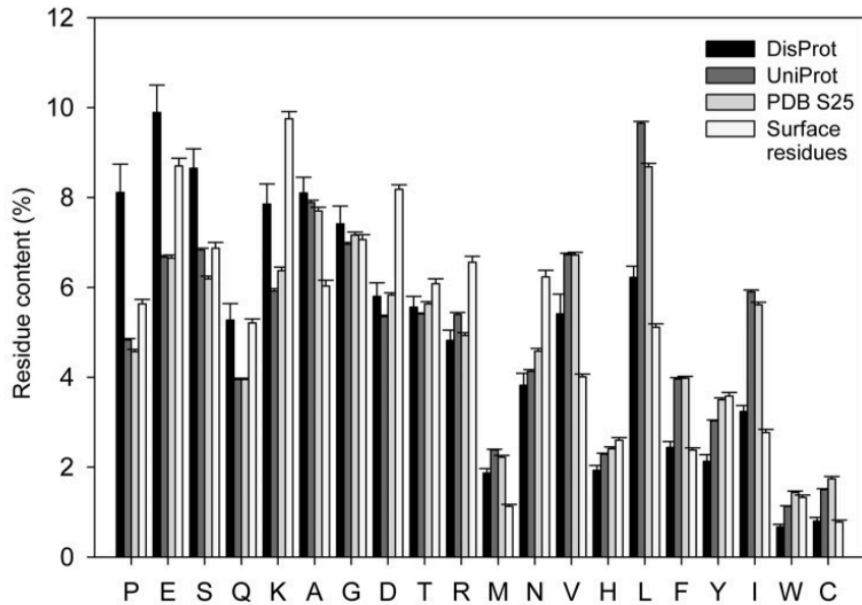
Protein chemistry was born and developed on the basis of the so-called “key-to-lock” paradigm whereby the functionality of a protein is dependent on its tridimensional structure<sup>2,3</sup>.

However, since the '50s, evidence of adaptability in protein structural arrangements emerged, shaping the concepts of configurational adaptability<sup>4</sup> and induced fit<sup>5</sup>. Many X-ray structures lacked information on electron density maps and other observables provided hints on the dynamics of some protein structural arrangements<sup>6-8</sup>.

In the following decades, the increased knowledge of genetic makeup together with the implementation of bioinformatics investigation<sup>9-12</sup>, databases<sup>13,14</sup> and improved spectroscopic techniques led to the identification of a class of natively unfolded species that are now known as Intrinsically Disordered Proteins and protein Regions (IDPs/IDRs)<sup>15</sup>.

Typical features of these macromolecules are the significant depletion in their composition of bulky hydrophobic amino acids and the specific electrostatic distribution pattern of charged residues along the primary

sequence (Figure 1)<sup>16</sup>.



**Figure 1.** The results of a statistical analysis of the amino acid compositions of proteins present in the standard data sets DisProt<sup>13</sup>, UniProt<sup>14</sup>, PDB Select 25<sup>1</sup> and surface residues<sup>12</sup>. Figure from reference 16.

The so-called “disorder promoting” residues proline, glycine and charged amino acids are preferred with respect to aromatic and hydrophobic amino acids, in order to establish a network of electrostatic interactions that prevent the formation of secondary and tertiary structural elements<sup>17</sup>. The poor variability of the amino-acidic composition allows to eliminate boundaries imposed by a fold, allowing sampling of larger conformational space and adopting different levels of compaction<sup>18,19</sup>.

Disorder is present from viruses to procaryotes and it increases while moving to eukaryotes because of the parallel increased complexity of

the Protein-Protein interaction (PPI) networks<sup>20</sup>. It is predicted that more than 30% of the proteome of eukaryotes is disordered<sup>21</sup>. IDPs act, indeed, as biological tools that can be handled for different aims; they are often involved in key regulatory processes for which the adaptability of the protein structure and dynamics represents a clear functional advantage (cellular signalling, translation and transcription, etc). IDPs and IDRs can fold upon binding<sup>22</sup>, condensate<sup>23</sup>, engage in “fuzzy” interactions<sup>24</sup> and they often exploit Post-Translational Modifications (PTMs) to modulate pivotal features such as their charge<sup>25,26</sup>. The lack of a stable structure guarantees malleability and plasticity of the polypeptide, which allows them to engage in multiple interactions with different partners and act as “hubs” in complex PPIs. In this light, the importance of the functional role of IDPs also becomes evident from the strong link identified between their misfunction and many pathologies<sup>27</sup>.

Many proteins are totally unfolded, such as  $\alpha$ -synuclein and prothymosin  $\alpha$ , but proteins that are exploited for critical fine-tuned roles usually present a modular organization. The tethering of folded and disordered domains is a simple and smart biological trick to compartmentalize specific functions into dedicated portions along the primary sequence. During the last two decades, it became more evident that many of the initially defined “flexible linkers” are actually functional domains whose effect is not merely confined to connecting folded regions or giving pliability to the structure<sup>28</sup>, expanding the concept of “allostery/cooperativity”<sup>29,30</sup>. This is the case of many TF<sup>31</sup> and nucleic acids binding proteins<sup>32</sup>.



In principle, following the concept of “inheritance through homology”, it should be possible to identify the IDRs’ functions from their primary sequence but it is not trivial because their amino-acidic composition is highly variable although they maintain the same biological role. However, the constant improvements of even more sophisticated algorithms dedicated to disorder prediction<sup>33,34</sup> allow the identification of Short Linear Motifs (SLIMs)<sup>35</sup> and Low Complexity Regions (LCR)<sup>36</sup>. They can help to determine novel micro-domains and to identify the effective role for those segments whose function is still mostly unknown.

The atomic-level characterization of IDPs has thus become a central topic, especially for the development of new drugs capable of interfering with them<sup>37</sup>. The obtainment of atom-resolved information on these biomolecules is quite challenging because of the absence of a structure and the repetitive nature of the primary sequence.

## **NMR: the golden technique**

Solution NMR is the most appropriate spectroscopic technique, if not the unique one, to obtain structural and dynamic information at atomic resolution, both *in-vitro* and *in-cell*.

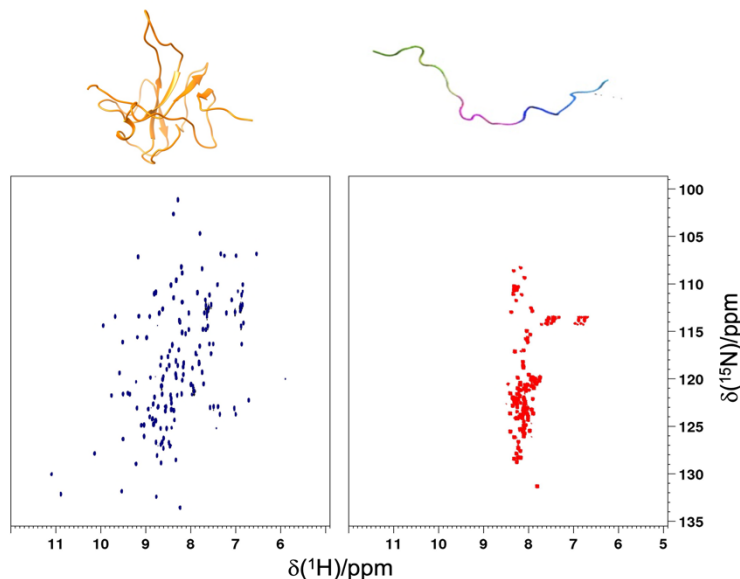
We are in a mature NMR era in which the instrumentation continuously becomes more sophisticated and allows us to perform fast multidimensional experiments or detect heteronuclei with excellent sensitivity. Thanks to these developments and to the design of novel NMR experiments, we are able to see proteins in their cellular

environment and examine the role of specific molecules in their cellular pathway.

The routine NMR experiments used for structure determination, such as 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC and 2D  $^1\text{H}$ - $^1\text{H}$  NOESY, were designed for folded proteins but the particular dynamic nature of IDPs/IDRs makes these experiments less informative. The main reason derives from the properties of the highly flexible, solvent-exposed surface that have a strong impact on NMR parameters. These cause, as a first consequence a moderate chemical shift dispersion, in particular for  $^1\text{H}$  nuclear spins (Figure 2).

The narrow range of chemical shift causes extensive signal overlap and thus complicates the spectra. The lack of a defined structure is also reflected in the decrease of the magnitude of 2D  $^1\text{H}$ - $^1\text{H}$  NOEs and amide protons experience efficient exchange processes with the solvent that can result in the loss of many signals. In this framework,  $^{13}\text{C}$  direct detection can overcome the problems and catch information even in these borderline conditions thanks to the non-exchangeable nuclei involved in the experiments.

Carbon chemical shifts are also well dispersed in a large spectral width, covering a large range of frequencies that becomes useful to resolve signals belonging from flexible domains.



**Figure 2.** Comparison of two  $^1\text{H}$ - $^{15}\text{N}$  spectra acquired on the globular N-terminal Domain of the N protein from SARS-CoV 2 (left) and the intrinsically disordered protein  $\alpha$ -synuclein (right). While the signals of the folded protein are well dispersed, the crosspeaks from the IDP are crowded and centered around 8-8.5 ppm .

Together with  $^{13}\text{C}$  direct detection, new horizons of investigation on IDPs are now accessible. Thanks to the recent improvements on NMR hardware ultra high-field magnets have recently become available, including the pioneer 1200 MHz (28.2 T) magnet installed at the Magnetic Resonance Center (CERM) of Florence. The combination with dedicated cryoprobes makes this instrument an incredible resource for the study of IDPs and complex modular proteins because the increased magnetic field allows to strongly augment resolution.

It is worth noting that also Transverse Relaxation Optimized Spectroscopy (TROSY)<sup>38</sup>, exploited for  $^1\text{H}$ -detection, benefits by the use of high-fields allowing to extend the Molecular Weight (MW) of systems that can be investigated with solution NMR.

Field dependent effects can be investigated to identify the challenges and opportunities provided by increasingly high magnetic fields. This shows the versatility and richness of information that NMR spectroscopy can reveal on complex macromolecules.

# CHAPTER 1

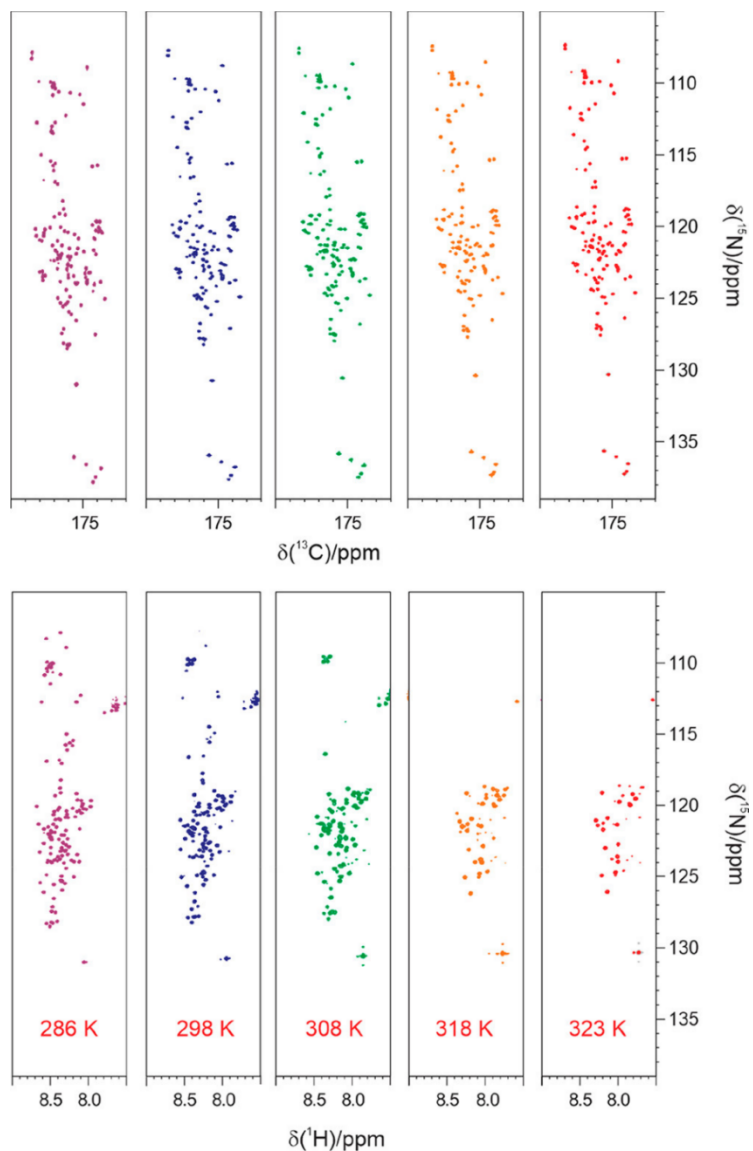
## Dealing with physiological conditions

### 1.1 Fingerprints of IDPs at physiological conditions

IDPs/IDRs are inherently sensitive to experimental conditions as their multi-functionality is strongly context dependent. Even a small modulation of the experimental conditions can alter their behavior. For this reason it is important to study their action *in-vitro* mimicking the physiological milieu in terms of salt, pH, temperature (T), and molecular crowding in order to obtain more realistic information on their physiological asset. High exchange rates of H<sup>N</sup> ( $k_{ex} \gg 100s^{-1}$ ) become a critical point when studying IDPs by NMR, especially when investigating their performance at model physiological conditions (pH 7.4, 310 K). <sup>13</sup>C direct detection is the perfect strategy to overcome this problem<sup>39</sup>. The approach enables to study the features of side-chains, which are expected to play a central role for the function of many IDPs. They are seldom studied because of the increased extensive resonances overlap of the NMR signals of their nuclei, as anticipated in the introduction. In particular, the 2D <sup>13</sup>C-<sup>15</sup>N CON experiment (here simply referred to as CON) is a fingerprint spectrum that shows signals belonging to the peptide bond nuclei (C'<sub>i</sub>-N<sub>i+1</sub>) and to Glutamine and

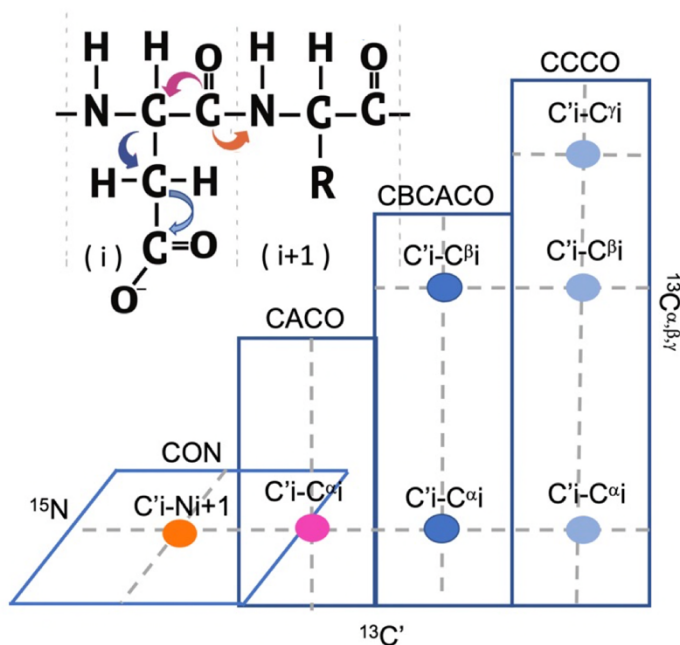
Asparagine side chain nuclei. The high chemical shift dispersion and the intrinsic features of the involved nuclei make CON an optimal alternative to 2D  $^1\text{H}^{15}\text{N}$ -HSQC (referred to as HSQC) when investigating IDPs/IDRs. The advantages are clearly visible comparing the two types of spectra as shown in Figure 3.

The set of spectra, acquired on  $\alpha$ -synuclein, shows the loss of  $\text{H}^{\text{N}}$  resonances while going up with T while CON maintains all the peaks. In addition, CON allows to obtain signals for proline residues<sup>40</sup>, not easily studied by  $\text{H}^{\text{N}}$  detection as they lack the amide hydrogen. Along the same lines,  $^{13}\text{C}$  CACO, CBCACO, and CCCO<sup>41-43</sup> provide cross-peaks belonging to  $\text{C}^{\alpha}$ ,  $\text{C}^{\beta}$ ,  $\text{C}^{\delta}$  and  $\text{C}^{\gamma}$  nuclei extending the information to the side-chains.



**Figure 3.** The spectra on the top are a series of CON spectra acquired at different temperatures (from 286 K to 323 K). The spectra on the bottom are a series of HN-HSQC spectra acquired on the same sample in the same experimental conditions. It is clearly visible how all the CON resonances are maintained while varying T with the respect of those from HSQC where many of the signals are lost already at 308 K. Picture from reference 39.

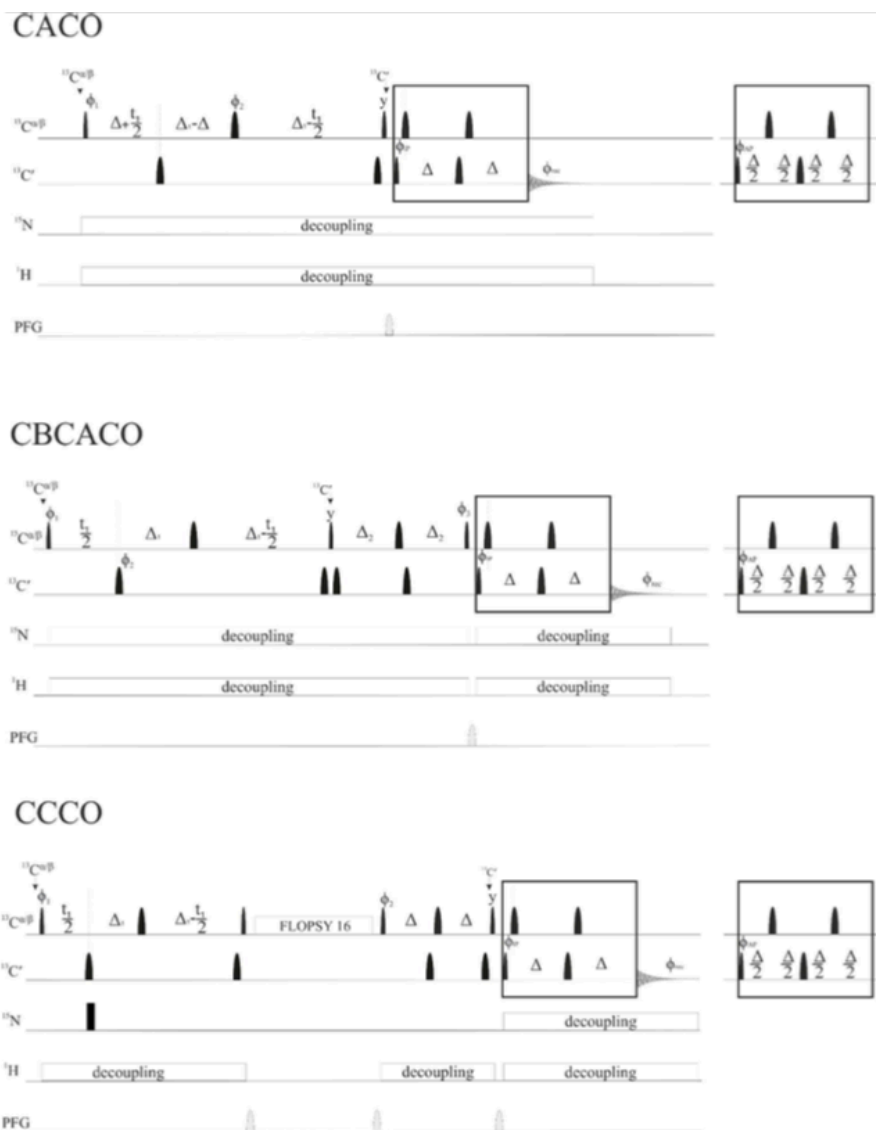
Figure 4 shows in a schematic way the magnetization pathway.



**Figure 4.** The magnetization pathway of the proposed  $^{13}\text{C}$  experiments with a scheme of the specific assignment of carbon nuclei.

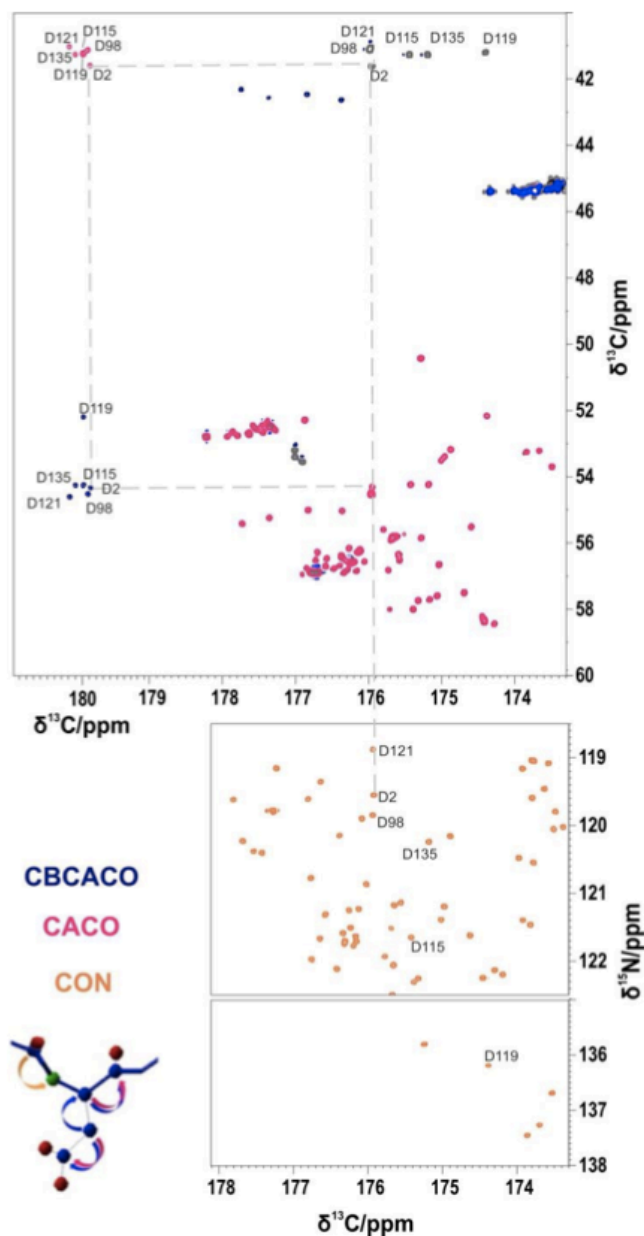
To obtain a complete fingerprint of an IDP when approaching high pH and T, a toolbox made of 2D  $^{13}\text{C}$ -direct detected experiments was created (Figure 5). The H-start versions of each experiment were also implemented. This escamotage allows increasing sensitivity, when needed as for low concentrated samples, thanks to the use of the non-exchangeable  $\text{H}^{\alpha}$  nucleus as a starting source of polarization.





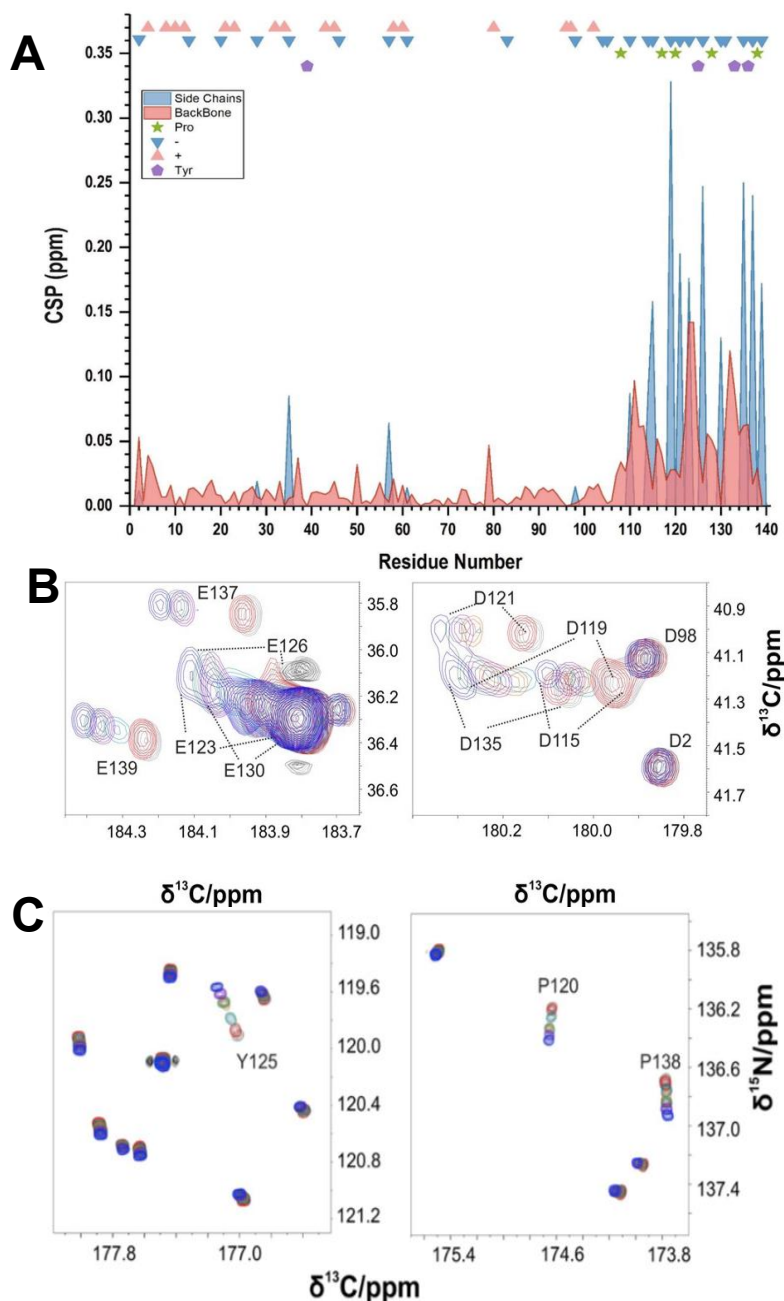
**Figure 5.** The pulse sequences used to acquire CACO, CBCACO, CCCO spectra. Narrow and wide black bars represent  $\pi/2$  and  $\pi$  non-selective pulses; narrow and wide rounded black bars represent  $\pi/2$  and  $\pi$  band-selective pulses. The pulse sequence elements reported in the boxes represent the two variants to acquire the in-phase (IP) and antiphase (AP) components of carbonyl signals needed to achieve  $^{13}\text{C}$  homonuclear decoupling through the IPAP approach<sup>42</sup>. Experimental details and H-start versions are provided in section the Supplementary Material of Article 1.1

The acquisition of the complete set of experiments allows us to assign the carboxylate and carbonyl side-chains (-CCOOH, -CCONH<sub>2</sub>) starting from CON spectrum. Figure 6, as an example, shows the sequential assignment of carbon resonances from Aspartate (D) of the natively disordered  $\alpha$ -synuclein. Moving from the backbone carbonyl C'<sub>i</sub> signal of a selected amino-acid (D2, in this case), this can be easily identified in the CACO and CBCACO spectra, correlating the connected C <sup>$\alpha$</sup> <sub>i</sub> and C <sup>$\beta$</sup> <sub>i</sub> nuclei. The latter are correlated to C <sup>$\gamma$</sup> <sub>i</sub> giving cross-peaks that belong to the carboxylate functional group and they fall in very isolated regions of the spectra. The same is possible for Asparagine (N) residues. Glutamate (E) and Glutamine (Q) residues need the additional CCCO spectrum to overcome ambiguities in the assignment of aliphatic C <sup>$\alpha$</sup> , C <sup>$\beta$</sup>  and C <sup>$\delta$</sup>  and the carbonyl C <sup>$\gamma$</sup> . In this way, a complete carbon-based assignment that includes side-chains can be achieved while adapting the deposited assignments to the new experimental conditions.



**Figure 6:** An illustration of the strategy used to obtain the sequence-specific assignment of the  $^{13}\text{C}'$  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),  $\text{C}^{\alpha}_i$  and  $\text{C}^{\beta}_i$  are identified in CACO (pink) and CBCACO (blue) spectra, superimposed in the Figure, and correlated to  $\text{C}^{\gamma}_i$  through the respective  $\text{C}^{\beta}_i - \text{C}^{\gamma}_i$  and  $\text{C}^{\alpha}_i - \text{C}^{\gamma}_i$  cross-peaks in a sequence-specific manner.

The set was used to follow a titration between  $\alpha$ -synuclein and calcium ions at model physiological conditions to monitor the effect of Aspartate and Glutamate side-chains, very abundant in the C-terminal domain (Figure 7), on the interaction.  $\alpha$ -synuclein is a natively disordered protein linked to Parkinson's disease and it is located close to the presynaptic terminus, where the micro-domains of high  $\text{Ca}^{2+}$  concentrations are linked to the release of neurotransmitters<sup>44</sup>. Some studies had explored the calcium-dependent behaviour of  $\alpha$ -synuclein like its secretion, both *in-vitro* and *in-vivo*, and the increased affinity for Small Unilamellar Vesicles (SUV)<sup>45-47</sup>. The interplay between metal ions and IDPs is crucial for the triggering of many IDPs dysfunctions linked to neurodegenerative diseases<sup>48,49</sup>. However this statement often remains speculative as it is difficult to determine the driving force of these interactions because of their "fuzziness". In addition, metal-dependent proteins are often characterized by a distinct amino-acidic pattern with well-defined interaction sites (i.e. calbindin) while there are still not many studies on possible flexible binding sites present in IDPs. The presented methodology can therefore provide a valuable tool of investigation to tackle these complicated studies.

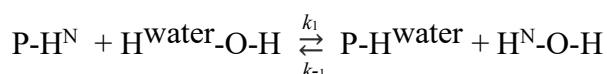


**Figure 7:** Chemical shift perturbation (difference in absolute value) of  $^{13}\text{C}$  resonances of  $\alpha$ -synuclein upon addition of a  $\text{Ca}^{2+}$  solution to the sample. Backbone carbonyl  $^{13}\text{C}$  values are plotted in red and Aspartate and Glutamate side-chain  $^{13}\text{C}$  are blue (panel A). The symbols over the graph depict the distribution of charged, tyrosine and proline residues to highlight the particular composition along the primary sequence: Aspartate and Glutamate (blue triangles), Lysine (red triangles), Proline (stars) and Tyrosine (pentagons). The lower panels show two regions of the CACO and CON spectra with cross-peaks of Aspartate and Glutamate side chains (B) and Tyrosine and Proline (C) residues within their shifts during the titration.

## 1.2 Spying chemical exchange with water

The study of amide proton exchange is essential for protein chemistry because it can give access to a lot of information about dynamic processes and structure. The first approach was conceived and implemented by Kaj Linderstrøm-Lang after Pauling's discovery of the secondary structure elements that he postulated to be stabilized by hydrogen bonds<sup>50</sup>. Lang realized that peptide H<sup>N</sup> hydrogens participate in a continuous exchange with the hydrogens of solvent and conceived the hydrogen-deuterium exchange to demonstrate Pauling's hypothesis<sup>51</sup>.

The kinetic of exchangeable protons with water can be treated as a pseudo-first order reaction:



Where P-H<sup>N</sup> indicates the protein's exposed amide hydrogen and H<sup>water-O-H</sup> the exchangeable hydrogen from the water molecules.

The rate of the reaction is

$$k_1[\text{P-H}^{\text{N}}][\text{H}^{\text{water-O-H}}] + k_{-1}[\text{P-H}^{\text{water}}][\text{H}^{\text{N-O-H}}]$$

Two consistent approximations can be applied.

Since the water concentration is higher than protein concentration, it is possible to ignore the reverse reaction:

$$[\text{P-H}^{\text{water}}][\text{H}^{\text{N-O-H}}] \ll [\text{P-H}^{\text{N}}][\text{H}^{\text{water-O-H}}]$$

Then, it is assumed the exchange constant is equal to

$$k_1[\text{H}^{\text{water-O-H}}] = k_{\text{ex}}$$

then the amide proton concentration is defined as

$$[P-H^N] = [P-H^N]_0 e^{-(k_{ex}t)}$$

And, as anticipated, the reaction follows a pseudo-first order kinetic rate rule

$$k_1 [H^{water-O-H}] [P-H^N] = k_{ex} [P-H^N]$$

Many HX-exchange methods are NMR-based and they are designed to measure different regimes.

Looking at the literature, it emerges that CLEANEX is the mostly used NMR experiment to determine solvent exchange data<sup>52</sup>.

It is based on the selective perturbation of water magnetization. The CLEANEX approach is implemented into 2D-HSQC based pulse sequence and it allows to avoid artefacts due to NOEs and ROEs with accurate quantification of the water-exchange constant.

However the difficult applicability at physiological-like conditions persists because of the very fast exchange kinetic of the exchange process.

In 2008, Segawa et al. published a method for the detection of invisible protons, such as those from Lysine (-RCNH<sub>3</sub>) and Arginine (-RCONH<sub>2</sub>) residues side chains, by exploiting <sup>13</sup>C-<sup>15</sup>N isotopically labelled proteins<sup>53</sup>. The idea is based on the measurement of the effect of scalar relaxation caused by the exchanging amide protons in the presence and absence of H-decoupling, measuring the decay of <sup>15</sup>N coherence under a refocusing Carr-Purcell-Meiboom-Gill (CPMG) pulse train.

A <sup>13</sup>C-detected alternative is suggested by Thakur et al. in 2013<sup>54</sup>. The

authors proposed the use of 2D-CON as a basis to measure a wide range of hydrogen-deuterium  $k_{ex}$  with the use of different approaches depending on the established regime. For fast regimes, they propose to monitor heteronuclear longitudinal two-spin order ( $2C'zNz$ ) exchange between CO-NH and CO-ND species as a function of time.

An alternative approach lies on the concept of “decorrelation” spectroscopy, firstly proposed by Skrynnikov and Ernst in 1999<sup>55</sup>. They suggested the possibility to quantify water- $H^N$  exchange with the measurement of the decay to equilibrium of the two-spin order operator  $2N_zH_z$  created during a  $^1H$   $^{15}N$  HMQC experiment. Along these lines, a novel CON version (namely DeCON) was firstly designed in order to achieve a  $k_{ex}$  measurement in an indirect way while exploiting the advantages of  $^{13}C$  detection.

The introduction of an additional building block into the CON pulse sequence allows to create a triple-spin order operator  $4C_zN_zH_z$  whose decay is monitored in different time frames ( $\tau_{decor}$ ).

Fitting the integrated peaks for each spectrum with a mono-exponential decay function, a  $k_{ex}$  ( $k_{zzz}$ ) value is obtained.

$$I_{zzz}(\tau_{decor}) = I_0 e^{-k_{zzz}\tau_{decor}}$$

In addition, the use of selective pulses allows to excite only  $H^N$  nuclei without perturbing water magnetization, in order to avoid radiation damping effect<sup>56</sup>. When pulsing on  $^1H$  magnetization with a radio frequency pulse, the water bulk magnetization creates an induced field  $B_1$  that rotates the magnetization of the solvent spins to its equilibrium, before any other relaxation mechanisms. As the water magnetization needs a long recovery time  $T_1$  to go back to equilibrium, it creates an



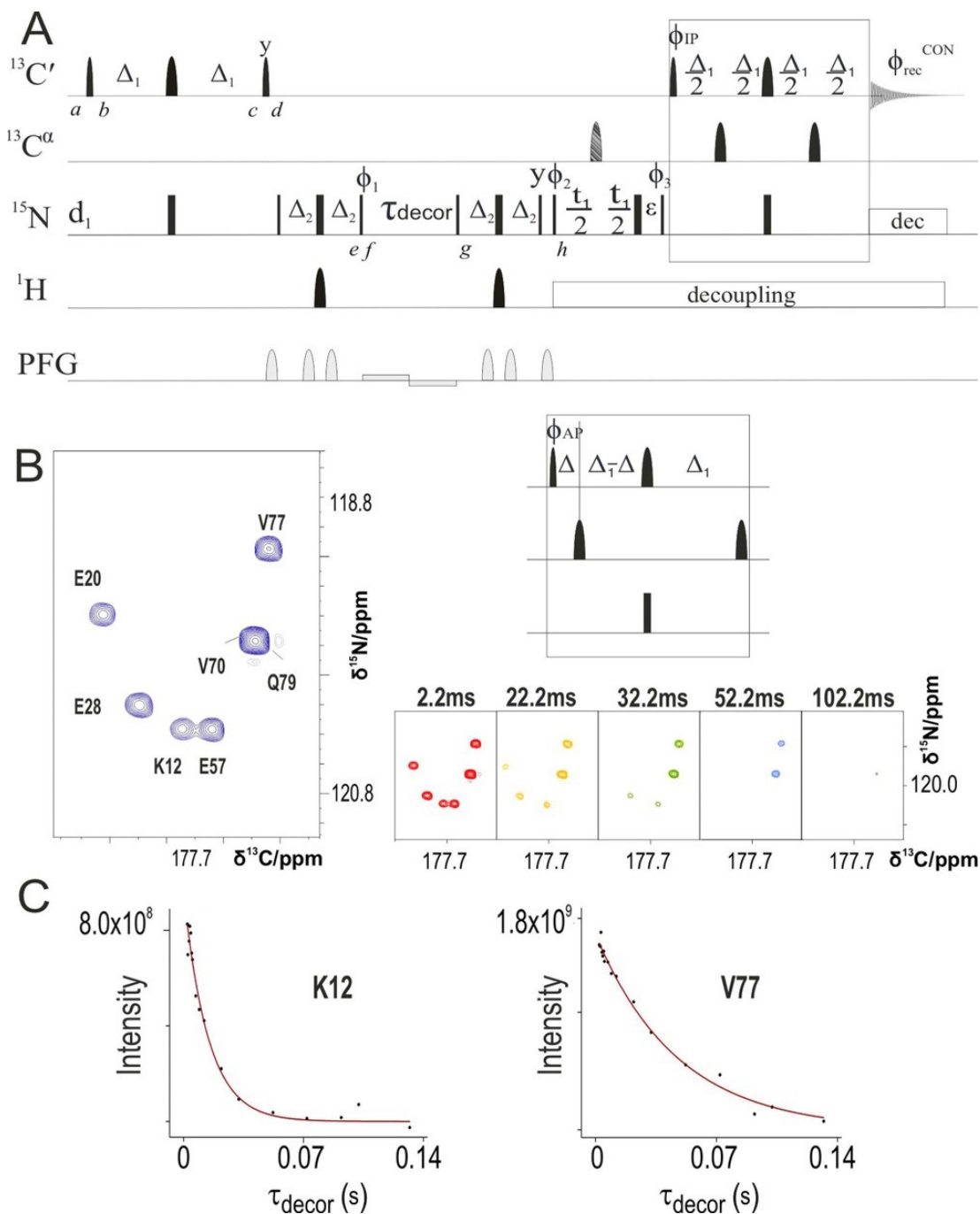
induced current in the coil that broadens the solute's  $^1\text{H}$  signals. The effect increases with the magnetic field because the higher is  $B_0$ , more effective will be the induced field  $B_1$ . For this reason, the recovery delay  $d_1$  is sometimes adjusted considering the radiation damping time in the selected experimental conditions.

The experiment was validated on  $\alpha$ -synuclein and then used to study the  $\alpha$ -synuclein-calcium ions interplay.

The DeCON was acquired in absence and presence of  $\text{Ca}^{2+}$  in pure  $\text{H}_2\text{O}$  solvent at 310 K, pH was adjusted to 7.4.

A 3D version of the experiment was also developed to increase the resolution when needed.

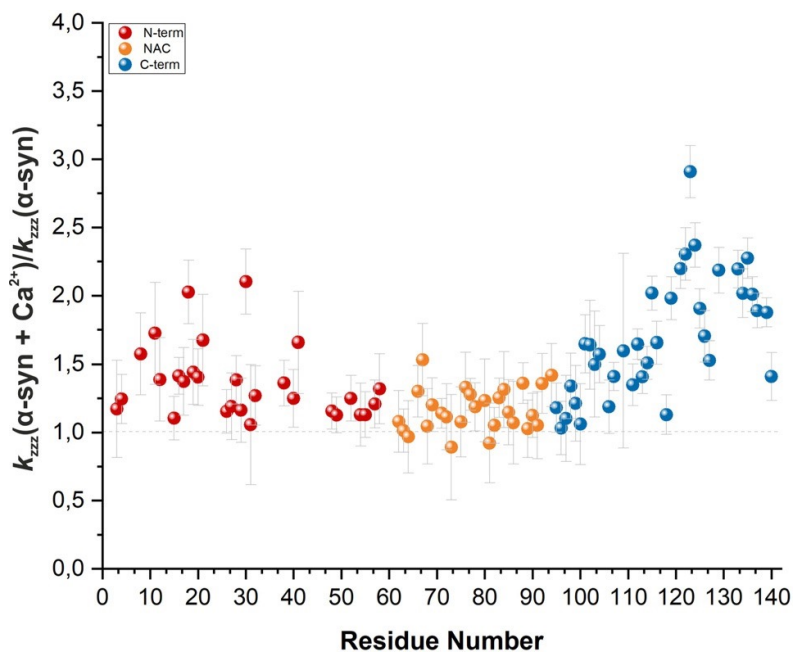
The pulse sequence and an example of the strategy used to extrapolate the  $k_{zzz}$  is presented in Figure 8.



**Figure 8** A) DeCON pulse sequence. 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  $\tau_{\text{decor}}$  are shown on the right. C) The intensities (arbitrary units) of two of these cross-peaks are reported as a function of  $\tau_{\text{decor}}$ .

### 1.3 $^{13}\text{C}$ toolbox highlights novel insights on $\alpha$ -synuclein

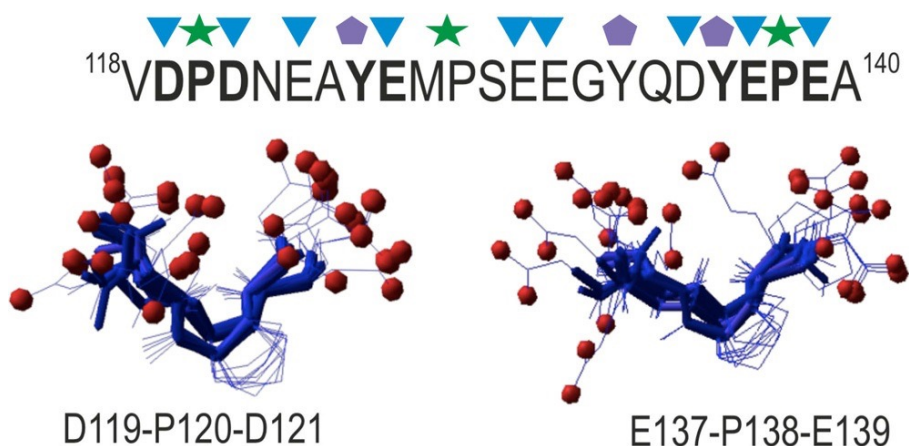
The application of the proposed experiment revealed novel insights on  $\alpha$ -synuclein. It emerged that its C-terminal part, rich of acidic residues, shows reduced exchange rates. The addition of  $\text{Ca}^{2+}$  causes an increase of the  $k_{\text{zzz}}$  (Figure 9). This could be due to a local change in electrostatic potential or to a local change in solvent accessibility (or both the effects), revealing how the interpretation of such effects demands more extensive investigation through the proposed approach.



**Figure 9.** The ratios between  $k_{\text{zzz}}$  measured in presence and absence of Calcium ions at pH 7.4, 310 K on a  $200\mu\text{M}$  sample of  $\alpha$ -synuclein in pure  $\text{H}_2\text{O}$  solvent. The values are plotted versus the residue number. The different colors evidence the three regions identified along the primary sequence of the protein: the N-terminus (red), the Non-Amyloidal Component (NAC, orange) and C-terminal tail (blue). The residues from this latter portion experience the stronger increase of  $k_{\text{zzz}}$  together with some amino-acids belonging from N-terminus.

The interaction is likely to disrupt some long-range contacts between the C-terminal region and N-terminus, resulting in the loss of compaction that is translated in an augmented exchange rate. In addition, the metal ions selectively affect not all the negatively charged amino-acids but only some specific residues, including two prolines of the total five and two tyrosines, all located in the C-tail.

Two distinct motifs emerged: a proline flanked by two acidic residues (DPD, EPE) and two tyrosine-glutamate pairs (YE), as presented in Figure 10.



**Figure 10.** Some possible models of the DPD (left) and EPE (right) motifs identified during the  $^{13}\text{C}$  experiments. The primary sequence where the most perturbed residues are located is represented on the top. The symbols identified the negative residues aspartate and glutamate (blue triangles), the disorder promoting proline (green stars) and the hydrophobic tyrosine (purple pentagon) in order to depict their distribution along the stretch.

While prolines could reduce local mobility and favor the proper relative orientation of negatively charged side chains for calcium binding, the bulky hydrophobic tyrosine could play a relevant role in reducing local motions and favoring interactions of highly flexible protein regions with  $\text{Ca}^{2+}$ , in particular if followed by an acidic  $\text{COO}^-$  group. These patterns

can be exploited as ion binding sites for fuzzy interactions, such as the interplay of  $\alpha$ -synuclein-calcium ions, and act as sensor motifs. It is important to stress on the fact that all the results were achieved at model physiological conditions thanks to the possibility of observing side-chains.

Related article:

**1.1. Article: Monitoring the Interaction of  $\alpha$ -Synuclein with Calcium Ions through Exclusively Heteronuclear Nuclear Magnetic Resonance Experiments**

# CHAPTER 2

## Viral modular proteins: the case of the N protein from SARS-CoV-2

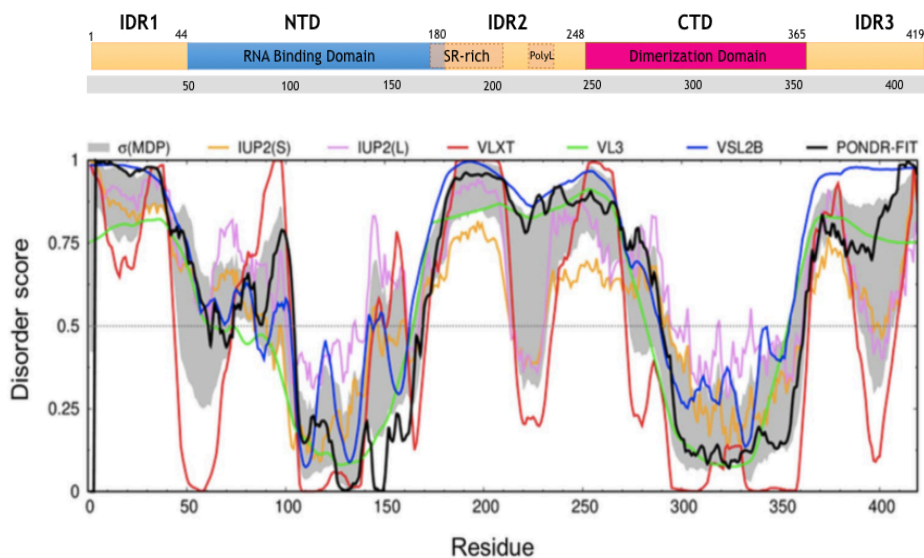
### 2.1 THE NUCLEOCAPSID PROTEIN

Viruses strategically exploit the structural disorder of their proteome to hijack host cells and to overcome many evolutionary obstacles. For instance, it has recently been proposed that the role of disorder in the protective shells of a virus is related to its resistance and transmissibility<sup>57,58</sup>. The presence and variation of the disorder in viruses appears to be an escamotage by which the pathogens modulate various functions of infected cells, including innate or acquired immunity<sup>59,60</sup>. The necessity to investigate viral disordered systems more in detail was denoted by several studies on different viruses such as HIV-1<sup>61</sup> and Human papilloma virus<sup>62</sup>. With the recent pandemic, we felt the urgency to focus on this field of research in order to delineate novel strategies to target viral proteins.

Many of these viral proteins present a modular organization to accomplish the different functions they need to perform during the infection and replication cycle<sup>63</sup>. The Nucleocapsid protein from SARS-

CoV 2 (N) is a brilliant example to show the versatility given by the modular-arrangement of domains. N is a structurally heterogeneous, 419 amino-acid-long, multi-domain protein that is found inside the viral envelope. It is one of the four structural proteins of the virus, crucial for the infection cycle, and the most expressed one upon viral infection. The main proposed role is to arrange the architecture of the RiboNucleoProtein (RNP) complex and to orchestrate PPis with the host proteome<sup>64–66</sup>.

N is organized in an RNA binding N-terminal domain (NTD), a dimerization C-terminal domain (CTD) and three intrinsically disordered regions (namely IDR1, IDR2, and IDR3) that comprise almost 40% of the protein primary sequence (Figure 11).



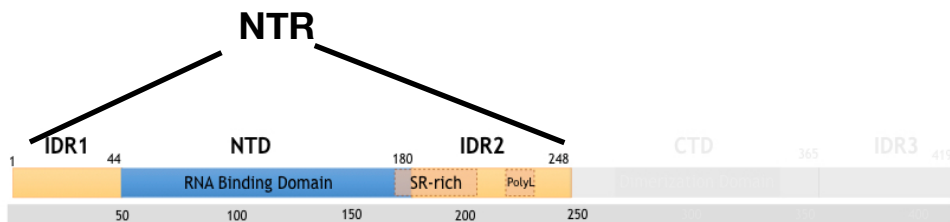
**Figure 11.** On the top, the sequence diagram of the N protein. On the bottom, a plot of the disorder score predicted using different dedicated algorithms (<http://www.pondr.com>).

Since the studies conducted on SARS, it emerged the IDRs are pivotal

to drive many mechanisms related to the N function<sup>66</sup>. However, their disordered nature complicates a detailed characterization of the molecular architecture because of the high dynamics and the motley behavior conferred by the high flexibility and disorder.

## 2.2 “Divide and conquer”: the NTR construct

One of the most used strategies to characterize a modular protein is to dissect its structure into single domains. However, this can lead to the loss of cooperation between some of these portions, in particular when IDRs are involved. A first evidence of the importance of the disordered regions of N emerged in a study conducted in 2009 by Chang et al. on the homologous protein from SARS CoV<sup>64</sup>. The authors studied the binding of a polynucleotide with several constructs, including the NTD alone, the 182-365 and the 45-365. The presence of the IDRs increased the apparent  $k_d$  together with the apparent Hill coefficient which is an indicator of the binding sites involved in a cooperative interaction. This consideration suggested us to start the characterization of N from an N-terminal region (1-248 construct, namely NTR) that includes IDR1, NTD and IDR2.



**Figure 12.** The sequence diagram of the NTR construct (1-248).

Model predictions of NTR were obtained using D-I-TASSER (Figure

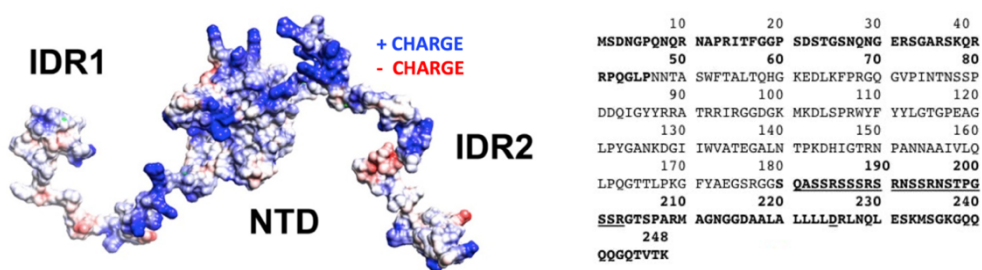


13), a high-accuracy protein structure and function predictor that exploits deep neural-network, deep-learning, and Monte Carlo simulations to obtain structure models and biological function predictions from the primary sequence of a protein. The protocol is based on the ab-initio folding of proteins lacking homology in the PDB.



**Figure 13.** Five models of the NTR and their superimposition obtained with D-I-TASSER.

A protocol for the recombinant protein expression was initially designed.  $^{13}\text{C}$ - $^{15}\text{N}$  NTR was expressed in *E.coli* (DE3) strain using a method that allows to increase the yield of isotopically labelled proteins while using low amounts of the necessary reagents<sup>67</sup>. For the purification, the high positive charge of the construct was exploited (pI= 10.15). The NTR, indeed, is rich of arginine residues, mostly located in the so-called SR-region, a portion of IDR2 targetable by PTMs<sup>68</sup> (Figure 14).



**Figure 14.** Electrostatic surface map of the NTR and its primary sequence. IDRs are evidenced in bold, the SR-region portion is underlined.

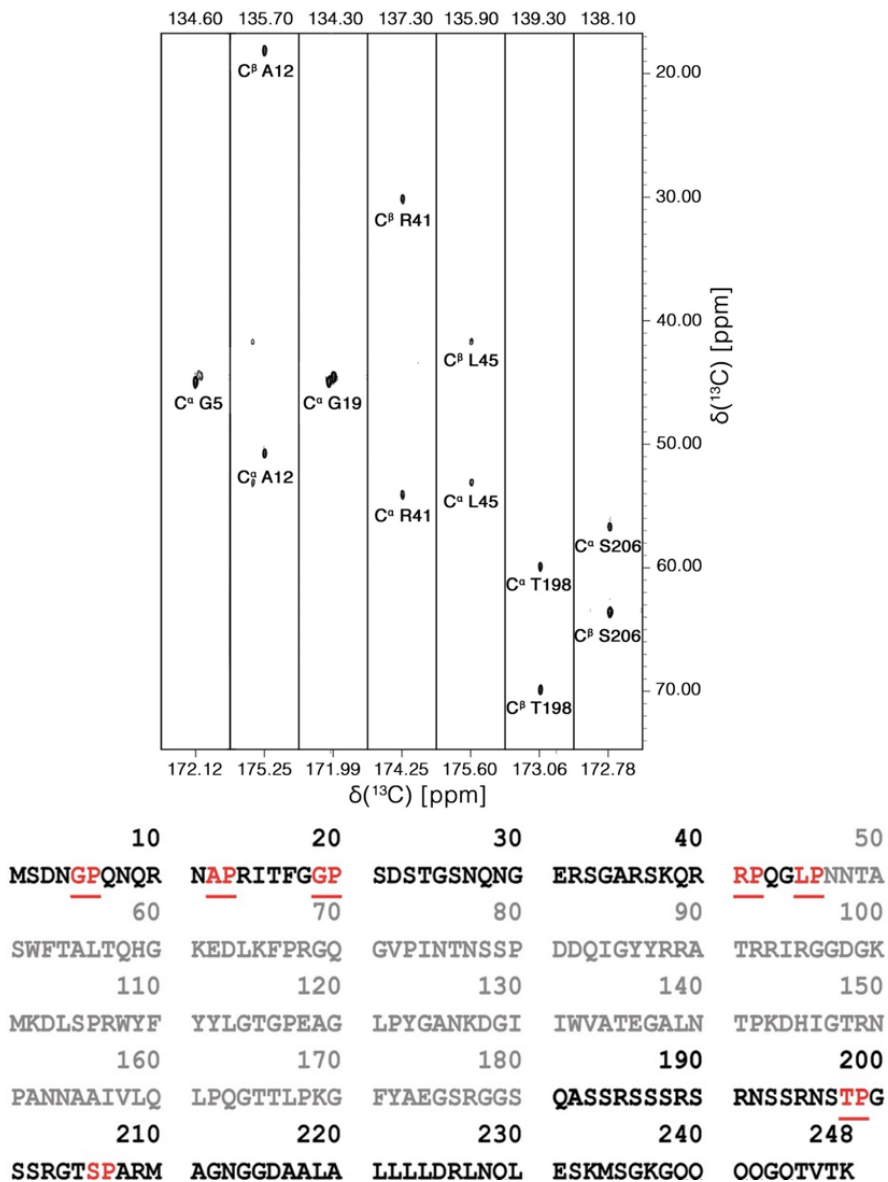
Ion Exchange Chromatography (IEC) was used as first step of purification to sequester the protein from the cellular content while Size Exclusion Chromatography (SEC) was used to obtain very pure samples needed to test protein's interactions. The strategy allows to obtain a "wild type" protein construct that can be used for structural biology applications ( Article 2.1).

The protein has limited stability since IDPs are susceptible to proteolysis. Many hints suggest an auto-cleavage mechanism<sup>69</sup>, firstly observed in SARS homologous protein both *in vivo* and *in vitro*<sup>70-72</sup>, that can explain the inefficacy of the addition of proteases' inhibitors used to prepare NMR samples. In parallel, another problem related to

the protein preparation regards its tendency to aggregate when the concentration exceeds 300  $\mu\text{M}$ , even at high ionic strength. The presence of high salt concentration is crucial to maintain the protein solubility and stability.

We opted for the  $^1\text{H}$ -start,  $^{13}\text{C}$  detected experiments. This approach is used to increase the sensitivity of  $^{13}\text{C}$  detected experiments, able to provide well resolved spectra for the IDRs signals, and at the same time reduce the experiment duration. This strategy allowed us to firstly characterize the NTR with increased resolution, providing experimental information about many of the proposed interaction sites. The large number of arginine, serine, glutamine, and glycine residues, very abundant in the two IDRs, could be detected and most of them were resolved. Several resonances in low complexity regions, such as the polyGlutamine stretch (238–242), were also resolved allowing their high-resolution investigation.

As previously discussed, exploiting carbon-detection gives the possibility to detect the proline residues. In this context, the related resonances were used as a starting point for sequence-specific assignment. The unambiguous identification of the  $X_{i-1}\text{-Pro}_i$  pairs was achieved thanks to the peculiar  $^{15}\text{N}$  chemical shifts of the proline nitrogen resonance that is correlated to the  $\text{C}'$ ,  $\text{C}^\alpha$ ,  $\text{C}^\beta$  of the preceding amino-acid through the 3D (H)CBCACON. The related spectrum strips are presented in Figure 15.



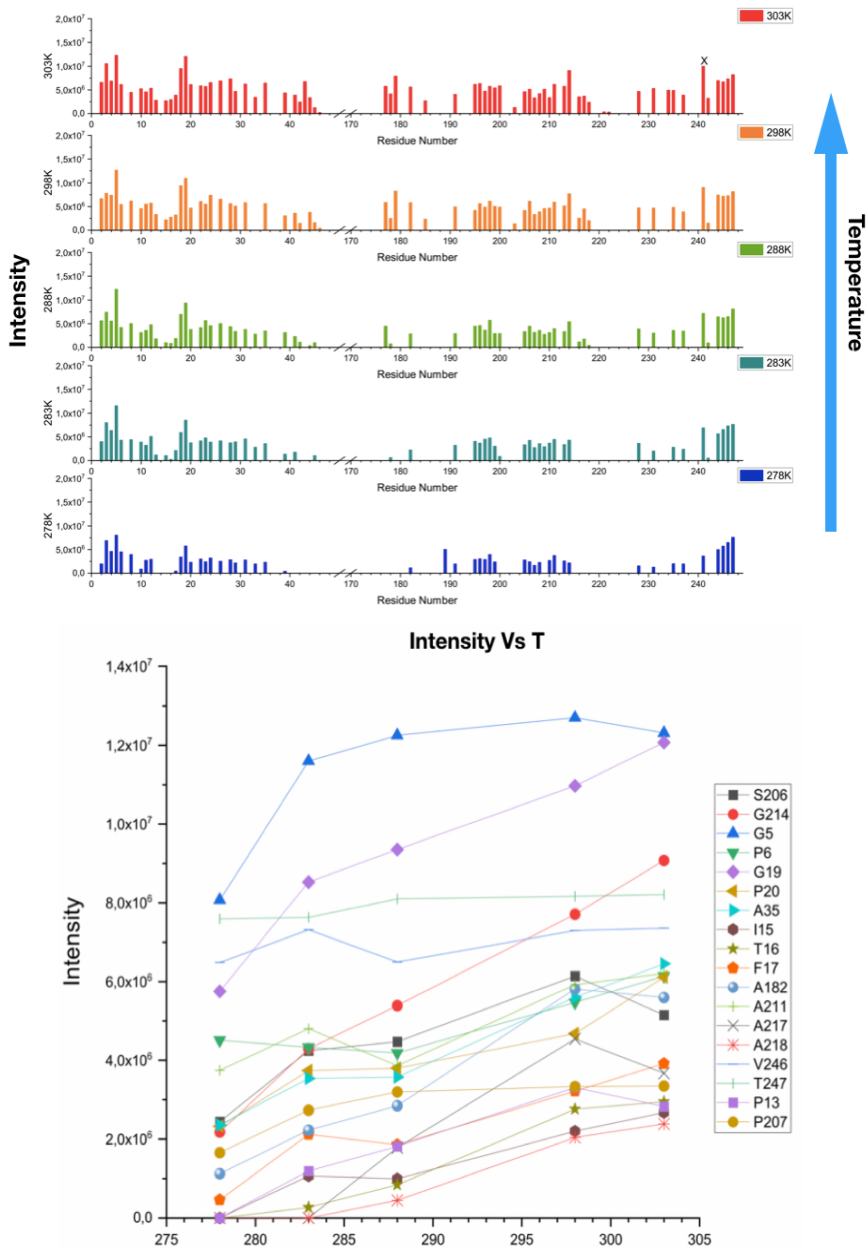
**Figure 15.** The 3D (H)CBCACON strips extracted of the  $^{15}\text{N}$ -chemical shift of proline residues, used to assign X-Pro pairs. Carbon shifts belong to the preceding X residue. The pairs are highlighted in red along the primary sequence. The IDRs are in bold while the folded domain (NTD) is reported in gray.

$C^\alpha$  and  $C^\beta$  resonances enable the identification of glycine, alanine, serine, and threonine residues the remaining X-Pro pairs can be easily identified as deriving from leucine and arginine residues by comparison with the primary sequence of the protein. A 95% complete assignment was achieved in 25 mM Tris-(hydroxymethyl)-aminomethan (TRIS), 450 mM NaCl, pH 6.5, 298 K by using NMR spectrometers operating at 700 MHz (16.4 T), 950 MHz (22.3 T) and 1200 MHz (28.2 T). It was deposited in the Biological Magnetic Resonance Data Bank (BMRB) with the entry numbers 50618 and 50619 (Article 2.2).

The use of 2D CACO at ultra-high field instrumentation also provided some pivotal details on the complex dynamics of the NTR. The  $C^\alpha$ - $C'$  bond is influenced by structure arrangements induced by the variation of the dihedral angle ( $\Psi$  and  $\phi$ ) so it reflects the variation of the backbone conformation.

The experiments were acquired at different temperature values using a 75  $\mu$ M sample in 25 mM TRIS and 450 mM NaCl, pH 7.2

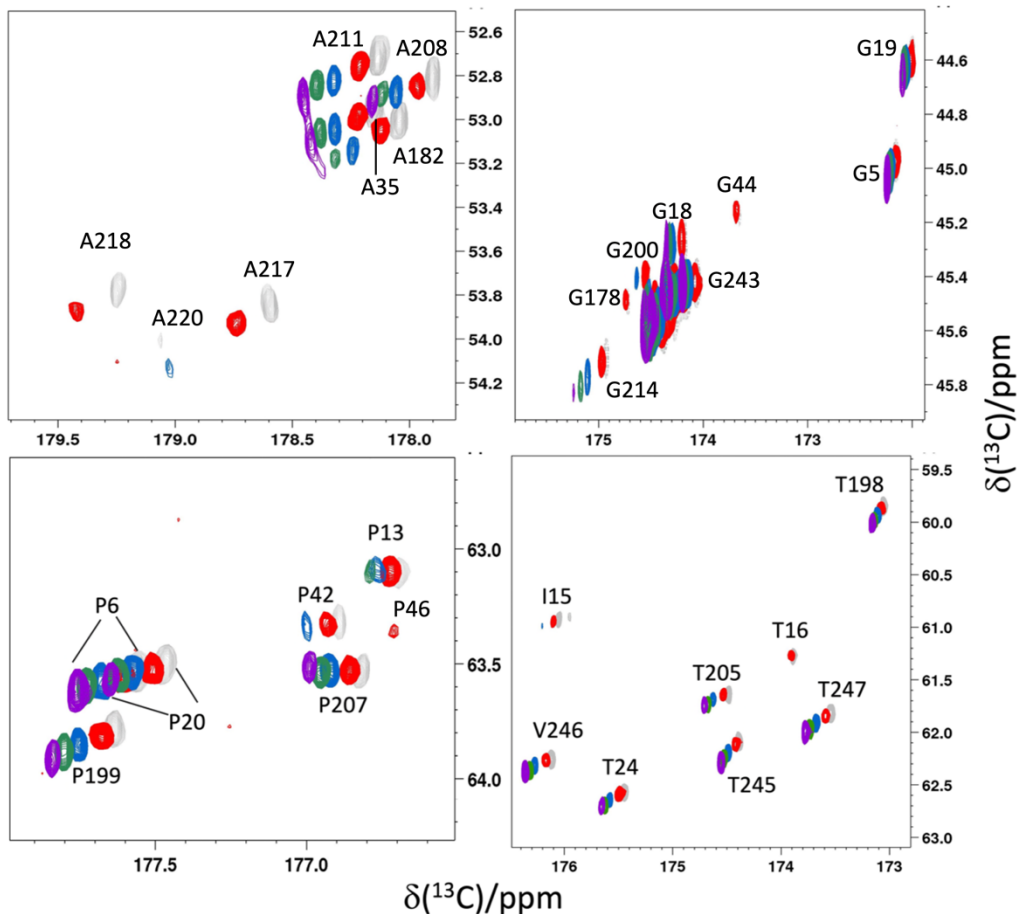
Reducing the temperature, the intensities decrease for all the residues, as shown in Figure 16.



**Figure 16.** On the top, the signals intensity vs residue number plots. They represent the variation of the CACO intensities while varying the T from 303 K (red) to 278 K (blue). On the bottom, the single residues intensity vs T graph, obtained with the variation of T. While some residues maintain their intensities (G5, A211), others (G214, A217) are more affected by the T decrease. The trend is related to the residue location along the IDRs.

The overall view suggests that there are additional portions of NTR that are not as flexible as expected. Indeed, the region comprising residues 13-19 ( $^{13}\text{PRITFGGP}^{19}$ ) disappears when reducing the temperature to 278 K, supporting the idea of a transient structure. This is characterized through different algorithms (see Figure 11) and suggested by the D-I-TASSER structural analysis. The proline residues, that fall in an isolated portion of the spectrum, are again pivotal residues that can be used as “structure sensors”. As visible in Figure 17, while most of the intensities are maintained, only residues from the 13-19 portion experience a strong decrease of their intensities suggesting a stiffening of this part, in line with the variation of the protein rotational correlation time  $\tau_r$ . It is worth noting that this stretch is limited by two proline residues (P13 and P19) and contains the only aromatic residue present in IDR1 (F17). The same trend is also shown by the leucine and alanine residues that belong to the polyLeucine region  $^{217}\text{AALALLL}^{224}$  (helical propensity).

The 2D CACO proves to be a simple alternative for a qualitative description of the a dynamic profile of a complex protein, in particular for challenging experimental conditions such as low protein concentration and high salt concentration, pH and T.



**Figure 17.** Four regions extracted from the CACO spectra acquired on NTR at different T: 303 K (gray), 298 K (red), 288 K (blue), 283 K (green), 278 K (purple). While most of the resonances present only chemical shift perturbation, some specific residues are perturbed both in chemical shift and intensity. This trend is particularly pronounced for the residues belonging to polyLeucine motif but it is also present for residues in the 13-19 tract suggesting this part has the tendency to be structured lowering T.



## 2.3 The promiscuous interactions of NTR

N, as various nucleocapsid proteins, is supposed to be involved in different functions<sup>65</sup> so its flexible arrangement, that gives to the structure the necessary plasticity, could be fundamental to engage in interactions with different partners. However, many studies presented in the literature addressed how the folded NTD is capable of distinguishing between various DNA/RNA based targets, in order to delineate the mechanisms at the basis of the N action<sup>73–77</sup>. The studies conducted in these years were not able to disentangle the atomic details that can describe the recognition mechanisms between N and the viral genomic RNA (gRNA) but it became evident how the presence of the IDRs are important in order to determine affinity and specificity<sup>64</sup>. The main proposed role for the N protein is to organize the gRNA inside the capsid<sup>78</sup> and to engage other viral and host proteins necessary to its translation<sup>79</sup>.

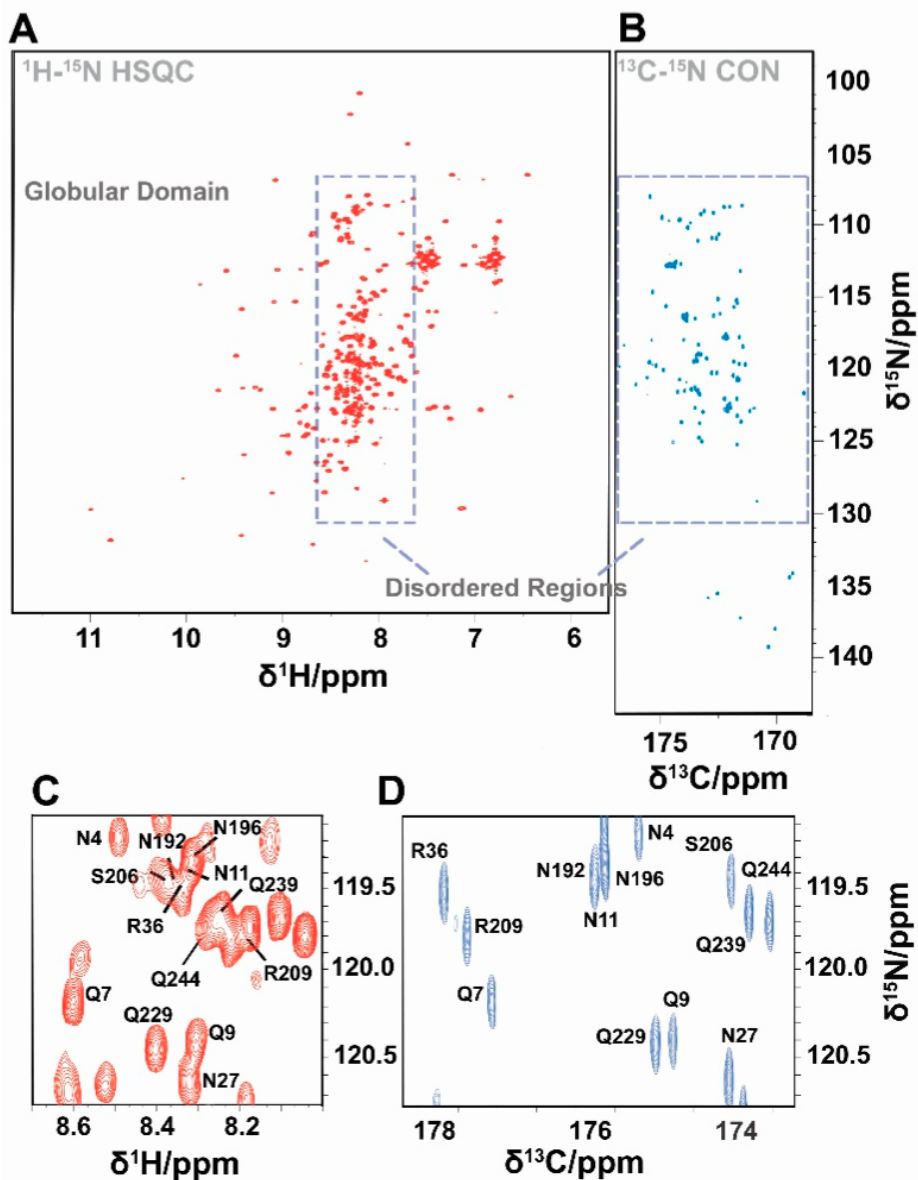
For eukaryotic RNPs, the presence of multiple cooperative sites is a common feature<sup>80–82</sup> so it is conceivable that a similar approach is also used by viral RNPs. To unveil the role of the IDRs in gRNA packaging, we tested the interaction of NTR and NTD with different viral gRNA elements. We opted to perform the experiments in potassium phosphate ( $\text{H}_2\text{PO}_4^-/\text{HPO}_4^{2-}$ ) buffer, 150 mM KCl, pH 6.5, T 298K, that are best suited for RNA interaction studies.

We thus chose the Stem-Loop 4 of the 5'-UTR (5\_SL4, nt 86–125) for a detailed analysis, since it is structurally conserved among the members of the betacoronavirus family and the related interaction is

targetable by small molecules<sup>83-85</sup>.

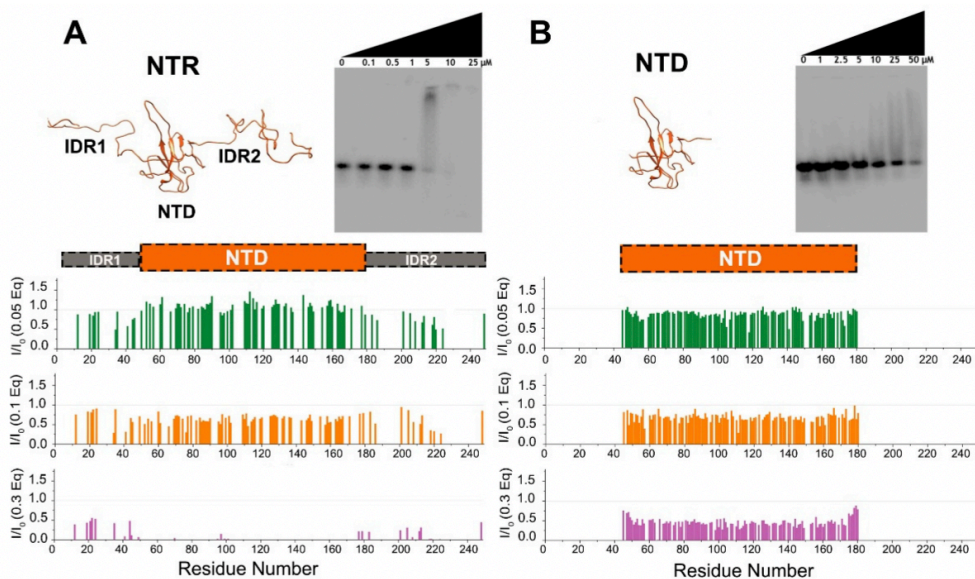
To achieve atom resolved details on both the globular and disordered parts while facing the limited timeframe imposed by protein degradation in the selected experimental conditions (in particular the low salt concentration), we exploited the HN//CON Multiple Receiver<sup>86,87</sup> approach (simply referred as HN//CON).

These experiments, which I contributed to develop during my master thesis<sup>86</sup>, enabled us to obtain a simultaneous view of the interplay between the folded and disordered protein modules through <sup>1</sup>H and <sup>13</sup>C detection thanks to the simultaneous acquisition of the two fingerprint spectra (HN-HSQC and CON). The FIDs of the HSQC are acquired during the <sup>13</sup>C magnetization recovery time ( $d_1$ ) of the CON experiment, exploiting this time to achieve additional information. While the HSQC is useful to follow signals belonging to the folded domain, the CON acts as a filter to monitor the IDRs (Figure 18).



**Figure 18.** The HN//CON acquired on the NTR construct. A and B are the resulting spectra acquired at the same time; C and D report a zoom on a region of the spectra to compare their information content. Panel C shows a crowded region where many IDRs resonances fall together with globular domains peaks. Panel D demonstrates the “filter action”, reporting only IDRs resonances that are well resolved.

It emerged that specific residues from IDRs are the first to be engaged when very low RNA amounts are added to the protein. These residues were identified and their characteristics in terms of amino acid type and distributions were discussed (Article 2.3). The results were also compared with those obtained on the NTD construct (Figure 19).

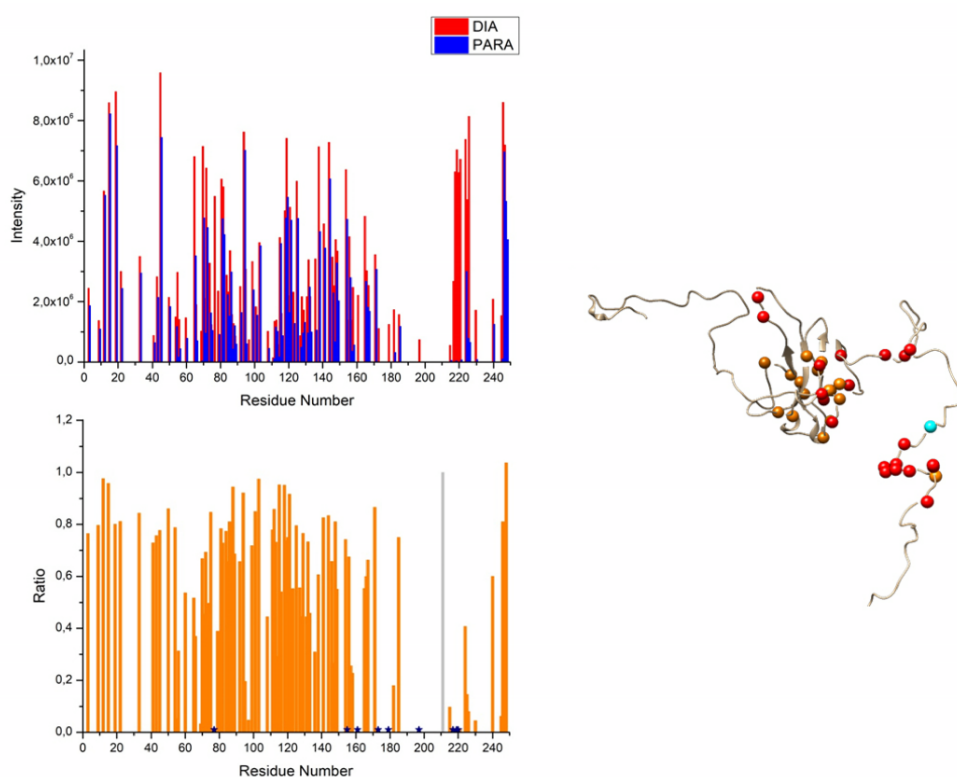


**Figure 19.** Intensity ratios plots of NTR (A) and NTD (B) construct using increasing amount of SL4. The Electrophoretic Mobility Shift Assay (EMSA) results are also shown.

We then exploited Paramagnetic Relaxation Enhancement (PRE)<sup>88–91</sup> experiments to identify long-range interactions between the IDRs and the NTD residues in the context of the modular NTR.

We produced two cysteine mutants of NTR (S23C, A211C) and then we introduced a paramagnetic tag (S-(1-oxyl-2,2,5,5,-tetramethyl-2,5,-dihydro-1H-pyrrol-3yl)methylmethane-sulfonylthiolate (MTSL)) that works as a probe for long range, transient interactions. The

paramagnetic effect acts with the broadening of the signals from residues that are spatially close to the tag. After the use of a reducing agent, the tag becomes diamagnetic and the intensities are recovered. The experiments confirmed our hypothesis of a cross-talk between the NTD and IDR2, as observed thanks to the results obtained on A211C mutant (Figure 20).



**Figure 20.** Plots of intensity and intensity ratio determined through PRE experiments conducted on A211C mutant of NTR.

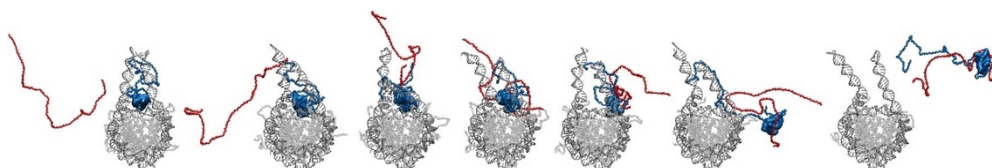
The analysis of the perturbed residues in NTR suggested an electrostatic-driven priming event for the binding, and the involvement of the hydrophobic poly-leucine tract <sup>217</sup>AALALLL<sup>224</sup> in the interaction, whose assignment was possible by exploiting the 3D (H)CBCACON

experiment, expanding the previously obtained sequence-specific assignment presented in Article 2.2. In addition, the titration experiments confirm the tendency of the system to form Liquid-Liquid Phase Separation (LLPS)<sup>83,92–95</sup>. We observed the phenomenon when more than 30% of RNA with respect to the protein concentration was added. The process of LLPS is a biological trick to temporarily compartmentalize molecules in membraneless organelles (nucleoli, stress-granules, Cajal bodies)<sup>96,97</sup>. The condensate formation could be driven by both hydrophobic and hydrophilic interactions, with the involvement of charged and/or aromatic residues from proteins and nucleic acids<sup>98,99</sup>.

In light of the N interactions, this is a very interesting insight because we observed the phenomenon only for the NTR construct but not for the NTD alone, an implicit demonstration that the disordered regions are responsible for the phase-separation, as suggested in the literature<sup>92,93,100</sup>. Further, the phenomenon limited our investigation because of the non-observability of most of the peaks (mainly those belonging from the globular domain) in the presence of the dense phase.

These results led us to investigate the effect of the electrostatic contribution to N interactions. The interactions that are based on merely electrostatic contributions are not as trivial to be investigated as they might seem to be. A very clear example of the functional activity of charge-only driven interactions among IDPs is given by the encounter of prothymosin  $\alpha$  and histone H1 protein related to its displacement from the nucleosome<sup>101,102</sup>. This fascinating competition mechanism is

based on a rapid electrostatically driven association–dissociation equilibrium that challenges the nano/picomolar affinity of the nucleosome-H1 interaction. The highly charged prothymosin  $\alpha$  invades the H1-nucleosome complex and facilitates its dissociation thanks to the interaction with H1 that is mediated by the disorder of both the proteins (Figure 21).



**Figure 21.** Simulation snapshots depicting the association of prothymosin  $\alpha$  (red) to the H1–nucleosome complex followed by the dissociation of prothymosin  $\alpha$ –H1 from the nucleosome. Picture from reference 102.

The study of the interactions of N with highly charged molecules such as polyanions can thus help to elucidate the electrostatic contribution on the key driving forces responsible for its function and to drive the design of tailored compounds for highly charged proteins.

Recently a study reported the presence of N on the surface of human transfected and infected cells's surfaces<sup>103</sup>. The protein binds with high specificity and nanomolar affinity only heparin/heparan sulfate among all the GlycosAmineGlycans (GAGs) present in the Extra Cellular Matrix (ECM) and the cellular surface. The literature shows various examples of viral RBPs that interact with cellular polysaccharides and chemokines to evade host immunity response<sup>104,105</sup>. This suggests that N could interfere with the cytokines-GAGs binding mechanism with a modulation of the innate and adaptive immunity of the host.

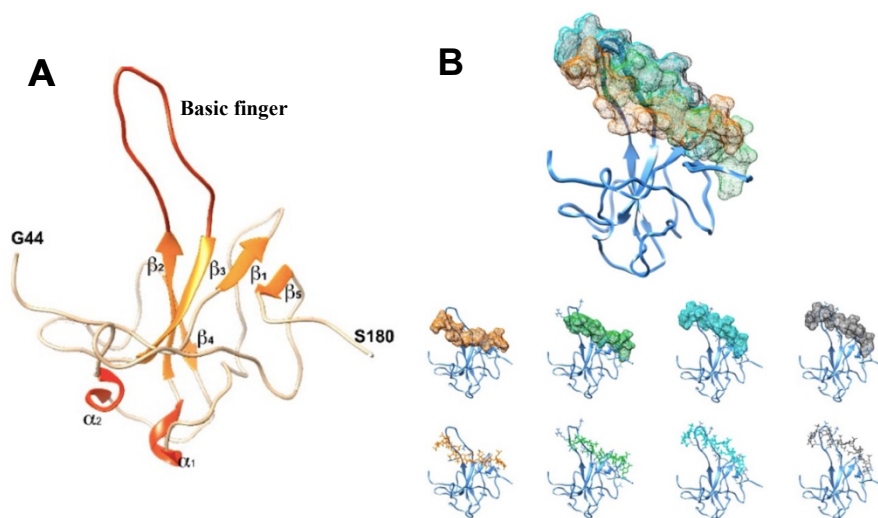
Heparin is an ubiquitous linear GAG characterized by different degrees

of sulfation which confer to it a high negative charge, it is a strong anticoagulant but also a fundamental component of the cell surface. Many studies about the effect of heparin and its derivatives were carried out<sup>106–109</sup>. Low molecular weight heparin was used in Covid-19 research treatments<sup>110</sup> and heparin-based resins were used for hemofiltration in critically ill patients in order to decrease N concentration<sup>111</sup>. The protein, indeed, was also found in human fluids<sup>112</sup>.

We then decided to test the interaction of N with enoxaparin, a low MW heparin used in clinical practice. In line with the investigation on the gRNA, we tested the differences in the binding with NTD and NTR (Article 2.4).

The dynamical studies conducted on the NTD alone showed the maintained flexibility of the flexible portions of the globular domain, such as the terminal arms and the so-called "basic finger", a loop comprising residues 92-107 (Figure 22, panel A). We exploited the HADDOCK web-server<sup>113</sup> (<https://wenmr.science.uu.nl/haddock2.4>) to obtain a model of the complex (Figure 22, panel B). It is conceived for the study of PPI but, in this case, it was possible to exploit the software for the enoxaparin docking thanks to some modification on the GAG's pdb file obtained with the help of Prof. Alexandre Bonvin.

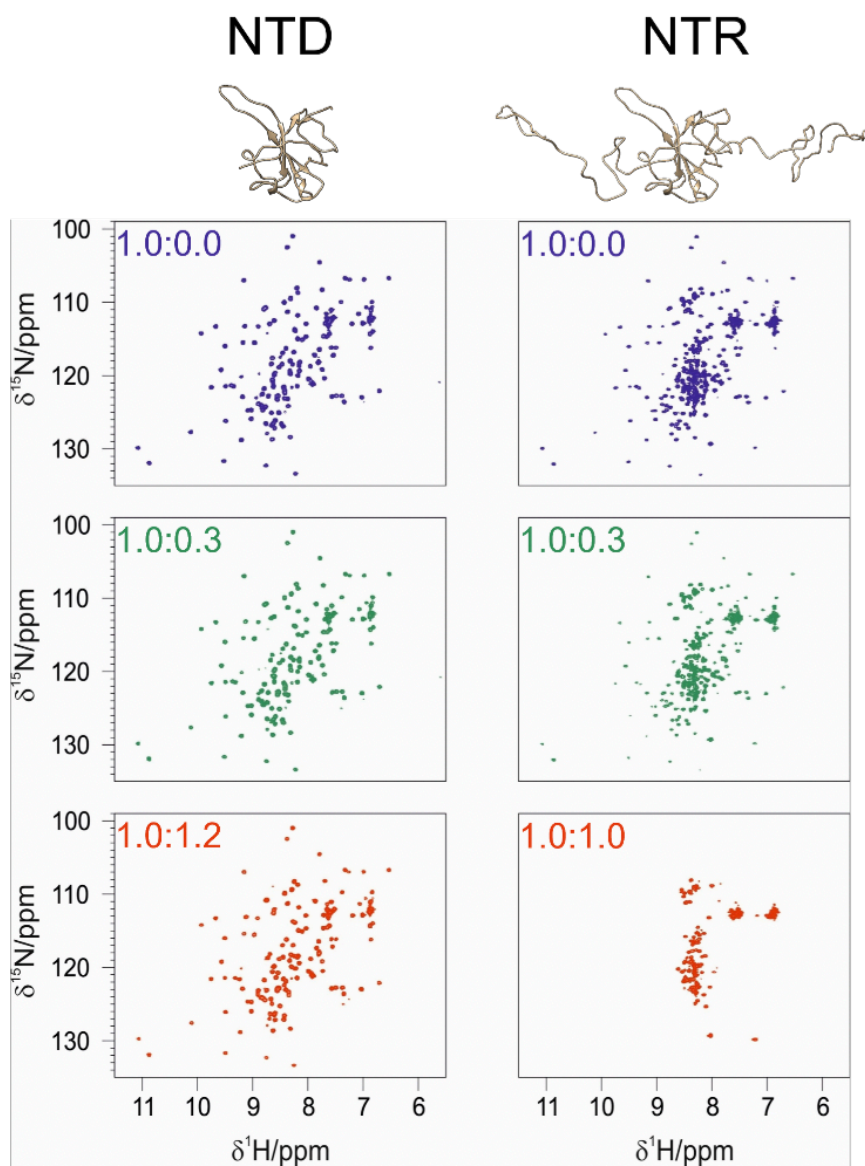




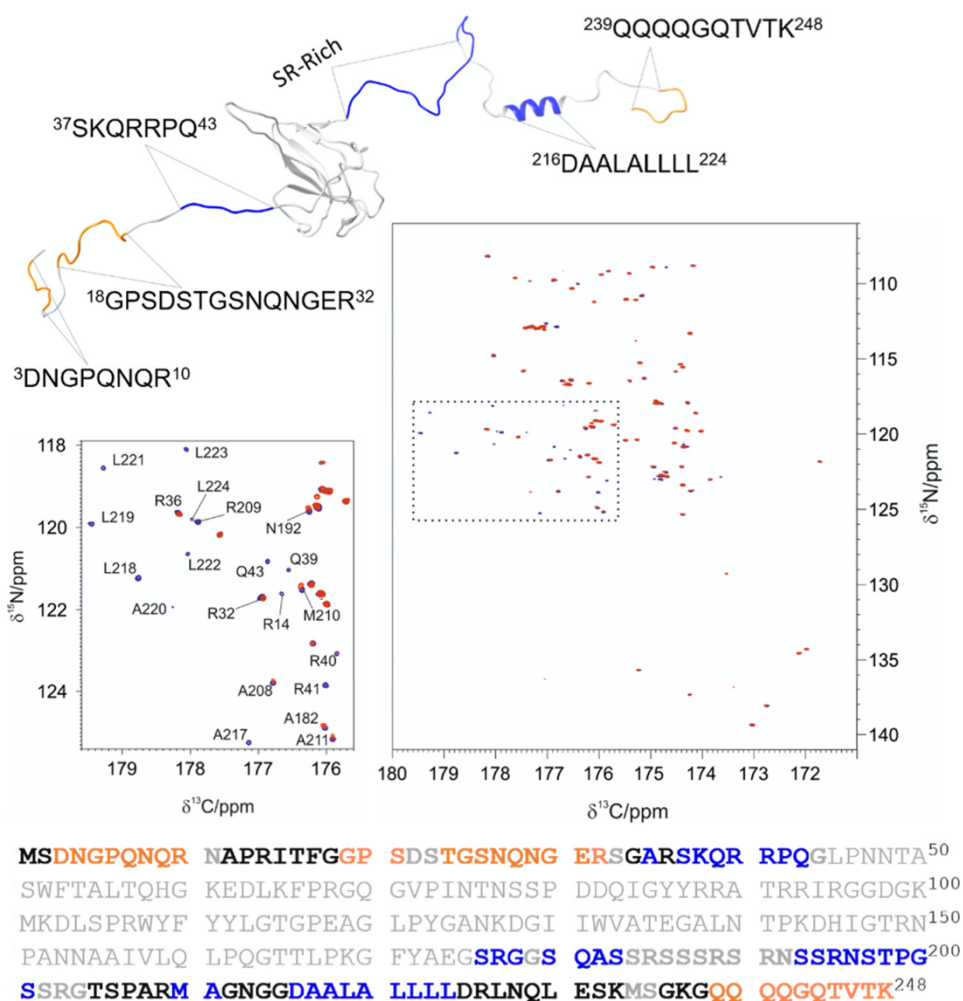
**Figure 22.** A) Secondary structural elements of the NTD domain. B) Some of the structures from the best cluster obtained with HADDOCK calculation. Enoxaparin is presented both with the mesh surface and the stick representation.

Moving to the NTR, the major detected difference was related to the binding regime experienced by the two constructs, as reported in Figure 23. The extensive broadening of the HSQC cross peaks of the globular domain can be ascribed to the augmented molecular mass of the NTR-enoxaparin complex and to the extended conformation adopted by the IDRs to accommodate the ligand. In particular, the CON mapping of the perturbed residues in the IDRs (both in chemical shifts and intensities) along NTR primary sequence, evidenced the involvement of stretch 37-43, the SR-rich and polyleucine motif (Figure 24). Other disorder promoting residues, however, are enhanced in terms of intensity, suggesting the displacement of the previously observed long-range interactions between IDR2 and NTD in favor of the enoxaparin binding. The extended protein surface, rich in arginine

residues, enhances the anchoring of the polyanion to the NTR with both coulombic interaction and H-bond formation. The IDRs could then create a platform that accommodates the long polysaccharide on NTD (or, alternatively, the negatively charged phosphodiester backbone of a nucleic acid molecule).



**Figure 23.** Comparison of HSQC spectra of NTD (left) and NTR (right) with increasing concentration of enoxaparin.



**Figure 24.** The CON mapping of the interaction between NTR and 0.3 equivalents of enoxaparin . The left panel shows a zoom in which the most perturbed residues are highlighted. The perturbed regions are depicted in blu on the NTR model in the top. The regions that are still observable in the final point of the titration are colored in orange. The same color coding is used in the primary sequence of the protein, reported on the bottom of the figure. The IDRs are reported in bold. The residues that are not detectable are reported in gray.

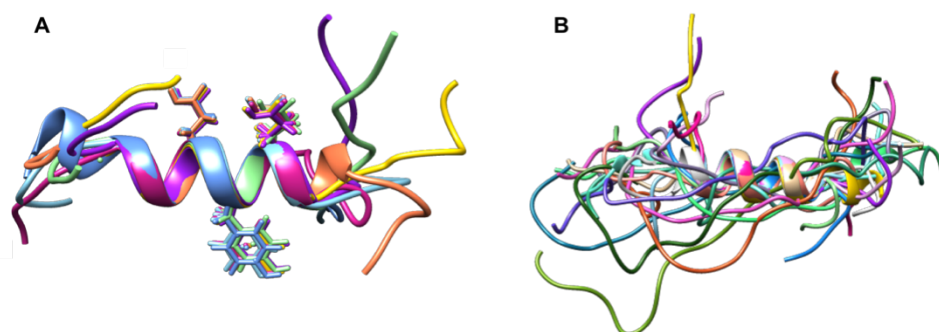
The obtained molecular details are useful to better understand the N action driven by the disordered domains and to help the drug-design of tailored compounds such as polyanions or other compounds.

The latter were the focus of a large NMR screening on a fragment-based library conducted by the Covid-19 NMR consortium (Article 2.5), a collaborative project to which we contributed. The NTD, NTD-SR and NTR constructs were involved in the screening. It emerged that NTR is targeted by 7 fragments, NTD-SR by 5 and NTD by 32. These differences suggest that the reduced number of compounds found for NTR with respect to NTD are related to some selection mechanisms mediated by the IDRs. It is thus interesting to investigate the effect of tailored ligands, such as peptides, on the NTD and NTR constructs.

Specifically, antiviral-peptide drugs are demonstrated to be an optimal candidate for novel therapies where the canonical ones are bound to fail<sup>114</sup>. Approved peptide-drugs are used against Herpes simplex, Influenza, HIV and Hepatitis virus and many others are under investigation because of the capability of these molecules to disrupt PPI both of viruses and host proteins, together with other important advantages such as non-immunogenicity, membrane permeability, safety, low cost<sup>115–118</sup>. Different studies on SARS-CoV 2 that propose the use of peptide-drugs, including peptide-vaccines, are available<sup>114</sup>. However, most of them target the Spike protein in order to avoid the hijacking of the cells.

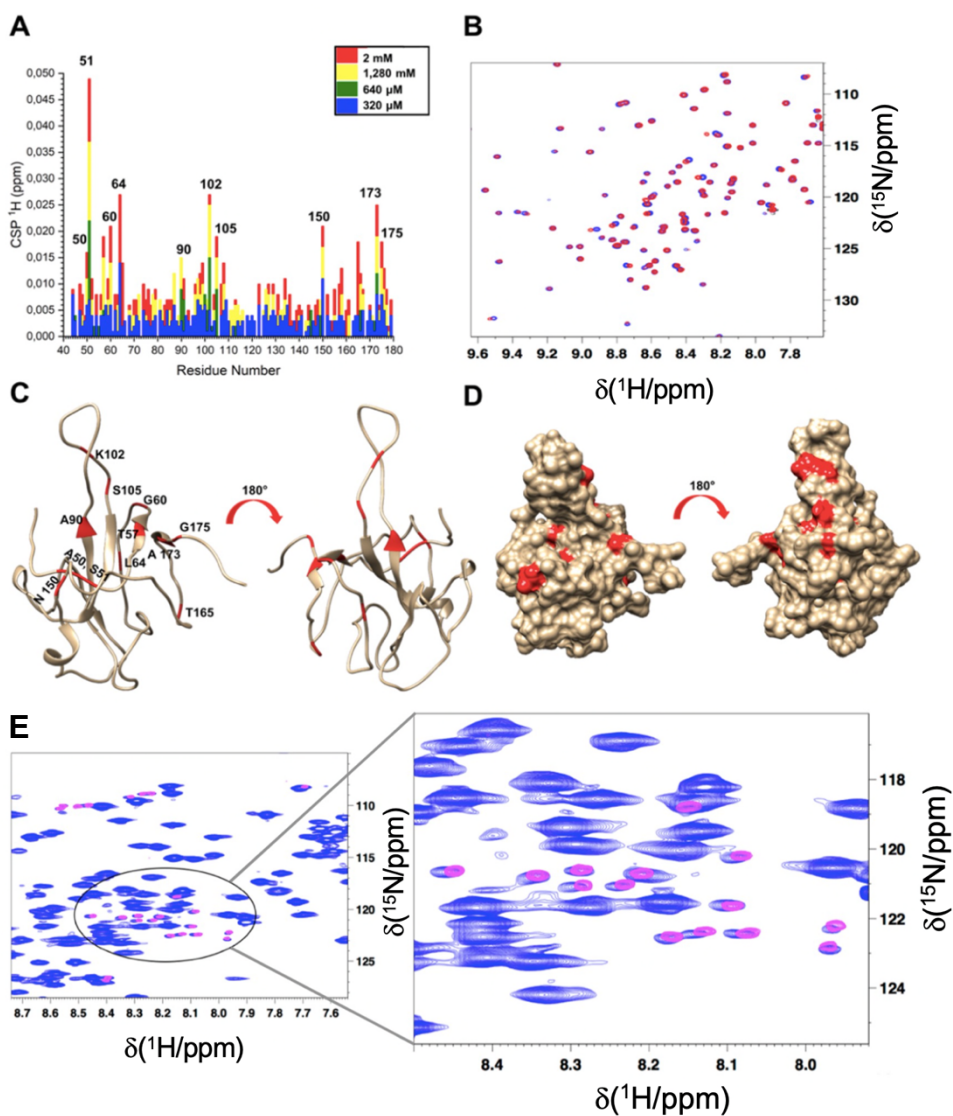
In this framework, we decided to explore this approach on the N protein starting with the design of a model-peptide able to interfere with the interaction of NTD with RNA. The first candidate, that we named AT2 (EGEGEGGLLELYLELLGGEGE( $\beta$ A)E), was designed and then obtained with Solid-Phase Peptide Synthesis (SPPS)<sup>119</sup>. The AT2 was then characterized through 2D-TOCSY and 2D-NOESY, HC-HSQC

and HN-HSQC spectra acquired at 950 MHz (22.3 T) in order to obtain a complete assignment of the molecule. Once the assignment was achieved, it was possible to compute an ensemble of conformers of the peptide based on the experimental data. We used the software Flexible Meccano<sup>120</sup> to generate ensembles based on the conformational sampling derived from the peptide primary sequence.



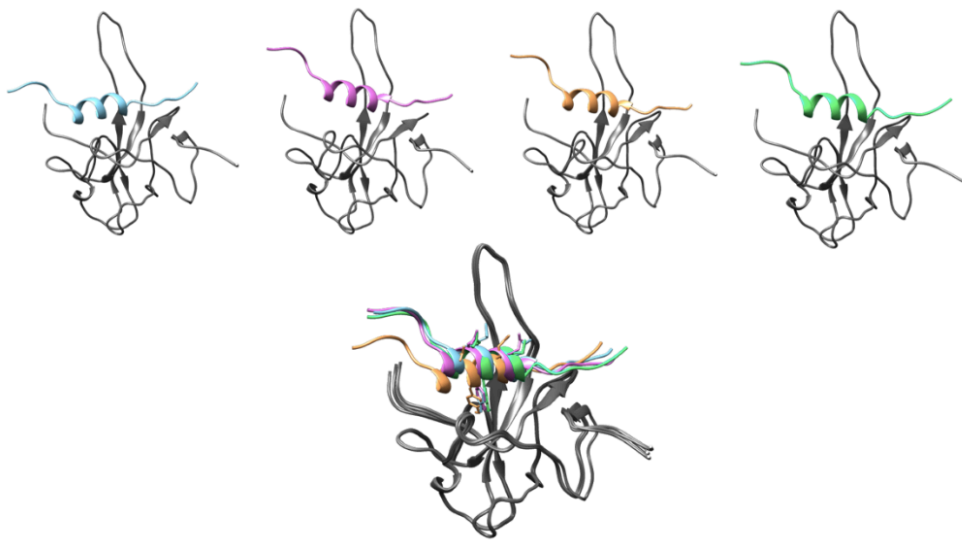
**Figure 25.** A) The superimposition of 8 structures, randomly selected from Flexible Meccano models ensemble, which form the helix. Y12 and E10, E14 residues are highlighted in all the models. B) Superimposition of different structured and unstructured models obtained with Flexible Meccano

The following step was to test the interaction of the peptide with NTD. We observed CSP only after the addition of 3.2 equivalents of AT2 with respect to the NTD concentration (100  $\mu$ M) (Figure 26, panel A). Interestingly, we observed the peaks belonging to AT2 (natural abundance) in the last point of the titration (1:20) (Figure 26, panel E). This was very useful because it allowed to observe the occurrence of the interaction also from the peptide side.



**Figure 26.** A) Chemical shift perturbations plot of the NTD-AT2 titration. B) Overlay of reference and 1:20 spectra. C) Mapping of the most perturbed residues on a NTD model. D) Space-filling representation of the same models presented in panel C. E) Superimposition and its zoom of 0.1 mM NTD+2mM AT2 (blue) and 2mM AT2 (pink) in the same experimental conditions. Some of the crosspeaks from AT2 are shifted.

We exploited HADDOCK for the docking calculation of AT2 and NTD based on the NMR data (Figure 27). We included the peptide residues that were identified from the comparison with the HSQC spectrum of the AT2 alone in the same experimental conditions.

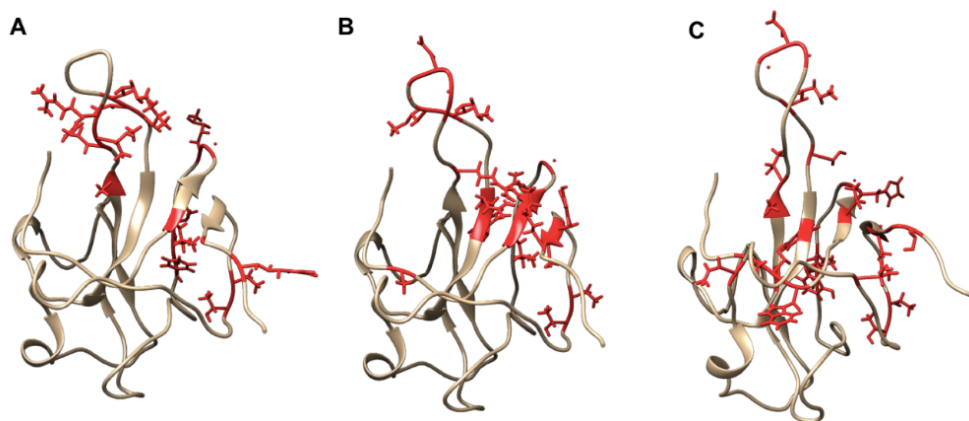


**Figure 27.** The four models of NTD:AT2 and their overlay obtained with HADDOCK calculation.

The obtained results were compared with those obtained with the gRNA and the enoxaparin molecule, showing that an extended region of NTD is affected upon interaction with the three polyanions, although to different extents (Figure 28). This could be due to the different dimensions and conformations of the ligands together with their different chemical nature. The small peptide, composed both by charged and hydrophobic residues can be a compromise between the RNA and enoxaparin. The possibility of a perturbation of the NTD folding, with the loss of hydrogen bonds and Van der Waals interactions,



cannot be ruled out.



**Figure 28.** Three models representing the CSP of NTD obtained with RNA (SL4, A), enoxaparin (B) and AT2 (C).

The very promising results open the doors to the possibility to design tailored peptides that can target the NTD. Before the latter step, it would be important to test the binding of AT2 with the NTR in order to identify the effect of the IDRs in the interactions. As emerged on the large screening discussed in paper 2.5, the presence of the IDRs results in a decrease in the number of potential ligands with the respect to the NTD alone.

In the frame of the investigation of the interaction of N and polyanions, we also investigated the interaction of NTR with a DNA fragment, the oligonucleotide ODN1 (12 mer, 5'-ATTAGGCCTAAT-3').

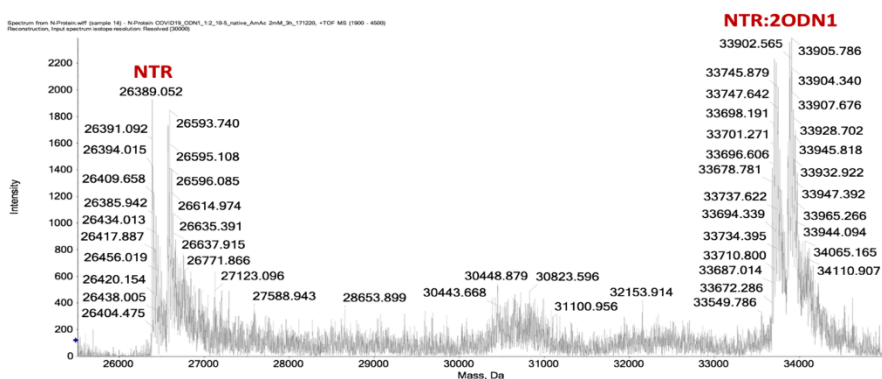
Nucleotide analogue inhibitors are optimal chemotherapeutic agents and are used for the therapy of different viruses such as HIV and hepatitis viruses<sup>121</sup>. The precursors are administered, metabolized and used to compete with natural nucleotide substrates. The strategy was

also used for Covid-19 treatment<sup>122,123</sup>. For this study, we also opted for the use of Native Mass Spectrometry (NMS), in collaboration with Prof. Luigi Messori and Doctor Lara Massai. With this approach, the native state of the investigated biomolecules is preserved thanks to the mild experimental conditions that are adopted.

NMS represents a useful approach to gaining insights into macromolecules and their interaction, also under physiological conditions. The technique was also applied to study the conformational characterization of IDPs<sup>124</sup>. An essential requirement for NMS is the use of a volatile buffer that is compatible with the ionization source to produce the gas phase from the liquid one. The main limit of the approach is related to the possible low number of available sites for acquiring a charge that results in a low number of visible charged states in the mass spectrum. In contrast to well-folded proteins, IDPs can take advantage of their flexibility which determines a high exposure of the residues.

Firstly, the NTR sample obtained after SEC step was dialyzed overnight against 2 mM ammonium acetate ( $\text{CH}_3\text{COONH}_4$ ), pH 7. Samples of 10  $\mu\text{M}$  of protein (MW 26390) were incubated with an excess of ODN1 (MW 3644) and then analyzed through the acquisition of NMS spectra after different times of incubation.

The spectrum presented in Figure 29 evidences the presence of a 1:2 NTR-ODN1 adduct in a 3 hour time-lapse. The high noise level could be related to the difficulty in the formation of the charged species through the ionization source as the association constant of the construct can be high. The adduct is no more detectable after 12h of incubation.



**Figure 29.** NMS spectrum acquired on 10 $\mu$ M NTR in 2mM CH<sub>3</sub>COONH<sub>4</sub> and an excess of ODN1 after 3 h of incubation. It is present a peak related to the formation of a 1:2 protein-oligonucleotides adduct.

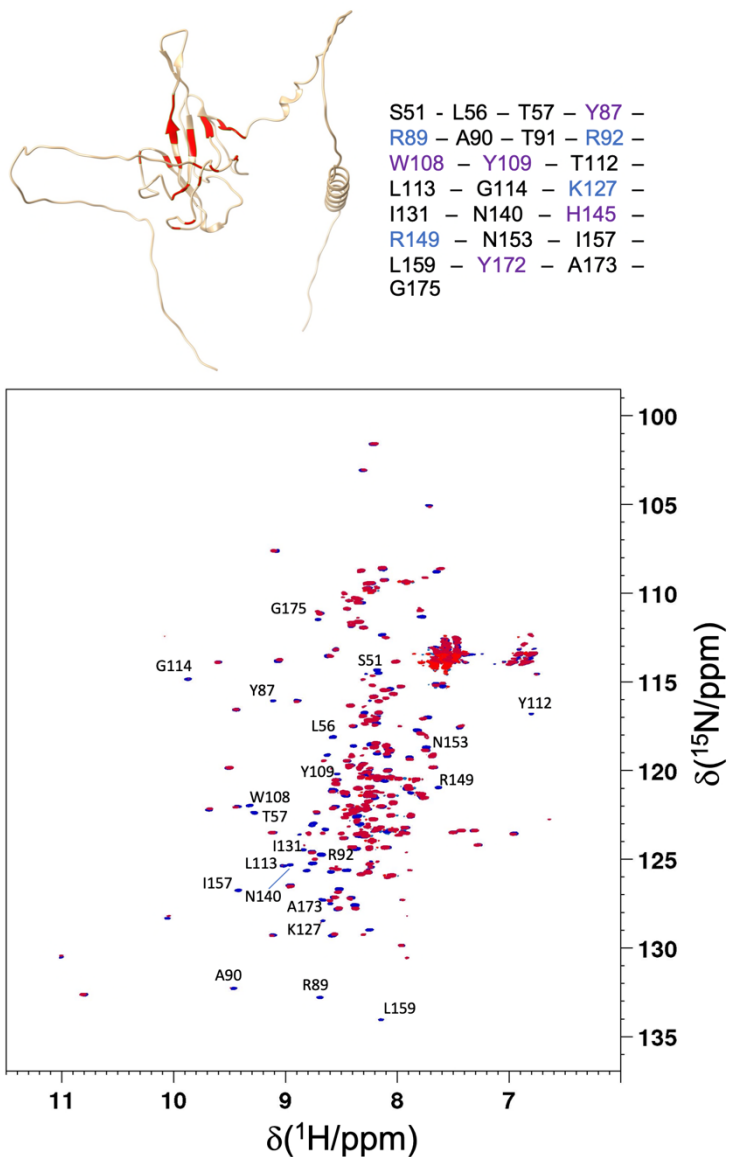
For the NMR investigation, ODN1 was firstly characterized and assigned through a suite of <sup>1</sup>H detected 1D and 2D NMR experiments acquired at high magnetic fields. The sequence-specific assignment strategy was achieved thanks to the analysis of 2D TOCSY and 2D NOESY spectra acquired at different temperatures and complemented by <sup>1</sup>H-<sup>13</sup>C HSQC experiments on 1 mM ODN1 samples. Two NOESY spectra with different mixing times, 120 ms and 250 ms, and one TOCSY spectrum, 60 ms spin-lock time, experiments were acquired at 298K in H<sub>2</sub>O solvent. Two additional NOESY spectra were also acquired dissolving the ODN1 in D<sub>2</sub>O in order to visualize more peaks that were invisible because of the H<sub>2</sub>O signal. This also allows a firm

identification of exchangeable protons that are not detectable in D<sub>2</sub>O, as the imino resonances.

For the titration, the oligonucleotide batch was obtained in 25mM TRIS, 270 mM NaCl, 0,03% NaN<sub>3</sub>, pH 6.5. Protein-DNA interaction was studied through a set of 2D <sup>1</sup>H <sup>15</sup>N-HSQC by adding the unlabeled oligonucleotide in a 80 μM NTR sample until a 1:4 stoichiometric ratio was achieved.

During the experiments, we observed both variations of intensities and chemical shifts (Figure 30). Most of the peaks that disappeared after the titration belong to the folded domain, many of them are aromatic residues (Y87, W108, Y109, H145, Y172) and positively charged residues (R89, R92, K127, R149). This combination of basic and aromatic residues may interact with the negatively charged phosphodiester backbone of ODN1, driving the binding mainly on the NTD. These preliminary experiments support the necessity to continue the investigation of the specific and non-specific interactions of the NTR and DNA/RNA fragments, to identify the key residues that are important for nucleic acid recognition in terms of affinity and specificity.

In order to follow up this study, it would be interesting to exploit both the oligonucleotide and peptide features by the synthesis of DNA mimicking molecules such as Peptide Nucleic Acids<sup>125,126</sup>. These molecules can provide an optimal tool to clarify the steps that lead to RNP complex formation in order to provide an alternative strategy for drug targeting.



**Figure 30.** Overlay of the two  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra acquired in the same conditions (25 mM TRIS, 270 mM NaCl, pH 7.2 of 80  $\mu\text{M}$  of  $^{15}\text{N}$ -labelled protein) in absence of ODN1 (blue) and containing protein:ODN1 in a molar ratio of 1:4 (red). In the figure the residues that disappears are listed. These are mapped on a NTR model and listed using the colour-coding as indicated: blue for basic residues, purple for aromatic residues and black for polar uncharged and hydrophobic ones.

## 2.4 Increasing the complexity: the full-length N protein

The characterization of large proteins remains one of the greatest challenges of NMR mainly because of the broadening of the signals due to the increased  $\tau_r$  related to the augmented molecular weight. This holds for folded proteins. The flexibility of IDRs allows us to visualize them through NMR also in the context of a large protein complex. Along this line, the filtering of the resonances of IDRs can help to elucidate the behaviour of large macromolecules giving the possibility to extract information even when challenging systems are studied. With the inspection of the flexible residues, it is possible to achieve a dynamic vision of the protein in solution and analyze the differences between the entire N protein with respect to the NTR, with the aim to describe the modular nature of N.

The Full Length (FL) N protein is dimeric, forming a 838-residues complex. The FL was obtained following the same protocol used for the NTR. The protein tends to precipitate in inclusion bodies so the slow expression at 288 K was chosen to facilitate the production of the soluble protein. It was recently proposed that the presence of nucleic acid contaminations strongly influences the dynamics and the oligomeric state of the protein<sup>127</sup>. For this reason, the use of DNase and RNase was indispensable because of the very strong affinity of N for nucleic acids. A one-step IEC purification was conducted.

This allowed obtaining samples of the FL with a good concentration (70  $\mu$ M). Protein solubility was maintained with the use of 450/500 mM

NaCl containing buffer.

We performed a structural prediction using D-I-TASSER as we did for the NTR construct (the complete results are available at this link <http://bit.ly/3gjzenj>). In addition, we obtained a second FL model with AlphaFold (AF), the new frontier of protein structure determination<sup>128</sup>. AF is an artificial intelligence-based algorithm capable to define protein folding with an accuracy that is competitive with the one obtained through experiments. However, as AF is conceived for folded proteins, it can force the folding of disordered regions.

The two predicted models are reported in Figure 31 together with the N primary sequence. Both the protein primary sequence and the models are colour-coded to highlight the elements that share the same structural features.

It is worth noting how IDR1 and IDR2 are predicted to be fully disordered (with the exception of poly-leucine stretch) by AF while D-I-TASSER predicts more structured elements, especially for IDR1. The latter model, indeed appears to predict a more compact asset for the NTR portion that could suggest the presence of some intramolecular interactions that involve IDR1 and IDR2.

In both models, IDR3 presents a helical structure. The same propensity is predicted by Fast Estimator of Latent Local Structure (FELLS), a tool that allows the visualisation of disordered protein features based on the primary sequence<sup>129</sup>. It is proposed that IDR3 interacts with the Membrane (M) protein, another viral structural protein, in order to anchor N to the inner part of the capsid<sup>95,130</sup>

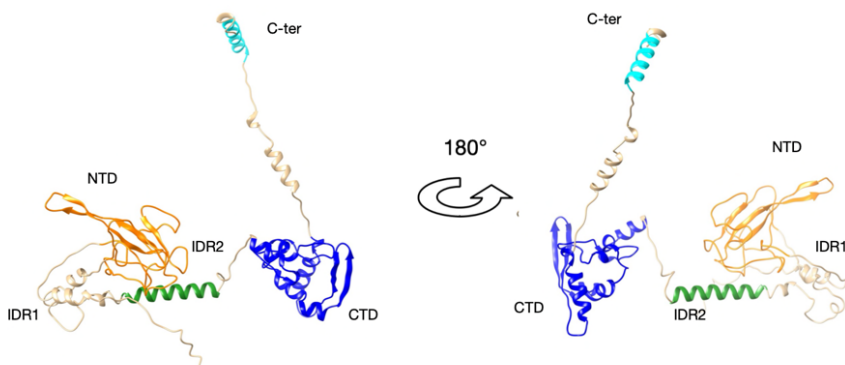
The results can be useful as a guide for the FL study but, in any case,

it is necessary not to over-interpret them, always keeping in mind that they are simulations obtained without the use of empirical information. For instance, the predictions ignore the dimeric form in solution and the

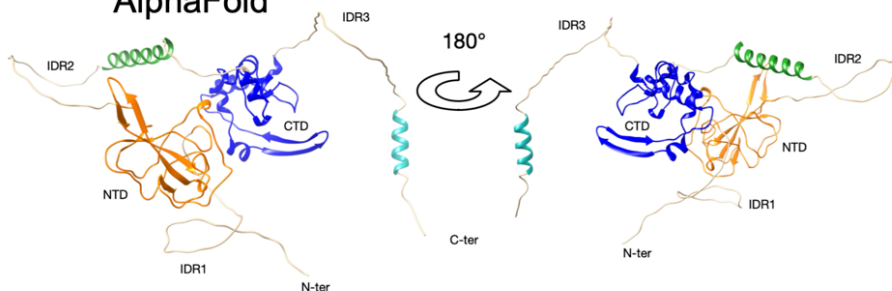


	10	20	30	40	50	60	70	80
	MSDNGPQNQR	NAPRITFGGP	SDSTGSNNG	ERSGARSQR	RPQGLPNNTA	SWFTALTQHG	KEDLKFFRQG	GVPINTNSSP
	90	100	110	120	130	140	150	160
	DDQIGYYRRA	TRIRGGDGK	MKDLSPRWYF	YYLGTGPEAG	LPYGANKDGI	IWVATEGALN	TPKDHIGTRN	PANNAIVLQ
	170	180	190	200	210	220	230	240
	LPQQTLPKG	FYAEGRSGGS	QASSRSSRS	RNSSRNSTPG	SSRGTSPARM	AGNGGDAALA	LLLLDLRLNQL	ESKMSGKGGQ
	250	260	270	280	290	300	310	320
	QQGQTVTKKS	AAEASKKPRQ	KRTATKAYNV	TQAFGRRGPE	QTQGNFGDQE	LIRQGTDYKH	WPQIAQFAPS	ASAFFGMSRI
	330	340	350	360	370	380	390	400
	GMEVTPSGTW	LTYTGAIKLD	DKDPNFKDQV	ILLNKHIDAY	KTFPPTEPKK	DKKKKADETQ	ALPQRQKKQQ	TVTLLPAADL
	410	419						
	DDFSKQLQQS	MSSADSTQA						

### D-I-TASSER



### AlphaFold

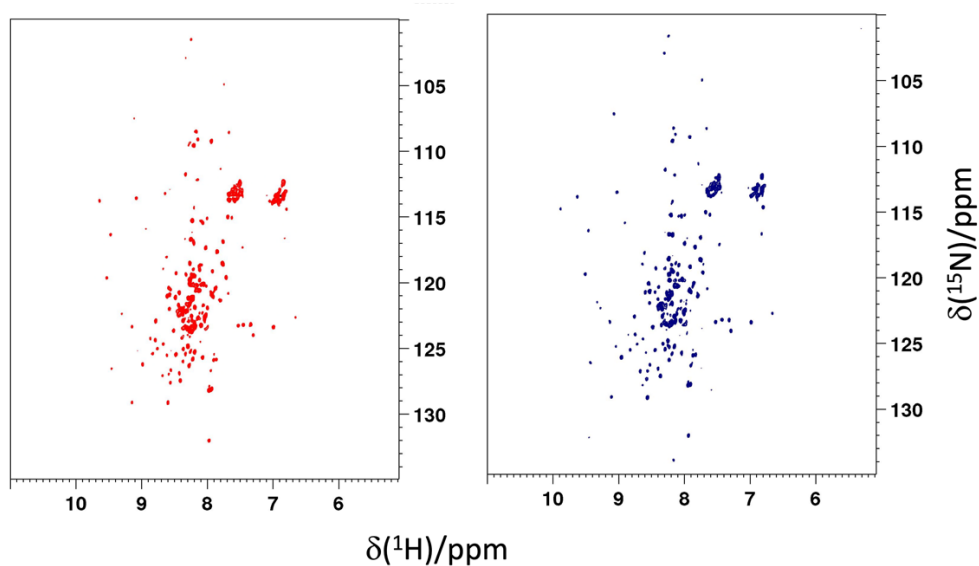


**Figure 31.** The two N models in different orientations as predicted by AlphaFold and D-I-TASSER. On the top, the primary sequence of N is colored according to the structural elements evidenced in the two models. In particular, they both present the helix conformation of the polyleucine tract and a helical portion in IDR3. The D-I-TASSER predicted model is more structured compared to the AlphaFold one.

dependence of the protein conformation on ionic strength.

The complete FL characterization was performed with the 28.7 T NMR

instrument. Figure 32 reports two HN-BEST TROSY (Transverse Relaxation Optimized spectroscopy)<sup>38</sup> spectra acquired at two different temperature values (298 K, red and 303 K, blue).

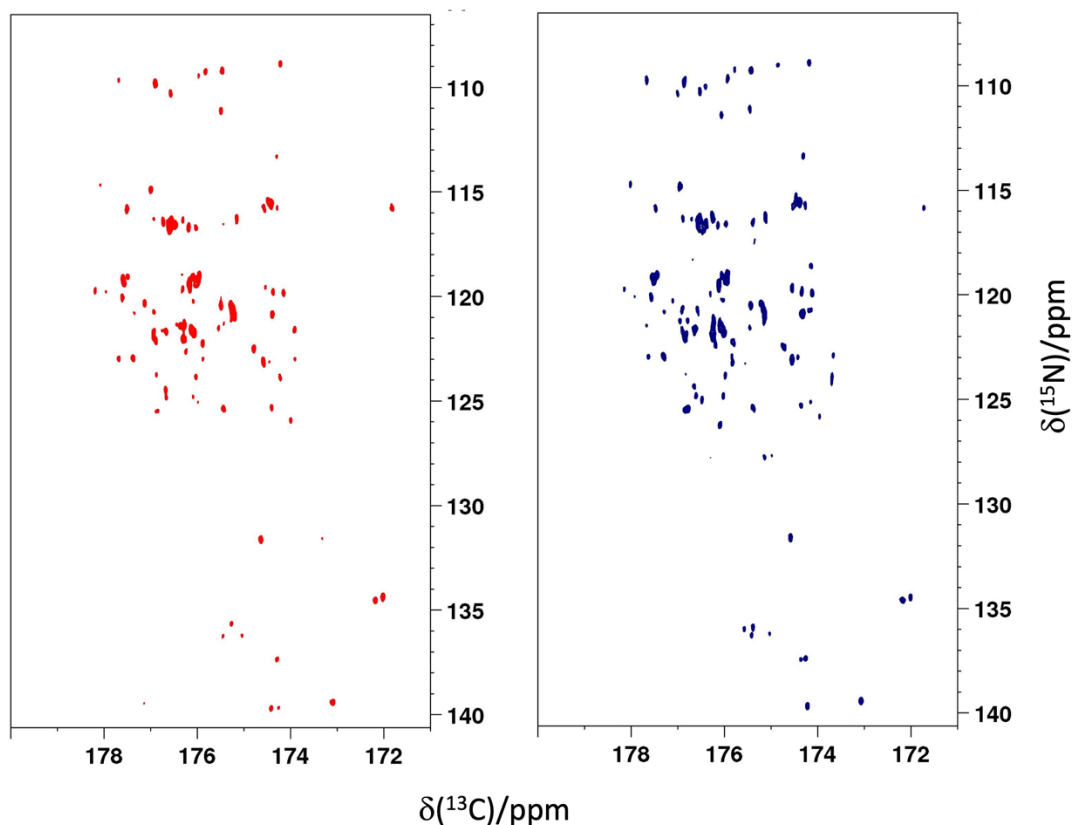


**Figure 32.** The two 2D HN-BEST-TROSY spectra acquired on a 70 $\mu$ M sample of the FL in 25mM TRIS, 450mM NaCl, 0.06% NaN<sub>3</sub>, pH 7.2. The red one was acquired at 298 K, the blue at 303K.

The TROSY strategy helps to acquire 2D HN spectra on large proteins with the selection of the sharper component of the obtained signal. The BEST-TROSY approach allows long FID acquisition times, an aspect that helps to resolve peaks from disordered residues.

Although the advantages of the TROSY effect at this high magnetic field, the results still highlight the benefits of <sup>13</sup>C direct detection for this challenging system. We opted for the H <sup>$\alpha$</sup> -start version of the CON experiment. The spectra at two temperatures are presented in Figure 33. It is possible to note that many more cross-peaks are present in the

blue spectrum acquired at 303 K. This is in line with the augmented flexibility of the protein when the temperature is increased. However, a modulation of the intensities of the resonances is also present. This suggest different dynamical behaviours of the three IDRs in the context of the FL protein.



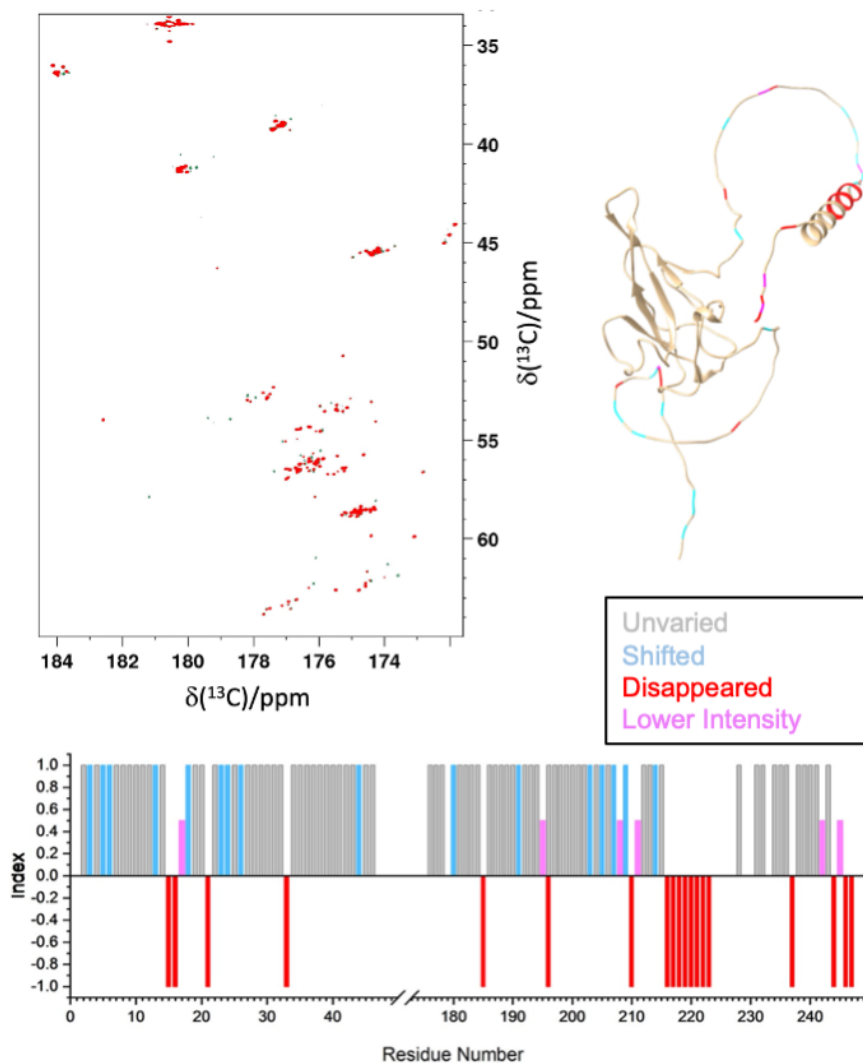
**Figure 33.** The 2D (HCA)CON spectra acquired on a 70  $\mu\text{M}$  sample of the FL in 25 mM TRIS, 450 mM NaCl, 0.06%  $\text{NaN}_3$ , pH 7.2. The red one was acquired at 298 K, the blue at 303 K.

The differences between NTR and FL were then investigated. Considering the very good results obtained with the use of 2D CACO on NTR, we decided to perform the same set-up on the FL while varying

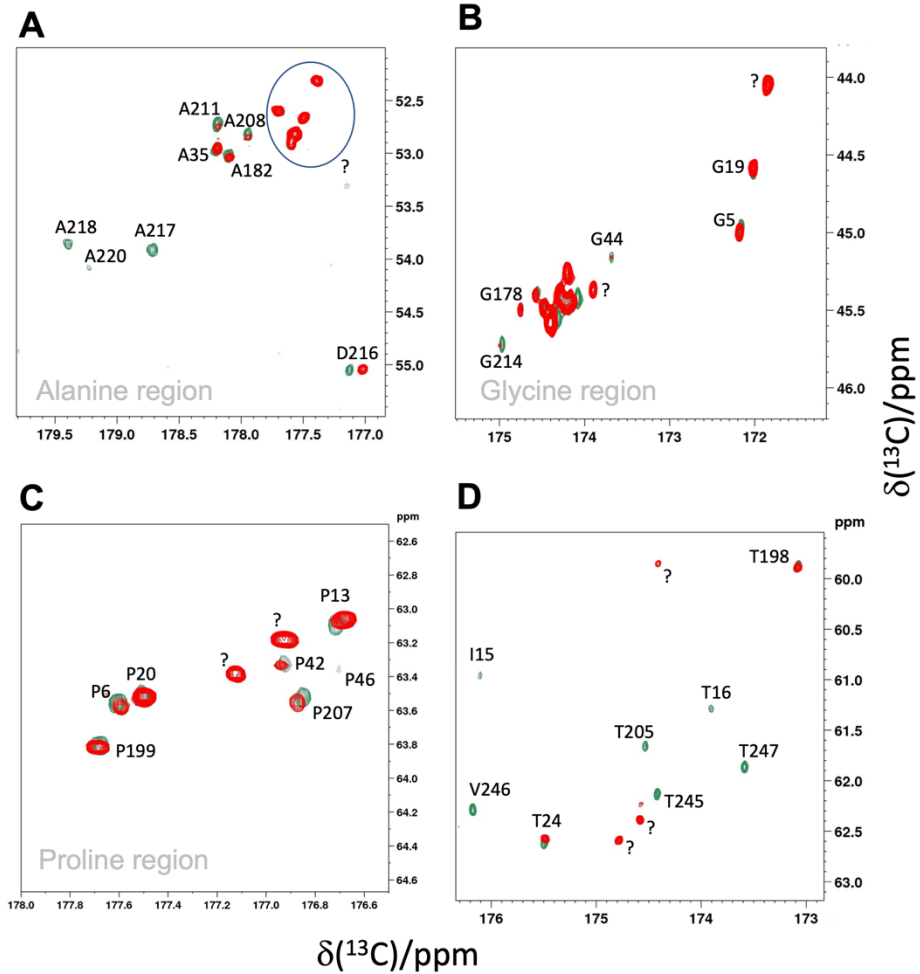
protein concentration and temperature. In this context, the IDR3 assignment was not trivial due to the short lifetime of the recombinant protein and the unsuitable concentration and salt conditions. However, a tentative assignment of some of the newly identified resonances of the FL is obtained thanks to the strategy proposed in Chapter 1. The superimposition of the 2D CACO acquired on the FL (red) and NTR (green) in the same experimental conditions (70  $\mu$ M protein concentration, 25 mM TRIS, 450 mM NaCl, pH 7.2, 0,06% NaN<sub>3</sub>, 298 K) is reported in Figure 34 together with the “observability plot”. The same results are plotted on an NTR model according to the colour code used for the plot. Four spectral regions extracted from the superimposed spectra are also presented in Figure 31 to highlight some novel intriguing observations.

Indeed, this simple analysis allowed us to easily discriminate new peaks belonging to the additional region present in the FL (249-419 residues) by comparison with the previously obtained results. This comparison also allowed us to identify relevant differences in terms of changes in cross peak intensities and chemical shifts. Most of the peaks from the <sup>13</sup>PRITFGGP<sup>19</sup> and <sup>217</sup>AALALLL<sup>224</sup> regions are not detectable in the FL context, in line with its dimeric asset and the possible presence of different dynamical properties of the FL protein. This is also supported by the consistent perturbation of the chemical shifts experienced by P13 (Figure 35, panel C) and D216 (Figure 35, panel A). The proline region is also useful to evidence the heterogeneity of the cross peak intensities (Figure 35, panel C). In addition, peaks belonging to the 207-212 tract are strongly broadened. Other perturbations are presented in

the NTR terminal part 244-248 which is predicted to be structured when tethered to CTD.



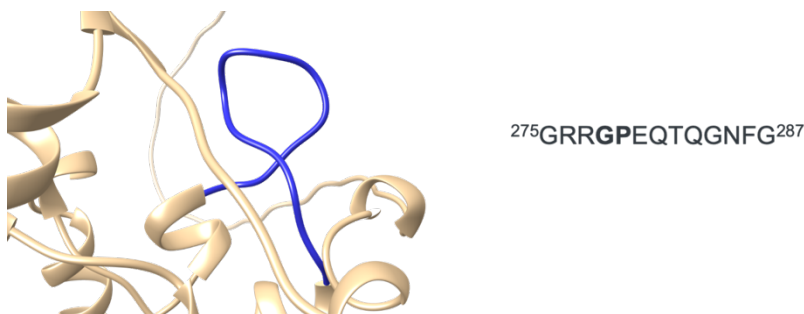
**Figure 34.** The overlay of the 2D CACO acquired on NTR (green) and the FL (red) in the same experimental conditions. On the bottom, the “observability plot”, qualitatively reporting the differences between the two experiments. A generic index is plotted against the primary sequence and reports unvaried (gray), shifted (light blue), disappeared (red) and lower intensity resonances (cyan). The residues are plotted on a NTR model according to the same color code.



**Figure 35.** The four regions extracted from the CACO spectra acquired on NTR (green) and the FL protein (red). Novel peaks belonging to the 249-419 region present in FL are indicated with the question mark symbol.

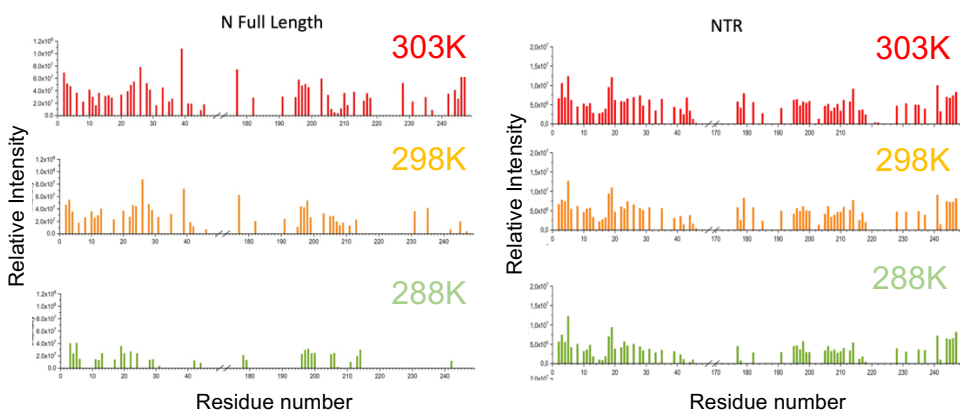
The detectable peaks belonging to the SR-region are not perturbed. A new peak resonates in the spectral region where Gly-Pro pairs fall but, interestingly, no GP pairs are present in the IDR3. This suggests additional flexibility of the CTD. Indeed, inspecting the primary sequence, a GP pair is present in position 278-279, where a flexible

loop is predicted by both AF and D-I-TASSER (Figure 36).



**Figure 36.** The flexible portion predicted by AF and its primary sequence .

The temperature dependence was also investigated together with the effect of varying protein concentrations. A set of CACO spectra were acquired at 288, 298 and 303 K (Figure 37).



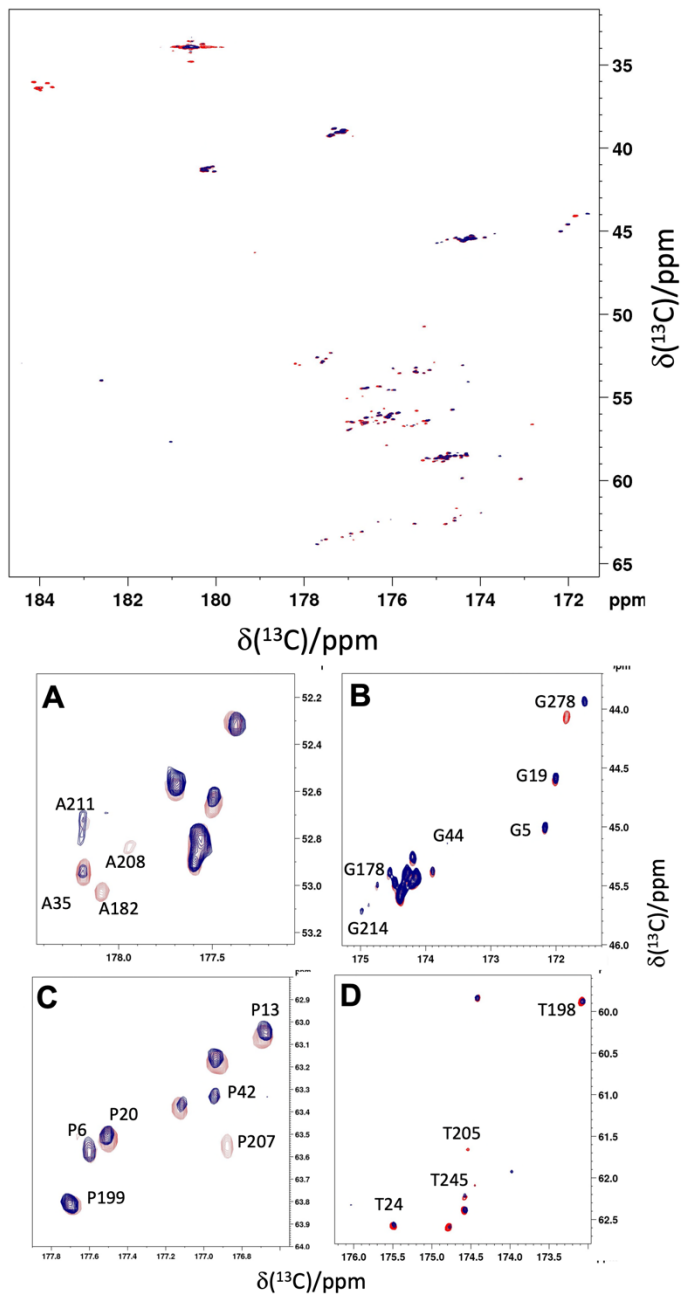
**Figure 37.** The intensity vs residue number plots of three different temperature obtained for the FL (left) and the NTR construct (right). Intensities of cross peaks were measured in CACO spectra.

The resonances from IDR2 seem to be more affected by the temperature variation but, overall, the FL behaviour is similar to those of the NTR alone.

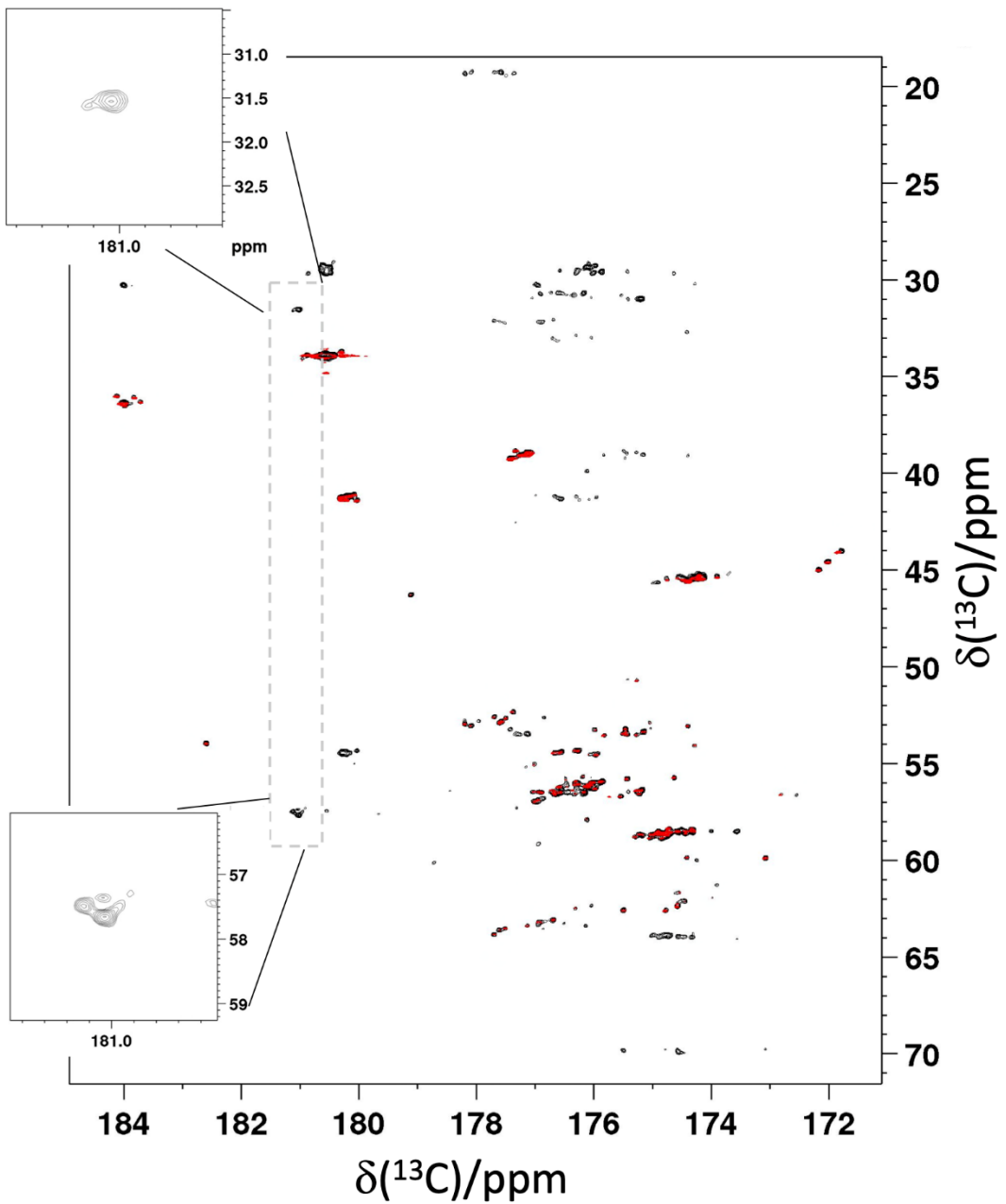


The behaviour of the protein is altered when the concentration is halved. The superimposition of 2D CACO recorded on samples containing 35  $\mu\text{M}$  (blue spectrum) and 70  $\mu\text{M}$  (red spectrum) FL protein is reported in Figure 38 together with the same zoomed regions proposed in the previous figures (Figure 17 and 35).

Many peaks experience CSP such as those that flank glutamine and arginine region (centered at about 175.6/56.7 ppm). This set of peaks is not present in the NTR, so it is reasonable to affirm that it derives from the CTD-IDR3 module. It is necessary to consider the possible presence of additional resonances related to highly flexible portions of the CTD. The resonance attributed to G278 experiences the strongest CSP while other peaks, such as P6, P42, G4, A211, G214 and a few unassigned ones, show higher intensities at 35  $\mu\text{M}$ . This is surprising considering the lower concentration of the sample and can only be explained invoking a variation in the local motional properties upon decreasing concentration. We can propose that they could be ascribed, at first instance, to the modulation of the PPI in the dimer or, possibly, to different dimer assets. Last, but not least, CBCACO and CACO show a subset of novel resonances in the region centered around 181.1 ppm/57.5 ppm/ 31.7 ppm whose observability is influenced both by temperature and concentration. The superimposed spectra, acquired on the 70  $\mu\text{M}$  sample, are presented in Figure 39. These preliminary results constitute an optimal starting point to characterize this large modular protein. They also demonstrate the need to configure the study of the chemo-physical properties of N under specific experimental conditions since even a slight variation can alter the results.



**Figure 38.** The superimposition of CACO spectra acquired on 35  $\mu\text{M}$  (blue) and 70  $\mu\text{M}$  (red) FL protein. On the bottom, the four spectral region zooms reported to evidence CSP and intensity modulation.



**Figure 39.** The CBCACO (black) and CACO (red) spectra acquired on 70  $\mu\text{M}$  FL protein at 298 K at 28.2 T. The CBCACO shows an additional set of peaks at 181/57,7/31.5 ppm that are not detectable in the NTR construct. They can belong to dimer's sidechains.

Related articles:

**2.1 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**

**2.2 Large-scale recombinant production of the SARS-CoV-2 proteome for high-throughput and structural biology applications**

**2.3 NMR reveals specific tracts within the intrinsically disordered regions of the SARS-CoV 2 Nucleocapsid protein involved in RNA encountering**

**2.4 The role of disordered Regions in orchestrating the properties of multidomain proteins: the SARS-CoV 2 Nucleocapsid protein and its interaction with Enoxaparin**

**2.5 Comprehensive Fragment Screening of the SARS-CoV-2 Proteome Explores Novel Chemical Space for Drug Development**

# Conclusion

It is now well established that IDPs and IDRs carved out a key role in protein chemistry. Their characterization continuously stimulates the development of novel strategies to investigate their biochemical and biophysical role, together with the development of medicinal chemistry approaches for drug discovery studies.

NMR spectroscopy is a pivotal technique to address all these topics as it allows to perform atomic-resolution investigations. In this framework, I conducted my PhD research with the aim of developing improved NMR approaches to face the study of IDPs and, in particular, of modular proteins.

As a first contribution, I developed a new strategy to obtain a fingerprint of an IDP when challenging experimental conditions are approached. With  $^{13}\text{C}$ -direct detected 2D experiments, it was possible to characterize  $\alpha$ -synuclein at physiological-like conditions (pH, salt, temperature), obtaining additional unprecedented information on the role of amino acid side chains. Indeed, the use of  $^{13}\text{C}$ -detected experiments allows us to explore high temperature, salt and pH conditions that usually limit  $^1\text{H}$ -detection.

The strategy proved to be optimal to investigate the fuzzy interplay with  $\text{Ca}^{2+}$  ions, giving a valuable tool to explore interactions that are of strong interest to understanding IDPs functional mechanisms.

Later, the conjunction of  $^{13}\text{C}$ -experiments and high-field hardware was demonstrated to be useful to characterize the heterogenous structure

of the SARS-CoV 2 N protein and to obtain atomic-level data on the flexible IDRs within complex constructs comprising globular and disordered domains.

The possibility to mimic physiological-like conditions became a tool for the *in-vitro* study of relevant biological mechanisms as the interplay with genomic RNA, LLPS phenomena and interactions with other molecules of interest such as organic compounds, small nucleic acid molecules, polyanions and synthetic peptides.

Last, but not least, the recombinant production of the large full-length N protein dimer was achieved together with its preliminary characterization obtained with the tools presented in the thesis.

The state-of-art of NMR makes now possible to explore “disorder” in more depth and to fill the gap of knowledge that persists in this fundamental field of research.

In the next future, it will allow addressing many unanswered questions related to molecular mechanisms at the basis of the function of IDPs, contributing to the development of alternative drug-discovery strategies and to a new vision of protein function, still too oriented on the concept of “structure”. The era of “unstructural” protein chemistry has just begun.

## Bibliography

1. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–42 (2000).
2. Fisher, E. Ueber den Einfluss der Konfiguration auf die Wirkung der Enzyme III. *Berichte der deutschen chemischen Gesellschaft* **28**, 2, 1429-1438 (1895)
3. Lemieux, R. U. & Spohr, U. How Emil Fischer was Led to the Lock and Key Concept for Enzyme Specificity. Presented at the symposium “Emil Fischer: 100 Years of Carbohydrate Chemistry,” 203rd National Meeting of the American Chemical Society, Division of Carbohydrate Chemistry, San Francisco, California, April 5–10, 1992. 1–20 (1994).
4. Karush, F. Heterogeneity of the Binding Sites of Bovine Serum Albumin . *J Am Chem Soc* **72**, 2705–2713 (1950).
5. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences* **44**, 98–104 (1958).
6. Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. & Klug, A. Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits. *Nature* **276**, 362–368 (1978).
7. Huber, R. Conformational flexibility and its functional significance in some protein molecules. *Trends Biochem Sci* **4**, 271–276 (1979).
8. Livnah, O., Bayer, E. A., Wilchek, M. & Sussman, J. L. Three-dimensional structures of avidin and the avidin-biotin complex. *Proceedings of the National Academy of Sciences* **90**, 5076–5080 (1993).
9. Romero, P. *et al.* Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput* 437–48 (1998).
10. Persson, B. Bioinformatics in protein analysis. *EXS* **88**, 215–31 (2000).
11. le Gall, T., Romero, P. R., Cortese, M. S., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in the Protein Data Bank. *J Biomol Struct Dyn* **24**, 325–42 (2007).
12. Vacic, V., Uversky, V. N., Dunker, A. K. & Lonardi, S. Composition Profiler: a tool for discovery and visualization of

- amino acid composition differences. *BMC Bioinformatics* **8**, 211 (2007).
13. Sickmeier, M. *et al.* DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* **35**, D786-93 (2007).
  14. Bairoch, A. *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33**, D154-9 (2005).
  15. Dunker, A. K. *et al.* Intrinsically disordered protein. *J Mol Graph Model* **19**, 26–59 (2001).
  16. Uversky, V. N. The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins. *Intrinsically Disord Proteins* **1**, e24684.
  17. Uversky, V. N. Intrinsically disordered proteins from A to Z. *Int J Biochem Cell Biol* **43**, 1090–1103 (2011).
  18. Uversky, V. N., Gillespie, J. R. & Fink, A. L. Why are ‘natively unfolded’ proteins unstructured under physiologic conditions? *Proteins* **41**, 415–27 (2000).
  19. Romero, P. *et al.* Sequence complexity of disordered protein. *Proteins* **42**, 38–48 (2001).
  20. Basile, W., Salvatore, M., Bassot, C. & Elofsson, A. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput Biol* **15**, e1007186 (2019).
  21. Dunker, A. K. *et al.* *Intrinsically disordered protein*. (2001).
  22. Daughdrill, G. W., Chadsey, M. S., Karlinsey, J. E., Hughes, K. T. & Dahlquist, F. W. The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28. *Nat Struct Biol* **4**, 285–91 (1997).
  23. Schuster, B. S. *et al.* Controllable protein phase separation and modular recruitment to form responsive membraneless organelles. *Nat Commun* **9**, 2985 (2018).
  24. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem Sci* **33**, 2–8 (2008).
  25. Bah, A. & Forman-Kay, J. D. Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. *J Biol Chem* **291**, 6696–705 (2016).
  26. Darling, A. L. & Uversky, V. N. Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter. *Front Genet* **9**, (2018).



27. Uversky, V. The triple power of D<sup>3</sup>: Protein intrinsic disorder in degenerative diseases. *Frontiers in Bioscience* **19**, 181 (2014).
28. Piai, A. *et al.* Just a Flexible Linker? The Structural and Dynamic Properties of CBP-ID4 Revealed by NMR Spectroscopy. *Biophys J* **110**, 372–381 (2016).
29. Kurzbach, D. *et al.* Cooperative Unfolding of Compact Conformations of the Intrinsically Disordered Protein Osteopontin. *Biochemistry* **52**, 5167–5175 (2013).
30. Tompa, P. Multiteric Regulation by Structural Disorder in Modular Signaling Proteins: An Extension of the Concept of Allostery. *Chem Rev* **114**, 6715–6732 (2014).
31. Cermakova, K. *et al.* A ubiquitous disordered protein interaction module orchestrates transcription elongation. *Science* (1979) **374**, 1113–1121 (2021).
32. Mackereth, C. D. & Sattler, M. Dynamics in multi-domain protein recognition of RNA. *Current Opinion in Structural Biology* **22**, 287–296 (2012).
33. Ferron, F., Longhi, S., Canard, B. & Karlin, D. A practical overview of protein disorder prediction methods. *Proteins* **65**, 1–14 (2006).
34. Necci, M., Piovesan, D. & Tosatto, S. C. E. Critical assessment of protein intrinsic disorder prediction. *Nat Methods* **18**, 472–481 (2021).
35. Edwards, R. J. & Palopoli, N. Computational Prediction of Short Linear Motifs from Protein Sequences. in 89–141 (2015).
36. Haerty, W. & Golding, G. B. Low-complexity sequences and single amino acid repeats: not just “junk” peptide sequences. *Genome* **53**, 753–762 (2010).
37. Metallo, S. J. Intrinsically disordered proteins are potential drug targets. *Curr Opin Chem Biol* **14**, 481–488 (2010).
38. Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated  $T_2$  relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proceedings of the National Academy of Sciences* **94**, 12366–12371 (1997).
39. Felli, I. C. & Pierattelli, R. <sup>13</sup>C Direct Detected NMR for Challenging Systems. *Chem Rev* **122**, 9468–9496 (2022).

40. Felli, I. C., Bermel, W. & Pierattelli, R. Exclusively heteronuclear NMR experiments for the investigation of intrinsically disordered proteins: focusing on proline residues. *Magnetic Resonance* **2**, 511–522 (2021).
41. Bermel, W., Bertini, I., Felli, I. C., Kümmerle, R. & Pierattelli, R. <sup>13</sup>C Direct Detection Experiments on the Paramagnetic Oxidized Monomeric Copper, Zinc Superoxide Dismutase. *J Am Chem Soc* **125**, 16423–16429 (2003).
42. Bermel, W., Bertini, I., Felli, I., Piccioli, M. Pierattelli & R. <sup>13</sup>C-detected protonless NMR spectroscopy of proteins in solution. *Prog Nucl Magn Reson Spectrosc* **48**, 25–45 (2006).
43. Bermel, W., Felli, I. C., Kümmerle, R. & Pierattelli, R. <sup>13</sup>C Direct-detection biomolecular NMR. *Concepts in Magnetic Resonance Part A* **32A**, 183–200 (2008).
44. Burré, J., Sharma, M. & Südhof, T. C. Cell Biology and Pathophysiology of  $\alpha$ -Synuclein. *Cold Spring Harb Perspect Med* **8**, a024091 (2018).
45. Pirc, K. & Ulrih, N. P.  $\alpha$ -Synuclein interactions with phospholipid model membranes: Key roles for electrostatic interactions and lipid-bilayer structure. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1848**, 2002–2012 (2015).
46. Rcom-H'cheo-Gauthier, A. N., Osborne, S. L., Meedeniya, A. C. B. & Pountney, D. L. Calcium: Alpha-Synuclein Interactions in Alpha-Synucleinopathies. *Front Neurosci* **10**, 570 (2016).
47. Lautenschläger, J. *et al.* C-terminal calcium binding of  $\alpha$ -synuclein modulates synaptic vesicle interaction. *Nat Commun* **9**, 712 (2018).
48. Leal, S. S., Botelho, H. M. & Gomes, C. M. Metal ions as modulators of protein conformation and misfolding in neurodegeneration. *Coord Chem Rev* **256**, 2253–2270 (2012).
49. Faller, P., Hureau, C. & la Penna, G. Metal Ions and Intrinsically Disordered Proteins and Peptides: From Cu/Zn Amyloid- $\beta$  to General Principles. *Acc Chem Res* **47**, 2252–2259 (2014).
50. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences* **37**, 205–211 (1951).

51. Englander, S. W., Mayne, L., Bai, Y. & Sosnick, T. R. Hydrogen exchange: The modern legacy of Linderstrøm-Lang. *Protein Science* **6**, 1101–1109 (1997).
52. 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**, 221–6 (1998).
53. Segawa, T., Kateb, F., Duma, L., Bodenhausen, G. & Pelupessy, P. Exchange Rate Constants of Invisible Protons in Proteins Determined by NMR Spectroscopy. *ChemBioChem* **9**, 537–542 (2008).
54. Thakur, A., Chandra, K., Dubey, A., D'Silva, P. & Atreya, H. S. Rapid Characterization of Hydrogen Exchange in Proteins. *Angewandte Chemie International Edition* **52**, 2440–2443 (2013).
55. Skrynnikov, N. R. & Ernst, R. R. Detection of intermolecular chemical exchange through decorrelation of two-spin order. *J Magn Reson* **137**, 276–80 (1999).
56. Krishnan, V. V. & Murali, N. Radiation damping in modern NMR experiments: Progress and challenges. *Prog Nucl Magn Reson Spectrosc* **68**, 41–57 (2013).
57. Goh, G. K.-M., Dunker, A. K., Foster, J. A. & Uversky, V. N. Rigidity of the Outer Shell Predicted by a Protein Intrinsic Disorder Model Sheds Light on the COVID-19 (Wuhan-2019-nCoV) Infectivity. *Biomolecules* **10**, (2020).
58. Goh, G. K.-M. & Uversky, V. N. Shell disorder and the HIV vaccine mystery: lessons from the legendary Oswald Avery. *J Biomol Struct Dyn* **40**, 5702–5711 (2022).
59. Davey, N. E., Travé, G. & Gibson, T. J. How viruses hijack cell regulation. *Trends Biochem Sci* **36**, 159–169 (2011).
60. Pushker, R., Mooney, C., Davey, N. E., Jacqué, J.-M. & Shields, D. C. Marked Variability in the Extent of Protein Disorder within and between Viral Families. *PLoS One* **8**, e60724 (2013).
61. Xue, B., Mizianty, M. J., Kurgan, L. & Uversky, V. N. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cellular and Molecular Life Sciences* **69**, 1211–1259 (2012).

62. Tamarozzi, E. R. & Giuliatti, S. Understanding the Role of Intrinsic Disorder of Viral Proteins in the Oncogenicity of Different Types of HPV. *Int J Mol Sci* **19**, (2018).
63. Shepley-McTaggart, A., Fan, H., Sudol, M. & Harty, R. N. Viruses go modular. *J Biol Chem* **295**, 4604–4616 (2020).
64. Chang, C.-K. *et al.* 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 (2009).
65. McBride, R., van Zyl, M. & Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **6**, 2991–3018 (2014).
66. Chang, C., Hou, M.-H., Chang, C.-F., Hsiao, C.-D. & Huang, T. The SARS coronavirus nucleocapsid protein – Forms and functions. *Antiviral Res* **103**, 39–50 (2014).
67. Marley, J., Lu, M. & Bracken, C. A method for efficient isotopic labeling of recombinant proteins. *J Biomol NMR* **20**, 71–5 (2001).
68. Fung, T. S. & Liu, D. X. Post-translational modifications of coronavirus proteins: roles and function. *Future Virol* **13**, 405–430 (2018).
69. Lutomski, C. A., El-Baba, T. J., Bolla, J. R. & Robinson, C. v. Multiple Roles of SARS-CoV-2 N Protein Facilitated by Proteoform-Specific Interactions with RNA, Host Proteins, and Convalescent Antibodies. *JACS Au* **1**, 1147–1157 (2021).
70. Ying, W. *et al.* Proteomic analysis on structural proteins of Severe Acute Respiratory Syndrome coronavirus. *Proteomics* **4**, 492–504 (2004).
71. Mark, J. *et al.* SARS coronavirus: unusual lability of the nucleocapsid protein. *Biochem Biophys Res Commun* **377**, 429–433 (2008).
72. Diemer, C. *et al.* Cell type-specific cleavage of nucleocapsid protein by effector caspases during SARS coronavirus infection. *J Mol Biol* **376**, 23–34 (2008).
73. Rangan, R. *et al.* RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA* **26**, 937–959 (2020).

74. Dinesh, D. C. *et al.* Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLoS Pathog* **16**, e1009100 (2020).
75. Cao, C. *et al.* The architecture of the SARS-CoV-2 RNA genome inside virion. *Nat Commun* **12**, 3917 (2021).
76. Redzic, J. S. *et al.* The Inherent Dynamics and Interaction Sites of the SARS-CoV-2 Nucleocapsid N-Terminal Region. *J Mol Biol* **433**, 167108 (2021).
77. Caruso, I. P. *et al.* Insights into the specificity for the interaction of the promiscuous SARS-CoV-2 nucleocapsid protein N-terminal domain with deoxyribonucleic acids. *Int J Biol Macromol* **203**, 466–480 (2022).
78. Forsythe, H. M. *et al.* Multivalent binding of the partially disordered SARS-CoV-2 nucleocapsid phosphoprotein dimer to RNA. *Biophys J* **120**, 2890–2901 (2021).
79. Bessa, L. M. *et al.* The intrinsically disordered SARS-CoV-2 nucleoprotein in dynamic complex with its viral partner nsp3a. *Sci Adv* **8**, (2022).
80. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* **8**, 479–490 (2007).
81. Mackereth, C. D. & Sattler, M. Dynamics in multi-domain protein recognition of RNA. *Curr Opin Struct Biol* **22**, 287–296 (2012).
82. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Mol Cell* **78**, 9–29 (2020).
83. Iserman, C. *et al.* Genomic RNA Elements Drive Phase Separation of the SARS-CoV-2 Nucleocapsid. *Mol Cell* **80**, 1078-1091.e6 (2020).
84. Miao, Z., Tidu, A., Eriani, G. & Martin, F. Secondary structure of the SARS-CoV-2 5'-UTR. *RNA Biol* **18**, 447–456 (2021).
85. Sreeramulu, S. *et al.* Exploring the Druggability of Conserved RNA Regulatory Elements in the SARS-CoV-2 Genome. *Angewandte Chemie International Edition* **60**, 19191–19200 (2021).
86. Schiavina, M. *et al.* Taking Simultaneous Snapshots of Intrinsically Disordered Proteins in Action. *Biophys J* **117**, 46–55 (2019).

87. Pons, M. A “Russian Doll” Approach to More Efficient Acquisition of IDP NMR Spectra. *Biophys J* **117**, 1–2 (2019).
88. Bloembergen, N. & Morgan, L. O. Proton Relaxation Times in Paramagnetic Solutions. Effects of Electron Spin Relaxation. *J Chem Phys* **34**, 842–850 (1961).
89. Battiste, J. L. & Wagner, G. Utilization of Site-Directed Spin Labeling and High-Resolution Heteronuclear Nuclear Magnetic Resonance for Global Fold Determination of Large Proteins with Limited Nuclear Overhauser Effect Data. *Biochemistry* **39**, 5355–5365 (2000).
90. Sharp, R., Lohr, L. & Miller, J. Paramagnetic NMR relaxation enhancement: recent advances in theory. *Prog Nucl Magn Reson Spectrosc* **38**, 115–158 (2001).
91. Venditti, V. & Fawzi, N. L. Probing the Atomic Structure of Transient Protein Contacts by Paramagnetic Relaxation Enhancement Solution NMR. *Methods Mol Biol* **1688**, 243–255 (2018).
92. Savastano, A., Ibáñez de Opakua, A., Rankovic, M. & Zweckstetter, M. Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. *Nat Commun* **11**, 6041 (2020).
93. Perdikari, T. M. *et al.* SARS-CoV-2 nucleocapsid protein phase-separates with RNA and with human hnRNPs. *EMBO J* **39**, (2020).
94. Cubuk, J. *et al.* The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat Commun* **12**, 1936 (2021).
95. Lu, S. *et al.* The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat Commun* **12**, 502 (2021).
96. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-Liquid Phase Separation in Biology. *Annu Rev Cell Dev Biol* **30**, 39–58 (2014).
97. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell* **176**, 419–434 (2019).
98. Yoshizawa, T., Nozawa, R.-S., Jia, T. Z., Saio, T. & Mori, E. Biological phase separation: cell biology meets biophysics. *Biophys Rev* **12**, 519–539 (2020).

99. Das, S., Lin, Y.-H., Vernon, R. M., Forman-Kay, J. D. & Chan, H. S. Comparative roles of charge,  $\pi$ , and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences* **117**, 28795–28805 (2020).
100. Cascarina, S. M. & Ross, E. D. Phase separation by the SARS-CoV-2 nucleocapsid protein: Consensus and open questions. *Journal of Biological Chemistry* **298**, 101677 (2022).
101. Sottini, A. *et al.* Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. *Nat Commun* **11**, 5736 (2020).
102. Heidarsson, P. O. *et al.* Release of linker histone from the nucleosome driven by polyelectrolyte competition with a disordered protein. *Nat Chem* **14**, 224–231 (2022).
103. López-Muñoz, A. D., Kosik, I., Holly, J. & Yewdell, J. W. Cell surface SARS-CoV-2 nucleocapsid protein modulates innate and adaptive immunity. *Sci Adv* **8**, (2022).
104. González-Motos, V., Kropp, K. A. & Viejo-Borbolla, A. Chemokine binding proteins: An immunomodulatory strategy going viral. *Cytokine Growth Factor Rev* **30**, 71–80 (2016).
105. Hernaez, B. & Alcamí, A. Virus-encoded cytokine and chemokine decoy receptors. *Curr Opin Immunol* **66**, 50–56 (2020).
106. Mycroft-West, C. J. *et al.* Heparin Inhibits Cellular Invasion by SARS-CoV-2: Structural Dependence of the Interaction of the Spike S1 Receptor-Binding Domain with Heparin. *Thromb Haemost* **120**, 1700–1715 (2020).
107. Tandon, R. *et al.* Effective Inhibition of SARS-CoV-2 Entry by Heparin and Enoxaparin Derivatives. *J Virol* **95(3)** e01987-20 (2021).
108. Yu, M. *et al.* Elucidating the Interactions Between Heparin/Heparan Sulfate and SARS-CoV-2-Related Proteins—An Important Strategy for Developing Novel Therapeutics for the COVID-19 Pandemic. *Front Mol Biosci* **7**, (2021).
109. Mangiafico, M., Caff, A. & Costanzo, L. The Role of Heparin in COVID-19: An Update after Two Years of Pandemics. *J Clin Med* **11**, 3099 (2022).
110. Shi, C. *et al.* Comprehensive Landscape of Heparin Therapy for COVID-19. *Carbohydr Polym* **254**, 117232 (2021).

111. Kielstein, J. T. *et al.* Hemofiltration with the Seraph® 100 Microbind® Affinity filter decreases SARS-CoV-2 nucleocapsid protein in critically ill COVID-19 patients. *Crit Care* **25**, 190 (2021).
112. Shan, D. *et al.* N-protein presents early in blood, dried blood and saliva during asymptomatic and symptomatic SARS-CoV-2 infection. *Nat Commun* **12**, 1931 (2021).
113. van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720–725 (2016).
114. Mahendran, A. S. K., Lim, Y. S., Fang, C.-M., Loh, H.-S. & Le, C. F. The Potential of Antiviral Peptides as COVID-19 Therapeutics. *Front Pharmacol* **11**, (2020).
115. Bruno, B. J., Miller, G. D. & Lim, C. S. Basics and recent advances in peptide and protein drug delivery. *Ther Deliv* **4**, 1443–1467 (2013).
116. Kovalainen, M. *et al.* Novel Delivery Systems for Improving the Clinical Use of Peptides. *Pharmacol Rev* **67**, 541–561 (2015).
117. Wang, L. *et al.* Therapeutic peptides: current applications and future directions. *Signal Transduct Target Ther* **7**, 48 (2022).
118. Agarwal, G. & Gabrani, R. Antiviral Peptides: Identification and Validation. *Int J Pept Res Ther* **27**, 149–168 (2021).
119. Merrifield, R. B. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *J Am Chem Soc* **85**, 2149–2154 (1963).
120. 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).
121. Ramesh, D., Vijayakumar, B. G. & Kannan, T. Advances in Nucleoside and Nucleotide Analogues in Tackling Human Immunodeficiency Virus and Hepatitis Virus Infections. *ChemMedChem* **16**, 1403–1419 (2021).
122. Pruijssers, A. J. & Denison, M. R. Nucleoside analogues for the treatment of coronavirus infections. *Curr Opin Virol* **35**, 57–62 (2019).
123. Chien, M. *et al.* Nucleotide Analogues as Inhibitors of SARS-CoV-2 Polymerase, a Key Drug Target for COVID-19. *J Proteome Res* **19**, 4690–4697 (2020).



124. Santambrogio, C., Natalello, A., Brocca, S., Ponzini, E. & Grandori, R. Conformational Characterization and Classification of Intrinsically Disordered Proteins by Native Mass Spectrometry and Charge-State Distribution Analysis. *Proteomics* **19**, 1800060 (2019).
125. Nielsen, P. E., Egholm, M. & Buchardt, O. Peptide nucleic acid (PNA). A DNA mimic with a peptide backbone. *Bioconjug Chem* **5**, 3–7 (1994).
126. Singh, K. R., Sridevi, P. & Singh, R. P. Potential applications of peptide nucleic acid in biomedical domain. *Engineering Reports* **2**, (2020).
127. Ribeiro-Filho, H. V. *et al.* Structural dynamics of SARS-CoV-2 nucleocapsid protein induced by RNA binding. *PLoS Comput Biol* **18**, e1010121 (2022).
128. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
129. Piovesan, D., Walsh, I., Minervini, G. & Tosatto, S. C. E. FIELDS: fast estimator of latent local structure. *Bioinformatics* **33**, 1889–1891 (2017).
130. Zhang, Z. *et al.* Structure of SARS-CoV-2 membrane protein essential for virus assembly. *Nat Commun* **13**, 4399 (2022).

## **Article 1.1.**

**Monitoring the Interaction of  $\alpha$ -Synuclein with Calcium Ions through Exclusively Heteronuclear Nuclear Magnetic Resonance Experiments**

VIP Protein Function Very Important Paper

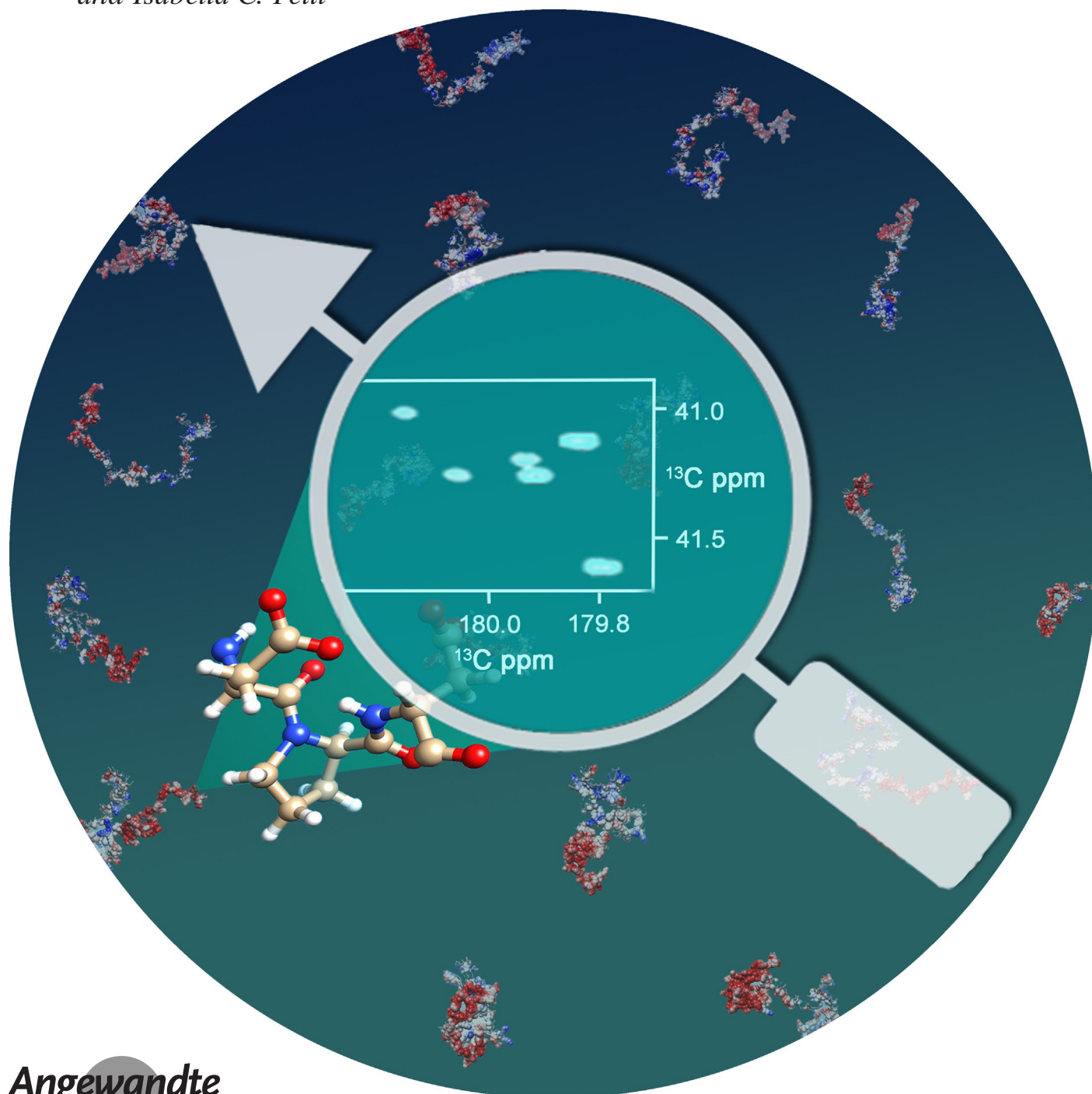
How to cite: *Angew. Chem. Int. Ed.* **2020**, *59*, 18537–18545

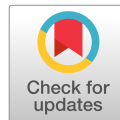
International Edition: doi.org/10.1002/anie.202008079

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

# Monitoring the Interaction of $\alpha$ -Synuclein with Calcium Ions through Exclusively Heteronuclear Nuclear Magnetic Resonance Experiments

Letizia Pontoriero<sup>+</sup>, Marco Schiavina<sup>+</sup>, Maria Grazia Murrari, Roberta Pierattelli,<sup>\*</sup> and Isabella C. Felli<sup>\*</sup>





# Angewandte GDCh Chemie

Eine Zeitschrift der Gesellschaft Deutscher Chemiker

[www.angewandte.de](http://www.angewandte.de)

## Akzeptierter Artikel

**Titel:** Zooming into the Interaction of  $\alpha$ -synuclein with Calcium Ions through Exclusively Heteronuclear NMR Experiments

**Autoren:** Letizia Pontoriero, Marco Schiavina, Maria Grazia Murrari, Roberta Pierattelli, and Isabella Caterina Felli

Dieser Beitrag wurde nach Begutachtung und Überarbeitung sofort als "akzeptierter Artikel" (Accepted Article; AA) publiziert und kann unter Angabe der unten stehenden Digitalobjekt-Identifizierungsnummer (DOI) zitiert werden. Die deutsche Übersetzung wird gemeinsam mit der endgültigen englischen Fassung erscheinen. Die endgültige englische Fassung (Version of Record) wird ehestmöglich nach dem Redigieren und einem Korrekturgang als Early-View-Beitrag erscheinen und kann sich naturgemäß von der AA-Fassung unterscheiden. Leser sollten daher die endgültige Fassung, sobald sie veröffentlicht ist, verwenden. Für die AA-Fassung trägt der Autor die alleinige Verantwortung.

**Zitierweise:** *Angew. Chem. Int. Ed.* 10.1002/anie.202008079

**Link zur VoR:** <https://doi.org/10.1002/anie.202008079>

# Zooming into the Interaction of $\alpha$ -synuclein with Calcium Ions through Exclusively Heteronuclear NMR Experiments

Letizia Pontoriero<sup>^</sup>, Marco Schiavina<sup>^</sup>, Maria Grazia Murralli<sup>#</sup>, Roberta Pierattelli<sup>\*</sup> and Isabella C. Felli<sup>\*</sup>

[\*] 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  
E-mail: felli@cerm.unifi.it

<sup>^</sup> These authors contributed equally to this work.

<sup>#</sup> Present address: Department of Chemistry and Biochemistry, University of California at Los Angeles, USA

Supporting information for this article is given via a link at the end of the document.

**Abstract:** Many properties of intrinsically disordered proteins (IDPs) or protein regions (IDRs) are expected to be modulated by the nature of amino acid side chains as well as by local solvent exposure. Here we propose a set of *exclusively heteronuclear* NMR experiments to investigate these features in different experimental conditions relevant for physiological function. The proposed approach is generally applicable to many IDPs/IDRs whose assignment is available in the BMRB to investigate how their properties are modulated by different, physiologically relevant conditions. The experiments are tested on  $\alpha$ -synuclein. They are then used to investigate how  $\alpha$ -synuclein senses  $\text{Ca}^{2+}$  concentration jumps associated with the transmission of nervous signals. Novel modules in the primary sequence of  $\alpha$ -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<sup>[1-4]</sup>. 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"<sup>[1]</sup> 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<sup>[5,6]</sup>. 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 malfunction. In this framework, NMR spectroscopy provides a unique investigation tool to access high resolution information<sup>[7,8]</sup>.

Human  $\alpha$ -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,  $\alpha$ -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- $\beta$  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<sup>[9]</sup>, stabilized by long range interactions between the N-terminal region and the C-terminal one<sup>[10,11]</sup>, it adopts helical conformations when interacting with membranes through the N-terminal region<sup>[12-15]</sup> and elongated conformations in amyloid fibrils<sup>[16]</sup>, just to name a few snapshots on the most studied, heterogeneous structural properties of  $\alpha$ -synuclein in different conditions<sup>[17]</sup>. 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"<sup>[18]</sup>, 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<sup>[19-25]</sup>, 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  $\alpha$ -synuclein and then used to focus on its interaction with  $\text{Ca}^{2+}$ , a potential trigger for the onset of Parkinson's disease. Mainly localized in presynaptic terminals,  $\alpha$ -synuclein is exposed to microdomains of high  $\text{Ca}^{2+}$  concentration associated with neurotransmitter release<sup>[26,27]</sup> and could be exposed to high extracellular  $\text{Ca}^{2+}$  concentration in cell-

to-cell secretion mechanisms<sup>[28]</sup>. New insights on how a structurally and dynamically heterogeneous protein linked to the onset of Parkinson's disease senses these  $\text{Ca}^{2+}$  concentration jumps become thus very relevant to describe  $\alpha$ -synuclein function.

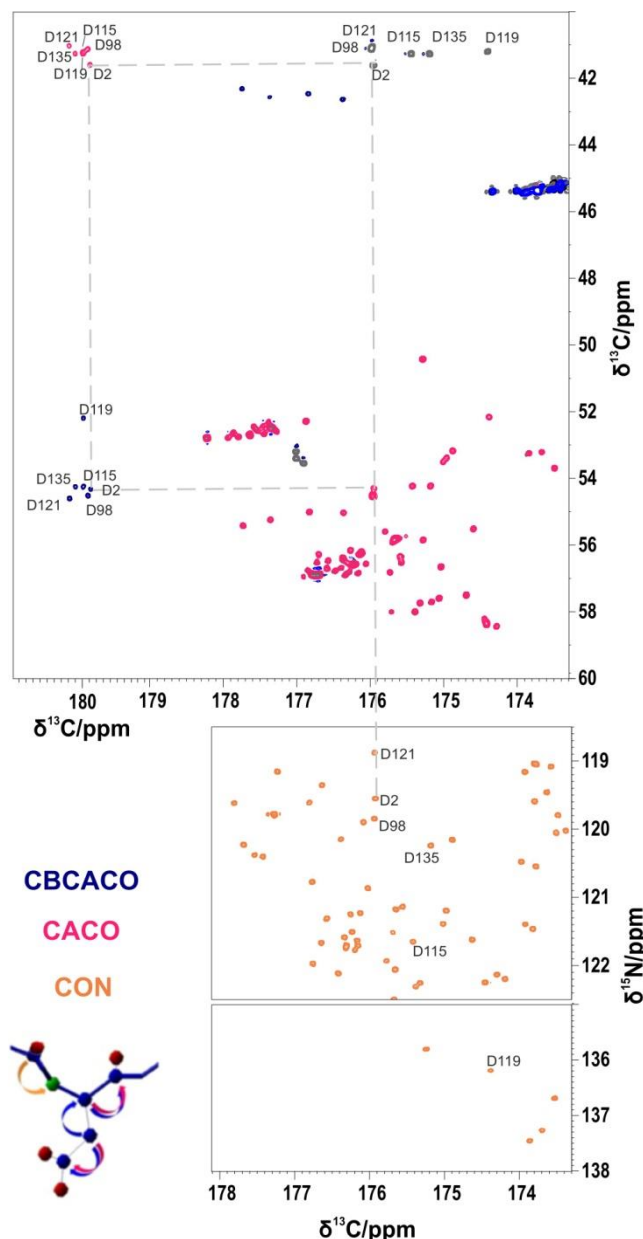
## Results and Discussion

### Fingerprint of an IDP at physiological pH and temperature: $^{13}\text{C}$ -detected 2D spectra

Amino acids' side chains 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  $^1\text{H}$ - $^{13}\text{C}$  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 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)<sup>[29,30]</sup>. 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 (Figure S1); experimental variants exploiting  $^1\text{H}$  polarization as a starting source ( $^1\text{H}$ -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 ( $-\text{COO}^-$ ,  $-\text{CONH}_2$ ). 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  $\alpha$ -synuclein is shown in Figure 1. As illustrated for Asp 2, starting from the backbone carbonyl ( $\text{C}'_i$ ) identified through the CON ( $\text{C}'_i\text{-N}_{i+1}$ ), the resonances of  $\text{C}^\alpha_i$  and  $\text{C}^\beta_i$  can be easily identified through inspection of the CACO and CBCACO. These are also correlated to the side chain carboxylate carbon resonance ( $\text{C}^\gamma_i$ ) through two additional cross peaks in a close but well isolated spectral region. Therefore, the  $\text{C}^\beta_i\text{-C}^\gamma_i$  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  $\text{C}^\alpha_i$ ,  $\text{C}^\beta_i$ ,  $\text{C}^\gamma_i$  aliphatic side chain resonances and finally to the  $\text{C}^\delta_i$  carbonyl/carboxylate one. The cross peaks assignment of the carboxylate/carbonyl functional groups for  $\alpha$ -synuclein ( $\text{C}^\beta_i\text{-C}^\gamma_i$  for Asp and Asn,  $\text{C}^\gamma_i\text{-C}^\delta_i$  for Glu and Gln), which fall in a very clean spectral region, is reported in Table S2.

It is worth noting that carbonyl/carboxylate side chain resonances are seldom assigned in general and for IDPs/IDRs in particular. They can be detected through triple resonance experiments based on amide proton detection<sup>[31-33]</sup>. However this approach is bound to fail in conditions in which amide protons are not detectable, such as for solvent exposed protein



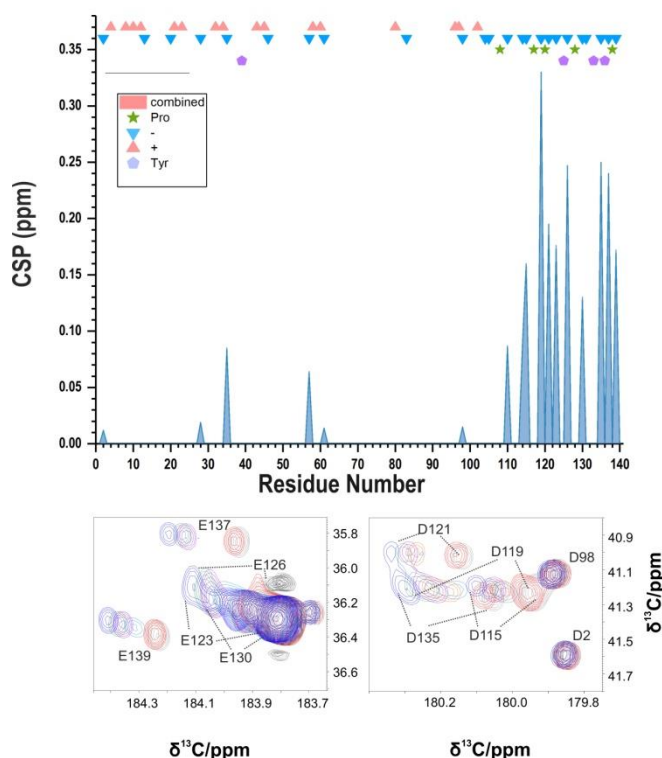
**Figure 1.** The figure illustrates the strategy used to obtain the sequence specific assignment of the  $^{13}\text{C}$  resonances of aspartate residues through 2D *exclusively heteronuclear* NMR experiments. As an example, gray dotted lines indicate the steps followed to assign side chains resonances of Asp 2. Starting from the carbonyl resonance identified in the CON spectrum (orange),  $\text{C}^\alpha_i$  and  $\text{C}^\beta_i$  are identified in CACO (red) and CBCACO (blue) spectra, superimposed in the figure and then correlated to  $\text{C}^\gamma_i$  through the respective  $\text{C}^\beta_i\text{-C}^\gamma_i$  and  $\text{C}^\alpha_i\text{-C}^\gamma_i$  cross-peaks in a sequence specific manner.

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<sup>[34]</sup> as well as to interact with metal ions<sup>[35–39]</sup>. Particularly relevant for  $\alpha$ -synuclein function is the interaction with  $\text{Ca}^{2+}$  involved in the transmission of nervous signals<sup>[39–41]</sup>. While intracellular  $\text{Ca}^{2+}$  concentrations are generally very low, microdomains of high  $\text{Ca}^{2+}$  concentrations are linked to the release of neurotransmitter from presynaptic terminals in all neurons<sup>[26]</sup>. Having the assignment in hand, it is now possible to “zoom-in” into the metal ion coordination sphere and access additional complementary information to that available through HN HSQC experiments<sup>[39]</sup>.

The set of 2D *exclusively heteronuclear* NMR experiments, in particular the CON and the CACO, was used to monitor the changes in  $\alpha$ -synuclein induced by the presence of  $\text{Ca}^{2+}$ . The chemical shift changes of Asp/Glu residues signals upon addition of  $\text{Ca}^{2+}$ , 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 resonances (C') (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).



**Figure 2.** Chemical Shift Perturbation (difference in absolute value) of aspartate and glutamate side chains  $^{13}\text{C}$  resonances upon addition of  $\text{Ca}^{2+}$ . The symbols over the graph depict the distribution of charged, tyrosine and proline residues to evidence the peculiar composition along the primary sequence: Asp and Glu (triangles), Lys (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, rich in negative charges.

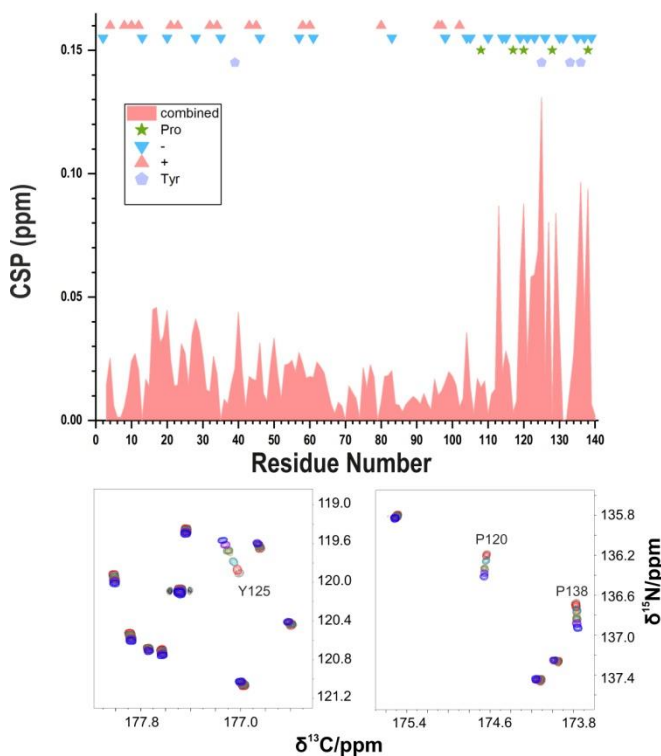
To assess the general applicability of this set of  $^{13}\text{C}$  detected experiments, the interaction studies with calcium ions were repeated using a sample with an order of magnitude lower concentration of  $\alpha$ -synuclein (50  $\mu\text{M}$ ). Despite the relatively low sensitivity of  $^{13}\text{C}$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

Shifting our attention to backbone nuclear spins by inspection of the combined chemical shift changes of C' and N chemical shifts<sup>[42]</sup>, not only direct but also indirect changes derived from the interaction with  $\text{Ca}^{2+}$  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  $\text{Ca}^{2+}$ , indicating that they are involved in the interaction of  $\alpha$ -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  $\text{Ca}^{2+}$  addition. These observations prompted us to inspect the 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  $\alpha$ -synuclein experiencing the most pronounced changes upon  $\text{Ca}^{2+}$  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<sup>[43,44]</sup>. It is thus interesting to investigate whether the interaction with  $\text{Ca}^{2+}$  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)<sup>[45]</sup>, has been one of the first NMR observables used in the past to identify amide protons protected from solvent exchange by globular protein folds<sup>[46]</sup>. On the other hand, only a little information is available so far on how solvent exchange is



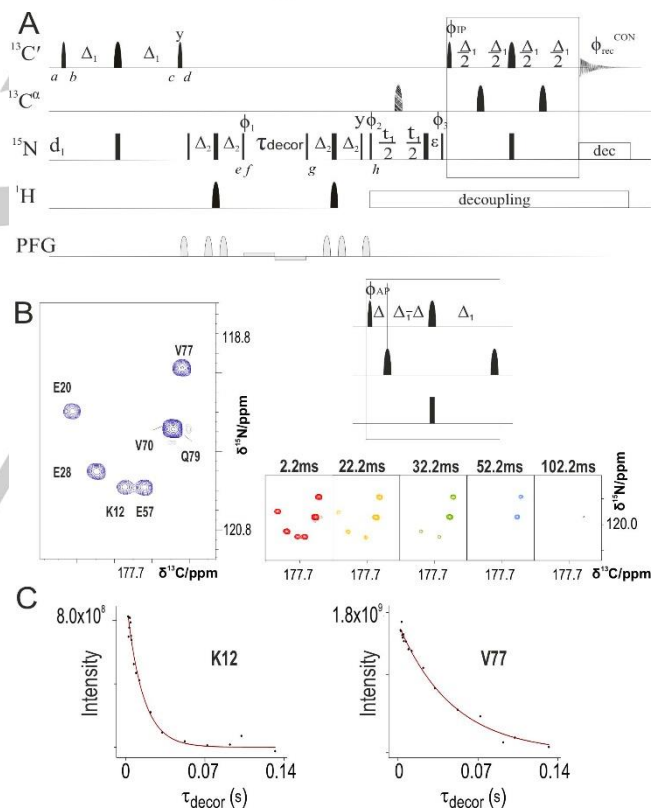
**Figure 3.** Chemical Shift Perturbation  $\sqrt{(0,15\delta(^{15}\text{N}_i))^2 + 0,34\delta(^{13}\text{C}_i)^2}$  [42] of CON signals upon addition of  $\text{Ca}^{2+}$ . The lower panels show the shift for the two proline residues (Pro 120 and Pro 138 (right)) and for tyrosine 125 (left), the most 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 evidences impact also for many residues in the N-terminal domain. The titration follows the pattern from blue ( $\alpha\text{-syn}:\text{Ca}^{2+}$ , 1:0) to purple ( $\alpha\text{-syn}:\text{Ca}^{2+}$ , 1:16).

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 ( $4\text{C}'_z\text{N}_z\text{H}_z$ ) and to monitor its decorrelation due to chemical exchange, along the lines of a method initially proposed by Skrynnikov and Ernst<sup>[47]</sup>.

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  $^1\text{H}$ - $^{15}\text{N}$ -based experiments. Starting from  $\text{C}'_z$  magnetization (a), transverse  $\text{C}'_y$  coherence is created (b) and converted into antiphase coherence  $2\text{C}'_x\text{N}_z$  (c). This is then converted into  $2\text{C}'_z\text{N}_y$  (d) and allowed to evolve under the effect of the  $^1\text{J}_{\text{HN}}$  coupling to generate  $4\text{C}'_z\text{N}_x\text{H}_z$  (e). In order to generate this latter operator a band-selective  $180^\circ$  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  $4\text{C}'_z\text{N}_z\text{H}_z$  (f) and its decay is monitored by introducing a free evolution delay ( $\tau_{\text{decor}}$ ). At the end of this delay the three spin order  $4\text{C}'_z\text{N}_z\text{H}_z$  is converted to  $4\text{C}'_z\text{N}_y\text{H}_z$  (g) which is picked up by the second part of the CON after conversion into  $2\text{C}'_z\text{N}_y$  (h). It is worth noting that through this approach a dependence on solvent exchange is reintroduced with minimal perturbation of the water resonance

avoiding 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  $\tau_{\text{decor}}$ , the time interval in which the three spin order  $4\text{C}'_z\text{N}_z\text{H}_z$  is allowed to evolve: while the signal of Val 77 is still observable with  $\tau_{\text{decor}} = 52.2$  ms, the one of Lys 12 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  $\tau_{\text{decor}}$  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  $^1\text{H}$  detected variant<sup>[47]</sup> (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  $^1\text{H}$  detected variant. This is in part due to the improved resolution deriving from the superior chemical shift dispersion of the  $\text{C}'_{i-1}\text{-N}_i$  correlations respect to the  $\text{H}_i\text{-N}_i$  ones, and in part to the different



**Figure 4.** (A) DeCON pulse sequence. The following phase cycling was employed:  $\phi_1 = 2(y)$ ,  $2(-y)$ ;  $\phi_2 = x$ ,  $-x$ ;  $\phi_3 = 4(x)$ ,  $4(-x)$ ,  $\phi^{\text{IP}} = x$ ;  $\phi^{\text{AP}} = -y$  and  $\phi_{\text{rec}} = x$ ,  $-x$ ,  $x$ ,  $-x$ ,  $x$ ,  $-x$ ,  $x$ ,  $-x$ . The length of the delays were:  $\Delta = 4.5$  ms  $\Delta_1 = 16.6$  ms;  $\Delta_2 = 2.7$  ms;  $\epsilon = \tau_1(0) + p180$  (500  $\mu\text{s}$ ). The striped pulse in the middle of the  $^{15}\text{N}$  evolution period is an adiabatic Chirp pulse that covers the whole  $^{13}\text{C}$  spectral region. Virtual decoupling of the  $\text{C}'\text{-C}^\alpha$  coupling was achieved by acquiring for each increment both the IP and AP component of the signals. 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-TPP1 approach incrementing phase  $\phi_2$ . (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  $\tau_{\text{decor}}$  are shown on the right. The intensities (arbitrary unit) of two of these cross peaks are reported as a function of  $\tau_{\text{decor}}$  in panel C.



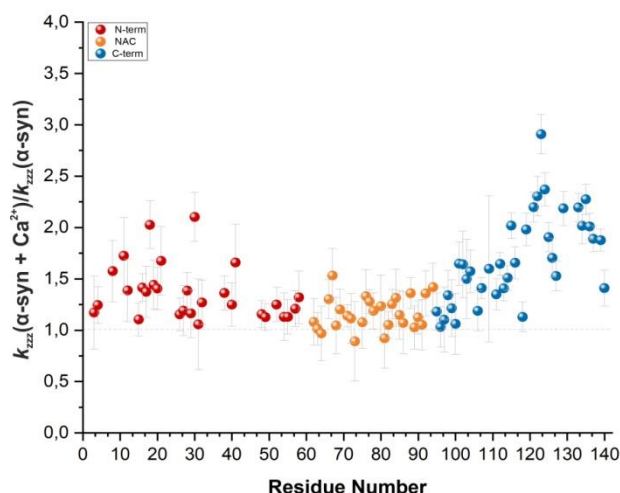
magnetization transfer pathway minimizing perturbation of  $^1\text{H}$  magnetization (in the HN coherence transfer pathway the proton magnetization is transverse both during the two INEPT steps and the acquisition, in the DeCON the proton magnetization is maintained along the z-axis). These properties 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  $^1\text{H}$  detected variant. The minor contribution of longitudinal relaxation to the observed decay was evaluated investigating the decay of the  $2\text{C}'_z\text{N}_z$  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  $\text{C}'$  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<sup>[9]</sup>. 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  $\text{Ca}^{2+}$  to the  $\alpha$ -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 x1.9 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  $\text{CaCl}_2$ . 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.

#### $\text{Ca}^{2+}$ sensing by $\alpha$ -synuclein: new insights

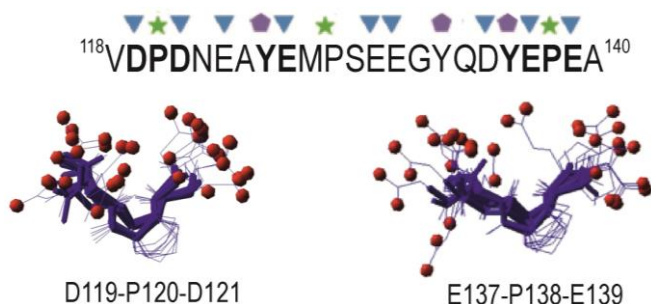
The high number of negatively charged amino acids in the final part of the primary sequence of  $\alpha$ -synuclein, in combination with the disordered nature of this protein that leaves this part of the backbone largely solvent exposed and easily accessible, provides a strong electrostatic negative 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



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

driving force, sensed at quite long distance, for the interaction with  $\text{Ca}^{2+}$  [35]. The negatively charged functional groups of amino acid side chains, such as carboxylate groups of aspartate and glutamate residues ( $\text{COO}^-$ ), in principle all have the potential to interact with positively charged metal ions. The disordered nature of the polypeptide that leaves  $\text{COO}^-$  groups largely exposed to the solvent would suggest an unspecific effect, with all  $\text{COO}^-$  sharing similar interaction properties. Instead, we find very specific differential effects for the  $\text{COO}^-$  groups in the different parts of the polypeptide chain. The sequence context thus has an important role in mediating interactions with  $\text{Ca}^{2+}$  even in the case of IDPs. The possibility to directly observe perturbations sensed by  $\text{COO}^-$  groups through  $^{13}\text{C}$  detected experiments allows us to zoom-in into the interaction site and identify the amino acids that are the most perturbed ones 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  $\text{Ca}^{2+}$  already has a strong preference for a subset of the  $\text{COO}^-$  groups in the C-terminal region in which  $\alpha$ -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 ones 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  $\pi$  electron density, two properties that could play a relevant role in reducing local motions and favoring interactions of highly flexible protein regions with  $\text{Ca}^{2+}$ , in particular if followed by an acidic residue providing a  $\text{COO}^-$  group. These could be key elements of specific motifs to modulate  $\text{Ca}^{2+}$  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.



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

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 chains 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  $\text{Ca}^{2+}$ . 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  $\text{Ca}^{2+}$  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  $\alpha$ -synuclein with  $\text{Ca}^{2+}$  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 motives 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.

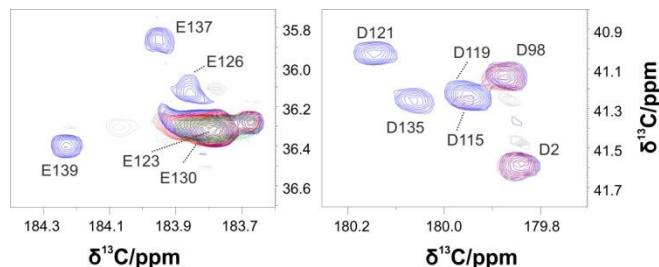
The interaction of  $\alpha$ -synuclein with  $\text{Ca}^{2+}$  is likely to disrupt the interaction of the C-terminal region with the N-terminal one, rich of lysine residues. This region adopts a helical conformation when bound to membranes<sup>[12-14]</sup> and was proposed to form long range interactions at the origin of a compact state of  $\alpha$ -synuclein populated in solution<sup>[10,11,48,50]</sup>. 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  $\text{COO}^-$  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  $\text{Ca}^{2+}$  addition would require additional experiments, for example exploiting paramagnetic relaxation enhancements induced by the presence of a paramagnetic tag, which might take advantage of  $^{13}\text{C}$ -NMR detection experiments<sup>[51]</sup>.

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  $\alpha$ -synuclein with  $\text{Ca}^{2+}$  reveals specific motives 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 motives have been investigated at high resolution and many more could be studied in detail with the tools proposed here. The characteristic features identified in  $\alpha$ -synuclein might also be useful as input for bioinformatics tools to search for similar  $\text{Ca}^{2+}$  binding patterns in disordered proteins.

$\alpha$ -synuclein has also been shown to interact with other metal ions<sup>[36-38,52]</sup>. Among them Cu(II), Fe(II), Co(II), Ni(II), Mn(II), and several lanthanide metal ions<sup>[38,53,54]</sup> 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<sup>[55]</sup>. Previous NMR investigations revealed that Mn(II) interacts with residues in the C-terminal region of  $\alpha$ -synuclein, in a very similar way to what observed for  $\text{Ca}^{2+}$ <sup>[38]</sup>. 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  $\text{Ca}^{2+}$  (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  $\alpha$ -synuclein amino acid side chains and to monitor local solvent exposure. These experiments allowed us to zoom into the metal ion coordination sphere, revealing novel motives 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 modulating protein function that can be studied through the approach presented here. The proposed experiments can thus become a tool of general interest to characterize properties of IDPs/IDRs in physiologically relevant conditions that have not been studied so far, 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 in part by the Italian Ministry for University and Research (FOE funding) and by a grant of the Fondazione CR Firenze to RP.

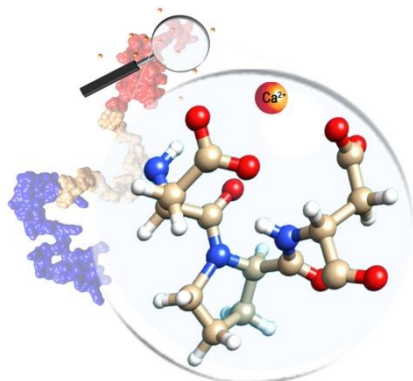
**Keywords:** IDPs •  $^{13}C$ -detection • Calcium binding • Side Chains • Water Exchange

- [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* **2014**, *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-Kazimierczuk, 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] I. C. Felli, R. Pierattelli, *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*, **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 Buntum, S. Vega, A. Webb, C. Luchinat, H. Schwalbe, L. Frydman, *Angew. Chemie - Int. Ed.* **2015**, *54*, 9162–9185.
- [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. Bertoncini, Y. Jung, C. O. Fernandez, W. Hoyer, C. Griesinger, T. M. Jovin, M. Zweckstetter, *Proc. Natl. Acad. Sci.* **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*.
- [16] M. D. Tuttle, G. Comellas, A. J. Nieuwkoop, D. J. Covell, D. A. Berthold, K. D. Kloepper, 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. Bernel, 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. Chemie - Int. Ed.* **2011**, *50*, 2339–2341.
- [21] S. Gil, T. Hošek, Z. Solyom, R. Kümmerle, B. Brutscher, R. Pierattelli, I. C. Felli, *Angew. Chemie* **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. Bouguchtouli, M. Julien, W. Bernel, R. Ghouil, S. Zinn-Justin, F. X. Theillet, *Angew. Chemie - Int. Ed.* **2020**, 1–6.
- [26] R. Llinás, M. Sugimori, R. B. Silver, *Science*. **1992**, *256*, 677–679.
- [27] R. R. E. M. Neuroglia, C. S. Harbor, *Nature* **2000**, *406*, 889–893.
- [28] H. J. Lee, E. J. Bae, S. J. Lee, *Nat. Rev. Neuro.* **2014**, *10*, 92–98.
- [29] W. Bernel, I. Bertini, L. Duma, I. C. Felli, L. Emsley, R. Pierattelli, P. R. Vasos, *Angew. Chemie - Int. Ed.* **2005**, *44*, 3089–3092.
- [30] W. Bernel, 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. U. S. A.* **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. Nespoivita, 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*.
- [40] A. Rcom-H'cheo-Gauthier, J. Goodwin, D. L. Pountney, *Biomolecules* **2014**, *4*, 795–811.
- [41] M. R. Post, O. J. Lieberman, E. V. Mosharov, *Front. Neurosci.* **2018**, *12*, 1–11.
- [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. Platzner, F. Tribuzio, R. Konrat, *J. Mol. Biol.* **2018**, *430*, 2439–2452.
- [45] M. Schiavina, M. G. Murrall, L. Pontoriero, V. Sainati, R. Kümmerle, W. Bermel, R. Pierattelli, I. C. Felli, *Biophys. J.* **2019**, 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* **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 Mol. Biol.* (Ed.: D. Eliezer), Springer New York, New York, NY, **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*, 1–10.
- [55] I. Bertini, C. Luchinat, G. Parigi, E. Ravera, *NMR of Paramagnetic Molecules*, Elsevier, **2017**.

## Entry for the Table of Contents



In this paper a set of *exclusively heteronuclear* NMR experiments, tailored for zooming into IDPs' side chains, is presented. These experiments are used to monitor the interaction of  $\alpha$ -synuclein with  $\text{Ca}^{2+}$  ions. Moreover, local solvent exposure, in presence and absence of  $\text{Ca}^{2+}$ , is studied with a  $^{13}\text{C}$ -direct detection approach. Novel modules, optimized for calcium sensing, are identified in the primary sequence of  $\alpha$ -synuclein.

Institute and/or researcher Twitter usernames: CERM institute twitter: @cerm\_cirmmp  
Letizia Pontoriero twitter: @letizia\_ponto  
Robera Pierattelli twitter: @rpierattelli

## Supporting Information

### **Monitoring the Interaction of $\alpha$ -Synuclein with Calcium Ions through Exclusively Heteronuclear Nuclear Magnetic Resonance Experiments**

*Letizia Pontoriero<sup>+</sup>, Marco Schiavina<sup>+</sup>, Maria Grazia Murralli, Roberta Pierattelli,\* and Isabella C. Felli\**

ange\_202008079\_sm\_miscellaneous\_information.pdf

**Table of Contents**

<b>Experimental Procedure</b> .....	3
Samples preparation.....	3
NMR Spectroscopy .....	3
<b>Supplementary Figures</b> .....	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
<b>Supplementary Tables</b> .....	17
Supplementary Table 1 .....	17
Supplementary Table 2 .....	18
<b>References</b> .....	19
<b>Author contributions</b> .....	19

## Experimental Procedures

**Samples preparation.** Three  $^{13}\text{C}$ ,  $^{15}\text{N}$  labelled  $\alpha$ -synuclein samples were prepared as previously described in the literature<sup>[1]</sup>, lyophilized in water and stocked at  $-20^\circ\text{C}$ . Experiments were acquired using a first sample in a 5 mm NMR tube containing 500  $\mu\text{L}$  of  $\alpha$ -synuclein with a concentration of 600  $\mu\text{M}$  in 20 mM Tris-Cl buffer, pH=7.4. 2%  $\text{D}_2\text{O}$  was added for the lock. For the titration experiments, a batch of  $\text{CaCl}_2$  solution was prepared with a final concentration of 100 mM in milliQ water. Seven additions of the  $\text{CaCl}_2$  stock were performed in order to reach the following ratios, in equivalents, of  $\alpha$ -synuclein to calcium ions: 1:0 – 1:1 – 1:2 – 1:4 – 1:6 – 1:8 – 1:10 – 1:16.

A second  $\alpha$ -synuclein sample was used to perform the  $\text{Mn}^{2+}$  titration. Experiments were acquired in a 5 mm NMR tube containing 500  $\mu\text{L}$  of  $\alpha$ -synuclein with a concentration of 600  $\mu\text{M}$  in 20 mM Tris-Cl buffer, pH=7.4. 2%  $\text{D}_2\text{O}$  was added for the lock. For the titration experiments, two batches of  $\text{MnCl}_2$  solution were prepared with a final concentration of 2.5 mM and 25 mM in milliQ water. Three additions of the  $\text{MnCl}_2$  stocks were performed in order to reach the following ratios, in equivalents, of  $\alpha$ -synuclein to manganese ions: 1:0 – 1:0.01 – 1:0.02 – 1:0.1.

A third more diluted  $\alpha$ -synuclein sample was prepared to perform  $\text{Ca}^{2+}$  titration. Experiments were acquired using a sample in a 5 mm NMR tube containing 500  $\mu\text{L}$  of  $\alpha$ -synuclein with a concentration of 50  $\mu\text{M}$  in 20 mM Tris-Cl buffer, pH=7.4. 2%  $\text{D}_2\text{O}$  was added for the lock. For the titration experiments, three batches of  $\text{CaCl}_2$  solution were prepared with a final concentration of 10 mM, 100 mM, and 1 M in buffer. Eight additions of the  $\text{CaCl}_2$  stocks were performed in order to reach the following ratios, in equivalents, of  $\alpha$ -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  $^1\text{H}$ , 176.05 MHz  $^{13}\text{C}$ , and 70.97 MHz  $^{15}\text{N}$  frequencies, equipped with a cryogenically cooled probehead optimized for  $^{13}\text{C}$ -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  $^1\text{H}$  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.  $^{13}\text{C}$  pulses were given at 176.1 ppm, 55.6 ppm and 42.6 ppm for  $\text{C}'$ ,  $\text{C}^\alpha$  and  $\text{C}^{\text{all}}$  regions.  $^{15}\text{N}$  pulses were given at 125.0 ppm (for CON experiments) or 118.0 ppm (for HN experiments). Q5 and Q3 shapes<sup>[2]</sup> of durations of 300 and 231  $\mu\text{s}$ , respectively, were used for  $^{13}\text{C}$  band-selective  $\pi/2$  and  $\pi$  flip angle pulses except for the  $\pi$  pulses that should be band-selective on the  $\text{C}^\alpha$  region (Q3, 900  $\mu\text{s}$ ), and for the adiabatic  $\pi$  pulse<sup>[3]</sup> to invert both  $\text{C}'$  and  $\text{C}^\alpha$  (smoothed Chirp 500  $\mu\text{s}$ , 20 % smoothing, 80 kHz sweep width, 11.3 kHz RF field strength). Composite pulse decoupling was applied on  $^1\text{H}$  (Waltz-16)<sup>[4]</sup> and  $^{15}\text{N}$  (Garp-4)<sup>[5]</sup> with an RF field strength of 3kHz and 1kHz respectively.  $^{13}\text{C}$  homonuclear decoupling was achieved through the IPAP virtual decoupling approach<sup>[6]</sup>. For DeCON experiments, both for the 2D and 3D version, a Reburp shape<sup>[7]</sup> of duration of 2076  $\mu\text{s}$  was used for  $^1\text{H}$  band-selective  $\pi$  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  $^1\text{H}$  and  $^{13}\text{C}$ ;  $^{15}\text{N}$  shifts were calibrated indirectly.



## Supplementary Figures

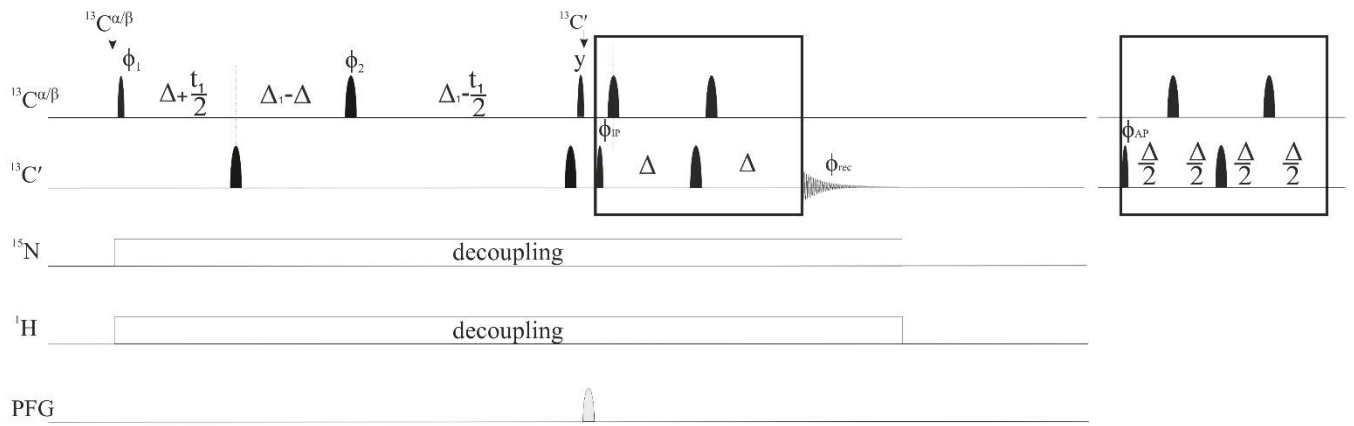
**Figure S1.** Pulse sequences used to acquire CACO, CBCACO, CCCO spectra. Narrow and wide black bars represent  $\pi/2$  and  $\pi$  non-selective pulses; narrow and wide rounded black bars represent  $\pi/2$  and  $\pi$  band-selective pulses. The pulse sequence elements reported in the boxes represent the two variants to acquire the in-phase (IP) and antiphase (AP) components of carbonyl signals needed to achieve  $^{13}\text{C}$  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  $\phi_1$ .

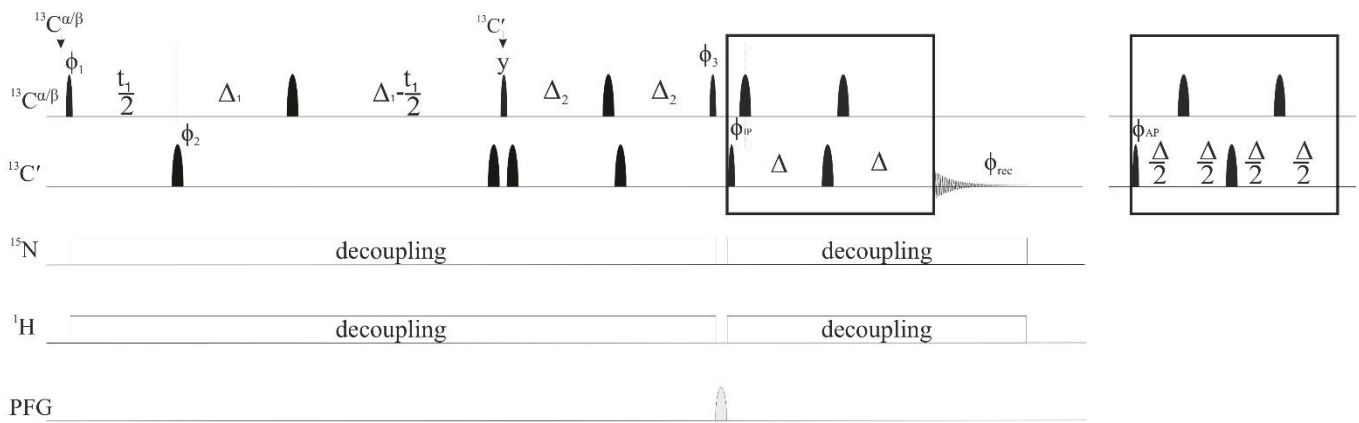
For CBCACO, the following phase cycling was employed:  $\phi_1 = x, -x$ ;  $\phi_2 = 8(x), 8(-x)$ ,  $\phi_3 = 2(y), 2(-y)$ ,  $\phi^{\text{IP}} = 4(x), 4(-x)$ ;  $\phi^{\text{AP}} = 4(-y), 4(y)$  and  $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x$ . The length of the delays was:  $\Delta = 4.5$  ms;  $\Delta_1 = 11.7$  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  $\phi_1$ .

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  $\phi_1$ . The  $^{13}\text{C}$  spinlock was applied with an RF field strength of 10kHz ( $1/4 \times 90$ ) with a FLOPSY sequence<sup>[8]</sup>. The grey pulse is a band-selective pulse on the  $\text{C}^{\alpha}$  region.

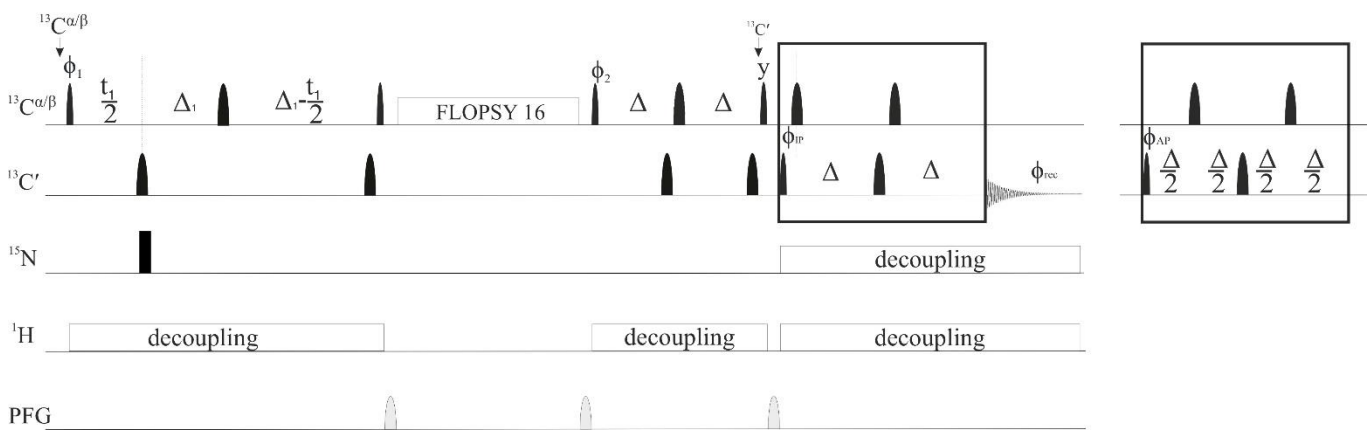
# CACO



# CBCACO



# CCCO



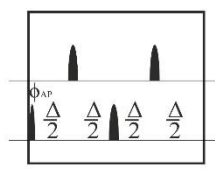
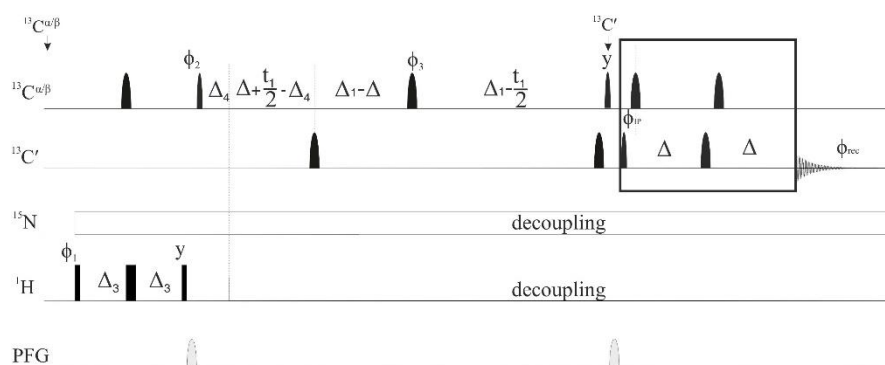
**Figure S2.** Pulse sequences used to acquire (H)CACO, (H)CBCACO, (H)CCCO spectra. Narrow and wide black bars represent  $\pi/2$  and  $\pi$  non-selective pulses; narrow and wide rounded black bars represent  $\pi/2$  and  $\pi$  band-selective pulses. The pulse sequence elements reported in the boxes represent the two variants to acquire the in-phase (IP) and antiphase (AP) components of carbonyl signals needed to achieve  $^{13}\text{C}$  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  $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x, -x, x, x, -x, x, -x, -x, x$ . 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  $\phi_2$ .

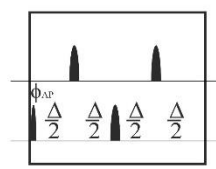
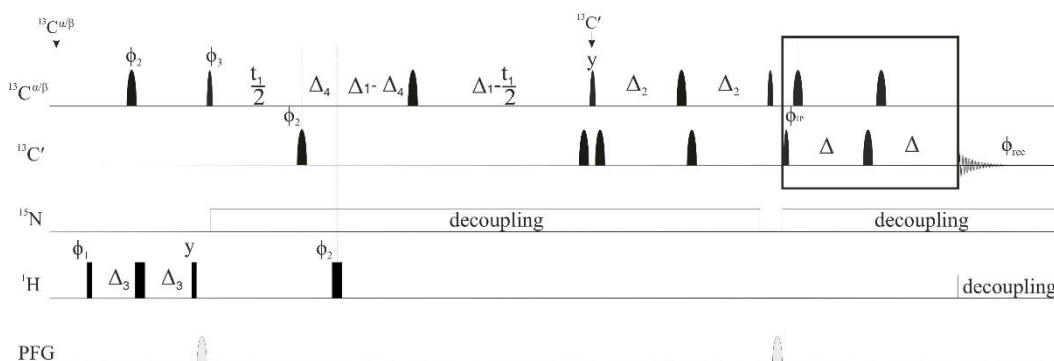
For (H)CBCACO, the following phase cycling was employed:  $\phi_1 = x, -x$ ;  $\phi_2 = 8(x), 8(-x)$ ,  $\phi_3 = 2(x), 2(-x)$ ;  $\phi_{\text{IP}} = 4(x), 4(-x)$ ;  $\phi_{\text{AP}} = 4(-y), 4(y)$  and  $\phi_{\text{rec}} = x, -x, -x, x, -x, x, x, -x$ . The length of the delays was:  $\Delta = 4.5$  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  $\phi_3$ .

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  $\phi_2$ . The  $^{13}\text{C}$  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 $\alpha$  region.

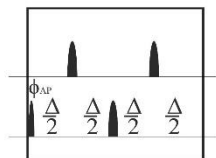
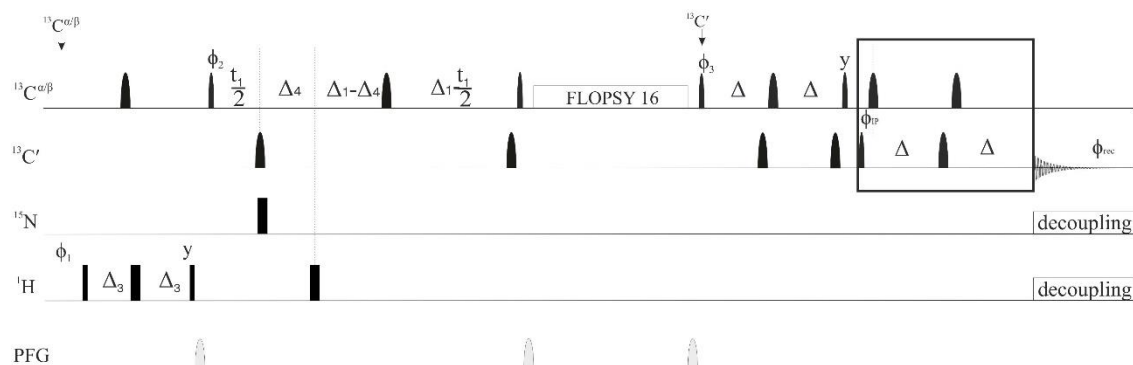
### (H)CACO



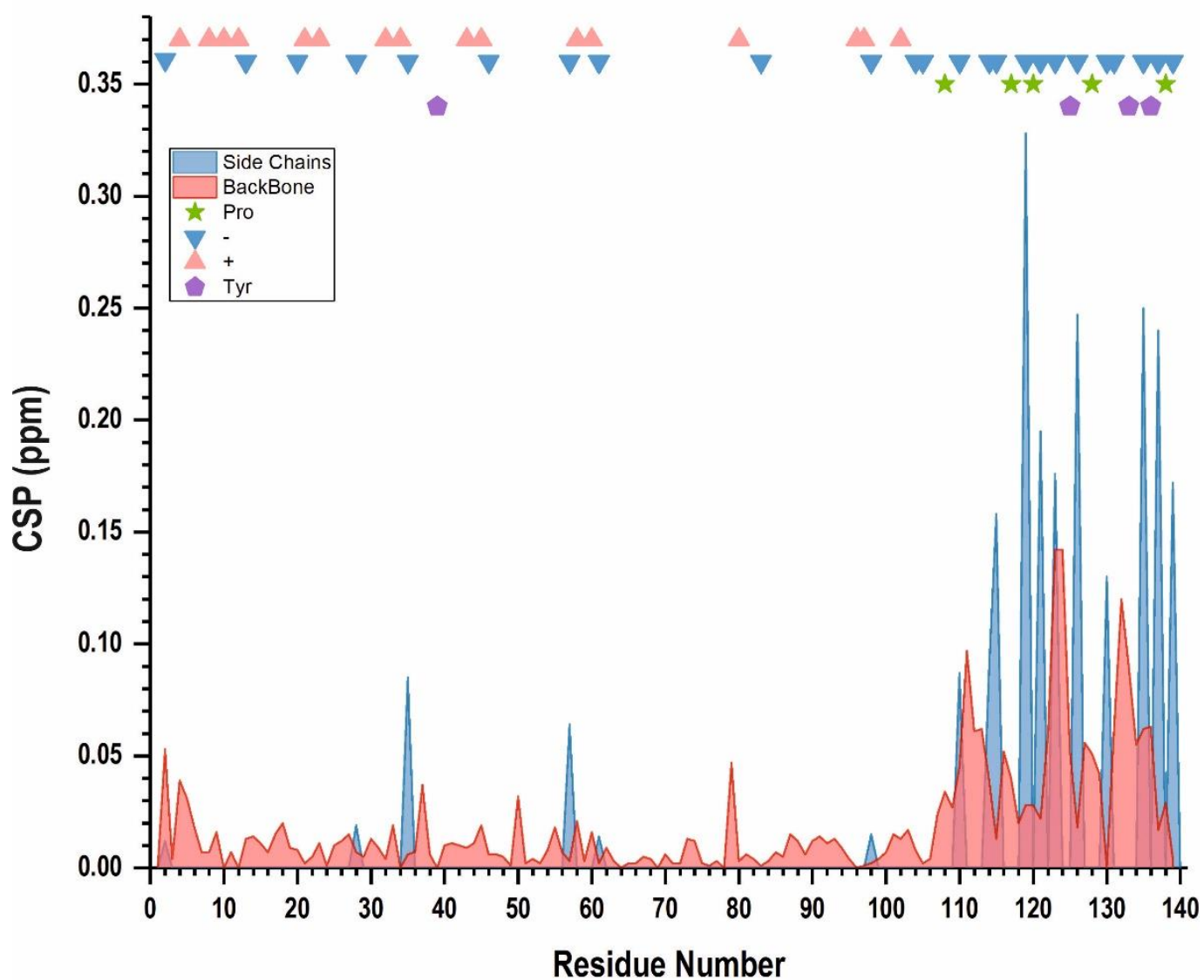
### (H)CBCACO



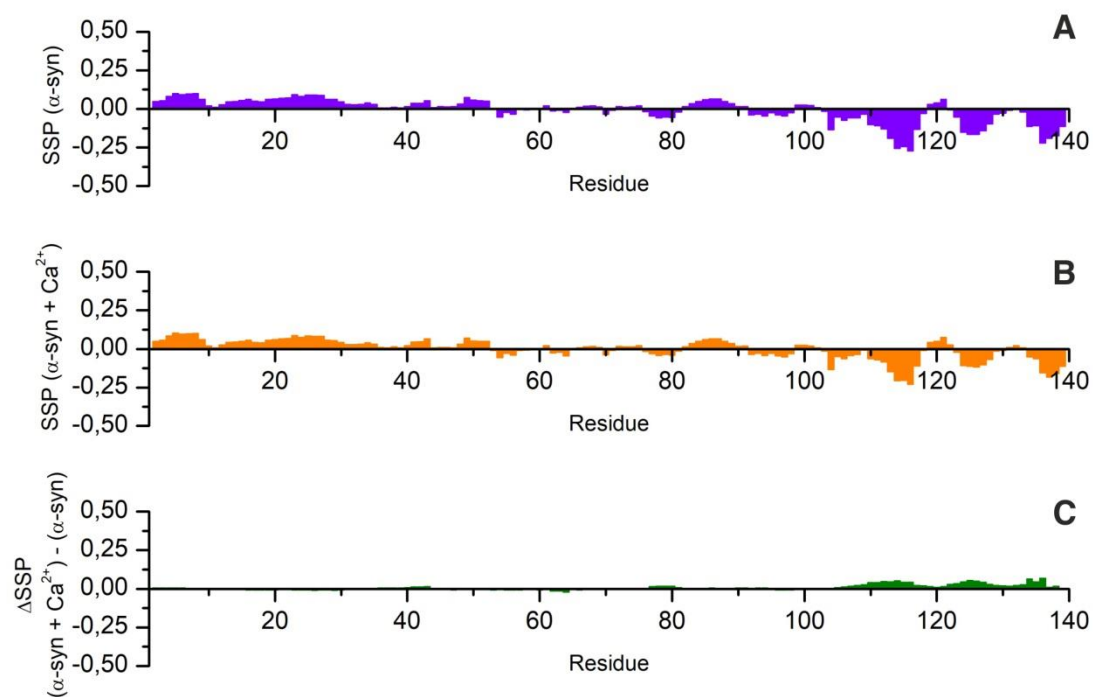
### (H)CCCO



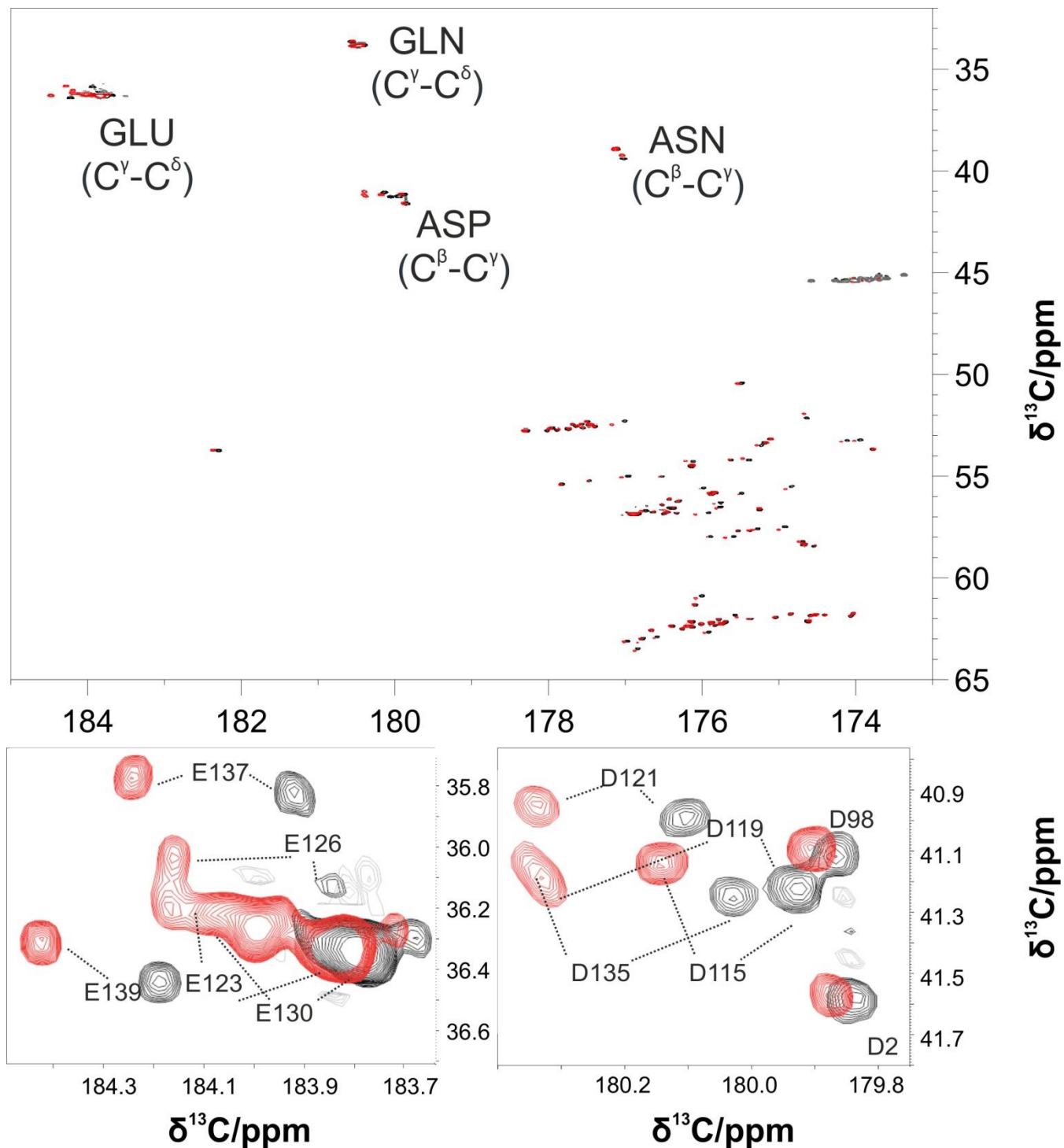
**Figure S3.** Comparison of chemical shift perturbations (CSP) of side chain carboxylate/carbonyl carbon chemical shifts (blue) with backbone carbonyl carbon chemical shifts (red) determined through from 2D-CACO and 2D-CON spectra ( $CSP = |\Delta(\delta^{13}C)|$ ). Backbone CSP values are smaller in 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.



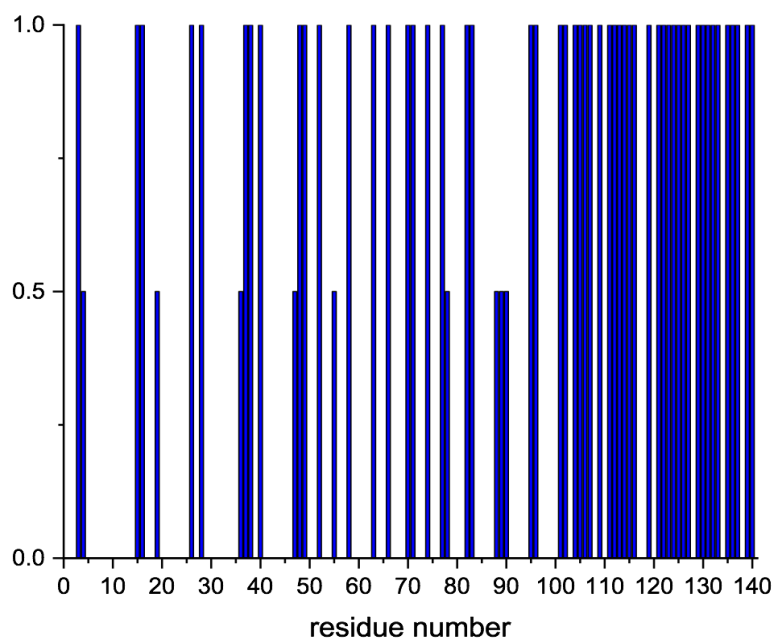
**Figure S4.** Secondary Structure Propensity (SSP) score obtained exploiting the Tamiola-Mulder approach<sup>[9]</sup> (<https://st-protein02.chem.au.dk/ncIDP/>) for  $\alpha$ -synuclein only (panel A) and  $\alpha$ -synuclein in the presence of calcium ions (panel B). Panel C reports the differences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



**Figure S5.** Superimposition of (H)CACO spectra recorded on a 50  $\mu\text{M}$   $\alpha$ -synuclein sample in absence (black) and presence (red) of  $\text{Ca}^{2+}$ . 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  $\text{Ca}^{2+}$ . The perturbation of Asp and Glu side chain resonances is analogous to the one observed with the more concentrated  $\alpha$ -synuclein sample. The titration follows the pattern from black ( $\alpha$ -syn: $\text{Ca}^{2+}$ , 1:0) to red ( $\alpha$ -syn: $\text{Ca}^{2+}$ , 1:256).

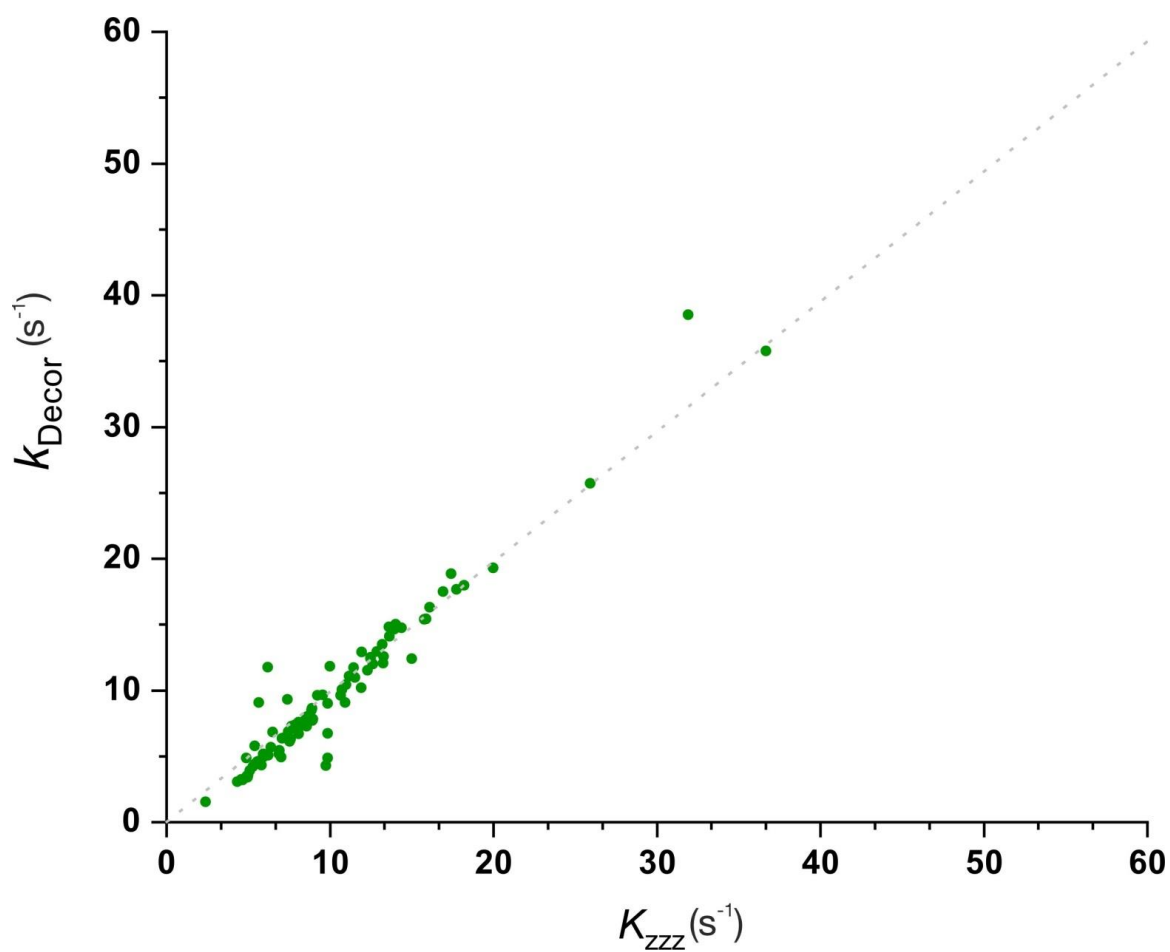


**Figure S6.** The figure reports in a schematic way the residues whose amide proton could be detected (0.5 indicates cross peaks close to the noise level). Backbone carbonyl carbon resonances could be detected for all peptide bonds.



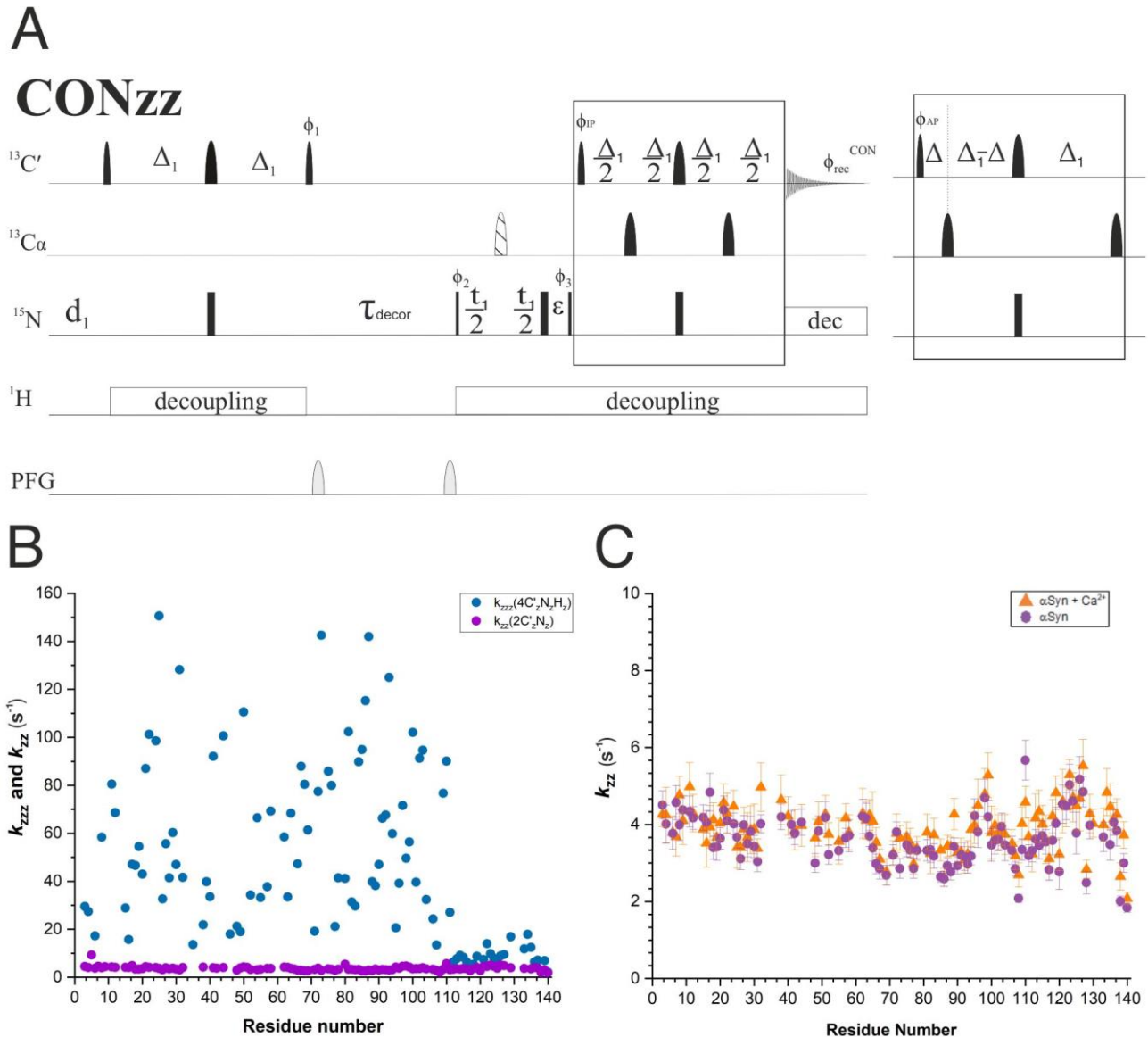


**Figure S7.** Comparison of DeCON and Decor performances. In both experiments the most pronounced effect leading to decorrelation of the longitudinal spin-order operators ( $4C'_{2}N_{x}H_{z}$  and  $2N_{x}H_{z}$  respectively) is the exchange of amide protons with the solvent. The agreement between 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  $\mu$ s). The striped pulse is an adiabatic Chirp pulse to invert  $^{13}\text{C}$  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  $\phi_2$ .

Intensities of cross-peaks determined in spectra acquired with the CONzz and DeCON pulse sequences as a function of  $\tau_{decor}$  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  $\alpha$ -synuclein before and after addition for  $\text{Ca}^{2+}$ .

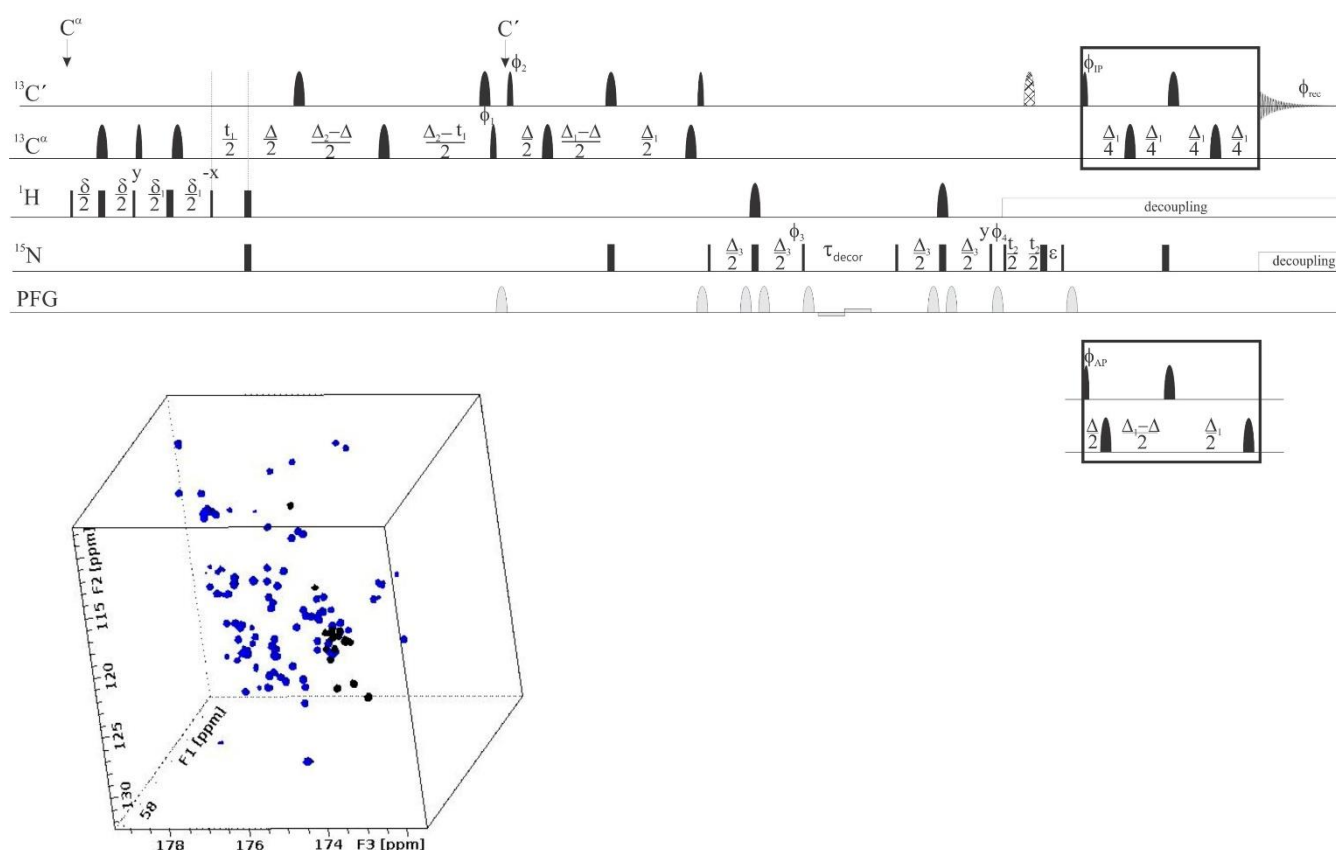


**Figure S9.** Pulse sequence used to acquire the 3D version of DeCON spectra. A screenshot of the acquired 3D spectrum is also shown. Narrow and wide black bars represent  $\pi/2$  and  $\pi$  non-selective pulses; narrow and wide rounded black bars represent  $\pi/2$  and  $\pi$  band-selective pulses. 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  $^{13}\text{C}$  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  $\phi_{\text{rec}} = x, -x, -x, x, -x, x, -x, -x, x, x, -x, -x, x, -x, -x, x$ . 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  $\mu\text{s}$ ). 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  $\phi_1$  for the  $^{13}\text{C}$  dimension and  $\phi_4$  for the  $^{15}\text{N}$  dimension.

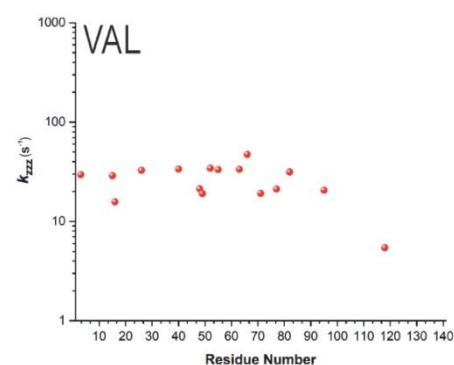
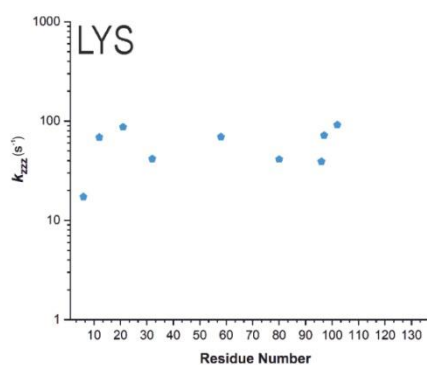
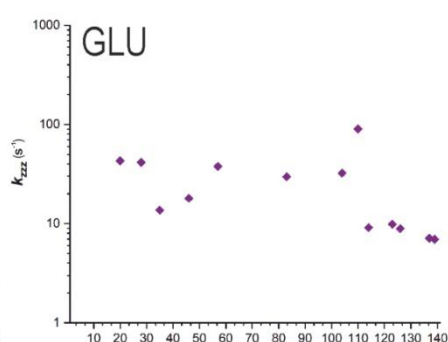
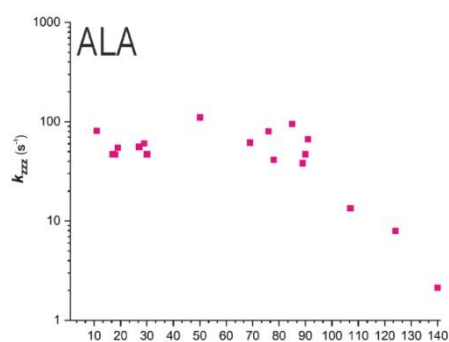
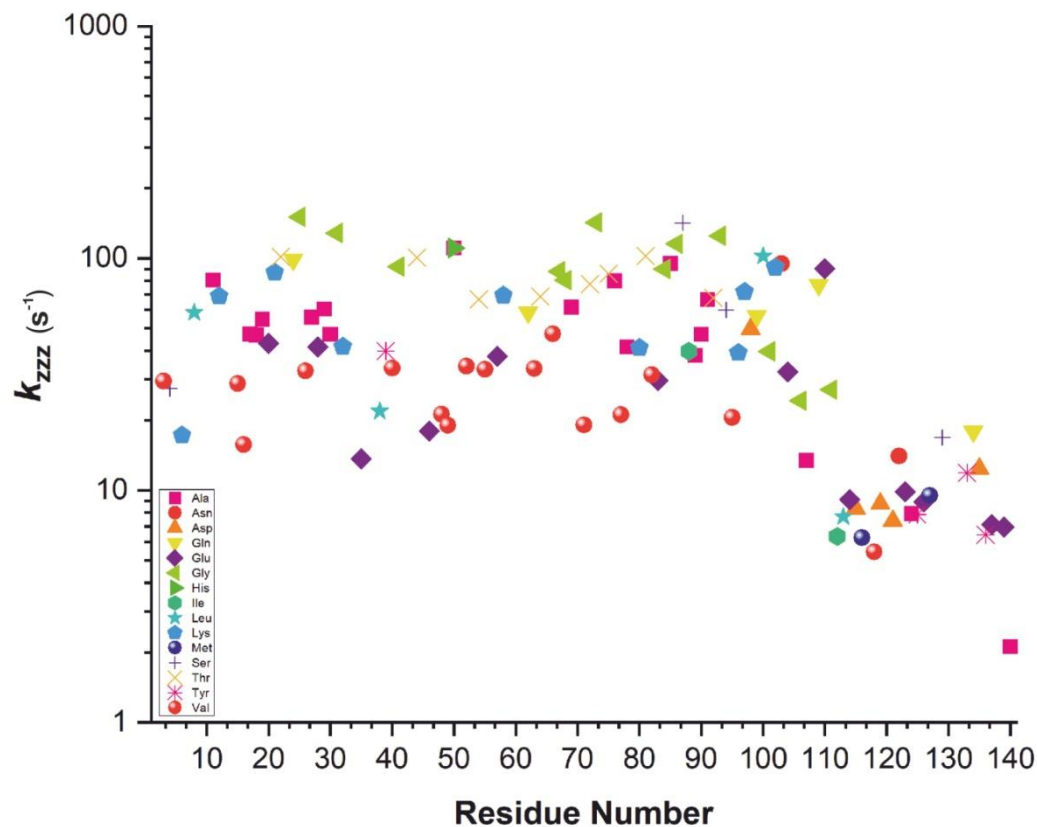
The 3D experiment was acquired on the 600  $\mu\text{M}$  sample of  $\alpha$ -synuclein using 8 scans per increment, a recovery delay of 0.9 s. 1024 points were acquired in the  $^{13}\text{C}'$  direct dimension using a sweep width of 5556 Hz (31.55 ppm), 96 points were used in both the indirect dimensions using a sweep width of 1785 Hz (25.17 ppm) for the  $^{15}\text{N}$  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.

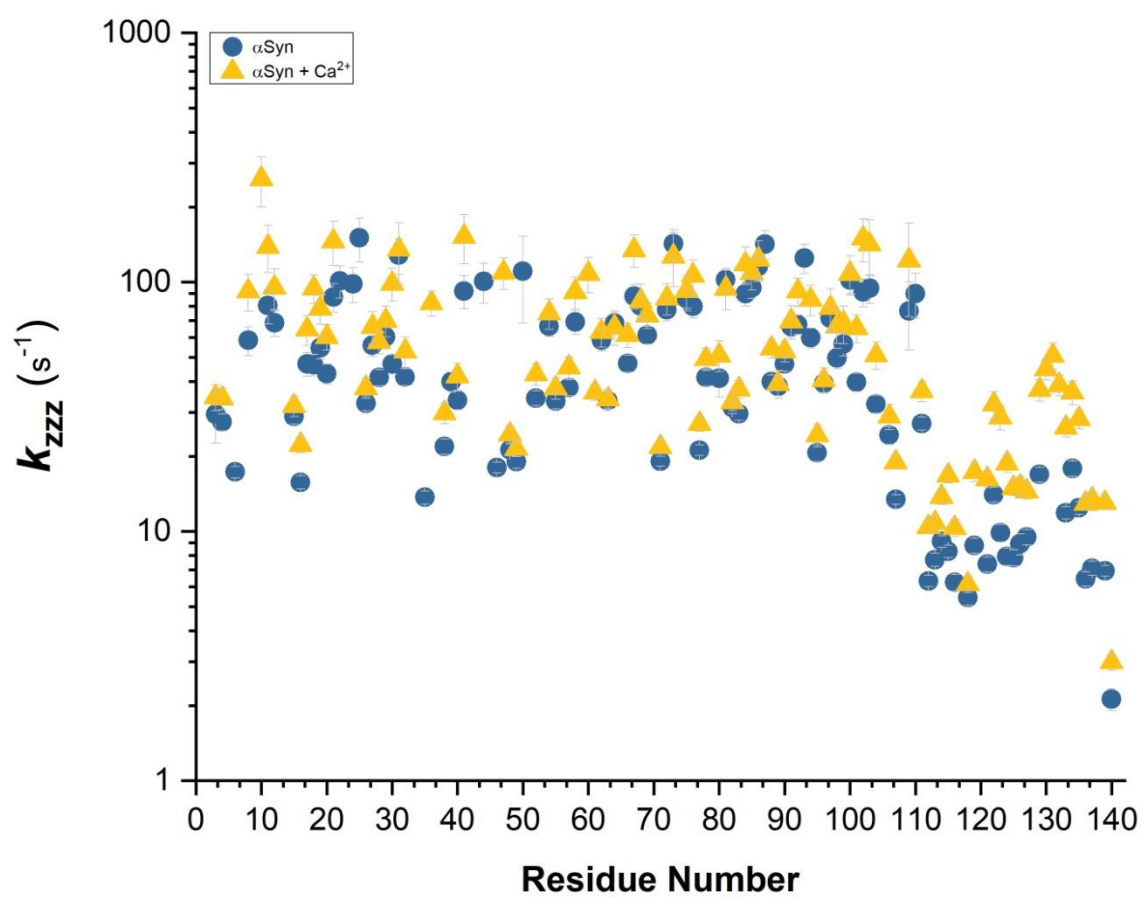
A series of  $^{13}\text{C}$ - $^{15}\text{N}$  planes of the 3D DeCON were acquired with different  $\tau_{\text{decor}}$  delays in order to monitor the decay of the signals. These experiments were performed with 8 scans per increment, a recovery delay of 0.9 s. 1024 points were acquired in the direct  $^{13}\text{C}'$  dimension using a sweep width of 5556 Hz (31.55 ppm) and 64 points in the  $^{15}\text{N}$  indirect dimension were used using a sweep width of 1785 Hz (25.17 ppm). The total duration of the experiment for the first  $\tau_{\text{decor}}$  used was 20 min. The used delays ( $\tau_{\text{decor}}$ ) were: 20  $\mu\text{s}$  - 600  $\mu\text{s}$  - 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  $\alpha$ -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  $\alpha$ -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 ( <sup>15</sup> N)	1024 ( <sup>13</sup> C)	2273 Hz (32.03 ppm)	5556 Hz (31.55 ppm)	2	2	2 hours 2 mins 6 s
2D-CACO	330 ( <sup>13</sup> C)	1024 ( <sup>13</sup> C)	5988 Hz (34.01 ppm)	5263 Hz (34.01 ppm)	2	2	48 mins 4 s
2D-CBCACO	476 ( <sup>13</sup> C)	1024 ( <sup>13</sup> C)	10417 Hz (59.16 ppm)	5263 Hz (34.01 ppm)	2	2	1 hour 9 mins 16 s
2D-CCCO	640 ( <sup>13</sup> C)	1024 ( <sup>13</sup> C)	10417 Hz (59.16 ppm)	5263 Hz (34.01 ppm)	2	2	1 hour 24 min 9 s
2D-DeCON	200 ( <sup>15</sup> N)	1024 ( <sup>13</sup> C)	1786 Hz (25.17 ppm)	5555 Hz (31.55 ppm)	4	4	2 hours 5 min 11 s
2D-CONzz	200 ( <sup>15</sup> N)	1024 ( <sup>13</sup> C)	1786 Hz (25.17 ppm)	5555 Hz (31.55 ppm)	4	4	2 hours 5 min 11 s
2D-Dècor	192 ( <sup>15</sup> N)	2048 ( <sup>1</sup> 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 ( <sup>15</sup> N)	1024 ( <sup>13</sup> C)	2273 Hz (32.03 ppm)	5556 Hz (31.55 ppm)	64	1.5	12 hours 11 mins 35 s
2D-(H)CACO	330 ( <sup>13</sup> C)	1024 ( <sup>13</sup> C)	5988 Hz (34.01 ppm)	5263 Hz (34.01 ppm)	16	0.9	3 hours 9 mins 4 s
2D- (H)CBCACO	476 ( <sup>13</sup> C)	1024 ( <sup>13</sup> C)	10417 Hz (59.16 ppm)	5263 Hz (34.01 ppm)	16	0.9	4 hours 33 mins 37 s

**Table S2.**  $^{13}\text{C}$  chemical shifts of Asp, Asn, Glu and Gln residues of  $\alpha$ -synuclein in 20 mM TRIS (tris-hydroxymethyl-aminomethane) buffer, 310 K, pH 7.4.

Type	Number	C'	C $\alpha$	C $\beta$	C $\gamma$	C $\delta$
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.



## **Article 2.1:**

**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**



# 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<sup>1,2</sup> · Letizia Pontoriero<sup>1,2</sup> · Vladimir N. Uversky<sup>3</sup> · Isabella C. Felli<sup>1,2</sup> · Roberta Pierattelli<sup>1,2</sup>

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 · <sup>13</sup>C 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  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$ , with the class  $\beta$ -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.

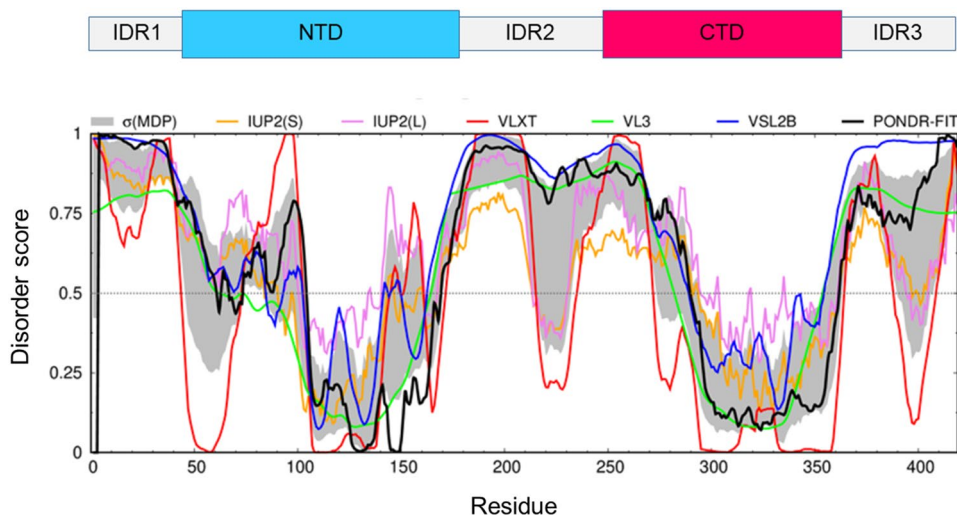
✉ Isabella C. Felli  
felli@cerm.unifi.it

✉ Roberta Pierattelli  
roberta.pierattelli@unifi.it

<sup>1</sup> Magnetic Resonance Center – CERM, University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, FI, Italy

<sup>2</sup> Department of Chemistry “Ugo Schiff”, University of Florence, Via della Lastruccia 3-13, 50019 Sesto Fiorentino, FI, Italy

<sup>3</sup> 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), PONDR® VLXT (red line), PONDR® VL3 (green line), PONDR® VSL2B (blue line), PONDR® FIT (black line). The gray shadow region signifies the error distribution  $\sigma$ (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](http://www.covid19-nmr.de)) enabling follow-up

calculated by averaging of the disorder profiles of individual predictors. Protein regions with a disordered score consistently larger than 0.5 are considered disordered, whereas 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

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  $^{13}\text{C}$ ,  $^{15}\text{N}$ -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  $\text{OD}_{600}$  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  $\mu$ 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 a 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  $\mu$ M IDR1-NTD-IDR2, 25 mM TRIS pH 6.5, 450 mM sodium chloride, 0.02% NaN<sub>3</sub>, 5% (v/v) D<sub>2</sub>O 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 <sup>1</sup>H, 176.05 MHz <sup>13</sup>C, and 70.97 MHz <sup>15</sup>N frequencies, equipped with a 5 mm cryogenically cooled probehead optimized for <sup>13</sup>C direct detection (TXO). Proton direct detected NMR experiments were acquired on a 28.3 T Bruker AVANCE NEO spectrometer operating at 1200.85 MHz <sup>1</sup>H, 301.97 MHz <sup>13</sup>C, and 121.70 MHz <sup>15</sup>N equipped with a 3 mm cryogenically cooled triple-resonance probehead (TCI).

Backbone assignment was performed by analyzing 2D and 3D <sup>1</sup>H and <sup>13</sup>C direct detected experiments. In particular, 2D-[<sup>1</sup>H, <sup>15</sup>N]-HSQC, 2D-[<sup>1</sup>H, <sup>15</sup>N]-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-<sup>13</sup>C 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 <sup>13</sup>C 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 <sup>15</sup>N 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 <sup>13</sup>C detected experiments and are summarized hereafter. The <sup>1</sup>H carrier was placed at 4.7 ppm for non-selective hard pulses. <sup>13</sup>C pulses were given at 176.7 ppm, 55.9 ppm, and 45.7 ppm for C', C $^{\alpha}$  and C<sup>ali</sup> regions, respectively.

**Table 1** Experimental parameters used to collect the NMR experiments

Experiments	Dimension of acquired data			Spectral width (ppm)			NS <sup>a</sup>	d1 + aq (s) <sup>b</sup>	Spectrometer frequency ( <sup>1</sup> H) (MHz)
	t1	t2	t3	F1	F2	F3			
<sup>1</sup> H detected									
<sup>1</sup> H- <sup>15</sup> N BEST-TROSY	512 ( <sup>15</sup> N)	9676 ( <sup>1</sup> H)		41	15		16	0.5	1200
BT-HNCACB	96 ( <sup>13</sup> C)	90 ( <sup>15</sup> N)	6144 ( <sup>1</sup> H)	75	41	14	96	0.2	1200
BT-HN(CO)CACB	96 ( <sup>13</sup> C)	80 ( <sup>15</sup> N)	6144 ( <sup>1</sup> H)	75	41	14	96	0.2	1200
HN(CA)CO	128 ( <sup>13</sup> C)	128 ( <sup>15</sup> N)	4096 ( <sup>1</sup> H)	7	28	18	8	1.0	1200
HNCO	128 ( <sup>13</sup> C)	220 ( <sup>15</sup> N)	4096 ( <sup>1</sup> H)	7	28	18	4	1.0	1200
<sup>13</sup> C detected									
CON	512 ( <sup>15</sup> N)	1024 ( <sup>13</sup> C)		34	31		32	1.6	700
(HCA)CON	220 ( <sup>15</sup> N)	1024 ( <sup>13</sup> C)		40	31		16	0.9	700
(H)CACO	330 ( <sup>13</sup> C)	1024 ( <sup>13</sup> C)		34	30		32	1.0	700
(H)CBCACO	476 ( <sup>13</sup> C)	1024 ( <sup>13</sup> C)		59	30		32	1.0	700
(H)CBCACON	128 ( <sup>13</sup> C)	96 ( <sup>15</sup> N)	1024 ( <sup>13</sup> C)	58	34	30	4	1.0	700
(H)CBCANCO	96 ( <sup>13</sup> C)	96 ( <sup>15</sup> N)	1024 ( <sup>13</sup> C)	58	34	30	16	1.0	700

<sup>a</sup>Number of acquired scans

<sup>b</sup>Relaxation delay (acquisition time plus recovery delay d1)

$^{15}\text{N}$  pulses were given at 124.0 ppm. Q5 and Q3 shapes (Emsley and Bodenhausen 1990) of durations of 300 and 231  $\mu\text{s}$ , respectively, were used for  $^{13}\text{C}$  band-selective  $\pi/2$  and  $\pi$  flip angle pulses except for the  $\pi$  pulses that should be band-selective on the  $\text{C}^\alpha$  region (Q3, 900  $\mu\text{s}$ ), and for the adiabatic  $\pi$  pulse (Böhlen and Bodenhausen 1993) to invert both  $\text{C}'$  and  $\text{C}^\alpha$  (smoothed Chirp 500  $\mu\text{s}$ , 20% smoothing, 80 kHz sweep width, 11.3 kHz RF field strength). Composite pulse decoupling was applied on  $^1\text{H}$  (Waltz-16) (Shaka et al. 1983) and  $^{15}\text{N}$  (Garp-4) (Shaka et al. 1985) with an RF field strength of 3 kHz and 1 kHz respectively.

$^1\text{H}$  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  $^{15}\text{N}$  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).  $\text{C}'$  and  $\text{C}^\alpha/\text{C}^\beta$  selective excitation was exploited through band selective pulses.

Carrier frequencies used for triple resonance experiments in  $^1\text{H}$  detected experiments were the same as for  $^{13}\text{C}$  detected experiments except for the  $^{15}\text{N}$  carrier placed at 118.0 ppm. Pulse shapes and lengths for  $^{13}\text{C}$  band-selective pulses were G4 (Emsley and Bodenhausen 1992) and Q3 (Emsley and Bodenhausen 1990) shapes of durations of 205 and 128  $\mu\text{s}$ , respectively, used for  $^{13}\text{C}$  band-selective  $\pi/2$  and  $\pi$  flip angle pulses except for the  $\pi$  pulses that should be band-selective on the  $\text{C}^\alpha$  region (Q3, 525  $\mu\text{s}$ ). The  $^1\text{H}$  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  $\pi$  pulses.

All the spectra were acquired, processed, and analysed by using Bruker TopSpin 4.0.8 software. Chemical shifts were referenced using the  $^1\text{H}$  and  $^{13}\text{C}$  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).

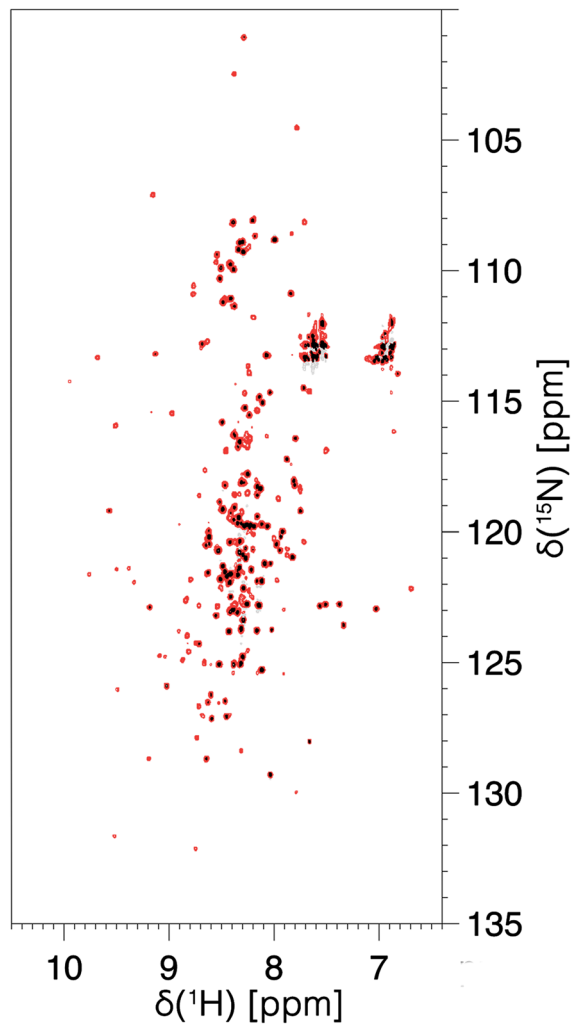
## Assignments and data deposition

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  $\text{C}'$  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  $^{13}\text{C}$ -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  $^1\text{H}$ , 301.97 MHz  $^{13}\text{C}$ , and 121.70 MHz  $^{15}\text{N}$  equipped with a 3 mm cryogenically cooled triple-resonance probehead (TCI)

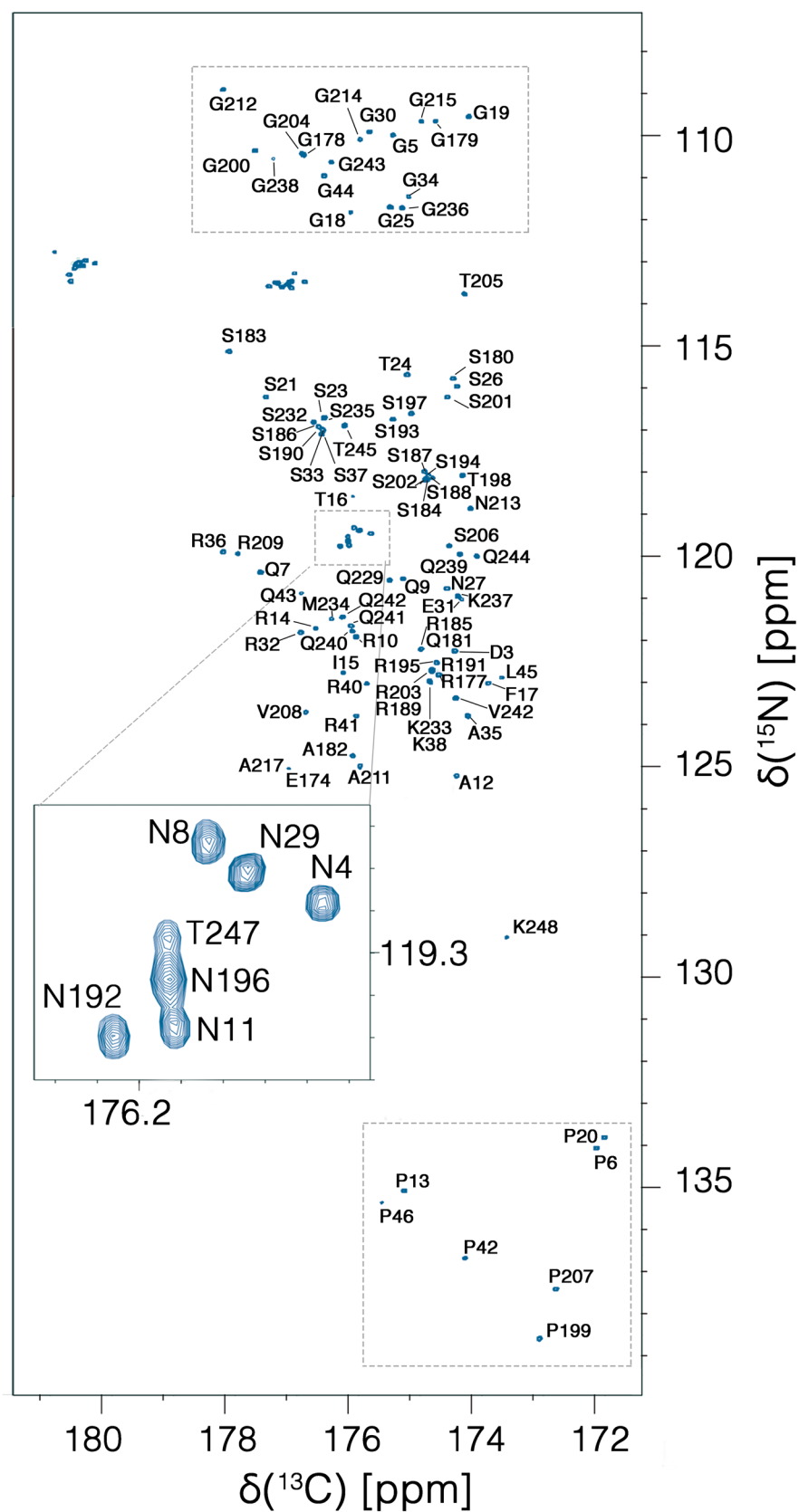
It is worth noting that proline resonances provide a useful starting point for sequence-specific assignment. The particular  $^{15}\text{N}$  chemical shift range expected for proline nitrogen signals ( $\text{N}_i$ ) and the fact that this is correlated to resonances of the preceding amino acid ( $\text{C}'_{i-1}$ ,  $\text{C}^\alpha_{i-1}$ ,  $\text{C}^\beta_{i-1}$ ) through the 2D CON and 3D (H)CBCACON spectra constitute two features that allow us to unambiguously identify the type of dipeptide ( $\text{X}_{i-1}$ -Pro $_i$  pair) that gives rise to specific signals as highlighted in Fig. 4. Indeed, the characteristic chemical shifts of  $\text{C}^\alpha$  and  $\text{C}^\beta$  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

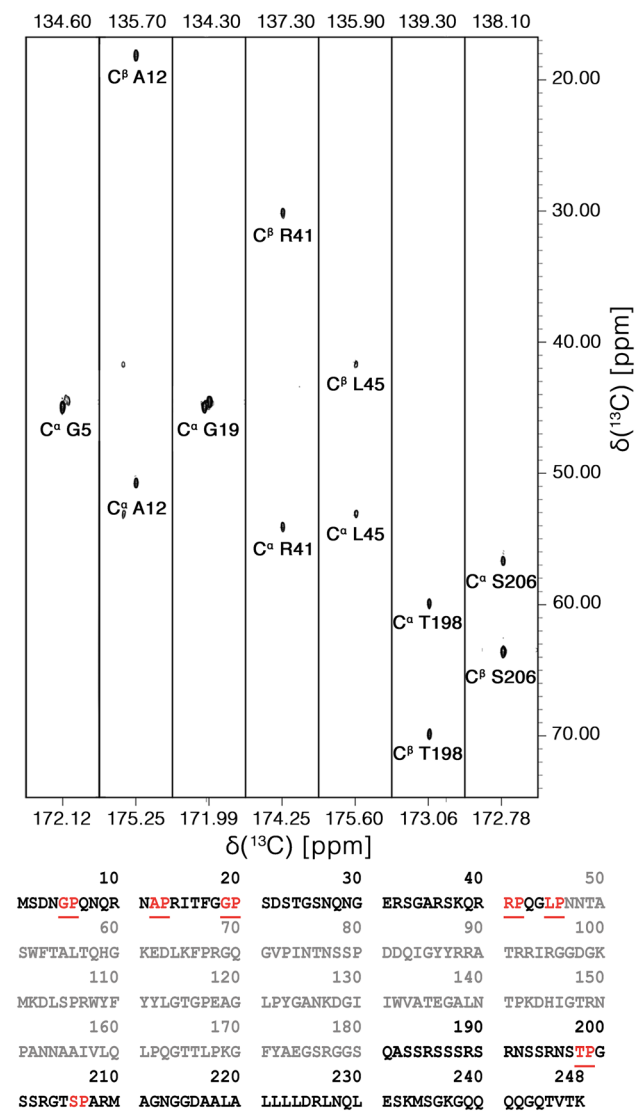
primary sequence of the protein. Therefore, already at this very early stage of the sequence-specific assignment process, most of the observed resonances in this region could be assigned to specific amino acids uniquely considering the type of X-Pro pairs present in the intrinsically disordered regions (all resonances could be unambiguously assigned except for the two Gly-Pro pairs). Similarly, inspecting the opposite region of the CON spectrum at low  $^{15}\text{N}$  chemical shifts (Fig. 3, top squared region) allows us to identify correlations involving  $^{15}\text{N}$  nuclear spins of glycine residues; correlation to the carbonyl carbon of the previous amino acid ( $\text{C}'_{i-1}$ - $\text{N}_i$ ) 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  $\text{X}_{i-1}$ -Gly $_i$  pairs achieved through inspection of the (H)CBCACON provides further input for their identification, as described above for the case of  $\text{X}_{i-1}$ -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  $^{15}\text{N}$  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 ( $\text{C}'$ ,  $\text{C}^\alpha$ ,  $\text{C}^\beta$ ,  $\text{N}$ ,  $\text{H}^{\text{N}}$ ) (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  $^1\text{H}$ , 176.05 MHz  $^{13}\text{C}$ , and 70.97 MHz  $^{15}\text{N}$  frequencies, equipped with a 5 mm cryogenically cooled probehead optimized for  $^{13}\text{C}$  direct detection (TXO)





**Fig. 4** Seven strips derived from the 3D-(H)CBCACON experiment extracted at the  $^{15}\text{N}$  chemical shift of proline residues. The  $\text{C}'$ ,  $\text{C}^\alpha$  and  $\text{C}^\beta$  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  $^1\text{H}$ , 176.05 MHz  $^{13}\text{C}$ , and 70.97 MHz  $^{15}\text{N}$  frequencies, equipped with a 5 mm cryogenically cooled probehead optimized for  $^{13}\text{C}$  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.

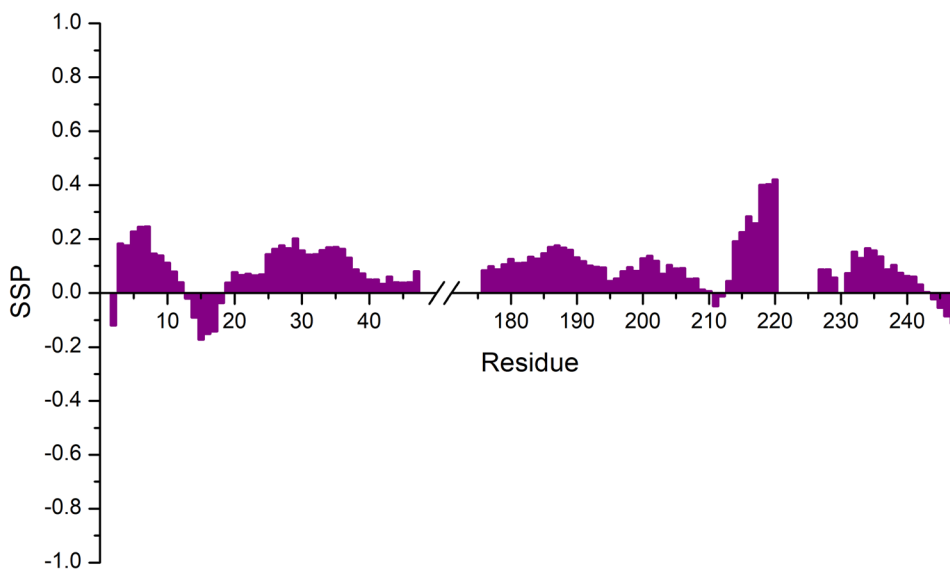
The first two disordered regions of the N protein from SARS-CoV-2 (IDR1 and IDR2) can now be investigated at atomic resolution providing experimental information regarding the many interaction sites that can be predicted through different approaches (Kumar et al. 2008; Giri et al. 2020). The resonances 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  $^{13}\text{C}$  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  $H^N$ ,  $N$ ,  $C'$ ,  $C^\alpha$ , and  $C^\beta$  nuclei were used. The two regions result to be highly disordered with a slight tendency to be in an  $\alpha$ -helix conformation for the residues 216–220



**Acknowledgements** The support and the use of resources of the CERM/CIRMMP 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.bmr.bwisc.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, Csizmok 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, Godhave 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 H<sub>2</sub>O 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>
- Kupce 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, Wrigth 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, Sonh 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  $\alpha$ -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>

**Article 2.2:**

**Large-scale recombinant production of the SARS-CoV-2 proteome for high-throughput and structural biology applications**

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  
(CINVESTAV), Mexico

\*Correspondence:

Anja Böckmann  
a.boeckmann@ibcp.fr  
Harald Schwalbe  
schwalbe@nmr.uni-frankfurt.de  
Martin Hengesbach  
hengesch@nmr.uni-frankfurt.de  
Andreas Schlundt  
schlundt@bio.uni-frankfurt.de

<sup>†</sup>These authors have contributed equally to this work  
and share first authorship

<sup>‡</sup>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:

Altincelik N, Korn 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, Bütikofer M, Camacho-Zarco AP, Carlomagno T, Caruso IP, Ceylan B, Chaikuad A, Chu F, Cole L, Crosby MG, de Jesus V, Dhamotharan K, Felli IC, Ferner J, Fleischmann Y, Fogeron M-L, Fourkotis 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, Henzler-Wildman K, Hoch JC, Hohmann KF, Hutchison MT, Jaudzems K, Jović K, Kaderli J, Kalniņš G, Kaņepe I, Kirchoerfer RN, Kirkpatrick J, Knapp S, Krishnathas R, Kutz F, zur Lage S, Lambert R, Lang A, Laurents D, Lecoq L, Linhard V, Löhner 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, Riesenra CM, Robertson A, Pinheiro AS, Sabbatella R, Salvi N, Saxena K, Schulte L, Schiavina M, Schwalbe H, Silber M, Almeida MdS, Sprague-Piercy MA, Spyroulas 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, Zehnder 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.* 8:653148. doi: 10.3389/fmolb.2021.653148

# Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications

Nadide Altincelik<sup>1,2†</sup>, Sophie Marianne Korn<sup>2,3†</sup>, Nusrat Shahin Qureshi<sup>1,2†</sup>, Marie Dujardin<sup>4†</sup>, Marti Ninot-Pedrosa<sup>4†</sup>, Rupert Abele<sup>5</sup>, Marie Jose Abi Saad<sup>6</sup>, Caterina Alfano<sup>7</sup>, Fabio C. L. Almeida<sup>8,9</sup>, Islam Alshamleh<sup>1,2</sup>, Gisele Cardoso de Amorim<sup>8,10</sup>, Thomas K. Anderson<sup>11</sup>, Cristiane D. Anobom<sup>8,12</sup>, Chelsea Anorma<sup>13</sup>, Jasleen Kaur Bains<sup>1,2</sup>, Adriaan Bax<sup>14</sup>, Martin Blackledge<sup>15</sup>, Julius Blechar<sup>1,2</sup>, Anja Böckmann<sup>4,\*†</sup>, Louis Brigandat<sup>4</sup>, Anna Bula<sup>16</sup>, Matthias Bütikofer<sup>6</sup>, Aldo R. Camacho-Zarco<sup>15</sup>, Teresa Carlomagno<sup>17,18</sup>, Icaro Putinhon Caruso<sup>8,9,19</sup>, Betül Ceylan<sup>1,2</sup>, Apirat Chaikuad<sup>20,21</sup>, Feixia Chu<sup>22</sup>, Laura Cole<sup>4</sup>, Marquise G. Crosby<sup>23</sup>, Vanessa de Jesus<sup>1,2</sup>, Karthikeyan Dhamotharan<sup>2,3</sup>, Isabella C. Felli<sup>24,25</sup>, Jan Ferner<sup>1,2</sup>, Yanick Fleischmann<sup>6</sup>, Marie-Laure Fogeron<sup>4</sup>, Nikolaos K. Fourkotis<sup>26</sup>, Christin Fuks<sup>1</sup>, Boris Fürtig<sup>1,2</sup>, Angelo Gallo<sup>26</sup>, Santosh L. Gande<sup>1,2</sup>, Juan Atilio Gerez<sup>6</sup>, Dhiman Ghosh<sup>6</sup>, Francisco Gomes-Neto<sup>8,27</sup>, Oksana Gorbatyuk<sup>28</sup>, Serafima Guseva<sup>15</sup>, Carolin Hacker<sup>29</sup>, Sabine Häfner<sup>30</sup>, Bing Hao<sup>28</sup>, Bruno Hargittay<sup>1,2</sup>, K. Henzler-Wildman<sup>11</sup>, Jeffrey C. Hoch<sup>28</sup>, Katharina F. Hohmann<sup>1,2</sup>, Marie T. Hutchison<sup>1,2</sup>, Kristaps Jaudzems<sup>16</sup>, Katarina Jović<sup>22</sup>, Janina Kaderli<sup>6</sup>, Gints Kalniņš<sup>31</sup>, Iveta Kaņepe<sup>16</sup>, Robert N. Kirchoerfer<sup>11</sup>, John Kirkpatrick<sup>17,18</sup>, Stefan Knapp<sup>20,21</sup>, Robin Krishnathas<sup>1,2</sup>, Felicitas Kutz<sup>1,2</sup>, Susanne zur Lage<sup>18</sup>, Roderick Lambert<sup>3</sup>, Andras Lang<sup>30</sup>, Douglas Laurents<sup>32</sup>, Lauriane Lecoq<sup>4</sup>, Verena Linhard<sup>1,2</sup>, Frank Löhner<sup>2,33</sup>, Anas Malki<sup>15</sup>, Luiza Mamigonian Bessa<sup>15</sup>, Rachel W. Martin<sup>13,23</sup>, Tobias Matzel<sup>1,2</sup>, Damien Maurin<sup>15</sup>, Seth W. McNutt<sup>22</sup>, Nathane Cunha Mebus-Antunes<sup>8,9</sup>, Beat H. Meier<sup>6</sup>, Nathalie Meiser<sup>1</sup>, Miguel Mompeán<sup>32</sup>, Elisa Monaca<sup>7</sup>, Roland Montserret<sup>4</sup>, Laura Mariño Perez<sup>15</sup>, Celine Moser<sup>34</sup>, Claudia Muhle-Goll<sup>34</sup>, Thais Cristina Neves-Martins<sup>8,9</sup>, Xiamonin Ni<sup>20,21</sup>, Brenna Norton-Baker<sup>13</sup>, Roberta Pierattelli<sup>24,25</sup>, Letizia Pontoriero<sup>24,25</sup>, Yulia Pustovalova<sup>28</sup>, Oliver Ohlenschläger<sup>30</sup>, Julien Orts<sup>6</sup>, Andrea T. Da Poian<sup>9</sup>, Dennis J. Pyper<sup>1,2</sup>, Christian Richter<sup>1,2</sup>, Roland Riek<sup>6</sup>, Chad M. Riesenra<sup>35</sup>, Angus Robertson<sup>14</sup>, Anderson S. Pinheiro<sup>8,12</sup>, Raffaele Sabbatella<sup>7</sup>, Nicola Salvi<sup>15</sup>, Krishna Saxena<sup>1,2</sup>, Linda Schulte<sup>1,2</sup>, Marco Schiavina<sup>24,25</sup>, Harald Schwalbe<sup>1,2,\*†</sup>, Mara Silber<sup>34</sup>, Marcus da Silva Almeida<sup>8,9</sup>, Marc A. Sprague-Piercy<sup>23</sup>, Georgios A. Spyroulias<sup>26</sup>, Sridhar Sreeramulu<sup>1,2</sup>, Jan-Niklas Tants<sup>2,3</sup>, Kaspars Tars<sup>31</sup>, Felix Torres<sup>6</sup>, Sabrina Töws<sup>3</sup>, Miguel Á. Treviño<sup>32</sup>, Sven Trucks<sup>1</sup>, Aikaterini C. Tsika<sup>26</sup>, Krisztina Varga<sup>22</sup>, Ying Wang<sup>17</sup>, Marco E. Weber<sup>6</sup>, Julia E. Weigand<sup>36</sup>, Christoph Wiedemann<sup>37</sup>, Julia Wimer-Bartoschek<sup>1,2</sup>, Maria Alexandra Wirtz Martin<sup>1,2</sup>, Johannes Zehnder<sup>6</sup>, Martin Hengesbach<sup>1,\*†</sup> and Andreas Schlundt<sup>2,3,\*†</sup>

<sup>1</sup>Institute for Organic Chemistry and Chemical Biology, Goethe University Frankfurt, Frankfurt am Main, Germany, <sup>2</sup>Center of Biomolecular Magnetic Resonance (BMRZ), Goethe University Frankfurt, Frankfurt am Main, Germany, <sup>3</sup>Institute for Molecular Biosciences, Goethe University Frankfurt, Frankfurt am Main, Germany, <sup>4</sup>Molecular Microbiology and Structural Biochemistry, UMR 5086, CNRS/Lyon University, Lyon, France, <sup>5</sup>Institute for Biochemistry, Goethe University Frankfurt, Frankfurt am Main, Germany, <sup>6</sup>Swiss Federal Institute of Technology, Laboratory of Physical Chemistry, ETH Zurich, Zurich, Switzerland, <sup>7</sup>Structural Biology and Biophysics Unit, Fondazione Ri.MED, Palermo, Italy, <sup>8</sup>National Center of Nuclear Magnetic Resonance (CNRMN, CENABIO), Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, <sup>9</sup>Institute of Medical Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, <sup>10</sup>Multidisciplinary Center for Research in Biology (NUMPEX), Campus Duque de Caxias Federal University of Rio de Janeiro, Duque de Caxias, Brazil, <sup>11</sup>Institute for Molecular Virology, University of Wisconsin-Madison, Madison, WI, United States, <sup>12</sup>Institute of Chemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, <sup>13</sup>Department of Chemistry, University of California, Irvine, CA, United States, <sup>14</sup>LCP, NIDDK, NIH, Bethesda, MD, United States, <sup>15</sup>Univ. Grenoble Alpes, CNRS, CEA, IBS, Grenoble, France, <sup>16</sup>Latvian Institute of Organic Synthesis, Riga, Latvia, <sup>17</sup>BMWZ and Institute of Organic

Chemistry, Leibniz University Hannover, Hannover, Germany, <sup>18</sup>Group of NMR-Based Structural Chemistry, Helmholtz Centre for Infection Research, Braunschweig, Germany, <sup>19</sup>Multiuser Center for Biomolecular Innovation (CMIB), Department of Physics, São Paulo State University (UNESP), São José do Rio Preto, Brazil, <sup>20</sup>Institute of Pharmaceutical Chemistry, Goethe University Frankfurt, Frankfurt am Main, Germany, <sup>21</sup>Structural Genomics Consortium, Buchmann Institute for Molecular Life Sciences, Frankfurt am Main, Germany, <sup>22</sup>Department of Molecular, Cellular, and Biomedical Sciences, University of New Hampshire, Durham, NH, United States, <sup>23</sup>Department of Molecular Biology and Biochemistry, University of California, Irvine, CA, United States, <sup>24</sup>Magnetic Resonance Centre (CERM), University of Florence, Sesto Fiorentino, Italy, <sup>25</sup>Department of Chemistry "Ugo Schiff", University of Florence, Sesto Fiorentino, Italy, <sup>26</sup>Department of Pharmacy, University of Patras, Patras, Greece, <sup>27</sup>Laboratory of Toxinology, Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, <sup>28</sup>Department of Molecular Biology and Biophysics, UConn Health, Farmington, CT, United States, <sup>29</sup>Signals GmbH & Co. KG, Frankfurt am Main, Germany, <sup>30</sup>Leibniz Institute on Aging—Fritz Lipmann Institute (FLI), Jena, Germany, <sup>31</sup>Latvian Biomedical Research and Study Centre, Riga, Latvia, <sup>32</sup>"Rocasolano" Institute for Physical Chemistry (IQFR), Spanish National Research Council (CSIC), Madrid, Spain, <sup>33</sup>Institute of Biophysical Chemistry, Goethe University Frankfurt, Frankfurt am Main, Germany, <sup>34</sup>IBG-4, Karlsruhe Institute of Technology, Karlsruhe, Germany, <sup>35</sup>Department of Biochemistry and National Magnetic Resonance Facility at Madison, University of Wisconsin-Madison, Madison, WI, United States, <sup>36</sup>Department of Biology, Technical University of Darmstadt, Darmstadt, Germany, <sup>37</sup>Institute 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](https://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  $\alpha$ - $\delta$  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) <sup>a</sup>	Trivial name construct expressed	Size (aa)	Boundaries	MW (kDa)	Homol. SCoV (%) <sup>b</sup>	Template PDB <sup>c</sup>	SCoV2 PDB <sup>d</sup>
<b>nsp1</b> 266–805	<b>Leader</b>	<b>180</b>		<b>19.8</b>	<b>84</b>		
	Full-length	180	1–180	19.8	83		
	Globular domain (GD)	116	13–127	12.7	85	2GDT	7K7P
<b>nsp2</b> 806–2,719		<b>638</b>		<b>70.5</b>	<b>68</b>		
	C-terminal IDR (CtDR)	45	557–601	4.9	55		
<b>nsp3</b> 2,720–8,554		<b>1,945</b>		<b>217.3</b>	<b>76</b>		
a	Ub-like (Ubl) domain	111	1–111	12.4	79	2IDY	7KAG
a	Ub-like (Ubl) domain + IDR	206	1–206	23.2	58		
b	Macrodomain	170	207–376	18.3	74	6VXS	6VXS
c	SUD-N	140	409–548	15.5	69	2W2G	
c	SUD-NM	267	409–675	29.6	74	2W2G	
c	SUD-M	125	551–675	14.2	82	2W2G	
c	SUD-MC	195	551–743	21.9	79	2KQV	
c	SUD-C	64	680–743	7.4	73	2KAF	
d	Papain-like protease PL <sup>PRO</sup>	318	743–1,060	36	83	6W9C	6W9C
e	NAB	116	1,088–1,203	13.4	87	2K87	
Y	CoV-Y	308	1,638–1,945	34	89		
<b>nsp5</b> 10,055–10,972	<b>Main protease (M<sup>PRO</sup>)</b>	<b>306</b>		<b>33.7</b>	<b>96</b>		
	Full-length <sup>e</sup>	306	1–306	33.7	96	6Y84	6Y84
<b>nsp7</b> 11,843–12,091		<b>83</b>		<b>9.2</b>	<b>99</b>		
	Full-length	83	1–83	9.2	99	6WIQ	6WIQ
<b>nsp8</b> 12,092–12,685		<b>198</b>		<b>21.9</b>	<b>98</b>		
	Full-length	198	1–198	21.9	97	6WIQ	6WIQ
<b>nsp9</b> 12,686–13,024		<b>113</b>		<b>12.4</b>	<b>97</b>		
	Full-length	113	1–113	12.4	97	6W4B	6W4B
<b>nsp10</b> 13,025–13,441		<b>139</b>		<b>14.8</b>	<b>97</b>		
	Full-length	139	1–139	14.8	97	6W4H	6W4H
<b>nsp13</b> 16,237–18,039	<b>Helicase</b>	<b>601</b>		<b>66.9</b>	<b>100</b>		
	Full-length	601	1–601	66.9	100	6ZSL	6ZSL
<b>nsp14</b> 18,040–19,620	<b>Exonuclease/ methyltransferase</b>	<b>527</b>		<b>59.8</b>	<b>95</b>		
	Full-length	527	1–527	59.8	95	5NFY	
	MTase domain	240	288–527	27.5	95		
<b>nsp15</b> 19,621–20,658	<b>Endonuclease</b>	<b>346</b>		<b>38.8</b>	<b>89</b>		
	Full-length	346	1–346	38.8	89	6W01	6W01
<b>nsp16</b> 20,659–21,552	<b>Methyltransferase</b>	<b>298</b>		<b>33.3</b>	<b>93</b>		
	Full-length	298	1–298	33.3	93	6W4H	6W4H
<b>ORF3a</b> 25,393–26,220		<b>275</b>		<b>31.3</b>	<b>72</b>		
	Full-length	275	1–275	31.3	72	6XDC	6XDC
<b>ORF4</b> 26,245–26,472	<b>Envelope (E) protein</b>	<b>75</b>		<b>8.4</b>	<b>95</b>		
	Full-length	75	1–75	8.4	95	5X29	7K3G
<b>ORF5</b> 26,523–27,387	<b>Membrane glycoprotein (M)</b>	<b>222</b>		<b>25.1</b>	<b>91</b>		
	Full-length	222	1–222	25.1	91		
<b>ORF6</b> 27,202–27,387		<b>61</b>		<b>7.3</b>	<b>69</b>		
	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) <sup>a</sup>	Trivial name construct expressed	Size (aa)	Boundaries	MW (kDa)	Homol. SCoV (%) <sup>b</sup>	Template PDB <sup>c</sup>	SCoV2 PDB <sup>d</sup>
<b>ORF7a</b> 27,394–27,759		<b>121</b>		<b>13.7</b>	<b>85</b>		
	Ectodomain (ED)	66	16–81	7.4	85	1XAK	6W37
<b>ORF7b</b> 27,756–27,887		<b>43</b>		<b>5.2</b>	<b>85</b>		
	Full-length	43	1–43	5.2	85		
<b>ORF8</b> 27,894–28,259		<b>121</b>		<b>13.8</b>	<b>32</b>		
	Full-length	121	1–121	13.8	32		
	w/o signal peptide	106	16–121	12	41	7JTL	7JTL
<b>ORF9a</b> 28,274–29,533	<b>Nucleocapsid (N)</b>	<b>419</b>		<b>45.6</b>	<b>91</b>		
	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
<b>ORF9b</b> 28,284–28,574		<b>97</b>		<b>10.8</b>	<b>72</b>		
	Full-length	97	1–97	10.8	72	6Z4U	6Z4U
<b>ORF14</b> 28,734–28,952		<b>73</b>		<b>8</b>	<b>n.a</b>		
	Full-length	73	1–73	8	n.a		
<b>ORF10</b> 29,558–29,674		<b>38</b>		<b>4.4</b>	<b>29</b>		
	Full-length	38	1–38	4.4	29		

<sup>a</sup>Genome position in nt corresponding to SCoV2 NCBI reference genome entry NC\_045512.2, identical to GenBank entry MN908947.3.

<sup>b</sup>Sequence identities to SCoV are calculated from an alignment with corresponding protein sequences based on the genome sequence of NCBI Reference NC\_004718.3.

<sup>c</sup>Representative PDB that was available at the beginning of construct design, either SCoV or SCoV2.

<sup>d</sup>Representative PDB available for SCoV2 (as of December 2020).

<sup>e</sup>Additional 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) <sup>a</sup> or (mg/ml) <sup>b</sup>	Results	Comments	BMRB	Supplementary Material
<b>nsp1</b>					
fl	5	NMR assigned	Expression only at >20°C; after 7 days at 25°C partial proteolysis	50620 <sup>d</sup>	SI1
GD	>0.5	HSQC	High expression; mainly insoluble; higher salt increases stability (>250 mM)		
<b>nsp2</b>					
CtDR	0.7–1.5	NMR assigned	Assignment with His-tag shown in (Mompéan et al., 2020)	50687 <sup>c</sup>	SI2
<b>nsp3</b>					
UBI	0.7	HSQC	Highly stable over weeks; spectrum overlays with Ubl + IDR		SI3
UBI + IDR	2–3	NMR assigned	Highly stable for >2 weeks at 25°C	50446 <sup>d</sup>	
Macrodomain	9	NMR assigned	Highly stable for >1 week at 25°C and > 2 weeks at 4°C	50387 <sup>d</sup> 50388 <sup>d</sup>	
SUD-N	14	NMR assigned	Highly stable for >10 days at 25°C	50448 <sup>d</sup>	
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 <sup>d</sup>	
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 <sup>d</sup>	
PL <sup>pro</sup>	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 <sup>d</sup>	
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		
<b>nsp5</b>					
fl	55	HSQC	Impaired dimerization induced by artificial N-terminal residues		SI4
<b>nsp7</b>					
fl	17	NMR assigned	Stable for several days at 35°C; stable for >1 month at 4°C	50337 <sup>d</sup>	SI5
<b>nsp8</b>					
fl	17	HSQC	Concentration dependent aggregation; low concentrations favor stability		SI6
<b>nsp9</b>					
fl	4.5	NMR assigned	Stable dimer for >4 months at 4°C and >2 weeks at 25°C	50621 <sup>d</sup> 50622 <sup>d</sup> 50513	SI7
<b>nsp10</b>					
fl	15	NMR assigned	Zn <sup>2+</sup> addition during expression and purification increases protein stability; stable for >1 week at 25°C	50392	SI8
<b>nsp13</b>					
fl	0.5	HSQC	Low expression; protein unstable; concentration above 20 µM not possible		SI9
<b>nsp14</b>					
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		
<b>nsp15</b>					
fl	5	HSQC	Tendency to aggregate at 25°C		SI11
<b>nsp16</b>					
fl	10	Pure protein	Addition of reducing agents; 5% (v/v) glycerol favorable; highly unstable		SI12
<b>ORF3a</b>					
fl	0.6	Pure protein	Addition of detergent during expression (0.05% Brij-58); stable protein		SI13
<b>E protein</b>					
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) <sup>a</sup> or (mg/ml) <sup>b</sup>	Results	Comments	BMRB	Supplementary Material
<b>M Protein</b>					SI15
fl	0.33	Pure protein	Addition of detergent during expression (0.05% Brij-58); stable protein		
<b>ORF6</b>					SI16
fl	0.27	HSQC	Soluble expression without detergent; stable protein; no expression with STREP-tag at N-terminus		
<b>ORF7a</b>					SI17
ED	0.4	HSQC	Unpurified protein tends to precipitate during refolding, purified protein stable for 4 days at 25°C		
<b>ORF7b</b>					SI18
fl	0.6	HSQC	Tendency to oligomerize; solubilizing agents needed		
fl	0.27	HSQC	Addition of detergent during expression (0.1% MNG-3); stable protein		
<b>ORF8</b>					SI19
fl	0.62	HSQC	Tendency to oligomerize		
ΔORF8	0.5	Pure protein			
<b>N protein</b>					SI20
IDR1-NTD- IDR2	12	NMR assigned	High salt (>0.4 M) for increased stability	50618, 50619, 50558, 50557 <sup>d</sup>	
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 <sup>d</sup>	
<b>ORF9b</b>					SI21
fl	0.64	HSQC	Expression without detergent, protein is stable		
<b>ORF14</b>					SI22
fl	0.43	HSQC	Addition of detergent during expression (0.05% Brij-58); stable in detergent but unstable on lipid reconstitution		
<b>ORF10</b>					SI23
fl	2	HSQC	Tendency to oligomerize; unstable upon tag cleavage		

<sup>a</sup>Yields from bacterial expression represent the minimal protein amount in mg/L independent of the cultivation medium. *Italic values indicate yields from CFPS.*

<sup>b</sup>Yields from CFPS represent the minimal protein amount in mg/ml of wheat-germ extract.

<sup>c</sup>COVID19-nmr BMRB depositions yet to be released.

<sup>d</sup>COVID19-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 <sup>15</sup>N-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

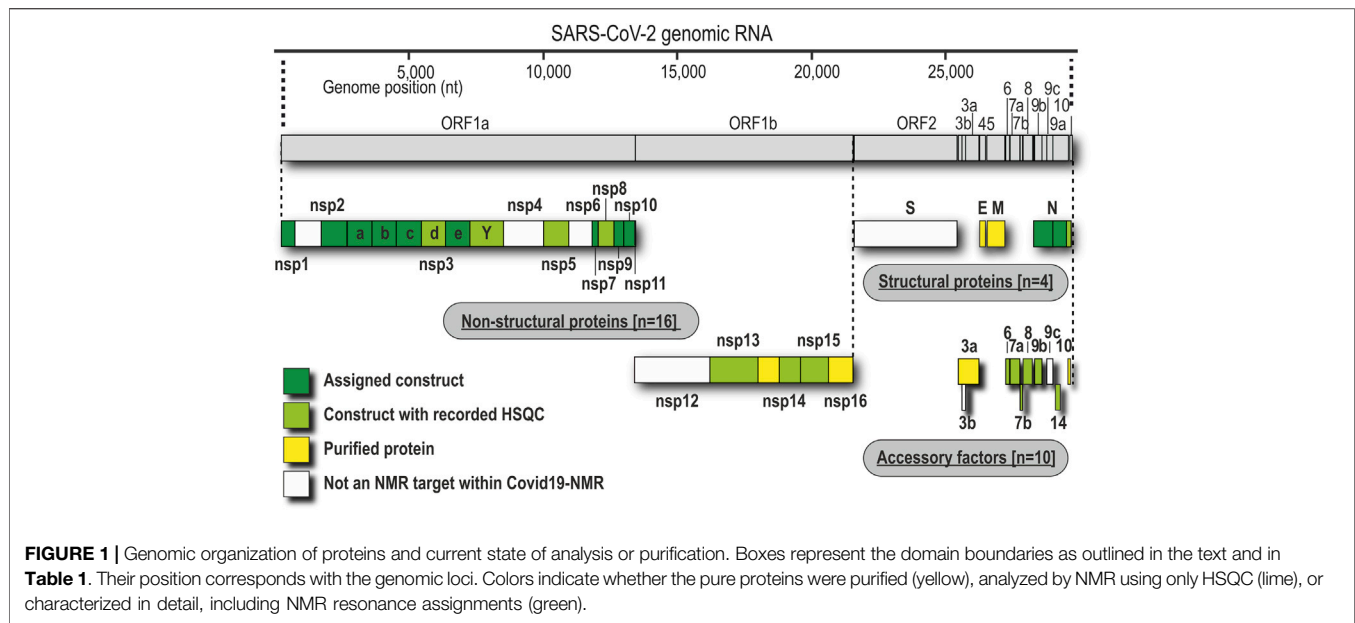
The rationale of construct design for all proteins can be found within the respective protocols in **Supplementary Tables SI1–SI23**. For bacterial production, *E. coli* strains and expression plasmids are given; for WG-CFPS, template

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 SI1–SI23**).

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,

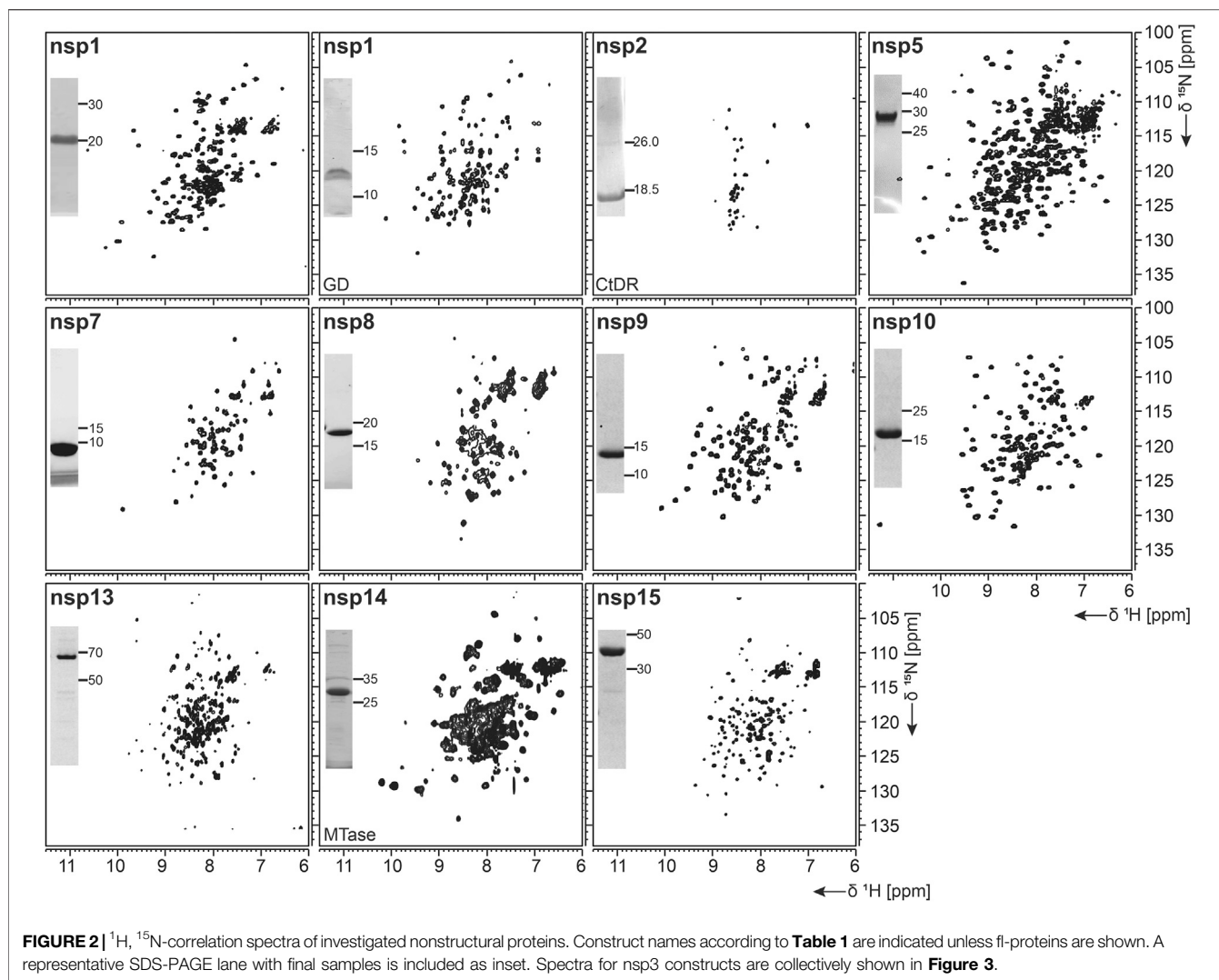
including detailed expression conditions, buffers, incubation times, supplements, storage conditions, yields, and stability, can be found in the respective **Supplementary Tables S11–S123** (see also **Supplementary Tables S1, S2**) and **Tables 1, 2**.

## 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  $\alpha$ - and  $\beta$ -CoV (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).

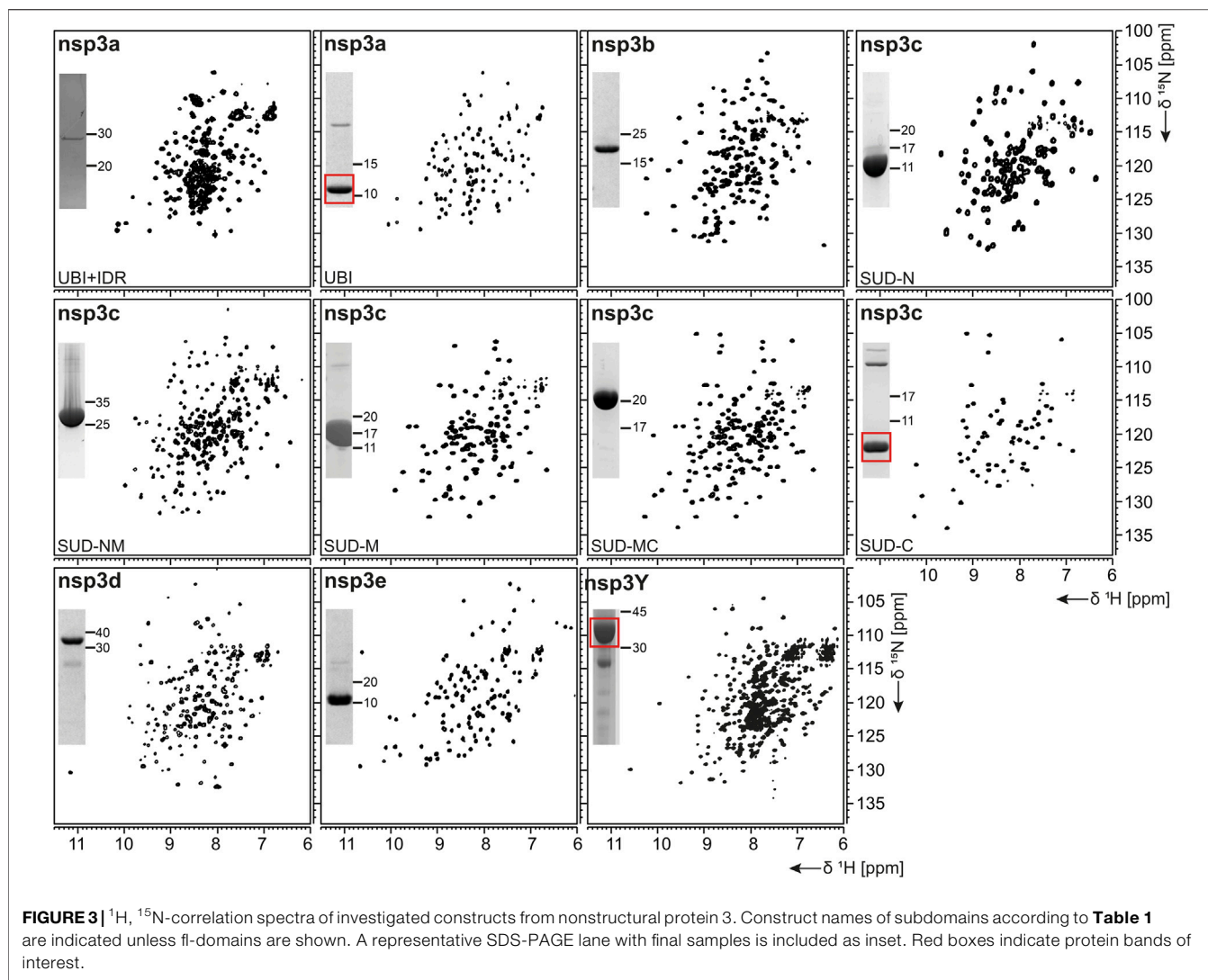
As summarized in **Table 1**, fl-domain boundaries of nsp1 were chosen to contain the first 180 amino acids, in analogy to its closest homolog from SCoV (Snijder et al., 2003). In addition, a shorter construct was designed, encoding only the globular core domain (GD, aa 13–127) suggested by the published 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*

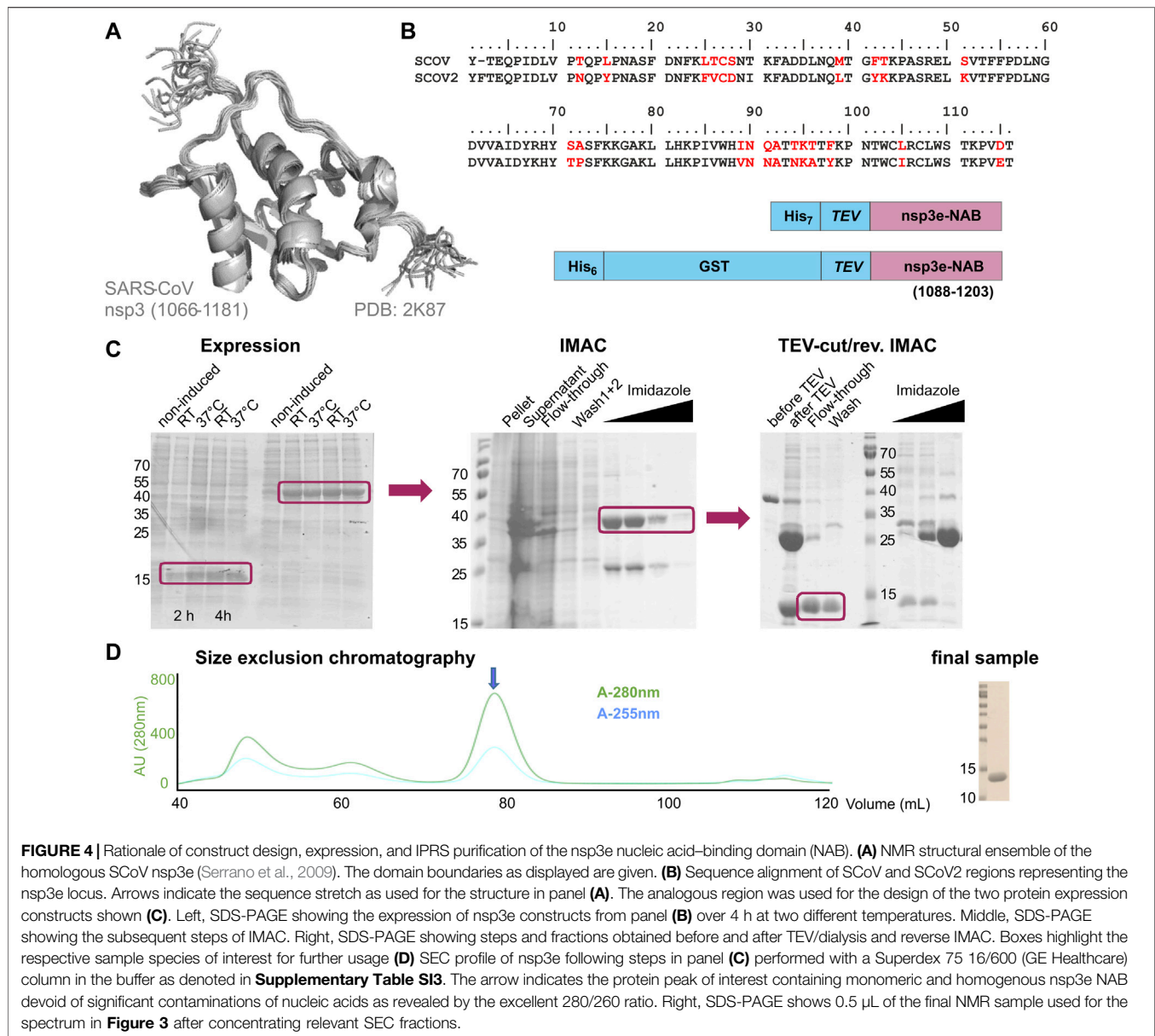
nsp3d comprises the papain-like protease (PL<sup>Pro</sup>) domain of nsp3 and, hence, is one of the two SCoV2 proteases that are responsible for processing the viral polypeptide chain and generating functional proteins (Shin et al., 2020). The domain boundaries of PL<sup>Pro</sup> within nsp3 are set 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<sup>Pro</sup>. 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<sup>Pro</sup>. The His-SUMO-tag allowed simple IMAC purification, followed by cleavage with Ulp1 and isolation of cleaved PL<sup>Pro</sup> via a second IMAC. A final purification step using gel filtration led to pure PL<sup>Pro</sup> 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



**FIGURE 4 |** Rationale of construct design, expression, and IPRES purification of the nsp3e nucleic acid-binding domain (NAB). **(A)** NMR structural ensemble of the homologous SCoV nsp3e (Serrano et al., 2009). The domain boundaries as displayed are given. **(B)** Sequence alignment of SCoV and SCoV2 regions representing the nsp3e locus. Arrows indicate the sequence stretch as used for the structure in panel **(A)**. The analogous region was used for the design of the two protein expression constructs shown **(C)**. Left, SDS-PAGE showing the expression of nsp3e constructs from panel **(B)** over 4 h at two different temperatures. Middle, SDS-PAGE showing the subsequent steps of IMAC. Right, SDS-PAGE showing steps and fractions obtained before and after TEV/dialysis and reverse IMAC. Boxes highlight the respective sample species of interest for further usage **(D)** SEC profile of nsp3e following steps in panel **(C)** performed with a Superdex 75 16/600 (GE Healthcare) column in the buffer as denoted in **Supplementary Table S13**. The arrow indicates the protein peak of interest containing monomeric and homogenous nsp3e NAB devoid of significant contaminations of nucleic acids as revealed by the excellent 280/260 ratio. Right, SDS-PAGE shows 0.5  $\mu$ L of the final NMR sample used for the spectrum in **Figure 3** after concentrating relevant SEC fractions.

assignment carried out at lower concentrations in a deuterated background (ongoing).

### nsp5

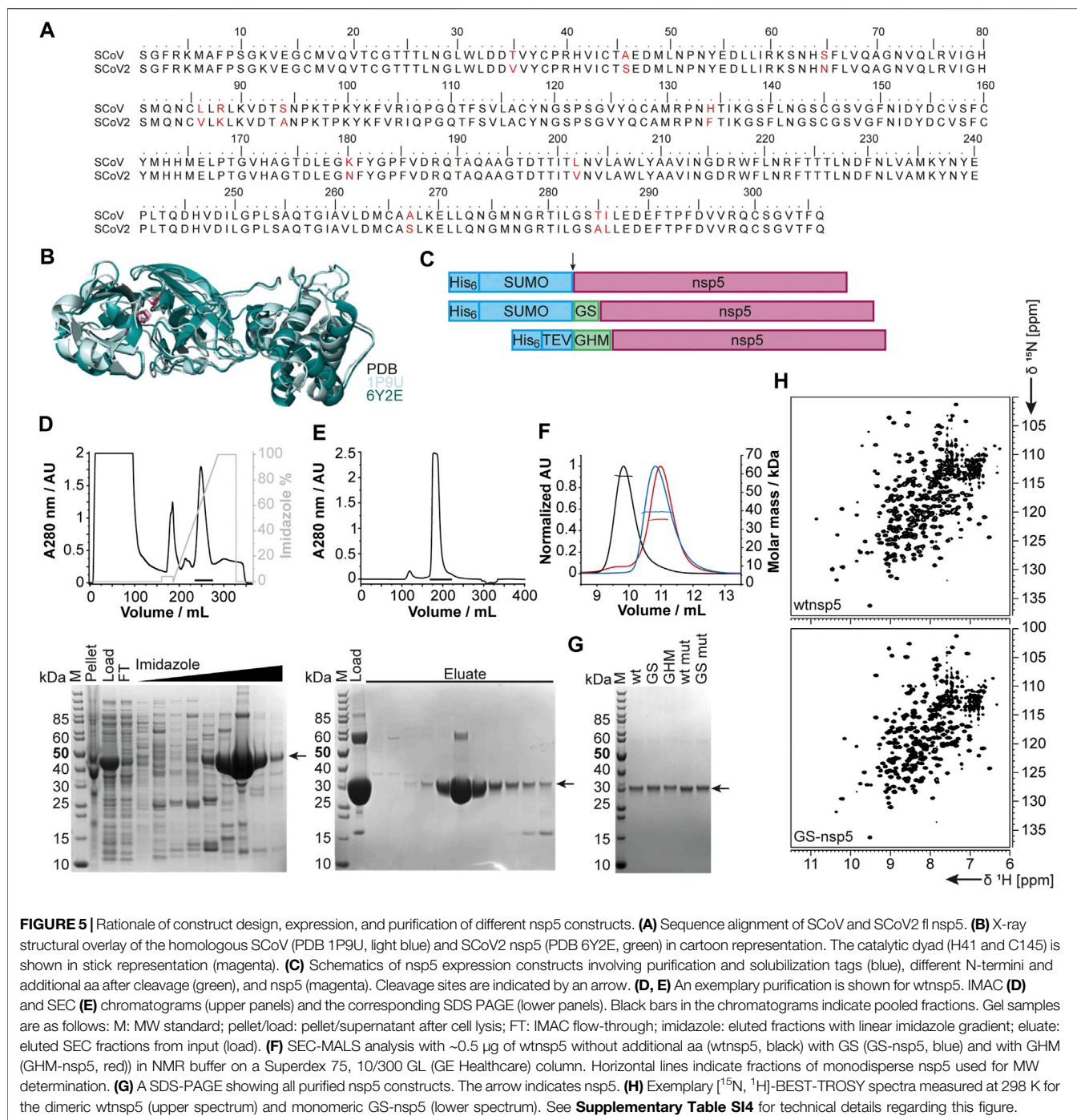
The functional main protease nsp5 ( $M^{Pro}$ ) is a dimeric cysteine protease (Ullrich and Nitsche, 2020). Amino acid sequence and 3D structure of SCoV [PDB 1P9U (Anand et al., 2003)] and SCoV2 (PDB 6Y2E [Zhang et al., 2020]) homologs are highly conserved (**Figures 5A,B**). The dimer interface involves 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 IPRES 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 <sup>13</sup>C- and <sup>15</sup>N-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 Zn<sup>2+</sup> 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.

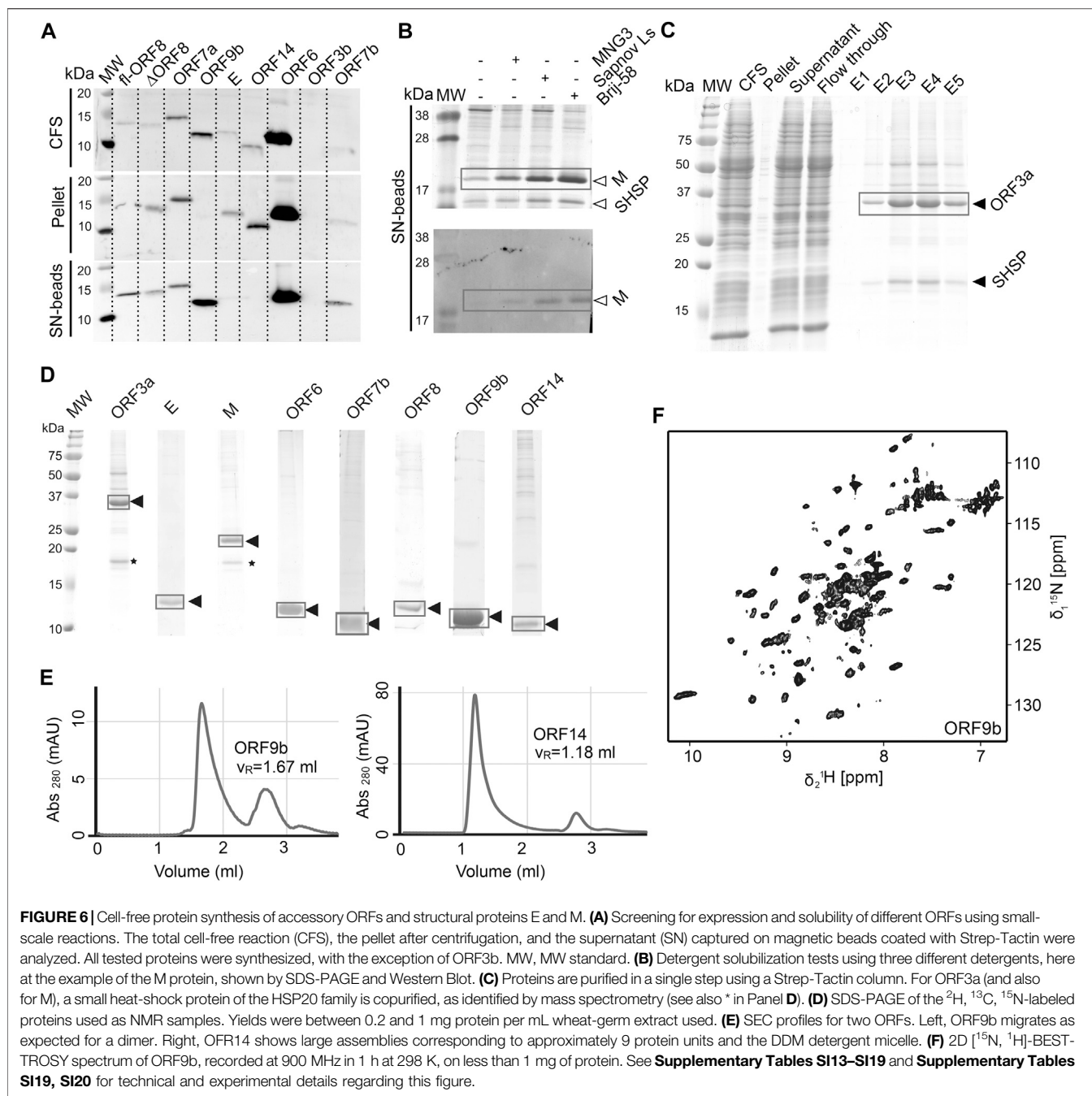
### nsp15

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<sup>pro</sup> 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  $\beta$ -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  $^2\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$  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)

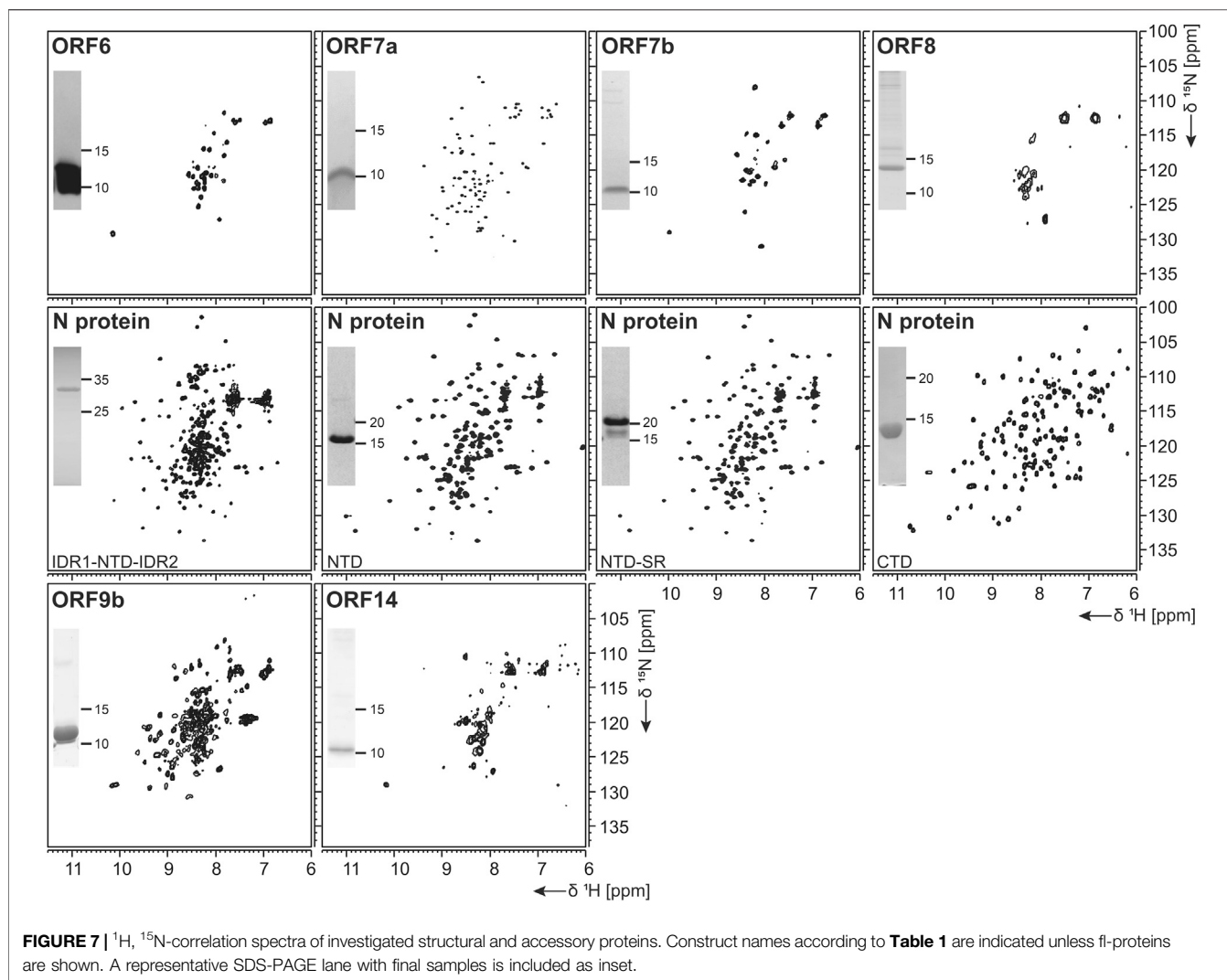
The M protein is the most abundant protein in the viral envelope and is believed to be responsible for maintaining the virion in its characteristic shape (Huang et al., 2004). M is a glycoprotein and sequence analyses predict three domains: A C-terminal endodomain, a TM domain with three predicted helices, 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 IPRS 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 ( $\Delta$ 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  $^{15}\text{N}$ -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  $^{13}\text{C}$ ,  $^{15}\text{N}$ -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  $\beta$ -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 nsps, 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<sup>Pro</sup> 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., Wadhwani, 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

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–CheMIST 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/CIRMMP 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>

- 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 macrodomain 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

- 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., Almasy, 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.0c00309
- Gallo, A., Tsika, A. C., Fourkiotis, N. K., Cantini, F., Banci, L., Sreeramulu, S., et al. (2020). <sup>1</sup>H, <sup>13</sup>C and <sup>15</sup>N 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). <sup>1</sup>H, <sup>13</sup>C and <sup>15</sup>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., Kokic, 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



- $\alpha$ -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 NS5A 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., Dhamotharan, 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
- Krafčikova, 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 macrodomain 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., Ardern, 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., Hagemeijer, 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). <sup>1</sup>H, <sup>13</sup>C and <sup>15</sup>N 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., Bredenbeek, 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.bbame.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., Vonnrhein, 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). <sup>1</sup>H, <sup>13</sup>C, and <sup>15</sup>N 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 adenyllyltransferase 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
- Vennema, 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). (<sup>1</sup>H), (<sup>13</sup>C), and (<sup>15</sup>N) 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  $\alpha$ -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, Tars, 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
- BEST** Band-selective excitation short-transient
- BMRB** Biomagnetic resonance databank
- CFPS** Cell-free protein synthesis
- CoV** Coronavirus
- CTD** C-terminal domain
- DEDD** Asp-Glu-Glu-Asp
- DMS** Dimethylsulfate
- E** Envelope protein
- ED** Ectodomain
- fl** Full-length
- GB1** Protein G B1 domain
- GD** Globular domain
- GF** Gel filtration
- GST** Glutathione-S-transferase
- His** Hisx-tag
- HSP** Heat-shock protein
- HSQC** Heteronuclear single quantum coherence
- IDP** Intrinsically disordered protein
- IDR** Intrinsically disordered region
- IEC** Ion exchange chromatography
- IMAC** Immobilized metal ion affinity chromatography
- IPRS** IMAC-protease cleavage-reverse IMAC-SEC;
- M** Membrane protein
- MERS** Middle East Respiratory Syndrome
- MHV** Murine hepatitis virus
- M<sup>pro</sup>** Main protease
- MTase** Methyltransferase
- N** Nucleocapsid protein
- NAB** Nucleic acid-binding domain
- nsp** Nonstructural protein
- NTD** N-terminal domain
- PL<sup>pro</sup>** Papain-like protease
- RdRP** RNA-dependent RNA polymerase
- S** Spike protein
- SARS** Severe Acute Respiratory Syndrome
- SEC** Size-exclusion chromatography
- SUD** SARS unique domain
- SUMO** Small ubiquitin-related modifier
- TEV** Tobacco etch virus
- TM** Transmembrane
- TROSY** Transverse relaxation-optimized spectroscopy
- Trx** Thioredoxin
- Ubl** Ubiquitin-like domain
- Ulp1** Ubiquitin-like specific protease 1
- WG** Wheat-germ.

Table S1: Overview of labs as assigned to protocols

<b>Protein</b>	<b>System</b>	<b>Main protocol from group(s)</b>	<b>Protocol as „additional information“ from group(s)</b>
<b>nsp1</b>	Bacterial	Carlomagno (fl), Schlundt (GD)	Schlundt (fl)
<b>nsp2</b>	Bacterial	Laurents (CtDR)	-
<b>nsp3a</b>	Bacterial	Blackledge (UBI+IDR), Schlundt (UBI)	-
<b>nsp3b</b>	Bacterial	Schwalbe (Macrodomain)	Alfano (Macrodomain)
<b>nsp3c</b>	Bacterial	Spyroulias (SUD-N, SUD-NM, SUD-M, SUD-MC, SUD-C)	-
<b>nsp3d</b>	Bacterial	Schwalbe (PL <sup>pro</sup> )	Schwalbe (PL <sup>pro</sup> )
<b>nsp3e</b>	Bacterial	Schlundt (NAB)	Schlundt (NAB)
<b>nsp3Y</b>	Bacterial	Hoch (CoV-Y)	-
<b>nsp5</b>	Bacterial	Schwalbe (fl)	Schwalbe (fl, <b>A-D</b> ), Orts (fl, <b>E</b> ), Varga (fl, <b>F</b> ), Bax ((fl, <b>G-H</b> ), Martin ((fl, <b>I</b> )
<b>nsp7</b>	Bacterial	Henzler-Wildman/Kirchdoerfer (fl)	-
<b>nsp8</b>	Bacterial	Henzler-Wildman/Kirchdoerfer (fl)	-
<b>nsp9</b>	Bacterial	Schlundt (fl)	Schlundt (fl, <b>A</b> ), Alfano (fl, <b>B</b> )
<b>nsp10</b>	Bacterial	Schwalbe (fl)	Jaudzems (fl)
<b>nsp13</b>	Bacterial	Schwalbe (fl)	Schwalbe (fl)
<b>nsp14</b>	Bacterial	Jaudzems (fl, MTase)	Schwalbe (fl)
<b>nsp15</b>	Bacterial	Schwalbe (fl)	-
<b>nsp16</b>	Bacterial	Jaudzems (fl)	Jaudzems (fl)
<b>ORF3a</b>	Cell-free	Böckmann (fl)	-
<b>Envelope (ORF4)</b>	Cell-free	Böckmann/Meier (fl)	-
<b>Membrane (ORF5)</b>	Cell-free	Böckmann/Meier (fl)	Böckmann/Meier (fl)
<b>ORF6</b>	Cell-free	Böckmann (fl)	Böckmann (fl)
<b>ORF7a</b>	Bacterial	Muhle-Goll (ED)	-
<b>ORF7b</b>	Bacterial	Schwalbe (fl)	Schwalbe (fl, <b>A-E</b> )
	Cell-free	Böckmann (fl)	-

<b>ORF8</b>	Bacterial	Wiedemann/Ohlenschläger (fl-L84S) Alfano (w/o signal peptide ( $\Delta$ ))	Wiedemann/Ohlenschläger (fl)
<b>Nucleo-capsid (ORF9a)</b>	Cell-free	Böckmann (fl, $\Delta$ )	-
	Bacterial	Pierattelli/Felli (IDR1-NTD-IDR2), Almeida (NTD-SR, NTD), Schlundt (CTD)	-
<b>ORF9b</b>	Cell-free	Böckmann (fl)	Böckmann (fl, <b>A-B</b> )
<b>ORF14</b>	Cell-free	Böckmann/Meier (fl)	-
<b>ORF10</b>	Bacterial	Schwalbe (fl)	Schwalbe (fl, <b>A-D</b> )

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''-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 $\beta$ -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 <sub>6</sub> (analog His <sub>7</sub> )	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- $\beta$ -d-thiogalactopyranoside
LB medium	Lysogeny broth medium
M9 medium	M9 minimal medium
MOPS	3-(N-morpholino)propanesulfonic acid, 4-morpholinepropanesulfonic acid
M <sup>pro</sup>	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 <sub>600</sub>	Optical density at 600 nm
ORF	Open reading frame
PDB	Protein Data Bank
PL <sup>pro</sup>	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp1
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp1 / Leader protein
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEKGVLPQ LEQPYVFIKRS DARTAPHGHVMVELVAELEGIQYGRSGETLGVLPVHVGEIPVAYRKVLLRKN NKGAGGHSYGADLKSFDLGDDELGTDPYEDFQENWNTKHSSGVTRELMRELNGG
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
fl	aa 1-180 (fl nsp1)
GD	aa 13-127 of fl nsp1
<b>5</b>	<b>Ratio for construct design</b>
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
<b>6</b>	<b>Sequence homology (to SCoV)</b>
fl	Identity: 83%; similarity: 89%
GD	Identity: 85%; similarity: 90%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PBD 2GDT, 2HSX SCoV2: PBD 7K3N, 7K7P, 6ZN5, 7JQC, 7K5I
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV: BMRB 7014 SCoV2: BMRB 50620

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
fl	pETM11 (Gunter Stier, EMBL Heidelberg)
GD	pKM263 (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
fl	N-terminal His <sub>6</sub>
GD	N-terminal His <sub>6</sub>
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>

fl	19.90 kDa / 12,950 M <sup>-1</sup> cm <sup>-1</sup> / 5.37
GD	12.93 kDa / 4,470 M <sup>-1</sup> cm <sup>-1</sup> / 6.22
<b>5</b>	<b>Comments on sequence of expressed construct</b>
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
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
fl	0.6 mM IPTG at OD <sub>600</sub> 0.7
GD	1 mM IPTG
<b>10</b>	<b>Cultivation temperature and time</b>
fl	16°C for 18-20 h
GD	16°C for 18-20 h

Table 3a: Protein Purification (fl nsp1)

<b>1</b>	<b>Buffer List</b>
A	50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 100 mM Na <sub>2</sub> SO <sub>4</sub> , 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 <sub>2</sub> SO <sub>4</sub> , 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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (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 <sup>2+</sup> -NTA (Cytiva)), washed first with buffer <b>1A</b> and then with buffer <b>1A</b> containing additional 2 M LiCl, before eluting with 300 mM imidazole in buffer <b>1A</b> .
C	Desalting and TEV-cleavage (0.5 mg TEV protease per 1 L culture) o.n. in buffer <b>1A</b> .
D	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer <b>1B</b> .
E	NMR sample preparation in buffer <b>1C</b> .

Table 3b: Protein Purification (GD nsp1)

<b>1</b>	<b>Buffer List</b>
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 <sub>3</sub> (SEC / final NMR buffer).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus 100 $\mu$ L protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni <sup>2+</sup> -NTA), Elution with 150-500 mM imidazole in buffer <b>1A</b> .
C	Dialysis o.n. in in buffer <b>1A</b> .
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer <b>1A</b> .
E	SEC on HiLoad SD 75 16/600 (GE Healthcare) in buffer <b>1B</b> .
F	NMR sample preparation in buffer <b>1B</b> .

Table 4: Final samples

<b>1</b>	<b>Yield</b>
fl	5 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
GD	< 0.5 mg/L <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
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.
<b>3</b>	<b>Comment on applicability</b>
fl	Suitable for NMR structure determination, fragment screening, interaction studies.
GD	purification needs optimization to obtain more soluble protein

## Additional information

<b>Constructs</b>	<b>Conditions</b>	<b>Comments</b>
aa 1-180 (fl nsp1); His <sub>7</sub> (pET-TEV-Nco (GenScript)), TEV-cleavage site, N-terminal 2 artificial residues "GA".	As above for GD nsp1.	Yields 2.4 mg/L <sup>15</sup> N, <sup>13</sup> 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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp2
<b>2</b>	<b>Region/Name/Further Specification</b>
	C-terminal IDR (CtDR)
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	AYTRYVDNDFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTE RSEKSYELQTPFEIKLAKKFDTFNGECPNFVPLNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASP NECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGATTCGYLPQNAVVKIYCP ACHNSEVGPEHSLAEYHNESGLKTILRKGGRITAFGGCVFSYVGCHNKCAWVPRASANIGCN HTGVVGESEGLNDNLEILQKEKVNINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFK QIVESCNGFKVTKGKAKKGAWNIGEQQSILSPLYAFASEAARVVRSIFSRTLETAQNSVRVLQK AAITILDGISQYSRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKLPVL DWLEEKFKEGVEFLRDGWEIVKFISTCACEIVGGQIVTCAKEIKESVQTFKLVNKFALCADSII IGGAKLKALNLGETFVTHSKGLYRKC VKSREETGLLMPLKAPKEIIFLEGETLPTEVLTEEVLK TGDLQPLEQPTSEAVEAPLVGTPVCINGLMLEIKDTEKYCALAPNMMVTNNTFTLKGK
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 557-601 of complete nsp2 (Ct-DR)
<b>5</b>	<b>Ratio for construct design</b>
	Based on disorder predictions (PrDOS (Ishida and Kinoshita, 2007))
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 55%; similarity: 68%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	-
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV: 50687

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	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.
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> -Trx
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>

	4.92 kDa / - / 3.9
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal „G“, one artificial residue due to TEV-cleavage.
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 star (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.5 mM IPTG at OD <sub>600</sub> 0.6
<b>10</b>	<b>Cultivation temperature and time</b>
	37°C until induction. Following induction, incubation at 25°C for 17 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell lysis in <b>1A</b> (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 <b>1A</b> .
C	Dialysis in buffer <b>1B</b> and TEV cleavage (4°C, 17 h).
D	IMAC2 (after TEV cleavage) (HisTrap crude 5 mL, Cytiva). Elution 10-500 mM imidazole in buffer <b>1A</b> (protein expected in flow-through).
E	Dialysis in buffer <b>1C</b> (4°C, 17 h).
F	Anionic IEC, elution 10-1,000 mM NaCl in buffer <b>1C</b> .
G	Dialysis in buffer <b>1C</b> (4°C, 48 h).
H	Cationic IEC. Elution 10-1,000 mM NaCl in buffer <b>1D</b> (protein expected in flow-through).

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	1.5 mg/L LB medium, 0.7-1.5 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
	No visible precipitation after two weeks at 4°C.
<b>3</b>	<b>Comment on applicability</b>
	Suitable for NMR structure determination, fragment screening, interaction studies.

## SI3: nsp3a

Table 1: General Information

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp3
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp3a Ubiquitin-like domain (Ubl) + IDR
<b>3</b>	<b>Sequence of “fl” protein (aa 1-206 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)</b>
	APTKVTFGDDTVIEVQGYKSVNITFELDERIDKVLNEKCSAYTVELGTEVNEFACVVADAVIKT LQPVSELLTPLGIDLDEWSMATYYLFDESGEFKLASHMYCSFYPPDEDEEEGDCEEEEFEPSTQY EYGTEDDYQGKPLEFGATSAAALQPEEEQEEDWLDDDSQQTVGQQDGSSEDNQTTTIQTIVEVQP QLEMELTPVVQTIE
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
Ubl+ IDR	aa 1-206 of complete nsp3
Ubl	aa 1-111 of complete nsp3
<b>5</b>	<b>Ratio for construct design</b>
Ubl+ IDR	Based on homologous structure from SCoV.
Ubl	Based on disorder prediction, folded domain and SCoV Ubl1.
<b>6</b>	<b>Sequence homology (to SCoV)</b>
Ubl+ IDR	Identity: 58%; Similarity: 75%
Ubl	Identity: 79%; Similarity: 89%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 2GRI; 2IDY
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV: BMRB 7019 SCoV2: BMRB 50446

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
Ubl+ IDR	pET-TEV-Nco (GenScript)
Ubl	pKM263 (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
Ubl+ IDR	N-terminal His <sub>6</sub>

Ubl	N-terminal His <sub>6</sub> -GST
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
Ubl+ IDR	23.50 kDa / 24,410 M <sup>-1</sup> cm <sup>-1</sup> / 3.62
Ubl	12.72 kDa / 14,440 M <sup>-1</sup> cm <sup>-1</sup> / 4.08
<b>5</b>	<b>Comments on sequence of expressed construct</b>
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.
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	1 mM IPTG at OD <sub>600</sub> 0.6-0.8
<b>10</b>	<b>Cultivation temperature and time</b>
Ubl+ IDR	37°C for 5 h
Ubl	18°C for 18 h

Table 3a: Protein Purification (Ubl + IDR)

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Resuspension of cell pellet in 50 mL per liter of culture of <b>1A</b> 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 <sup>2+</sup> -loaded IMAC resin (ThermoFisher scientific) pre-equilibrated with <b>1B</b> at 22°C.
E	Wash IMAC resin with 50 bed volumes of <b>1B</b> .
F	Elute protein from IMAC resin with 5 bed volumes of <b>1C</b> .
G	TEV cleavage with 1 mg TEV per 50 mg protein by dialysis against <b>1D</b> for 18 h at 4°C.
H	Removal of uncleaved protein and tag by elution through Ni <sup>2+</sup> -loaded IMAC resin pre-equilibrated with <b>1B</b> at 22°C.
I	Wash with 5 bed volumes of <b>1B</b> .
J	SEC with HiLoad SD 75 pg column (GE Healthcare) pre-equilibrated with <b>1E</b> at 4°C.

Table 3b: Protein Purification (Ubl)

<b>1</b>	<b>Buffer List</b>
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 <sub>3</sub> (dialysis after IMAC / TEV-cleavage)
C	25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl, 0.02% NaN <sub>3</sub> , pH7 (SEC / final NMR buffer)
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus 100 µL protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni <sup>2+</sup> -NTA), Elution with 150-500 mM imidazole in buffer <b>1A</b>
C	Dialysis o.n. in in buffer <b>1B</b>
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer <b>1B</b>
E	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer <b>1C</b>
F	NMR sample preparation in buffer <b>1C</b>

Table 4: Final sample

<b>1</b>	<b>Yield</b>
Ubl+ IDR	0.7 mg/L <sup>15</sup> N-M9 medium
Ubl	2-3 mg/L <sup>15</sup> N-M9 medium
<b>1b</b>	<b>A260/280 ratio</b>
Ubl+ IDR	0.57
Ubl	0.6
<b>2</b>	<b>Stability</b>
Ubl+	2 weeks at 25°C.

IDR	
Ubl	Very stable over weeks.
<b>3</b>	<b>Comment on applicability</b>
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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp3
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp3b / Macrodomain
<b>3</b>	<b>Sequence of “fl” protein (aa 207-376 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)</b>
	VNSFSGYLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATNNAMQV ESDDYIATNGPLKVGGSCLVLSGHNLAHKHCLHVVGPNVKNKGEDIQLLKSAYENFNQHEVLLAPL LSAGIFGADPIHSLRVCVDTVRTNVYLAVFDPKLNLYDKLVSSFLEMK
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 207-376 of complete nsp3
<b>5</b>	<b>Ratio for construct design</b>
	Based on homologous structure from SCoV (PDB 6VXS).
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 74%; similarity: 84%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV2: PDB 6W6Y, 6YWM, 6YWL, 6YWK, 6WEY, 7KG3, 6W02, 6WOJ, 6WEN, 6WCF, 6VXS, 7JME
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV2: BMRB 50387 (apo), 50388 (holo)

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pET28a(+) (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub>
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	18.65 kDa / 10,430 M <sup>-1</sup> cm <sup>-1</sup> / 7.20
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal “GHM” three artificial residues due to TEV-cleavage and construct design.

<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> T7 Express
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.2 mM IPTG at OD <sub>600</sub> 0.6-0.7
<b>10</b>	<b>Cultivation temperature and time</b>
	18-20°C for 16-18 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (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 <b>1A</b> .
C	TEV-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer <b>1B</b> .
D	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with 500 mM imidazole in buffer <b>1A</b> .
E	SEC (HiLoad 26/600 SD 200 pg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer <b>1C</b> (elution volume 245-290 mL).
F	NMR sample preparation in buffer <b>1C</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	94 mg/L <sup>15</sup> N-M9 medium, 9 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
	Stable throughout measurement (7 days, 298 K). No significant precipitation or degradation observed after storage at 4°C for 2 weeks.
<b>3</b>	<b>Comment on applicability</b>

Suitable for NMR structure determination, fragment screening, interaction studies.

#### Additional information

Constructs	Conditions	Comments
aa 206-374 of complete nsp3; His <sub>6</sub> -GST (mod pET9d), TEV-cleavage site, N-terminal "GAM" three artificial residues. Based on boundaries from crystal structure (PDB 6W6Y).	<b>IMAC buffer:</b> 50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 5% (v/v) glycerol, 50 mM imidazole, 1 mM DTT. <b>Cleavage buffer:</b> 50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 1 mM DTT. <b>SEC/final buffer:</b> 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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp3
<b>2</b>	<b>Region/Name/Further Specification</b>
<b>SUD-N</b>	nsp3c / SARS Unique Domain (SUD) -N
<b>SUD-NM</b>	nsp3c / SUD-NM
<b>SUD-M</b>	nsp3c / SUD-M
<b>SUD-MC</b>	nsp3c / SUD-MC
<b>SUD-C</b>	nsp3c / SUD-C
<b>3</b>	<b>Sequence of “fl” protein (aa 409-743 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)</b>
	QDDKKIKACVEEVTTTLEETKFLTENLLLYIDINGNLHPDSATLVSDIDITFLKKDAPYIVGDVV QEGVLTAVVIPTKKAGGTTEMLAKALRKVPTDNYITTPGQGLNGYTVVEEAKTVLKKCKSAFY ILPSIISNEKQEILGTVSWNLREMLAHAETRKLMPVCVETKAIVSTIQRKYKGIKIQEGVVDYG ARFYFYTSKTTVASLINTLNDLNETLVTMPLGYVTHGLNLEEAAARYMRSLKVPATVSVSSPDA VTAYNGYLTSSSKTPEEHFIETISLAGSYKDWYSYSGQSTQLGIEFLKRGDKSVYYTSPNPTTFHLD GEVITFDNLKTLLS
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
<b>SUD-N</b>	aa 409-548 of complete nsp3
<b>SUD-NM</b>	aa 409-675 of complete nsp3
<b>SUD-M</b>	aa 551-675 of complete nsp3
<b>SUD-MC</b>	aa 551-743 of complete nsp3
<b>SUD-C</b>	aa 680-743 of complete nsp3
<b>5</b>	<b>Ratio for construct design</b>
<b>SUD-N</b>	Based on X-ray structure of homologue nsp3c from SCoV (PDB 2W2G).
<b>SUD-NM</b>	Based on X-ray structure of homologue nsp3c from SCoV (PDB 2W2G).
<b>SUD-M</b>	Based on X-ray structure of homologue nsp3c from SCoV (PDB 2W2G).
<b>SUD-MC</b>	Based on NMR structure of homologue nsp3c from SCoV (PDB 2KQV, 2KQW).
<b>SUD-C</b>	Based on NMR structure of homologue nsp3c from SCoV (PDB 2KAF).
<b>6</b>	<b>Sequence homology (to SCoV)</b>

<b>SUD-N</b>	Identity: 69%, similarity: 81.6%
<b>SUD-NM</b>	Identity: 74%, similarity: 85.4%
<b>SUD-M</b>	Identity: 82%, similarity: 89.6%
<b>SUD-MC</b>	Identity: 79%, similarity: 88.7%
<b>SUD-C</b>	Identity: 73%, similarity: 87.7%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	-
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
<b>SUD-N</b>	SCoV2: BMRB 50448
<b>SUD-NM</b>	Ongoing
<b>SUD-M</b>	SCoV2: BMRB 50516 SUD-M
<b>SUD-MC</b>	Ongoing
<b>SUD-C</b>	SCoV2: BMRB 50517 SUD-C

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
<b>SUD-N</b>	pGEX4T1 (Addgene)
<b>SUD-NM</b>	pGEX4T1 (Addgene)
<b>SUD-M</b>	pET28a(+) (Addgene)
<b>SUD-MC</b>	pET28a(+) (Addgene)
<b>SUD-C</b>	pGEX4T1 (Addgene)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
<b>SUD-N</b>	N-terminal GST
<b>SUD-NM</b>	N-terminal GST
<b>SUD-M</b>	N-terminal His <sub>6</sub>
<b>SUD-MC</b>	N-terminal His <sub>6</sub>
<b>SUD-C</b>	N-terminal GST

<b>3</b>	<b>Cleavage Site</b>
	Thrombin
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
<b>SUD-N</b>	15.54 kDa / 8,940 M <sup>-1</sup> cm <sup>-1</sup> / 5.04
<b>SUD-NM</b>	29.60 kDa / 26,360 M <sup>-1</sup> cm <sup>-1</sup> / 6.03
<b>SUD-M</b>	14.27 kDa / 17,420 M <sup>-1</sup> cm <sup>-1</sup> / 8.71
<b>SUD-MC</b>	21.94 kDa / 28,880 M <sup>-1</sup> cm <sup>-1</sup> / 6.58
<b>SUD-C</b>	7.42 kDa / 11,460 M <sup>-1</sup> cm <sup>-1</sup> / 4.82
<b>5</b>	<b>Comments on sequence of expressed construct</b>
<b>SUD-N</b>	N-terminal „GS" two artificial residues due to thrombin-cleavage
<b>SUD-NM</b>	N-terminal „GS" two artificial residues due to thrombin-cleavage
<b>SUD-M</b>	N-terminal „GSHM" four artificial residues due to thrombin-cleavage and cloning
<b>SUD-MC</b>	N-terminal „GSHM" four artificial residues due to thrombin-cleavage and cloning
<b>SUD-C</b>	N-terminal „GS" two artificial residues due to thrombin-cleavage
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	1 mM IPTG at OD <sub>600</sub> 0.6-0.8
<b>10</b>	<b>Cultivation temperature and time</b>
	18°C for 18-20 h

Table 3a: Protein Purification (SUD-N and SUD-NM)

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus 25 $\mu$ L protease inhibitor cocktail (Sigma Aldrich P8849) and 2 mM DTT) by sonication, after sonication incubation with 25 $\mu$ L DNase (1 mg/mL) for 10 min on ice.
B	AC - GSTrap (GE Healthcare) (wash buffer <b>1A</b> ).
C	Cleavage on column (100 $\mu$ 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 <b>1A</b> , elution of GST with buffer <b>1C</b> and buffer exchange with Amicon Ultra 15 mL centrifugal filter membrane (10,000 MWCO) (Merck Millipore) to buffer <b>1B</b> .
E	SEC - SD Increase 75 10/300 GL (GE Healthcare) in buffer <b>1B</b> .
F	NMR sample preparation in buffer <b>1B</b> .

Table 3b: Protein Purification (SUD-M and SUD-MC)

<b>1</b>	<b>Buffer List</b>
A	50 mM Tris-HCl (pH 8.0), 500 mM NaCl (Cell disruption / IMAC).
B <b>SUD-M</b>	50 mM NaPi (pH 7.2), 50 mM NaCl, 2 mM EDTA, 2 mM DTT (SEC / NMR buffer).
B <b>SUD-MC</b>	50 mM NaPi (pH 7.6), 50 mM NaCl, 2 mM EDTA, 2 mM DTT (SEC / NMR buffer).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus 10 mM imidazole and 25 $\mu$ L protease inhibitor cocktail (Sigma Aldrich P8849) and 2 mM DTT) by sonication, before and after sonication incubation with 50 $\mu$ L DNase (1 mg/mL) for 15 min on ice.
B	IMAC - HisTrap (Ni <sup>2+</sup> ) (GE Healthcare), a step gradient elution of imidazole in buffer <b>1A</b> (10, 20, 40, 100, 200, 400 mM). <b>SUD-M</b> eluted mostly in 100 mM imidazole in buffer <b>1A</b> and a small amount in fraction 200 mM imidazole in buffer <b>1A</b> . <b>SUD-MC</b> eluted mostly in 100 mM imidazole in buffer <b>1A</b> and a small amount in 40 mM imidazole in buffer <b>1A</b> .
C	Buffer exchange with Amicon Ultra 15 mL centrifugal filter membrane (10,000 MWCO) (Merck Millipore) in buffer <b>1B SUD-M</b> and <b>SUD-MC</b> respectively.
D	Cleavage in solution (100 $\mu$ L thrombin (10 mg/mL) per 0.5 L culture) for <b>SUD-M</b> : 1 h at 4°C and then 1 h at rt; <b>SUD-MC</b> : 16 h at 4°C.
E	SEC - Superdex Increase 75 10/300 GL (GE Healthcare) in buffer <b>1C-SUD-M, 1C-SUD-MC</b> .
F	NMR sample preparation in buffer <b>1C-SUD-M, 1C-SUD-MC</b> .

Table 3c: Protein Purification (SUD-C)

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus 25 $\mu$ L protease inhibitor cocktail (Sigma Aldrich P8849) and 2 mM DTT) by sonication, after sonication incubation with 25 $\mu$ L DNase (1 mg/mL) for 10 min on ice.
B	AC with GSTrap (GE Healthcare) (wash buffer <b>1A</b> and then wash with buffer <b>1B</b> ).
C	Elution with buffer <b>1C</b> , buffer exchange with Amicon Ultra 15 mL centrifugal filter membrane (10,000 MWCO) (Merck Millipore) to buffer <b>1D</b> .
D	Cleavage in solution (350 $\mu$ 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 <b>1D</b> .
F	NMR sample preparation in buffer <b>1D</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
<b>SUD-N</b>	13.92 mg/L $^{15}\text{N}$ or $^{13}\text{C}$ , $^{15}\text{N}$ -M9 medium
<b>SUD-NM</b>	17.25 mg/L $^{15}\text{N}$ or $^{13}\text{C}$ , $^{15}\text{N}$ -M9 medium
<b>SUD-M</b>	8.47 mg/L $^{15}\text{N}$ or $^{13}\text{C}$ , $^{15}\text{N}$ -M9 medium
<b>SUD-MC</b>	12.06 mg/L $^{15}\text{N}$ or $^{13}\text{C}$ , $^{15}\text{N}$ -M9 medium
<b>SUD-C</b>	4.70 mg/L $^{15}\text{N}$ or $^{13}\text{C}$ , $^{15}\text{N}$ -M9 medium
<b>1b</b>	<b>A260/280 ratio</b>
<b>SUD-N</b>	0.55
<b>SUD-NM</b>	0.50
<b>SUD-M</b>	0.81
<b>SUD-MC</b>	0.62
<b>SUD-C</b>	0.71
<b>2</b>	<b>Stability</b>
<b>SUD-N</b>	Stable throughout NMR spectra acquisition (10 days, 298 K). No significant precipitation or degradation observed after thawing from -80°C. Very stable construct.
<b>SUD-NM</b>	Stable throughout measurement (7 days, 298 K). No significant precipitation or degradation observed after defrosting from -80°C.
<b>SUD-M</b>	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.

<b>SUD-MC</b>	Stable throughout measurement (7 days, 298 K). No significant precipitation or degradation observed after thawing from -80°C.
<b>SUD-C</b>	Stable throughout measurement (10 days, 298 K). No significant precipitation or degradation observed after thawing from -80°C. Stable construct.
<b>3</b>	<b>Comment on applicability</b>
	Suitable for NMR structure determination, fragment screening, interaction studies.

## SI3: nsp3d

Table 1: General Information

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp3
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp3d / papain-like protease / PL <sup>pro</sup>
<b>3</b>	<b>Sequence of “fl” protein (aa 743-1060 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)</b>
	SLREVRTIKVFTTVDNINLHTQVVDMSTYGGQFGPTYLDGADVTKIKPHNSHEGKTFYVLPN DDTLRVEAFEYYHTTDPSTFLGRYMSALNHTKKWKYPQVNGLTSTIKWADNNCYLATALLLTQQ IELKFNPPALQDAYRARAGEAANFCALILAYCNKTVGELGDVRETMSYLFQHANLDSCKRVL NVVCKTCGQQQTTLKGVEAVMYMGTLSEYQFKKGVQIPCTCGKQATKYLQQESPFVMMMSA PPAQYELKHGFTFCASEYTGNYQCGHYKHITSKETLYCIDGALLTKSSEYKGPITDVVFYKENSY TTTIK
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 743-1060 of complete nsp3
<b>5</b>	<b>Ratio for construct design</b>
	Based on homologous structure from SCoV (PDB 4M0W)
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 83%; similarity: 91%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 4M0W, 2FE8 SCoV2: PDB 6W9C
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pE-SUMO (LifeSensors)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> -SUMO
<b>3</b>	<b>Cleavage Site</b>
	Ulp1
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	35.99 kDa / 45,270 M <sup>-1</sup> cm <sup>-1</sup> / 8.17
<b>5</b>	<b>Comments on sequence of expressed construct</b>

	No artificial residues due to Ulp1-cleavage and construct design.
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.2 mM IPTG at OD <sub>600</sub> 0.6-0.7 (addition of 50 μM ZnCl <sub>2</sub> )
<b>10</b>	<b>Cultivation temperature and time</b>
	18-20°C for 16-18 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
A	20 mM Tris-HCl (pH 8.0), 300 mM NaCl, 10 mM imidazole, 50 μM ZnCl <sub>2</sub> , 10 mM bME (cell disruption / IMAC).
B	10 mM HEPES (pH 7.4), 100 mM NaCl, 50 μM ZnCl <sub>2</sub> , 10 mM bME (dialysis after IMAC / TEV-cleavage).
C	10 mM HEPES (pH 7.4), 100 mM NaCl, 50 μM ZnCl <sub>2</sub> , 5 mM DTT (SEC).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (addition of 50 μM ZnCl <sub>2</sub> ) by microfluidization.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer <b>1A</b> .
C	Ulp1-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer <b>1B</b> .
D	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with 500 mM imidazole in buffer <b>1A</b> .
E	SEC (HiLoad 26/600 SD 75 pg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer <b>1C</b> (elution volume 180-220 mL).

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	12 mg/L <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
	Tendency to aggregate.
<b>3</b>	<b>Comment on applicability</b>

Suitable for fragment screening, interaction studies.

#### Additional information

<b>Constructs</b>	<b>Conditions</b>	<b>Comments</b>
aa 743-1060 of complete nsp3; His <sub>6</sub> (pET28a(+)) (GenScript), TEV-cleavage site, N-terminal "GHM" three artificial residues.	Native (as above)	Weak expression, less protein.

## SI3: nsp3e

Table 1: General Information

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp3
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp3e / NAB globular domain
<b>3</b>	<b>Sequence of "fl" protein (aa 1080-1203 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)</b>
	YFTEQPIDLVPNQYPNASFDNFKFVCDNIKFADDLNQLTGYYKPPASRELKVTFPPDLNGDVVA IDYKHYTPSFKKGAKLLHKPIVWHVNNATNKATYKPNTWCIRCLWSTKPVET
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1088-1203 of complete nsp3
<b>5</b>	<b>Ratio for construct design</b>
	Based on boundaries from NMR structure of homologue nsp3e from SARS-CoV (2K87).
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 82%; similarity: 89%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 2K87
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV: BMRB 15723; SCoV2: BMRB 50334

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pKM263 (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> -GST
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	13.75 kDa / 25,565 M <sup>-1</sup> cm <sup>-1</sup> / 8.9
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal „GAMG" four artificial residues due to TEV-cleavage and construct design.
<b>6</b>	<b>Used expression strain</b>

	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	1 mM IPTG at OD <sub>600</sub> 0.7
<b>10</b>	<b>Cultivation temperature and time</b>
	20-22°C for 18-20 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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 <sub>3</sub> (dialysis after IMAC / TEV-cleavage).
C	25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl, 0.02% (w/v) NaN <sub>3</sub> (SEC / final NMR buffer).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus 100 µL protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni <sup>2+</sup> -NTA) (Carl Roth, Germany), elution with 150-500 mM imidazole in buffer <b>1A</b> .
C	Dialysis o.n. in buffer <b>1B</b> .
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer <b>1B</b> .
E	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer <b>1C</b> .
F	NMR sample preparation in buffer <b>1C</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	3.5 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
<b>2</b>	<b>A260/280 ratio</b>
	0.74
<b>3</b>	<b>Stability</b>
	Stable throughout measurement (7 days, 298 K). No significant precipitation or degradation observed after storage at 4°C for 5 weeks.
<b>4</b>	<b>Comment on applicability</b>



Suitable for NMR structure determination, fragment screening, interaction studies.

#### Additional information

<b>Constructs</b>	<b>Conditions</b>	<b>Comments</b>
NAB (aa 1088-1203) of complete nsp3; His <sub>7</sub> (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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp3
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp3-Y / Cov-Y
<b>3</b>	<b>Sequence of “fl” protein (aa 1638-1945 of complete nsp3, according to NCBI Reference Sequence NC_045512.2)</b>
	DTFCAGSTFISDEVARDLSLQFKRPINPTDQSSYIVDSVTVKNGSIHLYFDKAGQKTYERHSLSHF VNLDNLRANNTKGSLPINVIVFDGKSKCEESSAKSASVYYSQLMCQPILLDDQALVSDVGDSAE VAVKMFDAYVNTFSSTFNVPMEKLTAVATAEAELAKNVSLDNVLSSTFISAARQGFVDSVET KDVVECLKLSHQSDIEVTGDCSNMYMLTYNKVENMTPRDLGACIDCSARHINAQVAKSHNIAL IWNVKDFMSLSEQLRKQIRSAAKNNLFPKLTCAATTRQVVNVVTTKIALKGG
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1638-1945 of complete nsp3
<b>5</b>	<b>Ratio for construct design (detailed and comprehensible)</b>
	We took the C-terminal part of nsp3 after predicted transmembrane region and Y1 domain that consists of two sequential zinc finger motifs.
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 89%; similarity: 96%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	-
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pET28b(+) (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub>
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	34 kDa / 17,420 M <sup>-1</sup> cm <sup>-1</sup> / 6.66
<b>5</b>	<b>Comments on sequence of expressed construct</b>

	N-terminal „G" one artificial residue due to TEV-cleavage.
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labeling)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.5 mM IPTG at OD <sub>600</sub> 0.7
<b>10</b>	<b>Cultivation temperature and time</b>
	18°C for 15-16 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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.
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> by sonication.
B	IMAC (gravity flow Ni <sup>2+</sup> -NTA) (Thermo Scientific), wash with buffer <b>1B</b> and elution with 250 mM imidazole in buffer <b>1B</b> .
C	TEV-cleavage (5% (w/w) TEV protease per approximate amount of the protein) in buffer <b>1C</b> o.n. at rt.
D	Inv. IMAC (gravity flow Ni <sup>2+</sup> -NTA) in buffer <b>1C</b> .
E	SEC on 10/300 GL SD 200 (GE Healthcare) in buffer <b>1D</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	12 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
	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.
<b>3</b>	<b>Comment on applicability</b>

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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp5
<b>2</b>	<b>Region/Name/Further Specification</b>
	3C-like protease (3CL <sup>pro</sup> ) / main protease (M <sup>pro</sup> )
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSN HNFLVQAGNVQLRVIGHSMQNCVLKLVDTANPKTPKYKVFRIQPGQTFSVLACYNGSPSGVY QCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHMELPTGVHAGTDLEGNFYGPFVDRQ TAQAAGTDTTITVNLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILG PLSAQTGIAVLDMCASLKELLQNGMNGRTILGSALLEDEFTPFDDVVRQCSGVTFQ
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-306 (fl nsp5)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 96%; similarity: 99.7%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 1P9U, 6LU7 SCoV2: PDB 6Y2E, 5R7Y, 6Y84, 7K3T
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV: BMRB 17251

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pE-SUMO (LifeSensors)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> -SUMO
<b>3</b>	<b>Cleavage Site</b>
	Ulp1
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	33.80 kDa / 32,890 M <sup>-1</sup> cm <sup>-1</sup> / 5.95
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	No artificial residues due to TEV-cleavage and construct design.

<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.2 mM IPTG at OD <sub>600</sub> 0.6-0.7
<b>10</b>	<b>Cultivation temperature and time</b>
	18-20°C for 16-18 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (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 <b>1A</b> .
C	Ulp1-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer <b>1B</b> .
D	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with 500 mM imidazole in buffer <b>1A</b> .
E	SEC (HiLoad 26/600 SD 75 µg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer <b>1C</b> (elution volume 170-210 mL).
F	NMR sample preparation in buffer <b>1D</b> .

Table 4: Final sample

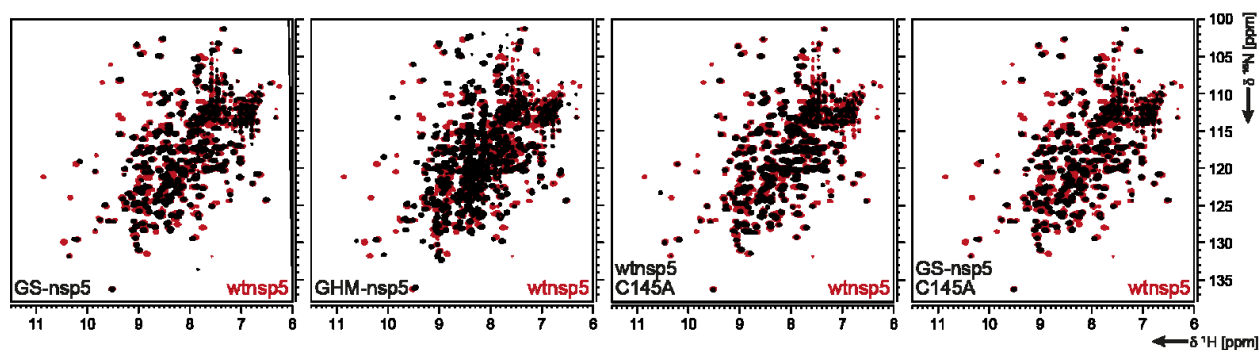
<b>1</b>	<b>Yield</b>
	55 mg/L <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
	No significant precipitation or degradation observed after storage at -80°C for a month.

<b>3</b>	<b>Comment on applicability</b>
	Suitable for NMR structure determination, fragment screening, interaction studies.

### Additional information

	<b>Constructs</b>	<b>Conditions</b>	<b>Comments</b>
<b>A</b>	aa 1-306 (fl nsp5) C145A mutation; His <sub>6</sub> -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 <sup>15</sup> N-M9 medium).
<b>B</b>	aa 1-306 (fl nsp5); His <sub>6</sub> -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 <sup>15</sup> N-M9 medium, 36 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium, 20 mg/L <sup>2</sup> H, <sup>13</sup> C, <sup>15</sup> N E. coli-OD2 CDN medium (Silantes)).
<b>C</b>	aa 1-306 (fl nsp5) C145A mutation; His <sub>6</sub> -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 <sup>15</sup> N-M9 medium).
<b>D</b>	aa 1-306 (fl nsp5); His <sub>6</sub> (pet28a+) (GenScript), TEV-cleavage site; N-terminal "GHM" three artificial residues.	Native (as above) <b>IMAC buffer (1A):</b> 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 <sup>15</sup> N-M9 medium, 10 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium).
<b>E</b>	aa 1-306 (fl nsp5); GST and His <sub>6</sub> -tag (pET-28a+) (GenScript), TEV and auto cleavage site for M <sup>pro</sup> , N-terminal „GS“ and C-terminal "GPHHHHHH" ten artificial residues.	<b>IMAC buffer:</b> 50 mM Tris-HCl (pH 8.0), 200 mM NaCl, 20 mM imidazole. <b>SEC-buffer:</b> 50 mM NaPi (pH 7.6), 50 mM NaCl, 0.02% (w/v) NaN <sub>3</sub> . <b>NMR-buffer:</b> 50 mM NaPi (pH 7.6), 50 mM NaCl, 0.02% (w/v) NaN <sub>3</sub> , 5 mM bME.	Yields 20 mg/L <sup>15</sup> 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.
<b>F</b>	aa 1-306 (fl nsp5); C-terminal His <sub>6</sub> -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.	<b>IMAC buffer:</b> 20 mM Tris-HCl (pH 7.33), 150 mM NaCl, 20 mM imidazole. <b>Storage buffer:</b> 20 mM Tris-HCl (pH 7.33), 150 mM NaCl.	Yields 5 mg/L <sup>15</sup> N-M9 medium. Stable for 2-3 weeks at 4°C at low micromolar concentration.
<b>G</b>	aa 1-306 (fl nsp5) C145A mutation; His <sub>6</sub> -GB1 (pET24a+) (GenScript), TEV-cleavage site, no artificial residues.	<b>IMAC buffer:</b> 20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 20 mM imidazole, 0.5 mM TCEP-HCl. <b>SEC/NMR buffer:</b> 10 mM NaPi (pH 7.0), 0.5 mM TCEP-HCl.	Yields ≥ 70 mg/L <sup>15</sup> N, <sup>2</sup> H, <sup>15</sup> N-M9, and <sup>2</sup> H, <sup>13</sup> C, <sup>15</sup> 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.

<b>H</b>	aa 1-306 (fl nsp5); His <sub>6</sub> -GB1 (pET24a(+)) (GenScript), TEV-cleavage site, no artificial residues.	As above (G).	Negligible expression when induced in <sup>15</sup> N-M9 medium at 25°C, 30°C, and 37°C, with 0.5-1 mM IPTG.
<b>I</b>	aa 1-306 (fl nsp5); His <sub>6</sub> -GST (pGEX-6p-1 (Genewiz)), autolytic and HRV 3C cleavage site, no artificial residues.	<b>IMAC buffer:</b> 25 mM Tris-HCl (pH 7.8), 150 mM NaCl, 5 mM imidazole, 1 mM bME. <b>Cleavage buffer:</b> 25 mM Tris-HCl (pH 7.8), 150 mM NaCl, 1 mM DTT. <b>SEC buffer:</b> 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 [<sup>15</sup>N, <sup>1</sup>H]-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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp7
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp7
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	SKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCVQLHNDILLAKDTTEAFEKMSVLLSVLLSMQ GAVDINKLCEEMLDNRATLQ
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-83 (fl nsp7)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 98.8%; similarity: 100%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	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
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV: PDB 1YSY, BMRB 6513, PDB 2KYS, BMRB 16981 SCoV2: BMRB 50337

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pET46
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> , enterokinase
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	9.24 kDa / 5500 cm <sup>-1</sup> M <sup>-1</sup> / 5.2
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal "G" an artificial residue due to TEV-cleavage.

<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> Rosetta2 pLysS
<b>7</b>	<b>Cultivation medium</b>
	M9 (uniformly <sup>15</sup> N, <sup>13</sup> C-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.5 mM IPTG at OD <sub>600</sub> 0.8
<b>10</b>	<b>Cultivation temperature and time</b>
	16°C for 14-16 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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.
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell lysis in buffer <b>1A</b> 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 <b>1A</b> . Protein was eluted with 5 column volumes of buffer <b>1B</b> .
B	Eluted protein was cleaved with 1% (w/w) TEV protease o.n. at rt while dialyzing the protein into 1 L buffer <b>1C</b> . 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 <b>1C</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	17 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
<b>1b</b>	<b>A260/280 ratio</b>
	0.5
<b>2</b>	<b>Stability</b>
	NMR sample stable at 4°C for a month, at 35°C for several days before degradation occurs.
<b>3</b>	<b>Comment on applicability</b>

Suitable for NMR-based screening applications.

## SI6: nsp8

Table 1: General Information

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp8
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp8
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	AIASEFSSLPSYAAFATAQEAYEQAVANGDSEVVLLKLLKSLNVAKSEFDRDAAMQRKLEKM ADQAMTQMYKQARSEDKRAKVTSAMQTMLFTMLRKLNDALNINIINNARDGCVPLNIPLTT AAKLMVVIPDYNTYKNTCDGTTFTYASALWEIQVVDADSKIVQLSEISMDNSPNLAWPLIVT ALRANSAVKLQ
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-198 (fl nsp8)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 97%; similarity: 98%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 6NUS, 6NUR, 2AHM, SCoV2: PDB 7C2K, 7BV2, 7BV1, 7CTT, 6M5I, 7BW4, 6XEZ, 7BZF, 6XQB, 6M71, 6YYT, 7BTF, 7JLT, 7AAP, 6WIQ, 6XIP, 6WQD, 6WTC, 6YHU
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pET46
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> , enterokinase
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	21.94 kDa / 19,940 cm <sup>-1</sup> M <sup>-1</sup> / 6.5
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal "G" an artificial residue due to TEV-cleavage.

<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> Rosetta2 pLysS
<b>7</b>	<b>Cultivation medium</b>
	M9 (uniformly <sup>15</sup> N-, <sup>13</sup> C-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.5 mM IPTG at OD <sub>600</sub> 0.8
<b>10</b>	<b>Cultivation temperature and time</b>
	16°C for 16-18 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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.
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
	Cell lysis in buffer <b>1A</b> 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 <b>1A</b> . Protein was eluted with 5 column volumes of buffer <b>1B</b> .
B	Eluted protein was cleaved with 1% (w/w) TEV protease o.n. at rt while dialyzing the protein into 1 L buffer <b>1C</b> . 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 <b>1D</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	17 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
<b>1b</b>	<b>A260/280 ratio</b>
	0.5
<b>2</b>	<b>Stability</b>
	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp9
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp9
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	NNELSPVALRQMSCAAGTTQACTDDNALAYYNTTKGGRFVLALLSDLQDLKWARFPKSDGT GTIYTELEPPCRFVTDTPKGPKVKYLYFIKGLNNLNRGMVLGSLAATVRLQ
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-113 (fl nsp9)
<b>5</b>	<b>Ratio for construct design (detailed and comprehensible)</b>
	In analogy to the available crystal structure (PDB 1QZ8) of nsp9 SCoV, fl sequence.
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 97%; similarity: 97%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 3EE7 (G104E), 1UW7, 1QZ8 SCoV2: PDB 6WXD, 6W4B, 6W9Q
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV: BMRB 6501 SCoV2: BMRB 50621, 50622

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pKM263 (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> -GST
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	12,7 kDa / 13,075 M <sup>-1</sup> cm <sup>-1</sup> / 9.1
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal „GAMG" four artificial residues due to TEV-cleavage and construct design
<b>6</b>	<b>Used expression strain</b>

	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	1 mM IPTG at OD <sub>600</sub> 0.7
<b>10</b>	<b>Cultivation temperature and time</b>
	20-22°C for 18-20 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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 <sub>3</sub> (SEC / final NMR buffer).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus 100 µL protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni <sup>2+</sup> -NTA (Carl Roth)), Elution with 150-500 mM imidazole in buffer <b>1A</b> .
C	Dialysis o.n. in in buffer <b>1A</b> .
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer <b>1A</b> .
E	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer <b>1B</b> . See relevant peak in attached SEC profile.
F	NMR sample preparation in buffer <b>1B</b> .

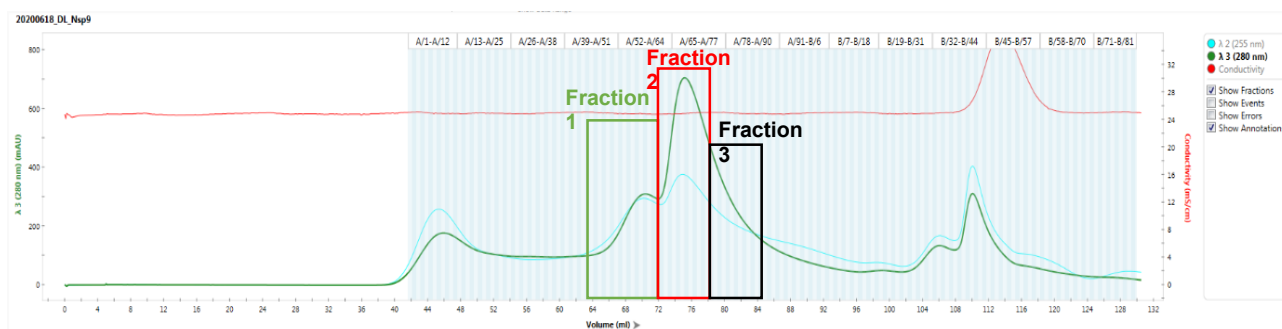
Table 4: Final sample

<b>1</b>	<b>Yield</b>
	4.5 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
<b>1b</b>	<b>A260/280 ratio</b>
	0.7
<b>2</b>	<b>Stability</b>
	Stable dimer. Storage at 4°C possible.
<b>3</b>	<b>Comment on applicability</b>
	Conditions for NMR structure determination may need to be optimized (concerning line width due to dimeric state). Backbone assignment and screening successful.

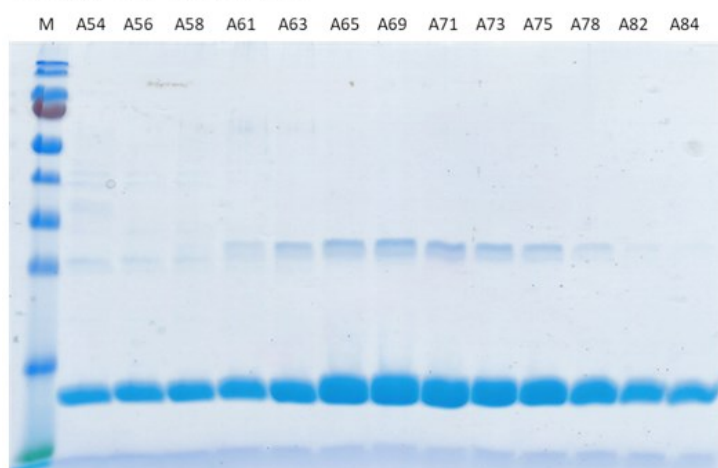


## Additional information

	Constructs	Conditions	Comments
<b>A</b>	aa 1-113 (fl nsp9); His <sub>7</sub> (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>B</b>		<p><b>IMAC buffer:</b> 25 mM NaPi (pH 7.4), 300 mM NaCl, 20 mM imidazole, 1 mM DTT.</p> <p><b>Cleavage buffer:</b> 25 mM NaPi (pH 7.4), 150 mM NaCl, 1 mM DTT.</p> <p><b>SEC/NMR buffer A:</b> 25 mM NaPi (pH 7.0), 150 mM NaCl, 1 mM DTT, 150 mM NaCl, 2 mM TCEP-HCl.</p> <p><b>SEC/NMR buffer B:</b> 25 mM NaAc (pH 5.0), 150 mM NaCl, 2 mM TCEP-HCl.</p>	3 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium. Sample in Buffer A looked degraded (from the <sup>15</sup> N HSQC) after 5 days of <sup>13</sup> C 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.



## H6-GST-TEV-Nsp9 (BL21)



**SEC profile of TEV-cleaved His<sub>6</sub>-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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1a and ORF1ab; nsp10
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp10
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	AGNATEVSPANSTVLSFCAFAVDAAKAYKDYLASGGQPITNCVKMLCTHTGTGQAITVTPEAN MDQESFGGASCCLYCRCHIDHPNPKGFCDLKGKYVQIPTTCANDPVGFTLKNVTCTVCGMWK GYGCSCDQLREPMLQ
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-139 (fl nsp10)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 97%; similarity: 99%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 5C8S, 5NFY, 2FYG, 2XYQ, 2XYV, 2XYV SCoV2: PDB 6W4H, 6W61, 7JYY, 7C2I, 7BQ7, 2G9T
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV2: BMRB 50392

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pET21b(+) (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub>
<b>3</b>	<b>Cleavage Site</b>
	-
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of protein</b>
	16.24 kDa / 12,950 M <sup>-1</sup> cm <sup>-1</sup> / 6.72
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal "MGSDKIHSHHHH" twelve artificial residues due to construct design
<b>6</b>	<b>Used expression strain</b>

	<i>E. coli</i> T7 Express
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.5 mM IPTG at OD <sub>600</sub> 0.6-0.7 (addition of 50 μM ZnCl <sub>2</sub> )
<b>10</b>	<b>Cultivation temperature and time</b>
	18-20°C for 16-18 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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)
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus one tablet of EDTA free protease inhibitor cocktail (Merck) and addition of 50 μM ZnCl <sub>2</sub> ) by microfluidization.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer <b>1A</b> .
C	SEC (HiLoad 26/600 SD 75 μg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer <b>1B</b> (elution volume 175-225 mL).
D	NMR sample preparation in buffer <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	25 mg/L <sup>15</sup> N-M9 medium, 15 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
	Stable throughout measurement (6 days, 298 K). No significant precipitation or degradation observed after storage at -80°C for 2 months.
<b>3</b>	<b>Comment on applicability</b>
	Suitable for NMR structure determination, fragment screening, interaction studies.

#### Additional information

Constructs	Conditions	Comments
aa 1-139 (fl nsp10); His <sub>6</sub> (pMCSG53 (BEI Resources, cat.	<b>IMAC-buffer:</b> 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 <sub>2</sub> , 0.1% (v/v) Triton X-100, 5-10% (v/v) glycerol, 50 mM imidazole. <b>SEC-buffer:</b> 20 mM HEPES (pH 8.5), 0.5 M NaCl, 10 mM bME, 2 mM MgCl <sub>2</sub> , 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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1ab; nsp13
<b>2</b>	<b>Region/Name/Further Specification</b>
	NTPase / helicase domain / RNA 5'-triphosphatase
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	AVGACVLCNSQTSRLRCGACIRRPFLCCKCCYDHSVISTSHKLVLSVNPYVCNAPGCDVTDVTQL YLGGMSSYYCKSHKPPISFPLCANGQVFGLYKNTCVGSDNVTDNFNAIATCDWTNAGDYILANTC TERLKLFAAETLKATEETFKLSYGIATVREVLSRELHLSWEVVGKPRPPLNRNYVFTGYRVTKN SKVQIGEYTFEKGDYGDVAVVYRGTTTYKLVNGDYFVLTSHTVMPLSAPTLVPQEHYVRITGLY PTLNISDEFSSNVANYQKVGMMQKYSTLQGGPGTGKSHFAIGLALYPSARIVYTACSHAAVDAL CEKALKYLPIDKCSRIIPARARVECFDKFKVNSTLEQYVFCTVNALPETTADIVVFDEISMATNY DLSVVNARLRAKHYYIGDPAQLPAPRTLLTKGTLEPEYFNSVCRMLKTIGPDMFLGTCRRCPA EIVDVTVSALVYDNKLLKAHKDKSAQCFKMFYKGVITHDVSSAINRPQIGVVREFLTRNPAWRKA VFISPYNSQNAVASKILGLPTQTVDSSQGSEYDYVIFTQTTETAHSCNVNRFNVAITRAKVGILCI MSDRDLYDKLQFTSLEIPRRNVATLQ
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	1-601 aa (fl nsp13)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 99.8%; similarity: 100%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV2: PDB 6ZSL, 6JYT, 6XEZ
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pE-SUMO (LifeSensors)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> -SUMO
<b>3</b>	<b>Cleavage Site</b>
	Ulp1
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>

	66.85 kDa / 67,160 M <sup>-1</sup> cm <sup>-1</sup> / 8.66
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	No artificial residues due to Ulp1-cleavage and construct design.
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.2 mM IPTG at OD <sub>600</sub> 0.6-0.7 (addition of 50 μM ZnCl <sub>2</sub> )
<b>10</b>	<b>Cultivation temperature and time</b>
	18-20°C for 16-18 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus one tablet of EDTA free protease inhibitor cocktail (Merck) and addition of 50 μM ZnCl <sub>2</sub> ) by microfluidization.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer <b>1A</b> .
C	SEC (HiLoad 26/600 SD 200 pg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer <b>1B</b> (elution volume 210-240 mL).
D	NMR sample preparation in buffer <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.5 mg/L <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
	Aggregation at > 20 μM under these conditions.
<b>3</b>	<b>Comment on applicability</b>
	Not suitable for NMR experiments.

#### Additional information

<b>Constructs</b>	<b>Conditions</b>	<b>Comments</b>
aa 1-601 (fl nsp13); His <sub>6</sub> (pET28a+) (GenScript), TEV-cleavage site, N-terminal "GHM" three artificial residues.	Native (as above)	Weak expression, instable protein.

## SI10: nsp14

Table 1: General Information

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1ab; nsp14
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp14 / 3'-to-5' exonuclease / guanine N7-methyltransferase (MTase)
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	AENVTLGLFKDCSKVITGLHPTQAPTHLSVDTKFKTEGLCVDIPGIPKDMTYRRLISMMGFKMNY QVNGYPNMFITREEAIRHVRAWIGFDVEGCHATREAVGTNLPLQLGFSTGVNLVAVPTGYVDT PNNTDFSRVSAKPPPG
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
<b>fl</b>	aa 1-527 (fl nsp14)
<b>MTase</b>	aa 288-527 (MTase domain)
<b>5</b>	<b>Ratio for construct design</b>
<b>fl</b>	fl protein
<b>MTase</b>	In analogy to SCoV structure (PDB 5C8U)
<b>6</b>	<b>Sequence homology (to SCoV)</b>
<b>fl</b>	Identity: 95%; similarity: 99%
<b>MTase</b>	Identity: 95%, similarity: 97%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 5C8U, 5C8S, 5C8T, 5NFY
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
<b>fl</b>	pRSF-Duet1 (Novagen)
<b>MTase</b>	pET28a (Novagen)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
<b>fl</b>	N-terminal His <sub>6</sub>
<b>MTase</b>	N-terminal His <sub>6</sub>
<b>3</b>	<b>Cleavage Site</b>
<b>fl</b>	TEV



<b>MTase</b>	Thrombin
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
<b>fl</b>	60.01 kDa / 91,660 M <sup>-1</sup> cm <sup>-1</sup> / 7.79
<b>MTase</b>	27.82 kDa / 48,970 M <sup>-1</sup> cm <sup>-1</sup> / 7.19
<b>5</b>	<b>Comments on sequence of expressed construct</b>
<b>fl</b>	N-terminal “GSM” three artificial residues due to construct design.
<b>MTase</b>	N-terminal “GSHM” four artificial residues due to construct design.
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	2xTY for protein production, LB for transformation and maintenance
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	1 mM IPTG at OD <sub>540</sub> 0.5-0.6
<b>10</b>	<b>Cultivation temperature and time</b>
	20°C for 18-20 h

Table 3: Protein Purification (fl nsp14 and nsp14 MTase)

<b>1</b>	<b>Buffer List</b>
A	50 mM Tris-HCl (pH 9.0), 0.5 M NaCl, 10 mM bME, 2 mM MgCl <sub>2</sub> , 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 <sub>2</sub> , 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 <sub>2</sub> , 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 <sub>2</sub> , 5% (v/v) glycerol, 20 mM imidazole (SEC).
E	20 mM potassium phosphate (pH 8.0), 0.25 M KCl (Screening).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> by sonication in pulse mode (0.5 s on /0.5 s off) for 10 min.
B	IMAC (gravity flow or batch Ni <sup>2+</sup> -NTA) (GE Healthcare), washing with buffer <b>1B</b> , elution with <b>1C</b> .
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- <b>MTase</b>	[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 <b>1D</b> (elution volume 75-95 mL).
E-fl	[Optional] Separation of TEV protease and uncleaved nsp14 material with IMAC, collection of flow through in buffer <b>1D</b> .
E- <b>MTase</b>	[Optional] Separation of thrombin protease and uncleaved MTase nsp14 material with IMAC, collection of flow through in buffer <b>1D</b> .
F	For fragment screening the buffer is exchanged to <b>1E</b> .
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

<b>1</b>	<b>Yield</b>
fl	6 mg/L 2xTY medium
<b>MTase</b>	~ 10 mg/L 2xTY medium
<b>1b</b>	<b>A260/280 ratio</b>
fl	0.6
<b>MTase</b>	0.6
<b>2</b>	<b>Stability</b>
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.
<b>MTase</b>	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.
<b>3</b>	<b>Comment on applicability</b>
	Suitable for fragment screening and enzymatic activity assays.

#### Additional information

Constructs	Conditions	Comments
Fl nsp14; His <sub>6</sub> (pETDuet (GenScript)), no cleavage site, N-terminal "MGSSHHHHHSQDP" 14 artificial residues.	<b>IMAC-buffer:</b> 25 mM Tris/HCl (pH 8.5), 300 mM NaCl, 5 mM imidazole, 10 mM bME, 5% (v/v) glycerol. <b>SEC-buffer:</b> 25 mM Tris/HCl (pH 8.5), 300 mM NaCl, 5 mM DTT, 5% (v/v) glycerol	Yields 14 mg/L <sup>15</sup> N-M9 medium. Tendency to aggregate.

# SI11: nsp15

Table 1: General Information

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1ab; nsp15
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp15 / NendoU / Endonuclease
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	<p>           SLENVAFNVVVKGHFDGQQGEVPVSIINNTVYTKVDGVDVELFENKTTLPVNVAFELWAKRNI            KPVPEVKILNNLGVDIAANTVIWDYKRDAPAHISTIGVCSMTDIAKKPTETICAPLTVFFDGRVD            GQVDLFRNARNGVLITEGSVKGLQPSVGPQASLNGVTLIGEAVKTQFNYYKKVDGVVQQLPE            TYFTQSRNLQEFKPRSQMEIDFLELAMDEFIERYKLEGYAFEHIVYGDFSHSQLGGLHLLIGLAK            RFKESPFLEDFIPMDSTVKNYFITDAQTGSSKCVCSVIDLLLDDFVEIIKSQDLSVVSQVVKVTI            DYTEISFMLWCKDGHVETFYPKLQ         </p>
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-346 (fl nsp15)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 89%; similarity: 98%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 2H85 SCoV2: PDB 6W01
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pET28a(+) (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub>
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	39.14 kDa / 32,890 M <sup>-1</sup> cm <sup>-1</sup> / 5.12
<b>5</b>	<b>Comments on sequence of expressed construct</b>

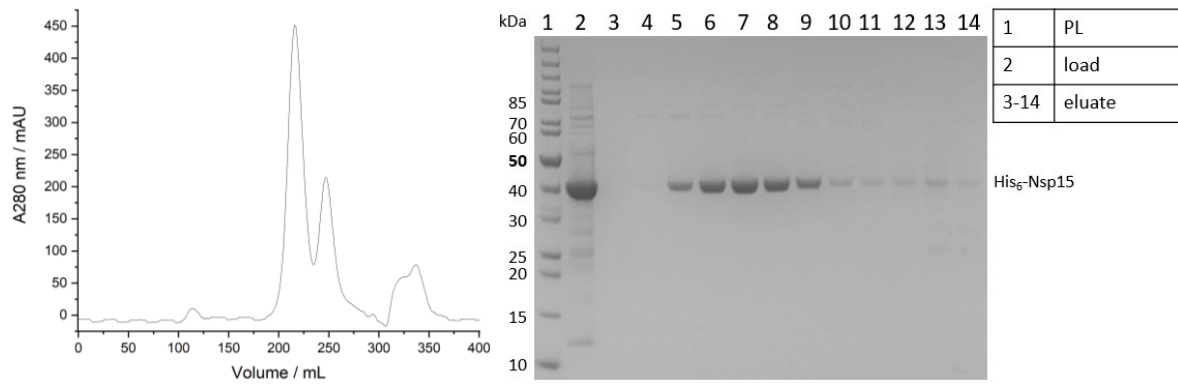
	N-terminal “GHM” three artificial residues due to TEV-cleavage and construct design
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.2 mM IPTG at OD <sub>600</sub> 0.6-0.7
<b>10</b>	<b>Cultivation temperature and time</b>
	18-20°C for 16-18 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (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 <b>1A</b> .
C	SEC (HiLoad 26/600 SD 200 µg (GE Healthcare), ÄKTApurifier (GE Healthcare)) in buffer <b>1B</b> (elution volume 200-260 mL).
D	NMR sample preparation in buffer <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	5 mg/L <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
	Tendency to aggregate at rt.
<b>3</b>	<b>Comment on applicability</b>
	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF1ab; nsp16
<b>2</b>	<b>Region/Name/Further Specification</b>
	nsp16 / 2'-O-ribose methyltransferase (2'-O-MTase)
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	SSQAWQPGVAMPNLYKMQRMLLEKCDLQNYGDSATLPKGIMMNVAKYTQLCQYLNTLTLAV PYNMRVIHFGAGSDKGVAPGTAVLRQWLPTGTLVSDLDLNDVSDADSTLIGDCATVHTANK WDLIISDMYDPKTKNVTKENDSKEGFFTYICGFIQKALGGSVAIKITEHSWNADLYKLMGHF AWWTAFVTNVNASSSEAFLLGICNYLGGKPREQIDGYVMHANYIFWRNTNPIQLSSYSLFDMSKFP LKLRTAVMSLKEGQINDMILSLLSKGRLIIRENNRVVVISSDVLVNN
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-298 (fl nsp16)
<b>5</b>	<b>Ratio for construct design</b>
	Based on fl annotation boundaries of YP_009725311.1 protein entry in NC_045512.2.
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 93%; similarity: 99%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 3R24, 2XYR, 2XYQ SCoV2: PDB 7JYY, 6W4H, 6YZ1, 7BQ7, 7C2I, 6W6I
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pRSF-Duet1 (Novagen)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub>
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	33.67 kDa / 55,790 M <sup>-1</sup> cm <sup>-1</sup> / 7.76
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal „GSMA" - four artificial residues due to TEV-cleavage and construct design.

<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21(DE3)
<b>7</b>	<b>Cultivation medium</b>
	2xTY
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	1 mM IPTG at OD <sub>540</sub> 0.5-0.6
<b>10</b>	<b>Cultivation temperature and time</b>
	20°C for 18-20 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
A	50 mM Tris-HCl (pH 9.0), 500 mM NaCl, 10 mM bME, 2 mM MgCl <sub>2</sub> , 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 <sub>2</sub> , 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 <sub>2</sub> , 5% (v/v) glycerol, 1 M imidazole (IMAC).
D	20 mM HEPES (pH 8.5), 500 mM NaCl, 10 mM bME, 2 mM MgCl <sub>2</sub> , 5% (v/v) glycerol, 20 mM imidazole (SEC).
E	20 mM KPi (pH 8.0), 200 mM KCl, 1 mM MgCl <sub>2</sub> , 2 mM DTT (Screening).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> by sonication in pulse mode (0.5 s on /0.5 s off) for 10 min.
B	IMAC (gravity flow or batch Ni <sup>2+</sup> -NTA) (GE Healthcare), washing with buffer <b>1B</b> , elution with <b>1C</b> .
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 <b>1D</b> (elution volume 90-100 mL).
E	[Optional] Separation of TEV protease and uncleaved nsp16 material with IMAC, collection of flow through in buffer <b>1D</b> .
F	nsp10 is added at 1:1 molar ratio – necessary for stability and activity.
G	For fragment screening the buffer is exchanged to <b>1E</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	~ 10-15 mg/L 2xTY medium.

<b>1b</b>	<b>A260/280 ratio</b>
	0.55
<b>2</b>	<b>Stability</b>
	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.
<b>3</b>	<b>Comment on applicability</b>
	Suitable for fragment screening.

#### Additional information

<b>Constructs</b>	<b>Conditions</b>	<b>Comments</b>
Fl nsp16; His <sub>6</sub> (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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF3a
<b>2</b>	<b>Region/Name/Further Specification</b>
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFQSASKIITLK KRWQLALSKGVHFCNLLLLFVTVYSHLLLVAAGLEAPFLYLYALVYFLQSINFVRIIMRLWLC WKCRSKNPLLYDANYFLCWHTNCYDYCIPYNSVTSSIVITSGDGTTSPISEHDYQIGGYTEKWE SGVKDCVVLHSYFTSDYYQLYSTQLSTDTGVEHVTFEYFNKIVDEPEEHVQIHTIDGSSGVVNPV MEPIYDEPTTTTSVPL
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-275 (fl ORF3a)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 72.4%; similarity: 90.2%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV2: PDB 6XDC
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Cell-free Protein Synthesis

<b>1</b>	<b>Expression vector</b>
	pEU-E01-MCS (Cell-Free Sciences)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	C-terminal Strep tag II (WSHPQFEK)
<b>3</b>	<b>Cleavage Site</b>
	-
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of protein</b>
	32.32 kDa / 64,205 M <sup>-1</sup> cm <sup>-1</sup> / 5.67
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.

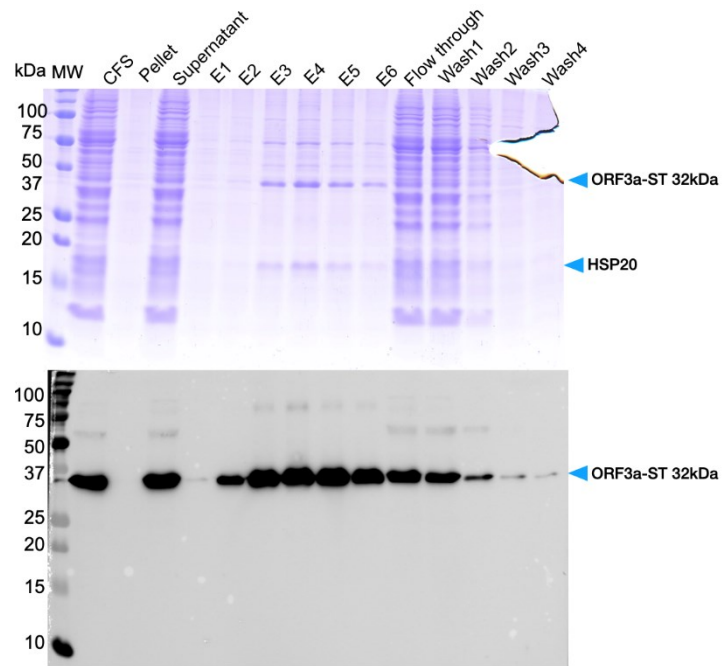
<b>6</b>	<b>Feeding buffer</b>
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix and 0.05% (w/v) Brij-58.
<b>7</b>	<b>Translation mix</b>
	50% (v/v) 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.
<b>8</b>	<b>Protein synthesis temperature and time</b>
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
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 <b>1A</b> (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of <b>1A</b> .
H	Elute the protein of interest with <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.6 mg/mL WGE
<b>1b</b>	<b>A260/280 ratio</b>
	1.08
<b>2</b>	<b>Stability</b>
	Stable at 4°C for at least 2 weeks.
<b>3</b>	<b>Comment on applicability</b>
	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF 4; Envelope (E) protein
<b>2</b>	<b>Region/Name/Further Specification</b>
	E protein
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MYSFVSEETGTLIVNSVLLFLAFVVLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYSRVKLNSSRVPDLLV
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-75 (fl ORF4)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 94.7%; similarity: 97.4%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 5X29
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV: BMRB 36049

Table 2: Cell-free Protein Synthesis

<b>1</b>	<b>Expression vector</b>
	pEU-E01-MCS (Cell-free Sciences)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	C-terminal Strep tag II (WSHPQFEK)
<b>3</b>	<b>Cleavage Site</b>
	-
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of protein</b>
	9.56 kDa / 11,460 M <sup>-1</sup> cm <sup>-1</sup> / 8.55
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.
<b>6</b>	<b>Feeding buffer</b>

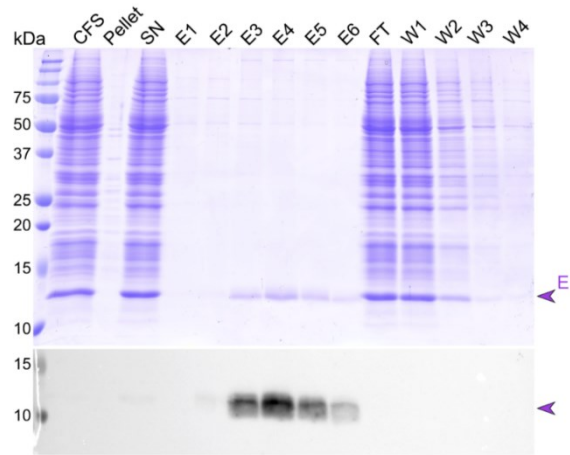
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix and 0.05% (w/v) Brij-58.
<b>7</b>	<b>Translation mix</b>
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix 0.05% (w/v) Brij-58.
<b>8</b>	<b>Protein synthesis temperature and time</b>
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
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 <b>1A</b> (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of <b>1A</b> .
H	Elute the protein of interest with <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.45 mg/mL WGE
<b>1b</b>	<b>A260/280 ratio</b>
	1.52
<b>2</b>	<b>Stability</b>
	Stable at least a few days at rt.
<b>3</b>	<b>Comment on applicability</b>
	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF5; Membrane glycoprotein (M)
<b>2</b>	<b>Region/Name/Further Specification</b>
	M protein
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MADSNGTITVEELKLLLEQWNLVIGFLFTWICLLQFAYANRNRFLYIIKLIFLWLLWPVTLACF VLAAYRINWITGGIAIAMAACLVGLMWLSYFIASFRLFARTRSMWSFNPETNILLNVPLHGTTILT RPLLESELVIGAVILRGHLRIAGHHLGRCDIKDLPKEITVATSRTLSYYKLGASQRVAGDSGFAA YSRYRIGNYKLNTDHSSSDNIALLVQ
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-222 (fl ORF5)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 90.5%; similarity: 98.2%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	-
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Cell-free Protein Synthesis

<b>1</b>	<b>Expression vector</b>
	pEU-E01-MCS (Cell-Free Sciences)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	C-terminal Strep tag II (WSHPQFEK)
<b>3</b>	<b>Cleavage Site</b>
	-
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of protein</b>
	26.35 kDa / 57,660 M <sup>-1</sup> cm <sup>-1</sup> / 9.48
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	C-terminal “SAWSHPQFEK” ten artificial residues due to construct design.

<b>6</b>	<b>Feeding buffer</b>
	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.
<b>7</b>	<b>Translation mix</b>
	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.
<b>8</b>	<b>Protein synthesis temperature and time</b>
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
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 <b>1A</b> (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of <b>1A</b> .
H	Elute the protein of interest with <b>1B</b> .

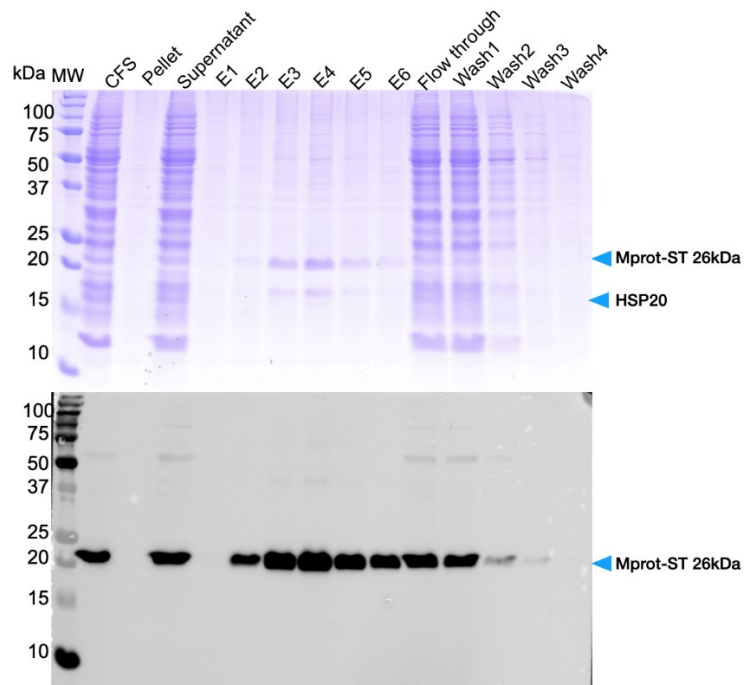
Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.33 mg/mL WGE
<b>1b</b>	<b>A260/280 ratio</b>
	1.16
<b>2</b>	<b>Stability</b>
	Stable at 4°C for at least 2 weeks.
<b>3</b>	<b>Comment on applicability</b>
	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 $\mu$ 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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF6
<b>2</b>	<b>Region/Name/Further Specification</b>
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MFHLVDFQVTIAEILLIMRTFKVSIWNLDYIINLIKNLSKSLTENKYSQLDEEQPMEID
<b>4</b>	<b>Protein boundaries - amino acid numbering (according to NCBI Reference Sequence NC_045512.2):</b>
	aa 1-61 (fl ORF6)
<b>5</b>	<b>Ratio for construct design (detailed and comprehensible)</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 68.9%; similarity: 93.4%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	-
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Cell-free Protein Synthesis

<b>1</b>	<b>Expression vector</b>
	pEU-E01-MCS (Cell-Free Sciences)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	C-terminal Strep tag II (WSHPQFEK)
<b>3</b>	<b>Cleavage Site</b>
	-
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	8470.85 kDa / 13,980 M <sup>-1</sup> cm <sup>-1</sup> / 4.89
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.
<b>6</b>	<b>Feeding buffer</b>

	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
<b>7</b>	<b>Translation mix</b>
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix
<b>8</b>	<b>Protein synthesis temperature and time</b>
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

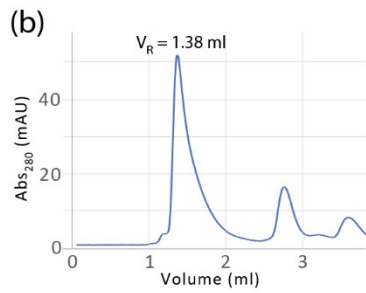
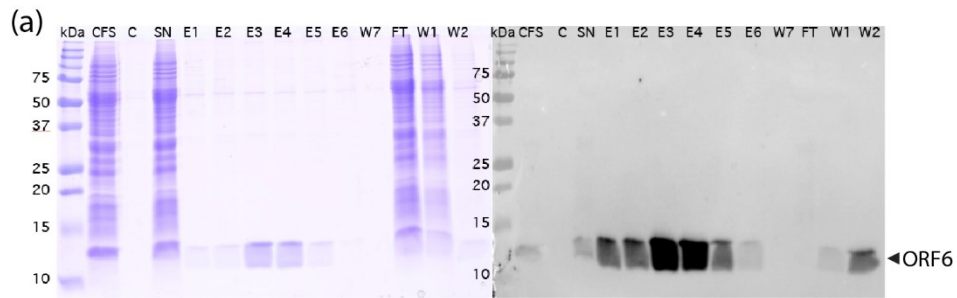
<b>1</b>	<b>Buffer List</b>
A	20 mM NaPi (pH 6.5), 50 mM NaCl (wash buffer).
B	20 mM NaPi (pH 6.5), 50 mM NaCl, 2.5 mM desthiobiotin (elution buffer).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Harvest total CFS.
B	Incubate with benzonase for 30 min on a 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 <b>1A</b> (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of <b>1A</b> .
H	Elute the protein of interest with <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.27 mg/mL of WGE and total production of 875 µg for NMR samples
<b>1b</b>	<b>A260/280 ratio</b>
	1.36
<b>2</b>	<b>Stability</b>
	stable
<b>3</b>	<b>Comment on applicability</b>
	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF7a
<b>2</b>	<b>Region/Name/Further Specification</b>
	Ectodomain (ED)
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MKIIILFLALITLATCELYHYQECVRGTTVLLKEPCSSGTYEGNSPFHPLADNKFALTCFSTQFAFA CPDGVKHHVYQLRARSVSPKLFIRQEEVQELYSPIFLIVAAIVFITLCFTLKRKTE
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 16-81 (ectodomain of ORF7a)
<b>5</b>	<b>Ratio for construct design (detailed and comprehensible)</b>
	Only the ectodomain without signaling peptide. Transmembrane helix is also not included in the construct.
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 85.3%; similarity: 95.9%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV: PDB 1XAK, 1YO4
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	SCoV: BMRB 6824

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pET24d-GB1 (Novagen, modified by G. Stier (Bogomolovas et al., 2009))
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> -GB1
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	7.49 kDa / 6,210 M <sup>-1</sup> cm <sup>-1</sup> / 6.99
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal „G" one artificial residue due to TEV-cleavage and construct design.
<b>6</b>	<b>Used expression strain</b>

	<i>E.coli</i> (DE3) BL21
<b>7</b>	<b>Cultivation medium</b>
	M9 (uniformly <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.2 mM IPTG at OD <sub>600</sub> 0.7
<b>10</b>	<b>Cultivation temperature and time</b>
	25°C for 18-20 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption and solubilization of pellet in buffer <b>1A</b> .
B	IMAC, gravity flow Ni <sup>2+</sup> -NTA (Qiagen), elution with 200 mM imidazole in buffer <b>1B</b> .
C	Dialysis against buffer <b>1C</b> .
D	TEV-cleavage (1 mg TEV protease per 10 mL protein solution) o.n. in buffer <b>1C</b> .
E	Inv. IMAC, elution with 200 mM imidazole in buffer <b>1C</b> .
F	Dialysis of flow-through of inv. IMAC against <b>1D</b> and concentrate (NMR-sample).

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.4 mg/L <sup>15</sup> N-M9 medium
<b>1b</b>	<b>A260/280 ratio</b>
	0.7
<b>2</b>	<b>Stability</b>
	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF7b
<b>2</b>	<b>Region/Name/Further Specification</b>
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MIELSLIDFY LCFLAFLFL VLIMLIIFWF SLELQDHNET CHA
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-43 (fl ORF7b)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 85.4%; similarity: 97.2%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	-
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

## Bacterial

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pThioRed (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> -Trx
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	5.37 kDa / 6,990 M <sup>-1</sup> cm <sup>-1</sup> / 4.17
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal "GA(M)G" three artificial residues due to TEV-cleavage and construct design.
<b>6</b>	<b>Used expression strain</b>



	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.2 mM IPTG at OD <sub>600</sub> 0.7
<b>10</b>	<b>Cultivation temperature and time</b>
	18-20°C for 16-18 h

Table 3: Protein Purification with detergent

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus one tablet of EDTA free protease inhibitor cocktail (Merck)) by microfluidization.
B	Solubilization of pellet after lysis <b>1B</b> (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 <b>1C</b> .
D	TEV-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer <b>1C</b>
E	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with 500 mM imidazole in buffer <b>1C</b> .
F	Rebuffer flow-through of inv. IMAC in buffer <b>1D</b> (NMR sample).
G	Analytical SEC (SD 75 Increase 10/300 GL (GE Healthcare), ÄKTA start (GE Healthcare)) in buffer <b>1D</b> .

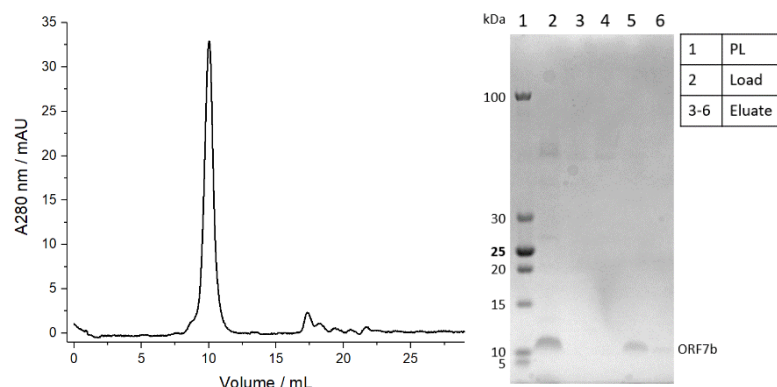
Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.6 mg/L <sup>15</sup> N-M9 medium
<b>2</b>	<b>Stability</b>
	Stable throughout measurement (2 days, 283/298 K). No significant precipitation or degradation observed after storage at 4°C for 3 months.

<b>3</b>	<b>Comment on applicability</b>
	Due to necessity of solubilizing agent and tendency to oligomerize structure determination, fragment screening, and interaction studies are hindered.

#### Additional information

	<b>Constructs</b>	<b>Conditions</b>	<b>Comments</b>
<b>A</b>	As above	Native <b>IMAC buffer:</b> 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME. <b>SEC buffer:</b> 25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl	Nearly no protein was extracted in soluble fraction.
<b>B</b>		Denaturing <b>Solubilizing buffer:</b> 25 mM Tris-HCl (pH 8.0), 6 M GdnHCl, 300 mM NaCl, 5 mM imidazole. <b>IMAC wash buffer:</b> 25 mM Tris-HCl (pH 8.0), 8 M urea, 300 mM NaCl, 5 mM imidazole. <b>Renaturing buffer:</b> 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME. <b>IMAC elution buffer:</b> 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 500 mM imidazole, 10 mM bME.	After refolding and cleavage degradation of protein.
<b>C</b>	F1 ORF7b; His <sub>6</sub> -SUMO (pE-SUMO (GenScript)), Ulp1-cleavage site, no artificial residues.	Native <b>IMAC buffer:</b> as above <b>SEC buffer:</b> 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.
<b>D</b>		Detergent <b>IMAC buffer:</b> 50 mM NaPi (pH 7.0), 200 mM NaCl, 0.1% (v/v) Triton X-100, 5 mM imidazole, 10 mM bME. <b>SEC buffer:</b> 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.
<b>E</b>		Semi-denaturing <b>IMAC buffer:</b> 50 mM Tris-HCl (pH 8.0), 2 M urea, 300 mM NaCl, 10 mM imidazole, 10 mM bME. <b>SEC buffer:</b> 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

<b>1</b>	<b>Expression vector</b>
	pEU-E01-MCS (Cell-Free Sciences)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	C-terminal Strep tag II (WSHPQFEK)
<b>3</b>	<b>Cleavage Site</b>
	-
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of protein</b>
	5.37 kDa / 6,990 M <sup>-1</sup> cm <sup>-1</sup> / 4.17
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	C-terminal “SAWSHPQFEK” ten artificial residues due to construct design.
<b>6</b>	<b>Feeding buffer</b>
	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.
<b>7</b>	<b>Translation mix</b>
	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.
<b>8</b>	<b>Protein synthesis temperature and time</b>
	22°C for 16 h without agitation (bilayer method).

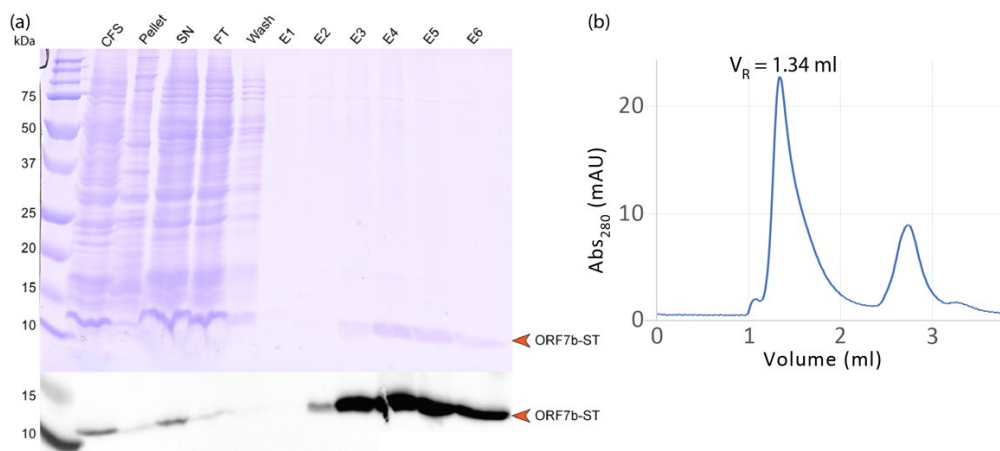
Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
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 <b>1A</b> (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of <b>1A</b> .
H	Elute the protein of interest with <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.27 mg/mL of WGE and total production of 880 µg for NMR samples
<b>1b</b>	<b>A260/280 ratio</b>
	1.36
<b>2</b>	<b>Stability</b>
	Stable in detergent over several days.
<b>3</b>	<b>Comment on applicability</b>
	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF8
<b>2</b>	<b>Region/Name/Further Specification</b>
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MKFLVFLGIITTVAAFHQECSLQSQCTQHQPYYVDDPCPIHFYSKWKYIRVVGARKSAPLIELCVDEA GSKSPIQYIDIGNYTVSCSPFTINCQEPKLGSLVVRCSFYEDFLEYHDVVRVLDLFI
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
<b>ORF8</b>	aa 1-121 (fl ORF8 = ORF8)
<b>ORF8<sub>m</sub></b>	aa 1-121 (fl ORF8) with L84S mutation (~ isolate 2019-nCoV_HKU-SZ-002a_2020).
<b>ΔORF8</b>	aa 16-121 (without signal peptide = ΔORF8)
<b>5</b>	<b>Ratio for construct design (detailed and comprehensible)</b>
<b>ORF8</b>	fl protein
<b>ΔORF8</b>	Protein after the hypothetical cleavage of the N-terminal Signal Peptide
<b>6</b>	<b>Sequence homology (to SCoV)</b>
<b>ORF8</b>	Identity: 31.7%; similarity: 70.7%
<b>ΔORF8</b>	Identity: 40.5%; similarity: 66.7%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV2: 7JTL, 7JX6
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

## Bacterial

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
<b>ORF8<sub>m</sub></b>	pPK1154 (GenScript)
<b>ΔORF8</b>	pET22b (+) (Merck/Novagen)
<b>2</b>	<b>Purification-/Solubility-Tag</b>

<b>ORF8<sub>m</sub></b>	N-terminal His <sub>6</sub> -SUMO
<b>ΔORF<sub>8</sub></b>	N-terminal His <sub>6</sub> -GST
<b>3</b>	<b>Cleavage Site</b>
<b>ORF8<sub>m</sub></b>	Ulp1
<b>ΔORF<sub>8</sub></b>	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
<b>ORF8<sub>m</sub></b>	13.80 kDa / 15,930 M <sup>-1</sup> cm <sup>-1</sup> / 5.42
<b>ΔORF<sub>8</sub></b>	12.54 kDa / 15,930 M <sup>-1</sup> cm <sup>-1</sup> / 5.15
<b>5</b>	<b>Comments on sequence of expressed construct</b>
<b>ORF8<sub>m</sub></b>	No artificial residues due to Ulp1-cleavage and construct design.
<b>ΔORF<sub>8</sub></b>	N-terminal “GAMG” three artificial residues due to TEV-cleavage and construct design.
<b>6</b>	<b>Used expression strain</b>
<b>ORF8<sub>m</sub></b>	<i>E. coli</i> BL21 (DE3)
<b>ΔORF<sub>8</sub></b>	<i>E. coli</i> BL21 (DE3) pLysS
<b>7</b>	<b>Cultivation medium</b>
<b>ORF8<sub>m</sub></b>	LB / M9 (uniformly <sup>15</sup> N-labelled)
<b>ΔORF<sub>8</sub></b>	LB
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
<b>ORF8<sub>m</sub></b>	0.5 mM IPTG at OD <sub>600</sub> 0.6
<b>ΔORF<sub>8</sub></b>	0.5 mM IPTG at OD <sub>600</sub> 0.6-0.7
<b>10</b>	<b>Cultivation temperature and time</b>
<b>ORF8<sub>m</sub></b>	16-20°C for 16-18 h
<b>ΔORF<sub>8</sub></b>	18°C for 16-18 h

Table 3a: Protein Purification (ORF8)

<b>1</b>	<b>Buffer List</b>
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.
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (plus one tablet of EDTA free protease inhibitor cocktail (Merck)) by French-press.
B	Solubilization of pellet after lysis <b>1B</b> (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 <b>1C</b> .
D	Ulp1-cleavage (Protein/Ulp1 ratio 10:1) o.n. at 21°C in buffer <b>1D</b> .
E	Rebuffer in buffer <b>1E</b> .

Table 3b: Protein Purification ( $\Delta$ ORF8)

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> (supplemented with 0.5 mg/mL lysozyme, 10 $\mu$ g/mL DNaseI, 5 mM MgCl <sub>2</sub> , 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 periplasmatic 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 $\mu$ m syringe filters.
B	IMAC (HisTrap FF Crude (GE Healthcare), ÄKTA Pure 25 M1 (GE Healthcare)), binding with buffer <b>1A</b> supplemented with 50 mM imidazole, elution with imidazole gradient up to 500 mM in buffer <b>1A</b> .
C	TEV-cleavage (Protein/TEV ratio 1:10) at 4°C, o.n. in buffer <b>1B</b> .
D	Inv. IMAC (HisTrap FF Crude (GE Healthcare), ÄKTA Pure 25 M1 (GE Healthcare)), binding with buffer <b>1A</b> supplemented with 50 mM imidazole, elution with imidazole gradient up to 500 mM in buffer <b>1A</b> .
E	SEC on Increase 10/300 S75 (GE Healthcare) at 4°C in buffer <b>1C</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
<b>ORF8 m</b>	<0.5 mg/L LB mg/mL <sup>15</sup> N-M9 medium

$\Delta$ ORF8	0.5 mg/L LB medium
<b>2</b>	<b>Stability</b>
ORF8m	Not determined.
$\Delta$ ORF8	No significant precipitation or degradation observed after storage at 4°C for 1 week.
<b>3</b>	<b>Comment on applicability</b>
ORF8m	Weak expression into soluble fraction, 30%/70% soluble/inclusion bodies. After purification extremely low yield for NMR studies.
$\Delta$ 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 <sub>6</sub> (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

<b>1</b>	<b>Expression vector</b>
	pEU-E01-MCS (Cell-Free Sciences)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
ORF8	C-terminal Strep tag II (WSHPQFEK)
$\Delta$ ORF8	N-terminal Strep tag II (WSHPQFEK)
<b>3</b>	<b>Cleavage Site</b>
	-
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
ORF8	15.00 kDa / 21,805 M <sup>-1</sup> cm <sup>-1</sup> / 5.64
$\Delta$ ORF8	13.53 Da / 21,805 M <sup>-1</sup> cm <sup>-1</sup> / 5.39
<b>5</b>	<b>Comments on sequence of expressed construct</b>
ORF8	C-terminal “SAWSHPQFEK” ten artificial residues due to construct design.
$\Delta$ ORF8	N-terminal “M” and C-terminal “SAWSHPQFEK” eleven artificial residues due to construct design.



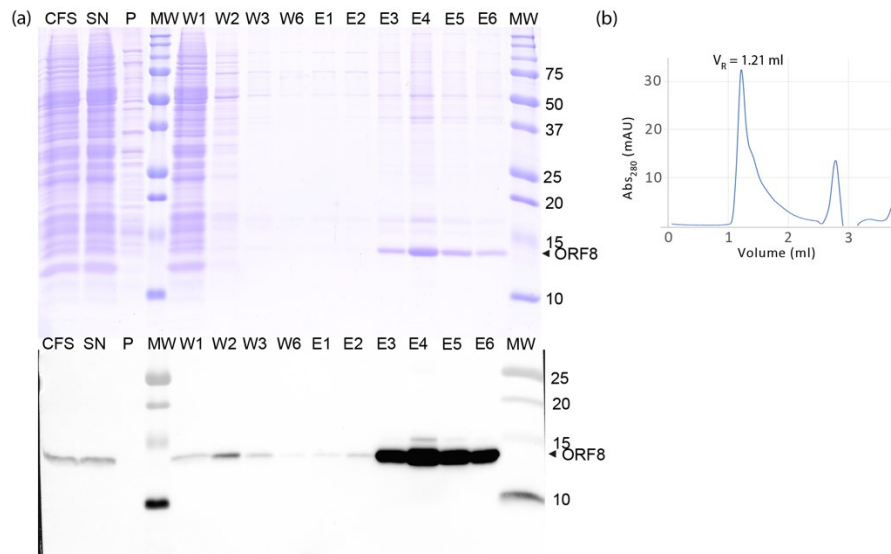
<b>6</b>	<b>Feeding buffer</b>
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix and 0.05% (w/v) Brij-58.
<b>7</b>	<b>Translation mix</b>
	50% (v/v) 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.
<b>8</b>	<b>Protein synthesis temperature and time</b>
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification (ORF8a and ORF8b)

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
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 <b>1A</b> (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of <b>1A</b> .
H	Elute the protein of interest with <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.62 mg/mL WGE after purification. Total of 683 µg for the NMR samples
<b>1b</b>	<b>A260/280 ratio</b>
	0.7
<b>2</b>	<b>Stability</b>
	Stable at 4°C for weeks.
<b>3</b>	<b>Comment on applicability</b>
	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF9a; Nucleocapsid (N) phosphoprotein
<b>2</b>	<b>Region/Name/Further Specification</b>
	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)
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MSDNGPQNQRNAPRITFGGPSDSTGNSQNGERSGARSKQRRPQGLPNNTASWFTALTQHGKED LKFPRGQGVPINTNSSPDDQIGYYRRATRIRGGDGKMKDLSRWYFYLLGTGPEAGLPYGAN KDGIIWVATEGALNTPKDHIGTRNPANNAIVLQLPQGTTLPKGFYAEGSRGGGQASSRSSRSR NSSRNSTPGSSRGTSPARMAGNGGDAALALLLDRLNQLESKMSGKGQQQGGQTVTKKSAE ASKKPRQKRTATKAYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKHWPQIAQFAPSASAFFG MSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTPEPKKDKKKKADET QALPQRQKKQQTVLLPAADLDDFSKQLQQSMSSADSTQA
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
<b>IDR1- NTD- IDR2</b>	aa 1-248 (of fl ORF9a)
<b>NTD- SR</b>	aa 44-212 (of fl ORF9a)
<b>NTD</b>	aa 44-180 (of fl ORF9a)
<b>CTD</b>	aa 247-364 (of fl ORF9a)
<b>5</b>	<b>Ratio for construct design (detailed and comprehensible)</b>
<b>IDR1- NTD- IDR2</b>	Based on boundaries from SCoV homolog.
<b>NTD- SR</b>	In analogy to the available NMR (PDB 6YI3) and crystal (6M3M) structures of N-NTD SCoV2.
<b>NTD</b>	In analogy to the available NMR (PDB 6YI3) and crystal (6M3M) structures of N-NTD SCoV2.
<b>CTD</b>	In analogy to the available NMR structure (PDB 2JW8) of N-CTD from SCoV.
<b>6</b>	<b>Sequence homology (to SCoV)</b>
<b>IDR1- NTD- IDR2</b>	Identity: 90%; similarity: 94%
<b>NTD- SR</b>	Identity: 92%; similarity: 96%
<b>NTD</b>	Identity: 93%; similarity: 97%
<b>CTD</b>	Identity: 96%; similarity: 98%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>

	SCoV: PDB 2JW8, 2CJR SCoV2: PDB 6YI3, 6M3M, 6VYO, 6WKP, 6WZO, 6WJI, 6YUN, 6ZCO, 7CE0, 7C22
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	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

<b>1</b>	<b>Expression vector</b>
<b>IDR1- NTD- IDR2</b>	pET29a(+) (Twistbioscience)
<b>NTD- SR</b>	pET-28a(+) (GenScript)
<b>NTD</b>	pET-28a(+) (GenScript)
<b>CTD</b>	pKM263 (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
<b>IDR1- NTD- IDR2</b>	-
<b>NTD- SR</b>	N-terminal His <sub>6</sub>
<b>NTD</b>	N-terminal His <sub>6</sub>
<b>CTD</b>	N-terminal His <sub>6</sub> -GST
<b>3</b>	<b>Cleavage Site</b>
<b>IDR1- NTD- IDR2</b>	-
<b>NTD- SR</b>	TEV
<b>NTD</b>	TEV
<b>CTD</b>	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
<b>IDR1- NTD- IDR2</b>	26.52 kDa / 26,930 M <sup>-1</sup> cm <sup>-1</sup> / 10.57
<b>NTD- SR</b>	18.10 kDa / 26,930 M <sup>-1</sup> cm <sup>-1</sup> / 10.35
<b>NTD</b>	14.85 kDa / 26,930 M <sup>-1</sup> cm <sup>-1</sup> / 9.60
<b>CTD</b>	13.56 kDa / 16,960 M <sup>-1</sup> cm <sup>-1</sup> / 9.77

<b>5a</b>	<b>Comments on sequence of expressed construct</b>
<b>IDR1-NTD-IDR2</b>	No artificial residues due to construct design.
<b>NTD-SR</b>	No artificial residues due to TEV-cleavage and construct design.
<b>NTD</b>	No artificial residues due to TEV-cleavage and construct design.
<b>CTD</b>	N-terminal „GAMG" four artificial residues due to TEV-cleavage and construct design.
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)
<b>7</b>	<b>Cultivation medium</b>
<b>IDR1-NTD-IDR2</b>	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>NTD-SR</b>	LB / M9 (uniformly <sup>15</sup> N-labelled)
<b>NTD</b>	LB / M9 (uniformly <sup>15</sup> N-labelled)
<b>CTD</b>	LB / M9 (uniformly <sup>15</sup> N or <sup>13</sup> C, <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
<b>IDR1-NTD-IDR2</b>	0.2 mM IPTG at OD <sub>600</sub> 0.8
<b>NTD-SR</b>	0.2 mM IPTG at OD <sub>600</sub> 0.7
<b>NTD</b>	0.2 mM IPTG at OD <sub>600</sub> 0.7
<b>CTD</b>	1 mM IPTG at OD <sub>600</sub> 0.7
<b>10</b>	<b>Cultivation temperature and time</b>
<b>IDR1-NTD-IDR2</b>	Cells are grown at 37°C in 1 L LB until OD <sub>600</sub> 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 <sub>600</sub> 0.9.
<b>NTD-SR</b>	16-18°C for 16-18 h
<b>NTD</b>	16-18°C for 16-18 h
<b>CTD</b>	20-22°C for 18-20 h

Table 3a: Protein Purification (IDR1-NTD-IDR2)

<b>1</b>	<b>Buffer List</b>
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 <sub>3</sub> (NMR buffer).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell lysis in Buffer <b>1A</b> 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 <b>1B</b> for buffer exchange.
C	Ion Exchange chromatography with HiTrap SP FF 5 mL column (GE Healthcare), gradient elution with buffer <b>1C</b> . The protein eluted at 45-50% gradient.

Table 3b: Protein Purification (NTD and NTD-SR)

<b>1</b>	<b>Buffer List</b>
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 <sub>2</sub> 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 <sub>2</sub> HPO <sub>4</sub> (pH 6.5), 50 mM NaCl, 500 µM PMSF, 3 mM NaN <sub>3</sub> , 3 mM EDTA (final NMR buffer).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Cell disruption in buffer <b>1A</b> by sonication.
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer <b>1B</b> and <b>1C</b> .
C	TEV-cleavage (1:10 (v/v) TEV:protein solution) during dialysis o.n. in buffer <b>1D</b> .
D	Inv. IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), elution with imidazole gradient up to 500 mM in buffer <b>1B</b> and <b>1C</b> .
E	NMR sample preparation in buffer <b>1E</b> .

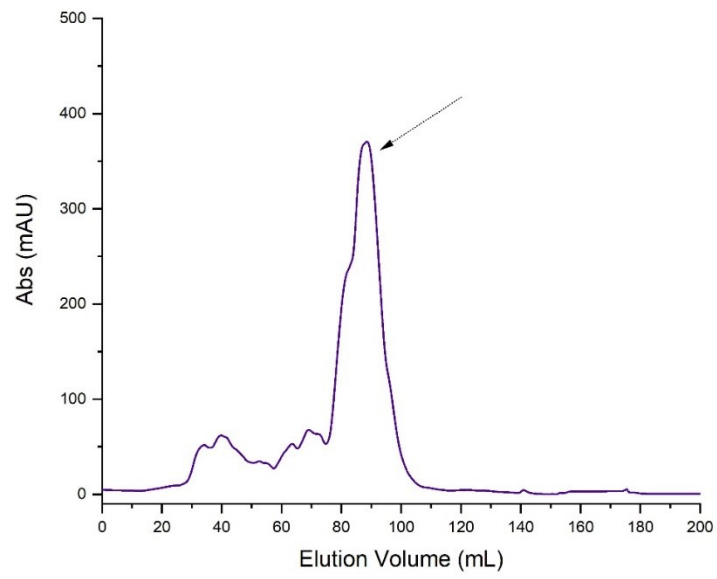
Table 3c: Protein Purification (CTD)

<b>1</b>	<b>Buffer List</b>
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 <sub>3</sub> (SEC / final NMR buffer).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>

A	Cell disruption in buffer <b>1A</b> (plus 100 $\mu$ L protease inhibitor (Serva)) by sonication.
B	IMAC (gravity flow Ni <sup>2+</sup> -NTA), Elution with 150-500 mM imidazole in buffer <b>1A</b> .
C	Dialysis o.n. in in buffer <b>1A</b> .
D	TEV-cleavage (0.5 mg TEV protease per 1 L culture) in buffer <b>1A</b> .
E	SEC on HiLoad 16/600 SD 75 (GE Healthcare) in buffer <b>1B</b> .
F	NMR sample preparation in buffer <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
<b>IDR1-NTD-IDR2</b>	12 mg/L <sup>13</sup> C, <sup>15</sup> N M9 medium
<b>NTD-SR</b>	3 mg/L <sup>15</sup> N M9 medium
<b>NTD</b>	3 mg/L <sup>15</sup> N M9 medium
<b>CTD</b>	2 mg/L <sup>13</sup> C, <sup>15</sup> N-M9 medium
<b>1b</b>	<b>A260/280 ratio</b>
<b>IDR1-NTD-IDR2</b>	0.63
<b>NTD-SR</b>	0.7
<b>NTD</b>	0.7
<b>CTD</b>	0.55
<b>2</b>	<b>Stability</b>
<b>IDR1-NTD-IDR2</b>	Protein is stable for at least one1 week at working conditions (298 K).
<b>NTD-SR</b>	Stable throughout measurement (15 days, 298 K). No significant precipitation or degradation observed after storage at 4°C for 5 weeks.
<b>NTD</b>	Stable throughout measurement (15 days, 298 K). No significant precipitation or degradation observed after storage at 4°C for 5 weeks.
<b>CTD</b>	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.
<b>3</b>	<b>Comment on applicability</b>
	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

<b>1</b>	<b>Protein Name</b>
	ORF9b
<b>2</b>	<b>Region/Name/Further Specification</b>
<b>3</b>	<b>Sequence of fl protein</b>
	MDPKISEMHP ALRLVDPQIQ LAVTRMENAV GRDQNNVGPK VYPIILRLGS PLSLNMARKT LNSLEDKAFQ LTPIAVQMTK LATTEELPDE FVVVTVK
<b>4</b>	<b>Protein boundaries of expressed construct</b>
	aa 1-97 (fl ORF9b)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity : 72.4%; similarity: 95.0%
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	SCoV2: PDB 6Z4U
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Cell-free Protein Synthesis

<b>1</b>	<b>Expression vector</b>
	pEU-E01-MCS (Cell-Free Sciences)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	C-terminal Strep tag II (WSHPQFEK)
<b>3</b>	<b>Cleavage Site</b>
	-
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of protein</b>
	11.99 kDa / 6,990 M <sup>-1</sup> cm <sup>-1</sup> / 6.73
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	C-terminal "SAWSHPQFEK" ten artificial residues due to construct design.
<b>6</b>	<b>Feeding buffer</b>

	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
<b>7</b>	<b>Translation mix</b>
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix
<b>8</b>	<b>Protein synthesis temperature and time</b>
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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.
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
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 <b>1A</b> (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of <b>1A</b> .
H	Elute the protein of interest with <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.64 mg/mL WGE and total production of 1338 µg for NMR samples.
<b>1b</b>	<b>A260/280 ratio</b>
	0.76
<b>2</b>	<b>Stability</b>
	Stable at 4°C for a week.
<b>3</b>	<b>Comment on applicability</b>
	Protein studied at pH 6, 7.5 and pH 8. Methionine gets oxidized without DTT in the buffer.

Additional information

	<b>Constructs</b>	<b>Conditions</b>	<b>Comments</b>
<b>A</b>	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>B</b>		As above without DTT	Methionines get oxidated.

## SI22: ORF14

Table 1: General Information

<b>1</b>	<b>Protein Name</b>
	ORF14
<b>2</b>	<b>Region/Name/Further Specification</b>
<b>3</b>	<b>Sequence of fl protein</b>
	MLQSCYNFLKEQHCQKASTQKGAEAAVKPLLVPHHVVATVQEIQLQAAVGELELLLEWLAMAVMLLLLCCCLTD
<b>4</b>	<b>Protein boundaries of expressed construct</b>
	aa 1-73 (fl ORF14)
<b>5</b>	<b>Ratio for construct design</b>
	fl protein
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: NA; similarity: NA
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	-
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Cell-free Protein Synthesis

<b>1</b>	<b>Expression vector</b>
	pEU-E01-MCS (Cell-Free Sciences)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal Strep tag II (WSHPQFEK)
<b>3</b>	<b>Cleavage Site</b>
	-
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of protein</b>
	9.26 kDa / 12,490 M <sup>-1</sup> cm <sup>-1</sup> / 6.01
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal “WSHPQFEKGGG” eleven artificial residues due to construct design.
<b>6</b>	<b>Feeding buffer</b>

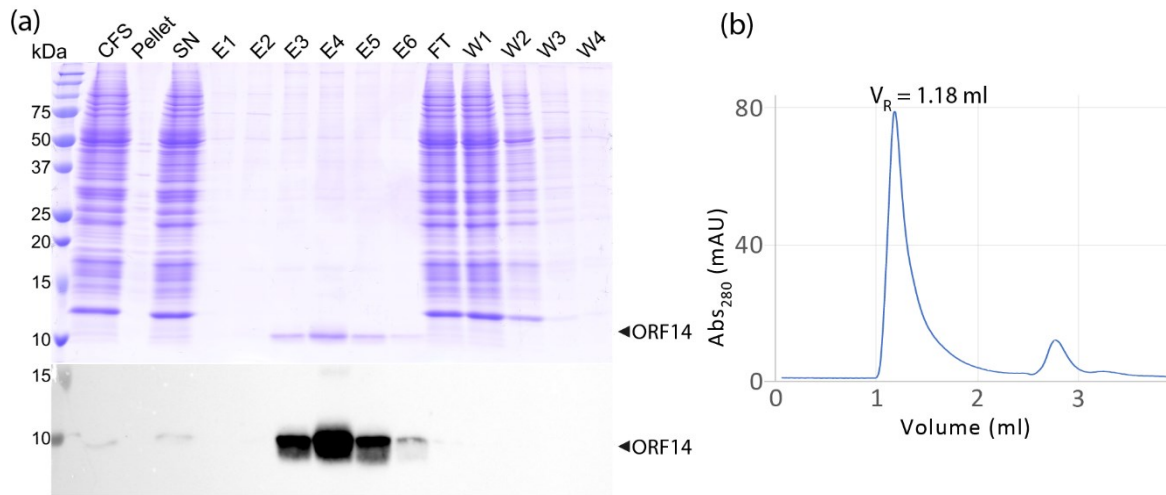
	30 mM HEPES-KOH (pH 7.6), 100 mM potassium acetate, 2.7 mM magnesium acetate, 16 mM creatine phosphate, 0.4 mM spermidine, 1.2 mM ATP, 0.25 mM GTP, 4 mM DTT and 6 mM (average concentration) amino acid mix and 0.05% (w/v) Brij-58.
<b>7</b>	<b>Translation mix</b>
	50% (v/v) mRNA, 50% (v/v) home-made WGE, 40 µg/mL creatine kinase, and 6 mM (average concentration) amino acid mix 0.05% (w/v) Brij-58.
<b>8</b>	<b>Protein synthesis temperature and time</b>
	22°C for 16 h without agitation (bilayer method).

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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).
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
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 <b>1A</b> (all steps performed on the bench by gravity).
F	Load SN onto the column.
G	Wash the column with 5 CV of <b>1A</b> .
H	Elute the protein of interest with <b>1B</b> .

Table 4: Final sample

<b>1</b>	<b>Yield</b>
	0.43 mg/mL WGE
<b>1b</b>	<b>A260/280 ratio</b>
	1.06
<b>2</b>	<b>Stability</b>
	protein has proved unstable during lipid insertion using cyclodextrin for detergent removal
<b>3</b>	<b>Comment on applicability</b>
	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

<b>1</b>	<b>Protein Name (according to NCBI Reference Sequence NC_045512.2)</b>
	ORF10
<b>2</b>	<b>Region/Name/Further Specification</b>
<b>3</b>	<b>Sequence of fl protein (according to NCBI Reference Sequence NC_045512.2)</b>
	MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT
<b>4</b>	<b>Protein boundaries of expressed construct (according to NCBI Reference Sequence NC_045512.2)</b>
	aa 1-38 (fl ORF10)
<b>5</b>	<b>Ratio for construct design</b>
	Hypothetical fl protein.
<b>6</b>	<b>Sequence homology (to SCoV)</b>
	Identity: 29%; similarity: 52% with ORF9b
<b>7</b>	<b>Published structures (SCoV2 or homologue variants)</b>
	-
<b>8</b>	<b>(Published) assignment (SCoV2 or homologue variants)</b>
	-

Table 2: Protein Expression

<b>1</b>	<b>Expression vector</b>
	pThioRed (GenScript)
<b>2</b>	<b>Purification-/Solubility-Tag</b>
	N-terminal His <sub>6</sub> -Trx
<b>3</b>	<b>Cleavage Site</b>
	TEV
<b>4</b>	<b>Molecular weight / Extinction coefficient / pI - of cleaved protein</b>
	4.45 kDa / 4,470 M <sup>-1</sup> cm <sup>-1</sup> / 7.93
<b>5</b>	<b>Comments on sequence of expressed construct</b>
	N-terminal "GA" two artificial residues due to TEV-cleavage and construct design
<b>6</b>	<b>Used expression strain</b>
	<i>E. coli</i> BL21 (DE3)

<b>7</b>	<b>Cultivation medium</b>
	LB / M9 (uniformly <sup>15</sup> N-labelled)
<b>8</b>	<b>Induction system</b>
	IPTG inducible T7 promoter
<b>9</b>	<b>Induction of protein expression</b>
	0.2 mM IPTG at OD <sub>600</sub> 0.6-0.7
<b>10</b>	<b>Cultivation temperature and time</b>
	18-20°C for 16-18 h

Table 3: Protein Purification

<b>1</b>	<b>Buffer List</b>
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.
<b>2</b>	<b>Purification steps (with corresponding buffer(s) and incubation times)</b>
A	Solubilization of cell pellet and inclusion bodies in <b>1A</b> (plus one tablet of EDTA free protease inhibitor cocktail (Merck)).
B	IMAC (HisTrap HP (GE Healthcare), ÄKTA start (GE Healthcare)), washed with buffer <b>1B</b> , refolded on column in buffer <b>1C</b> , elution with imidazole gradient up to 500 mM in buffer <b>1C</b> .
C	Analytic TEV-cleavage (1 mg TEV protease per 50 mL protein solution) o.n. in buffer <b>1C</b> .
D	Analytical SEC (SD 75 Increase 10/300 GL (GE Healthcare), ÄKTA start (GE Healthcare)) in buffer <b>1D</b> .

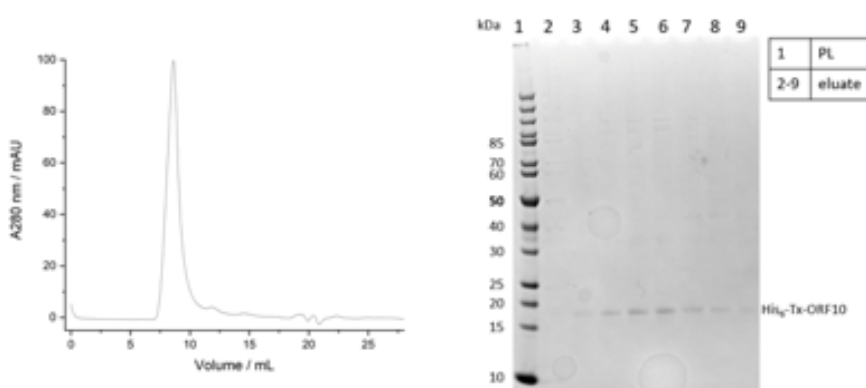
Table 4: Final sample

<b>1</b>	<b>Yield</b>
	2 mg/L ( <sup>15</sup> N-M9) His <sub>6</sub> -SUMO-fused
<b>2</b>	<b>Stability</b>
	Degrades after cleavage
<b>3</b>	<b>Comment on applicability</b>
	Tendency to oligomerize (exclusion volume of SD 75 column).



Additional information

	Constructs	Conditions	Comments
<b>A</b>	As above	Native <b>IMAC buffer:</b> 25 mM Tris-HCl (pH 8.0), 300 mM NaCl, 5 mM imidazole, 10 mM bME. <b>SEC buffer:</b> 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>B</b>	F1 ORF10; His <sub>6</sub> -SUMO (pE-SUMO (GenScript)), Ulp1-cleavage site, no artificial residues.	Native <b>IMAC buffer:</b> as above <b>SEC buffer:</b> 25 mM NaPi (pH 7.0), 150 mM NaCl, 2 mM TCEP-HCl.	Protein is mostly soluble with fusion, partial degradation (copurification of His <sub>6</sub> -SUMO), runs in exclusion volume of SD 200 columns, degrades after cleavage. NMR shows SUMO is mostly unfolded.
<b>C</b>		Detergent <b>IMAC buffer:</b> 50 mM NaPi (pH 7.0), 200 mM NaCl, 0.1% (v/v) Triton X-100, 5 mM imidazole, 10 mM bME. <b>SEC buffer:</b> 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.
<b>D</b>		Semi-denaturing <b>IMAC buffer:</b> 50 mM Tris-HCl (pH 8.0), 2 M urea, 300 mM NaCl, 10 mM imidazole, 10 mM bME. <b>SEC buffer:</b> 25 mM NaPi (pH 6.5), 50 mM NaCl, 2 M urea, 5 mM DTT.	Degrades after cleavage.



Analytical SEC of His<sub>6</sub>-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).

## **Article 2.3:**

**NMR reveals specific tracts within the intrinsically disordered regions of the SARS-CoV 2 Nucleocapsid protein involved in RNA encountering**

## Article

# NMR Reveals Specific Tracts within the Intrinsically Disordered Regions of the SARS-CoV-2 Nucleocapsid Protein Involved in RNA Encountering

Letizia Pontoriero <sup>1,†</sup>, Marco Schiavina <sup>1,†</sup>, Sophie M. Korn <sup>2,†</sup>, Andreas Schlundt <sup>2,\*</sup> , Roberta Pierattelli <sup>1,\*</sup>  and Isabella C. Felli <sup>1,\*</sup>

<sup>1</sup> Magnetic Resonance Center (CERM) and Department of Chemistry “Ugo Schiff”, University of Florence, Via L. Sacconi 6, Sesto Fiorentino, 50019 Florence, Italy; pontoriero@cerm.unifi.it (L.P.); schiavina@cerm.unifi.it (M.S.)

<sup>2</sup> Center for Biomolecular Magnetic Resonance (BMRZ), Institute for Molecular Biosciences, Johann Wolfgang Goethe-University, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany; bochmann@bio.uni-frankfurt.de

\* Correspondence: schlundt@bio.uni-frankfurt.de (A.S.); roberta.pierattelli@unifi.it (R.P.); felli@cerm.unifi.it (I.C.F.)

† These authors contributed equally to the work.

**Abstract:** The SARS-CoV-2 nucleocapsid (N) protein is crucial for the highly organized packaging and transcription of the genomic RNA. Studying atomic details of the role of its intrinsically disordered regions (IDRs) in RNA recognition is challenging due to the absence of structure and to the repetitive nature of their primary sequence. IDRs are known to act in concert with the folded domains of N and here we use NMR spectroscopy to identify the priming events of N interacting with a regulatory SARS-CoV-2 RNA element. <sup>13</sup>C-detected NMR experiments, acquired simultaneously to <sup>1</sup>H detected ones, provide information on the two IDRs flanking the N-terminal RNA binding domain (NTD) within the N-terminal region of the protein (NTR, 1–248). We identify specific tracts of the IDRs that most rapidly sense and engage with RNA, and thus provide an atom-resolved picture of the interplay between the folded and disordered regions of N during RNA interaction.

**Keywords:** SARS-CoV-2; COVID-19; IDP; RNA; NMR



**Citation:** Pontoriero, L.; Schiavina, M.; Korn, S.M.; Schlundt, A.; Pierattelli, R.; Felli, I.C. NMR Reveals Specific Tracts within the Intrinsically Disordered Regions of the SARS-CoV-2 Nucleocapsid Protein Involved in RNA Encountering. *Biomolecules* **2022**, *12*, 929. <https://doi.org/10.3390/biom12070929>

Academic Editors: Stefania Brocca, Keith Dunker, Sonia Longhi and Prakash Kulkarni

Received: 6 June 2022

Accepted: 29 June 2022

Published: 2 July 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The nucleocapsid protein N of SARS-CoV-2 plays a pivotal role in the viral life cycle. The protein is organized in five different modular domains, two folded and three disordered ones, with the latter comprising almost 40% of the whole protein sequence (Supplementary Figure S1) [1,2]. It exerts various functions including packaging of genomic RNA (gRNA) inside the viral capsid [3–8] but the structural and mechanistic details of packaging remain enigmatic. The SARS-CoV-2 genome comprises a multitude of highly conserved structured *cis* regulatory RNA elements [4], which have been suggested as target sites for N in the context of packaging [9]. It is thus important to study how the disordered protein regions modulate the interaction with RNA. Recent work showed the potential of solution NMR [10–17] to describe the structural and dynamic features of different N constructs and how they interact with RNA fragments. Here we would like to explore how <sup>13</sup>C detection can contribute to this field.

<sup>13</sup>C-NMR emerged as a key technique to study intrinsically disordered proteins (IDPs) [18]. The large chemical shift dispersion of heteronuclei (<sup>13</sup>C, <sup>15</sup>N) is crucial to obtaining highly resolved spectra in the absence of a stable 3D structure. Solvent exchange often leads to the broadening of amide proton signals, in particular for exposed protein backbones, when approaching physiological pH and temperature. <sup>13</sup>C-detected heteronuclear NMR experiments allow us to overcome this limitation. For these reasons, they

constitute a valuable tool to investigate highly flexible polypeptide chains also when part of a multi-domain protein.

The contribution of the flexible regions of N to the interaction with RNA is investigated here by selecting a construct comprising the folded N-terminal domain NTD (44–180) and the flanking intrinsically disordered regions, IDR1 (1–43) and IDR2 (181–248). This allows us to focus on the IDRs while linked to the NTD, that is the domain deputed to bind gRNA [1]. The interaction between this N construct (1–248, referred to as N-terminal region, NTR) with RNA was studied by selecting a highly conserved *cis* element of the gRNA, namely the 5'-UTR-contained stem-loop 4 (5\_SL4) [19]. This is centrally located within the 5'-UTR, has very recently been found targetable by small molecules [20] and thus represents a potential drug target to disrupt its interactions with abundant viral proteins such as N. It is described as stable [5] and is chemically versatile comprising a pentaloop, two internal loops, a bulge, and a good mix of nucleotides and types of base pairs (Supplementary Figure S1). It thus represents a bona fide example RNA for this study.

## 2. Materials and Methods

### 2.1. Protein Sample Preparation

The NTD and the NTR samples were prepared as previously described [13,21] and briefly summarized hereafter.

For the NTR construct, the gene of the N protein comprising residues 1–248 was designed based on the boundaries determined from the SARS-CoV homologue [1]. The codon-optimized gene was synthesized by Twist Bioscience and cloned into pET29b(+) vector between NdeI and XhoI restriction sites.

Uniformly  $^{13}\text{C}$ ,  $^{15}\text{N}$ -labeled NTR protein was expressed in *E. coli* strain BL21 (DE3) following the Marley method [22]. The cells were grown in 1 L Luria Bertani medium at 37 °C until an optical density ( $\text{OD}_{600}$ ) of 0.8 was reached. Then, the culture was transferred in 250 mL of labeled minimal medium supplemented with 0.25 g/L  $^{15}\text{NH}_4\text{Cl}$  (Cambridge Isotope Laboratories) and 0.75 g/L  $^{13}\text{C}_6\text{-D-glucose}$  (Eurisotop). After 1 h of unlabeled metabolite clearance, the culture was induced with 0.2 mM isopropyl-beta-thiogalactopyranoside (IPTG) at 16 °C for 18 h. The pellet was harvested and stored at –20 °C overnight. The cell pellet was then resuspended in 25 mM 2-amino-2-(hydroxymethyl)-1,3-propanediol (TRIS), 1.0 M NaCl, 10% glycerol, and protease inhibitor cocktail (SIGMA) at pH 8.0. Cells were disrupted by sonication and the lysate was centrifuged at  $30,000\times g$  for 50 min at 4 °C.

The soluble fraction was dialyzed overnight against a solution of 25 mM TRIS, pH 7.2 at 4 °C. The protein solution was then loaded on a HiTrap SP FF 5 mL column and eluted in 25 CV with a 70% gradient of 25 mM TRIS and 1.0 M NaCl. Fractions containing the protein were pooled, concentrated, and loaded on a HiLoad 16/1000 Superdex 75 pg column equilibrated with 25 mM potassium phosphate, 450 mM KCl, pH 6.5. The fractions containing the protein were pooled and concentrated using centrifugal concentrators (molecular weight cut-off 10 kDa).

The gene of the single cysteine A211C mutant of the NTR protein was synthesized by Twist Bioscience and cloned into the pET29b(+) vector between NdeI and XhoI restriction sites. Uniformly  $^{15}\text{N}$ -labeled A211C protein was expressed and purified following the same protocol used for the NTR construct, with the addition of 5 mM dithiothreitol (DTT) in the lysis and purification buffers.

The soluble fraction was dialyzed overnight against a solution of 25 mM TRIS and 5 mM DTT, pH 7.2 at 4 °C. The protein solution was then loaded on a HiTrap SP FF 5 mL column and eluted in 25 CV with a 70% gradient of 25 mM TRIS, 1.0 M NaCl, and 5 mM DTT, pH 7.2. Fractions containing the protein were pooled and concentrated to a final concentration of 25  $\mu\text{M}$ .

The sequence of the NTD (44–180) was based on SARS-CoV-2 NCBI reference genome entry NC\_045512.2, identical to GenBank entry MN90894 [23]. Domain boundaries for the core NTD were defined in analogy to the available NMR structure (PDB 6YI3) [10]. An *E. coli* codon-optimized DNA construct was obtained from Eurofins Genomics and

sub-cloned into the pET-21-based vector pET-Trx1a, containing an *N*-terminal His<sub>6</sub>-tag, a thioredoxin-tag and a tobacco etch virus (TEV) cleavage site. After proteolytic TEV cleavage, the produced 14.9 kDa protein contains one artificial *N*-terminal residue (Gly0), before the start of the native protein sequence at Gly1 which corresponds to Gly44 in the full-length *N* protein sequence.

Uniformly <sup>15</sup>N-labeled NTD was expressed in *E. coli* strain BL21 (DE3) in M9 minimal medium containing 1.0 g/L <sup>15</sup>NH<sub>4</sub>Cl (Cambridge Isotope Laboratories) and 25 µg/mL kanamycin. Protein expression was induced at an OD<sub>600</sub> of 0.8 with 1 mM IPTG for 18 h at room temperature. Cell pellets were resuspended in 50 mM TRIS/HCl pH 8.0, 300 mM NaCl, 10 mM imidazole, and 100 µL protease inhibitor mix (SERVA) per 1.0 L of culture. Cells were disrupted by sonication. The supernatant was cleared by centrifugation (30 min, 9000 × *g*, 4 °C). The cleared supernatant was passed over a Ni<sup>2+</sup>-NTA gravity flow column (Sigma-Aldrich) and the His<sub>6</sub>-Trx-tag was cleaved overnight at 4 °C with 0.5 mg of TEV protease per 1.0 L of culture and dialyzed into fresh buffer (50 mM TRIS/HCl pH 8.0, 300 mM NaCl, 10% glycerol). TEV protease and the cleaved tag were removed via a second Ni<sup>2+</sup>-NTA gravity flow column, and core NTD was further purified via size exclusion on a HiLoad 16/600 SD 75 (Cytiva) in 25 mM potassium phosphate, 150 mM KCl, 2 mM Tris-(2-carboxyethyl)-phosphin (TCEP), 0.02% NaN<sub>3</sub>, pH 6.5. Pure NTD protein-containing fractions were determined by SDS-PAGE, pooled and concentrated using Amicon centrifugal concentrators (molecular weight cut-off of 10 kDa).

## 2.2. RNA Production

The 40 nucleotides (nt) SARS-CoV-2 genomic RNA element stem loop 4 (SL4) located within the 5'UTR (nt 86 to 125), extended 5' by two guanine residues and 3' by two cytidine residues, yielded the 44-nt sequence 5'-GGGUG UGG CUG UCA CUC GGC UGC AUG CUU AGU GCA CUC ACGC CC-3' [19]. The DNA template for 5\_SL4 was kindly provided in a HDV ribozyme vector by the COVID19-nmr consortium. The unlabeled RNA was produced by in-house optimized *in vitro* transcription and purified as described previously [5]. Final RNA samples were buffer-exchanged to 25 mM potassium phosphate, 150 mM KCl, pH 6.5, and sample quality, homogeneity and long-term stability were verified by native and denaturing PAGE as well as 1D-NMR experiments by means of the characteristic imino proton pattern.

## 2.3. Spin-Labeling Reaction for PRE Experiments

The A211C protein solution was purified from DTT using a PD-10 desalting column and then incubated with a ten-fold excess of S-(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl) methylmethane-sulfonothiolate (MTSL) relative to the protein concentration. The reaction was performed overnight in absence of light at 4 °C while gently stirring. Then, the unreacted spin-label was eliminated using two steps of purification with a PD-10 desalting column. The protein eluted in 25 mM TRIS and 150 mM NaCl.

To reduce MTSL and obtain the diamagnetic sample, a five-fold excess of ascorbate with respect to the protein concentration was added.

## 2.4. Protein NMR Samples

For NTR, experiments were acquired using two 500-µL-samples of 140 µM <sup>13</sup>C,<sup>15</sup>N NTR solution in 25 mM potassium phosphate at pH 6.5, 150 mM KCl, 0.01% NaN<sub>3</sub> in H<sub>2</sub>O with 5% D<sub>2</sub>O. The titration was performed in 5 mm NMR tubes. A highly concentrated batch of 5\_SL4 solution in 25 mM potassium phosphate, 150 mM KCl, 0.01% NaN<sub>3</sub>, pH 6.5 was prepared as previously described and added to a protein solution sample in small aliquots to reach NTR:RNA ratios of 1:0.01, 1:0.025, and 1:0.05. A second identical protein sample was used to reach NTR:RNA ratios of 1:0.1, 1:0.3, and 1:0.6.

For NTD, experiments were acquired using one 500-µL-sample of 70 µM <sup>15</sup>N NTD solution in 25 mM potassium phosphate at pH 6.5, 150 mM KCl, 2 mM TCEP, and 0.02% NaN<sub>3</sub> in H<sub>2</sub>O with 5% D<sub>2</sub>O. A highly concentrated batch of 5\_SL4 solution in 25 mM

potassium phosphate, 150 mM KCl, 0.02% NaN<sub>3</sub>, 2 mM TCEP, and pH 6.5 was prepared as previously described and added to a protein solution sample in small aliquots to reach NTD:RNA ratios of 1:0.1, 1:0.3, 1:1.2, and 1:2.4.

### 2.5. NMR Experiments

To follow the interaction between NTR and 5\_SL4, the mr\_CON//HN experiment [24] was used. To complete the available assignment [13], a 3D-(H)CBCACON experiment [25] was also acquired on a 100 μM <sup>13</sup>C,<sup>15</sup>N NTR sample.

These NMR experiments were acquired on a Bruker AVANCE NEO spectrometer operating at 700.06 MHz <sup>1</sup>H, 176.05 MHz <sup>13</sup>C, and 70.97 MHz <sup>15</sup>N frequencies equipped with a cryogenically cooled probehead optimized for <sup>13</sup>C-direct detection (TXO) at 298 K. Standard radiofrequency pulses and carrier frequencies for triple resonance experiments were used and are summarized hereafter. <sup>13</sup>C pulses were given at 176.7 ppm, 55.9 ppm, and 45.7 ppm for C', C<sup>α</sup> and C<sup>ali</sup> spectral regions, respectively. <sup>15</sup>N pulses were given at 124.0 ppm. The <sup>1</sup>H carrier was placed at 4.7 ppm. Q5- and Q3-shaped pulses [26] of durations of 300 and 231 μs, respectively, were used for <sup>13</sup>C band-selective π/2 and π flip angle pulses except for the π pulses that should be band selective on the C<sup>α</sup> region (Q3, 1200 μs) and for the adiabatic π pulse to invert both C' and C<sup>α</sup> (smoothed chirp 500 μs, 20% smoothing, 80 kHz sweep width, 11.3 kHz radio frequency field strength) [27]. Decoupling of <sup>1</sup>H and <sup>15</sup>N was achieved with waltz65 (100 μs) and garp4 (250 μs) decoupling sequences, respectively [26,28]. All gradients employed had a smoothed square shape.

The mr\_CON//HN was acquired with an interscan delay of 1.6 s; during this delay, the HN experiment was acquired as discussed in [24]. Solvent suppression was achieved through the 3:9:19 pulse scheme [29]. For each increment of the CON experiment, acquired with 16 scans, the in-phase (IP) and antiphase (AP) components were recorded and properly combined to achieve IPAP virtual decoupling [30]. The CON spectrum was acquired with sweep widths of 5263 Hz (<sup>13</sup>C) × 2840 Hz (<sup>15</sup>N) and 1024 × 400 real points in the two dimensions, respectively. The HN spectrum was acquired with 32 scans, with sweep widths of 20869 Hz (<sup>1</sup>H) × 3194 Hz (<sup>15</sup>N) and 4096 × 400 real points in the two dimensions, respectively.

The 3D-(H)CBCACON was acquired with an interscan delay of 1 s, with 8 scans, with sweep widths of 5263 Hz (<sup>13</sup>C') × 2415 Hz (<sup>15</sup>N) × 10,204 Hz (<sup>13</sup>C<sub>ali</sub>) and 1024 × 96 × 110 real points in the three dimensions, respectively.

To follow the interaction between NTD and 5\_SL4 the 2D HN fingerprint spectra were acquired with the Fast-HSQC experimental variant [31] using a Bruker AVANCE III HD spectrometer operating at 700.17 MHz <sup>1</sup>H, 176.05 MHz <sup>13</sup>C, and 70.95 MHz <sup>15</sup>N frequencies equipped with a quadruple-resonance cryo-probehead optimized for <sup>1</sup>H-direct detection (QCI) at 298 K. The <sup>1</sup>H carrier was placed at 4.7 ppm for non-selective hard pulses and the one for <sup>15</sup>N at 117 ppm. The pulse scheme includes a 60 μs delay for binomial water suppression flanking the reverse INEPT step and calculated for the H<sup>N</sup> central region and field strength. Decoupling of <sup>15</sup>N was achieved with garp4 (250 μs) [26]. The HN experiments were acquired with an interscan delay of 1 s with 32 scans with sweep widths of 11904 Hz (<sup>1</sup>H) × 2412 Hz (<sup>15</sup>N) and 2048 × 128 real points in the two dimensions, respectively.

For the Paramagnetic Relaxation Enhancement experiments (PRE), sensitivity improvement 2D HN HSQC [32] spectra were acquired on a Bruker AVANCE NEO spectrometer operating at 900.06 (<sup>1</sup>H) and 91.20 (<sup>15</sup>N) MHz equipped with a cryogenically cooled probehead (TCI). The experiments were acquired with 32 scans, with an interscan delay of 6 s, with sweep widths of 20833 Hz (<sup>1</sup>H) × 3289 Hz (<sup>15</sup>N) and 4096 × 400 points in the two dimensions. <sup>15</sup>N pulses were given at 117.0 ppm and the <sup>1</sup>H carrier was placed at 4.7 ppm. Decoupling of <sup>15</sup>N was achieved with garp (250 μs) decoupling sequences [26]. All gradients employed had a smoothed square shape.



## 2.6. Protein Visualization

The images and the surface potential of the proteins were created and calculated using Chimera 1.14 [33] by adding to the experimental NTD structure (PDB: 6YI3 [10]) an arbitrary conformer for IDR1 and IDR2 obtained through Flexible Meccano [34].

## 2.7. NMR Spectral Analysis

All the spectra were acquired and processed by using Bruker TopSpin 4.0.8 software. Calibration of the spectra was achieved using 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) as a standard for  $^1\text{H}$  and  $^{13}\text{C}$ ;  $^{15}\text{N}$  shifts were calibrated indirectly [35].

The NTR and NTD spectra were analyzed with the aid of CARRA [36] and its tool NEASY [37]. All the spectra were integrated manually with NEASY taking into consideration only the well-resolved peaks. The volume of each peak, from each titration point, was divided by the volume measured in the reference spectrum acquired. The obtained ratios were plotted against the residue number. The missing values in the ratio intensity plots belong to proline residues (in the case of HN spectra), or to peaks that overlap with others, unless otherwise specified.

The Chemical Shift Perturbation (CSP) analysis was performed comparing two HN-HSQC acquired on the NTR and NTD at the same temperature and in the very same buffer (the one used for the RNA titration). The peak lists were manually inspected and only the well-resolved peaks were used to obtain the CSP values reported in the plot. The CSP values were calculated using the following equation:  $\text{CSP} = \sqrt{\frac{1}{2}(\delta_H^2 + 0.1 \cdot \delta_N^2)}$ , where  $\delta_H$  and  $\delta_N$  represent the variation in the chemical shift of the  $^1\text{H}$  and  $^{15}\text{N}$  nuclei, respectively.

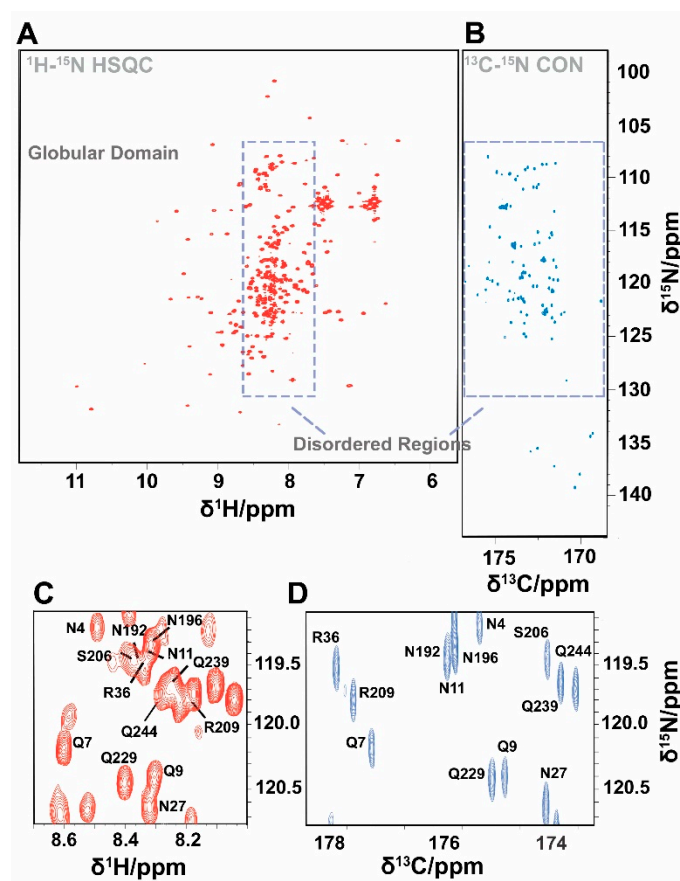
## 2.8. Electromobility Shift Assay (EMSA)

Radioactive EMSAs were performed according to [38] with the following modifications: RNA transcripts (30 pmol) were dephosphorylated using Quick CIP (NEB) following the manufacturer's protocol and finally resuspended in  $\text{H}_2\text{O}$ . Subsequently, 5' end-labelling of 15 pmol SL4 RNA with  $[\gamma\text{-}^{32}\text{P}]\text{-ATP}$  was accomplished with T4 polynucleotide kinase (NEB). Labeled RNA was separated from unincorporated  $[\gamma\text{-}^{32}\text{P}]\text{-ATP}$  by column purification (NucAway) and adjusted with binding buffer (25 mM potassium phosphate, 150 mM KCl, pH 6.5) to 0.03 pmol/ $\mu\text{L}$ . A master mix containing tRNA,  $^{32}\text{P}$ -labeled SL4 RNA, and reaction buffer was prepared and then mixed with dilutions of the NTR or NTD, respectively, to achieve the indicated protein concentrations. Binding was performed for 10 min at RT in 20- $\mu\text{L}$  reaction volume in the presence of 0.6  $\mu\text{g}$  tRNA from baker's yeast (Sigma), 3 nM  $^{32}\text{P}$ -labeled SL4 RNA, 25 mM potassium phosphate, 150 mM KCl, pH 6.5, and 1 mM  $\text{MgCl}_2$ . After the addition of 3  $\mu\text{L}$  loading buffer (30% glycerol, bromphenol blue, xylene cyanol), the RNP complexes were resolved by PAGE (6% polyacrylamide, 5% glycerol, and  $1 \times \text{TBE}$ ) at 80 V for 75 min at RT. Gels were fixed and dried and subsequently exposed to a phosphor imager screen and visualized using a GE Typhoon laser scanner under "phosphorimager" settings.

## 3. Results and Discussion

The interaction of NTR with 5\_SL4 (referred to as RNA hereafter) was studied through the  $^{13}\text{C}$ -detected  $^{13}\text{C}$ - $^{15}\text{N}$  CON (2D CON) experiment. Due to the very different structural and dynamic properties of the globular NTD domain and the flanking disordered regions, with the chosen setup, the NMR signals of the NTD are very weak or absent in the 2D CON. This allows to selectively pick up information about the disordered regions of NTR, yielding well-resolved NMR spectra, which reveal also information about seven proline residues (Figure 1). It thus provides highly complementary information to that available through a  $^1\text{H}$ -detected  $^1\text{H}$ - $^{15}\text{N}$  HSQC (2D HN) experiment. The latter allows monitoring of most of the residues belonging to the folded domain, while those of the flexible regions suffer from extensive spectral overlap or line broadening (Figure 1). The combined use of the two NMR experiments thus provides a complete picture of NTR upon interaction with

RNA. The two experiments can also be collected simultaneously [24] without compromises in the quality of either of them. This experimental variant, referred to as mr\_CON//HN, is particularly useful when dealing with multi-domain proteins constituted by globular domains and flexible regions. More than for time-saving, the approach is useful to achieve simultaneous snapshots of the protein which allow us to monitor the occurrence of the interaction from two different points of view. The two spectra obtained contain information about three different nuclei, one of them ( $^{15}\text{N}$ ) common to the two spectra. Moreover, the 2D HN can be collected with high S/N without increasing the experimental time, just exploiting the relaxation delay of the 2D CON experiment. The NMR spectra obtained through this approach on NTR are reported in Figure 1.

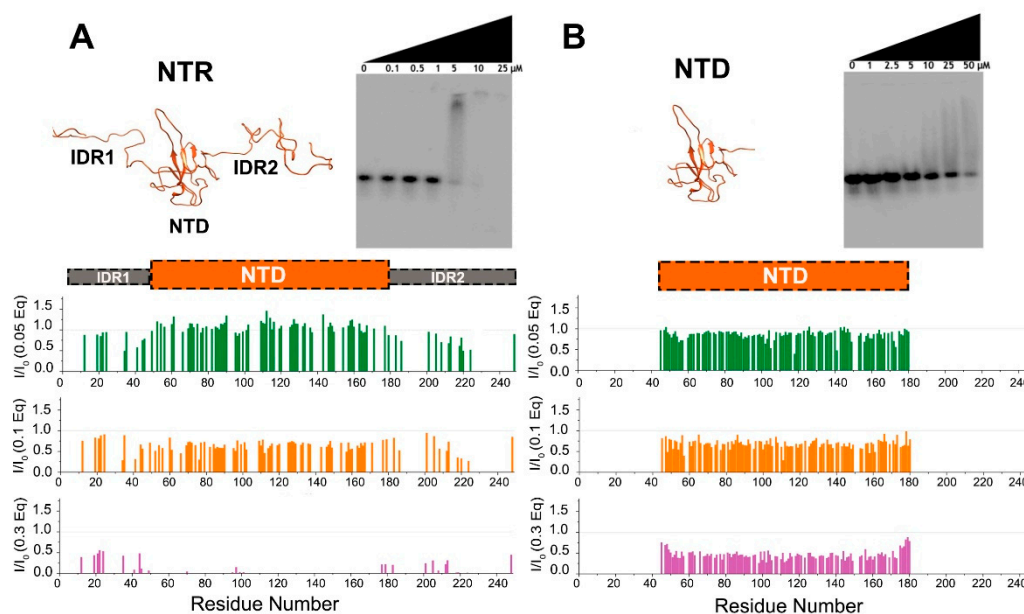


**Figure 1.** Panels A and B report the spectra obtained through the mr\_CON//HN experiment. The 2D HN spectrum (A) shows a set of well-isolated signals deriving from the globular NTD domain as well as a number of signals, clustered in a narrow central region of the spectrum, deriving from the IDRs. The 2D CON spectrum (B) allows achieving the necessary resolution to investigate resonances from IDRs, including signals of proline residues. While IDR peaks fall in a very crowded region of the HN spectrum (1.1 ppm on  $^1\text{H}$  dimension), they are well dispersed in the CON spectrum (7.2 ppm on  $^{13}\text{C}$  dimension), as indicated by the two boxes. A zoom of a region of the two spectra centered at 120 ppm for  $^{15}\text{N}$  is reported in panels (C,D) to stress this concept.

NMR spectroscopy reveals at the residue level the importance of the two disordered regions for the interaction with RNA. This is already evident when a sub-stoichiometric RNA concentration (0.05 equivalents) is added to NTR (Figure 2A). Inspection of the 2D HN spectra of NTR show variations in cross peak intensities, reported in Figure 2 as intensity ratios upon addition of increasing RNA equivalents, while shift changes are negligible (Supplementary Figure S3). In the very first points of the titration, a remarkable decrease in intensity is observed for the few resolved resonances of the HN signals from IDRs. In contrast, the signals that arise from the globular domain of the construct, seem

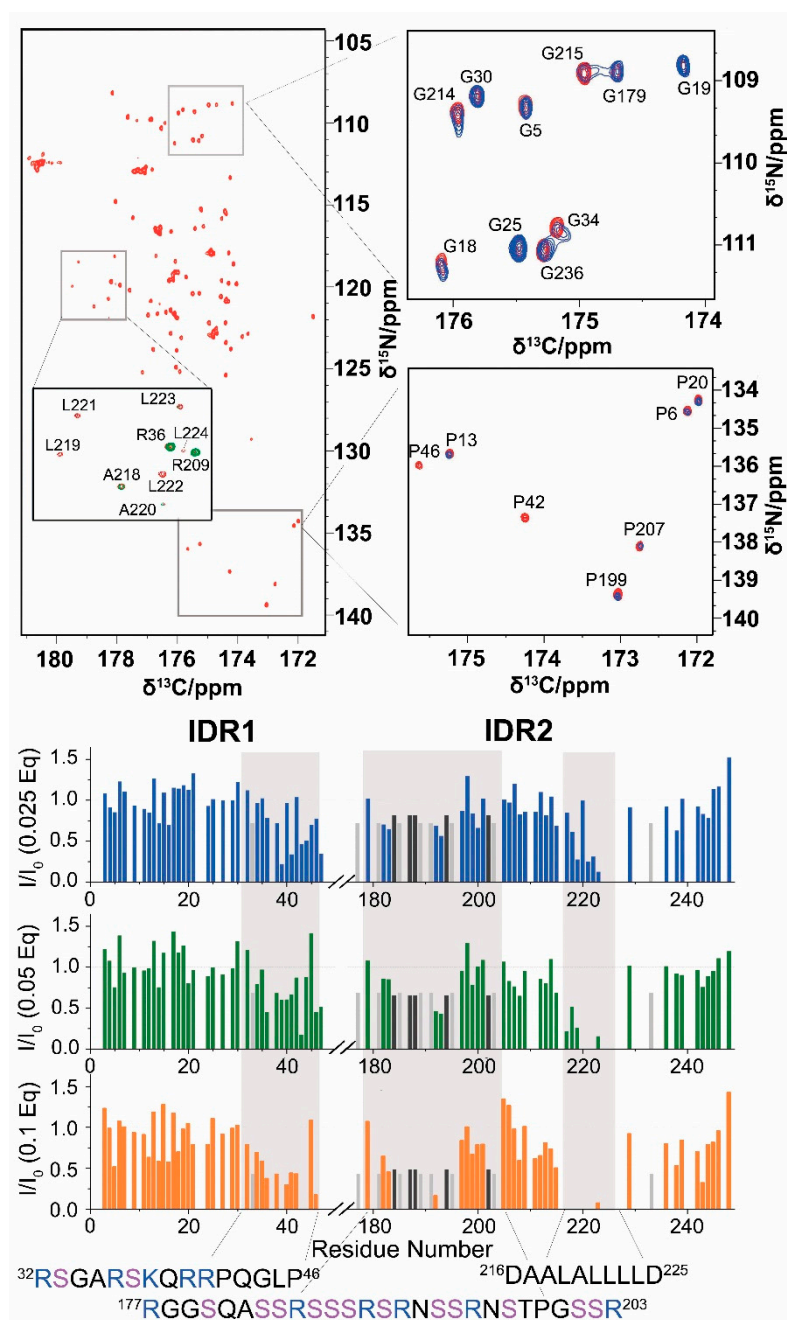


to be less perturbed by the addition of a small RNA quantity. A further increase in RNA concentration leads to a measurable signal reduction of the NTD residues, with the complete disappearance of the signals upon the addition of 0.3 equivalents of RNA. In our experimental conditions, upon further addition of RNA, we observed liquid–liquid phase separation [11,39–41], not further investigated here. In contrast, the addition of RNA to the NTD (lacking the IDRs) at the same equivalent concentrations had smaller effects on line-broadening, suggesting a reduced affinity of the isolated domain (Figure 2B). This is confirmed by Electrophoretic Mobility Shift Assay (EMSA) experiments (Figure 2 and Supplementary Figure S2). The results indicate that the NTR construct has a higher affinity towards RNA compared to the NTD alone as indicated by gel shifts observed at lower concentrations. While both NTD-containing proteins show binding to RNA, the two IDRs flanking the NTD visibly increase affinity to RNA.



**Figure 2.** Differences in the interaction of NTR (A) and NTD (B) with 5\_SL4 followed by NMR and EMSA. Upper panels show the two constructs and their different binding affinities for RNA as demonstrated by EMSA experiments. The binding of NTR to RNA occurs at a lower concentration as compared to that of NTD alone. The lower panels show plots of the HN HSQC peak intensity ratios versus residue number after the addition of increasing amounts of 5\_SL4 (with equivalents as indicated) relative to protein. The structural models were obtained as described in the experimental part.

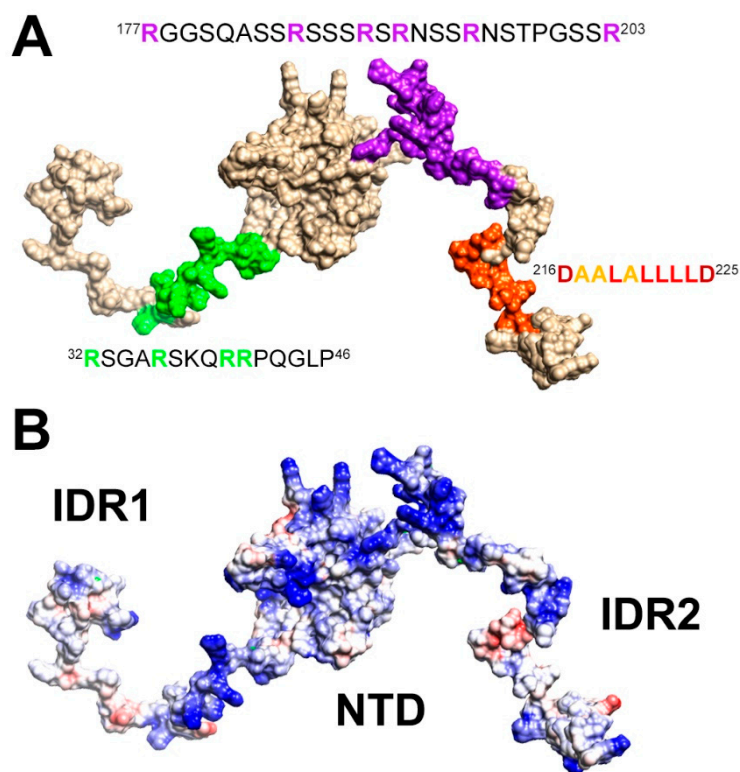
A zoom into the IDRs can be achieved through the analysis of the 2D CON spectrum. This allowed us to monitor most of the residues belonging to the highly flexible IDRs. As an example, Figure 3 shows the enlargement of selected portions of the 2D CON in diagnostic spectral regions such as that of glycine (top) and proline residues (bottom). Addition of 0.1 equivalents of RNA shows intensity changes for specific cross-peaks, suggesting the presence of preferred IDR sites for the interaction with RNA. Intensity ratios of the CON cross-peaks, obtained upon subsequent addition of RNA are reported versus the residue number in Figure 3. The most perturbed regions, indicated in the gray areas in Figure 3, comprise three different tracts (32–46, 177–203, and 216–225). These feature peculiar signatures in terms of amino acid composition as it often happens for interactions involving intrinsically disordered protein regions [42–51].



**Figure 3.** 2D CON experiments reveal differential effects of RNA-binding in specific regions of IDR1 and IDR2. The CON spectrum acquired on NTR is reported in red (**top, left**). The inset shows the superposition of the reference spectrum with NTR upon the addition of 0.01 eq of RNA (green). The enlargements of two portions of the spectra reported on the right panels (namely, the typical Gly and Pro regions) show the spectrum acquired on the NTR upon the addition of RNA (0.1 equivalents, blue) superimposed to the spectrum acquired in the absence of RNA (reference, red). The intensity ratios of CON cross-peaks are reported in the lower panel versus the residue number; spectra were acquired simultaneously to the HN spectra. Light and dark gray bars represent the intensity ratio of the envelope of signals centered at 176.6 ppm ( $^{13}\text{C}$ )–116.5 ppm ( $^{15}\text{N}$ ) and 174.8 ppm ( $^{13}\text{C}$ )–117.9 ppm ( $^{15}\text{N}$ ), respectively. Gray shaded areas highlight the protein regions most perturbed upon the addition of RNA.

Two of the tracts of NTR perturbed by the addition of RNA are very rich in positively charged residues: four arginine and one lysine residues in the region  $^{32}\text{RSGARSKQRRPQGLP}^{46}$ , and six arginine residues in the  $^{177}\text{RGGSQASSRSSSRSRNSSRNSTPGSSR}^{203}$  region

(“SR-rich region”, Supplementary Figure S1). These segments are mapped on a conformer of NTR in Figure 4A, while Figure 4B highlights the distribution of positively charged amino acids. The two tracts extend the large patch of basic residues located in the flexible, arginine-rich loop of the NTD [52], forming an extended, yet adaptable, positively charged region. These charged residues may contribute to the interaction with the RNA backbone in a priming event driven by electrostatic interactions sensed at long-distance [53]. Notably, these two regions are likely targets of regulatory post-translational modifications, such as the phosphorylation of the serine residues within the SR-rich portion that alters the overall charge of this tract (Supplementary Figure S4) [11,54].



**Figure 4.** A cartoon of the NTR construct illustrating (A) the most perturbed regions upon the addition of RNA resulting from this study and (B) the large positive patch spanning both the IDRs and the globular domain. The two models were obtained as described in the experimental section.

The third region that is perturbed by the addition of RNA (216–225) has completely different properties. This region possesses a peculiar amino acid composition (<sup>216</sup>DAALALLLD<sup>225</sup>, Figure 4A) and the NMR signals of the hydrophobic residues are weak, likely due to a helical propensity of this segment, which is reflected in signal broadening due to exchange with the protein-free conformation. Indeed, sequence-specific assignment of resonances in this region posed challenges to different NMR approaches before [13,14,16]. We obtained the assignment of the resonances belonging to these residues by exploiting a 3D (H)CBCACON experiment (Figure S5), thus extending the previously obtained sequence-specific assignment [13].

Differently from the two arginine-rich regions involved in the interaction with RNA (32–46 and 177–203), the 216–225 region does not present positively charged amino acids but has a highly hydrophobic nature resulting from branched-chain amino acids such as leucine [thus referred to as the poly-leucine (poly-L) region]. This hydrophobic stretch of 8 amino acids flanked by two negatively charged residues (Asp 216 and Asp 225) is likely to be engaged in transient interactions with other portions of NTR. A comparison of chemical shifts observed for the isolated NTD with those of the same nuclei in the NTR construct supports this hypothesis, and the insertion of a spin-label at position 211 indeed confirms

a cross-talk between the IDR and the NTD domain (Supplementary Figure S6). Of note, the potency of the poly-L stretch to mediate protein-protein interactions has very recently been manifested in its complex with the SARS-CoV-2 nsp3 Ubl domain, while, interestingly, this interaction competes with RNA-binding of N [17]. Our data support this picture in which the poly-L region serves as an interactive hub. From our data, the observed intensity changes upon the addition of RNA in the poly-L region could derive both from direct interactions with RNA as well as from weak/fuzzy intra-molecular interactions involving different domains of NTR that are disrupted by the interaction with RNA. The latter effect might alter the dynamic properties of NTR and account for the slight increase in relative signal intensities of the globular domain observed when sub-stoichiometric amounts of RNA are added (Figure 2A). Judging by our and the previous data [17], the poly-L region might act as a regulatory motif that, within N, releases the NTD in presence of RNA and/or guides the protein to functionally relevant RNP complexes via protein-protein interactions.

Summarizing, the present results indicate that electrostatics is the main driving force for molecular recognition and the arginine-rich regions, that were found to be perturbed at the early stages of the titration, are key players to promote binding with the negatively charged RNA backbone [55,56]. Interactions between disordered protein regions with complementary charges have indeed been shown to lead to high-affinity complexes [50]. The involvement of the flexible linkers is however not limited to the arginine-rich regions but also includes the poly-L region preset in IDR2. Altogether, this suggests a complex interplay between various parts of the NTR construct.

The experimental investigation of the highly dynamic properties of N is by no means a trivial task but is of crucial importance to identifying novel approaches to interfere with SARS-CoV-2. Several insights have been recently obtained on its dynamic heterogeneity [16], on the key role of the SR-rich [11] region, on the interaction with a viral chaperone, nsp3 [17]. The interaction of NTD with different RNA fragments has been studied [10,15]. In many cases, detection of NMR signals required the use of short constructs [11,14,16,17] or changes in pH and T [16]. Increasing the complexity of the system [12] revealed very interesting insights although at the expense of residue-resolved information on the disordered regions. The proposed approach offers a tool to overcome these limitations and observe in a clean way highly flexible disordered regions within multi-domain protein constructs. As an example, the 210–248 region that comprises 56% of the IDR2 residues is challenging to observe unless smaller fragments are studied, but deletion of this region from the full-length protein has been shown to significantly alter protein function [41]. It is worth noting that this portion (219–230) shares many physicochemical properties with nucleocapsids from related coronaviruses [1–3].

#### 4. Conclusions

In conclusion,  $^{13}\text{C}$ -detected NMR experiments such as the 2D CON allow us to access residue-resolved information on IDRs also when part of a multi-domain protein. They can be added to any high-resolution investigation performed through NMR, often based on the analysis of 2D HN NMR spectra only. The mr\_CON//HN approach allows their simultaneous acquisition, providing a complete picture at residue level not only for the flexible regions but at the same time for the globular NTD domain. This complementary information is highly valuable as it reflects all components in their native context.

The NMR data, supported by EMSA data, demonstrated that the flanking disordered regions of the SARS-CoV-2 NTD initiate and enhance the binding of the protein to RNA. They revealed specific tracts of the IDRs involved in the interaction within a multi-domain, cleavage prone, structurally and dynamically complex protein as NTR is.

This represents a first step necessary to unravel the detailed molecular determinants of the N protein for specific RNA encountering and subsequent complex formation, e.g., during viral genome packaging. It paves the way for further studies with increasingly complex protein constructs, ultimately with the full-length protein, as well as with other relevant elements of the SARS-CoV-2 RNA.



**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/biom12070929/s1>, Figure S1: molecular components used in this study [57,58]; Figure S2: Triplicates of EMSA gels. Figure S3: 2D HN spectra of NTR upon addition of RNA; Figure S4: NetPhos results [59]; Figure S5: 3D-(H)CBCACON strips; Figure S6: CSP and PRE results. It also includes the additional references.

**Author Contributions:** A.S., R.P. and I.C.F. conceived the project and planned the experiments; L.P. and S.M.K. produced the protein samples; S.M.K. produced the RNA samples and performed the EMSA experiments together with A.S.; L.P. and M.S. acquired and analyzed the NMR spectra on NTR together under the guidance of I.C.F. and R.P.; S.M.K. acquired and analyzed the NMR spectra on NTD together with A.S.; A.S., R.P. and I.C.F. wrote the manuscript with the contribution from all the other authors. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded in part by a grant from the Italian Ministry of University and Research (FISR2020IP\_02112, ID-COVID), by Fondazione CR Firenze and with the support of the Italian government program “MIUR Dipartimenti di Eccellenza 2018–2022” (58503\_DIPECC). This work was also supported by the Goethe Corona Funds, by the German Research Council (DFG) within CRC902 (“Molecular Principles of RNA-based regulation”) project part B18, through DFG grant number SCHL2062/2-1 to A.S., and by the Johanna Quandt Young Academy at Goethe through the financial support of A.S. (stipend number 2019/AS01).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available upon request from the corresponding authors.

**Acknowledgments:** The support of the CERM/CIRMMP center of Instruct-ERIC is gratefully acknowledged. The Covid19-NMR consortium is acknowledged for providing the DNA template plasmid for the 5\_SL4 RNA and for stimulating discussions. We would like to thank Volodya for his creativity and for his continuous inspiration to change the way we look at proteins, a strong motivation to improve experimental tools to characterize protein disorder.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Chang, C.K.; Hou, M.H.; Chang, C.F.; Hsiao, C.D.; Huang, T.H. The SARS Coronavirus Nucleocapsid Protein—Forms and Functions. *Antivir. Res.* **2014**, *103*, 39–50. [[CrossRef](#)] [[PubMed](#)]
2. Giri, R.; Bhardwaj, T.; Shegane, M.; Gehi, B.R.; Kumar, P.; Gadhave, K.; Oldfield, C.J.; Uversky, V.N. Understanding COVID-19 via Comparative Analysis of Dark Proteomes of SARS-CoV-2, Human SARS and Bat SARS-like Coronaviruses. *Cell. Mol. Life Sci.* **2021**, *78*, 1655–1688. [[CrossRef](#)] [[PubMed](#)]
3. Chang, C.-K.; Hsu, Y.-L.; Chang, Y.-H.; Chao, F.-A.; Wu, M.-C.; Huang, Y.-S.; Hu, C.-K.; Huang, T.-H. Multiple Nucleic Acid Binding Sites and Intrinsic Disorder of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein: Implications for Ribonucleocapsid Protein Packaging. *J. Virol.* **2009**, *83*, 2255–2264. [[CrossRef](#)] [[PubMed](#)]
4. Rangan, R.; Zheludev, I.N.; Hagey, R.J.; Pham, E.A.; Wayment-Steele, H.K.; Glenn, J.S.; Das, R. RNA Genome Conservation and Secondary Structure in SARS-CoV-2 and SARS-Related Viruses: A First Look. *RNA* **2020**, *26*, 937–959. [[CrossRef](#)] [[PubMed](#)]
5. Wacker, A.; Weigand, J.E.; Akabayov, S.R.; Altincekic, N.; Bains, J.K.; Banijamali, E.; Binas, O.; Castillo-Martinez, J.; Cetiner, E.; Ceylan, B.; et al. Secondary Structure Determination of Conserved SARS-CoV-2 RNA Elements by NMR Spectroscopy. *Nucleic Acids Res.* **2020**, *48*, 12415–12435. [[CrossRef](#)]
6. Cao, C.; Cai, Z.; Xiao, X.; Rao, J.; Chen, J.; Hu, N.; Yang, M.; Xing, X.; Wang, Y.; Li, M.; et al. The Architecture of the SARS-CoV-2 RNA Genome inside Virion. *Nat. Commun.* **2021**, *12*, 3917. [[CrossRef](#)]
7. de Tavares, R.C.A.; Mahadeshwar, G.; Wan, H.; Huston, N.C.; Pyle, A.M. The Global and Local Distribution of RNA Structure throughout the SARS-CoV-2 Genome. *J. Virol.* **2021**, *95*, e02190-20. [[CrossRef](#)]
8. Bai, Z.; Cao, Y.; Liu, W.; Li, J. The SARS-CoV-2 Nucleocapsid Protein and Its Role in Viral Structure, Biological Functions, and a Potential Target for Drug or Vaccine Mitigation. *Viruses* **2021**, *13*, 1115. [[CrossRef](#)]
9. Iserman, C.; Roden, C.A.; Boerneke, M.A.; Sealfon, R.S.G.; McLaughlin, G.A.; Jungreis, I.; Fritch, E.J.; Hou, Y.J.; Ekena, J.; Weidmann, C.A.; et al. Genomic RNA Elements Drive Phase Separation of the SARS-CoV-2 Nucleocapsid. *Mol. Cell* **2020**, *80*, 1078–1091.e6. [[CrossRef](#)]
10. Dinesh, D.C.; Chalupska, D.; Silhan, J.; Koutna, E.; Nencka, R.; Veverka, V.; Boura, E. Structural Basis of RNA Recognition by the SARS-CoV-2 Nucleocapsid Phosphoprotein. *PLoS Pathog.* **2020**, *16*, e1009100. [[CrossRef](#)]

11. Savastano, A.; de Opakua, A.I.; Rankovic, M.; Zweckstetter, M. Nucleocapsid Protein of SARS-CoV-2 Phase Separates into RNA-Rich Polymerase-Containing Condensates. *Nat. Commun.* **2020**, *11*, 6041. [[CrossRef](#)] [[PubMed](#)]
12. Forsythe, H.M.; Rodriguez Galvan, J.; Yu, Z.; Pinckney, S.; Reardon, P.; Cooley, R.B.; Zhu, P.; Rolland, A.D.; Prell, J.S.; Barbar, E. Multivalent Binding of the Partially Disordered SARS-CoV-2 Nucleocapsid Phosphoprotein Dimer to RNA. *Biophys. J.* **2021**, *120*, 2890–2901. [[CrossRef](#)] [[PubMed](#)]
13. Schiavina, M.; Pontoriero, L.; Uversky, V.N.; Felli, I.C.; Pierattelli, R. 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.* **2021**, *15*, 219–227. [[CrossRef](#)] [[PubMed](#)]
14. Guseva, S.; Perez, L.M.; Camacho-Zarco, A.; Bessa, L.M.; Salvi, N.; Malki, A.; Maurin, D.; Blackledge, M. <sup>1</sup>H, <sup>13</sup>C and <sup>15</sup>N Backbone Chemical Shift Assignments of the N-Terminal and Central Intrinsically Disordered Domains of SARS-CoV-2 Nucleoprotein. *Biomol. NMR Assign.* **2021**, *15*, 255–260. [[CrossRef](#)]
15. Caruso, Í.P.; Sanches, K.; Da Poian, A.T.; Pinheiro, A.S.; Almeida, F.C.L. Dynamics of the SARS-CoV-2 Nucleoprotein N-Terminal Domain Triggers RNA Duplex Destabilization. *Biophys. J.* **2021**, *120*, 2814–2827. [[CrossRef](#)]
16. Redzic, J.S.; Lee, E.; Born, A.; Issaian, A.; Henen, M.A.; Nichols, P.J.; Blue, A.; Hansen, K.C.; D’Alessandro, A.; Vögeli, B.; et al. The Inherent Dynamics and Interaction Sites of the SARS-CoV-2 Nucleocapsid N-Terminal Region. *J. Mol. Biol.* **2021**, *433*, 167108. [[CrossRef](#)]
17. Bessa, L.M.; Guseva, S.; Camacho-Zarco, A.R.; Salvi, N.; Maurin, D.; Perez, L.M.; Botova, M.; Malki, A.; Nanao, M.; Jensen, M.R.; et al. The Intrinsically Disordered SARS-CoV-2 Nucleoprotein in Dynamic Complex with Its Viral Partner Nsp3a. *Sci. Adv.* **2022**, *8*, eabm4034. [[CrossRef](#)]
18. Felli, I.C.; Pierattelli, R. <sup>13</sup>C Direct Detected NMR for Challenging Systems. *Chem. Rev.* **2022**, *122*, 9468–9496. [[CrossRef](#)]
19. Vögele, J.; Ferner, J.-P.; Altincekic, N.; Bains, J.K.; Ceylan, B.; Fürtig, B.; Grün, J.T.; Hengesbach, M.; Hohmann, K.F.; Hymon, D.; et al. <sup>1</sup>H, <sup>13</sup>C, <sup>15</sup>N and <sup>31</sup>P Chemical Shift Assignment for Stem-Loop 4 from the 5'-UTR of SARS-CoV-2. *Biomol. NMR Assign.* **2021**, *15*, 335–340. [[CrossRef](#)]
20. Sreeramulu, S.; Richter, C.; Berg, H.; Wirtz Martin, M.A.; Ceylan, B.; Matzel, T.; Adam, J.; Altincekic, N.; Azzaoui, K.; Bains, J.K.; et al. Exploring the Druggability of Conserved RNA Regulatory Elements in the SARS-CoV-2 Genome. *Angew. Chem. Int. Ed.* **2021**, *60*, 19191–19200. [[CrossRef](#)]
21. Altincekic, N.; Korn, S.M.; Qureshi, N.S.; Dujardin, M.; Ninot-Pedrosa, M.; Abele, R.; Abi Saad, M.J.; Alfano, C.; Almeida, F.C.L.; Alshamleh, I.; et al. Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications. *Front. Mol. Biosci.* **2021**, *8*, 653148. [[CrossRef](#)]
22. Marley, J.; Lu, M.; Bracken, C. A Method for Efficient Isotopic Labeling of Recombinant Proteins. *J. Biomol. NMR* **2001**, *20*, 71–75. [[CrossRef](#)] [[PubMed](#)]
23. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; et al. A New Coronavirus Associated with Human Respiratory Disease in China. *Nature* **2020**, *579*, 265–269. [[CrossRef](#)] [[PubMed](#)]
24. Schiavina, M.; Murrall, M.G.; Pontoriero, L.; Sainati, V.; Kümmerle, R.; Bermel, W.; Pierattelli, R.; Felli, I.C. Taking Simultaneous Snapshots of Intrinsically Disordered Proteins in Action. *Biophys. J.* **2019**, *117*, 46–55. [[CrossRef](#)] [[PubMed](#)]
25. Bermel, W.; Bertini, I.; Csizmok, V.; Felli, I.C.; Pierattelli, R.; Tompa, P. H-Start for Exclusively Heteronuclear NMR Spectroscopy: The Case of Intrinsically Disordered Proteins. *J. Magn. Reson.* **2009**, *198*, 275–281. [[CrossRef](#)]
26. Emsley, L.; Bodenhausen, G. Optimization of Shaped Selective Pulses for NMR Using a Quaternion Description of Their Overall Propagators. *J. Magn. Reson.* **1992**, *97*, 135–148. [[CrossRef](#)]
27. Böhlen, J.M.; Bodenhausen, G. Experimental Aspects of Chirp NMR Spectroscopy. *J. Magn. Reson. Ser. A* **1993**, *102*, 293–301. [[CrossRef](#)]
28. Geen, H.; Freeman, R. Band-Selective Radiofrequency Pulses. *J. Magn. Reson.* **1991**, *93*, 93–141. [[CrossRef](#)]
29. Piotto, M.; Saudek, V.; Sklenar, V. Gradient-Tailored Excitation for Single-Quantum NMR Spectroscopy of Aqueous Solutions. *J. Biomol. NMR* **1992**, *2*, 661–665. [[CrossRef](#)]
30. Felli, I.C.; Pierattelli, R. Spin-State-Selective Methods in Solution- and Solid-State Biomolecular <sup>13</sup>C NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **2015**, *84–85*, 1–13. [[CrossRef](#)]
31. Mori, S.; Abeygunawardana, C.; Johnson, M.O.; Vanzijl, P.C.M. 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* **1995**, *108*, 94–98. [[CrossRef](#)] [[PubMed](#)]
32. Palmer, A.G.; Cavanagh, J.; Wright, P.E.; Rance, M. Sensitivity Improvement in Proton-Detected Two-Dimensional Heteronuclear Correlation NMR Spectroscopy. *J. Magn. Reson.* **1991**, *93*, 151–170. [[CrossRef](#)]
33. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera: A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [[CrossRef](#)] [[PubMed](#)]
34. Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J.R.; Jensen, M.R.; Segard, S.; Bernadó, P.; Charavay, C.; Blackledge, M. Flexible-Meccano: A Tool for the Generation of Explicit Ensemble Descriptions of Intrinsically Disordered Proteins and Their Associated Experimental Observables. *Bioinformatics* **2012**, *28*, 1463–1470. [[CrossRef](#)]
35. Markley, J.L.; Bax, A.; Arata, Y.; Hilbers, C.W.; Kaptein, R.; Sykes, B.D.; Wright, P.E.; Wuethrich, K. Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. *Pure Appl. Chem.* **1998**, *70*, 117–142. [[CrossRef](#)]
36. Keller, R. *The Computer Aided Resonance Assignment Tutorial*; Cantina Verlag: Goldau, Switzerland, 2004; pp. 1–81.

37. Bartels, C.; Xia, T.H.; Billeter, M.; Güntert, P.; Wüthrich, K. The Program XEASY for Computer-Supported NMR Spectral Analysis of Biological Macromolecules. *J. Biomol. NMR* **1995**, *6*, 1–10. [[CrossRef](#)]
38. Ryder, S.P.; Recht, M.I.; Williamson, J.R. Quantitative Analysis of Protein-RNA Interactions by Gel Mobility Shift. *Methods Mol. Biol.* **2008**, *488*, 99–115.
39. Perdikari, T.M.; Murthy, A.C.; Ryan, V.H.; Watters, S.; Naik, M.T.; Fawzi, N.L. SARS-CoV-2 Nucleocapsid Protein Phase-separates with RNA and with Human HnRNPs. *EMBO J.* **2020**, *39*, e106478. [[CrossRef](#)]
40. Cubuk, J.; Alston, J.J.; Incicco, J.J.; Singh, S.; Stuchell-Brereton, M.D.; Ward, M.D.; Zimmerman, M.I.; Vithani, N.; Griffith, D.; Wagoner, J.A.; et al. The SARS-CoV-2 Nucleocapsid Protein Is Dynamic, Disordered, and Phase Separates with RNA. *Nat. Commun.* **2021**, *12*, 1936. [[CrossRef](#)]
41. Lu, S.; Ye, Q.; Singh, D.; Cao, Y.; Diedrich, J.K.; Yates, J.R.; Villa, E.; Cleveland, D.W.; Corbett, K.D. The SARS-CoV-2 Nucleocapsid Phosphoprotein Forms Mutually Exclusive Condensates with RNA and the Membrane-Associated M Protein. *Nat. Commun.* **2021**, *12*, 502. [[CrossRef](#)]
42. Tompa, P.; Fuxreiter, M. Fuzzy Complexes: Polymorphism and Structural Disorder in Protein-Protein Interactions. *Trends Biochem. Sci.* **2008**, *33*, 2–8. [[CrossRef](#)] [[PubMed](#)]
43. Mittag, T.; Kay, L.E.; Forman-Kay, J.D. Protein Dynamics and Conformational Disorder in Molecular Recognition. *J. Mol. Recognit.* **2009**, *23*, 105–116. [[CrossRef](#)] [[PubMed](#)]
44. Kurzbach, D.; Schwarz, T.C.; Platzer, G.; Höfler, S.; Hinderberger, D.; Konrat, R. Compensatory Adaptations of Structural Dynamics in an Intrinsically Disordered Protein Complex. *Angew. Chem. Int. Ed.* **2014**, *53*, 3840–3843. [[CrossRef](#)] [[PubMed](#)]
45. Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V.N. Introducing Protein Intrinsic Disorder. *Chem. Rev.* **2014**, *114*, 6561–6588. [[CrossRef](#)]
46. Fuxreiter, M.; Tóth-Petróczy, Á.; Kraut, D.A.; Matouschek, A.; Matouschek, A.T.; Lim, R.Y.H.; Xue, B.; Kurgan, L.; Uversky, V.N. Disordered Proteinaceous Machines. *Chem. Rev.* **2014**, *114*, 6806–6843. [[CrossRef](#)]
47. Contreras-Martos, S.; Piai, A.; Kosol, S.; Varadi, M.; Bekesi, A.; Lebrun, P.; Volkov, A.N.; Gevaert, K.; Pierattelli, R.; Felli, I.C.; et al. Linking Functions: An Additional Role for an Intrinsically Disordered Linker Domain in the Transcriptional Coactivator CBP. *Sci. Rep.* **2017**, *7*, 4676. [[CrossRef](#)]
48. Arbesú, M.; Iruela, G.; Fuentes, H.; Teixeira, J.M.C.; Pons, M. Intramolecular Fuzzy Interactions Involving Intrinsically Disordered Domains. *Front. Mol. Biosci.* **2018**, *5*, 39. [[CrossRef](#)]
49. Spreitzer, E.; Usluer, S.; Madl, T. Probing Surfaces in Dynamic Protein Interactions. *J. Mol. Biol.* **2020**, *432*, 2949–2972. [[CrossRef](#)]
50. Sottini, A.; Borgia, A.; Borgia, M.B.; Bugge, K.; Nettels, D.; Chowdhury, A.; Heidarsson, P.O.; Zosel, F.; Best, R.B.; Kragelund, B.B.; et al. Polyelectrolyte Interactions Enable Rapid Association and Dissociation in High-Affinity Disordered Protein Complexes. *Nat. Commun.* **2020**, *11*, 5736. [[CrossRef](#)]
51. Murralli, M.G.; Felli, I.C.; Pierattelli, R. Adenoviral E1A Exploits Flexibility and Disorder to Target Cellular Proteins. *Biomolecules* **2020**, *10*, 1541. [[CrossRef](#)]
52. Clarkson, M.W.; Lei, M.; Eisenmesser, E.Z.; Labeikovsky, W.; Redfield, A.; Kern, D. Mesodynamics in the SARS Nucleocapsid Measured by NMR Field Cycling. *J. Biomol. NMR* **2009**, *45*, 217–225. [[CrossRef](#)] [[PubMed](#)]
53. 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* **2013**, *110*, 13392–13397. [[CrossRef](#)]
54. Carlson, C.R.; Asfaha, J.B.; Ghent, C.M.; Howard, C.J.; Hartooni, N.; Safari, M.; Frankel, A.D.; Morgan, D.O. Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for Its Dual Functions. *Mol. Cell* **2020**, *80*, 1092–1103.e4. [[CrossRef](#)] [[PubMed](#)]
55. Calabretta, S.; Richard, S. Emerging Roles of Disordered Sequences in RNA-Binding Proteins. *Trends Biochem. Sci.* **2015**, *40*, 662–672. [[CrossRef](#)] [[PubMed](#)]
56. Järvelin, A.I.; Noerenberg, M.; Davis, I.; Castello, A. The New (Dis)Order in RNA Regulation. *Cell Commun. Signal.* **2016**, *14*, 9. [[CrossRef](#)]
57. Popena, M.; Szachniuk, M.; Antczak, M.; Purzycka, K.J.; Lukasiak, P.; Bartol, N.; Blazewicz, J.; Adamiak, R.W. Automated 3D Structure Composition for Large RNAs. *Nucleic Acids Res.* **2012**, *40*, e112. [[CrossRef](#)]
58. Hofacker, I.L. Vienna RNA Secondary Structure Server. *Nucleic Acids Res.* **2003**, *31*, 3429–3431. [[CrossRef](#)]
59. Blom, N.; Sicheritz-Pontén, T.; Gupta, R.; Gammeltoft, S.; Brunak, S. Prediction of Post-Translational Glycosylation and Phosphorylation of Proteins from the Amino Acid Sequence. *Proteomics* **2004**, *4*, 1633–1649. [[CrossRef](#)]

# NMR Reveals Specific Tracts within the Intrinsically Disordered Regions of the SARS-CoV-2 Nucleocapsid Protein Involved in RNA Encountering

Letizia Pontoriero <sup>1,†</sup>, Marco Schiavina <sup>1,†</sup>, Sophie M. Korn <sup>2,†</sup>, Andreas Schlundt <sup>2,\*</sup>, Roberta Pierattelli <sup>1,\*</sup> and Isabella C. Felli <sup>1,\*</sup>

<sup>1</sup> Magnetic Resonance Center (CERM) and Department of Chemistry “Ugo Schiff”, University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino (Florence), Italy; pontoriero@cerm.unifi.it (L.P.); schiavina@cerm.unifi.it (M.S.)

<sup>2</sup> Institute for Molecular Biosciences, Center for Biomolecular Magnetic Resonance (BMRZ), Johann Wolfgang Goethe-University, Max-von-Laue-Str. 9, 60438 Frankfurt a. M., Germany; bochmann@bio.uni-frankfurt.de (S.M.K.)

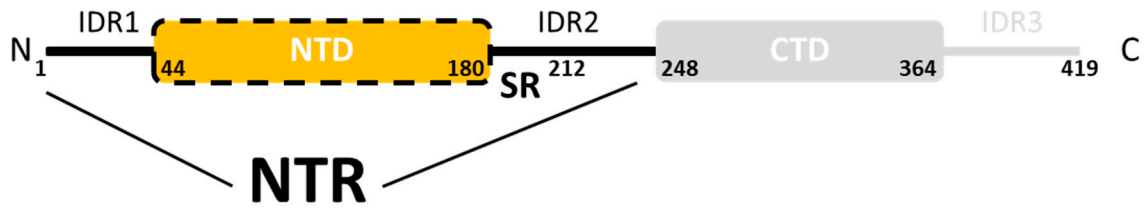
\* Correspondence: schlundt@bio.uni-frankfurt.de (A.S.); roberta.pierattelli@unifi.it (R.P.); felli@cerm.unifi.it (I.C.F.)

† These authors contributed equally to the work.

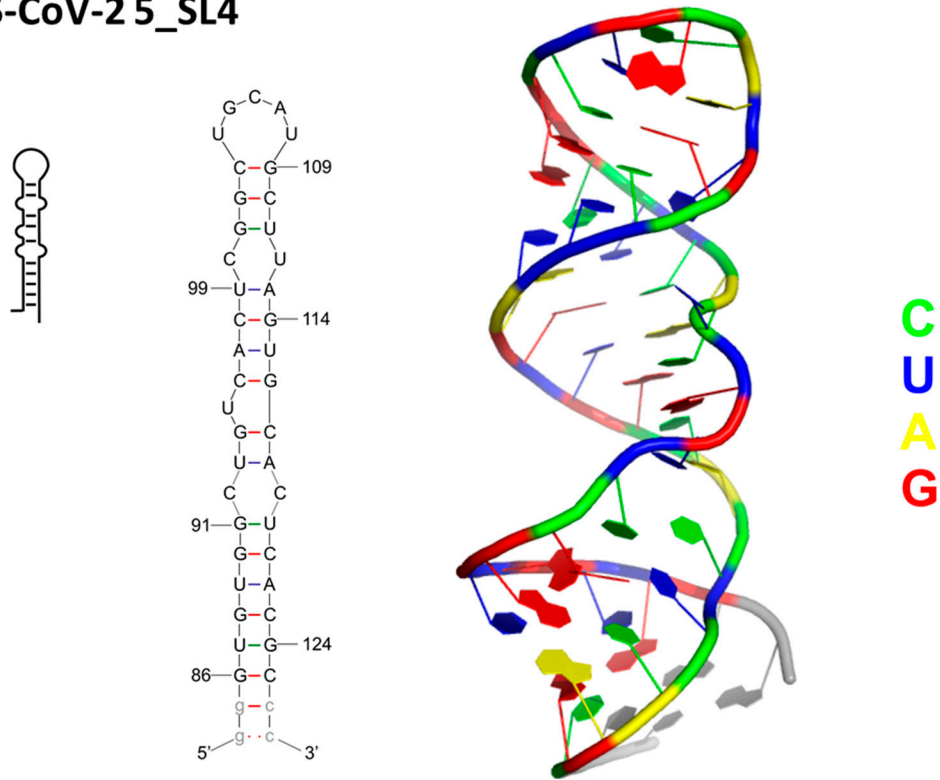


Supplementary Figures

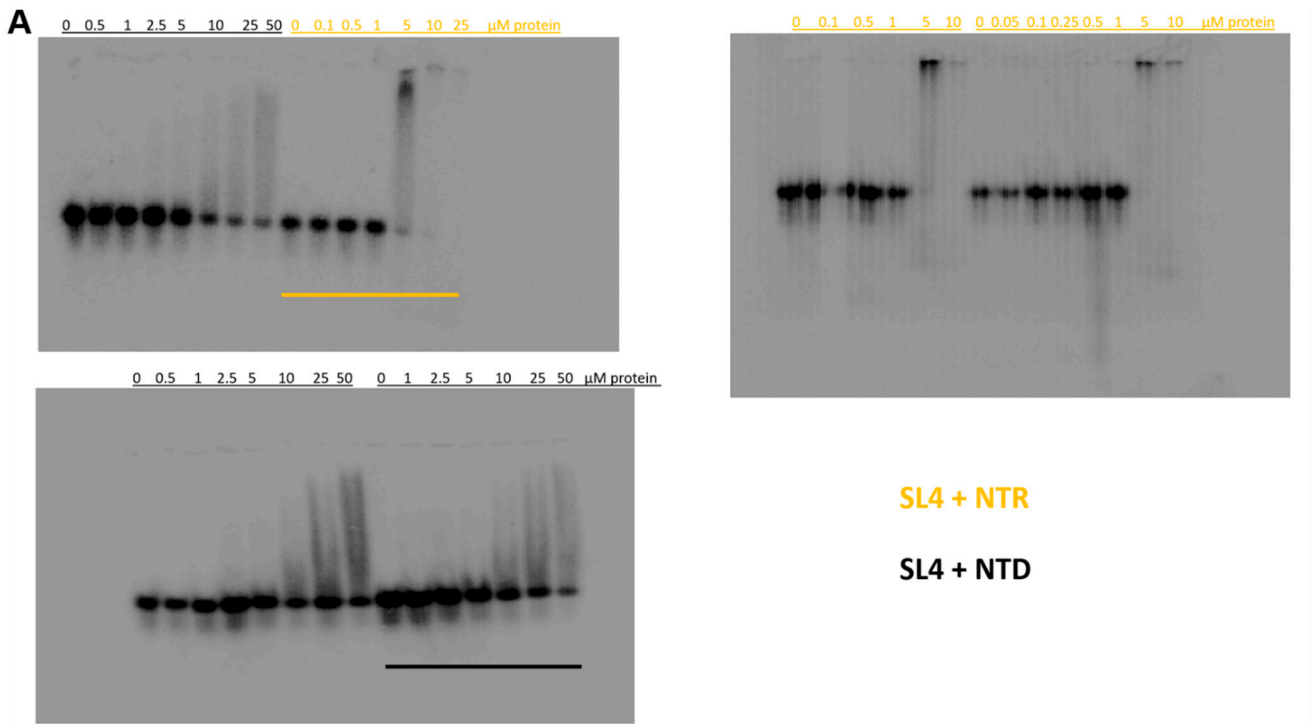
**A SARS-CoV-2 nucleocapsid protein (N)**



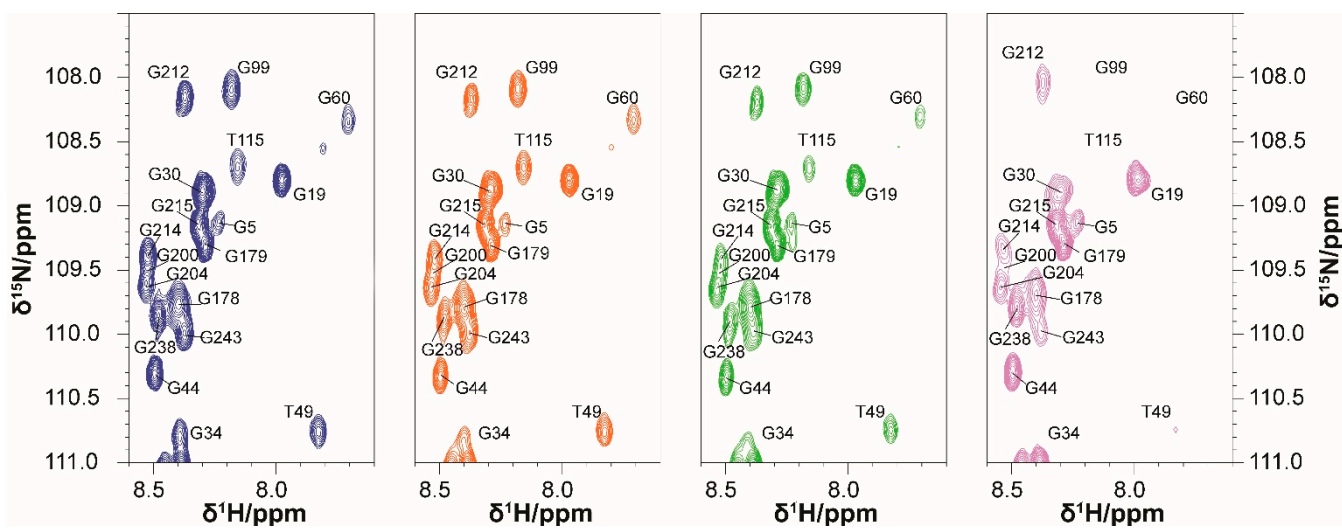
**B SARS-CoV-2 5\_SL4**



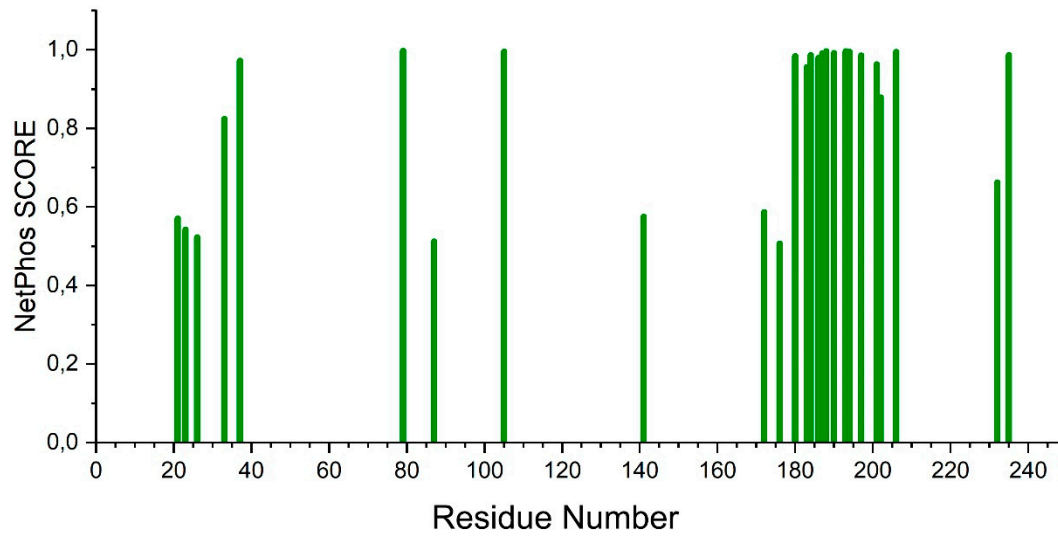
**Figure S1:** Overview of the molecular components used in this study. A) Domains organization of the SARS-CoV-2 full-length N protein. Numbers indicate the herein considered boundaries between IDRs and globular domains. Grey-shaded regions beyond residue 248 are not part of this study. The IDR2-embedded SR region is shown for convenience. B) Secondary structure and model of 5\_SL4 as obtained from RNAcomposer [57] and based on RNAfold [58]. Color-coded nucleotides are used to visualize the base distribution. The secondary structure is based on [3] including genomic numbering. Grey bases indicate artificial nucleotides in the construct used for RNA in vitro transcription.



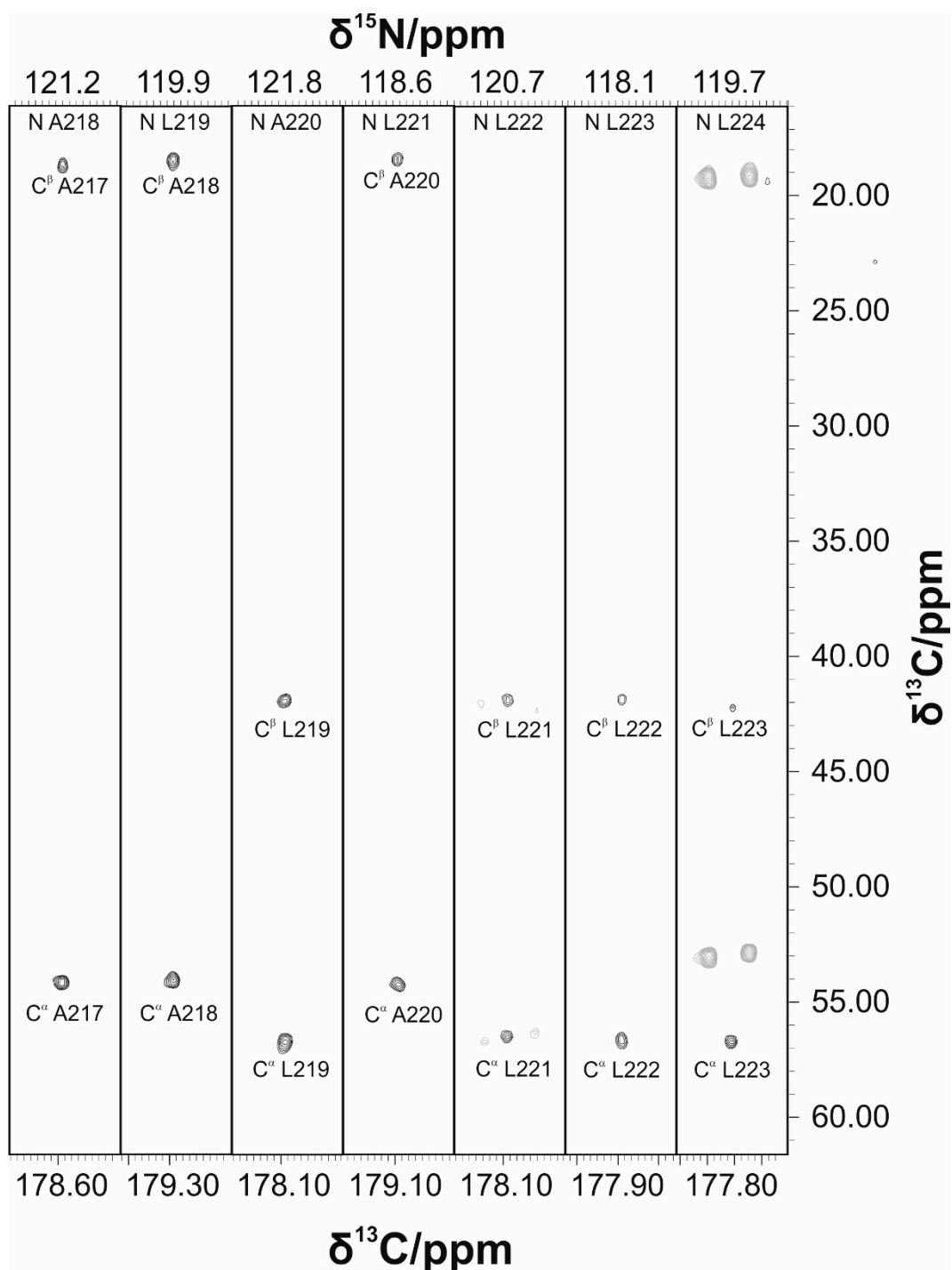
**Figure S2:** Interaction between NTR and NTD with SL4 RNA. Triplicates of EMSAs as shown in main text Figure 2. Presented are uncropped images as obtained from the Phosphoimager (see Methods). Concentrations of proteins added to SL4 RNA are given above respective lanes (orange, NTR and black, NTD). Bars indicate the replicates that have been used for the main figure panel, respectively.



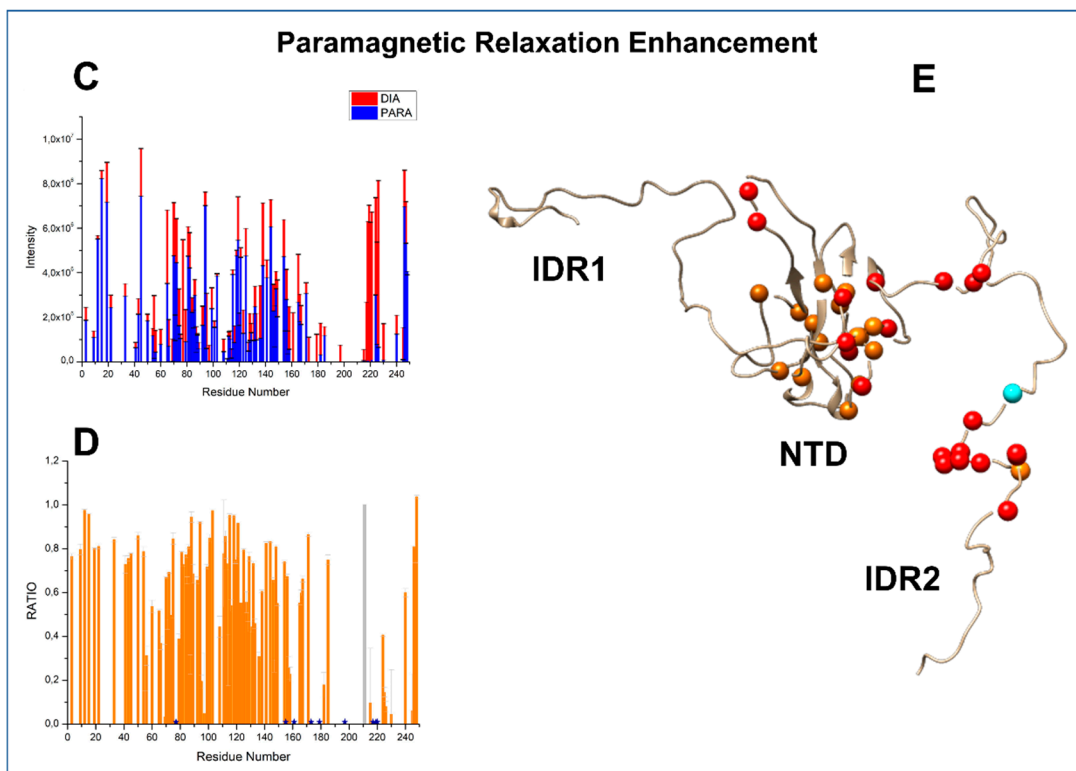
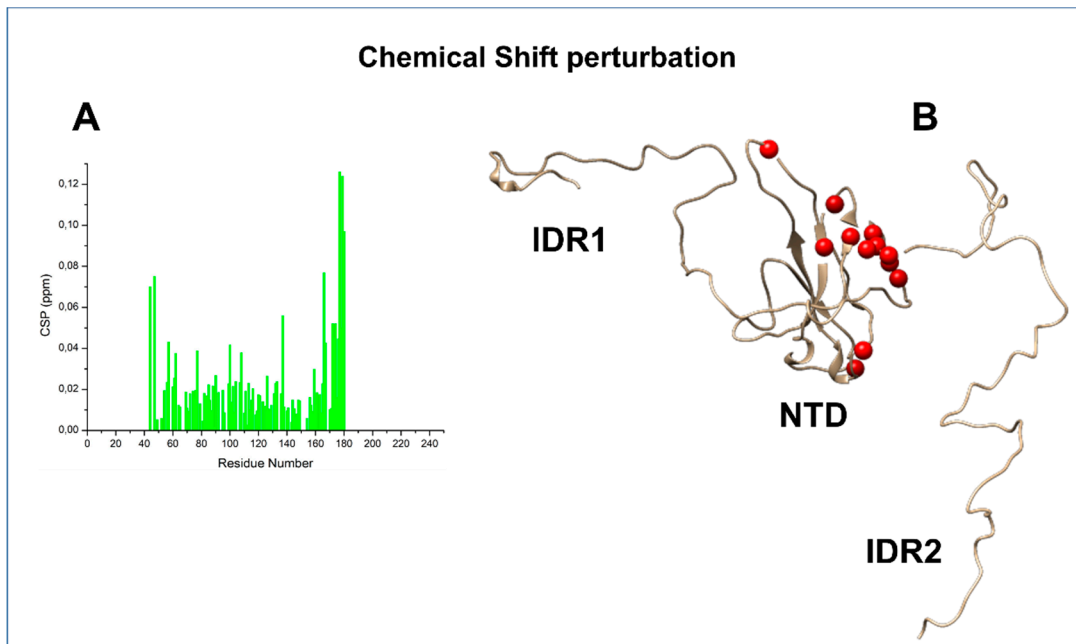
**Figure S3:** Series of 2D HN spectra of NTR upon addition of different RNA equivalents are reported in Blue (Reference), Orange (0.05 RNA equivalents), Green (0.10 RNA equivalents) and Pink (0.30 RNA equivalents). The overall picture displays negligible chemical shift perturbations, but a drastic decrease in intensity (see e.g., G99 and T49).



**Figure S4:** NetPhos [59] results for the serine residues of the NTR construct. Only the serine residues displaying a NetPhos score higher than 0.5 are reported.



**Figure S5:** Strips extracted from the 3D (H)CBCACON used to identify the resonances of the poly-Leucine stretch.  $^{13}\text{C}$ -Resonances of the fourth leucine residue in the four-amino acid sequence ( $^{221}\text{LLLL}^{224}$ ) were identified in the 2D CACO and 2D CBCACO experiments and assigned to Leu 224 by exclusion.



**Figure S6:** Chemical Shift Perturbation (CSP) and Paramagnetic Relaxation Enhancement (PRE) results.

Panel A reports the CSP values as obtained comparing the chemical shift values of  $H^N$  and  $N$  resonances of the NTR construct with those of the NTD alone. The residues with a CSP value higher than 0.03 (average + 1 standard deviation) are mapped on a protein model in panel B. The four terminal residues of NTD (44-47 and 177-180) were excluded as their chemical environment, and thus their CSPs, are mainly influenced by the absence vs. presence of the IDR residues themselves.

Panel C reports the intensity values as obtained from the diamagnetic (red) and the paramagnetic (blue) spectra of the A211C mutant. We decided to place the spin label at position 211, prior to the  $^{216}\text{DAALALLLLD}^{225}$  region but still quite distant from it to avoid perturbing such a crucial region, and also distant from the SR-rich region ( $^{177}\text{RGGSQASSRSSSRSRNSSRNSTPGSSR}^{203}$ ). The ratio between the two forms was calculated and reported in panel D against the residue number. The asterisks represent the residues whose peaks are broadened beyond detection in the paramagnetic spectrum. This ratio is mapped in panel E on a protein model. The residues displaying an intensity ratio between 0 to 25% are reported in red and those from 25.1% to 50% are reported in orange. The position of the spin label is depicted in cyan.

## References

3. Wacker, A.; Weigand, J.E.; Akabayov, S.R.; Altincekic, N.; Bains, J.K.; Banijamali, E.; Binas, O.; Castillo-Martinez, J.; Cetiner, E.; Ceylan, B.; et al. Secondary Structure Determination of Conserved SARS-CoV-2 RNA Elements by NMR Spectroscopy. *Nucleic Acids Res.* **2020**, *48*, 12415–12435, doi:10.1093/nar/gkaa1013.
57. Popena, M.; Szachniuk, M.; Antczak, M.; Purzycka, K.J.; Lukasiak, P.; Bartol, N.; Blazewicz, J.; Adamiak, R.W. Automated 3D Structure Composition for Large RNAs. *Nucleic Acids Res.* **2012**, *40*, e112–e112, doi:10.1093/nar/gks339.
58. Hofacker, I.L. Vienna RNA Secondary Structure Server. *Nucleic Acids Res.* **2003**, *31*, 3429–3431, doi:10.1093/nar/gkg599.
59. Blom, N.; Sicheritz-Pontén, T.; Gupta, R.; Gammeltoft, S.; Brunak, S. Prediction of Post-Translational Glycosylation and Phosphorylation of Proteins from the Amino Acid Sequence. *Proteomics* **2004**, *4*, 1633–1649, doi:10.1002/pmic.200300771.

## **Article 2.4:**

**The role of disordered Regions in orchestrating the properties of multidomain proteins: the SARS-CoV 2 Nucleocapsid protein and its interaction with Enoxaparin**



## Article

# The Role of Disordered Regions in Orchestrating the Properties of Multidomain Proteins: The SARS-CoV-2 Nucleocapsid Protein and Its Interaction with Enoxaparin

Marco Schiavina <sup>†</sup>, Letizia Pontoriero <sup>†</sup> , Giuseppe Tagliaferro, Roberta Pierattelli <sup>\*</sup>  and Isabella C. Felli <sup>\*</sup>

Magnetic Resonance Center (CERM) and Department of Chemistry “Ugo Schiff”, University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

<sup>\*</sup> Correspondence: roberta.pierattelli@unifi.it (R.P.); felli@cerm.unifi.it (I.C.F.)<sup>†</sup> These authors contributed equally to the work.

**Abstract:** Novel and efficient strategies need to be developed to interfere with the SARS-CoV-2 virus. One of the most promising pharmaceutical targets is the nucleocapsid protein (N), responsible for genomic RNA packaging. N is composed of two folded domains and three intrinsically disordered regions (IDRs). The globular RNA binding domain (NTD) and the tethered IDRs are rich in positively charged residues. The study of the interaction of N with polyanions can thus help to elucidate one of the key driving forces responsible for its function, i.e., electrostatics. Heparin, one of the most negatively charged natural polyanions, has been used to contrast serious cases of COVID-19 infection, and we decided to study its interaction with N at the molecular level. We focused on the NTR construct, which comprises the NTD and two flanking IDRs, and on the NTD construct in isolation. We characterized this interaction using different nuclear magnetic resonance approaches and isothermal titration calorimetry. With these tools, we were able to identify an extended surface of NTD involved in the interaction. Moreover, we assessed the importance of the IDRs in increasing the affinity for heparin, highlighting how different tracts of these flexible regions modulate the interaction.

**Keywords:** SARS-CoV-2; COVID-19; IDP; viral proteins; enoxaparin; NMR



**Citation:** Schiavina, M.; Pontoriero, L.; Tagliaferro, G.; Pierattelli, R.; Felli, I.C. The Role of Disordered Regions in Orchestrating the Properties of Multidomain Proteins: The SARS-CoV-2 Nucleocapsid Protein and Its Interaction with Enoxaparin. *Biomolecules* **2022**, *12*, 1302. <https://doi.org/10.3390/biom12091302>

Academic Editors: Elisar Barbar, Nathalie Sibille and Vladimir N. Uversky

Received: 27 July 2022

Accepted: 11 September 2022

Published: 15 September 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Since the COVID-19 pandemic impacted our lives, the development of novel and robust pharmacological strategies to contrast the SARS-CoV-2 virus became a priority worldwide. This pushed biomedical researchers to explore different alternatives to face the spreading of the infection [1]. The main results were the development of innovative mRNA-based vaccines and the use of monoclonal antibodies as a therapy [2,3]. These techniques have been fundamental to smoothing out the emergency. Nevertheless, the circulation of the virus is not over yet, and drug discovery studies are continuously in progress.

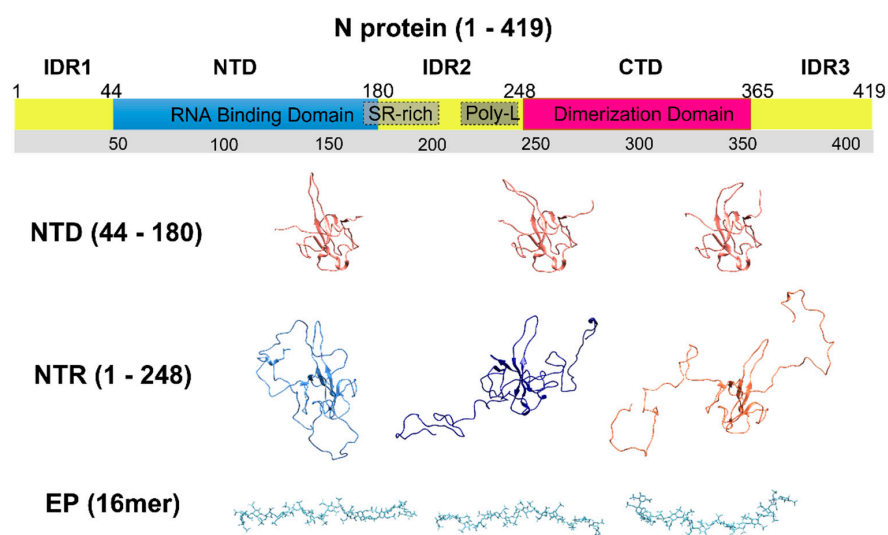
Nowadays, the most common pharmaceutical approaches target the Spike protein (S) [4], which is the access key to the host's cells. It is strongly affected by mutations [5,6], some of which are of concern since they affect the transmissibility and antigenicity of the disease. However, other viral proteins have emerged as potential drug candidates and they are now under investigation [7,8]. One of the most promising targets is the Nucleocapsid protein (N), the most expressed protein within the SARS-CoV-2 proteome [9].

N shares 90% of its homology with related proteins from other coronaviruses, and its mutations occur in limited regions of the sequence [8,10]. The main function of N is to package genomic RNA, but it is also involved in pivotal mechanisms for the viral replication cycle [11]. This multi-functional role is possible thanks to the modular organization of its structure (Figure 1). N is composed of two folded domains (N-terminal Domain, NTD, and C-terminal Domain, CTD) and three intrinsically disordered regions (IDR1, IDR2, and IDR3) [12–15]. These latter portions are necessary both for the formation

of the RiboNucleoProtein (RNP) complex and for recruiting partners necessary for the transcription of the viral genome [16–19]. While the importance of the flexible regions for N protein function has long been recognized [16], their atomic resolution investigation still remains a challenge, in particular, when IDRs are part of a multidomain protein [20]. NMR resonance assignments of the first two IDRs (IDR1 and IDR2) have recently become available [15,21], opening the way to the investigation at atomic resolution of their role in modulating protein function [17–19,22–25].

Both the globular RNA binding domain (NTD) and the tethered IDRs are rich in positively charged residues that drive the interaction between the N protein and its partners, like the negatively charged RNA fragments [17,19,22–25]. The study of the interaction of N with molecules that mimic nucleic acids' charge, such as polyanions, can thus help to elucidate one of the key driving forces responsible for its function, electrostatic contribution. The aim of this study is, thus, to investigate the interaction of N with one of the naturally occurring polyanions, heparin. This is a ubiquitous linear glycosaminoglycan (GAG), characterized by different degrees of sulfation which confer it a high negative charge. It is a component of the cell surface and of the extracellular matrix. It is also often used as a drug for its anticoagulant properties.

Low molecular weight heparin is used in clinical protocols to contrast serious cases of COVID-19 infection [26]. The literature reports about the interaction between heparin and the N protein detected in human samples, such as blood and saliva [27], and heparin-based resins have been used for hemofiltration in crucially ill COVID-19 patients, demonstrating a reduction in the N blood concentration after the treatment [28]. Furthermore, it was recently shown that N is not strictly confined to the cytosol but it is also found on the infected/transfected cells' surface, where it binds the heparin of the extracellular matrix [29]. Other viral RNA-binding proteins were found to adopt similar mechanisms to the N one [30,31]; it is suggested that N exploits these properties to interfere with the binding of cytokines to the GAGs.



**Figure 1.** The scheme reported on top shows the modular organization of the nucleocapsid protein [12–15]. The IDRs are colored in orange (IDR1, IDR2, and IDR3), the NTD is in blue, and the CTD is in red. Some regions of IDR2 important for the discussion are also highlighted (SR-rich; Poly-L). The molecules studied in the present work (NTD, NTR, and EP) are illustrated below the scheme. Three NTD conformers from the 6YI3 [22] PDB entry were selected to show the structural heterogeneity adopted by some parts of NTD. Several NTR conformers were generated using the EOM software (version 3.0. EMBL, Hamburg, Germany) [32,33] based on the NTD conformers. Three of them are reported here to sketch the conformational space that can be sampled by the protein. The EP conformers were selected from the PDB entry 3IRI [34].

On these grounds, we studied the interplay between two different N protein constructs and enoxaparin (EP, 16mer, 4.5 kDa, Figure 1), a low molecular weight heparin. In particular, we focused on the N-terminal region of the protein using a construct that comprises residues 1-248 (IDR1-NTD-IDR2, referred to as NTR), and on the NTD construct, (residues 44-180). The two different protein constructs thus differ in the presence of the two positively charged disordered regions that are expected to be relevant for the protein behavior in binding highly negatively charged partners and are central in the present study. We characterized the NTD and NTR interaction with EP using different Nuclear Magnetic Resonance (NMR) approaches and we complemented the analysis with Isothermal Titration Calorimetry (ITC). The high-resolution mapping of the binding obtained in this work could help the design of tailored polyelectrolytes able to interfere with N protein function.

## 2. Materials and Methods

### 2.1. Protein Sample Preparation

The NTD and NTR samples were prepared as previously described [35] and briefly summarized hereafter.

The sequence of the NTD (44-180) was based on SARS-CoV-2 NCBI reference genome entry NC\_045512.2, identical to GenBank entry MN90894 [36]. The gene inserted into pET28a(+) containing an N-terminal His<sub>6</sub>-tag, a tobacco etch virus (TEV) cleavage site vector, was kindly provided by Prof. Fabio Almeida from the University of Rio De Janeiro. After proteolytic TEV cleavage, the produced 14.85 kDa protein does not contain any artificial residue.

Uniformly <sup>15</sup>N-labeled and <sup>13</sup>C, <sup>15</sup>N-labeled NTD was expressed in *E. coli* strain BL21 (DE3) in M9 minimal medium containing 1.0 g/L ammonium chloride (<sup>15</sup>NH<sub>4</sub>Cl) (Cambridge Isotope Laboratories, Tewksbury, MA, USA) and, for <sup>13</sup>C labeling, 3 g/L <sup>13</sup>C<sub>6</sub>-D-glucose (Eurisotop, Cambridge Isotope Laboratories, Tewksbury, MA, USA). Protein expression was induced at an Optical Density measured at 600 nm (OD<sub>600</sub>) of 0.7 with 0.2 mM isopropyl-beta-thiogalactopyranoside (IPTG) for 18 h at 16 °C. Cell pellets were resuspended in 50 mM 2-amino-2-(hydroxymethyl)-1,3-propanediol (TRIS) at pH 8.0, 500 mM sodium chloride (NaCl), 20 mM imidazole, 10% *v/v* glycerol, and a protease inhibitor cocktail (SIGMAFAST). The cells were disrupted by sonication. The supernatant was cleared by centrifugation for 30 min at 30,000 × *g* at 4 °C.

The cleared supernatant was passed over a Ni<sup>2+</sup>-NTA HisTrap HP (GE Healthcare, Chicago, IL, USA), and the His<sub>6</sub>-Trx-tag was cleaved overnight at 4 °C with 1:10 *v/v* of TEV protease:protein solution, while dialyzing into fresh buffer composed of 50 mM TRIS at pH 8.0, 500 mM NaCl, and 1 mM dithiothreitol (DTT). TEV protease and the cleaved tag were removed via a second Ni<sup>2+</sup>-NTA HisTrap HP. The fractions containing the pure NTD protein were determined by SDS-PAGE, pooled, and concentrated. Buffer exchange was performed through a PD-10 desalting column (GE Healthcare) or through dialysis, with a final buffer containing 25 mM potassium phosphate (KH<sub>2</sub>PO<sub>4</sub>/K<sub>2</sub>HPO<sub>4</sub>) 150 mM potassium chloride (KCl), and 0.02% sodium azide (NaN<sub>3</sub>) at pH 6.5.

For the NTR (1-248), the gene of the N protein construct comprising residues 1-248 was designed based on the boundaries determined from the SARS-CoV homologue. The codon-optimized gene was synthesized by Twist Bioscience and cloned into the pET29b(+) vector between NdeI and XhoI restriction sites.

Uniformly <sup>15</sup>N and <sup>13</sup>C, <sup>15</sup>N-labelled NTR protein was expressed in *E. coli* strain BL21 (DE3) following the Marley protocol [37]. The cells were grown in 1 L of Luria Bertani medium at 37 °C until OD<sub>600</sub> of 0.8. Then, the culture was transferred in 250 mL of labeled M9 minimal medium supplemented with 1.0 g/L <sup>15</sup>NH<sub>4</sub>Cl and, for <sup>13</sup>C labeling, 3.0 g/L of <sup>13</sup>C<sub>6</sub>-D-glucose. After 1 h of unlabeled metabolite clearance, the culture was induced with 0.2 mM IPTG at 16 °C for 18 h. The pellet was harvested and stored at −20 °C overnight. The cell pellet was then dissolved in 25 mM TRIS, 1.0 M NaCl, 10% *v/v* glycerol, and protease inhibitor cocktail (SIGMAFAST) at pH 8.0 and centrifuged at 30,000 × *g* for 50 min at 4 °C.

The soluble fraction was dialyzed overnight against a solution of 25 mM TRIS, pH 7.2, at 4 °C. The protein solution was then loaded onto a HiTrap SP FF 5 mL column and eluted with a 70% gradient of 25 mM TRIS and 1.0 M NaCl. Fractions containing the protein were pooled, concentrated, and loaded onto a HiLoad 16/1000 Superdex 75 pg column equilibrated with 25 mM (KH<sub>2</sub>PO<sub>4</sub>/K<sub>2</sub>HPO<sub>4</sub>), 150 mM KCl, and 0.02% NaN<sub>3</sub> at pH 6.5.

Regarding the NTD construct, <sup>1</sup>H detected experiments were acquired using a 500-μL-sample of 70 μM <sup>15</sup>N-labelled NTD protein. The titration was performed in 5 mm NMR tubes. Proper aliquots of a 22 mM stock solution of commercially available enoxaparin sodium salt (CLEXANE, Sanofi S.p.A.) were added to the protein solution to reach NTD:EP ratios of 1:0.01, 1:0.025, 1:0.10, 1:0.30, 1:0.60, 1:0.90, 1:1.20, 1:2.40, 1:4.80, 1:9.60, and 1:19.20.

Briefly, <sup>13</sup>C detected experiments were acquired using a 500-μL-sample of 200 μM <sup>13</sup>C-<sup>15</sup>N-labelled NTD protein. The titration was performed in 5 mm NMR tubes. Proper aliquots of the stock solution of EP were added to the protein solution to reach NTD:EP ratios of 1:0.10, 1:0.30, 1:0.45, 1:0.60, 1:0.9, 1:1.20, 1:2.4, and 1:4.8. Moreover, 2D HN experiments were also collected during this titration as control.

Furthermore, <sup>15</sup>N Relaxation experiments were acquired using a 500-μL-sample of 200 μM <sup>15</sup>N-labelled NTD protein. The same experiments were recorded after the addition of 1.2 EP equivalents.

Regarding the NTR construct, <sup>1</sup>H and <sup>13</sup>C detected experiments were acquired using a 500-μL-sample of 70 μM <sup>13</sup>C, <sup>15</sup>N-labelled NTR protein. The titration was performed in 5 mm NMR tubes. Proper aliquots of the stock solution of EP were added to a protein solution sample to reach NTR:EP ratios of 1:0.10, 1:0.30, and 1:1.00.

Moreover, <sup>1</sup>H detected experiments were repeated using a 500-μL-sample of 70 μM <sup>15</sup>N-labelled NTR protein. The titration was performed in 5 mm NMR tubes. Proper aliquots of the stock solution of EP were added to a protein solution sample to reach NTR:EP ratios of 1:0.01, 1:0.05, 1:0.10, 1:0.30, 1:0.60, 1:1.20, and 1:6.00.

Diffusion Orderd Spectroscopy (DOSY) experiments were acquired using a 500-μL-sample of 70 μM <sup>15</sup>N-labelled NTD protein. The same experiments were recorded after the addition of 9.6 EP equivalents.

## 2.2. NMR Experiments

The interaction between the N constructs and EP was followed at 298 K, exploiting a series of 2D HN HSQC [38], 2D HC HSQC [38,39], 2D CACO [40], 2D (H)CBCACO [40], 2D (HCA)CON [41], and mr\_HN//CON [42] experiments.

The following spectrometers (Bruker, Billerica, MA, USA) were used:

- a Bruker AVANCE III spectrometer operating at 950.20 MHz <sup>1</sup>H, 238.93 MHz <sup>13</sup>C, and 96.28 MHz <sup>15</sup>N frequencies, equipped with a cryogenically cooled probe head optimized for <sup>1</sup>H-direct detection (TCI). Namely, 950.
- a Bruker AVANCE NEO spectrometer operating at 700.06 MHz <sup>1</sup>H, 176.03 MHz <sup>13</sup>C, and 70.94 MHz <sup>15</sup>N frequencies equipped with a cryogenically cooled probe head optimized for <sup>13</sup>C-direct detection (TXO). Namely, 700C.
- a Bruker Avance NEO spectrometer operating at 700.13 MHz <sup>1</sup>H, 176.05 MHz <sup>13</sup>C, and 70.94 MHz <sup>15</sup>N equipped with a cryogenically cooled triple resonance probe head optimized for <sup>1</sup>H-direct detection (TXI). Namely, 700H.
- a Bruker AVANCE III-HD spectrometer operating at 600.13 MHz <sup>1</sup>H, 120.90 MHz <sup>13</sup>C, and 60.81 MHz <sup>15</sup>N frequencies equipped with a probe head optimized for <sup>1</sup>H-direct detection (TXI). Namely, 600.

Standard radiofrequency pulses were used. The decoupling of <sup>1</sup>H and <sup>15</sup>N was achieved with waltz65 and garp4 decoupling sequences, respectively [43,44]. All gradients employed had a smoothed square shape.

The 2D HN HSQC [38] experiments recorded to follow the titration of NTD and NTR with EP were acquired at 950. The carrier frequency for <sup>1</sup>H was set at 4.7 ppm; for <sup>15</sup>N, the carrier was set at 120 ppm for standard HN spectra and at 80 ppm for spectra tailored to detect the arginine side-chain's correlations.

The 2D HC HSQC [38,39], 2D CACO [40], 2D (H)CBCACO [40], and 2D (HCA)CON [41] experiments were acquired at 700C. Briefly,  $^{13}\text{C}$  pulses were centered at 176.7 ppm, 49.7 ppm, 45.7 ppm, and 122.7 ppm for the  $\text{C}'$ ,  $\text{C}^\alpha$ ,  $\text{C}^{\text{ali}}$ , and  $\text{C}^{\text{aro}}$  regions. Further,  $^{15}\text{N}$  pulses were given at 121.0 ppm. The  $^1\text{H}$  carrier was placed at 4.7 ppm. Q5- and Q3-shaped pulses [44] of durations of 300 and 231  $\mu\text{s}$ , respectively, were used for  $^{13}\text{C}$  band-selective  $\pi/2$  and  $\pi$  flip angle pulses, except for the  $\pi$  band-selective pulses on the  $\text{C}^\alpha$  region (Q3, 1200  $\mu\text{s}$ ) and for the adiabatic  $\pi$  pulse to invert both  $\text{C}'$  and  $\text{C}^\alpha$  (smoothed chirp 500  $\mu\text{s}$ , 20% smoothing, 80 kHz sweep width, 11.3 kHz radio frequency field strength).

The interaction between  $^{13}\text{C}$ - and  $^{15}\text{N}$ -labelled NTR and EP were followed, exploiting a series of  $\text{mr\_CON//HN}$  [42] experiments acquired at 700C. The  $^{13}\text{C}$  pulses were centered at 176.7 ppm and 55.9 ppm for  $\text{C}'$  and  $\text{C}^\alpha$ . Further,  $^{15}\text{N}$  pulses were centered at 122.5 ppm for the CON experiment and at 118 ppm for the HN one. The  $^1\text{H}$  carrier, shapes, and duration of the  $^{13}\text{C}$  selective pulses were the same as reported for the experiments acquired on NTD.

The  $\text{mr\_CON//HN}$  [42] was acquired with an interscan delay of 1.9 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 [45] virtual decoupling. In the  $\text{mr\_CON//HN}$  [42] experiment, solvent suppression was achieved through the 3:9:19 pulse scheme [46].

The acquisition parameters are reported in Table 1.

**Table 1.** Acquisition parameters for the recorded spectra.

Construct	Experiment	Data Points		Spectral Width (Hz)		Number of Scans	Interscan Delay (s)	Field ( $^1\text{H}$ MHz)
		F1	F2	F1	F2			
NTD	2D CACO	128	1024	7407 ( $^{13}\text{C}^\alpha$ )	5263 ( $^{13}\text{C}'$ )	32	1.6	700
NTD	2D (H)CBCACO	174	1024	11,628 ( $^{13}\text{C}^{\text{ali}}$ )	5263 ( $^{13}\text{C}'$ )	32	1.0	700
NTD	2D (HCA)CON	128	1024	3413 ( $^{15}\text{N}$ )	5000 ( $^{13}\text{C}'$ )	96	1.1	700
NTD	2D HC	256	1024	10,638 ( $^{13}\text{C}^{\text{aro}}$ )	11,364 ( $^1\text{H}$ )	4	1.1	700
NTD	2D HN	256	2048	4347 ( $^{15}\text{N}$ )	19,132 ( $^1\text{H}$ )	16	1.0	950
NTD	2D $\text{H}^\epsilon\text{N}^\epsilon$	256	4096	11,627 ( $^{15}\text{N}^\epsilon$ )	19,132 ( $^1\text{H}^\epsilon$ )	8	1.0	950
NTR	$\text{mr\_CON//HN}$	400	1024	2840 ( $^{15}\text{N}$ )	5263 ( $^{13}\text{C}$ )	16	1.9	700
NTR	<b><math>\text{mr\_CON//HN}</math></b>	400	4096	3195 ( $^{15}\text{N}$ )	20,833 ( $^1\text{H}$ )	32	1.9	700

For the  $\text{mr\_CON//HN}$ , the experiment to which the parameters are referred is in bold.

To complete the available NTD assignment (BMRB 34511 [22]), a 3D (H)CBCACON experiment [41] was also performed on a 450  $\mu\text{M}$   $^{13}\text{C}$ ,  $^{15}\text{N}$  NTD sample at 950. Pulses were centered at 176.2 ppm, 56.1 ppm, 45.7 ppm, 122.0 ppm, and 4.7 for  $\text{C}'$ ,  $\text{C}^\alpha$ ,  $\text{C}^{\text{ali}}$ , N, and H regions, respectively. Q5- and Q3-shaped pulses [44] of durations of 259 and 162  $\mu\text{s}$ , respectively, were used for  $^{13}\text{C}$  band-selective  $\pi/2$  and  $\pi$  flip angle pulses, except the adiabatic  $\pi$  pulse to invert both  $\text{C}'$  and  $\text{C}^\alpha$  (smoothed chirp 500  $\mu\text{s}$ , 20% smoothing, 80 kHz sweep width, 11.3 kHz radio frequency field strength).

The 3D (H)CBCACON was acquired with an interscan delay of 1.1 s. This spectrum was acquired with 16 scans, with sweep widths of 9566 Hz ( $^{13}\text{C}'$ )  $\times$  4830 Hz ( $^{15}\text{N}$ )  $\times$  19,118 Hz ( $^{13}\text{C}^{\text{ali}}$ ) and 1024  $\times$  64  $\times$  96 real points in the three dimensions, respectively. The obtained resonances' assignment is reported in Supplementary Table S1 and deposited in BMRB (51,620) together with the rest of the assignment obtained in our experimental condition.

The NMR experiments to determine the  $^{15}\text{N}$  relaxation values [38,47] ( $^{15}\text{N}$   $R_1$ ,  $^{15}\text{N}$   $R_2$ , and  $^1\text{H}$ - $^{15}\text{N}$  NOEs) were recorded at 700H. The  $^{15}\text{N}$   $R_1$  and  $R_2$  experiments were performed using the standard Bruker pulse sequences, with 16 scans and sweep widths of 10,869 Hz



( $^1\text{H}$ )  $\times$  2551 Hz ( $^{15}\text{N}$ ) acquiring  $2048 \times 192$  real points in the two dimensions. A relaxation delay of 3.0 s has been used. To determine the  $^{15}\text{N}$   $R_1$  values, the following delays were used: 20 ms, 100 ms, 200 ms, 300 ms, 400 ms, 500 ms, 600 ms, 800 ms, 1000 ms, 1200 ms, 1500 ms, and 2000 ms. The 200 ms point was acquired twice for statistical analysis. To determine the  $^{15}\text{N}$   $R_2$  values, the following delays were used: 16 ms, 32 ms, 48 ms, 64 ms, 80 ms, 96 ms, 112 ms, 128 ms, 160 ms, 192 ms, 240 ms, and 320 ms. The 32 ms point was acquired twice for statistical analysis. The  $^1\text{H}$ - $^{15}\text{N}$  NOE experiments were performed with 96 scans with sweep widths of 10,869 Hz ( $^1\text{H}$ )  $\times$  2551 Hz ( $^{15}\text{N}$ ) and  $2048 \times 128$  real points in the two dimensions. A relaxation delay of 6.0 s was used.

DOSY experiments were performed at 600. The stimulated echo version [48] has been exploited using bipolar gradient pulses for diffusion. Solvent suppression was achieved through the 3:9:19 pulse scheme [46]. Both the experiments conducted in the presence and absence of 1.2 equivalents of EP were acquired with an interscan delay of 3.8 s. The gradient distance  $\Delta$  was set to 150 ms, and the bipolar gradient length  $\delta$  was set to 3 ms. The gradient ramp was linear with 128 steps applying a gradient strength from 2% to 95%, with a full power strength of 5.65 G/mm.

All the spectra were processed with TopSpin 4.0.6 and analyzed using CARA [49] and its tool, NEASY [50].

Chemical shifts were referenced using the  $^1\text{H}$  and  $^{13}\text{C}$  shifts of DSS. The  $^{15}\text{N}$  chemical shifts were referenced indirectly [51].

### 2.3. $K_d$ Estimation

The dissociation constant ( $K_d$ ) for the interaction between the two N constructs and EP was determined through NMR spectroscopy measuring the variation of chemical shift for each peak in a series of  $^1\text{H}$ - $^{15}\text{N}$  HSQC spectra recorded at increasing concentrations of EP. The data were fitted using the following equation:

$$\frac{\Delta_{obs}}{\Delta_{max}} = \frac{C_P + C_{EP} + K_d - \sqrt{(C_P + C_{EP} + K_d)^2 - 4C_P \cdot C_{EP}}}{2C_P}$$

where  $\Delta_{obs}$  is the observed chemical shift perturbation at the different titration points,  $\Delta_{max}$  is the maximum value obtained at the end of the titration,  $C_P$  is the total protein concentration (NTD or NTR),  $C_{EP}$  is the EP concentration at the different titration points, and  $K_d$  is the dissociation constant.

The CSP values of those peaks displaying a perturbation higher than the average were used as inputs in the calculation to estimate the  $K_d$ . The residues used to calculate the  $K_d$  for the NTD construct were A50, T57, R92, G96, G97, K102, W108, T166, Y172, and A173. The residues used to calculate the  $K_d$  for the NTR construct were K38, L45, S176, S180, S183, S194, and T205.

A  $K_d$  for the NTR:EP was obtained from isothermal calorimetry (ITC) as well. An NTR sample of 30  $\mu\text{M}$  was dialyzed overnight against the working buffer (25 mM  $\text{KH}_2\text{PO}_4/\text{K}_2\text{HPO}_4$ , 150 mM KCl, pH 6.5). The same buffer was used to prepare a batch of EP 300  $\mu\text{M}$  that was used to titrate the protein. Measurements were carried out with a VP-ITC microcalorimeter instrument (MicroCal, Inc., GE Healthcare, Chicago, IL, USA) at 298 K and analyzed using the ITC version of Origin 7.0 with embedded calorimetric fitting routines.

### 2.4. Protein-Ligand Docking

We performed the molecular docking of EP and NTD using the HADDOCK server (version 2.4 Bonvin Lab, Utrecht, The Netherlands) [52,53]. The protein structural coordinates used as input were obtained by selecting one of the models deposited in the Protein Data Bank (PDB) under access code 6YI3 [22]. Protonation states of histidine residues 59 and 145 at pH 7.0 were set accordingly to the HADDOCK standard protocol.

The EP structural coordinates have been derived from the PDB under the access code 3IRI [34]. We selected one of the models, properly renumbering and renaming the different atoms to encode 10 monomers according to HADDOCK's formalism.

The protein active residues were selected as those showing a CSP upon interaction with EP, taking into consideration all the acquired spectra (49, 50, 56, 57, 58, 59, 60, 62, 63, 88, 92, 93, 94, 96, 97, 98, 99, 100, 101, 102, 103, 104, 107, 108, 154, 162, 165, 166, 167, 169, 172, 173, and 174). The passive residues were automatically selected by the HADDOCK server.

In addition, NTD's flexible region composing the "finger" (92–106) was defined as a fully flexible segment for the advanced stages of the docking calculation.

In total, 1000 complex structures of rigid-body docking were calculated by using the standard HADDOCK protocol with an optimized potential for liquid simulation parameters (OPLSX). The final 200 lowest-energy structures were selected for subsequent explicit solvent (water) and semi-flexible simulated annealing.

The final structures were clustered using the fraction of common contacts (FCC) with a cutoff of 0.6 and a minimal cluster size of five.

The 191 resulting structures were sorted into six clusters and the one with the best HADDOCK score was selected for the analysis as discussed later. The latter is composed by 130 structures (68%) while the other clusters contain, respectively 26 (13%), 9 (5%), 14 (7%), 7 (4%), and 5 (3%) structures.

### 3. Results

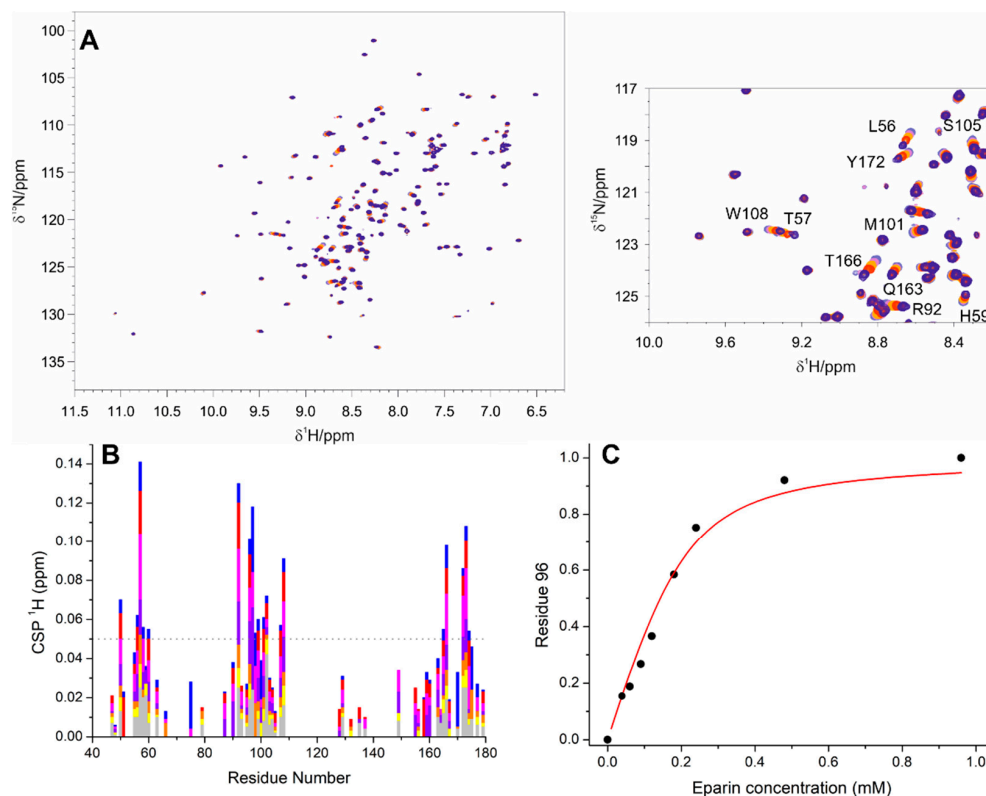
Different NMR approaches were used to focus on the globular domain (NTD) and on the disordered regions (IDRs) present in the NTR construct. These allowed us to achieve atom-resolved information on the interaction with EP, as described in detail hereafter.

#### 3.1. The Interaction of EP with NTD

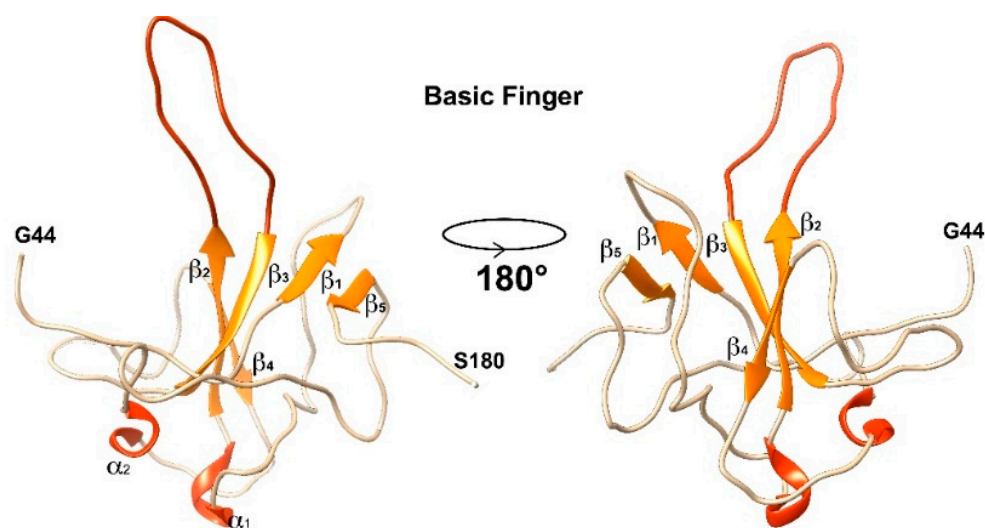
Two-dimensional NMR spectra were used to identify at atomic resolution, which are the regions of the protein that are perturbed upon the addition of increasing amounts of EP. As a first step to characterize this interaction, we decided to focus on the NTD construct following changes in the 2D  $^1\text{H}$ - $^{15}\text{N}$  HSQC NMR spectra (2D HN hereafter) upon the addition of EP. The results are reported in Figure 2A. The interaction is in a fast exchange regime on the NMR time scale, and the observed spectral changes upon the addition of up to 4.8 equivalents of EP to the protein are plotted in Figure 2B. Monitoring the  $^1\text{H}$  chemical shift values upon titration (Figure 2B) allowed us to estimate a dissociation constant ( $K_d$ ) of  $44 \pm 9 \mu\text{M}$  (G96; Figure 2C).

As extensively discussed in the literature [22,24] and shown in Figure 3, NTD is organized into five  $\beta$ -strands ( $\beta 1$ ,  $\beta 2$ ,  $\beta 3$ ,  $\beta 4$ , and  $\beta 5$ ), two short  $\alpha$ -helices ( $\alpha 1$  and  $\alpha 2$ ), and a flexible hairpin. The secondary structural elements  $\beta 2$ – $\beta 3$  compose the core of the protein fold, very rich in aromatic residues, and extend into the flexible hairpin (the "finger"), rich in positively charged residues. The antiparallel  $\beta$ -sheet formed by  $\beta 1$ – $\beta 5$  is, instead, a junction between the two domain's ends.

Looking at the  $^1\text{H}^{\text{N}}$  chemical shift, the most perturbed residues upon EP interaction are clustered mainly in two regions: the basic finger (R92, G96, G97, D98, G99, M101, and K102) and the  $\beta 1$ – $\beta 5$  antiparallel sheet (L56, T57, Q58, G60, Y172, A173, and E174). Other few residues external to these regions (A50, R107, W108, T165, and T166) were also affected.



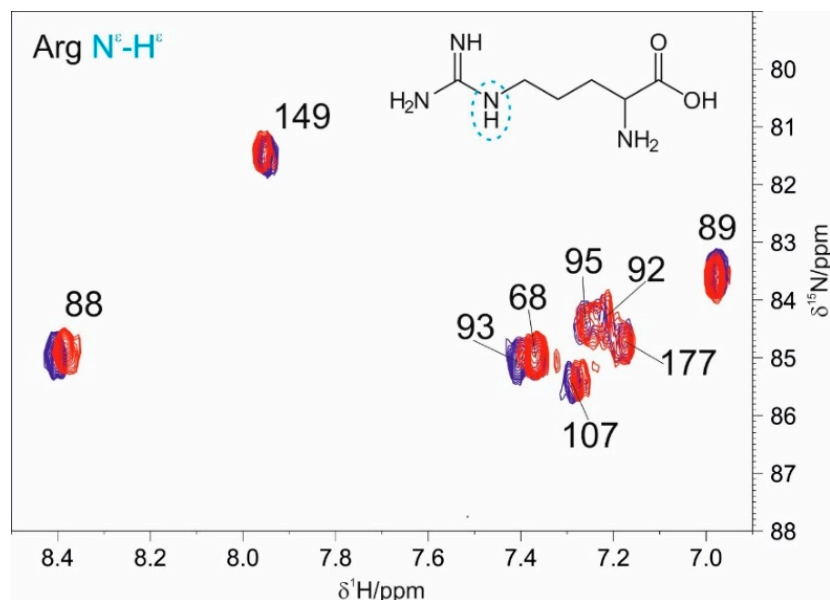
**Figure 2.** Panel (A) reports the overlay of NTD 2D HN spectra upon the addition of EP. Light blue, pink, orange, red, and blue represent the 1:0, 1:0.3, 1:0.6, 1:1.2, and 1:2.40 molar ratios of NTD:EP, respectively (the protein concentration was 200  $\mu$ M). A zoom in a spectral region where several peaks are perturbed is reported on the right. The assignment of the most perturbed peaks is shown. Panel (B) reports the variations in chemical shifts of  $^1\text{H}$  nuclei (CSP) against the residue number at 1:0.30, 1:0.45, 1:0.60, 1:0.90, 1:1.20, 1:2.40, and 1:4.80 of the NTD:EP ratios (grey, yellow, orange, violet, magenta red, and blue, respectively). Panel (C) reports the fittings and the obtained  $K_d$  values for G96.



**Figure 3.** Representation of the secondary structural elements forming the fold of the NTD.  $\beta$  sheets are reported in orange together with the loop composing the finger.  $\alpha$  helices are reported in red. The elements that are not comprised in any secondary structural element are reported in grey.



To assess the importance of positively charged residues in the interaction between NTD and EP, we also acquired a series of 2D HN-HSQC spectra centered in the region where the arginine side chain nitrogen nuclei are expected to resonate ( $\delta^{15}\text{N} \approx 80$  ppm). As reported in Figure 4, it is possible to observe that three out of nine  $\text{H}^\epsilon\text{-N}^\epsilon$  signals were found to be perturbed (R88, R93, and R107) upon the addition of 1.2 equivalents of EP to the protein.

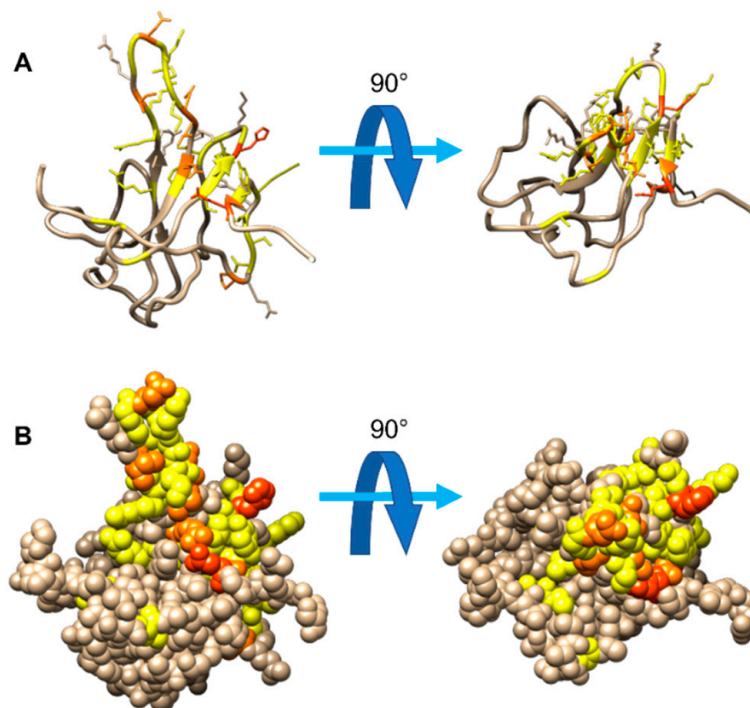


**Figure 4.** Overlay of 2D HN spectra recorded focusing on the arginine side chain region upon the addition of 1.2 equivalents of EP to NTD (blue reference, red addition of 1.2 equivalents of EP).

The protein fingerprinting can be expanded by performing a set of 2D  $^{13}\text{C}$ -detected NMR experiments (2D CON, 2D CACO, and 2D CBCACO) [54]. Preliminary to this, a 3D (H)CBCACON experiment [41] was also performed to complete the available assignment of NTD (BMRB: 34511 [22]). These experiments allowed us to assign 100% of the  $\text{C}^\alpha$ ,  $\text{C}^\beta$ ,  $\text{C}'$  99.2%  $\text{H}^\text{N}$  (G44 missing), and 99.2% N (including those from the 11 proline residues, 8% of the total protein composition, G44 missing) resonances in our experimental conditions. Regarding side chains, we assigned also 100% of the  $\text{H}^\epsilon$  and  $\text{N}^\epsilon$  from arginine residues, and 24 out of 25 resonances arising from the side chains of glutamate, glutamine, aspartate, and asparagine residues [40] (Supporting Table S1, BMRB 51620).

The analysis of the 2D CON spectrum shows a high heterogeneity in the intensities of the cross peaks (Figure S1). The most intense ones are those of the residues composing the initial and final protein's regions (45–50 and 175–180) as well as part of the finger (92–106). This provides a qualitative but firm indication of the high flexibility of the basic finger, almost comparable to the initial and final residues within this domain. Analysis of the chemical shift perturbations (CSP) induced by EP confirms the picture achieved through 2D HN, highlighting a few additional peaks (L45, T49, H59, I94, D103, L104, N154, P162, L167, L169, and A173). Most importantly, the flexibility of the mobile tracts is maintained in the complex as one can verify by the intensities of the cross peaks in these regions also when 1.2 EP equivalents are added (Figure S1). The CACO/CBCACO experiments [40] provide information also on the  $\text{C}^\beta$  and  $\text{C}^\alpha$  nuclei and on side chains containing carbonyl/carboxylate functional groups [40]. Major perturbations were identified for the residues, H59, I94, K100, L104, P162, and E174 ( $\text{C}^\beta$  and  $\text{C}^\alpha$  resonances). Interestingly,  $\text{H}^\delta\text{-C}^\delta$  and  $\text{H}^\epsilon\text{-C}^\epsilon$  of H59 were found to be perturbed also in the 2D  $^1\text{H}\text{-}^{13}\text{C}$  HSQC spectra acquired to monitor changes for the aromatic regions (data not shown). The region of carboxylate resonances of aspartate and glutamate residues in CACO spectra also shows interesting variations for the residues, E62, D63, D98, D103, and E174 (data not shown).

The collective analysis of these 2D NMR spectra provides a comprehensive view of the interaction of NTD with EP, reporting information on all backbone resonances and on selected side chain ones [40,54–56]. An overview of the most perturbed residues, considering all the analyzed spectra, is reported in Figure 5.



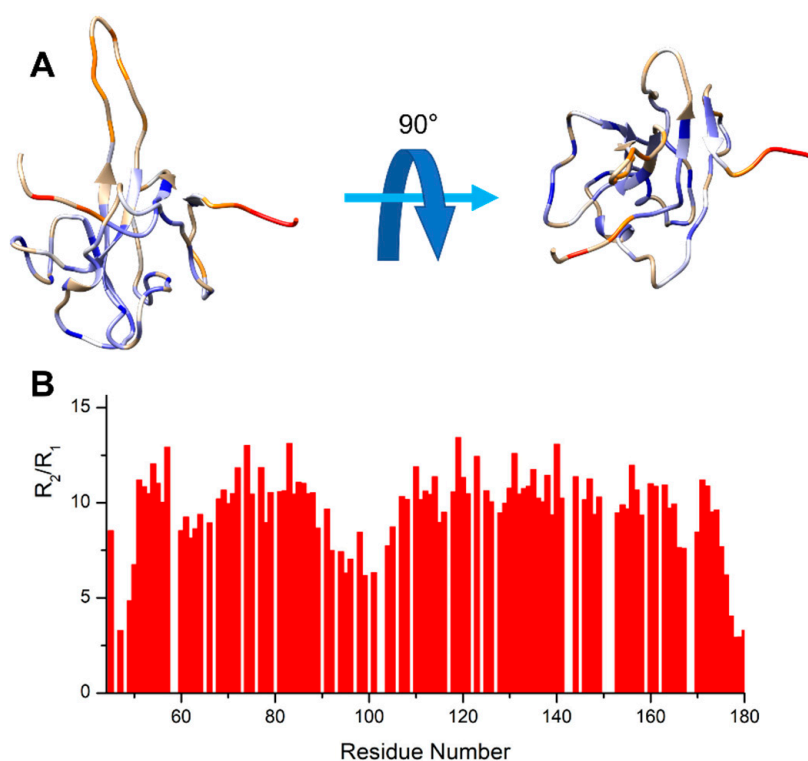
**Figure 5.** Mapping of the residues displaying the strongest perturbation in different 2D spectra (2D HN (backbone region), 2D HN (arginine region), 2D CON, 2D CACO, 2D CBCACO, and 2D HC (aromatic region)) at a molar ratio of 1:1.2 of NTD:EP. Panel (A) reports the protein in two different orientations, rotated by 90°, one with respect to the other; the heavy atoms of the perturbed residues are displayed as well. Panel (B) shows the same protein orientations with the models represented in a space-filling way. The color-coding is the following: residues which are found to be perturbed in a single experiment (yellow), in two experiments (orange), and in three or more experiments (red).

A more quantitative picture of the dynamic properties of NTD in the complex can be obtained through the analysis of the  $^{15}\text{N}$  relaxation rates ( $^{15}\text{N}$   $R_1$ ,  $^{15}\text{N}$   $R_2$ , and  $^1\text{H}$ - $^{15}\text{N}$  NOEs) for the isolated protein and upon the addition of 1.2 equivalents of EP (Figure S2).

The  $R_2/R_1$  ratio (Figure 6) provides an initial estimation of the global correlation time. These values are mapped on the protein 3D model and reveal a more rigid core of the protein fold (blue, high  $R_2/R_1$  values). On the other hand, several regions show higher flexibility (red, low  $R_2/R_1$  ratios). These comprise the finger (residues 92–106), a few external loops, and the residues at the edges of the construct, as also previously reported in the literature [24,57,58].

The  $R_2/R_1$  ratios can be used to estimate the local correlation time ( $\tau_r$ ), as described in [47]. Focusing on the residues in the globular protein fold core (the blue ones in Figure 6), these are characterized by an average  $R_2/R_1$  ratio value of 11.2 that provides a correlation time of 9 ns.

Upon interaction with EP, a homogeneous increase in the  $^{15}\text{N}$   $R_2$  and  $^1\text{H}$ - $^{15}\text{N}$  NOE values is observed along with a reduction in the  $^{15}\text{N}$   $R_1$  values (Figure S2). In this case, the  $R_2/R_1$  ratio of the most rigid portion of the protein is 16.9. These variations are consistent with slower tumbling due to an increased molecular mass, which corresponds to a correlation time of 11.5 ns. Notably, even upon interaction, the flexibility of the finger, the loops, and the edges is maintained with lower  $R_2/R_1$  ratio values with respect to the rest of the protein construct (Figure S2 Panel D).



**Figure 6.** The NTD protein construct from two different views is reported in panel (A). The protein construct is colored on the basis of the  $R_2/R_1$  values, which gives a first estimation of the correlation time. The residues with lower  $R_2/R_1$  values are reported in red, while those with higher values are reported in blue. Panel (B) shows the  $R_2/R_1$  values against the residue number.

Further evidence about the interaction can be achieved through DOSY experiments performed on the free and bound form of NTD, in the presence of 1.2 equivalents of EP (Figure S3). The obtained diffusion coefficients are  $D^{\text{FREE}}: 1.5 \pm 0.2 \cdot 10^{-10} \text{ m}^2/\text{s}$  and  $D^{\text{BOUND}}: 1.3 \pm 0.1 \cdot 10^{-10} \text{ m}^2/\text{s}$ . The smaller diffusion coefficient upon the addition of EP is in line with a reduced diffusion of the active species in solution.

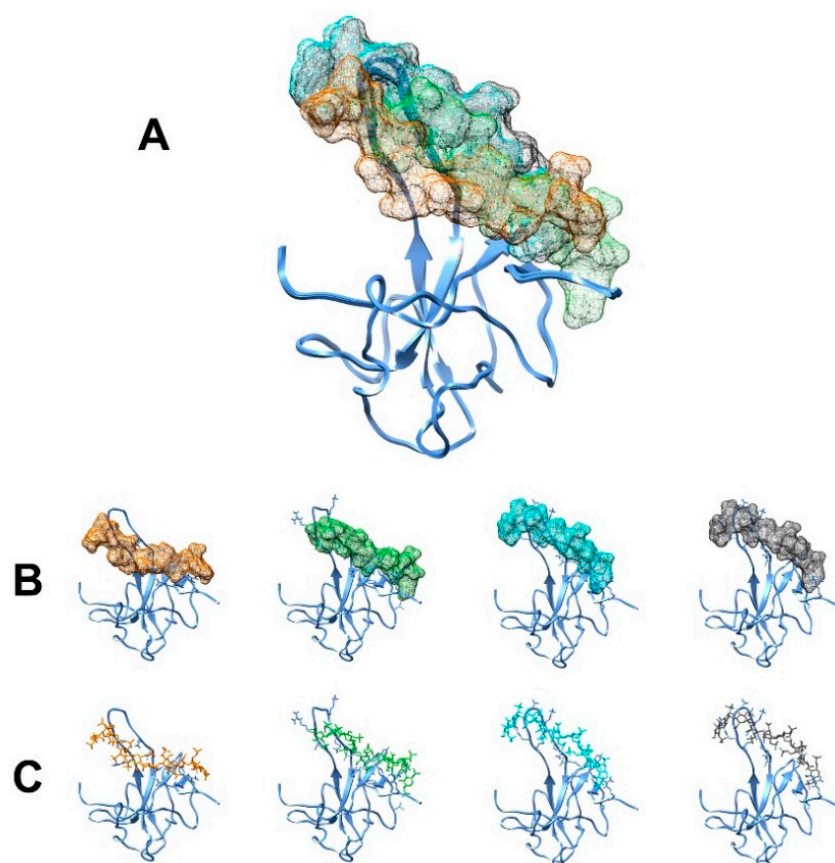
Collectively, these observations support the presence of a quite extended surface of the interaction of NTD with EP and that the flexibility of the finger is retained in the complex.

To visualize the possible scenarios, we performed a docking calculation between the NTD construct and the EP molecule using the HADDOCK server [52,53]. The active residues were identified from all the previously mentioned observed CSP values (see the Materials and Methods section for details).

Among the final 200 lowest-energy structures, 191 of them were divided into six clusters, with the one having the best HADDOCK score being selected for the analysis. This cluster, composed of 130 structures (68% of the total), possesses the best HADDOCK score ( $-49.7 \pm 3.1$ ) and provides the least violations of experimental restraints ( $144.2 \pm 37.5 \text{ Kcal} \cdot \text{mol}^{-1}$ ). The four best representative structures of this cluster are reported in Figure 7.

As can be seen from panel A of Figure 7, the EP seems to surround the protein from the side of the  $\beta_5$  and  $\beta_1$  sheets, being in contact also with the region of the flexible finger. Looking in detail at the four best structural models (Figure 7, panels B and C), the residues computed to contribute most to the interaction with EP are I94, R95, G96, G97, K102, L104, and Y172.

All the other clusters are found to have a lower HADDOCK score and higher violations of experimental restraints. In these clusters (Figure S4), the positive finger is always involved in the interaction. However, the core of the protein is computed to interact quite differently from cluster to cluster.

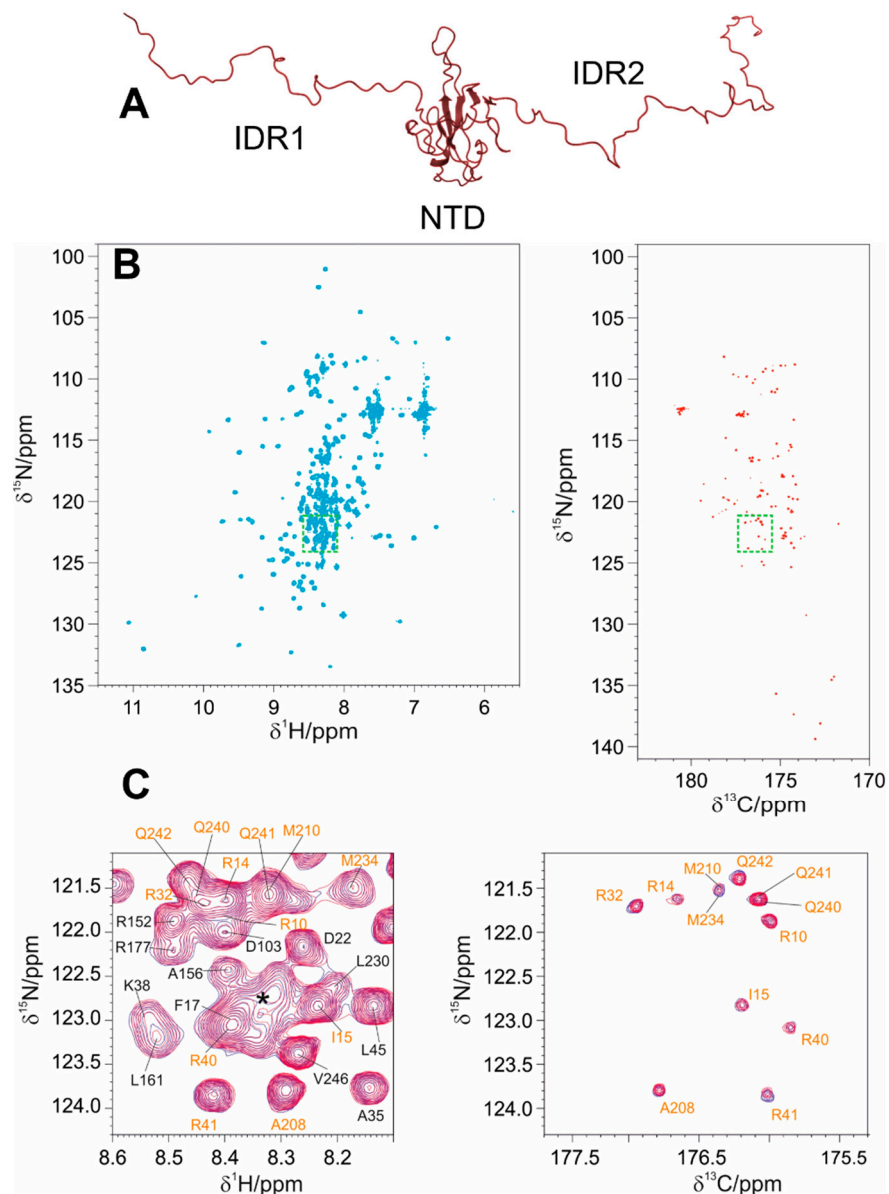


**Figure 7.** The results for the docking performed using HADDOCK are reported in the picture. The four structures derived from the best cluster are reported in Panels (A,B) in a superimposed and separate view, respectively. The protein structures are represented in the ribbon view while the mesh surface of EP is shown. The same four complexes with EP structure presented in stick view are displayed in Panel (C). Moreover, the side chains of NTD's residues computed to be in close contact with EP are also shown in Panels (B,C).

### 3.2. The Interaction of EP with NTR: The Role of the Intrinsically Disordered Regions

Previous studies demonstrated the importance of the IDRs in enhancing the interaction potential of the N protein with its partners, such as RNA [16,17,19,23,24,59–62]. RNA can be considered as a polymer composed both of a negatively charged component (phosphodiester backbone groups) and an aromatic component (base groups). EP is also a linear polyanion with a strong negative charge and, in principle, it might mimic the charge properties of the RNA backbone.

We thus decided to assess how the two disordered regions flanking the globular NTD domain influence the interaction with EP by exploiting the NTR construct (1-248, IDR1-NTD-IDR2). To this end we opted for the mr\_CON//HN [42] multiple-receiver NMR experiment, which allowed us to acquire two simultaneous NMR spectra of the protein, providing highly resolved information both for the globular domain and for the IDRs when part of the NTR construct (Figure 8). This experimental set-up is conceived to exploit the longitudinal recovery time necessary to restore the equilibrium of  $^{13}\text{C}$  magnetization for the 2D CON experiment to acquire the 2D HN FIDs needed for the 2D HN experiment. This experimental approach thus allows us to acquire the two spectra simultaneously, a key aspect to access experimental information on both globular domains and IDRs when part of a multidomain protein construct.



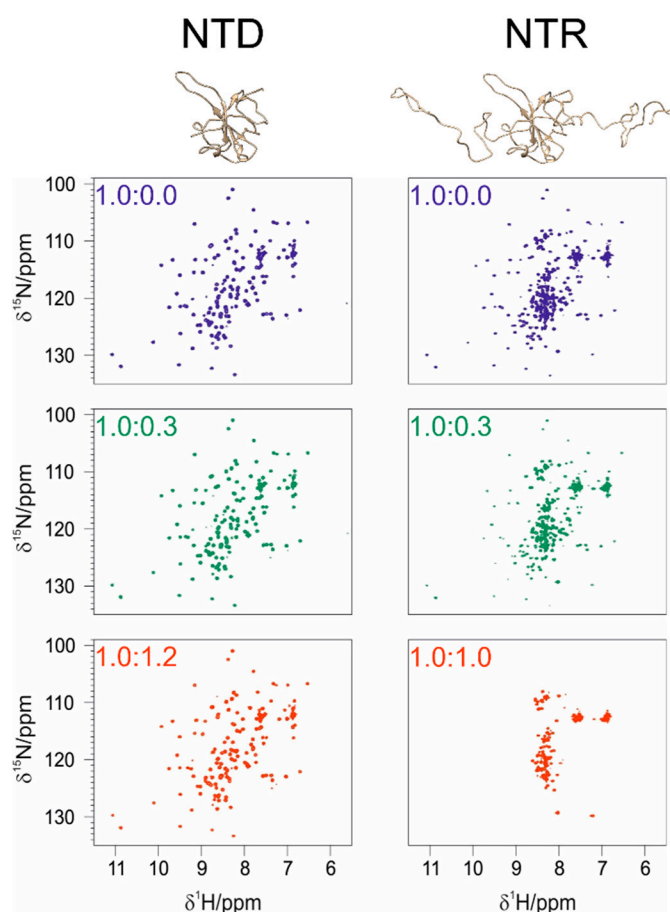
**Figure 8.** The picture reports the results of the mr\_CON/HN [42] experiment Panels (B,C) performed on the NTR construct Panel (A). The 2D HN spectrum is reported on the left in Panel (B); the 2D-CON spectrum is reported on the right in the same panel. Panel (C) shows a zoom of the superimposed spectra (HN in the left and CON on the right) acquired through the multiple receiver approach in the absence (blue) and with the addition of 0.1 equivalents of EP (magenta). These two regions are highlighted by green boxes in Panel (B). The CON provides superior resolution with respect to the HN one, which provides a higher number of signals, complicating the analysis.

From a more technical point of view, this strategy combines the sensitivity of the 2D HN experiment to pick up the signals arising from the globular domain with the high resolution provided by the 2D CON for the study of the IDRs. Indeed, this latter experiment acts as a relaxation filter that allows us to monitor the signals of the highly flexible regions in a clean way, enabling the study of IDRs within this modular construct rather than in isolation.

A comparison of the 2D HN spectra of NTD and NTR with a comparable protein:EP molar ratio is reported in Figure 9 and shows that the IDRs have a marked effect on EP binding. Focusing on the well-dispersed signals of the globular domain in the NTR construct, these show similar chemical shift perturbations as those observed when studying



the isolated NTD construct (Figure S5). However, a pronounced decrease in the intensities of the cross peaks of the globular domain is also observed even in the presence of low amounts of EP (1:0.3 NTR:EP, Figures 9 and S5). This leads to the complete disappearance of the cross-peaks from the globular domain at the molar ratio of 1:1, while a set of cross peaks deriving from the IDRs is still observed. The extensive broadening of the cross peaks of the globular domain is probably due to the increased molecular mass and structural heterogeneity of the NTR construct with respect to the NTD one, which implies a slower tumbling upon interaction, with the IDRs still retaining their flexibility [56]. Indeed, the addition of 110 flexible amino acids further increases the structural complexity of the protein. IDR1 and IDR2 highly extend the conformational space sampled by the protein. The occurrence of intermolecular interactions mediated by EP promoting the increase in the molecular mass cannot be ruled out [57].

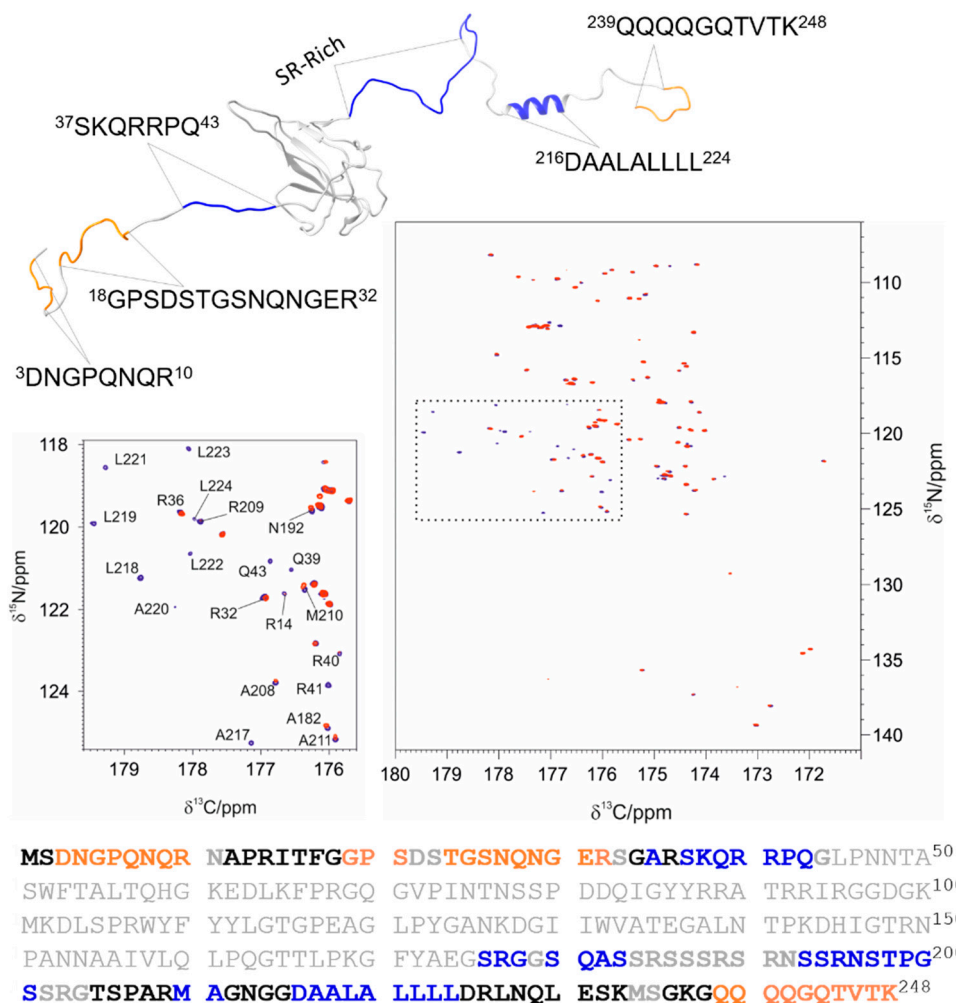


**Figure 9.** Comparison between the NTD and the NTR constructs at different molar ratios of protein:EP as reported in the top left corner of each spectrum.

It is interesting to inspect the perturbations sensed by the IDRs upon interaction with EP at the residue level. Most of the resonances arising from the IDRs fall in crowded regions in the 2D HN spectra, complicating the analysis (Figure 8 panel C). However, some cross peaks show measurable CSP values, as shown in Figure S6, including residues in the proximity of the NTD (e.g., S183 in Figure S6). The fitting of the CSP values measured for the few resolved peaks observed all along the titration provides  $K_d = 8 \pm 3 \mu\text{M}$  (L45 and S176). A value in the same range,  $K_d = 10.6 \pm 0.4 \mu\text{M}$ , was obtained by Isothermal Titration Calorimetry (ITC) (Figure S7) which was used to corroborate the NMR-derived result. The IDRs appear thus responsible for a higher affinity of the NTR construct for heparin.

Inspection of the CON allows us to clearly focus on the signals of the residues in the IDRs. Figure 10 shows an overlay of the CON spectra before and after the addition of 0.3 equivalents of EP. The spectra clearly show that a subset of cross peaks experiences a reduction in intensity

(a few cross peaks experience also minor chemical shift changes). A plot of the intensity ratios versus the residue number is shown in Figure S8. There are two main regions showing a significant decrease in signal intensities. The first is the portion  $^{36}\text{RSKQRRPQ}^{43}$ , whose signals completely disappear. A second interesting region is the so-called poly-Leu region, characterized by the residues  $^{216}\text{DAALALLL}^{224}$ . All the peaks that belong to this region disappear upon the addition of 0.3 EP equivalents. Several residues in the initial part of IDR2 are also perturbed. It is also interesting to note which residues are still observable upon the final addition of EP. These are mainly in the final region of IDR2 ( $^{239}\text{QQQGQTVTK}^{248}$ ) and two regions of IDR1 ( $^3\text{DNGPQNQR}^{10}$  and  $^{24}\text{TGSNQNGE}^{31}$ ).



**Figure 10.** The picture reports the superimposition of the two CON spectra of NTR: the spectrum before the addition of EP (blue) and the one acquired after the addition of 0.3 equivalents of EP (red). The positively charged regions of the protein and the poly-Leu region are found to be the most affected by the interaction. This is highlighted in the expansion reported on the left, where the cross peaks from nuclei in the stretches 36–43, 208–211, and 216–224 are shown. A structural model of NTR is reported in the upper part of the figure and the primary sequence of the protein in the lower part. The IDRs are reported in bold, and the following color coding has been used to highlight the different behavior of specific tracts: the residues that are still observable at the end of the titration are reported in orange, while the residues that show a variation in chemical shift and/or a reduction in intensity are reported in blue.

Interestingly a subset of cross peaks shows a higher intensity in the presence of EP (Figure S7). These are due to the nuclei of residues in regions that remain highly flexible in the complex and are almost all “disorder-promoting” amino acids [63–67], with a large

share of glycine and glutamine residues. The increase in intensity upon binding could be related to the increased mobility of these residues in the complex with respect to that in the isolated protein [68].

## 4. Discussion

### 4.1. The Dynamical Binding Modes of NTD

The combined analysis of the CSP determined through the 2D NMR spectra based on  $^1\text{H}^{\text{N}}$ - and  $^{13}\text{C}'$ -detection delineates a clustering in two main regions on the NTD, the basic finger and the  $\beta 1$ – $\beta 5$  antiparallel sheet. The basic finger is mainly characterized by positively charged residues (5 out of 15 residues in the 92–106 stretch) and possesses an amino acid pattern characteristic of a glycosaminoglycan (GAG)-binding domain (an X-BXB motif in the  $^{104}\text{LDKMKG}^{99}$  region) [69,70]. This region is found to be perturbed in our analysis (99–101–102 perturbed in the 2D HN spectra, 101–102–103–104 perturbed in the 2D CON spectra, and 100–104 perturbed in the 2D CBCACO). Interestingly, this region contains two lysine residues (K100, K102) but does not possess any arginine residue, usually the primary actors in a protein–GAG interaction. However, the side chains of arginine residues very close to this main interaction site (99–104) were found to be perturbed. Indeed, the resonances of  $\text{H}^{\epsilon}$ - $\text{N}^{\epsilon}$  of R93 and R107 are affected upon the addition of EP to the protein solution. The involvement of the finger in the interaction with EP is thus in line with predictions/expectations. On the contrary, it is interesting to note that most of the residues forming the  $\beta 1$  and  $\beta 5$  secondary structure elements are not positively charged ( $\beta 1$ :  $^{56}\text{LTQ}^{58}$  and  $\beta 5$ :  $^{171}\text{FYA}^{173}$ ), nor is the region preceding  $\beta 5$  ( $^{165}\text{TTLPK}^{169}$ ), which is also perturbed upon the addition of EP. In addition, E62 and D63, in the loop following strand  $\beta 1$  and E174 at the end of strand  $\beta 5$  are also perturbed. These are negatively charged and are likely to be engaged in intramolecular electrostatic interactions. Therefore, the observed changes in these regions upon the addition of EP could also be due to perturbations that are propagated throughout the 3D structure. Indeed, the  $\beta 1$  and  $\beta 5$  strands are very short, and  $\beta 5$  is close to the terminal amino acid of the NTD domain, two aspects that render this region quite sensitive to perturbations, a change that could then be easily propagated to the preceding residues (165–169).

The highly negatively charged compound EP could be driven to the positively charged basic finger of NTD thanks to the strong electrostatic attraction. However, its dimension (16mer, 4.5 KDa) and the absence of hydrophobicity limit the contact with the protein core. On the other hand, the bulkiness of EP could play an important role in perturbing the structure close to the domains' ends, also interfering with the network of intramolecular electrostatic interactions. Thus, even residues located far from the initial interaction surface can be perturbed due to structural fluctuations.

The docking analysis supports this picture. Considering the best results of the docking, the interaction region is overall positively charged comprising the highly charged region of the finger. HADDOCK computes electrostatic force as the main contribution of the interaction ( $-380.6 \pm 83.4 \text{ Kcal}\cdot\text{mol}^{-1}$ ) with respect to the Van der Waals energy ( $-38.6 \pm 5.7 \text{ Kcal}\cdot\text{mol}^{-1}$ ). This is in line with the opposite charges of the two interacting partners. Additionally, from the docking point of view, the interaction between EP and the protein core is hindered, with EP placed on the edge of the NTD's surface capable of disrupting intramolecular interactions that eventually occur, as well as possible intermolecular interactions with other partner molecules such as RNA.

The binding affinity between the two molecules and the peculiar folding topology of the protein limit the representation of the binding with a unique, well-defined binding model and indicate an extended perturbed surface. In this representation, the basic residues are the main drivers of the interaction and imply structural modifications sensed far from the binding region.

This is also supported by the analysis of the dynamic properties of NTD. In the presence of EP, the relaxation properties are indicative of a species with higher molecular mass



in solution, with increased  $R_2/R_1$  ratios; flexibility in the finger is retained upon binding. This is a typical behavior of modular proteins in a transient complex with RNA [71].

It is interesting to compare our results with recent studies focusing on the interaction of NTD with different fragments of nucleic acids. Indeed, NMR spectra were used to follow CSPs of NTD upon the addition of increasing concentrations of nucleic acid fragments to map interaction surfaces [19,22,25,58,72]. The region of the basic finger is generally extensively perturbed. Interestingly, the mutation of R92 was found to abrogate the interaction with DNA [73]. Another common feature monitored in these studies is that the perturbed residues are not limited to a specific region of the protein, but generally, large surface areas are found to be perturbed upon interaction. The interaction with EP significantly resembles this general behavior, indicating that it shares common features with the interplay of NTD with different kinds of polyanions, such as nucleic acid fragments. The identification of specific features linked to the different types of partner molecules (RNA, DNA), to whether they are single or double strand, to how the length of the fragment and its conformation affect the interaction are still a matter of debate [22,25,58].

#### 4.2. The Role of IDRs in Orchestrating NTR-EP Interaction

The important role of the flexible linkers in modulating the properties of N has been pointed out in the literature since early studies on the SARS-CoV-1 variant that showed how the linkers promote an increase in the affinity of the NTD for fragments of RNA [16]. However, atom-resolved information about their role has remained elusive, as only recently a sequence-specific assignment of the linkers in the context of the NTR construct has become available [15,21]. Briefly,  $^{13}\text{C}$ -direct detection has been recently demonstrated to be an effective tool to monitor the effect of IDRs in the interplay between gRNA and NTR [19]. This now allows us to inspect in detail the effect of the IDRs also on the interaction with EP. The IDRs comprise the majority of the basic amino acids distributed along the primary sequence of the protein (13 arginine and four lysine residues). This class of amino acids can be the first interacting partners to a negatively charged molecule such as EP [74]. Moreover, the high mobility typical of IDRs facilitates the encounter between the two molecules with a much higher sampled space with respect to the NTD finger [75,76]. This is in line with the increased affinity for EP observed when the two IDRs flank the NTD in the NTR construct.

Zooming into the IDRs through 2D NMR spectra, in particular through the 2D CON experiments, which reveal atom-resolved information about the IDRs in a very clean way [19], it is interesting to note that different regions of the IDRs are perturbed to different extents. In particular, two arginine-rich regions are significantly perturbed, in agreement with the electrostatic sensing of negatively charged EP. Interestingly the most perturbed segment in IDR1 ( $^{37}\text{SKQRRPQ}^{43}$ ) has a characteristic EP interaction motif [69]. However, the overall picture is more complex than that, as expected from a structurally and dynamically heterogeneous protein such as the NTR [76–78]. For example, the addition of EP also influences the resonances in the 216–225 region, a quite remote one from the NTD. This region, mainly composed of leucine residues ( $^{217}\text{AALALLL}^{224}$ ), adopts a helical conformation and is flanked by two aspartate residues (D216 and D225), all features that render it quite inappropriate for direct interaction with the highly negatively charged “ligand”. The observed changes could thus be due to the disruption of intra-molecular interactions that are perturbed by the interaction with EP, as also observed upon interaction with RNA [19]. Finally, other observed features upon binding are that a subset of the residues of the IDRs remain highly flexible; actually, in a few cases, they even seem to increase their mobility upon interaction. This could result from the perturbation of the ensemble of the conformers describing the NTR when in the presence of EP. The amino acids whose mobility is enhanced upon the addition of EP (residues Q4, P6, Q7 N8, Q9, F17, P20, T24, G44, G200, L201, G238, and Q239) mainly belong to the so-called “disordered promoting” class (A, R, G, Q, S, P, and E), in particular, with several glycine and glutamine residues involved in the peptide bond (6 and 5, respectively) [79–81]. Interestingly, compensatory adaptations in

different regions of an IDP, osteopontin, were previously observed upon interaction with heparin [82], reminiscent of the observations in the present study.

These new insights provide a hint of the stronger anchoring of EP on the extended protein surface of NTR. The IDRs, together with the basic finger, seem to act as *sensors* for negatively charged molecules. They might create a platform to accommodate the long polysaccharide on NTD while exploiting a major surface area given by the disordered regions. Moreover, the structural and dynamic features induced by IDRs further complicate the binding mode landscape, as often happens when IDPs/IDRs are involved in binding [75,81,83–85].

In particular, the exposed arginine residues scattered on the primary sequence of flexible regions not only establish strong coulombic interactions but they can also participate in hydrogen bonds with the sulfate groups of the EP. The distribution of the charged residues, in particular in the SR-motif, could simulate the effect of GAG-binding motifs, determining the stronger affinity observed for NTR. Serine residues are indeed the most frequent amino acids which intercalate the cluster of basic residues typically observed in many heparin-binding motifs [69].

## 5. Conclusions

We have shown that the N-terminal region of N from SARS-CoV-2 (NTR, 1-248) interacts with enoxaparin. This interaction was initially investigated by focusing on the NTD globular domain (44-180). This allowed us to map on the 3D structure of this domain an extended region perturbed upon the addition of EP, with the core of the interaction being the flexible basic finger, rich in positively charged residues. As a following step, we showed that two disordered regions flanking the globular domain, IDR1 (1-45) and IDR2 (181-248), contribute to an increase in the affinity of EP to the protein. NMR allowed us to access atom-resolved information on the two IDRs part of the whole NTR construct, revealing a complex interplay between different regions of this multi-domain protein construct and highlighting the importance of these flexible segments for the protein behavior when an interaction occurs. Selected motifs on the IDRs, rich in arginine residues, were shown to be involved in the interaction. Interestingly the data also reveal protein regions that remain highly flexible in the complex.

These molecular details on the interaction of N with EP may contribute to understanding the possible interactions of N with endogenous heparin/glycosaminoglycans, as well as to reveal unpredicted roles exerted by low molecular weight heparin used in the treatment of COVID-19. The perspective of this work is the investigation of the full-length N protein, which can provide further insights into understanding the key mechanism of the interaction of the protein with polyanions able to interfere with its function.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/biom12091302/s1>, Table S1: assigned resonances; Figure S1: CON peaks intensity for the NTD construct, Figure S2:  $^{15}\text{N}$   $R_2$ ,  $R_1$ , NOE values, Figure S3: DOSY spectra, Figure S4: Cluster 2 to 6 obtained from HADDOCK, Figure S5: CSP and Intensity ratio for NTD and NTR, Figure S6: ITC titration, Figure S7: zoom of HN spectra for the NTR construct, Figure S8: Intensity ratios of CON's peaks.

**Author Contributions:** R.P. and I.C.F. conceived the project and planned the experiments; L.P. and G.T. produced the protein samples; L.P., M.S. and G.T. acquired and analyzed the NMR spectra under the guidance of I.C.F. and R.P.; L.P. and M.S. acquired the ITC data. R.P. and I.C.F. wrote the manuscript with the contribution of all the other authors. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded in part by a grant from the Italian Ministry of University and Research (FISR2020IP\_02112, ID-COVID), by Fondazione CR Firenze and with the support of the Italian government program “MIUR Dipartimenti di Eccellenza 2018–2022” (58503\_DIPECC).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available upon request from the corresponding authors. The obtained NMR assignment has been deposited on the BMRB under the accession code 51620.

**Acknowledgments:** The support of the CERM/CIRMMP center of Instruct-ERIC is gratefully acknowledged. We thank Fabio Almeida from the University of Rio de Janeiro, who provided us the gene to express the NTD domain. Alexandre M. J. J. Bonvin from the University of Utrecht and Robert Konrat from the University of Vienna are gratefully acknowledged for the stimulating discussion and the precious comments. Marco Fragai is acknowledged for his useful suggestions for the ITC measurements. The Covid19-NMR consortium is acknowledged for stimulating discussion. Finally, we would like to thank Sonia Longhi for opening the field of intrinsically disordered viral proteins.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Tarighi, P.; Eftekhari, S.; Chizari, M.; Sabernavaei, M.; Jafari, D.; Mirzabeigi, P. A Review of Potential Suggested Drugs for Coronavirus Disease (COVID-19) Treatment. *Eur. J. Pharmacol.* **2021**, *895*, 173890. [[CrossRef](#)]
2. Tabll, A.A.; Shahein, Y.E.; Omran, M.M.; Elnakib, M.M.; Ragheb, A.A.; Amer, K.E. A Review of Monoclonal Antibodies in COVID-19: Role in Immunotherapy, Vaccine Development and Viral Detection. *Hum. Antibodies* **2021**, *29*, 179–191. [[CrossRef](#)]
3. Zheng, C.; Shao, W.; Chen, X.; Zhang, B.; Wang, G.; Zhang, W. Real-World Effectiveness of COVID-19 Vaccines: A Literature Review and Meta-Analysis. *Int. J. Infect. Dis.* **2022**, *114*, 252–260. [[CrossRef](#)] [[PubMed](#)]
4. Faraji, S.N.; Raee, M.J.; Hashemi, S.M.A.; Daryabor, G.; Tabrizi, R.; Dashti, F.S.; Behboudi, E.; Heidarnajad, K.; Nowrouzi-Sohrabi, P.; Hatam, G. Human Interaction Targets of SARS-CoV-2 Spike Protein: A Systematic Review. *Eur. J. Inflamm.* **2022**, *20*, 1721727X2210953. [[CrossRef](#)]
5. Molina-Mora, J.A. Insights into the Mutation T1117I in the Spike and the Lineage B.1.1.389 of SARS-CoV-2 Circulating in Costa Rica. *Gene Rep.* **2022**, *27*, 101554. [[CrossRef](#)] [[PubMed](#)]
6. Harvey, W.T.; Carabelli, A.M.; Jackson, B.; Gupta, R.K.; Thomson, E.C.; Harrison, E.M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S.J.; et al. SARS-CoV-2 Variants, Spike Mutations and Immune Escape. *Nat. Rev. Microbiol.* **2022**, *19*, 409–424. [[CrossRef](#)]
7. Gordon, D.E.; Jang, G.M.; Bouhaddou, M.; Xu, J.; Obernier, K.; White, K.M.; O’Meara, M.J.; Rezelj, V.V.; Guo, J.Z.; Swaney, D.L.; et al. A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing. *Nature* **2020**, *583*, 459–468. [[CrossRef](#)] [[PubMed](#)]
8. Thorne, L.G.; Bouhaddou, M.; Reuschl, A.-K.; Zuliani-Alvarez, L.; Polacco, B.; Pelin, A.; Batra, J.; Whelan, M.V.X.; Hosmillo, M.; Fossati, A.; et al. Evolution of Enhanced Innate Immune Evasion by SARS-CoV-2. *Nature* **2022**, *602*, 487–495. [[CrossRef](#)]
9. Syed, A.M.; Taha, T.Y.; Tabata, T.; Chen, I.P.; Ciling, A.; Khalid, M.M.; Sreekumar, B.; Chen, P.-Y.; Hayashi, J.M.; Soczek, K.M.; et al. Rapid Assessment of SARS-CoV-2-Evolved Variants Using Virus-like Particles. *Science* **2021**, *374*, 1626–1632. [[CrossRef](#)]
10. Quaglia, F.; Salladini, E.; Carraro, M.; Minervini, G.; Tosatto, S.C.E.; Le Mercier, P. SARS-CoV-2 Variants Preferentially Emerge at Intrinsically Disordered Protein Sites Helping Immune Evasion. *FEBS J.* **2022**, *289*, 4240–4250. [[CrossRef](#)]
11. Chang, C.K.; Hou, M.H.; Chang, C.F.; Hsiao, C.D.; Huang, T.H. The SARS Coronavirus Nucleocapsid Protein-Forms and Functions. *Antivir. Res.* **2014**, *103*, 39–50. [[CrossRef](#)]
12. Xue, B.; Blocquel, D.; Habchi, J.; Uversky, A.V.; Kurgan, L.; Uversky, V.N.; Longhi, S. Structural Disorder in Viral Proteins. *Chem. Rev.* **2014**, *114*, 6880–6911. [[CrossRef](#)] [[PubMed](#)]
13. Goh, G.K.-M.; Dunker, A.K.; Foster, J.A.; Uversky, V.N. Rigidity of the Outer Shell Predicted by a Protein Intrinsic Disorder Model Sheds Light on the COVID-19 (Wuhan-2019-NCov) Infectivity. *Biomolecules* **2020**, *10*, 331. [[CrossRef](#)]
14. Giri, R.; Bhardwaj, T.; Shegane, M.; Gehi, B.R.; Kumar, P.; Gadhave, K.; Oldfield, C.J.; Uversky, V.N. Understanding COVID-19 via Comparative Analysis of Dark Proteomes of SARS-CoV-2, Human SARS and Bat SARS-like Coronaviruses. *Cell. Mol. Life Sci.* **2021**, *78*, 1655–1688. [[CrossRef](#)] [[PubMed](#)]
15. Schiavina, M.; Pontoriero, L.; Uversky, V.N.; Felli, I.C.; Pierattelli, R. 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.* **2021**, *15*, 219–227. [[CrossRef](#)] [[PubMed](#)]
16. Chang, C.-K.; Hsu, Y.-L.; Chang, Y.-H.; Chao, F.-A.; Wu, M.-C.; Huang, Y.-S.; Hu, C.-K.; Huang, T.-H. Multiple Nucleic Acid Binding Sites and Intrinsic Disorder of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein: Implications for Ribonucleocapsid Protein Packaging. *J. Virol.* **2009**, *83*, 2255–2264. [[CrossRef](#)]
17. Forsythe, H.M.; Rodriguez Galvan, J.; Yu, Z.; Pinckney, S.; Reardon, P.; Cooley, R.B.; Zhu, P.; Rolland, A.D.; Prell, J.S.; Barbar, E. Multivalent Binding of the Partially Disordered SARS-CoV-2 Nucleocapsid Phosphoprotein Dimer to RNA. *Biophys. J.* **2021**, *120*, 2890–2901. [[CrossRef](#)]
18. Bessa, L.M.; Guseva, S.; Camacho-Zarco, A.R.; Salvi, N.; Maurin, D.; Perez, L.M.; Botova, M.; Malki, A.; Nanao, M.; Jensen, M.R.; et al. The Intrinsically Disordered SARS-CoV-2 Nucleoprotein in Dynamic Complex with Its Viral Partner Nsp3a. *Sci. Adv.* **2022**, *8*, eabm4034. [[CrossRef](#)]

19. Pontoriero, L.; Schiavina, M.; Korn, S.M.; Schlundt, A.; Pierattelli, R.; Felli, I.C. NMR Reveals Specific Tracts within the Intrinsically Disordered Regions of the SARS-CoV-2 Nucleocapsid Protein Involved in RNA Encountering. *Biomolecules* **2022**, *12*, 929. [[CrossRef](#)]
20. Peng, Y.; Du, N.; Lei, Y.; Dorje, S.; Qi, J.; Luo, T.; Gao, G.F.; Song, H. Structures of the SARS-CoV-2 Nucleocapsid and Their Perspectives for Drug Design. *EMBO J.* **2020**, *39*, e105938. [[CrossRef](#)]
21. Guseva, S.; Perez, L.M.; Camacho-Zarco, A.; Bessa, L.M.; Salvi, N.; Malki, A.; Maurin, D.; Blackledge, M. <sup>1</sup>H, <sup>13</sup>C and <sup>15</sup>N Backbone Chemical Shift Assignments of the n-Terminal and Central Intrinsically Disordered Domains of SARS-CoV-2 Nucleoprotein. *Biomol. NMR Assign.* **2021**, *15*, 255–260. [[CrossRef](#)] [[PubMed](#)]
22. Dinesh, D.C.; Chalupska, D.; Silhan, J.; Koutna, E.; Nencka, R.; Veverka, V.; Boura, E. Structural Basis of RNA Recognition by the SARS-CoV-2 Nucleocapsid Phosphoprotein. *PLoS Pathog.* **2020**, *16*, e1009100. [[CrossRef](#)] [[PubMed](#)]
23. Savastano, A.; Ibáñez de Opakua, A.; Rankovic, M.; Zweckstetter, M. Nucleocapsid Protein of SARS-CoV-2 Phase Separates into RNA-Rich Polymerase-Containing Condensates. *Nat. Commun.* **2020**, *11*, 6041. [[CrossRef](#)] [[PubMed](#)]
24. Redzic, J.S.; Lee, E.; Born, A.; Issaian, A.; Henen, M.A.; Nichols, P.J.; Blue, A.; Hansen, K.C.; D'Alessandro, A.; Vögeli, B.; et al. The Inherent Dynamics and Interaction Sites of the SARS-CoV-2 Nucleocapsid N-Terminal Region. *J. Mol. Biol.* **2021**, *433*, 167108. [[CrossRef](#)]
25. Caruso, I.P.; dos Santos Almeida, V.; do Amaral, M.J.; de Andrade, G.C.; de Araújo, G.R.; de Araújo, T.S.; de Azevedo, J.M.; Barbosa, G.M.; Bartkevicius, L.; Bezerra, P.R.; et al. Insights into the Specificity for the Interaction of the Promiscuous SARS-CoV-2 Nucleocapsid Protein N-Terminal Domain with Deoxyribonucleic Acids. *Int. J. Biol. Macromol.* **2022**, *203*, 466–480. [[CrossRef](#)]
26. Shi, C.; Tingting, W.; Li, J.-P.; Sullivan, M.A.; Wang, C.; Wang, H.; Deng, B.; Zhang, Y. Comprehensive Landscape of Heparin Therapy for COVID-19. *Carbohydr. Polym.* **2021**, *254*, 117232. [[CrossRef](#)]
27. Shan, D.; Johnson, J.M.; Fernandes, S.C.; Suib, H.; Hwang, S.; Wuelfing, D.; Mendes, M.; Holdridge, M.; Burke, E.M.; Beauregard, K.; et al. N-Protein Presents Early in Blood, Dried Blood and Saliva during Asymptomatic and Symptomatic SARS-CoV-2 Infection. *Nat. Commun.* **2021**, *12*, 1931. [[CrossRef](#)]
28. Kielstein, J.T.; Borchina, D.-N.; Fühner, T.; Hwang, S.; Mattoon, D.; Ball, A.J. Hemofiltration with the Seraph<sup>®</sup> 100 Microbind<sup>®</sup> Affinity Filter Decreases SARS-CoV-2 Nucleocapsid Protein in Critically Ill COVID-19 Patients. *Crit. Care* **2021**, *25*, 190. [[CrossRef](#)]
29. López-Muñoz, A.D.; Kosik, I.; Holly, J.; Yewdell, J.W. Cell Surface SARS-CoV-2 Nucleocapsid Protein Modulates Innate and Adaptive Immunity. *Sci. Adv.* **2022**, *8*, eabp9770. [[CrossRef](#)]
30. González-Motos, V.; Kropp, K.A.; Viejo-Borbolla, A. Chemokine Binding Proteins: An Immunomodulatory Strategy Going Viral. *Cytokine Growth Factor Rev.* **2016**, *30*, 71–80. [[CrossRef](#)]
31. Hernaez, B.; Alcamí, A. Virus-Encoded Cytokine and Chemokine Decoy Receptors. *Curr. Opin. Immunol.* **2020**, *66*, 50–56. [[CrossRef](#)] [[PubMed](#)]
32. Bernadó, P.; Mylonas, E.; Petoukhov, M.V.; Blackledge, M.; Svergun, D.I. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* **2007**, *129*, 5656–5664. [[CrossRef](#)] [[PubMed](#)]
33. Tria, G.; Mertens, H.D.T.; Kachala, M.; Svergun, D.I. Advanced Ensemble Modelling of Flexible Macromolecules Using X-ray Solution Scattering. *IUCr* **2015**, *2*, 207–217. [[CrossRef](#)]
34. Khan, S.; Gor, J.; Mulloy, B.; Perkins, S.J. Semi-Rigid Solution Structures of Heparin by Constrained X-ray Scattering Modelling: New Insight into Heparin–Protein Complexes. *J. Mol. Biol.* **2010**, *395*, 504–521. [[CrossRef](#)]
35. Altincekic, N.; Korn, S.M.; Qureshi, N.S.; Dujardin, M.; Ninot-Pedrosa, M.; Abele, R.; Abi Saad, M.J.; Alfano, C.; Almeida, F.C.L.; Alshamleh, I.; et al. Large-Scale Recombinant Production of the SARS-CoV-2 Proteome for High-Throughput and Structural Biology Applications. *Front. Mol. Biosci.* **2021**, *8*, 653148. [[CrossRef](#)]
36. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; et al. A New Coronavirus Associated with Human Respiratory Disease in China. *Nature* **2020**, *579*, 265–269. [[CrossRef](#)] [[PubMed](#)]
37. Marley, J.; Lu, M.; Bracken, C. A Method for Efficient Isotopic Labeling of Recombinant Proteins. *J. Biomol. NMR* **2001**, *20*, 71–75. [[CrossRef](#)]
38. Palmer, A.G.; Cavanagh, J.; Wright, P.E.; Rance, M. Sensitivity Improvement in Proton-Detected Two-Dimensional Heteronuclear Correlation NMR Spectroscopy. *J. Magn. Reson.* **1991**, *93*, 151–170. [[CrossRef](#)]
39. Schleucher, J.; Schwendinger, M.; Sattler, M.; Schmidt, P.; Schedletzky, O.; Glaser, S.J.; Sorensen, O.W.; Griesinger, C. A General Enhancement Scheme in Heteronuclear Multidimensional NMR Employing Pulsed Field Gradients. *J. Biomol. NMR* **1994**, *4*, 301–306. [[CrossRef](#)]
40. Pontoriero, L.; Schiavina, M.; Murralli, M.G.; Pierattelli, R.; Felli, I.C. Monitoring the Interaction of A-Synuclein with Calcium Ions through Exclusively Heteronuclear Nuclear Magnetic Resonance Experiments. *Angew. Chem. Int. Ed.* **2020**, *59*, 18537–18545. [[CrossRef](#)]
41. Bermel, W.; Bertini, I.; Csizmok, V.; Felli, I.C.; Pierattelli, R.; Tompa, P. H-Start for Exclusively Heteronuclear NMR Spectroscopy: The Case of Intrinsically Disordered Proteins. *J. Magn. Reson.* **2009**, *198*, 275–281. [[CrossRef](#)] [[PubMed](#)]
42. Schiavina, M.; Murralli, M.G.; Pontoriero, L.; Sainati, V.; Kümmerle, R.; Bermel, W.; Pierattelli, R.; Felli, I.C. Taking Simultaneous Snapshots of Intrinsically Disordered Proteins in Action. *Biophys. J.* **2019**, *117*, 46–55. [[CrossRef](#)] [[PubMed](#)]
43. Geen, H.; Freeman, R. Band-Selective Radiofrequency Pulses. *J. Magn. Reson.* **1991**, *93*, 93–141. [[CrossRef](#)]
44. Emsley, L.; Bodenhausen, G. Optimization of Shaped Selective Pulses for NMR Using a Quaternion Description of Their Overall Propagators. *J. Magn. Reson.* **1992**, *97*, 135–148. [[CrossRef](#)]



45. Felli, I.C.; Pierattelli, R. Spin-State-Selective Methods in Solution- and Solid-State Biomolecular  $^{13}\text{C}$  NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **2015**, *84*, 1–13. [[CrossRef](#)]
46. Piotto, M.; Saudek, V.; Sklenar, V. Gradient-Tailored Excitation for Single-Quantum NMR Spectroscopy of Aqueous Solutions. *J. Biomol. NMR* **1992**, *2*, 661–665. [[CrossRef](#)] [[PubMed](#)]
47. Farrow, N.A.; Muhandiram, R.; Singer, A.U.; Pascal, S.M.; Kay, C.M.; Gish, G.; Shoelson, S.E.; Pawson, T.; Forman-Kay, J.D.; Kay, L.E. Backbone Dynamics of a Free and a Phosphopeptide-Complexed Src Homology 2 Domain Studied by  $^{15}\text{N}$  NMR Relaxation. *Biochemistry* **1994**, *33*, 5984–6003. [[CrossRef](#)]
48. Johnson, C.S. Diffusion Ordered Nuclear Magnetic Resonance Spectroscopy: Principles and Applications. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 203–256. [[CrossRef](#)]
49. Keller, R. *The Computer Aided Resonance Assignment Tutorial*; Cantina Verlag: Goldau, Switzerland, 2004; pp. 1–81.
50. Bartels, C.; Xia, T.H.; Billeter, M.; Güntert, P.; Wüthrich, K. The Program XEASY for Computer-Supported NMR Spectral Analysis of Biological Macromolecules. *J. Biomol. NMR* **1995**, *6*, 1–10. [[CrossRef](#)] [[PubMed](#)]
51. Markley, J.L.; Bax, A.; Arata, Y.; Hilbers, C.W.; Kaptein, R.; Sykes, B.D.; Wright, P.E.; Wüthrich, K. Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. *Pure Appl. Chem.* **1998**, *70*, 117–142. [[CrossRef](#)]
52. Dominguez, C.; Boelens, R.; Bonvin, A.M.J.J. HADDOCK: A Protein–Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.* **2003**, *125*, 1731–1737. [[CrossRef](#)] [[PubMed](#)]
53. Honorato, R.V.; Koukos, P.I.; Jiménez-García, B.; Tsaregorodtsev, A.; Verlatto, M.; Giachetti, A.; Rosato, A.; Bonvin, A.M.J.J. Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. *Front. Mol. Biosci.* **2021**, *8*, 729513. [[CrossRef](#)]
54. Felli, I.C.; Pierattelli, R.  $^{13}\text{C}$  Direct Detected NMR for Challenging Systems. *Chem. Rev.* **2022**, *122*, 9468–9496. [[CrossRef](#)] [[PubMed](#)]
55. Bertini, I.; Felli, I.C.; Gonnelli, L.; Pierattelli, R.; Spyraniti, Z.; Spyroulias, G.A. Mapping Protein–Protein Interaction by  $^{13}\text{C}'$ -Detected Heteronuclear NMR Spectroscopy. *J. Biomol. NMR* **2006**, *36*, 111–122. [[CrossRef](#)]
56. Alik, A.; Bougouchtoui, C.; Julien, M.; Bermel, W.; Ghoul, R.; Zinn-Justin, S.; Theillet, F. Sensitivity-Enhanced  $^{13}\text{C}$ -NMR Spectroscopy for Monitoring Multisite Phosphorylation at Physiological Temperature and pH. *Angew. Chem. Int. Ed.* **2020**, *59*, 10411–10415. [[CrossRef](#)]
57. Clarkson, M.W.; Lei, M.; Eisenmesser, E.Z.; Labeikovsky, W.; Redfield, A.; Kern, D. Mesodynamics in the SARS Nucleocapsid Measured by NMR Field Cycling. *J. Biomol. NMR* **2009**, *45*, 217–225. [[CrossRef](#)]
58. Korn, S.M.; Dhamotharan, K.; Schlundt, A. The Preference Signature of the SARS-CoV-2 Nucleocapsid NTD for Its 5'-Genomic RNA Elements. *Res. Sq.* **2022**. [[CrossRef](#)]
59. Perdikari, T.M.; Murthy, A.C.; Ryan, V.H.; Watters, S.; Naik, M.T.; Fawzi, N.L. SARS-CoV-2 Nucleocapsid Protein Phase-separates with RNA and with Human HnRNPs. *EMBO J.* **2020**, *39*, e106478. [[CrossRef](#)]
60. Carlson, C.R.; Asfaha, J.B.; Ghent, C.M.; Howard, C.J.; Hartooni, N.; Safari, M.; Frankel, A.D.; Morgan, D.O. Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for Its Dual Functions. *Mol. Cell* **2020**, *80*, 1092–1103.e4. [[CrossRef](#)]
61. Cubuk, J.; Alston, J.J.; Incicco, J.J.; Singh, S.; Stuchell-Brereton, M.D.; Ward, M.D.; Zimmerman, M.I.; Vithani, N.; Griffith, D.; Wagoner, J.A.; et al. The SARS-CoV-2 Nucleocapsid Protein Is Dynamic, Disordered, and Phase Separates with RNA. *Nat. Commun.* **2021**, *12*, 1936. [[CrossRef](#)]
62. Lu, S.; Ye, Q.; Singh, D.; Cao, Y.; Diedrich, J.K.; Yates, J.R.; Villa, E.; Cleveland, D.W.; Corbett, K.D. The SARS-CoV-2 Nucleocapsid Phosphoprotein Forms Mutually Exclusive Condensates with RNA and the Membrane-Associated M Protein. *Nat. Commun.* **2021**, *12*, 502. [[CrossRef](#)]
63. Xue, B.; Dunbrack, R.L.; Williams, R.W.; Dunker, A.K.; Uversky, V.N. PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids. *Biochim. Biophys. Acta-Proteins Proteom.* **2010**, *1804*, 996–1010. [[CrossRef](#)] [[PubMed](#)]
64. Romero, P.; Obradovic, Z.; Li, X.; Garner, E.C.; Brown, C.J.; Dunker, A.K. Sequence Complexity of Disordered Protein. *Proteins* **2001**, *42*, 38–48. [[CrossRef](#)]
65. Uversky, V.N.; Oldfield, C.J.; Dunker, A.K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. *Annu. Rev. Biophys.* **2008**, *37*, 215–246. [[CrossRef](#)] [[PubMed](#)]
66. Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337. [[CrossRef](#)]
67. Dosztányi, Z.; Csizmók, V.; Tompa, P.; Simon, I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* **2005**, *347*, 827–839. [[CrossRef](#)]
68. Kurzbach, D.; Platzer, G.; Schwarz, T.C.; Henen, M.A.; Konrat, R.; Hinderberger, D. Cooperative Unfolding of Compact Conformations of the Intrinsically Disordered Protein Osteopontin. *Biochemistry* **2013**, *52*, 5167–5175. [[CrossRef](#)]
69. Hileman, R.E.; Fromm, J.R.; Weiler, J.M.; Linhardt, R.J. Glycosaminoglycan-Protein Interactions: Definition of Consensus Sites in Glycosaminoglycan Binding Proteins. *BioEssays* **1998**, *20*, 156–167. [[CrossRef](#)]
70. Capila, I.; Linhardt, R.J. Heparin-Protein Interactions. *Angew. Chem. Int. Ed.* **2002**, *41*, 390–412. [[CrossRef](#)]
71. Deka, P.; Rajan, P.K.; Perez-Canadillas, J.M.; Varani, G. Protein and RNA Dynamics Play Key Roles in Determining the Specific Recognition of GU-Rich Polyadenylation Regulatory Elements by Human Cstf-64 Protein. *J. Mol. Biol.* **2005**, *347*, 719–733. [[CrossRef](#)]

72. Huang, Q.; Yu, L.; Petros, A.M.; Gunasekera, A.; Liu, Z.; Xu, N.; Hajduk, P.; Mack, J.; Fesik, S.W.; Olejniczak, E.T. Structure of the N-Terminal RNA-Binding Domain of the SARS CoV Nucleocapsid Protein. *Biochemistry* **2004**, *43*, 6059–6063. [[CrossRef](#)] [[PubMed](#)]
73. Caruso, Í.P.; Sanches, K.; Da Poian, A.T.; Pinheiro, A.S.; Almeida, F.C.L. Dynamics of the SARS-CoV-2 Nucleoprotein N-Terminal Domain Triggers RNA Duplex Destabilization. *Biophys. J.* **2021**, *120*, 2814–2827. [[CrossRef](#)] [[PubMed](#)]
74. Mukrasch, M.D.; Biernat, J.; von Bergen, M.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. Sites of Tau Important for Aggregation Populate  $\beta$ -Structure and Bind to Microtubules and Polyanions. *J. Biol. Chem.* **2005**, *280*, 24978–24986. [[CrossRef](#)] [[PubMed](#)]
75. Sottini, A.; Borgia, A.; Borgia, M.B.; Bugge, K.; Nettels, D.; Chowdhury, A.; Heidarsson, P.O.; Zosel, F.; Best, R.B.; Kragelund, B.B.; et al. Polyelectrolyte Interactions Enable Rapid Association and Dissociation in High-Affinity Disordered Protein Complexes. *Nat. Commun.* **2020**, *11*, 5736. [[CrossRef](#)]
76. Teilum, K.; Olsen, J.G.; Kragelund, B.B. On the Specificity of Protein–Protein Interactions in the Context of Disorder. *Biochem. J.* **2021**, *478*, 2035–2050. [[CrossRef](#)]
77. Arbesú, M.; Pons, M. Integrating Disorder in Globular Multidomain Proteins: Fuzzy Sensors and the Role of SH3 Domains. *Arch. Biochem. Biophys.* **2019**, *677*, 108161. [[CrossRef](#)]
78. Arbesú, M.; Iruela, G.; Fuentes, H.; Teixeira, J.M.C.; Pons, M. Intramolecular Fuzzy Interactions Involving Intrinsically Disordered Domains. *Front. Mol. Biosci.* **2018**, *5*, 39. [[CrossRef](#)]
79. Uversky, V.N.; Gillespie, J.R.; Fink, A.L. Why Are “natively Unfolded” Proteins Unstructured under Physiologic Conditions? *Proteins Struct. Funct. Genet.* **2000**, *41*, 415–427. [[CrossRef](#)]
80. Dunker, A.K.; Babu, M.M.; Barbar, E.; Blackledge, M.; Bondos, S.E.; Dosztányi, Z.; Dyson, H.J.; Forman-Kay, J.; Fuxreiter, M.; Gsponer, J.; et al. What’s in a Name? Why These Proteins Are Intrinsically Disordered. *Intrinsically Disord. Proteins* **2013**, *1*, e24157. [[CrossRef](#)]
81. Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V.N. Introducing Protein Intrinsic Disorder. *Chem. Rev.* **2014**, *114*, 6561–6588. [[CrossRef](#)]
82. Kurzbach, D.; Schwarz, T.C.; Platzer, G.; Höfler, S.; Hinderberger, D.; Konrat, R. Compensatory Adaptations of Structural Dynamics in an Intrinsically Disordered Protein Complex. *Angew. Chem.* **2014**, *53*, 3840–3843. [[CrossRef](#)] [[PubMed](#)]
83. Sibille, N.; Sillen, A.; Leroy, A.; Wieruszkeski, J.-M.; Mulloy, B.; Landrieu, I.; Lippens, G. Structural Impact of Heparin Binding to Full-Length Tau As Studied by NMR Spectroscopy. *Biochemistry* **2006**, *45*, 12560–12572. [[CrossRef](#)] [[PubMed](#)]
84. Tompa, P.; Fuxreiter, M. Fuzzy Complexes: Polymorphism and Structural Disorder in Protein-Protein Interactions. *Trends Biochem. Sci.* **2008**, *33*, 2–8. [[CrossRef](#)]
85. Murralli, M.G.; Felli, I.C.; Pierattelli, R. Adenoviral E1A Exploits Flexibility and Disorder to Target Cellular Proteins. *Biomolecules* **2020**, *10*, 1541. [[CrossRef](#)] [[PubMed](#)]

# The Role of Disordered Regions in Orchestrating the Properties of Multidomain Proteins: The SARS-CoV-2 Nucleocapsid Protein and Its Interaction with Enoxaparin

Marco Schiavina †, Letizia Pontoriero †, Giuseppe Tagliaferro, Roberta Pierattelli \* and Isabella C. Felli \*

Magnetic Resonance Center (CERM) and Department of Chemistry “Ugo Schiff”, University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino (Florence), Italy

† These authors contributed equally to the work.

\* Correspondence: [roberta.pierattelli@unifi.it](mailto:roberta.pierattelli@unifi.it) (R.M.); [felli@cerm.unifi.it](mailto:felli@cerm.unifi.it) (I.C.F.)

## Supplementary Tables

**Table S1:** Chemical shift values (ppm) for the newly assigned  $^{13}\text{C}$  and  $^{15}\text{N}$  resonances of the NTD construct in 25 mM  $\text{KH}_2\text{PO}_4/\text{K}_2\text{HPO}_4$  buffer, 150 mM KCl, pH 6.5 at 298K. This chemical shift values has been deposited on the BMRB code under the accession code 51620.

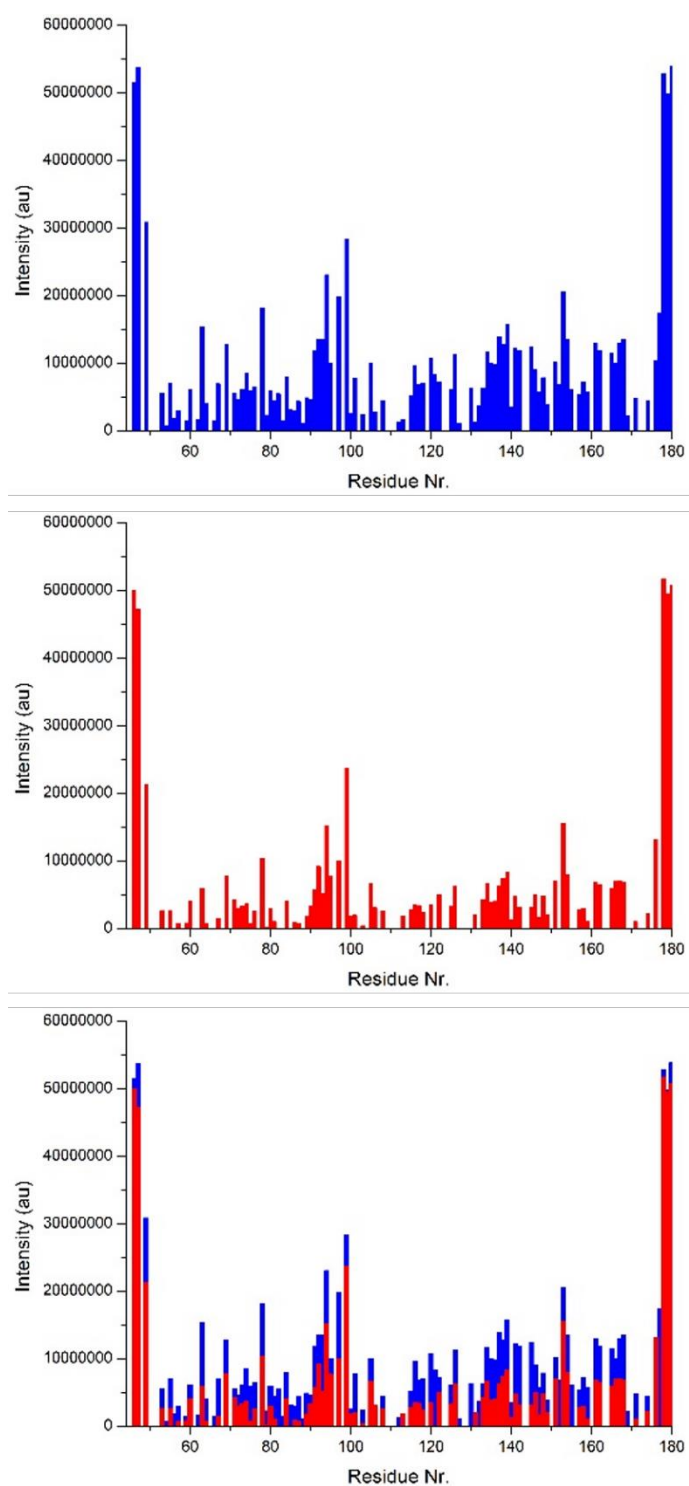
Assignment	$\delta$ (ppm)
N46-PRO	135.90
C $^{\gamma}$ 47-ASN	177.03
C $^{\gamma}$ 48-ASN	177.30
C $^{\gamma}$ 62-GLU	33.59
C $^{\delta}$ 62-GLU	183.93
C $^{\gamma}$ 63-ASP	179.34
N67-PRO	135.50
C $^{\gamma}$ 70-GLN	33.44
C $^{\delta}$ 70-GLN	178.77
N73-PRO	140.31
C $^{\gamma}$ 75-ASN	177.29
C $^{\gamma}$ 77-ASN	177.69
N80-PRO	140.43
C $^{\gamma}$ 81-ASP	178.52
C $^{\gamma}$ 82-ASP	179.18
C $^{\gamma}$ 83-GLN	32.44
C $^{\delta}$ 83-GLN	180.38
C $^{\gamma}$ 98-ASP	180.15
C $^{\gamma}$ 103-ASP	180.18
N106-PRO	138.21
N117-PRO	134.29
C $^{\gamma}$ 118-GLU	33.989
C $^{\delta}$ 118-GLU	181.22
N122-PRO	137.26
C $^{\gamma}$ 126-ASN	176.86
C $^{\gamma}$ 128-ASP	179.87
C $^{\gamma}$ 136-GLU	35.79
C $^{\delta}$ 136-GLU	183.74
C $^{\gamma}$ 140-ASN	176.06
N142-PRO	132.72
C $^{\gamma}$ 144-ASP	179.13
C $^{\gamma}$ 150-ASN	177.39
N151-PRO	136.92
C $^{\gamma}$ 153-ASN	175.91
C $^{\gamma}$ 154-ASN	177.19
C $^{\gamma}$ 160-GLN	33.97
C $^{\delta}$ 160-GLN	180.91
N162-PRO	133.48



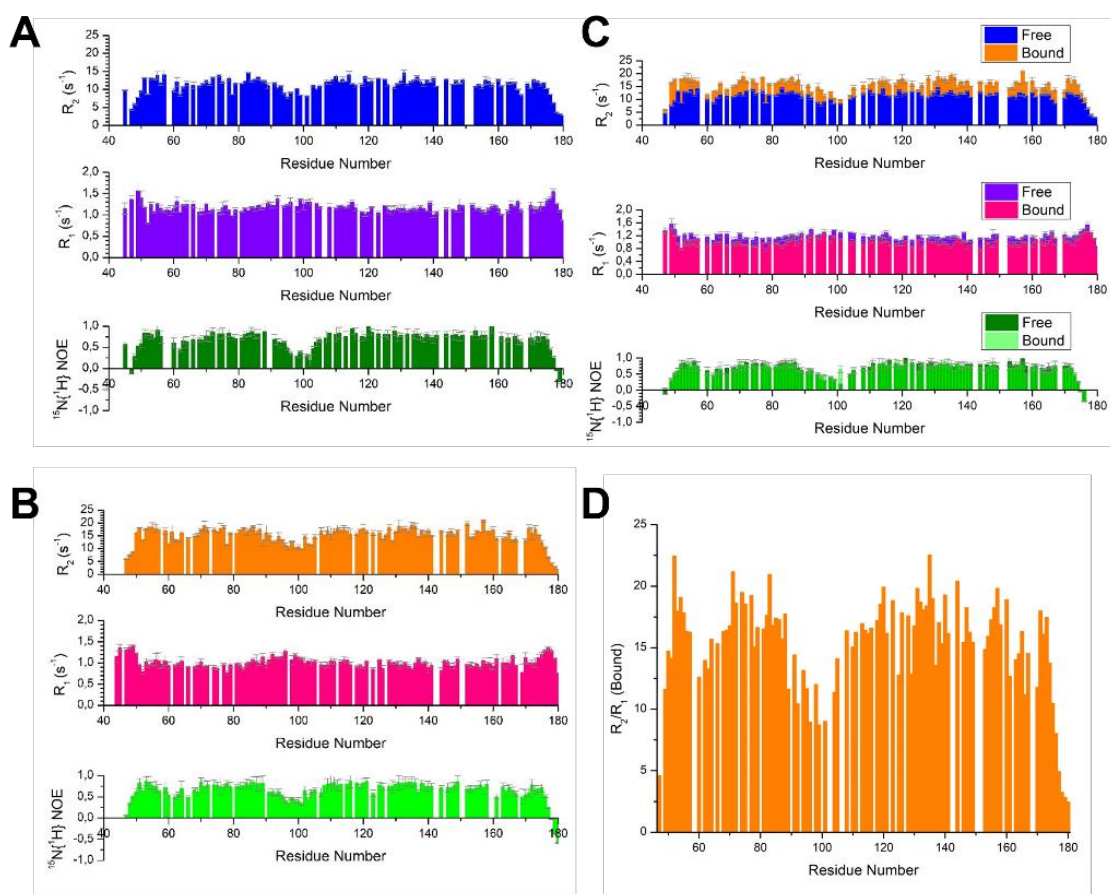
C <sup>γ</sup> 163-GLN	33.47
C <sup>δ</sup> 163-GLN	180.25
N168-PRO	136.06
C <sup>γ</sup> 174-GLU	36.29
C <sup>δ</sup> 174-GLU	183.55

## Supplementary Figures

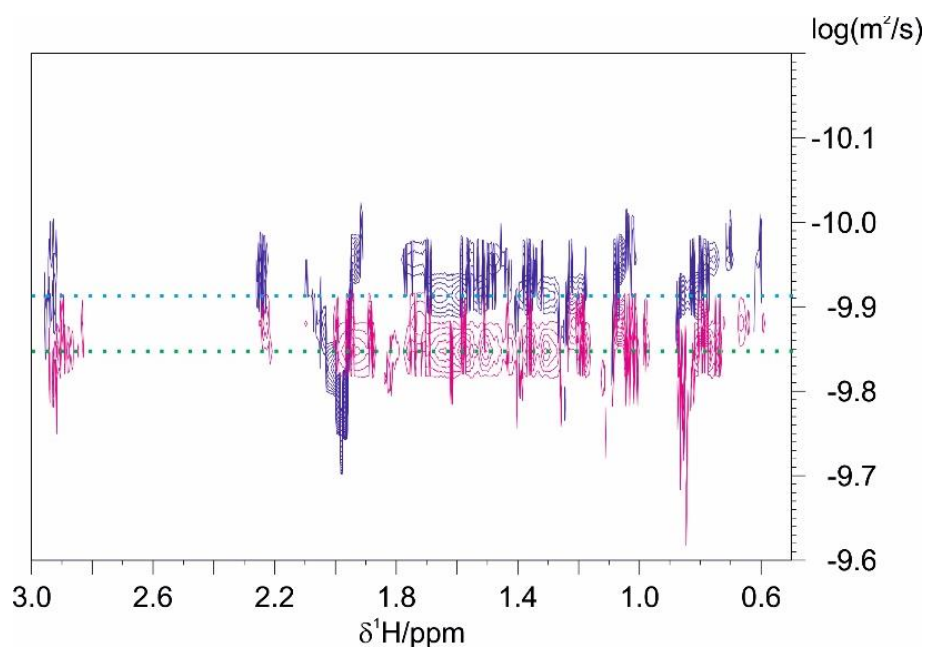
**Figure S1:** The intensities of the cross peaks in CON spectra of NTD are reported versus the residue number for the isolated protein (blue) and after addition of 0.3 equivalents of EP (red).



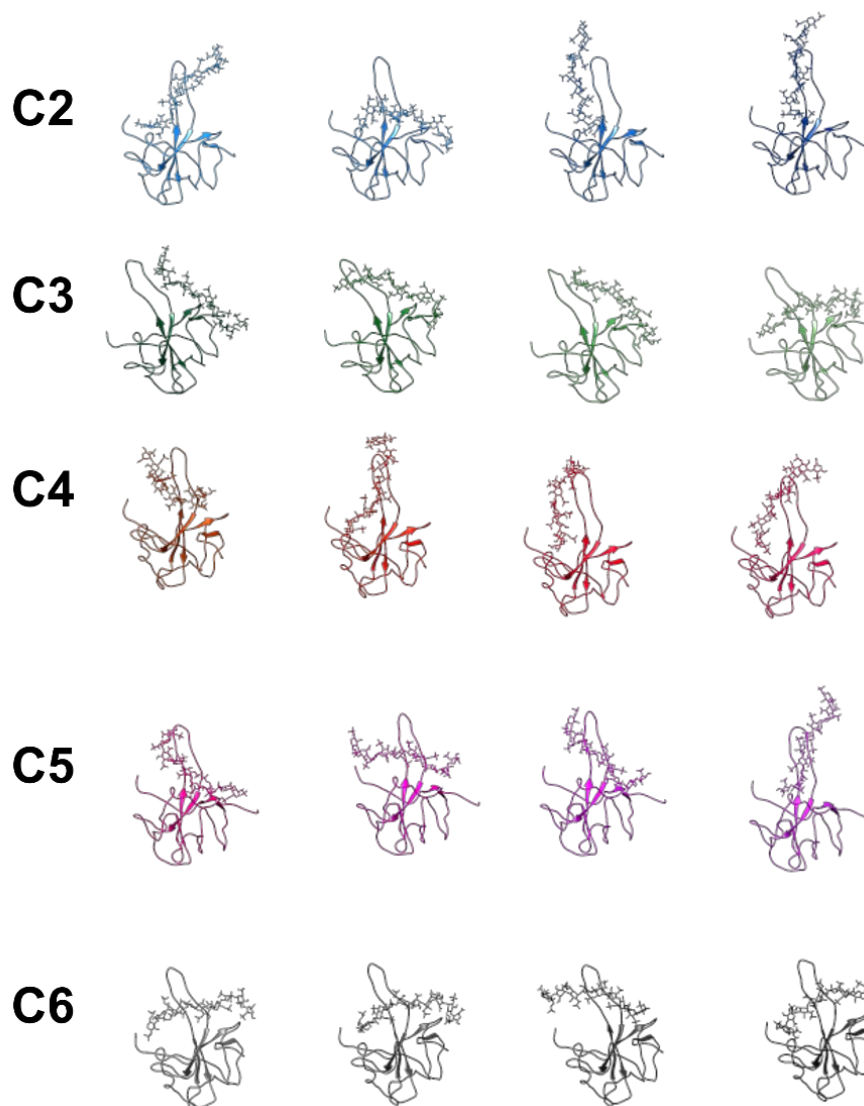
**Figure S2:** The central region involved in the interaction with EP is the most flexible one in the NTD construct. Upon EP binding, the mobility is overall reduced with the flexible regions that, however, still retain their flexibility. The protein possesses a flexible loop, the basic finger, which spans from residue 92 to 106. The  $^{15}\text{N}$   $R_2$ ,  $R_1$ , NOE values are reported in panel A for the free form. Panel B reports the same values for the bound form. In panel C is reported the overlay of the  $^{15}\text{N}$   $R_2$ ,  $R_1$  and NOE values.  $R_2/R_1$  values are reported for the bound form against the primary sequence in panel D.



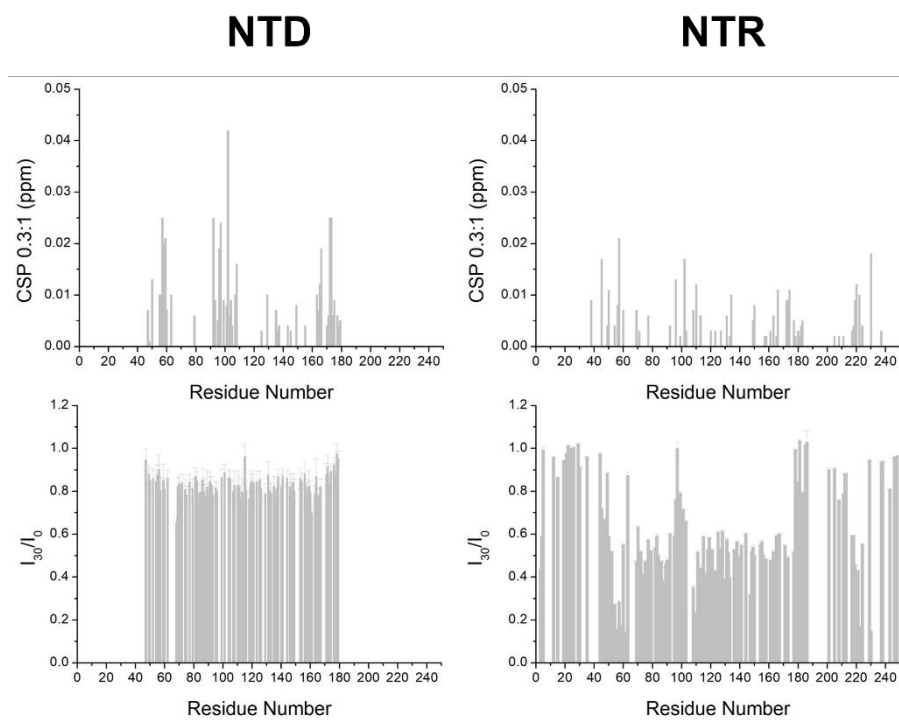
**Figure S3:** DOSY experiments [48] confirm the binding of heparin to NTD. The data obtained for the NTD protein alone are reported in blue, the ones determined for the NTD:EP adduct (1:9.6) are reported in magenta. The NTD:EP adduct has a slower diffusion time with respect to NTD as it can be appreciated comparing the dotted lines (light blue for the free form (upper), green for the bound form (lower)).



**Figure S4:** The four best structures of clusters 2 – 6 derived from the docking are reported in the picture. All these clusters present a HADDOCK [52,53] score which is lower with respect to Cluster 1.

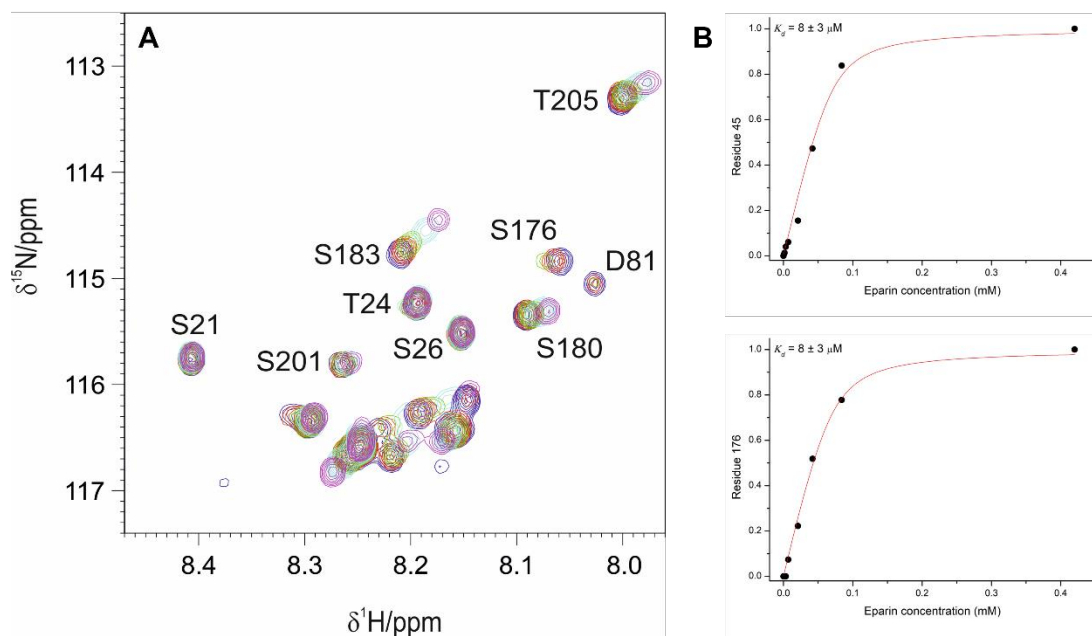


**Figure S5:** Comparison of CSP and intensity ratio for the two N constructs at the same protein:EP ratio (equal to 1:0.3). The data for NTD are reported on the left, the ones for NTR on the right.

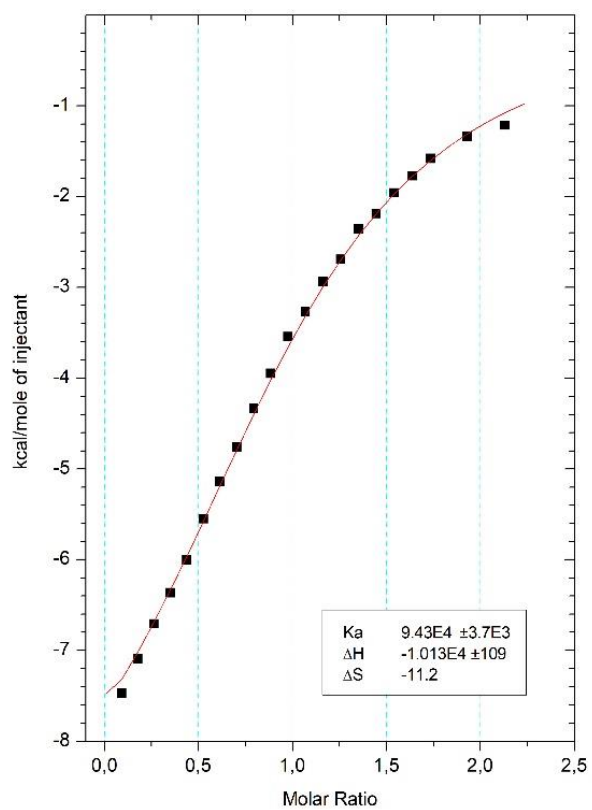


**Figure S6:** Panel A reports the overlay of NTR 2D HN spectra upon addition of EP. Blue, red, green, cyan and pink represent the 1:0, 1:0.1, 1:0.3, 1:06 and 1:1.2 molar ratio of NTR:EP respectively. The residues closest to the globular domain (NTD) experience the strongest perturbation in term of CSP and decrease in intensity (S176, S180, S183). Other resonances from residues belonging to the IDR2 are found to be perturbed as well (S201, T205). On the other hand, peaks in the initial part of IDR1 are not perturbed at all (S21, T24, S26). A subset of peaks falls in a very crowded region (ca. 116.5 ppm  $^{15}\text{N}$ ), where the resolution of 2D HN spectra is not enough to achieve information at the residue level.

Panel B reports the fitting of the  $K_d$  obtained from the CSP analysis for L45 and S176, presenting a  $K_d$  of  $8 \pm 3 \mu\text{M}$

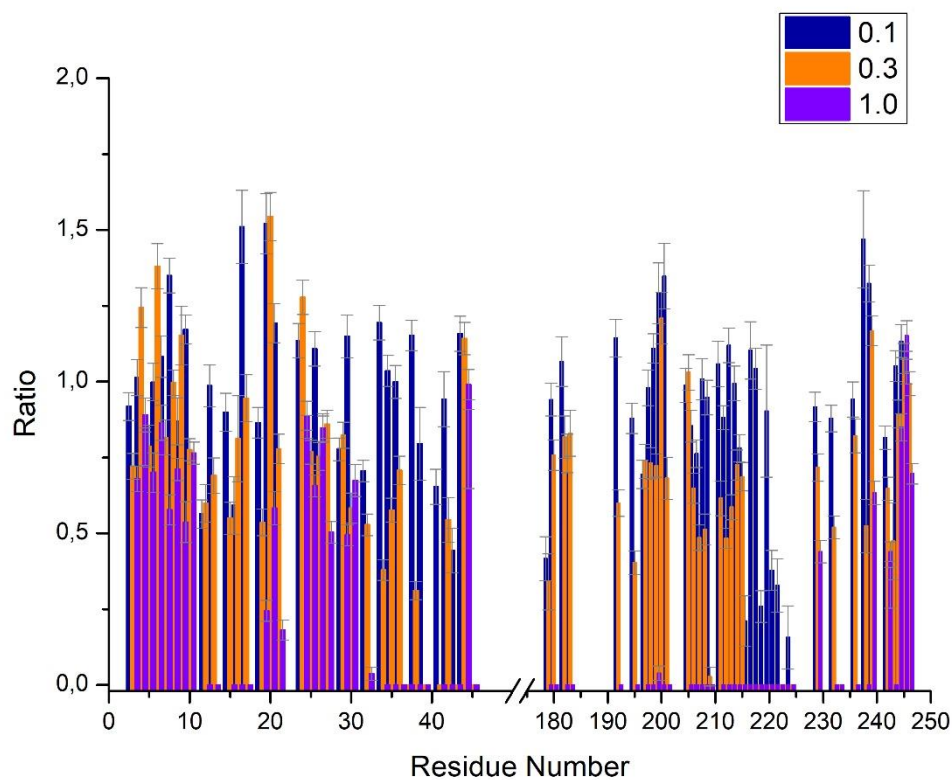


**Figure S7:** ITC experiments confirms a higher affinity for NTR with respect to NTD. The  $K_a$  of the interaction can be estimated in the order of  $10 \mu\text{M}$ .





**Figure S8:** Intensity ratios of cross peaks observed in 2D CON NMR spectra acquired with increasing amounts of EP. A subset of cross peaks shows a higher intensity in the presence of EP, in particular  $^3\text{DQ}^4$ ,  $^5\text{GP}^6$ ,  $^8\text{NQ}^9$ ,  $^{19}\text{GP}^{20}$ ,  $^{43}\text{QG}^{44}$ ,  $^{199}\text{PC}^{200}$ ,  $^{204}\text{GT}^{205}$ ,  $^{237}\text{KG}^{238}$  and  $^{238}\text{CQ}^{239}$  considering that the CON experiment reveals  $\text{C}'_{i-1}\text{-N}_i$  correlations. These are almost all “disorder-promoting” amino acids, with a large share of glycine and glutamine residues [79-81].



**Article 2.5:**

**Comprehensive Fragment Screening of the SARS-CoV-2 Proteome Explores Novel Chemical Space for Drug Development**

## SARS-CoV-2

How to cite:

International Edition: doi.org/10.1002/anie.202205858

German Edition: doi.org/10.1002/ange.202205858

# Comprehensive Fragment Screening of the SARS-CoV-2 Proteome Explores Novel Chemical Space for Drug Development

Hannes Berg<sup>+</sup>, Maria A. Wirtz Martin<sup>+</sup>, Nadide Altincekic<sup>+</sup>, Islam Alshamleh, Jasleen Kaur Bains, Julius Blechar, Betül Ceylan, Vanessa de Jesus, Karthikeyan Dhamotharan, Christin Fuks, Santosh L. Gande, Bruno Hargittay, Katharina F. Hohmann, Marie T. Hutchison, Sophie Marianne Korn, Robin Krishnathas, Felicitas Kutz, Verena Linhard, Tobias Matzel, Nathalie Meiser, Anna Niesteruk, Dennis J. Pyper, Linda Schulte, Sven Trucks, Kamal Azzaoui, Marcel J. J. Blommers, Yojana Gadiya, Reagon Karki, Andrea Zaliani, Philip Gribbon, Marcius da Silva Almeida, Cristiane Dinis Anobom, Anna L. Bula, Matthias Bütikofer, Ícaro Putinhon Caruso, Isabella Caterina Felli, Andrea T. Da Poian, Gisele Cardoso de Amorim, Nikolaos K. Fourkiotis, Angelo Gallo, Dhiman Ghosh, Francisco Gomes-Neto, Oksana Gorbatyuk, Bing Hao, Vilius Kurauskas, Lauriane Lecoq, Yunfeng Li, Nathane Cunha Mebus-Antunes, Miguel Mompeán, Thais Cristtina Neves-Martins, Martí Ninot-Pedrosa, Anderson S. Pinheiro, Letizia Pontoriero, Yulia Pustovalova, Roland Riek, Angus J. Robertson, Marie Jose Abi Saad, Miguel Á. Treviño, Aikaterini C. Tsika, Fabio C. L. Almeida, Ad Bax, Katherine Henzler-Wildman, Jeffrey C. Hoch, Kristaps Jaudzems, Douglas V. Laurents, Julien Orts, Roberta Pierattelli, Georgios A. Spyroulias, Elke Duchardt-Ferner, Jan Ferner, Boris Fürtig, Martin Hengesbach, Frank Löhr, Nusrat Qureshi, Christian Richter, Krishna Saxena, Andreas Schlundt, Sridhar Sreeramulu, Anna Wacker, Julia E. Weigand, Julia Wirmer-Bartoschek, Jens Wöhnert, and Harald Schwalbe\*

**Abstract:** SARS-CoV-2 (SCoV2) and its variants of concern pose serious challenges to the public health. The variants increased challenges to vaccines, thus necessitating for development of new intervention strategies including anti-virals. Within the international Covid19-NMR consortium, we have identified binders targeting the RNA genome of SCoV2. We established protocols for the production and NMR characterization of more than 80 % of all SCoV2 proteins. Here, we performed an NMR screening using a fragment library for binding to 25 SCoV2 proteins and identified hits also against previously unexplored SCoV2 proteins. Computational mapping was used to predict binding sites and identify functional moieties (chemotypes) of the ligands occupying these pockets. Striking consensus was observed between NMR-detected binding sites of the main protease and the computational procedure. Our investigation provides novel structural and chemical space for structure-based drug design against the SCoV2 proteome.

## Introduction

SARS-CoV-2 (SCoV2) is the cause for the COVID-19 pandemic resulting in more than 5 million deaths across the world and continues to pose serious challenges to public health and safety.<sup>[1]</sup> Countering the continuously evolving virus has not only seen an unprecedented success in the vaccine development but also given birth to several novel campaigns for anti-viral drug discovery,<sup>[2,3]</sup> including the recently approved oral antivirals paxlovid (Pfizer) and molnupiravir (Merck & Co.).<sup>[4-6]</sup>

The extensively mutated and highly infective variant of SCoV2, Omicron,<sup>[7]</sup> is resistant to several therapeutic antibodies,<sup>[8,9]</sup> evades double immunization,<sup>[8,10]</sup> and dominates the pandemic in 2022, calling for the development of new therapeutic strategies in combating the virus, specifically, by exploiting the conserved features.<sup>[11,12]</sup>

The SCoV2 genome consists of an ≈29.9 kb long positive-sense single-stranded RNA,<sup>[13]</sup> two-thirds of which comprises the open-reading frames (ORF) 1a and 1ab. Both ORFs encode polyproteins, which are proteolytically processed into 16 different non-structural proteins (nsp1-

- [\*] H. Berg,<sup>†</sup> M. A. Wirtz Martin,<sup>†</sup> N. Altincekic,<sup>†</sup> Dr. I. Alshamleh, J. Kaur Bains, J. Blechar, B. Ceylan, V. de Jesus, C. Fuks, Dr. S. L. Gande, B. Hargittay, K. F. Hohmann, M. T. Hutchison, R. Krishnathas, F. Kutz, V. Linhard, T. Matzel, N. Meiser, Dr. A. Niesteruk, Dr. D. J. Pyper, Dr. L. Schulte, Dr. S. Trucks, Dr. J. Ferner, Dr. B. Fürtig, Dr. M. Hengesbach, Dr. N. Qureshi, Dr. C. Richter, Dr. K. Saxena, Dr. S. Sreeramulu, Dr. A. Wacker, Dr. J. Wirmer-Bartoschek, Dr. Prof. Dr. H. Schwalbe  
Institute for Organic Chemistry and Chemical Biology, Goethe University Frankfurt, 60438 Frankfurt am Main (Germany)
- H. Berg,<sup>†</sup> M. A. Wirtz Martin,<sup>†</sup> N. Altincekic,<sup>†</sup> Dr. I. Alshamleh, J. Kaur Bains, B. Ceylan, V. de Jesus, K. Dhamotharan, Dr. S. L. Gande, B. Hargittay, K. F. Hohmann, M. T. Hutchison, Dr. S. Marianne Korn, R. Krishnathas, F. Kutz, V. Linhard, T. Matzel, Dr. A. Niesteruk, Dr. D. J. Pyper, Dr. L. Schulte, E. Duchardt-Ferner, Dr. J. Ferner, Dr. B. Fürtig, Dr. F. Löhr, Dr. N. Qureshi, Dr. C. Richter, Dr. K. Saxena, Dr. A. Schlundt, Dr. S. Sreeramulu, Dr. A. Wacker, Dr. J. Wirmer-Bartoschek, Prof. Dr. J. Wöhnert, Dr. Prof. Dr. H. Schwalbe  
Center of Biomolecular Magnetic Resonance (BMRZ), Goethe University Frankfurt, Frankfurt am Main (Germany)  
Max-von-Laue-Strasse 7+9, 60438 Frankfurt am Main (Germany)  
E-mail: schwalbe@nmr.uni-frankfurt.de
- K. Dhamotharan, Dr. S. Marianne Korn, E. Duchardt-Ferner, Dr. F. Löhr, Dr. A. Schlundt, Prof. Dr. J. Wöhnert  
Institute for Molecular Biosciences, Goethe University Frankfurt, Max-von-Laue-Strasse 7+9, 60438 Frankfurt am Main (Germany)
- Dr. K. Azzaoui, Dr. M. J. J. Blommers  
Saverna Therapeutics, Pumpmattenweg 3, 4105 Biel-Benken (Switzerland)
- Y. Gadiya, R. Karki, A. Zaliani, Dr. P. Gribbon  
Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Screening Port, Schnackenburgallee 114, 22525 Hamburg (Germany)
- Y. Gadiya, R. Karki, A. Zaliani, Dr. P. Gribbon  
Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor-Stern-Kai 7, 60596 Frankfurt am Main (Germany)
- M. da Silva Almeida, Í. Putinhon Caruso, A. T. Da Poian, G. Cardoso de Amorim, N. Cunha Mebus-Antunes, T. Cristina Neves-Martins, F. C. L. Almeida  
Institute of Medical Biochemistry, Federal University of Rio de Janeiro, 21941-902 Rio de Janeiro (Brazil)
- M. da Silva Almeida  
Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor-Stern-Kai 7, 60596 Frankfurt am Main (Germany)
- C. Dinis Anobom, F. Gomes-Neto\*, F. C. L. Almeida  
National Center of Nuclear Magnetic Resonance (CNRMN), CENABIO, Federal University of Rio de Janeiro, 21941-902 Rio de Janeiro (Brazil)
- C. Dinis Anobom, A. S. Pinheiro  
Department of Biochemistry, Institute of Chemistry, Federal University of Rio de Janeiro, 21941-902 Rio de Janeiro (Brazil)
- A. L. Bula, K. Jaudzems  
Latvian Institute of Organic Synthesis, Aizkraukles 21, LV-1006 Riga (Latvia)
- M. Bütikofer, D. Ghosh, R. Riek  
ETH, Swiss Federal Institute of Technology, Laboratory of Physical Chemistry, HCI F217, Vladimir-Prelog-Weg 2, 8093 Zürich (Switzerland)
- Í. Putinhon Caruso  
Multiuser Center for Biomolecular Innovation (CMB), Department of Physics, São Paulo State University (UNESP), 01049-010 São José do Rio Preto (Brazil)
- I. Caterina Felli, L. Pontoriero, R. Pierattelli  
Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, Sesto Fiorentino, 50019, Florence (Italy)
- I. Caterina Felli, L. Pontoriero, R. Pierattelli  
Department of Chemistry "Ugo Schiff", University of Florence, Via della Lastruccia 3-13, Sesto Fiorentino, 50019, Florence (Italy)
- G. Cardoso de Amorim  
Multidisciplinary Center for Research in Biology (NUMPEX), Campus Duque de Caxias Federal University of Rio de Janeiro, 25.250-470 Duque de Caxias (Brazil)
- N. K. Fourkiotis, A. Gallo, A. C. Tsika, G. A. Spyroulias  
Department of Pharmacy, University of Patras, 26504 Patras (Greece)
- A. Gallo  
Department of Chemistry, University of Torino IT-10126 Torino (Italy)
- F. Gomes-Neto\*  
Laboratory of Toxinology, Oswaldo Cruz Foundation (FIOCRUZ), 21040-900 Rio de Janeiro (Brazil)
- O. Gorbatyuk, B. Hao, Y. Li, Y. Pustovalova, J. C. Hoch  
Department of Molecular Biology and Biophysics UConn Health 263 Farmington Ave., Farmington, CT 06030-3305 (USA)
- V. Kurauskas, K. Henzler-Wildman  
Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706 (USA)
- L. Lecoq, M. Ninot-Pedrosa  
Molecular Microbiology and Structural Biochemistry, UMR5086 CNRS/Université Lyon 1, 7, passage du Vercors, 69367 Lyon (France)
- M. Mompeán, M. Á. Treviño, D. V. Laurents  
"Rocasolano" Institute for Physical Chemistry, CSIC 28006 Madrid (Spain)
- A. J. Robertson, A. Bax  
Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney  
20892-0520 Bethesda (USA)
- M. Jose Abi Saad, J. Orts  
University of Vienna, Department of Pharmaceutical Sciences, Josef-Holaubek-Platz 2, A-1090 Vienna (Austria)
- Dr. J. E. Weigand  
Department of Biology, Technical University of Darmstadt, 64289 Darmstadt (Germany)
- [†] These authors contributed equally to this work.
- © 2022 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

nsp16).<sup>[14,15]</sup> Four structural proteins: spike (S), envelope (E), membrane (M) and nucleocapsid (N) and nine additional accessory factors are expressed from the 13 ORFs located at the 3' end of the viral genome. In total, the viral genome encodes for at least 28 peptides or proteins.<sup>[16–18]</sup> Repurposing of (approved) drugs has been actively pursued as a strategy to counter SCoV2 infections,<sup>[19–22]</sup> however, with little clinical success.<sup>[23]</sup> Most of repurposed drugs were primarily an outcome of structure-based virtual screening campaigns and solely focused on a small fraction of the proteome, namely proteases (nsp3d, nsp5) or polymerase (nsp12) as targets.<sup>[24–30]</sup> Within the viral life cycle, the enzymes nsp3 (papain-like protease), nsp5 (main protease), nsp7·nsp8 (primase complex), nsp12 (primary RNA-dependent RNA polymerase (RdRp)), nsp13 (helicase), nsp14 (exoribonuclease) and the methyltransferases nsp14/nsp16 are important components of the replicase-transcriptase complex and hence are also listed as attractive drug targets.<sup>[16,31]</sup> X-ray crystallography and NMR have been successfully used to screen either fragments, approved drugs, or drugs in clinical trials, against a subset of key SCoV2 protein drug targets like nsp5, nsp3b, nsp13 and nsp14.<sup>[32–40]</sup>

The current drug development has typically focused its efforts around the two key viral proteins, a protease (nsp5) and a polymerase (nsp12, RdRp), and soon such a monotherapy can result in the virus developing resistance against the first-generation antivirals, thus warranting us to develop new antivirals involving different targets.<sup>[41]</sup> Recently, using a range of biochemical assays, several drugs were identified as inhibitors against a total of seven enzymes of SCoV2.<sup>[42–49]</sup> Therefore, developing drugs or synergistic combinations involving multiple viral targets appears as a viable therapeutic strategy for the treatment of COVID-19.<sup>[2,3,50]</sup>

Within the Covid19-NMR consortium, we undertook a massive NMR-based ligand screening with the aim of identifying fragments as new chemical entities targeting SCoV2 proteins. Previously, via consorted efforts between NMR groups worldwide we have successfully developed protocols for large-scale production of more than 80 % of all SCoV2 proteins.<sup>[51]</sup> Soon, the availability of proteins and the experience gained from the completion of >20 screens with the DSI-PL fragment library for binding against the viral RNA<sup>[52]</sup> positioned us to embark on this massive screening campaign. For this purpose, >20 SCoV2 proteins (nsp1, nsp2 (CtDR), nsp3a, nsp3b, nsp3b·GS-441524, nsp3c (SUD-N), nsp3c (SUD-MC), nsp3d, nsp3e, nsp3Y, nsp5, <sub>GHM</sub>nsp5, <sub>GS</sub>nsp5, nsp7, nsp8, nsp9, nsp10, nsp10·nsp14, <sub>His6</sub>nsp15, nsp10·nsp16, ORF9a (IDR1-NTD-IDR2), ORF9a (NTD), ORF9a (NTD-SR), ORF9a (CTD), ORF9b; (for definitions see Supporting Information Table 1) were produced in NMR groups at sites all over the world and subsequently shipped to the Frankfurt NMR center (BMRZ) for conducting the NMR screening. We applied ligand-observed <sup>1</sup>H NMR experiments and identified 311 binders across the 25 screened SCoV2 proteins. Further, we used FTMap,<sup>[53]</sup> a computational mapping server which has been proven to be more accurate than the conventional GRID and MCSS methods to identify binding sites (or hot spots) on macromolecules (protein, DNA or RNA). Active sites in enzymes

are usually concave surfaces that are suitable for ligand binding and therefore, in our study, binding site, hot spot, and active site are used interchangeably. FTMap predicts chemical scaffolds and functional units occupying these binding pockets. A comparison of the predicted scaffolds and functional units with the constitution of the experimental fragment hits for which we detected binding in our experimental screens showed striking correlation, as exemplified by comparing predicted and experimentally determined binding pockets for the main protease nsp5, the latter obtained both from crystallographic screens<sup>[54]</sup> as well as NMR protein-based screens conducted here. We thus propose this novel methodology for the analysis of ligand binding capability across multiple protein targets as provided in this work. Such methodology bears excellent potential to act as a unique resource for developing novel inhibitors.

## Results and Discussion

We conducted fragment-based screenings for a large number of SCoV2 viral proteins (Table 1 and Supporting Information Table 1). The viral proteins can be classified broadly into three different classes, namely, (i) proteases, (ii) replicase-transcriptase (RT) complex proteins and (iii) other accessory proteins. The main protease (nsp5, Mpro, CLpro) and the Papain-like protease (nsp3d, PLpro) are two important viral proteases that play a functionally important role in viral maturation.<sup>[55,56]</sup> Nsp5 is responsible for the cleavage of 12 nsp5 (nsp4-nsp16) and therefore represents one of the most attractive drug targets. We screened three different constructs (nsp5, <sub>GS</sub>nsp5 and <sub>GHM</sub>nsp5) of nsp5. The two (<sub>GS</sub>nsp5) or three (<sub>GHM</sub>nsp5) additional amino acids in the N-terminus resulted from cloning. SEC-MALS analysis of these two proteins revealed that they are monomeric in solution compared to the dimeric wildtype nsp5.<sup>[51]</sup> Recently, it has been shown that the monomer-dimer equilibrium is coupled to the catalytic activity of nsp5, with maximum activity associated with the dimeric state.<sup>[57]</sup> Therefore, identifying small molecules that interfere with the dimer formation is considered as an alternative strategy to impair catalytic activity<sup>[58]</sup> and so screening of both monomeric and dimeric states of the proteins may act as a valuable tool in identifying and developing allosteric ligands. Nsp3d is responsible for the cleavage of the N-terminus of the polyprotein, releasing nsp1, nsp2 and nsp3 and is therefore also a potential drug target. The RT-complex is composed of multiple enzymes, and we screened the SCoV2 putative primases (nsp7 and nsp8) and the methyltransferases (nsp14 and nsp16) in complex with its co-factor nsp10 (nsp10·nsp14, nsp10·nsp16). The other screened set of proteins included several nsp5, various domain constructs of nsp3 and structural and accessory proteins (ORF9a (N-protein) and ORF9b). The molecular weight of the screened proteins ranged between 5 kDa (nsp2 (CtDR)) to 78 kDa (nsp10·nsp14). Further, the 25 screened proteins also included intrinsically disordered proteins (nsp2 (CtDR)), proteins with intrinsically disordered regions (N-protein),

**Table 1:** SCoV2 protein constructs screened by NMR.

Protein <i>genome</i> <i>position (nt)</i> <sup>[a]</sup>	Trivial name Construct expressed	Size (aa) <sup>[b]</sup>	Boundaries	MW [kDa]	PDB code used for FTMap	Number of binders identified	Crossclusters in Cleft1	Crossclusters in Cleft2
nsp1 266–805	<i>Leader</i>	180		19.8				
	Globular Domain (GD)	116	13–127	12.7	7k7p	5	0, 7	1, 2, 3
nsp2 806–2,719		638		70.5				
	C-terminal IDR (CtDR)	45	557–601	4.9	-	19	-	-
nsp3 2,720–8,554		1,945		217.3				
	Ub-like (UBI) domain	111	1–111	12.4	7kag	14	3, 6	0, 5, 8, 10
	nsp3b (Macro domain)	170	207–376	18.3	6vxs	10	0, 1, 2, 3, 5	-
	nsp3b·GS- 441524	170	207–376	18.3	6vxs	5	-	-
c	SUD-N	140	409–548	15.4	2w2 g	10	0, 2, 4, 5, 6, 9	-
c	SUD-MC	193	551–743	21.5	2kqv	154	1, 2, 3, 4, 6	0
d	Papain-like protease PL <sup>pro</sup>	318	743–1,060	36	6w9c	150	5, 7	1, 2, 4
	NAB	116	1,088–1,203	13.4	2k87	21	1, 4 (Cleft 3)	-
e Y nsp5 10,055– 10,972		286		31.5		81	-	-
	<i>Main protease (M<sup>pro</sup>)</i>	306		33.8				
	<sub>CS</sub> nsp5	306	1–306	33.8	-	12	-	-
	<sub>CHM</sub> nsp5	306	1–306	33.8	-	38	-	-
	Full-length	306	1–306	33.8	5r83	78	3, 4, 6	1, 2, 7, 8
nsp7 11,843– 12,091		83		9.2				
	Full-length	83	1–83	9.2	2kys	92	0, 1, 3, 6	-
nsp8 12,092– 12,685		198		21.9				
	Full-length	198	1–198	21.9	6wiq	35	1, 3, 4, 5, 6	-
nsp9 12,686– 13,024		113		12.4				
	Full-length	113	1–113	12.4	6w4b	2	1, 3	0, 2, 4
nsp10 13,025– 13,441		139		14.8				
	Full-length	139	1–139	14.8	6zpe	38	0, 3, 5, 6	-
nsp15 19,621– 20,658	<i>Endonuclease</i>	346		38.8				
	<sub>HIS6</sub> nsp15	346	1–346	38.8	6w01	42	1, 2	4
nsp10·nsp16 20,659– 21,552	<i>Methyltransferase</i>	298		33.3				
	nsp10·nsp16	298	1–298 (nsp16)	33.3	6w4 h	92	3, 4, 5, 7, 8, 10	0
nsp10·nsp14 18,040– 19,620	<i>Exoribonuclease</i>	527		61.4				
	nsp10·nsp14	527	7–527 (nsp14)	61.4	modelled	44	2, 5, 9 (Cleft 3)	-
ORF9a 28,274– 29,533	<i>Nucleocapsid (N)</i>	419		45.6				
	IDR1-NTD-IDR2	248	1–248	26.5	6yi3	7	-	-



Table 1: (Continued)

	NTD-SR	169	44–212	18.1	6yi3	5	–	–
	NTD	136	44–180	14.9	6yi3	32	0, 1, 3, 5, 6, 7	2
	CTD	118	247–364	13.3	7c22	9	1, 2, 6, 8	–
ORF9b 28,284– 28,574		97		10.8				
	Full-length	97	1–97	10.8	6z4u	8	0, 3, 5 (Cleft 3)	–

[a] Genome position in nucleotide (nt) corresponding to SCoV2 NCBI reference genome entry NC\_045512.2, identical to GenBank entry MN908947.3. [b] number of amino acids excluding the additional residues due to cloning

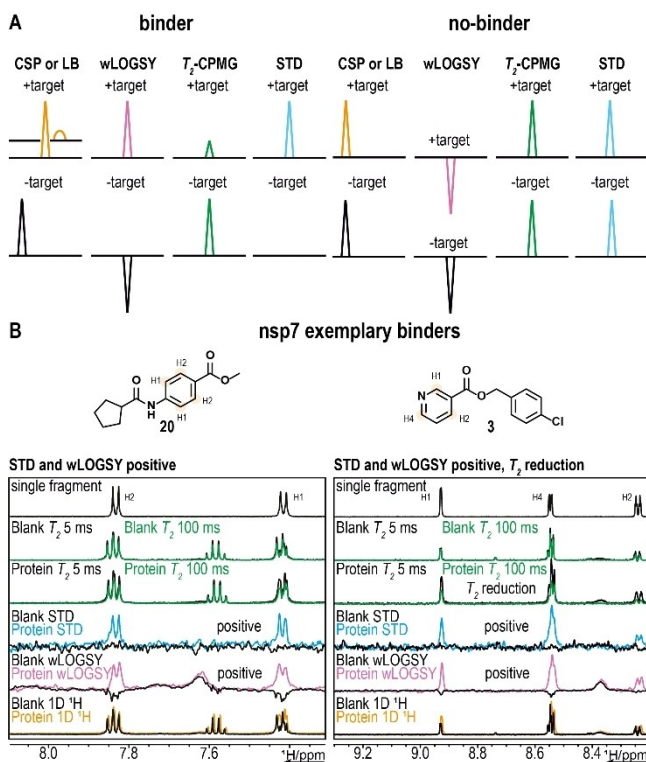
and even a protein-inhibitor complex (nsp3b·GS-441524) with the quest to identify ligands binding in close proximity to the nucleotide binding pocket as starting point for fragment growth medicinal chemistry.

The DSI-poised library (DSI-PL, Supporting Information, excel sheet 1 DSI PL Poised Library.xlsx)<sup>[59–61]</sup> has already been successfully used to screen the druggability of the RNA regulatory elements and the main protease nsp5 from SCoV2.<sup>[52,54]</sup> This library is composed of 768 highly diverse and poised fragments specifically designed to

facilitate easy downstream synthesis. We applied ligand-observed <sup>1</sup>H NMR experiments and performed the screening with 64 mixes containing 12 fragments each as described previously.<sup>[52]</sup> In these screening experiments, changes in the <sup>1</sup>H signals of the ligand in the presence and absence of the protein served as readout for binding.

For identifying binders within the mixtures, we first compared spectra from four different NMR experiments and analyzed differences by visual inspection. As criteria, chemical shift perturbations (CSPs) or severe line broadening, sign change in the waterLOGSY (wLOGSY), STD signal or significant decrease of signal intensity in a  $T_2$ -relaxation experiment were used to identify binders (Figure 1A). Ligands were assigned as a binder if one of the four criteria was satisfied. For example, binder 20 qualifies as a binder, showing changes in wLOGSY and STD, while only minor CSP and change in  $T_2$  (Figure 1B, left). Similarly, binder 3 qualifies as a hit, displaying changes in wLOGSY, STD and  $T_2$ , but no CSP (Figure 1B, right).

NMR-based screening resulted in 311 binders across the 25 screened SCoV2 proteins (Figure 2). Our results show that the overall binders identified against a target ranged from 2 (nsp9) to 154 (nsp3c (SUD-MC)). No correlation was observed between the molecular weight of the target and the number of binders (Supporting Information Figure 1). Strikingly, the intrinsically disordered domain of nsp2 (CtDR) shows 19 binders. By contrast, the well folded protein nsp3b has only 3 binders. The protease nsp3d and the nsp3c (SUD) as a didomain with its middle and C-terminus (MC), are amongst those with the largest number of binders (Supporting Information Table 2). The nsp3b (macro domain) is evolutionarily conserved and regarded as

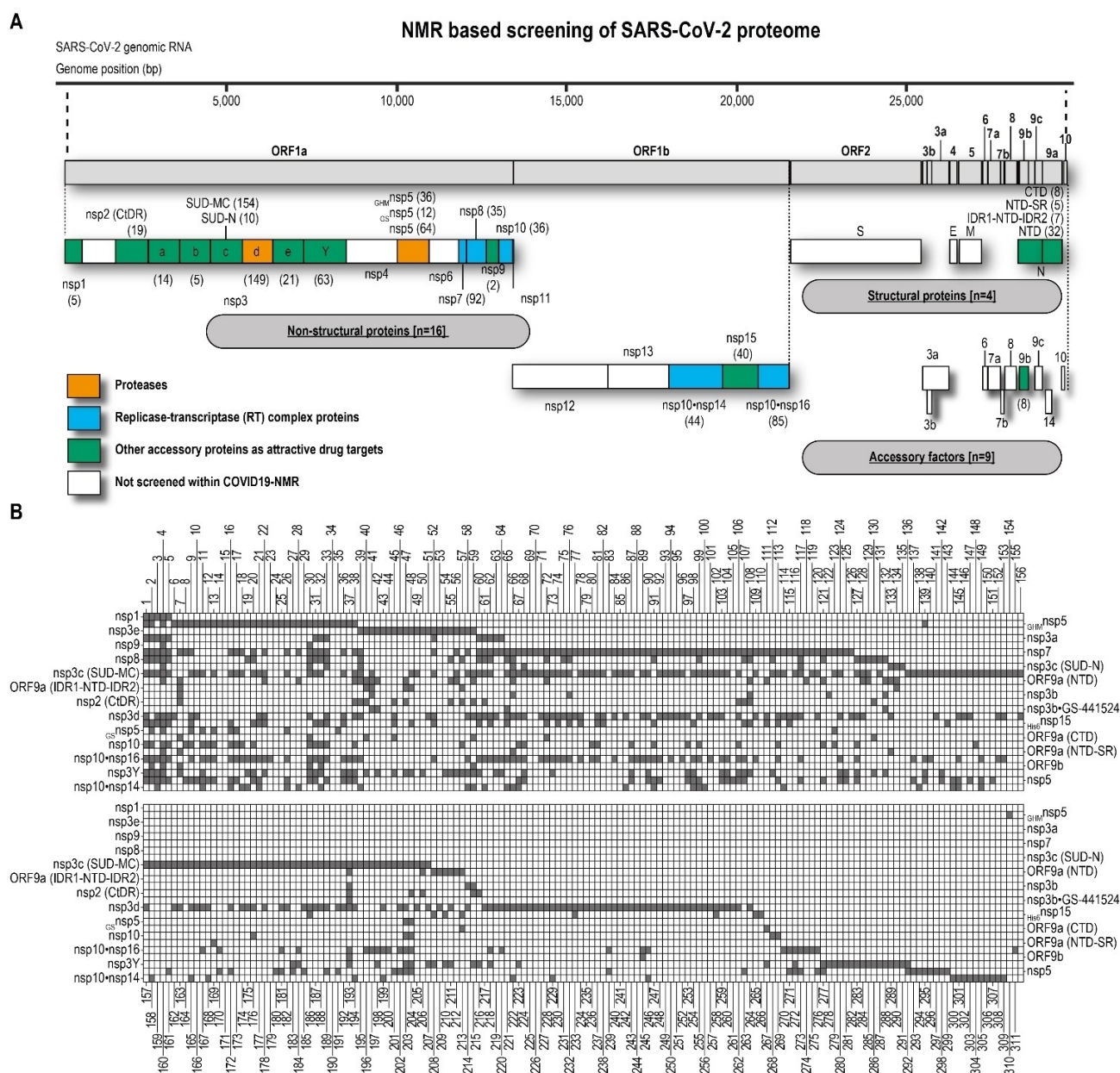


**Figure 1.** NMR based identification of binding fragments. A) Schematic representation of all NMR experiments used in the screening that show exemplary effects indicating binding events in the presence of ligand compared to ligand free spectra. B) NMR spectra (1D <sup>1</sup>H, wLOGSY, STD, and  $T_2$ -CPMG (5 ms and 100 ms) and chemical structure (binder 20 and binder 3) of two binding fragments identified for nsp7. Single fragment spectra (top) are used for chemical shift deconvolution in the mixture. Binder 20 shows clear sign changes in the STD and wLOGSY in presence of nsp7 protein. Binder 3 also shows signal in the STD and a sign change in the wLOGSY, as well as a  $T_2$  reduction of approximately 50% in presence of nsp7 protein.

Table 2: Affinities of the SCoV2 protein binders.

Ligand observed	nsp3c (SUD-MC) <sup>[a]</sup>	nsp5 <sup>[a]</sup>	Protein Observed	Bindersnsp5 <sup>[a]</sup>	nsp10 <sup>[a]</sup>
BinderORF9a (NTD) <sup>[a]</sup>					
40	> 5	–	–	21	0.46 ± 0.04
129	> 5	–	–	32	–
209	> 5	–	–	2	1.70 ± 0.54
68	–	0.45 ± 0.71	–	–	–
30	–	> 5	–	–	–
13	–	–	0.02 ± 0.007	–	–
26	–	–	> 5	–	–

[a]  $K_D$  in millimolar.



**Figure 2.** 311 binding fragments identified for SCoV2 proteins from NMR based fragment screening. A) Schematic representation of the SCoV2 genome (adapted from<sup>[6]</sup>). B) The two tables summarize all binding fragments identified in the NMR screening for their corresponding protein (grey). The first table shows binder 1 to 156 (columns) and the corresponding bound proteins (right and left rows). The second table shows binder 157 to 311 and the corresponding proteins (left and right rows).

a potential drug target. We conducted screening in its apo/free state and in the presence of GS-441524, the active drug and metabolite of remdesivir. We observed one common binder (binder 41) and two and four unique binders, respectively (Supporting Information Table 2). The main protease nsp5 is a dimeric cysteine protease and its N-terminus forms a part of the dimer interface. Subtle changes in the amino acid sequence at the N-terminus influence the oligomeric state ( $G_S$ nsp5 and  $G_{HM}$ nsp5, monomeric; nsp5, dimeric) of the protein.<sup>[51]</sup> For the three (nsp5,  $G_S$ nsp5 and  $G_{HM}$ nsp5) screened constructs we identified 78, 12, and 38 binders, respectively. Only 8 binders overlapped (Supporting Information Figure 2) between the three constructs,

suggesting that indeed there are differential surfaces exposed for ligand binding, which in turn stems from the monomer/dimer state of the protein constructs.<sup>[51]</sup> Previously, using the DSI-PL, nsp5 and nsp14 have been screened by crystallography identifying 39<sup>[54]</sup> and 41<sup>[38]</sup> binders, respectively. In contrast, 78 binders were identified by NMR for the identical construct of nsp5, and for a subset of these identified binders crystallization could be reproduced in house.

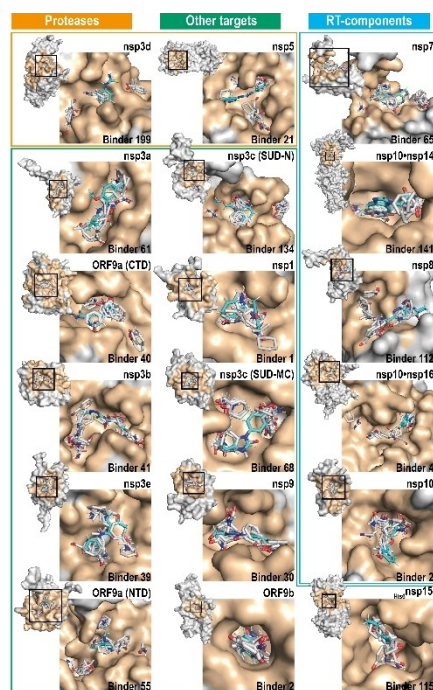
A comparison of the binders revealed 6 common binders including two 3-aminopyrimidine-like compounds (21 and 26) that form the chemical starting points within the COVID moonshot initiative.<sup>[40]</sup> The twice as large number of binders



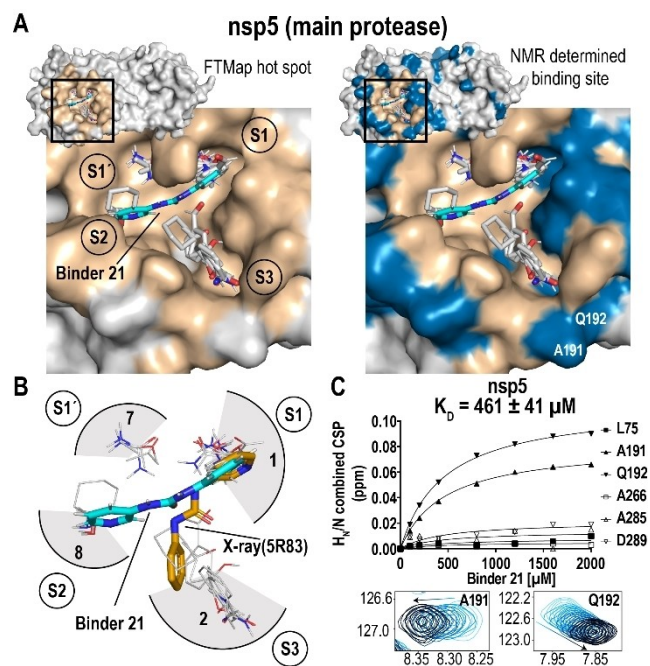
identified by NMR is potentially attributed either to the presence of multiple stable conformations of nsp5 in solution<sup>[62]</sup> or to the fact that the different NMR-based screening experiments can identify binders within different affinity regimes (low micromolar to high millimolar). For nsp10·nsp14, we identified 44 binders with only one binder (binder 168) overlapping with the X-ray hits, wherein the screening was performed in the absence of nsp10. Further, 7 overlapping binders were found between nsp10·nsp14 and nsp10 NMR screens (Supporting Information Figure 2). Given the fact that significant conformational differences exist between nsp14 and nsp10·nsp14 structures,<sup>[38]</sup> it is not surprising that different sets of binders are identified in X-ray and NMR screens. Further, NMR competition experiments with sinefungin, a methyltransferase inhibitor and structural analog of s-adenosyl methionine (SAM), identified that binder 141 and 146 bind to the SAM binding site.

The relatively diverse and varying number of binders across the screened SCoV2 proteins in this work is likely correlated to the accessible surface of a given protein. In general, proteins that routinely bind to either small molecules or substrates to perform their function have well-defined cavities and pockets. For example, the cysteine protease nsp5 and the nsp3b (macro domain) each have a substrate or endogenous ligand binding cleft that both are currently exploited for designing functional inhibitors. Traditionally, ligand binding pockets in proteins are deter-

mined experimentally either by X-ray crystallography or NMR. Such experimental identification of binding pockets for large sets of binders across several targets of SCoV2 reported here would be very time-consuming and sample intensive. Thus, we deduced the ligand binding sites of the SCoV2 proteins using FTMap.<sup>[53]</sup> FTMap uses 16 small organic molecules (Supporting Information Figure 3) as probes to scan the surface of the protein target and to identify regions that bind multiple of these probes, thus forming a probecluster. Several probeclusters which are in close proximity on the protein surface form one crosscluster, thus defining a consensus site or hot spot. We performed the FTMap analysis for the 18 of the 25 screened proteins for which structural coordinates were available (Supporting Information Table 3). Except for nsp3e, the pdb structures for all proteins were from SCoV2. Further, for structures with multiple chains but with the same sequence (for example: dimer) the FTMap protocol recommends each chain to be independently mapped and therefore a single monomer unit was used for all the proteins except for nsp5, ORF9a (CTD) and ORF9b that is known to exist as a stable dimer in solution and both monomeric and dimeric state were analyzed. Typically, one to three binding sites (Supporting Information Figure 4 to 21) were identified for each of the proteins. For example, the binding sites in monomeric nsp5 clustered mainly around three distinct

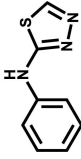
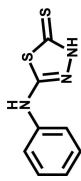
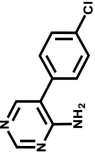
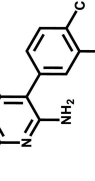
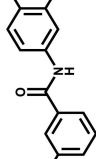
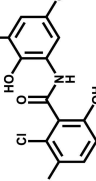
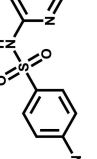
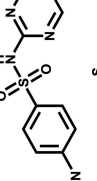
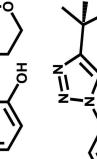
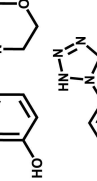
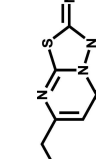
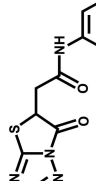
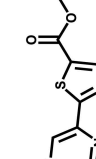
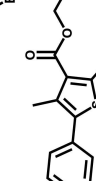
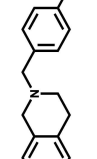
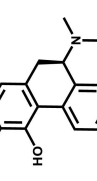




**Figure 3.** Hot spots identified using FTMap along with the docking of the NMR identified binders for 18 SCoV2 proteins. Proteases are highlighted with an orange box, RT-components with a blue box, and other targets with a green box. Zoom-ins show one of the identified clefts (beige colored) from PDBSum with its corresponding hot spots (and probes in grey sticks) from the FTMap analysis. For each of the targets, one of the binders was docked using SwissDock, shown in cyan.



**Figure 4.** Agreement between bioinformatic and experimental mapping of the binding site. A) The FTMap identified hot spot for nsp5. The subsites of the active site are labeled as S1, S1', S2 and S3. The crossclusters (1, 2, 7, and 8) occupying the binding site are shown in grey sticks. The docked pose of binder 21 is shown in cyan. Mapping of the CSPs (in blue) on to the structure of nsp5. B) Active site of nsp5 with an overlay of a docking (cyan) and X-ray determined (orange) structure of binder 21. C) The interaction of binder 21 and nsp5 was monitored via NMR titration. Binder 21 binds to nsp5 with a  $K_D$  of 461  $\mu$ M. The inset shows two shifting peaks (A191 and Q192) with increasing concentration of binder 21 (light blue-low to black-high).

**Table 3:** Fragment hits from NMR-based screening and related analogues identified as biologically active compounds in SCoV2 related assays in public databases.

Binder Structure	Binding targets detected by NMR	Bioactive analogue	CHEMBL Compound ID	Name	Tanimoto score	IC <sub>50</sub> related BioAssay (nM)	Assay Description
	nsp7, nsp3c (SUD-MC), nsp3d, nsp10, nsp10·nsp16, nsp5		CHEMBL289356	CL-17107	0.88	390	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC <sub>50</sub> FRET format with a peptide substrate
	nsp7, nsp3c (SUD-MC), nsp3d, nsp15, nsp10·nsp16		CHEMBL264373	Metoprine	0.85	2340	Cell based, SCoV2 induced cytotoxicity of VERO-E6 cells after 48 hours exposure to 0.01 MOI SCoV2 virus by high content imaging
	nsp7, nsp3c (SUD-MC), nsp3d, nsp15, nsp10·nsp16		CHEMBL2105450	Oxyclozanide	0.82	3710	Cell based, Antiviral activity against SCoV2 (viral titer) measured by plaque assay in Vero cells at MOI 0.0125 after 24 hr
	nsp3c (SUD-MC)		CHEMBL1382627	silver-sulfadiazine	0.81	750	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC <sub>50</sub> FRET format with a peptide substrate
	nsp3e, nsp3y		CHEMBL1380480	VANITOLIDE	0.74	1320	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC <sub>50</sub> FRET format with a peptide substrate
	nsp5, nsp3d		CHEMBL226652	4-DAMP	0.72	2360	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC <sub>50</sub> FRET format with a peptide substrate
	nsp3c (SUD-MC), nsp3d		CHEMBL243652	PD096194	0.72	2040	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC <sub>50</sub> FRET format with a peptide substrate
	nsp5, nsp3y, nsp10·nsp16, nsp3e, nsp3y		CHEMBL566136	PD121351	0.71	3190	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC <sub>50</sub> FRET format with a peptide substrate
	nsp3e, nsp7, nsp3c (SUD-MC), ORF9a (NTD), nsp3y		CHEMBL1616	APOMORPHINE HYDROCHLORIDE	0.71	520	Biochemical, nsp5 (SCoV2 3CL-Pro protease inhibition) IC <sub>50</sub> FRET format with a peptide substrate

regions of the protein, including the already known catalytic active site (Supporting Information Figure 20). However, FTMap analysis performed on the dimeric nsp5 does not identify the catalytic site (Supporting Information Figure 22 and Supporting Information Figure 23), which is in line with one of the limitations of FTMap that it works best for single domains. Therefore, monomeric form of nsp5 was utilized for the analysis of druggability. For nsp3b, hot spots clustered mainly in the ADPr binding site (Supporting Information Figure 13). Similarly, we observed the same (previously known and additional binding pockets) trend of hot spot clustering in the other proteins of SCoV2, which facilitated the definition of the relevant clefts on the protein. We used PDBsum<sup>[63]</sup> to calculate the cleft regions and ranked the clefts according to their volume. Integration of the PDBsum derived cleft information and the FTMap-identified binding sites strikingly revealed that for 13 out of 18 proteins, the hot spots identified by FTMap overlapped with cleft 1, for three proteins with cleft 2 and for three proteins with cleft 3 as identified by PDBsum (Figure 3 and Supporting Information Table 4). Importantly, FTMap analysis together with the cleft analysis for each of the SCoV2 proteins investigated here revealed that indeed, the 18 proteins contain defined potential ligand binding sites and are thus druggable. As a next step, for a given hot spot, we compared and correlated the types of FTMap probes predicted to bind in the binding sites with the chemical substructures present in the experimentally identified fragments in the DSI-PL. For this purpose, we scanned and extracted the number of occurrences of the 16 FTMap probes for all the 768 compounds from the DSI-PL using cheminformatic tools (Supporting Information excel sheet 2 DSI PL Poised Library Characterized into the 16 Probes of FTMap.xlsx). As a next step, for each of the identified binder for a given target, we quantified the overlap of probes between the hits and FTMap probes (Supporting Information excel sheets). We then selected one binder for each target, for which binding effects were observed in one or more NMR experiment. Mapping of the ligand-derived functional units revealed that for 14 out of 18 of these ligands, a 100 % correlation was observed with the probes found within one or more of the crossclusters spanning the predicted cleft (Figure 3 and Supporting Information Table 4). For example, binder 21 showed positive binding effects in both wLOGSY and STD NMR experiments for nsp5 and was hence chosen as ligand of choice for this target. FTMap and cleft analysis of nsp5 suggested that crossclusters 1, 2, 7, and 8 were situated within the known active site (cleft 2) of the protease. Binder 21 is composed of mainly three (methanamine, benzene and urea) FTMap probes, and all of them are present in the crosscluster 1 (100 %). The crossclusters 2, 7 and 8 each consist of one of the three probes (33 %). These observations show that there is a good overlap between the chemical substructures of the FTMap ligands and those experimental fragments that occupy the hot spots, suggesting a likely binding site for this ligand. Further, in order to gain insight into the binding site of the ligand, we performed molecular docking using the Swissdock web server.<sup>[64,65]</sup> For 50 % of the targets, we

observed that the top-ranked pose (i.e., the ligand with the lowest binding free energy) of the ligand docks onto the binding site (Figure 3, docked ligand shown in cyan).

In order to test the validity of our predicted ligand binding sites, we performed ligand-observed (ORF9a (NTD), nsp3 (SUD-MC) and nsp5) and/or protein observed (nsp5 and nsp10) titrations and determined the dissociation constants for a subset of targets by NMR. In general, the dissociation constants  $K_D$  for the fragments ranged from 50 to 2000  $\mu\text{M}$  (Table 2 and Supporting Information Figure 24). Binder 13 (Z979145504) bound to nsp5 with the highest affinity. In addition, we also performed protein-observed titrations for ligands that bind to nsp5 and nsp10. An advantage of protein-observed NMR titrations is that apart from obtaining information on the dissociation constants, it is also possible to visualize the binding site of the ligand by mapping the CSPs, provided the backbone amides are assigned. Previously, within the Covid19-NMR consortium we have achieved the near-to-complete backbone assignments of nsp10 and nsp5.<sup>[36,62,66]</sup> Binder 21 was titrated to nsp5 and bound with a  $K_D$  of  $\approx 500 \mu\text{M}$  (Figure 4, bottom right). Mapping of the CSPs revealed that apart from remote CSP effects, the residues involved in the binding mainly clustered around the active site (Figure 4, top right, blue regions), which was in good agreement with the binding cleft identified by FTMap. Moreover, FTMap and cleft analysis of nsp5 not only identified the same two sites (S1 and S3) in line with the crystal structure of binder 21 in complex with nsp5 (Figure 4, lower left, orange stick), but also reveals two additional sites (S1' and S2). A similar analysis performed for a weak binder (binder 2,  $K_D$  of  $\approx 2000 \mu\text{M}$ ) of nsp10 (Supporting Information Figure 25) reveals a striking correlation between the binding site mapped based on NMR CSPs and the FTMap-detected hot spot, thus supporting the robustness and validity of our analysis. Further, FTMap analysis of the 6 and 8 overlapping binders for X-ray/NMR screening and three nsp5 constructs, respectively, suggests, that the active site (cleft 2) is their putative binding site (Supporting Information Table 5 and Supporting Information Table 6). Moreover, the 6 X-ray/NMR overlapping binders revealed identical docking poses for single chains of either monomeric (5r83) or dimeric (7khp) structures as documented in Supporting Information Figure 26.

The NMR-based fragment hit structures were compared to >2 million molecules contained in the ChEMBL,<sup>[67]</sup> PubChem<sup>[68]</sup> and NCATS (<https://opendata.ncats.nih.gov/covid19/>) associated data resources of bioactive compounds. 2D Tanimoto scoring<sup>[69]</sup> was used to identify analogues annotated as active in SCoV2 bioassays. To capture “weak associations” between hits and bioactive analogues, a cut-off of 0.65 was set, which revealed 35 hit fragments associated with 50 analogues identified as active in 16 different SCoV2 assays, representing a total of 154 distinct bioactivities (Supporting Information excel sheet 3 Hits to Bioactivities.xlsx). A knowledge graph additionally annotated with links to public SCoV2 assay information and relevant metadata on the bioactivities and primary targets of the 154 compounds can be accessed at <https://github.com/Fraunhofer-ITMP/COVID-NMR-KG>. At a more stringent



Tanimoto cut-off of 0.70, a group of 9 hit fragments representing 9 analogues were identified (Table 3). Seven of the analogues, with  $IC_{50}$  values between 390 nM and 3190 nM, were identified as inhibitors of protease activity, in the study by Kuzikov et al.,<sup>[70]</sup> who screened a compound repurposing collection in a FRET-based biochemical assay against full-length nsp5. Although the fragment hits binding to nsp5 also binds to at least one additional protein, three (binder 6, 37 and 67) have analogues that inhibit nsp5 activity. Two analogue compounds were also active in phenotypic assays monitoring the anti-cytopathic effect of SCoV2 in Vero E6 cell models (Metoprine,  $IC_{50}$  = 2340 nM and Oxyclozanide,  $IC_{50}$  = 3710 nM.<sup>[71]</sup> The NMR hit (binder 74) related to Metoprine, binds multiple proteins (nsp7, nsp3c (SUD-MC), nsp3d, His6nsp15, nsp10 and nsp16) whilst the Oxyclozanide related compound (binder 79) targets a smaller group of viral proteins, namely nsp7 and nsp3c (SUD-MC).

## Conclusion

Covid19 has triggered enormous research efforts. For the less than 30 viral proteins and 15 conserved RNA regulatory elements, holistic approaches screening almost all viral components can be pursued. X-ray crystallography with recently introduced automatization of fragment screening approaches<sup>[33,54]</sup> has spearheaded medicinal chemistry approaches focusing on a subset of the viral protein targets. Previously (Sreeramulu, Richter et al.) and here, we exploit the unique advances of NMR spectroscopy for screening of structured elements of the RNA genome as well as the soluble parts of the proteome. The work described thus provides information for  $25 \times 768 = 19200$  possible protein-ligand interactions monitored by 4 different ligand-based NMR experiments. The 768 ligands come from a highly privileged fragment library. They have been assembled previously and validated by NMR for their chemical purity and solubility.<sup>[59,60]</sup>

The screening identifies 311 hits (1.5% overall hit rate). The work goes, however, beyond reporting these screening results. We delineate a procedure to combine computational methods to validate binding site prediction from FTMap and PDBsum with the experimentally detected binding ligands. This procedure relies on the prediction of chemical submoieties essential for binding and the similarity of these substructures in the set of experimental binders. The thus identified and prioritized binding sites allow application of focused docking protocols and further, the experimental cross-validation by protein-based NMR experiments. From these protein-based NMR experiments, we show that dissociation constants of these fragments with proteins range from 80  $\mu$ M to several millimolar. The determination of binding affinities can be used to prioritize medicinal chemistry campaigns. Using bioinformatics, identification of fragment binders also serves as starting point for database searches of known binders, using chemical similarity scores between fragments and known inhibitors as selection criterion. Thus, the herein developed workflow allows for

holistic screening of the majority of the viral proteome. It provides highly valuable data for the day-to-day support of medicinal chemistry campaigns aiming at developing novel drugs applying fragment-based drug discovery. These data will also serve development of artificial intelligence (AI) based algorithms to inform hit-to-lead campaigns.

## Acknowledgements

The work has been conducted in the international consortium of Covid19-NMR (covid19-nmr.de). We would like to thank Roderick Lambertz and Katharina Targaczewski for technical assistance. We thank Peter Maas and his team at SPECS for assembling the fragment library. Work at BMRZ is supported by the state of Hesse. Work in Covid19-NMR was supported by the Goethe Corona Funds, by the IWB-EFRE-program 20007375 of state of Hesse, the DFG through CRC902: "Molecular Principles of RNA-based regulation." and through infrastructure funds (project numbers: 277478796, 277479031, 392682309, 452632086, 70653611) and by European Union's Horizon 2020 research and innovation program iNEXT-discovery under grant agreement No 871037. BY-COVID receives funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement number 101046203. "INSPIRED" (MIS 5002550) project, implemented under the Action "Reinforcement of the Research and Innovation Infrastructure," funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014–2020) and co-financed 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). The support of the CERM/CIRMMMP center of Instruct-ERIC is gratefully acknowledged. This work has been funded in part by a grant of the Italian Ministry of University and Research (FISR2020IP\_02112, ID-COVID) and by Fondazione CR Firenze. A.S. is supported by the Deutsche Forschungsgemeinschaft [SFB902/B16, SCHL2062/2-1] and the Johanna Quandt Young Academy at Goethe [2019/AS01]. M.H. and C.F. thank SFB902 and the Stiftung Polytechnische Gesellschaft for the Scholarship. L.L. work was supported by the French National Research Agency (ANR, NMR-SCoV2-ORF8), the Fondation de la Recherche Médicale (FRM, NMR-SCoV2-ORF8), FINOVI and the IR-RMN-THC Fr3050 CNRS. Work at UConn Health was supported by grants from the US National Institutes of Health (R01 GM135592 to B.H., P41 GM111135 and R01 GM123249 to J.C.H.) and the US National Science Foundation (DBI 2030601 to J.C.H.). Latvian Council of Science Grant No. VPP-COVID-2020/1-0014. National Science Foundation EAGER MCB-2031269. This work was supported by the grant Krebsliga KFS-4903-08-2019 and SNF-311030\_192646 to J.O. P.G. (ITMP) The EOSC Future project is co-funded by the European Union Horizon Programme call INFRAEOSC-03-2020—Grant Agreement Number 101017536. Open Access funding enabled and organized by Projekt DEAL.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are openly available in BMRBbig (bmrbig45 to bmrbig69) at <https://bmrbig.bmrb.io/released/>.<sup>[172–96]</sup>

**Keywords:** COVID19-NMR · Drug Discovery · Fragment Screening · NMR Spectroscopy · Protein · SARS-CoV-2

- [1] D. Adam, *Nature* **2022**, *601*, 312–315.
- [2] G. J. Kontoghiorghes, S. Fetta, C. N. Kontoghiorghes, *Front. Biosci.* **2021**, *26*, 1723–1736.
- [3] Z. A. Shyr, K. Gorshkov, C. Z. Chen, W. Zheng, *J. Pharmacol. Exp. Ther.* **2020**, *375*, 127–138.
- [4] T. T. Le, J. P. Cramer, R. Chen, S. Mayhew, *Nat. Rev. Drug Discovery* **2020**, *19*, 667–668.
- [5] M. Mei, X. Tan, *Front. Mol. Biosci.* **2021**, *8*, <https://doi.org/10.3389/fmolb.2021.671263>.
- [6] M. Cully, *Nat. Rev. Drug Discovery* **2022**, *21*, 3–5.
- [7] E. Petersen, F. Ntoumi, D. S. Hui, A. Abubakar, L. D. Kramer, et al., *Int. J. Infectious Diseases* **2022**, *114*, 268–272.
- [8] M. Hoffmann, N. Krüger, S. Schulz, A. Cossmann, C. Rocha, et al., *Cell* **2022**, *185*, 447–456.
- [9] L. A. VanBlargan, J. M. Errico, P. J. Halfmann, S. J. Zost, J. E. Crowe, et al., *Nat. Med.* **2022**, *28*, 490–495.
- [10] W. F. Garcia-Beltran, K. J. St Denis, A. Hoelzemer, E. C. Lam, A. D. Nitido, et al., *Cell* **2022**, *185*, 457–466.
- [11] Y.-W. Zhou, Y. Xie, L.-S. Tang, D. Pu, Y.-J. Zhu, et al., *Signal Transduction Targeted Ther.* **2021**, *6*, 317.
- [12] B. Malone, N. Urakova, E. J. Snijder, E. A. Campbell, *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 21–39.
- [13] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, et al., *Nature* **2020**, *579*, 265–269.
- [14] A. R. Fehr, S. Perlman, *Methods Mol. Biol.* **2015**, *1282*, 1–23.
- [15] E. J. Snijder, P. J. Bredenbeek, J. C. Dobbe, V. Thiel, J. Ziebuhr, et al., *J. Mol. Biol.* **2003**, *331*, 991–1004.
- [16] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, et al., *Nature* **2020**, *583*, 459–468.
- [17] C. W. Nelson, Z. Arden, T. L. Goldberg, C. Meng, C. H. Kuo, et al., *eLife* **2020**, *9*, e59633.
- [18] A. Pavesi, *Virology* **2020**, *546*, 51–66.
- [19] P. Venkatesan, *Lancet Respir. Med.* **2021**, *9*, e63.
- [20] B. Cao, Y. Wang, D. Wen, W. Liu, J. Wang, et al., *N. Engl. J. Med.* **2020**, *382*, 1787–1799.
- [21] M. Wang, R. Cao, L. Zhang, X. Yang, J. Liu, et al., *Cell Research* **2020**, *30*, 269–271.
- [22] W.-C. Ko, J.-M. Rolain, N.-Y. Lee, P.-L. Chen, C.-T. Huang, et al., *Int. J. Antimicrob. Agents* **2020**, *55*, 105933.
- [23] M. A. Martinez, *Front. Immunol.* **2022**, *12*, <https://doi.org/10.3389/fimmu.2021.635371>.
- [24] J. O. Ogidigo, E. A. Iwuchukwu, C. U. Ibeji, O. Okpalefe, M. E. S. Soliman, *J. Biomol. Struct. Dyn.* **2022**, *40*, 2284–2301.
- [25] C. Liu, X. Zhu, Y. Lu, X. Zhang, X. Jia, T. Yang, *J. Pharm. Anal.* **2021**, *11*, 272–277.
- [26] M. A. White, W. Lin, X. Cheng, *J. Phys. Chem. Lett.* **2020**, *11*, 9144–9151.
- [27] M. T. J. Quimque, K. I. R. Notarte, R. A. T. Fernandez, M. A. O. Mendoza, R. A. D. Liman, et al., *J. Biomol. Struct. Dyn.* **2021**, *39*, 4316–4333.
- [28] A. Carino, F. Moraca, B. Fiorillo, S. Marchianò, V. Sepe, et al., *Front. Chem.* **2020**, *8*, <https://doi.org/10.3389/fchem.2020.572885>.
- [29] S. Barage, A. Karthic, R. Bavi, N. Desai, R. Kumar, et al., *J. Biomol. Struct. Dyn.* **2022**, *40*, 2557–2574.
- [30] M. Macchiagodena, M. Pagliai, P. Procacci, *Chem. Phys. Lett.* **2020**, *750*, 137489.
- [31] S. Yazdani, N. de Maio, Y. Ding, V. Shahani, N. Goldman, M. Schapira, *J. Proteome Res.* **2021**, *20*, 4212–4215.
- [32] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, et al., *Science* **2020**, *368*, 409–412.
- [33] M. Schuller, G. J. Correy, S. Gahbauer, D. Fearon, T. Wu, et al., *Sci. Adv.* **2021**, *7*, eabf8711.
- [34] S. Günther, P. Y. A. Reinke, Y. Fernández-García, J. Lieske, T. J. Lane, et al., *Science* **2021**, *372*, 642–646.
- [35] J. A. Newman, A. Douangamath, S. Yazdani, Y. Yosaatmadja, A. Aimon, et al., *Nat. Commun.* **2021**, *12*, 4848.
- [36] F. Cantrelle, E. Boll, L. Brier, D. Moschidi, S. Belouzard, et al., *Angew. Chem. Int. Ed.* **2021**, *60*, 25428–25435; *Angew. Chem.* **2021**, *133*, 25632–25639.
- [37] A. L. Kantsadi, E. Cattermole, M.-T. Matsoukas, G. A. Spyroulias, I. Vakonakis, *J. Biomol. NMR* **2021**, *75*, 167–178.
- [38] N. Imprachim, Y. Yosaatmadja, J. A. Newman, *bioRxiv* **2022**, <https://doi.org/10.1101/2022.03.11.483836>.
- [39] V. Napolitano, A. Dabrowska, K. Schorpp, A. Mourão, E. Barreto-Duran, et al., *Cell Chem. Biol.* **2022**, *29*, 774–784.
- [40] T. C. M. Consortium, H. Achdout, A. Aimon, E. Bar-David, H. Barr, et al., *bioRxiv* **2022**, <https://doi.org/10.1101/2020.10.29.339317>.
- [41] M. Kozlov, *Nature* **2022**, *601*, 496.
- [42] J. F. X. Diffley, *Biochem. J.* **2021**, *478*, 2533–2535.
- [43] C. T. Lim, K. W. Tan, M. Wu, R. Ulferts, L. A. Armstrong, et al., *Biochem. J.* **2021**, *478*, 2517–2531.
- [44] J. C. Milligan, T. U. Zeisner, G. Papageorgiou, D. Joshi, C. Soudy, et al., *Biochem. J.* **2021**, *478*, 2499–2515.
- [45] S. Basu, T. Mak, R. Ulferts, M. Wu, T. Deegan, et al., *Biochem. J.* **2021**, *478*, 2481–2497.
- [46] B. Canal, R. Fujisawa, A. W. McClure, T. D. Deegan, M. Wu, et al., *Biochem. J.* **2021**, *478*, 2465–2479.
- [47] B. Canal, A. W. McClure, J. F. Curran, M. Wu, R. Ulferts, et al., *Biochem. J.* **2021**, *478*, 2445–2464.
- [48] A. P. Bertolin, F. Weissmann, J. Zeng, V. Posse, J. C. Milligan, et al., *Biochem. J.* **2021**, *478*, 2425–2443.
- [49] J. Zeng, F. Weissmann, A. P. Bertolin, V. Posse, B. Canal, et al., *Biochem. J.* **2021**, *478*, 2405–2423.
- [50] P. Ren, W. Shang, W. Yin, H. Ge, L. Wang, et al., *Acta Pharm. Sin.* **2022**, *43*, 483–493.
- [51] N. Altincekic, S. M. Korn, N. S. Qureshi, M. Dujardin, M. Ninot-Pedrosa, et al., *Front. Mol. Biosci.* **2021**, *8*, 89.
- [52] S. Sreeramulu, C. Richter, H. Berg, M. A. Wirtz Martin, B. Ceylan, et al., *Angew. Chem. Int. Ed.* **2021**, *60*, 19191–19200; *Angew. Chem.* **2021**, *133*, 19340–19349.
- [53] D. Kozakov, L. E. Grove, D. R. Hall, T. Bohnuud, S. E. Mottarella, et al., *Nature Protocols* **2015**, *10*, 733–755.
- [54] A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacic, et al., *Nat. Commun.* **2020**, *11*, 5047.
- [55] D. Shin, R. Mukherjee, D. Grewe, D. Bojkova, K. Baek, et al., *Nature* **2020**, *587*, 657–662.
- [56] K. Anand, J. Ziebuhr, P. Wadhvani, J. R. Mesters, R. Hilgenfeld, *Science* **2003**, *300*, 1763–1767.
- [57] N. T. Nashed, A. Aniana, R. Ghirlando, S. C. Chiliveri, J. M. Louis, *Commun. Biol.* **2022**, *5*, 160.
- [58] B. Goyal, D. Goyal, *ACS Comb. Sci.* **2020**, *22*, 297–305.
- [59] O. B. Cox, T. Krojer, P. Collins, O. Monteiro, R. Talon, A. Bradley, O. Fedorov, J. Amin, B. D. Marsden, J. Spencer, F. von Delft, P. E. Brennan, *Chem. Sci.* **2016**, *7*, 2322–2330.

- [60] S. Sreeramulu, C. Richter, T. Kuehn, K. Azzaoui, M. J. J. Blommers, et al., *J. Biomol. NMR* **2020**, *74*, 555–563.
- [61] H. Berg, M. A. Wirtz Martin, A. Niesteruk, C. Richter, S. Sreeramulu, H. Schwalbe, *J. Visualized Exp.* **2021**, *172*, e62262.
- [62] A. J. Robertson, J. Ying, A. Bax, *Magn. Reson.* **2021**, *2*, 129–138.
- [63] R. A. Laskowski, J. Jabłońska, L. Pravda, R. S. Vařeková, J. M. Thornton, *Protein Sci.* **2018**, *27*, 129–134.
- [64] A. Grosdidier, V. Zoete, O. Michielin, *J. Comput. Chem.* **2011**, *32*, 2149–2159.
- [65] A. Grosdidier, V. Zoete, O. Michielin, *Nucleic Acids Res.* **2011**, *39*, W270–W277.
- [66] N. Kubatova, N. S. Qureshi, N. Altincekic, R. Abele, J. K. Bains, et al., *Biomol. NMR Assign.* **2021**, *15*, 65–71.
- [67] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, et al., *Nucleic Acids Res.* **2019**, *47*, D930–D940.
- [68] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, et al., *Nucleic Acids Res.* **2021**, *49*, D1388–D1395.
- [69] X. Chen, C. H. Reynolds, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407–1414.
- [70] M. Kuzikov, E. Costanzi, J. Reinshagen, F. Esposito, L. Vangeel, et al., *ACS Pharmacol. Transl. Sci.* **2021**, *4*, 1096–1110.
- [71] S. Jeon, M. Ko, J. Lee, I. Choi, S. Y. Byun, S. Park, D. Shum, S. Kim, *Antimicrob. Agents Chemother.* **2020**, *64*, <https://doi.org/10.1128/AAC.00819-20>.
- [72] nsp1, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp1, BMRbig, bmrbig48, <https://bmrbig.bmr.io/released/bmrbig48>.
- [73] \_GHM\_nsp5, H. Berg, 2022, DSI-PL\_COVID19-NMR\_GHM\_nsp5, BMRbig, bmrbig45, <https://bmrbig.bmr.io/released/bmrbig45>.
- [74] nsp3e, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp3e, BMRbig, bmrbig56, <https://bmrbig.bmr.io/released/bmrbig56>.
- [75] nsp3a, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp3a, BMRbig, bmrbig50, <https://bmrbig.bmr.io/released/bmrbig50>.
- [76] nsp9, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp9, BMRbig, bmrbig61, <https://bmrbig.bmr.io/released/bmrbig61>.
- [77] nsp7, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp7, BMRbig, bmrbig59, <https://bmrbig.bmr.io/released/bmrbig59>.
- [78] nsp8, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp8, BMRbig, bmrbig60, <https://bmrbig.bmr.io/released/bmrbig60>.
- [79] nsp3c (SUD-N), H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp3c\_SUD-N, BMRbig, bmrbig54, <https://bmrbig.bmr.io/released/bmrbig54>.
- [80] nsp3c (SUD-MC), H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp3c\_SUD-MC, BMRbig, bmrbig53, <https://bmrbig.bmr.io/released/bmrbig53>.
- [81] ORF9a (NTD), H. Berg, 2022, DSI-PL\_COVID19-NMR\_ORF9a\_NTD, BMRbig, bmrbig67, <https://bmrbig.bmr.io/released/bmrbig67>.
- [82] ORF9a (IDR1-NTD-IDR2), H. Berg, 2022, DSI-PL\_COVID19-NMR\_ORF9a\_IDR1-NTD-IDR2, BMRbig, bmrbig66, <https://bmrbig.bmr.io/released/bmrbig66>.
- [83] nsp3b, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp3b, BMRbig, bmrbig51, <https://bmrbig.bmr.io/released/bmrbig51>.
- [84] nsp2 (CtDR), H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp2\_CtDR, BMRbig, bmrbig49, <https://bmrbig.bmr.io/released/bmrbig49>.
- [85] nsp3b-GS-441524, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp3b-GS-441524, BMRbig, bmrbig52, <https://bmrbig.bmr.io/released/bmrbig52>.
- [86] nsp3d, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp3d, BMRbig, bmrbig55, <https://bmrbig.bmr.io/released/bmrbig55>.
- [87] \_His6\_nsp15, H. Berg, 2022, DSI-PL\_COVID19-NMR\_His6\_nsp15, BMRbig, bmrbig46, <https://bmrbig.bmr.io/released/bmrbig46>.
- [88] \_GS\_nsp5, H. Berg, 2022, DSI-PL\_COVID19-NMR\_GS\_nsp5, BMRbig, bmrbig47, <https://bmrbig.bmr.io/released/bmrbig47>.
- [89] ORF9a (CTD), H. Berg, 2022, DSI-PL\_COVID19-NMR\_ORF9a\_CTD, BMRbig, bmrbig65, <https://bmrbig.bmr.io/released/bmrbig65>.
- [90] nsp10, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp10, BMRbig, bmrbig62, <https://bmrbig.bmr.io/released/bmrbig62>.
- [91] ORF9a (NTD-SR), H. Berg, 2022, DSI-PL\_COVID19-NMR\_ORF9a\_NTD-SR, BMRbig, bmrbig68, <https://bmrbig.bmr.io/released/bmrbig68>.
- [92] nsp10\_nsp16, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp10\_nsp16, BMRbig, bmrbig63, <https://bmrbig.bmr.io/released/bmrbig63>.
- [93] ORF9b, H. Berg, 2022, DSI-PL\_COVID19-NMR\_ORF9b, BMRbig, bmrbig69, <https://bmrbig.bmr.io/released/bmrbig69>.
- [94] nsp3y, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp3y, BMRbig, bmrbig57, <https://bmrbig.bmr.io/released/bmrbig57>.
- [95] nsp5, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp5, BMRbig, bmrbig58, <https://bmrbig.bmr.io/released/bmrbig58>.
- [96] nsp10\_nsp14, H. Berg, 2022, DSI-PL\_COVID19-NMR\_nsp10\_nsp14, BMRbig, bmrbig64, <https://bmrbig.bmr.io/released/bmrbig64>.

Manuscript received: April 21, 2022

Accepted manuscript online: September 15, 2022

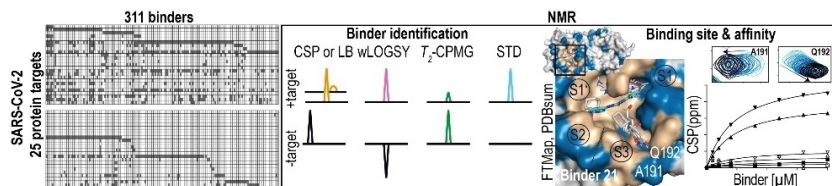
Version of record online: ■■■, ■■■

## Research Articles

## SARS-CoV-2

H. Berg, M. A. Wirtz Martin, N. Altincekic, I. Alshamleh, J. Kaur Bains, J. Blechar, B. Ceylan, V. de Jesus, K. Dhamotharan, C. Fuks, S. L. Gande, B. Hargittay, K. F. Hohmann, M. T. Hutchison, S. Marianne Korn, R. Krishnathas, F. Kutz, V. Linhard, T. Matzel, N. Meiser, A. Niesteruk, D. J. Pyper, L. Schulte, S. Trucks, K. Azzaoui, M. J. J. Blommers, Y. Gadiya, R. Karki, A. Zaliani, P. Gribbon, M. da Silva Almeida, C. Dinis Anoborn, A. L. Bula, M. Bütikofer, Í. Putinhon Caruso, I. Caterina Felli, A. T. Da Poian, G. Cardoso de Amorim, N. K. Fourkiotis, A. Gallo, D. Ghosh, F. Gomes-Neto, O. Gorbatyuk, B. Hao, V. Kurauskas, L. Lecoq, Y. Li, N. Cunha Mebus-Antunes, M. Mompeán, T. Cristtina Neves-Martins, M. Ninot-Pedrosa, A. S. Pinheiro, L. Pontoriero, Y. Pustovalova, R. Riek, A. J. Robertson, M. Jose Abi Saad, M. Á. Treviño, A. C. Tsika, F. C. L. Almeida, A. Bax, K. Henzler-Wildman, J. C. Hoch, K. Jaudzems, D. V. Laurents, J. Orts, R. Pierattelli, G. A. Spyroulias, E. Duchardt-Ferner, J. Ferner, B. Fürtig, M. Hengesbach, F. Löhr, N. Qureshi, C. Richter, K. Saxena, A. Schlundt, S. Sreeramulu, A. Wacker, J. E. Weigand, J. Wirmer-Bartoschek, J. Wöhnert, H. Schwalbe\* — **e202205858**

Comprehensive Fragment Screening of the SARS-CoV-2 Proteome Explores Novel Chemical Space for Drug Development



Using a fragment-based screening strategy by NMR, we identified 311 small molecule binders of 25 SARS-CoV-2 proteins, thus expanding the previously unexplored chemical and target space.

Further, using experimental and bioinformatic analysis we identify potential binding sites. This comprehensive data would greatly assist medicinal chemistry efforts even beyond COVID-19.