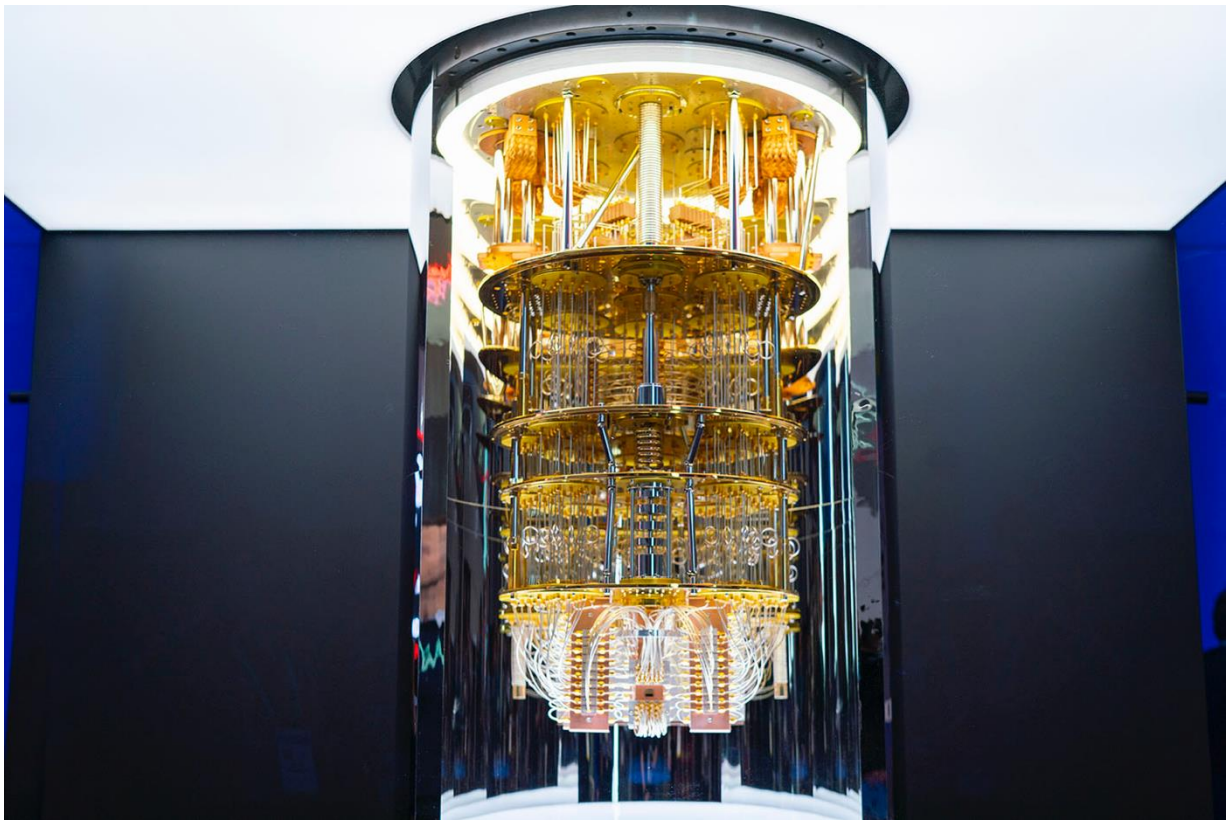
Italian National Agency for New Technologies,
Energy and Sustainable Economic Development

High Performance Computing on CRESCO infrastructure: research activities and results 2020



December 2021

High Performance Computing on CRESCO Infrastructure:
research activity and results 2020

Contributions provided by a selection of users of the CRESCO infrastructure
Scientific Editors: Davide De Chiara and Simone Giusepponi, ENEA, TERIN-ICT-HPC
Cover: Amedeo Trolese, ENEA, TERIN-ICT, CR Frascati
ISBN: 978-88-8286-429-3

Contents

FOREWORD

During the year 2020, the CRESCO high performance computing clusters have provided 108 million hours of "core" computing time, at a high availability rate, to about 150 users, supporting ENEA research and development activities in many relevant scientific and technological domains. In the framework of joint programs with ENEA researchers and technologists, computational services have been provided also to academic and industrial communities.

This report, the twelfth of a series started in 2008, is a collection of 51 papers illustrating the main results obtained during 2020 using the CRESCO/ENEAGRID HPC facilities. The significant number of contributions proves the importance of the HPC facilities in ENEA for the research community. The topics cover various fields of research, such as materials science, efficient combustion, climate research, nuclear technology, plasma physics, biotechnology, aerospace, complex systems physics, geophysical flow, renewable energies, environmental issues, HPC technology. In particular, cause to the pandemic situation related to COVID-19, in 2020 part of the computational resources were devoted to work related to this considerable issue. These researchers employed about 15 % of the overall computing time and yielded 8 works. The report shows the wide spectrum of applications of high performance computing, which has become an all-round enabling technology for science and engineering.

Since 2008, the main ENEA computational resources are located near Naples, in Portici Research Centre. This is a result of the CRESCO Project (Computational Centre for Research on Complex Systems), co-funded, in the framework of the 2001-2006 PON (European Regional Development Funds Program), by the Italian Ministry of Education, University and Research (MIUR).
The CRESCO Project provided the financial resources to set up the first HPC x86_64 Linux cluster in ENEA; a major computing installation for both the Italian and the International context: it ranked 126 in the HPC Top 500 June 2008 world list, with 17.1 Tflops and 2504 cpu cores. It was later decided to keep CRESCO as the name for all the Linux clusters in the ENEAGRID infrastructure, which integrates all ENEA scientific computing systems, and is currently distributed in six Italian sites.
CRESCO computing resources were later upgraded in the framework of PON 2007-2013 with the project TEDAT and the cluster CRESCO4, 100 Tflops computing power. In 2020 the ENEAGRID computational resources consist of ~25000 computing cores and a raw data storage of about 5 PB.
In 2015 ENEA and CINECA, the main HPC institution in Italy, signed a collaboration agreement to promote joint activities and projects in HPC. In this framework, CINECA and ENEA participated successfully to a selection launched by EUROfusion, the European Consortium for the Development of Fusion Energy, for the procurement of a several PFlops HPC system, beating the competition of 7 other institutions. The new system MARCONI-FUSION started operation in July 2016 at 1 Pflops computation power level which has been increased to 5 Pflops in the summer of 2017 and thanks to a new awarded tender, the HPC services of MARCONI Fusion have been extend until 2023 with a power peak of 8 PFlops of conventional processors Intel Skylake and 2 PFlops of accelerated GPU partition of CINECA Marconi 100.
The ENEA-CINECA agreement is the key basis for ENEA HPC developments. The CRESCO6 cluster has been installed in 2018 and its own 1.4 PFlops peak computing power, ranked 420th in November 2018 Top500 list. CRESCO6 has been a challenge in HPC co-design system thanks to implement a multi-fabric network able for working Infiniband and Omni-Path on a single GPFS cluster using the same storage systems of CRESCO data centre.

The success and the quality of the results produced by CRESCO stress the role that HPC facilities can play in supporting science and technology for all ENEA activities, national and international collaborations, and the ongoing renewal of the infrastructure provides the basis for an upkeep of this role in the forthcoming years.

Dipartimento Tecnologie Energetiche e Fonti Rinnovabili,
Divisione per lo Sviluppo di Sistemi per l'Informatica e l'ICT - CRESCO Team

BLIND PREDICTIONS FOR THE SAMPL9 HOST-GUEST CHALLENGE

Piero Procacci*

University of Florence, Chemistry Department, Via Lastruccia n. 3, 50019, Sesto Fiorentino, Italy

ABSTRACT. We present our blind predictions for the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL), 9th challenge, focusing on binding of WP6 (carboxy-pillar[6]arene) with ammonium/diammonium cationic guests. Host-guest binding free energies have been calculated using the recently developed virtual double system single box approach, based on the enhanced sampling of the bound and unbound end-states followed by fast switching nonequilibrium alchemical simulations [M Macchiagodena, M Pagliai, M Karrenbrock, G Guarnieri, F Iannone, Procacci, J Chem Theory Comput, 16, 7260, 2020]. Experimental binding free energies for the thirteen host-guest complexes will be disclosed by November 2021.

1 Introduction

The Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) is a set of community-wide blind challenges aimed at advancing the computational techniques in drug design [1-3]. New experimental data, such as dissociation free energies, hydration free energies, acid-base dissociation constants or partition coefficients are withheld from participants until the prediction submission deadline, so that the true predictive power of methods can be assessed.

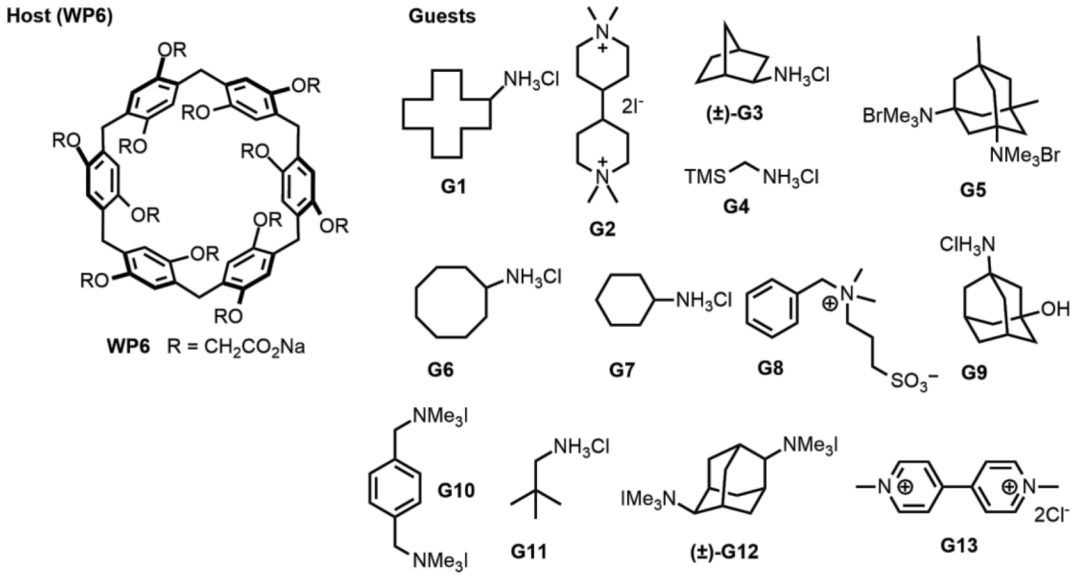


Fig. 1. Host-guest SAMPL9 challenge

In the latest 9th challenge, participants are required to predict the binding affinities of thirteen ammonium/diammonium cationic ligands (guests) *vs* WP6, a water-soluble toroidal macrocyclic host molecule (see Figure 1). Host-guest experimental binding affinities, measured by Isothermal Titration Calorimetry (ICT) at pH=7.4 will be disclosed by early November 2021.

*Corresponding author. E-mail: procacci@unifi.it

The host WP6, yet to be featured in previous SAMPL challenges, is structurally similar to cucurbit[n]urils (CBn) used in SAMPL6-SAMPL8[2,3] with important differences. Unlike the neutral CBn host, the quasi-D_{6h} WP6 bears six carboxylated moieties on the upper and lower rims that can be in part protonated at pH=7.4[4,5]. As the pKa's of WP6 are not available experimentally, such a feature represents an important challenge in the blind prediction of WP6-guest binding affinities. Participants will in fact need to deal with this complication.
In this report, we present the dissociation free energies of the fully anionic $WP6^{-12}$ for the thirteen cationic guests of Figure 1 plus five additional guest molecules with known dissociation constants, namely methylene blue (G14)[6], GNF-Pf-3194 (G15)[7], M2 (G16)[7], choline (G17)[8] and betaine (G18)[8]. The calculations have been performed on the CRESCO6-ENEA cluster[9] in Portici (Italy) using the so-called virtual Double System Single Box (vDSSB) method[10], based on a production of a swarm of concurrent nonequilibrium (NE) alchemical simulations that are started from end-state canonical configurations sampled using Hamiltonian Replica Exchange (HREM).

2 Methods

As stated previously, the twelve pKa's of WP6 host are not known. Strictly speaking, in presence of multiple host-guest complexes with various protonation states, the overall observed association constant for Gn-WP6 complexes is given by

$$K_a = \sum \; W_k K_a^{(k)} \qquad (1)$$

where $K_a^{(k)}$is the association constant for the k-th protonated WP6 species and $W_k$is the corresponding normalized weight. The Gn-WP6 dissociation free energyΔG_d should be hence calculated as

$$\Delta G_d = RTln\left(\sum \; W_k e^{\beta \Delta G_a^{(k)}}\right) \qquad (2)$$

where$\Delta G$is the dissociation free energy with the host in the k-th protonation state. The predicted pKa of the WP6 template monobasic acid 2-(2,5-dimethylphenoxy) acetic acid is 3.23 [11]. Assuming a pKa distribution of the twelve equivalent protonation states modulated by the so-called statistical factor [12], we obtain that the prevalent species at pH=7.4 is the 12-anion with all deprotonated carboxylated groups ($W = 0.98$). We hence computed the dissociation constant for the $WP6^{12-}$ species only.

The vDSSB methodology is thoroughly described in Ref. [10]. In brief, the method consists of two massively parallel computational steps, the HREM stage and the nonequilibrium alchemical stage. The *HREM stage* relies on the enhanced sampling of the host-guest bound state in explicit water and of the isolated guest molecule. The initial configurations for the unbound state are obtained by combining the HREM-sampled (decoupled) ligand gas-phase snapshots with a pre-equilibrated box filled with explicit water. H-REM uses n=16 and n=8 replicas for the bound and unbound states respectively, with a maximum scaling factor of S=0.1 (corresponding to a temperature of 3000 K) involving only the intra-solute interactions. The scaling factors along the replica progression are computed according to the protocol$s_m = S^{\frac{(m-1)}{n}}$with *m=1..n*. Only scaling factors are exchanged among neighboring replicas to minimize the communication overhead on the MPI layer. HREM simulations are run for 48 ns and 16 ns for the bound and gas-phase state, respectively.
In the *NE alchemical stage,* the bound state leg of the alchemical cycle is performed by rapidly decoupling the bound ligand in a swarm of 720 independent NE simulations, each lasting for 0.72 ns. The unbound leg of the cycle is done by growing (recoupling) the ghost ligand in the solvent in 480 NE alchemical simulations lasting 0.36 ns. For further details on the ligand coupling/decoupling protocols, we refer to Ref. [10]. The number of concurrent alchemical simulations in a parallel job corresponds to the number of requested MPI processes.
The final bound and unbound resulting alchemical work distribution are combined in the convolution

$$P(W) = P_b * P_u(W) = \int \; dw P_b(W) P_u(W - w) \qquad (3)$$

The dissociation free estimate can be performed by way of the Jarzynski identity[13]

$$\Delta G_{vDSSB} = -RTln\left(\int \; P(W)e^{-\beta W}dW\right) \qquad (4)$$

or using the Crooks theorem in the assumption that the convolution $P(W)$can be described by a mixture of three normal distribution [10,14], i.e

$$\Delta G_{vDSSB} = -RTln\left(\sum \; c_i e^{-\beta\left(\mu_i - \frac{\beta\sigma_i^2}{2}\right)}\right) \qquad (5)$$

where $c_i, \mu_i, \sigma_i^2$are the normalized weight, mean and variance of the i-th normal component, determined via the expectation-maximization algorithm [15]. Estimates based on eq. (4) or eq. (5) are used depending on the width and character of the work distribution as assessed by the Anderson-Darling normality test[16].
The free energy estimates are corrected for a volume and a charge term (for non-neutral ligand) given by

$$\Delta G_{Vol} = RTln\left(4\pi\frac{(2\sigma)^3}{(3V_0)}\right) \qquad (6)$$

$$\Delta G_q = \frac{\pi}{\alpha^2}\left[\frac{q_G^2+2q_Gq_H}{V_b} - \frac{q_G^2}{V_u}\right] \qquad (7)$$

where $\sigma$is the standard deviation of the host-guest COM-COM distance distribution in the bound state, $V_0$is the standard state volume, $q_G$and $q_H$are the net charges of the guest and the host, and $V_b, V_u$are the mean volume of the MD box for the bound and unbound state respectively, and α is the Ewald convergence parameter. The final blind prediction for the standard dissociation free energy is given by

$$\Delta G_d = \Delta G_{vDSSB} + \Delta G_{vol} + \Delta G_q \qquad (8)$$

The Force Field (FF) parameters and topology of the host and guests molecules were prepared using the PrimaDORAC interface [17] based on the GAFF2[18] parameter set. The initial bound state was prepared using the Autodock Vina code[19]. The bound complexes and the ghost ligands were solvated in about 1600 OPC3[20] and 512 water molecules, respectively. A background neutralizing plasma was assumed within the PBC Ewald method (PME[21]). All simulations, HREM or nonequilibrium, were performed in the NPT ensemble in standard conditions using an isotropic Parrinello-Rahman Langrangian[22] and a series of Nose' thermostats[23] for pressure and temperature control, respectively. All simulations have been performed using the hybrid OpenMP-MPI program ORAC[24] on the CRESCO6 cluster[9].

3 Results and Discussion

In Table 1 we show the results obtained for the dissociation free energy for the eighteen host-guest pairs. The correlation coefficient and the Kendall rank coefficient for the guests with known dissociation free energy are R=0.83 and τ=0.71, respectively with a mean *signed* error of -3.3 kcal/mol, indicating a systematic overestimate of the dissociation free energies.

Table 1: Blind predictions for the host-guest SAMPL9 set dissociation free energies. Free energies are given in kcal/mol.$\sigma_b$and $\sigma_u$are the standard deviation in kcal/mol of bound and unbound work distribution. "Type" refers to the estimate type (JR, eq (6); GM*n* eq (7) with *n* normal components). Experimental dissociation free energies, ΔG_{exp}, have been taken from the Refs. in the square bracket.

Host-guest	ΔG_d	$\Delta G_{vol} + \Delta G_q$	σ_b	σ_u	Type	$\Delta G_{exp.}$
G1-WP6	7.7∓1.9	-4.6	3.9	1.9	JR	-
G2-WP6	17.0∓0.4	-7.2	2.8	1.1	JR	-
G3-WP6	11.5∓0.6	-6.5	3.1	0.8	JR	-
G4-WP6	7.0∓0.9	-6.2	2.6	0.8	GM1	-
G5-WP6	9.5∓0.5	-8.8	3.0	1.3	JR	-
G6-WP6	14.7∓1.3	-6.3	3.1	0.9	GM1	-
G7-WP6	12.2∓1.7	-7.0	3.2	0.8	GM3	-
G8-WP6	12.2∓1.6	-3.3	3.9	1.6	GM3	-
G9-WP6	12.5∓0.6	-5.6	3.3	1.0	JR	-
G10-WP6	13.5∓1.1	-6.9	2.8	1.1	GM1	-
G11-WP6	8.7∓0.6	-6.2	2.6	0.8	GM3	-
G12-WP6	18.6∓1.2	-7.7	2.9	1.1	GM3	-
G13-WP6	8.3∓0.8	-7.3	3.0	0.9	GM3	8.5 [5]
G14-WP6	17.2∓0.4	-4.8	2.5	1.1	GM3	9.7 [6]
G15-WP6	11.2∓1.6	-4.8	2.8	0.9	GM3	8.4 [7]
G16-WP6	14.5∓1.5	-4.7	2.6	1.1	GM3	10.6 [7]
G17-WP6	7.9∓0.8	-6.3	2.1	0.7	GM3	6.5 [8]
G18-WP6	4.1∓1.3	-3.7	2.2	0.9	GM3	< 1 [8]

Such discrepancies might be due to systematic force field deficiencies in the ammonium/diammonium guests or in the host, or to missing contributions to the dissociation free energies due to species with protonation state differing from the fully anionic WP6. Predicted WP6-Gn dissociation free energies using vDSSB are quite high and comparable to those observed for CB8 in past SAMPL challenges with similar cationic guests. Nonetheless, due to the smaller cavity of the host, WP6 affinities are expected to be sensibly smaller than that of the structurally related (but larger) CB8 host with similar cationic guests. We may thus infer that our blind prediction of the standard Gn-WP6 dissociation free energies can be affected by a systematic *positive* bias, due to the presence of partially protonated WP6 species at pH=7.4. SAMPL9 organizers are working to obtain pKa values of the WP6 carboxy groups via pH-metric titration so it is possible this information may be available at some point, potentially before the challenge close. In that case, we may try to update our blind prediction by computing the dissociation free energies of the existing WP6 protonated species at the experimental pH using eq (2).

Acknowledgments

The computing resources and the related technical support used for this work have been provided by CRESCO/ENEAGRID High-Performance Computing infrastructure and its staff[9]. CRESCO/ENEAGRID High-Performance Computing infrastructure is funded by ENEA, the Italian National Agency for New Technologies, Energy and Sustainable Economic Development and by Italian and European research programmes (see www.cresco.enea.it for information).

References

[1] Andrea Rizzi, Steven Murkli, John N. McNeill, Wei Yao, Matthew Sullivan, Michael K. Gilson, Michael W. Chiu, Lyle Isaacs, Bruce C. Gibb, David L. Mobley, and John D. Chodera. Overview of the sampl6 host–guest binding affinity prediction challenge. *J. Comput. Aided Mol. Des.*, **32** pp. 937-963, (2018).

[2] M. Amezcua, L. El Khoury, and David L. Mobley. Sampl7 host-guest challenge overview: assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. *J. Comput. Aided Mol. Des*., **35**, pp. 1-35 ,(2021)

[3] SAMPL8 analysis for CB8 and GDCC host. https://github.com/samplchallenges/SAMPL8, DOI: 10.5281/zenodo.4029560

[4] Guocan Yu 1, Xiangyan Zhou, Zibin Zhang, Chengyou Han, Zhengwei Mao, Changyou Gao, Feihe Huang, Pillar[6]arene/paraquat molecular recognition in water: high binding strength, pH-responsiveness, and application in controllable self-assembly, controlled release, and treatment of paraquat poisoning, *J. Am. Chem Soc.*, **134**, pp. 19489-97, (2012)

[5] Nicolas H, Yuan B, Xu J, Zhang X, Schönhoff M. pH-Responsive Host-Guest Complexation in Pillar[6]arene-Containing Polyelectrolyte Multilayer Films. *Polymers (Basel),* **9**, 719, (2017)

[6] Yang K , Wen J , Chao S , Liu J , Yang K , Pei Y , Pei Z . A supramolecular photosensitizer system based on the host-guest complexation between water-soluble pillar[6]arene and methylene blue for durable photodynamic therapy. *Chem. Commun.,* **54**, pp. 5911-5914, (2018)

[7] Hessz D, Bádogos S, Bojtár M, Bitter I, Drahos L, Kubinyi M. Complexes of carboxylato pillar[6]arene with Brooker-type merocyanines: Spectral properties, pKa shifts and the design of a displacement assay for trimethyl lysine. *Spectrochim Acta A Mol Biomol Spectrosc.,* **252**:119455, (2021)

[8] Hua, B.; Zhang, Z.H.; Sun, J.F.; Yang, J. Pillar[6]arene/acridine orange host–guest complexes as colorimetric and fluorescence sensors for choline compounds and further application in monitoring enzymatic reactions. *Sens. Actuators B*, **255**, pp. 1430–1435, (2018)

[9] Iannone F.; Ambrosino F.; Bracco G.; De Rosa M.; Funel A.; Guarnieri G.; Migliori S.; Palombi F.; Ponti G.; Santomauro G.; Procacci P. CRESCO ENEA HPC clusters: a working example of a multifabric GPFS spectrum scale layout. *International Conference on High Performance Computing Simulation (HPCS)* 2019, pp 1051–1052.

[10] Macchiagodena M, Pagliai M, Karrenbrock M, Guarnieri G, Iannone F, Procacci P. Virtual Double-System Single-Box: A Nonequilibrium Alchemical Technique for Absolute Binding Free Energy Calculations: Application to Ligands of the SARS-CoV-2 Main Protease. *J Chem Theory Comput.,* **16**, pp. 7160-7172. (2020)

[11] SciFinder; Chemical Abstracts Service: Columbus, OH; carbon-13 NMR spectrum; spectrum ID CC-03-C_SPC-3734; RN 50-52-2; https://scifinder.cas.org (accessed June 9, 2010).

[12]D. D. Perrin, Boyd Dempsey and E. P. Serjeant. Molecular Factors that Modify pKa Values. In pKa Prediction for Organic Acids and Bases. Netherlands: Springer (1981)

[13] Jarzynski, C., Nonequilibrium equality for free energy differences, *Phys. Rev. Lett.*, **78** , pp. 2690, (1997)

[14] Procacci, P., Unbiased free energy estimates in fast nonequilibrium transformations using Gaussian mixtures, *J. Chem. Phys.* **142**, pp. 154117, (2015)

[15] Dempster, A.P.; Laird, N.M.; Rubin, D.B., Maximum Likelihood from Incomplete Data via the EM Algorithm, *J. Royal Stat. Soc. B.,* **39**, pp. 1-38, (1977)

[16] Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons. J. Am. Stat. Ass., **69**, pp. 730-737, (1974)

[17] Procacci, P., PrimaDORAC: A Free Web Interface for the Assignment of Partial Charges, Chemical Topology, and Bonded Parameters in Organic or Drug Molecules, *J. Chem. Inf. Model.*, **57**, pp. 1240-1245, (2017)

[18] Wang, J., Wang, W., Kollman P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations, *J. Mol. Graph. Model.* , **25**, pp. 247260, (2006)

[19] Trott, O., Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**, pp. 455-461, (2010)

[20] Izadi, Saeed; Onufriev, Alexey V., Accuracy limit of rigid 3-point water models *J. Chem. Phys.*, **145**, 074501, (2016)

[21] Essmann, Ulrich, Perera, Lalith, Berkowitz, Max L., Darden, Tom, Lee, Hsing, Pedersen, Lee G., A smooth particle mesh Ewald method, *J. Chem. Phys.,* **103**, pp.8577-8593, (1995)
[22] M Parrinello, A Rahman., *J. Appl. Phys.* **52**, pp. 7182-7190, (1981)
[23] S. Nose, A unified formulation of the constant temperature molecular dynamics methods *J. Chem. Phys.,* **81**, pp. 511, (1984).
[24] Procacci, P., Hybrid MPI/OpenMP Implementation of the ORAC Molecular Dynamics Program for Generalized Ensemble and Fast Switching Alchemical Simulations, *J. Chem. Inf. Model*.,**56**, pp. 1117-1121, (2016)