

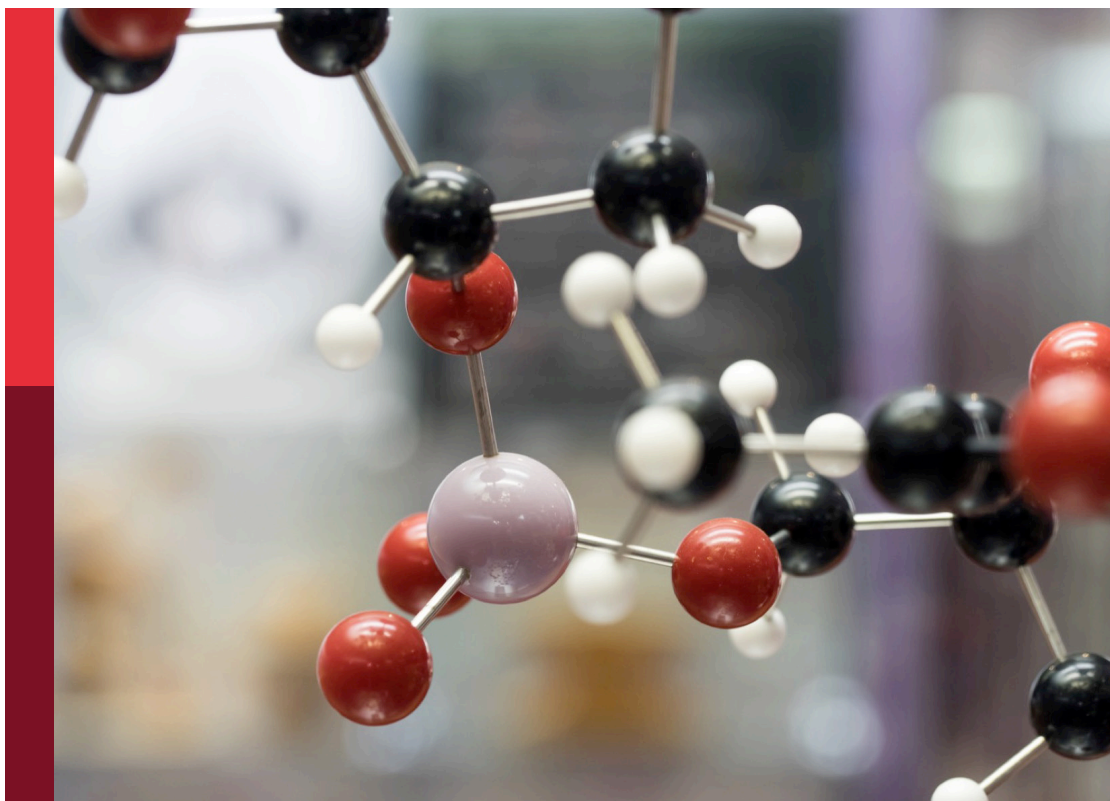
Recent advances in computational modelling of biomolecular complexes

Edited by

Zhongjie Liang, Adolfo Poma, Kurt Kremer, Kei-ichi Okazaki,
Sergio Pantano and Simón Poblete

Published in

Frontiers in Chemistry



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-2254-7
DOI 10.3389/978-2-8325-2254-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Recent advances in computational modelling of biomolecular complexes

Topic editors

Zhongjie Liang — Soochow University, China

Adolfo Poma — Institute of Fundamental Technological Research, Polish Academy of Sciences, Poland

Kurt Kremer — Max Planck Society, Germany

Kei-ichi Okazaki — Institute for Molecular Science (NINS), Japan

Sergio Pantano — Pasteur Institute of Montevideo, Uruguay

Simón Poblete — Universidad Austral de Chile, Chile

Citation

Liang, Z., Poma, A., Kremer, K., Okazaki, K.-i., Pantano, S., Poblete, S., eds. (2023). *Recent advances in computational modelling of biomolecular complexes*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2254-7

Table of contents

- 04 **Editorial: Recent advances in computational modelling of biomolecular complexes**
Simón Poblete, Sergio Pantano, Kei-ichi Okazaki, Zhongjie Liang, Kurt Kremer and Adolfo B. Poma
- 07 **Study of tyramine-binding mechanism and insecticidal activity of oil extracted from *Eucalyptus* against *Sitophilus oryzae***
Farshid Zargari, Zahra Nikfarjam, Ebrahim Nakhaei, Masoumeh Ghorbanipour, Alireza Nowroozi and Azam Amiri
- 24 **Nuclear quantum effects in fullerene–fullerene aggregation in water**
Sara Panahian Jand, Zahra Nourbakhsh and Luigi Delle Site
- 32 **Modelling eNvironment for Isoforms (MoNvIso): A general platform to predict structural determinants of protein isoforms in genetic diseases**
Francesco Oliva, Francesco Musiani, Alejandro Giorgetti, Silvia De Rubeis, Oksana Sorokina, Douglas J. Armstrong, Paolo Carloni and Paolo Ruggerone
- 40 **Generating a conformational landscape of ubiquitin chains at atomistic resolution by back-mapping based sampling**
Simon Hunkler, Teresa Buhl, Oleksandra Kukharenko and Christine Peter
- 52 **Modeling the molecular fingerprint of protein-lipid interactions of MLKL on complex bilayers**
Ricardo X. Ramirez, Oluwatoyin Campbell, Apoorva J. Pradhan, G. Ekin Atilla-Gokcumen and Viviana Monje-Galvan
- 65 **Influence of ionic conditions on knotting in a coarse-grained model for DNA**
Sarah Wettermann, Ranajay Datta and Peter Virnau
- 71 **Molecular dynamics simulation of an entire cell**
Jan A. Stevens, Fabian Grünewald, P. A. Marco van Tilburg, Melanie König, Benjamin R. Gilbert, Troy A. Brier, Zane R. Thornburg, Zaida Luthey-Schulten and Siewert J. Marrink
- 80 **The coexistence region in the Van der Waals fluid and the liquid-liquid phase transitions**
Dinh Quoc Huy Pham, Mateusz Chwastyk and Marek Cieplak
- 86 **May the force be with you: The role of hyper-mechanostability of the bone sialoprotein binding protein during early stages of *Staphylococci* infections**
Priscila S. F. C. Gomes, Meredith Forrester, Margaret Pace, Diego E. B. Gomes and Rafael C. Bernardi
- 95 **Assessing a computational pipeline to identify binding motifs to the $\alpha 2\beta 1$ integrin**
Qianchen Liu and Alberto Perez



OPEN ACCESS

EDITED AND REVIEWED BY

Hai Lin,
University of Colorado Denver,
United States

*CORRESPONDENCE

Adolfo B. Poma,
✉ apoma@ippt.pan.pl

RECEIVED 04 April 2023

ACCEPTED 06 April 2023

PUBLISHED 11 April 2023

CITATION

Poblete S, Pantano S, Okazaki K-i, Liang Z,
Kremer K and Poma AB (2023), Editorial:
Recent advances in computational
modelling of biomolecular complexes.
Front. Chem. 11:1200409.
doi: 10.3389/fchem.2023.1200409

COPYRIGHT

© 2023 Poblete, Pantano, Okazaki, Liang,
Kremer and Poma. 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.

Editorial: Recent advances in computational modelling of biomolecular complexes

Simón Poblete^{1,2}, Sergio Pantano³, Kei-ichi Okazaki⁴,
Zhongjie Liang⁵, Kurt Kremer⁶ and Adolfo B. Poma^{7*}

¹Instituto de Ciencias Físicas y Matemáticas, Universidad Austral de Chile, Valdivia, Chile, ²Computational Biology Lab, Fundación Ciencia & Vida, Santiago, Chile, ³Institut Pasteur de Montevideo, Montevideo, Uruguay, ⁴Department of Theoretical and Computational Molecular Science, Institute for Molecular Science, National Institutes of Natural Sciences, Okazaki, Japan, ⁵Center for Systems Biology, Department of Bioinformatics, School of Biology and Basic Medical Sciences, Soochow University, Suzhou, China, ⁶Max Planck Institute for Polymer Research, Mainz, Germany, ⁷Biosystems and Soft Matter Division, Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland

KEYWORDS

coarse-grained method, machine learning, multiscale approach, biopolymers, aggregation, GōMartini approach, Martini 3, nanomechanics

Editorial on the Research Topic

Recent advances in computational modelling of biomolecular complexes

The spatiotemporal description of the molecular interactions that rule the biological world poses tremendous challenges to current modeling and molecular dynamics (MD) simulation methods (Pantano, 2022). The deep intricacies of interactions spanning several orders of magnitude in time and space have prompted the scientific community to develop novel methods to enhance our understanding of biomolecular complexes.

We present a Research Topic illustrating state-of-the-art applications to study key constituents of biological matter. The modeling of large complexes demands the development of new approaches which are derived based on statistical and thermodynamic principles, such as the case of coarse-grained (CG) methods (Ingólfsson et al., 2022). Some CG studies in this Research Topic deal with the nanomechanics of protein complexes by the GōMartini approach (Liu et al., 2021; Mahmood et al., 2021), the first-ever CG modeling of an entire cell, coupling of different molecular resolutions (i.e., CG and all-atom) by the AdResS method and the study of double-stranded DNA. Figure 1 shows the integration of different methodologies for the study of biomolecular complexes.

A first example of the power of CG descriptions is found in the work of Wettermann et al., where a bead-stick model is used to represent double strands of DNA under different ionic conditions and study their topological features. The analysis of the probability of knot formations for systems of hundreds of thousands of base pairs can be directly compared with data from nanopore experiments. Moreover, their analysis predicts a scenario where the knotting probability is extremely low and therefore, useful for setting up experiments where the knots are undesired.

Molecular modeling plays a crucial role in identifying binding motifs in large protein systems such as integrins, yet it is limited by system sizes. At cellular scale, significant conformational changes led to mechanotransduction, cell adhesion, differentiation, etc. In such context, the perspective article by Liu and Perez shows traditional routes for identifying collagen-like motifs that bind the I-domain of the $\alpha 2 \beta 1$ integrin, addressing their limitations

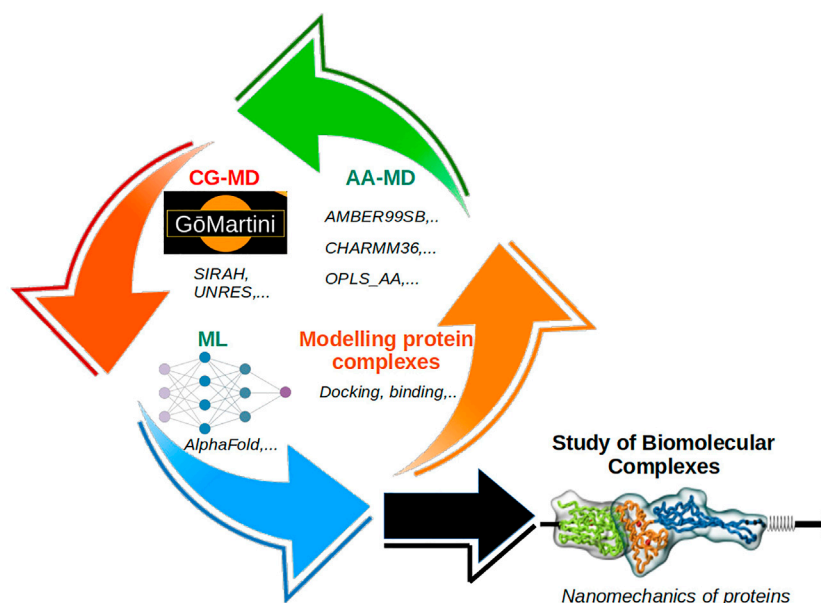


FIGURE 1

Schematic description of the integration of different methodologies employed in the study of biomolecular complexes: i) all-atom MD, ii) coarse-grained MD, iii) machine learning, and iv) modeling tools. Together they can cope with the study of complex molecular systems, such as biomechanics of protein complexes and other systems at the cellular level.

and presenting alternative solutions by machine learning approaches (i.e., AlphaFold). Also, [Oliva et al.](#) propose another tool for protein structure modeling, Modelling eNvironment for Isoforms (MoNvIso), which aims to discover protein structures for genetic diseases. The method was tested on 70 proteins which correspond to 257 human isoforms. This procedure can handle large sets of proteins, but it can only model protein regions where structural templates are given. A comparison with AlphaFold supports validation of the MoNvIso approach for large search determination of protein-protein interactions.

Beyond the description and prediction of structures and interactions, the dynamic interplay between biological partners has a fundamental role in describing biomolecular complexes, especially when quantum and classical levels are required. The work by [Jand et al.](#) employed the multiscale approach, denoted as Adaptive Resolution Scheme (AdResS), to study the quantum delocalization in space of the water molecules during the aggregation process of two fullerene molecules. Using path-integral MD for the quantum part and all-atom/CG MD description for the classical region, they show the relevance of quantum effects in the free energy profiles with consequences in the formation of fullerene complexes.

Classical all-atom MD simulations are key for understanding biomolecular complexes, as they capture local conformations enabling the calculation of free energy profiles. [Ramirez et al.](#) elucidated the fingerprints of necroptotic pathways driven by the electrostatic interactions in protein-lipid complexes. [Zargari et al.](#) report on free energy calculation of protein-ligand by funnel metadynamics using all-atom MD. Their results successfully combine all-atom MD and enhanced sampling in docking studies. The work by [Pham et al.](#) combines statistical mechanics and MD simulation to obtain a better understanding of liquid-liquid phase transitions in cellular organelles.

They propose several quantities to characterize the metastability regime, such as specific heat, surface tension, feature in molecular clusters, etc. They employed a Lennard-Jones model for the analysis of liquid-liquid transitions.

Similarly, [Gomes et al.](#) employed the Martini 3 force field combined with the GōMartini approach to capture the mechanical stability of bone sialoprotein binding protein in the early stages of Staphylococci infections, namely, the Bbp:Fga complex. It required sampling significant conformational changes in protein complexes by means of steered CG and all-atoms MD simulations. The approach accurately described the stabilization mechanism of the Bbp:Fga complex. The high force-loads present during the initial stages of bacterial infection stabilize β -sheet motifs in both proteins that, due to their position in the complex, cannot be peeled as in another bacterial system.

Multiscale modeling usually requires reintroducing all-atom details onto the CG trajectories to generate a complete atomistic picture. This task can be as challenging as the design of the CG model. Moreover, the simplification introduced by the CG simulation might generate conformations that have no correspondence in an all-atom representation. These important Research Topic are addressed by [Hunkler et al.](#), extending the Back-mapping Based Sampling (BMBS) to large systems, by applying it to the simulation of K48-linked triubiquitin. The authors discuss how to correct the inaccuracies generated by the exploration in the CG level and distinguish relevant regions on a low-dimension projection of the conformational space. Their approach allows them to confidently access, with the all-atom resolution, parts of the conformation space that are very difficult or nearly impossible to explore by plain MD simulations.

Finally, a remarkable example of the capability of integrating coarse-grained representations, molecular modeling, and simulation techniques is provided by [Stevens et al.](#) They combined a large volume of

experimental data (i.e., cryoEM, cryoET, -omics data) to produce a model of a JCVI-syn3A cell. The integrative approach required the development of mesoscopic models integrated into the Martini 3 ecosystem by standard toolkits such as Polyply, Martinize2, and TS2CG. The length scale is nearly half a micrometer, with 561 million CG beads representing more than 6 billion atoms. The size and architecture of this cellular model represent a milestone in building a particle-based whole-cell model.

Certainly, subsequent multiscale simulations combining the techniques illustrated in this Research Topic will be instrumental in leading us to the next level of understanding and integration in cellular and structural biology.

Author contributions

SPO, SPA, and AP contributed to the writing of this editorial. All authors have read the last version of the manuscript.

Funding

SPO acknowledges support from FONDECYT grant 1231071. AP acknowledges financial support from the National Science Center, Poland, under grant 2022/45/B/NZ1/02519 and computational resources were supported by the PL-GRID

References

- Ingólfsson, H. I., Lopez, C. A., Uusitalo, J. J., de Jong, D. H., Gopal, S. M., Periole, X., et al. (2014). The power of coarse graining in biomolecular simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 4 (3), 225–248. doi:10.1002/wcms.1169
- Liu, Z., Moreira, R. A., Dujmović, A., Liu, H., Yang, B., Poma, A. B., et al. (2021). Mapping mechanostable pulling geometries of a therapeutic anticalin/CTLA-4 protein complex. *Nano Lett.* 22 (1), 179–187. doi:10.1021/acs.nanolett.1c03584
- Mahmood, M. I., Poma, A. B., and Okazaki, K. I. (2021). Optimizing gō-MARTINI coarse-grained model for F-bar protein on lipid membrane. *Front. Mol. Biosci.* 8, 619381. doi:10.3389/fmolb.2021.619381
- Pantano, S. (2022). Back and forth modeling through biological scales. *Biochem. Biophysical Res. Commun.* 633, 39–41. doi:10.1016/j.bbrc.2022.09.037

infrastructure. This work was partially funded by FOCEM (MERCOSUR Structural Convergence Fund), COF 03/11.

Acknowledgments

AP dedicates this work to Tania, David and Isabel for their unconditional support and inspiring science on me. SPO dedicates this work to Delba and Juan for their loving support.

Conflict of Interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



OPEN ACCESS

EDITED BY

Kei-Ichi Okazaki,
Institute for Molecular Science (NINS),
Japan

REVIEWED BY

Stefano Pieraccini,
University of Milan, Italy
Rodrigo Azevedo Moreira da Silva,
Basque Center for Applied Mathematics,
Spain

*CORRESPONDENCE

Zahra Nikfarjam,
nikfarjam.zahra14@gmail.com

SPECIALTY SECTION

This article was submitted to Theoretical
and Computational Chemistry,
a section of the journal
Frontiers in Chemistry

RECEIVED 08 June 2022

ACCEPTED 18 August 2022

PUBLISHED 23 September 2022

CITATION

Zargari F, Nikfarjam Z, Nakhaei E,
Ghorbanipour M, Nowroozi A and
Amiri A (2022), Study of tyramine-
binding mechanism and insecticidal
activity of oil extracted from *Eucalyptus*
against *Sitophilus oryzae*.
Front. Chem. 10:964700.
doi: 10.3389/fchem.2022.964700

COPYRIGHT

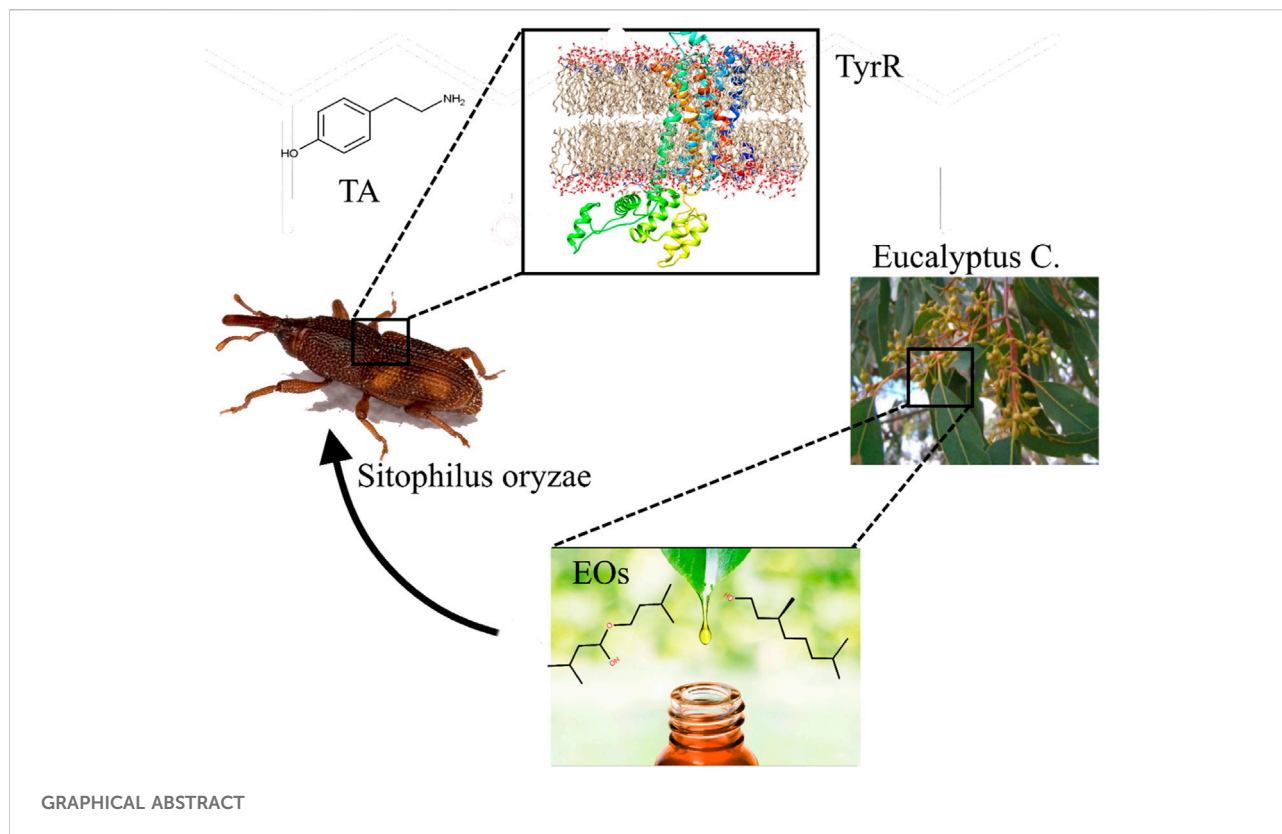
© 2022 Zargari, Nikfarjam, Nakhaei,
Ghorbanipour, Nowroozi and Amiri.
This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright 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.

Study of tyramine-binding mechanism and insecticidal activity of oil extracted from *Eucalyptus* against *Sitophilus oryzae*

Farshid Zargari^{1,2}, Zahra Nikfarjam^{3*}, Ebrahim Nakhaei²,
Masoumeh Ghorbanipour⁴, Alireza Nowroozi² and Azam Amiri⁵

¹Pharmacology Research Center, Zahedan University of Medical Sciences, Zahedan, Iran, ²Department of Chemistry, Faculty of Science, University of Sistan and Baluchestan (USB), Zahedan, Iran, ³Department of Physical & Computational Chemistry, Chemistry and Chemical Engineering Research Center of Iran, Tehran, Iran, ⁴Department of Physical Chemistry, Faculty of Chemistry, University of Tabriz, Tabriz, Iran, ⁵College of Geography and Environmental Planning, University of Sistan and Baluchestan, Zahedan, Iran

The rice weevil, *Sitophilus oryzae* (L.), is a major pest of stored grains throughout the world, which causes quantitative and qualitative losses of food commodities. *Eucalyptus* essential oils (EOs) possess insecticidal and repellent properties, which make them a potential option for insect control in stored grains with environmentally friendly properties. In the current study, the binding mechanism of tyramine (TA) as a control compound has been investigated by funnel metadynamics (FM) simulation toward the homology model of tyramine1 receptor (TyrR) to explore its binding mode and key residues involved in the binding mechanism. EO compounds have been extracted from the leaf and flower part of *Eucalyptus camaldulensis* and characterized by GC/MS, and their effectiveness has been evaluated by molecular docking and conventional molecular dynamic (CMD) simulation toward the TyrR model. The FM results suggested that Asp114 followed by Asp80, Asn91, and Asn427 are crucial residues in the binding and the functioning of TA toward TyrR in *Sitophilus Oryzae*. The GC/MS analysis confirmed a total of 54 and 31 constituents in leaf and flower, respectively, where most of the components (29) are common in both groups. This analysis also revealed the significant concentration of *Eucalyptus* and α -pinene in leaves and flower EOs. The docking followed by CMD was performed to find the most effective compound in *Eucalyptus* EOs. In this regard, butanoic acid, 3-methyl-, 3-methyl butyl ester (B12) and 2-Octen-1-ol, 3,7-dimethyl- (B23) from leaf and trans- β -Ocimene (G04) from flower showed the maximum dock score and binding free energy, making them the leading candidates to replace tyramine in TyrR. The MM-PB/GBSA and MD analysis proved that the B12 structure is the most effective compound in inhibition of TyrR.



KEYWORDS

eucalyptus camaldulensis, sitophilus oryzae, GC/MS, tyramine binding mechanism, funnel metadynamics

1 Introduction

Stored-product insect pests are very popular in cereal industries, as their metabolic wastes and body parts have a detrimental impact on buyer satisfaction (Neethirajan et al., 2007; Chang et al., 2017). It has been reported that 10%–15% of grains are damaged during postharvest in developing countries (Kumar and Kalita, 2017). The rice weevil, *Sitophilus oryzae* (L.) (Coleoptera: Curculionidae), is a significant insect affecting cereals worldwide (Ahmed, 2001). Its eggs are laid on cereal grain, and the incubated hatchlings dig out a complete grain, where they pupate till they mature into adult weevils (Sharifi and Mills, 1971). Getting into grains of rice, weevil causes quantitative and qualitative alterations and losses of nutritional value and germination, acts as contamination in food commodities with insect bodies and excrement, and most importantly, encourages the growth of storage fungi (Mondal et al., 2016; Seada et al., 2016). While phosphine or even other compounds actually have been employed as fumigants to control rice weevils (Hymavathi et al., 2011), resistance to

phosphine administration has been a significant challenge in rice weevil management (Nayak et al., 2007; Hossain et al., 2014). In order to protect stored grain products, additional antiinsect pest techniques are required. Plant-derived essential oils (EOs), derived through nonwoody portions of the plants, mainly foliage, have insecticidal and repellent qualities and can be used to control insects such as *Sitophilus oryzae* in stored grains (Taylor et al., 2007; Kiran and Prakash, 2015; Hossain et al., 2017). These compositions determine the characteristics of plants and supply plants with a crucial defense plan, especially against herbivorous insect pests and harmful fungus (Dhifi et al., 2016). The genus *Eucalyptus* is one of the most planted species, which include volatile oils in their leaves (Brooker and Kleinig, 1990). For years, essential oils from several *Eucalyptus* species have been employed in the medicinal, beauty, and food fields (Marzoug et al., 2011). Based on previous studies, *Eucalyptus* EO exhibited the highest toxicity to the rice weevil across a variety of EO treatments (Hossain et al., 2019; Ebadollahi and Setzer, 2020). Furthermore, components including 1,8-cineole, citronellal,

citronellol, citronellyl acetate, p-cymene, eucamalol, limonene, linalool, -pinene, -terpinene, -terpineol, alloocimene, and aromadendrene have been linked to the insecticidal activity of Eucalypt (Batish et al., 2008). Among various species of *Eucalyptus*, *Eucalyptus camaldulensis* has the most comprehensive natural distribution, and its essential oils (EOs) have a more complex makeup, with third components accounting for 95% of the total leaf oil found (Dhakad et al., 2018).

According to the literature, the presence of two biogenic amines, octopamine (OA) and tyramine (TA), in high concentrations in the nervous systems of invertebrates shows their pivotal role in neurotransmission, neuromodulation, and neurohormones in insects (Ohta and Ozoe, 2014). As the appearance of octopamine (OctR) and tyramine (TyrR) is limited to invertebrates, two seven-transmembrane protein receptors belonging to a class A G protein-coupled receptor (GPCR) family are the targeting receptors for the introduction of the new bioactive compounds against insects (Degen et al., 2000).

Despite pesticides available for targeting OctR and TyrR in insects (Kostyukovsky et al., 2002), the atomic-level understanding is still in demand to shed light on the OA and TA mechanism of action toward specific target receptors to find and develop new drugs.

Demands for accurate in silico techniques lead researchers to find a visually appealing and cost-effective method to convey valuable, relevant data on protein–ligand binding such as ligand-binding mode, ligand binding free energy, and binding kinetics properties (Broomhead and Soliman, 2017; Raniolo and Limongelli, 2020). Funnel metadynamics (FM) (Limongelli et al., 2013) is a binding free-energy method that attempts to simulate a bias potential flexibly created as a combination of Gaussian functions in the region of chosen degrees of freedom termed collective variables (CVs) to model the binding process of a ligand from its own completely solubilized form to the eventual binding site (Laio and Parrinello, 2002). When the approximate position of the binding site in the protein structure is known but there is little or no information about the ligand-binding mechanism, these strategies come in handy. This method can identify the ligand-binding mode, clarify the dynamics of the ligand-binding mechanism, and calculate the absolute protein–ligand binding free energy (Limongelli et al., 2013; Hsiao and Söderhjelm, 2014; Troussicot et al., 2015; Comitani et al., 2016; Saleh et al., 2017; Saleh et al., 2018; Yuan et al., 2018; Wang et al., 2021). So far, according to the authors' knowledge, the binding mechanism between TA and *Sitophilus oryzae* TyrR and the interaction models between this receptor and some monoterpenes have been studied by some researchers (Braza et al., 2019; Ocampo et al., 2020), but there is no systematic approach to this issue. On the other hand, a detailed examination of the interactions between the components of EO *E. camaldulensis* as a bioinsecticide and molecular targets of *Sitophilus oryzae* has been published;

therefore, in the current research, the interaction between *E. camaldulensis* essential oil as a control agent and molecular targets of *Sitophilus oryzae* as stored product pests has been explored to 1) determine the chemical composition of *E. camaldulensis* EOs, 2) to identify the EO components with the highest affinity to insect molecular targets, and 3) analyze the mechanism of action of more stable EO components on insect molecular targets.

2 Material and methods

2.1 Preparation of *eucalyptus camaldulensis* material and extraction

First, the aerial parts of *Eucalyptus camaldulensis* (leaf and blossoms) were collected from the botanic farm of the University of Sistan and Baluchestan (USB). Fresh leaves and flowers were disinfected, dried in the sun, and afterward made into a fine powder in a blender. Hydrodistillation with Clevenger (Unividros®) equipment and a heated mantle are used to extract the EO. After removing the organic matter, 100 g of plant material was weighed and transferred to a 1 L flask, which was half-filled using distilled water. The extraction took nearly 3 h. Anhydrous sodium sulfate (Na_2SO_4 , Synth®) was used to eliminate trace water from the oil, which was collected in a container. This technique was repeated to extract roughly 3 ml of pure essential oil, and the extracts were subjected to GC/MS studies.

2.2 Gas chromatography-mass spectrometry analysis

The analysis of extracted phytochemicals compounds was done with GC-MS (Agilent Technologies 7890B—GC systems 5977A MSD) using the electron impact (EI) mode (ionizing capability 70 eV) and a capillary column (VF-5 ms) (50 m × 0.25 mm, film thickness 0.25 μm) filled with 5% phenyl dimethyl silicone, and the ion supply temperature used was 250°C. In addition, the GC/MS settings are as follows: the preliminary column temperature was set at 35°C and maintained for 5 min; the temperature was increased to 260°C at a rate of 5°C/min, and the split ratio was 1:10. The fraction composition of the samples was computed from the GC peak regions (Supplementary Figures S1, S2). The molecular structure of chemical compounds was approved using the WILEY8, NIST08s, and FAME libraries and is listed in Supplementary Tables S1, S2. The chemical composition of *Eucalyptus camaldulensis* oil revealed the 54/31 constituents for leaves/flowers (Supplementary Figures S3, S4). Among the components of leaves, eucalyptol (22.50%), α-pinene (14.33%), 1H-Cycloprop[e]azulene, decahydro-1,1,7-

trimethyl-4-methylene (9.01%), β -pinene (6.32%), and (-)-globulol (5.01%) are the most abundant species, while the major constituents of flowers are eucalyptol (26.5%), α -pinene (16.24%), globulol((-)-globulol) (5.93%), β -pinene (5.80%), and γ -terpinene (5.23%). Evaluating the chemical structure of these species show that most of them (29) are common, while the bicyclo[3.1.0]hex-2-ene,2-methyl-5-(1-methylethyl)- and 1H-Indene,1-ethylideneoctahydro-7a-methyl-,(1E,3a α ,7a β) are only observed in the flower oil. In contrast, all of the other compounds (25) are only related to the oil of leaves.

2.3 Homology modeling

It is necessary to find the crystal structure with high-sequence similarity to the TyrR of *Sitophilus oryzae* in the homology modeling process. The amino acid sequence came from the UniProt database (ID A0A0S1VX60) (Masson et al., 2015). The CLUSTALX program was also used instantly from its website at <https://www2.ebi.ac.uk/CLUSTALX> to align the sequence of the TyrR receptor to that of the D2 dopamine receptor as the template (PDB ID:6CM4) (Thompson et al., 1997). MODELLER (Šali and Blundell, 1993) version 10.1 is used to create homology models of TyrR using the D2 dopamine receptor crystallographic structure and the methods implemented in MODELLER. The 3D models all comprising nonhydrogen atoms were automatically generated from the alignments. The model with the lowermost probability density function (pdf) and the fewest constraint violations was chosen out of 1,000 for further investigation. To improve loops of the chosen model, an ab initio method implemented in the MODELLER was applied. The MODELLER was used to determine the root means square (RMS) deviations of the models concerning the template (6CM4) and determine the R differences using template geometry for bond lengths and angles. MODELLER was also used to determine the R differences using template geometry for bond lengths and angles. The software PROCHECK evaluated the overall stereochemical value of the results and produced a model for each tyramine receptor type (Laskowski et al., 1996). PROCHECK was used to determine the G-factor for the proposed model. In addition, the Verify-3D is also used to validate the environmental profile of the final generated model (Lüthy et al., 1992).

2.4 Docking studies

Protein–ligand docking was initiated using LeDock software (<http://lephar.com>). The initial structure of all compounds, including a total of 54/31 constituents of leaves/flowers that were obtained from gas chromatography-mass spectrometry analysis, was sketched using HyperChem (Teppen, 1992). Then, the geometry optimization and calculation of electronic

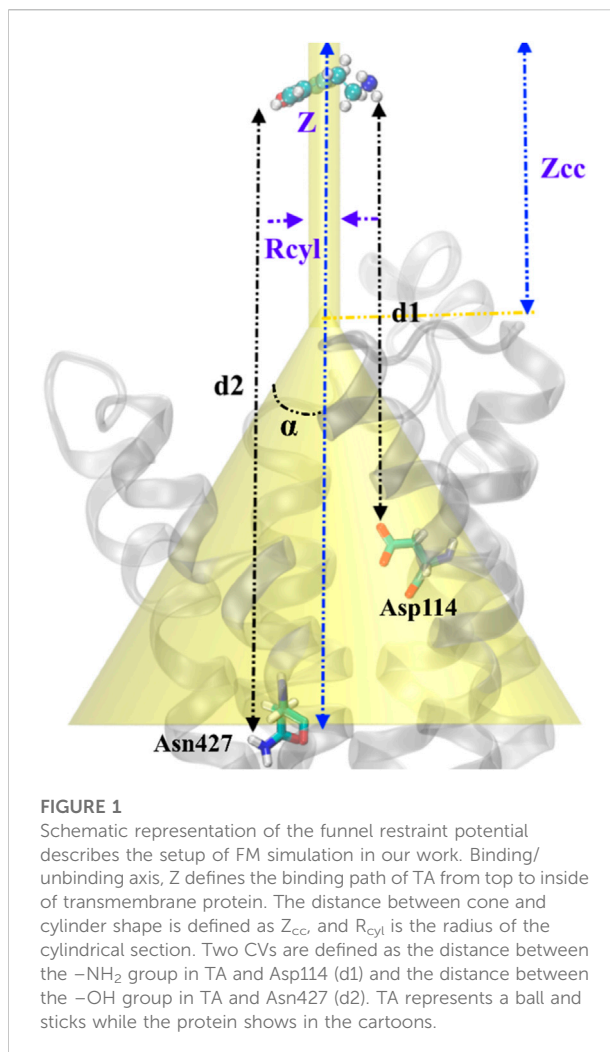


FIGURE 1

Schematic representation of the funnel restraint potential describes the setup of FM simulation in our work. Binding/unbinding axis, Z defines the binding path of TA from top to inside of transmembrane protein. The distance between cone and cylinder shape is defined as Z_{cc} , and R_{cyl} is the radius of the cylindrical section. Two CVs are defined as the distance between the $-NH_2$ group in TA and Asp114 (d1) and the distance between the $-OH$ group in TA and Asn427 (d2). TA represents a ball and sticks while the protein shows in the cartoons.

energy of the benchmark systems were performed using ORCA software (Hočevar and Demšar, 2016) at the DFT, B3LYP/cc-pvdz level of theory. The homology model of *Sitophilus oryzae* TyrR was selected for docking and subjected to the LePro module (<http://lephar.com>) for pretreatment of the macromolecule. The docking parameters, including the active site of the protein, were set so that the box with the dimensions of $16 \times 16 \times 16 \text{ \AA}$ was placed in the center of D114 and N427, as these are the most critical residues in the active site of TyrR. The number of binding poses and the spacing value are set to 100 and 1.0 \AA , respectively. The conformation with the lowest binding energy and the most interacting residues were chosen as the best.

2.5 Funnel-metadynamics simulation setup

Funnel metadynamics (FM) (Limongelli et al., 2013) simulations were performed using well-tempered

metadynamics (Barducci et al., 2008). The funnel parameters are properly defined based on a previous study on GPCRs (Saleh et al., 2017). The PLUMED plugin, the master version (Bonomi et al., 2009), coupled with GROMACS 2020.1 (Pronk et al., 2013), was employed to carry out ~360 ns of metadynamics simulations in the NPT ensemble. The computational protocol was built by setting the initial Gaussians height at 1.0 kJ/mol and their width at 0.01 Å for the distance between the nitrogen/oxygen atom of tyramine with Asp114 (d1)/TyrR427 (d2) CVs. Gaussians were added every 500 steps (1 ps) so that the deposition rate was equal to 1 kJ/mol.ps. The bias factor was set to 20; consequently, ΔT was 3600 K. The cluster analysis of the conformations found in basin A was performed using the GROMOS algorithm (Daura et al., 1999) of the g-cluster tool implemented in the GROMACS. The absolute TA/TyrR binding free energy was calculated using the following equation (Limongelli et al., 2013):

$$\Delta G_b^0 = -\frac{1}{\beta} \ln(k_b) \quad (1)$$

where K_b represents the equilibrium binding constant and can be computed as follows:

$$K_b = C^0 \pi R_{\text{cyl}}^2 \int dze^{-\beta(w(z)-w_{\text{ref}})} \quad (2)$$

where C^0 is the standard concentration of 1 M and is equal to $1/1.660 \text{ Å}^{-3}$ and πR_{cyl}^2 is the surface of the cylinder used as a restraint potential in the unbound state. In contrast, the potential $W(z)$ and its value in the unbound state, W_{ref} , can be derived from the potential of mean force (PMF) obtained through FM calculations. β is a constant and equal to $1/k_B T$, where k_B and T are the Boltzmann constant and the system's temperature, respectively. Considering cylinder radius $R = 1 \text{ Å}$, a schematic of the setup related to the funnel metadynamics is depicted in Figure 1.

2.6 Conventional molecular dynamics (CMD) simulations

2.6.1 System setup

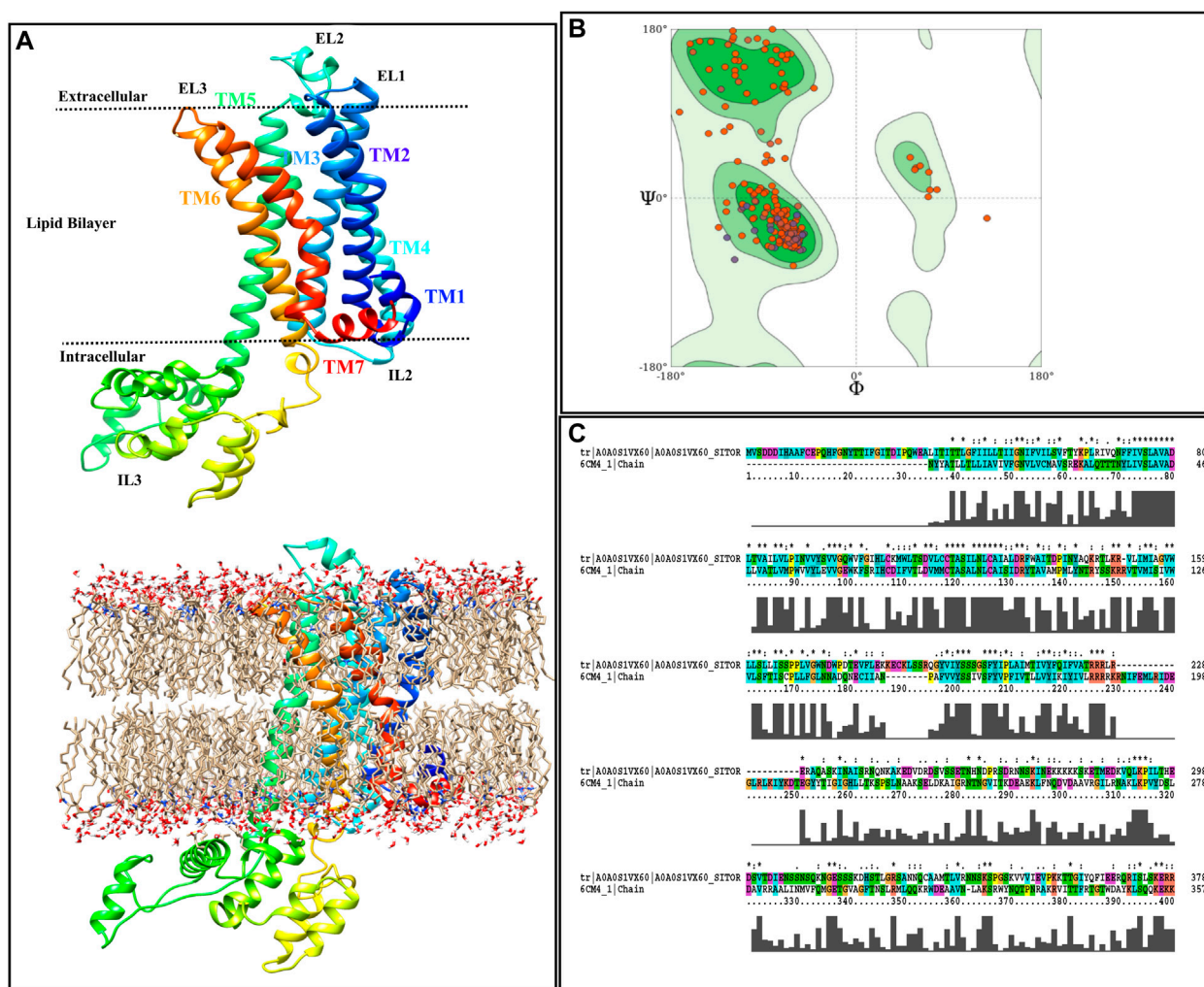
Compounds with the best docking pose were chosen to study the interactions of ligands with the active site of the TyrR and to examine the inhibitory efficacy of the ligands. The homology model of TyrR was embedded in the Sphingo and Ceramide Lipid model, which was suggested to be the central part of membrane proteins in insects (Zhou et al., 2013). The membrane was therefore oriented toward the XY plane, bringing the GPCR main axis and the Z-axis near to parallel. Also, the VDW and bonded parameters of the TA and the general amber force field (GAFF) were used to detect specified compounds following docking using AmberTools' antechamber program (Wang et al., 2004), while the protein was modeled by the AMBERff14SB force field (Maier et al., 2015). The partial

atomic charges are also calculated by considering the RESP charge model (Vanquelef et al., 2011). The TIP3P water model was used for full solvation, and 0.15 M KCL was employed to neutralize the system. In three dimensions, the periodic boundary condition (PBC) was employed, and all MD simulations were done using a parallel version of SANDER in the AmberTools 19 software package (Case et al., 2018). It is worth noting that, before the MD simulation of protein–ligand complexes, the steepest descent approach was employed to reduce their efficiency and energy, as well as a leap-frog algorithm to integrate their movements (Hockney and Eastwood, 1988). In this procedure, to figure out the effect of long-range electrostatic interactions of molecules, the particle mesh Ewald (PME) method, much like the preceding studies, was implemented (Darden et al., 1993). In addition, the constraints applied on H-bonds using the LINCS algorithm in both equilibration and production run (Hess et al., 1997). The cutoff for nonbonded interactions was set to 12.0 nm. After the optimization of the energy of the system, it was simulated for 200 ps within the canonical ensemble (NVT) and with a 1 ns time-step within the NPT ensemble. Moreover, two models, including the Langevin dynamic model (Goga et al., 2012) and the Parrinello–Rahman one (Parrinello and Rahman, 1981), were served using coupling constants of 0.1 and 0.5 ps to couple the temperature and pressure of the system.

6.2.2 Free energy calculations, energy decomposition, and clustering

Molecular docking is the most popular method in structure-based drug design (Hu and Shelver, 2003), which is applied chiefly to predict the binding pose of candidate drugs in the predefined active site of the protein. However, the accuracy of free energy calculation by docking score might be argued in terms of its reliability in distinguishing between compounds with a comparable binding affinity (Hu and Shelver, 2003).

Among the several methods being used for calculation of binding free energy of ligand–protein complex, the molecular mechanic energies coupled with the Poisson–Boltzmann surface area (MM/PBSA) or generalized Born surface area (MM/GBSA) are proposed in terms of their accuracy and efficacy (Genheden and Ryde, 2012). Here, we used these methods to calculate the relative binding free energies of selected compounds extracted from *Eucalyptus*'s Eos. We used a single MD trajectory of the bound complex in our calculations, and 1,500 snapshots were employed from 15 replicas to estimate the binding free energy of each ligand. To obtain binding free energy of the ligands bound to TyrR, the MMPBSA.py package has been employed (Miller et al., 2012). We used the modified GB model which is consistent with PB behavior in the electrostatic part of the solvation energy (Feig et al., 2004). The *saltcon* parameter was set to 0.15 M for reconciliation between PB and GB solvation energies, as previously described (Srinivasan et al., 1999).

**FIGURE 2**

Homology modeling procedure of TyrR is shown. **(A)** The TyrR model describing TM and loop sections and showing the intra/extracellular boundaries of the protein lustration as implicit (top) and explicit (bottom) lipid bilayer, **(B)** Ramachandran diagram is presented to the stereochemical quality of the model made **(C)** alignment of the target sequence to the dopamine D2 receptor with the CLUSTALX program is shown.

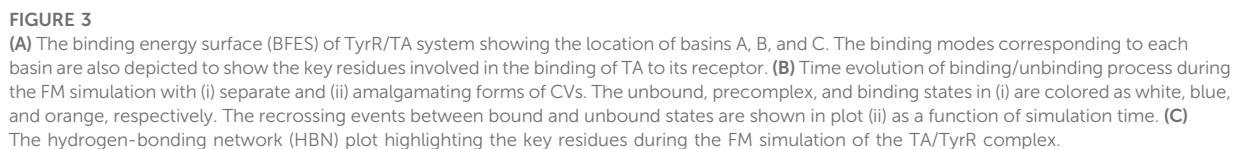
Decomposition of energy for each residue is defined as the most significant contribution of each residue to the ligand binding, and is classified as the polar, nonpolar, VDW, and electrostatic part of energy for every single residue. We used the water swap residue-wise binding energy decompositions in our work (Kiani et al., 2019).

Clustering of MD frames is, in particular, beneficial for molecular docking simulations. In step with some standards, MD frames that can be positioned within the identical group are just like each other. Consequently, one may want to assume that the alike clusters will behave similarly if a receptor in a cluster interacts agreeably with a selected ligand. The most conventional and regarded degree of similarity is the root mean square deviation (RMSD) values obtained for partitioning MD

trajectories, which can be received through pairwise or matrix error distances (De Paris et al., 2015).

2.7 Building of the TyrR model and molecular docking

The absence of the crystal structure of TyrR forces us to construct its 3D homology model. Hence, the MODELLER (Šali and Blundell, 1993) was used, employing the D2 Dopamine Receptor as a template to build the structure. It is suggested that structural and sequence similarity within TM regions, in terms of its quality and importance in ligand binding, is preferable to those within



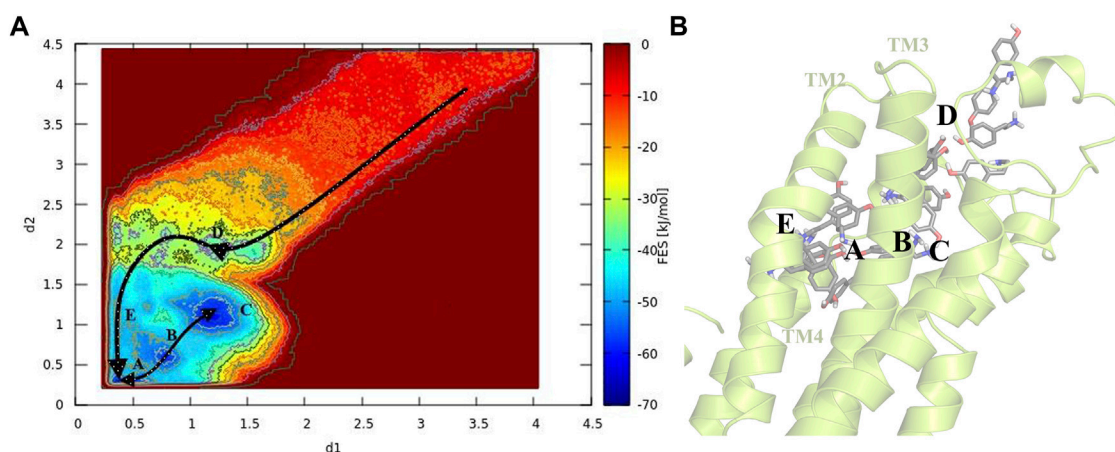


FIGURE 4

Schematic illustration of the binding pathway trajectory of TA against TyrR is depicted on (A) the FES and (B) the corresponding 3D structure of the protein. The trajectory showing the binding path is shown as arrows connecting each basin.

intra- or extracellular loops (Mirzadegan et al., 2003; Kinoshita and Okada, 2015).

GMQE (Global Model Quality Estimate) is a quality estimate, which combines properties from the target–template alignment and the template structure. This property for our model is 0.41, which is expected for the model. Despite the low percentage of sequence similarity between target and template (36.7%), it can still be stated that the obtained TyrR model possesses this quality, especially in the transmembrane region where the natural ligand binds (Cavasotto and Phatak, 2009).

Figure 2A illustrates the acquired model annotating TM regions, intra- and extracellular loops, as well as showing the boundaries in which each part is placed. Moreover, the starting model, inserted in the Sphingo and ceramide lipid, constructed by the CHARMM-GUI membrane builder module (Jo et al., 2007), is also represented in Figure 2A. The stereochemical quality of the constructed model is also reported as a Ramachandran plot, and the results are shown in Figure 2B. According to this figure, by the majority of residues in the allowed region, the quality of the model can also be confirmed for further analysis. Figure 2C represents the alignment of the target sequence on the D2 Dopamine Receptor with the CLUSTALX program. The essential residues that play a vital role in the binding of TA in *Bombyx mori*, including Asp114 in TM3, Ser200, and Ser204 in TM5 (Ohta et al., 2004), are conserved in target and template.

Molecular docking is an essential device in structural biology and computer-aided drug design (CADD), in which two molecules fit together in a 3D area [9]. In the present work, the 3D model of TyrR has been constructed as previously described: the active site of the insect's TyrA receptors including Asp114 residue in TM3 and Ser200 and Ser204 in TM5 (Ohta et al., 2004). The leaf and flower ligands were obtained from PubChem

databases and saved in a structure-data file (SDF) format. The ligands were docked onto the TA receptor using LeDock, and the obtained docking energies are depicted in Supplementary Table S1.

According to Supplementary Table S1, the binding affinity of ligands with the receptor active site can be easily discussed by comparing the docking scores. It has been seen that butanoic acid, 3-methyl-, 3-methylbutyl ester (−2.88 kcal/mol), 2-octen-1-ol, 3,7-dimethyl- (−2.86 kcal/mol), citronellol (−2.82 kcal/mol), trans- β -Ocimene (−2.48 kcal/mol), 1,3,6-Octatriene, 3,7-dimethyl-, (Z)- (−2.42 kcal/mol), 1,4-eicosadiene (−2.39 kcal/mol), and 3-eicosyne (−2.14 kcal/mol), respectively, had high binding affinity on the TA receptor than the other compounds in leaf. Also, the docking score of *E. camaldulensis* flower oil structures with TyrR shows that high binding affinity relies upon butanoic acid, 3-methyl-, 3-methyl butyl ester (−2.87 kcal/mol) and β -ocimene (−2.47 kcal/mol) compounds. By taking the docking score of tyramine as a reference binding energy (−2.84 kcal/mol), one can conveniently interpret the binding affinity competition of leaf and flower ligands with TA receptor against tyramine. The results indicate that butanoic acid, 3-methyl-, 3-methylbutyl ester, 2-Octen-1-ol- 3,7-dimethyl-, citronellol, and trans- β -ocimene with the maximum dock score are the main candidates to be replaced instead of tyramine in TA receptor and disrupt its function, the result of may lead to the insect's death. Finally, for better analyses of these interactions, the 3D structures of TA receptor, tyramine, and the mentioned compounds were selected to proceed toward ligand–protein molecular dynamics studies.

2.8 Molecular dynamics simulation

It is important to note that even if based on the analysis of the docking results, it is stated that the ligand is placed in

TABLE 1 Affinity and binding free energy of TA against TyrR were obtained from FM simulation and compared with available experimental data. The binding free energies and the binding affinities are calculated for basins A and C.

	IC ₅₀ , exp	Basin A		Basin C	
		ΔG_{calc}	K _i ^b	ΔG_{calc}	K _i
TA affinity	5.19 ¹	-11.0 ± 0.8	8.64	-11.2 ± 0.8	6.17

Experimental affinity data for *Bombyx mori* was obtained from Ohta et al. (2004) and was converted with the relation $\Delta G = -RT \ln (K_i)$. ^b The IC₅₀, K_i, and ΔG units are in nM, nM, and kcal.mol⁻¹, respectively.

a suitable binding state, it should be kept in mind that in the results obtained from the docking, the effects related to the solvent and temperature are not included. In this regard, the more accurate results related to the binding of the ligand in the activator of the studied protein have been made reliable using MD simulations, and after that, relevant analyses have been performed on the necessary and effective molecular interactions for ligand–protein binding. They also showed the dynamic behavior of the complex at the atomic level in a flexible manner that treated the ligand–receptor complex.

3 Result

3.1 Identifying the binding pose of TA by FM

The funnel-shaped restraining potential was set in a way that its cone was placed on a region surrounding all crucial residues in the proposed binding site to avoid the influence of the restraining potential on the ligand-binding mode. We chose and optimized the dimension for the cone to boost the convergence (Figure 3). As a first choice for the CVs (Figure 3B), we selected the distance between the oxygen atom in TA and the CG atom in Asp114 as d1 and the distance between the nitrogen atom of TA and the ND2 atom in TyrR427 as d2. The convergence was observed after 0.36 μ s when the ligand started from the unbound state where it was fully solvated in the water phase, at the extracellular region, and finally found its way to explore the binding site. Several recrossing events were achieved in this trajectory, thereby providing a quantitatively well-characterized FES and an accurate estimation of TyrR-TA binding free energy. The three lowest energy minima (basins) have been detected from the FES corresponding to point A, point B, and point C in Figure 3A. In basin A, TA adopted a configuration in which a hydrogen bond between the –OH group of TA and ND2 atom of Asn427 and two hydrogen bonds between the –NH₂ group of TA and O and OD atoms of Asp114 occurred.

This basin corresponds to the free energy of –62.7 kJ/mol. The configuration of TA in basin C has the same binding energy of –63.0 kJ/mol. The TA is involved in hydrogen bonds between its –OH, –NH₂ moieties, OD1 atom of Asp114 and OD1 atom of Asp 80, respectively. These findings suggested that Asp 114 is a crucial residue in the binding and the function of TA toward TyrR in *Sitophilus Oryzae* (Figure 3C).

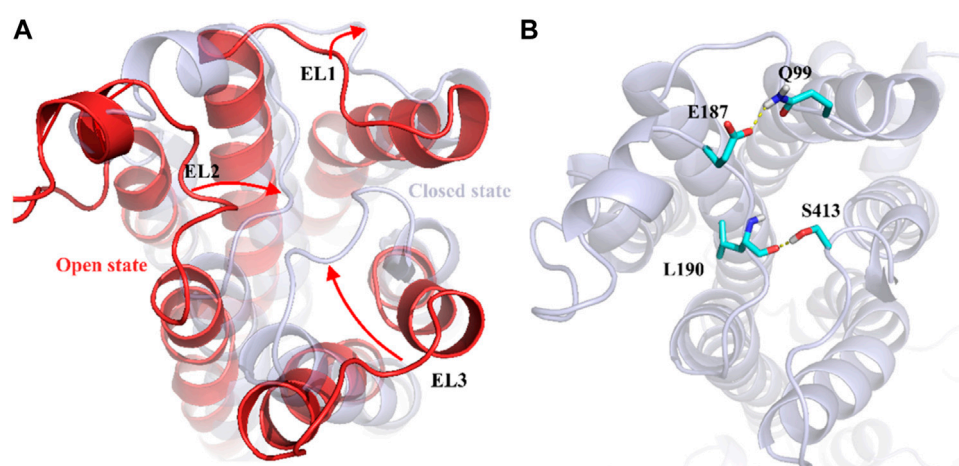
In basin B, which corresponds to the –58.3 kJ/mol in FES, we observed the water-mediated binding mode, which sheds light on the water's role in binding structures of TA. In this mode, we can see the hydrogen bond formed between the –NH₂ group of TA and Asn91, a water-mediated hydrogen bond between the –OH moiety in TA and Asp80 (Figure 3C).

3.2 TA binding path and evaluating the binding free energy

To gain a better perception of binding events of TA on the receptor, a rigorous method was required to sample the path of binding/unbinding and produce an exact FES. Hence, we exploited the FM simulation to obtain a quantitatively well-described free energy landscape of ligand binding and calculate the binding free energy of TA against TyrR. In this regard, we track the binding events during the binding process of TA, and the results are depicted in Figure 4. However, as mentioned before in Figure 3A which points to the ligand-binding pathway in reconstruction of a full energy landscape, the arrows are used to illustrate the path constructed by each basin and also the path the ligand adopted during the binding pathway. Figure 4B depicts the frames containing the ligand obtained from the FM trajectory corresponding to each basin in the free energy landscape. The ligand enters when ELs are in the open state (see the next section) and reaches the binding site cleft among the TMs after several binding/unbinding events. In this stage, the ligand dropped in basin D and then tried to find its pathway toward basin E, which corresponds to the cleft between TM2 and TM4. The ligand spent some time in this basin and then found its absolute binding modes corresponding to basins A and C with an intermediate binding mode (basin B), which facilitates the conversion between them (Figure 4).

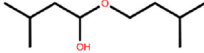
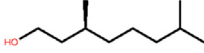
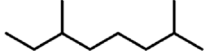
For the evaluation of the absolute binding free energy of TA, the two minima A and C in FES were considered as bound states, and the ligand at the starting point of the simulation submerging in the bulk water was deemed to be the unbound state. The TA binding free energy is calculated initially between these two states. Table 1 represents the binding energy for two basins, A and C, concerning the unbound state, considering the analytical correction.

With the lack of experimental data for the binding affinity of TA toward *Sitophilus Oryzae*, we used the data for *Bombyx mori*

**FIGURE 5**

Demonstration of ELs's role in the mechanism of TA binding obtained from FM simulation. **(A)** Transformation of the protein from the open to the closed state, involving EL1, EL2, and EL3. The corresponding movement of each EL is depicted as red arrows **(B)** The effective hydrogen bond between residues Glu187 and Gln 99 and also between Leu190 and Ser413, holds EL2 and EL3 together in the closed state.

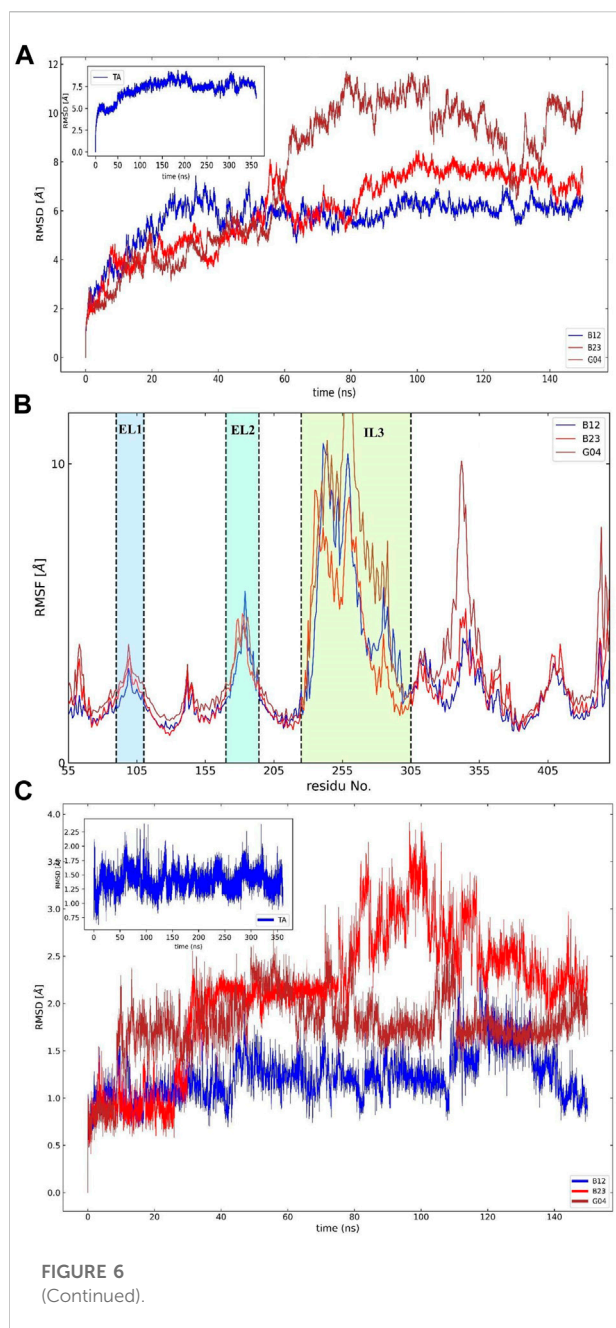
TABLE 2 Binding free energy and its components obtained by MM-PB/GBSA calculation for all ligands.

		 B12	 B23	 G04
MM/PBSA	ΔE_{VDW}	-24.42	-19.48	-21.66
	ΔE_{ele}	-0.64	-1.59	-0.09
	ΔE_{PB}	3.96	4.02	2.55
	ΔE_{NP}	-21.26	-18.92	-18.44
	ΔG_{solv}	20.05	16.89	17.01
	ΔG_{gas}	-25.06	-21.07	-21.76
	ΔG_{Bind}	-5.01 ± 3.2	-4.17 ± 2.35	-4.74 ± 2.92
MM/GBSA	ΔE_{VDW}	-24.42	-19.48	-21.66
	ΔE_{elec}	-2.56	-6.37	-0.38
	ΔE_{GB}	10.62	14.59	8.88
	ΔE_{Surf}	-3.77	-3.20	-3.29
	ΔG_{solv}	6.84	11.39	5.58
	ΔG_{gas}	-26.99	-25.85	22.04
	ΔG_{Bind}	-20.1 ± 3.5	-14.45 ± 3.83	-16.46 ± 2.97

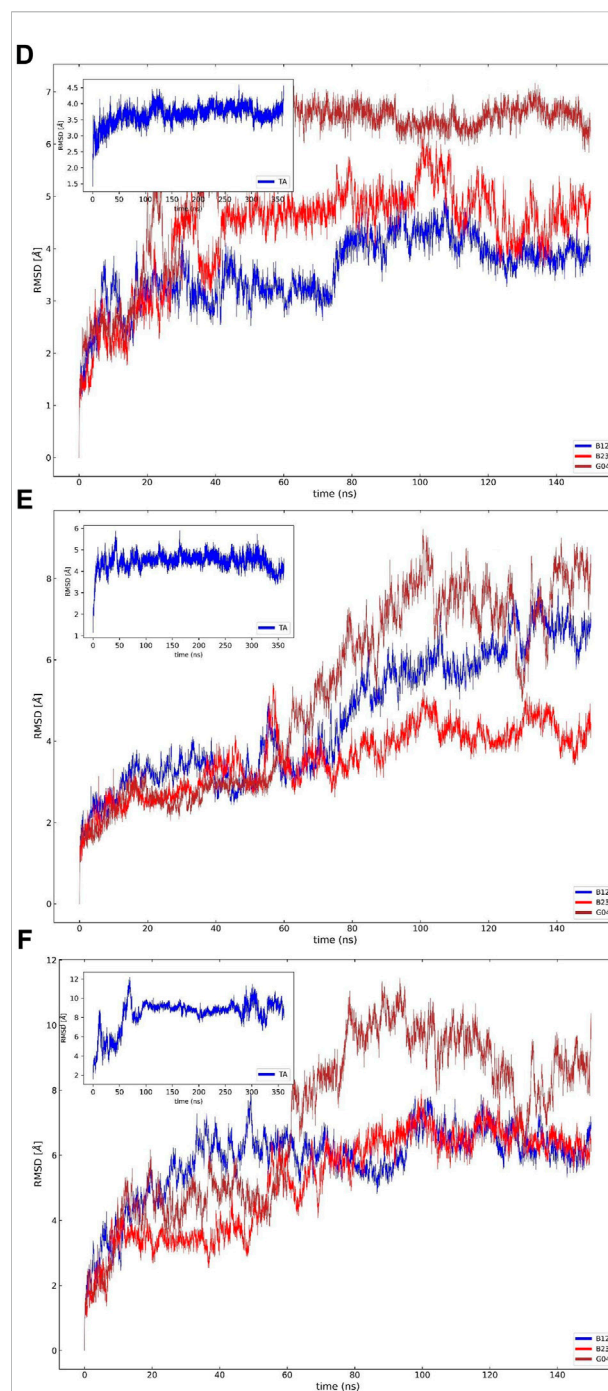
to compare our results to the available experimental data (Ohta et al., 2004). In addition, to provide more structural details on the TA binding, using a reweighting algorithm (Tiwary and Parrinello, 2015), the FES is remapped as a function of the position along the funnel line and distance from the funnel line, producing the FES from the WT-MetaD trajectory above. The WT-MetaD simulations' binding mechanism is validated mainly through the consistency of the minima found on the two FES (Supplementary Figure S5).

3.3 Role of extracellular loops in the binding of TA

To understand the physiological action of the TyrR receptor, it is pivotal to characterize the molecular mechanism of TA recognized by the TyrR receptor. The recognition mechanism of peptide and nonpeptide ligands by G protein-coupled receptors (GPCRs) has a different type where peptide ligands prefer to interact primarily with amino



acid residues in the extracellular loops (ELs), but nonpeptide ligands such as TA interact predominantly with binding site cleft among the TMs (Baldwin, 1993). Here, we discuss the possible involvement of the ELs in the binding mechanism of TA to the TyrR receptor. With a visual inspection, obtained from FM simulation, we observed that the ligand induces conformational changes in ELs at the early stage of approaching the binding site cleft among the TMs. In addition, Figure 5A illustrates the conformational changes



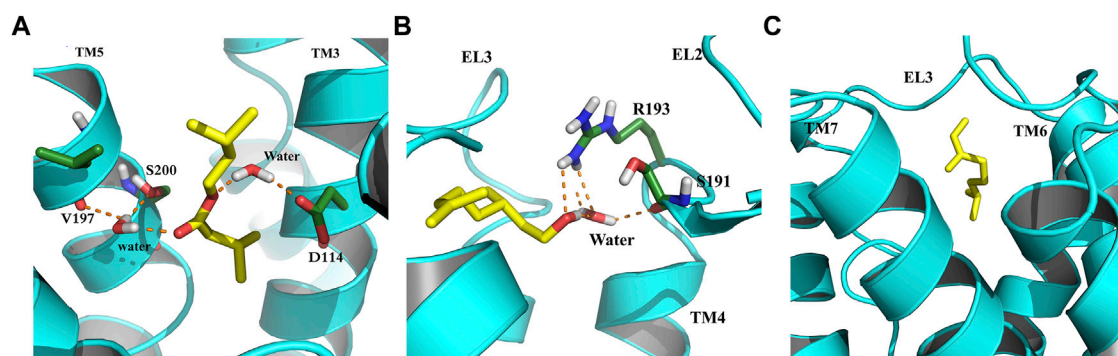


FIGURE 7

Binding modes of selected EO ligands include (A) B12, (B) B23, and (C) G04 bound to the TyrR after 150 ns of CMD. The protein is represented in cyan, the ligand is yellow, and the hydrogen bonds are represented as orange dashed lines.

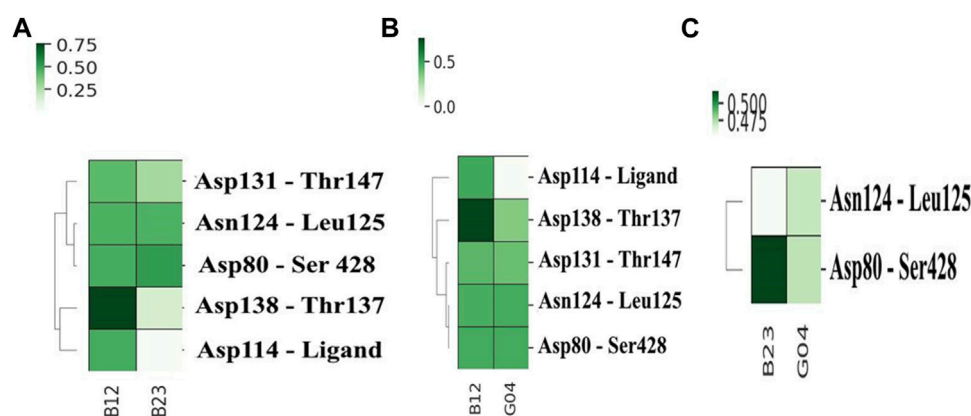


FIGURE 8

Heat map contacts comparing hydrogen bond formation frequencies in (A) B12 and B23, (B) B12 and G04, and (C) B23 and G04 in the active site of the protein. The related HBN is also shown in each plot.

in ELs that occur during TA recognition. A moment after entry of ligand to the binding site cleft among TMs, the EL2 and EL3 start to move inside toward the perpendicular axis of the protein. Meanwhile, the EL1 moves outside toward the vertical axis of the protein; these movements change the conformation of the protein from “open state” to the “closed state,” which curbs the ligand from going back again outside of the protein channel (Figure 5). In the conformation of the protein, changing from an open to a closed state in the TA binding process, we observed that two couples of residues were involved in a strong hydrogen bond to make this conformational change happen. We also showed that the interaction between Glu187 and Gln 99 side chains on the one hand and the hydrogen bond between Leu190 and Ser41, on the other hand, are responsible for keeping EL2 and EL3 close to each other, forming the closed state (Figure 5).

3.4 Screening of the EO's components of *E. camaldulensis*

In the first step of discovering the affinity of essential oil's components against TyrR and discriminating the effectiveness of compounds in flower and leaf, it is of interest to find the possible binding modes of small molecules in the active site of the protein. The docking was performed as described in the material and method section. To screen all compounds, including 54 in leaf and 31 in flower. The results of the docking are represented in [Supplementary Table S1](#). To make docking results more reliable, it is necessary to evaluate the chosen program in terms of its reproducibility of native ligand (TA) in the TyrR, which is supposed to be the binding mode obtained from FM simulation. Redocking results of TA in the protein have been shown in [Supplementary Figure S1A](#). As shown in the related figure, there is a good match between the LeDock result and

the FM binding mode. Therefore, after ensuring the performance of the program, all of the extracted structures, as well as TA, were docked on the active site of the TyrR, the results of which are given in [Supplementary Table S1](#). The two compounds from the leaf and the one from the flower with a high docking score were chosen for further analysis. [Figure 5](#) shows the most effective compounds in the EOs.

3.5 Molecular affinity of EO's components toward TyrR

In this study, three ligand–protein structures have been selected to perform 150 ns of MD simulation, and 1,500 snapshots from 15 replicas have been taken from the MD trajectories to calculate the MM-GB/PBSA binding energies. This may guarantee the accuracy of binding energy obtained from these methods ([Sadiq et al., 2010](#)). For this set of ligands, the standard error of the mean provided in this table is expected to be around 1 kcal/mol on average. [Table 2](#) shows the MM-PB/GBSA binding energies for three ligands. The ranks for the abovementioned ligands are demonstrated by the relevant PB/GB binding energies.

It is of great interest to rank EO's selected compounds in terms of their binding energy toward the TyrR. According to [Table 2](#), B12 is the most effective compound in the inhibition of the protein. However, it should be noted that the binding energy results are very close to each other, and this indicates that to obtain a more accurate result, further analysis such as decomposition analysis should be performed along with the RMSD, RMSF, and binding modes from cluster analysis and hydrogen bond (H-bond) frequency plots for all three compounds. This result led us to further investigate how B12 could be a viable candidate for inhibiting the protein.

3.6 Conformational analysis of the TyrR-EO systems

3.6.1 RMSD analysis

To assess the effective compound in the *Eucalyptus Camaldulensis* EO, we need to inspect the conformational changes of receptors bound to each compound during the MD course and compare them to the conformational pattern we observed from the TA dynamic during FM simulation. The first frame, as a reference conformation, has been used to measure structural changes based on the root mean square deviation (RMSD). We focused on the central regions in the protein whose conformational changes have a significant impact on the function of the protein, that is, transmembrane (TM) helices and intra-/extracellular (IL/EL) loops. [Figure 6A](#) showed the overall RMSD of the protein bound to the EO compounds and TA as a subplot for visual comparison in which, considering the first 150 ns of FM simulation, the B12 and B23 compounds from the flower showed a similar RMSD pattern to TA. We also found that the TMs have negligible contributions to the overall RMSD of the protein due to restricted movement in the

membrane bilayer (~ 3 Å). Therefore, we concentrated on EL and IL motion to compare the movements of the loops when the compounds B12, B23, and G04 bound to the receptor with those we observed for TA in FM simulation. [Figures 6C,D](#) show the RMSD for backbone atoms of EL1 and EL2 loops, respectively. As can be seen, the average RMSD of backbone atoms in EL1 and EL2 in the binding/unbinding process of TA is ~ 1.5 Å and 3.5 Å, respectively. Among the selected compounds from EO, only B12 shows the same pattern when it binds to the receptor in terms of EL1 and EL2 movements.

The long intracellular loop 3 (IL3) is a 150-amino-acid loop located between the TM5 and TM6 domains. Moreover, research suggested that IL2 and IL3 consist of important interaction areas in GPCRs as well as other cytoplasmic effectors ([Gacasan et al., 2017](#)). The RMSD of IL2 and IL3 is given in [Figures 6E,F](#). As can be seen, the flexibilities of IL2 and IL3 have been affected by each EO compound, but it is B12 that asserts the same signal of movements to the IL2 and 3 loops on its binding state.

3.6.2 RMSF values

The influence of screening ligands on the flexibility of the protein structure was studied using root-mean-square fluctuations (RMSFs). [Figure 6B](#) shows that three regions, that is, EL1, EL2, and IL3, fluctuate the most in the presence of B12, B23, and G04 compounds. B12 and B23 show the nearly same pattern of fluctuation in all regions except IL3.

3.6.3 Clustering analysis

In an attempt to elucidate the binding mode of the selected compounds from EOs, the cluster analysis has been done, and the midpoint structure from the most populated cluster has been determined as a representative structure for each ligand–protein complex. [Figure 7](#) shows the representative structures of B12, B23, and G04 bound to TyrR. This can be further evidence for claiming the effectiveness of B12 since, as can be seen in [Figure 7A](#), this ligand is involved in H-bond interactions with Asp114 and Ser200 with water intervention. This is in line with our findings of TA binding mode from FM simulation. The proposed binding modes for B23 and G04 are thoroughly different, indicating different mechanisms of action for these ligands. As it is illustrated in [Figure 7B](#), the B23 involved residues including Arg193 and Ser193 in the EL2 loop. We previously discussed the crucial role of the EL2 loop in the binding of TA, and adaptation of such a binding mode by this ligand could affect the binding mechanism of TA in insects. The same scenario can be imagined for G04 where these ligands also intend to occupy a cleft under the EL3 loop and between TM6 and TM7, as shown in [Figure 7C](#).

3.6.4 Hydrogen bonds analysis

To compare the stability of each selected ligand, it is a prerequisite to evaluate the contacts it made during the MD simulation. The get contacts (<https://getcontacts.github.io/>) have been used to make such a comparison between chosen ligands,

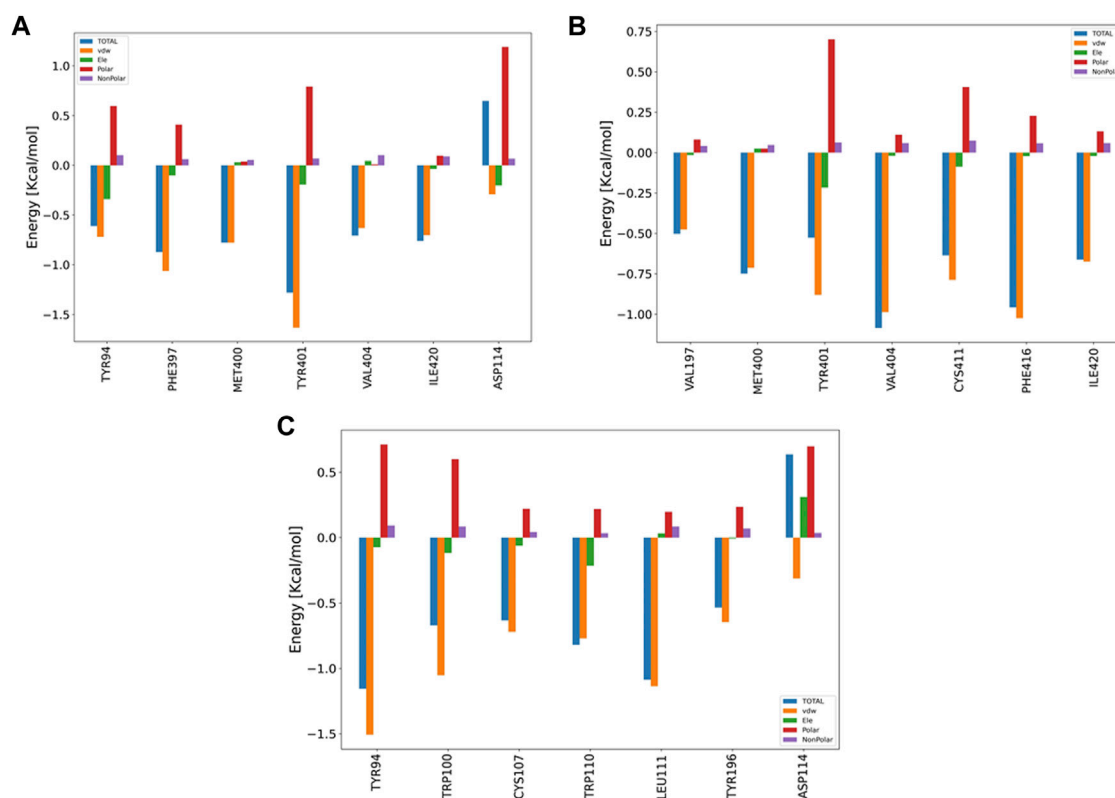


FIGURE 9

Bar plot depiction of energy decomposition as the VDW, electrostatic, polar solvation, and total energies for B12, B23, and G04.

and the results are shown in Figure 8. In this figure, the most frequent hydrogen bonds are calculated for pairwise combinations of each ligand. These plots not only offer the ligand contacts but also give some information about the hydrogen-bonding network (HBN) in the presence of each ligand. Figure 8A depicts the heat map contacts comparing B12 and B23, and we can see that B12 shows more frequent contacts with Asp114 compared to B23. Likewise, this can be seen in Figures 8B,C, where the heat contact maps compare B12–G04 and B23–G04 in the same fashion, and we see the same trend indicating the effectiveness of B12 involving more hydrogen bonds with critical residues such as Asp114. In addition, we can see some shared contacts in these ligand–protein contact maps, such as Asn124 in contact with Leu125 and Asp80 in contact with Ser428. However, it suggests that the communications between these residues are crucial for HBN and the function of the protein in the presence of these ligands.

3.6.5 Decomposition analysis

The pair-wise decomposition analysis can reveal the contribution of energy terms of each residue in the binding energy of the ligand–protein system. Figure 9 illustrates such an analysis for three compounds: B12, B23, and G04. As seen in Figure 9A, we can track down the contribution of Asp114, one of

the most essential residues in the binding of TA stressed by experimental and FM simulation, in B12 and G04's decomposition plots. According to the figure, although the total energy in the decomposition of Asp114 is an adverse effect on the binding energy of B12, the VDW and electrostatic interaction can favor the binding; the necessary information related to the decomposition analysis of B23 and B24 structures are provided in Figures 9B,C, respectively. In the case of G04, as can be seen in Figure 6C, we also observed the contribution of Asp114 in energy binding of this ligand, but lacking a heteroatom in the structure makes it convenient to have merely VDW interaction.

4 Discussion

As noted, previous experimental studies showed that *Eucalyptus* essential oil exhibited the highest toxicity to the rice weevil among the variety of EO treatments, and the results show that *E. camaldulensis* essential oils, rich in insecticidal terpenes, can be alternative candidates to synthetic chemicals in the management of *S. oryzae*. In this regard, the EOs can interfere with neurotransmission by blocking the mechanism

of action of OA/TA, which in insects, causes paralysis and may be followed by death (Jankowska et al., 2018). Therefore, in the current study, first, the TA binding mechanism of action toward TyrR has been investigated as a reference to, second, shed light on the interactions between the EO components of *E. camaldulensis* and Sitophilus oryzae tyramine receptor (SoTyrR) with a view toward a detailed analysis of this insecticidal. For this aim, funnel metadynamics and molecular docking, followed by conventional molecular dynamics (CMD) simulation of the ligand–protein complexes, were employed.

In this study, after extracting the EOs from leaves and flowers of *Eucalyptus camaldulensis* using the GC/MS technique, we performed relevant analysis related to the experimental phase. The GC/MS analysis revealed a total of 54/31 constituents for leave/flower chemical composition of *E. camaldulensis* oil, in which most of the components (29) are common. Among the total components, eucalyptol and α -pinene for both chemical groups were the major constituents. Following the experimental phase, computational studies were further investigated. At first, after performing the homology modeling and determining the protein 3D structure with the least error and the most accuracy, the molecular docking method was used to select the appropriate compounds. The docking results show that butanoic acid, 3-methyl-, 3-methylbutyl ester, 2-Octen-1-ol, 3, 7-dimethyl-, citronellol, and trans- β -Ocimene with the maximum dock score are the main candidates to replace instead of tyramine in TA receptor. Free energy methods, which play a pivotal role in drug design research, use two main approaches to calculate free energy. One is to calculate the bound and unbound states separately, in approaches such as the MM/PBSA, and the other is to evaluate the free energy difference between bound and unbound states, which we can term absolute binding free energy. The latter can be executed in two aspects: by decoupling the interactions between the ligand and its receptor, by giving a nonphysical pathway, and by displacing the ligand along a physical pathway of binding. The immediate output of a binding-pathway free energy method is not a free energy difference but a potential of mean force (PMF), which is defined as the negative logarithm of the probability of being at a given value of a specified reaction coordinate (Eqs 1, 2). Funnel metadynamics (FM) is a kind of binding-pathway free energy that calculated the PMF alongside the funnel-shaped pathway.

Therefore, in the current research, using the FM simulation method with high sensitivity and in 360 ns, the mechanism of action and the binding mode of the reference ligand, TA, have been performed. The FM results suggested that Asp114 followed by Asp80, Asn91, and Asn427 are crucial residues in the binding and the function of TA toward TyrR in Sitophilus Oryzae. Finally, in order to explore the effective compounds in EOs, the binding free energies of the selected ligands were investigated from 150 ns of CMD. The two compounds of the leaf (B12 and

B23) and the one structure (G04) from a flower with high potential inhibition of the TyrR were chosen for MD analysis. The shreds of evidence from the RMSD, RMSF, hydrogen bonding, clustering, and decomposition analysis indicate that the B12 structure has a higher ability to intervene in the biological function of the TA in the insect.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

FZ: literature search and data analysis, writing—original draft, visualization, computational modeling, funnel metadynamics, conceptualization, and final editing. ZN: literature search and data analysis, writing—original draft, visualization, computational modeling, conceptualization, supervision, and final editing. EN: literature search and data analysis, writing—original draft, experimental phase, and final editing. MG: literature search and writing—original draft. AN: writing original draft. AA: writing original draft.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fchem.2022.964700/full#supplementary-material>

References

- Ahmed, M. (2001). *Food irradiation principles and applications*. Disinfestation of stored grains, pulses, dried fruits and nuts, and other dried foods. New York: Wiley, 77–112.
- Baldwin, J. M. (1993). The probable arrangement of the helices in G protein-coupled receptors. *EMBO J.* 12, 1693–1703. doi:10.1002/j.1460-2075.1993.tb05814.x
- Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* 100, 020603. doi:10.1103/physrevlett.100.020603
- Batish, D. R., Singh, H. P., Kohli, R. K., and Kaur, S. (2008). Eucalyptus essential oil as a natural pesticide. *For. Ecol. Manag.* 256, 2166–2174. doi:10.1016/j.foreco.2008.08.008
- B. Gacasan, S. B., Baker, D. L., L. Baker, A. L., and L. Parrill, A. (2017). G protein-coupled receptors: The evolution of structural insight. *AIMS Biophys.* 4, 491–527. doi:10.3934/biophys.2017.3.491
- Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., et al. (2009). Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* 180, 1961–1972. doi:10.1016/j.cpc.2009.05.011
- Braza, M. K. E., Gazmen, J. D. N., Yu, E. T., and Nellas, R. B. (2019). Ligand-induced conformational dynamics of A tyramine receptor from *Sitophilus oryzae*. *Sci. Rep.* 9, 16275. doi:10.1038/s41598-019-52478-x
- Brooker, M. I. H., and D.A., Kleinig, (1990). *Field Guide to Eucalypts*. Vol 2. South-Western and Southern Australia. Inkata Press., Melbourne, 1990.
- Broomhead, N. K., and Soliman, M. E. (2017). Can we rely on computational predictions to correctly identify ligand binding sites on novel protein drug targets? Assessment of binding site prediction methods and a protocol for validation of predicted binding sites. *Cell biochem. Biophys.* 75, 15–23. doi:10.1007/s12013-016-0769-y
- Case, D., Ben-Shalom, I., Brozell, S., Cerutti, D., Cheatham, T., III, Cruzeiro, V., et al. (2018)., 2018. San Francisco: AMBER.
- Cavasotto, C. N., and Phatak, S. S. (2009). Homology modeling in drug discovery: Current trends and applications. *Drug Discov. today* 14, 676–683. doi:10.1016/j.drudis.2009.04.006
- Chang, Y., Lee, S. H., Na, J. H., Chang, P. S., and Han, J. (2017). Protection of grain products from *sitophilus oryzae* (L.) contamination by anti-insect pest repellent sachet containing allyl mercaptan microcapsule. *J. food Sci.* 82, 2634–2642. doi:10.1111/1750-3841.13931
- Comitani, F., Limongelli, V., and Molteni, C. (2016). The free energy landscape of GABA binding to a pentameric ligand-gated ion channel and its disruption by mutations. *J. Chem. Theory Comput.* 12, 3398–3406. doi:10.1021/acs.jctc.6b00303
- Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: AnN-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397
- Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. (1999). Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed.* 38, 236–240. doi:10.1002/(sici)1521-3773(19990115)38:1/2<236::aid-anie236>3.0.co;2-m
- De Paris, R., Quevedo, C. V., Ruiz, D. D., Norberto de Souza, O., and Barros, R. C. (2015). Clustering molecular dynamics trajectories for optimizing docking experiments. *Comput. Intell. Neurosci.* 2015. doi:10.1155/2015/916240
- Degen, J., Gewecke, M., and Roeder, T. (2000). Octopamine receptors in the honey bee and locust nervous system: Pharmacological similarities between homologous receptors of distantly related species. *Br. J. Pharmacol.* 130, 587–594. doi:10.1038/sj.bjp.0703338
- Dhakad, A. K., Pandey, V. V., Beg, S., Rawat, J. M., and Singh, A. (2018). Biological, medicinal and toxicological significance of *Eucalyptus* leaf essential oil: A review. *J. Sci. Food Agric.* 98, 833–848. doi:10.1002/jsfa.8600
- Dhifi, W., Bellili, S., Jazi, S., Bahloul, N., and Mnif, W. (2016). Essential oils' chemical characterization and investigation of some biological activities: A critical review. *Medicines* 3, 25. doi:10.3390/medicines3040025
- Ebadollahi, A., and Setzer, W. N. (2020). Analysis of the essential oils of *Eucalyptus camaldulensis* dehn. And *E. Viminalis* labill. As a contribution to fortify their insecticidal application. *Nat. Product. Commun.* 15, 1934578X2094624–10. doi:10.1177/1934578X20946248
- Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A., and Brooks, C. L., III (2004). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* 25, 265–284. doi:10.1002/jcc.10378
- Genheden, S., and Ryde, U. (2012). Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies. *Proteins* 80, 1326–1342. doi:10.1002/prot.24029
- Goga, N., Rzeplia, A., de Vries, A., Marrink, S., and Berendsen, H. (2012). Efficient algorithms for Langevin and DPD dynamics. *J. Chem. Theory Comput.* 8, 3637–3649. doi:10.1021/ct3000876
- Hess, B., Bekker, H., Berendsen, H. J., and Fraaije, J. G. (1997). Lincs: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463–1472. doi:10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h
- Hočevar, T., and Demčar, J. (2016). Computation of graphlet orbits for nodes and edges in sparse graphs. *J. Stat. Soft.* 71, 1–24. doi:10.18637/jss.v071.i10
- Hockney, R., and Eastwood, J. (1988). *Computer simulation using particles*. Boca Raton: CRC Press.
- Hossain, F., Follett, P., Salmieri, S., Vu, K. D., Jamshidian, M., and Lacroix, M. (2017). Perspectives on essential oil-loaded nanodelivery packaging technology for Controlling stored cereal and grain pests. Boca Raton: CRC Press, 487–508. doi:10.1201/9781315153131-26
- Hossain, F., Follett, P., Salmieri, S., Vu, K. D., Harich, M., and Lacroix, M. (2019). Synergistic effects of nanocomposite films containing essential oil nanoemulsions in combination with ionizing radiation for control of rice weevil *Sitophilus oryzae* in stored grains. *J. food Sci.* 84, 1439–1446. doi:10.1111/1750-3841.14603
- Hossain, F., Lacroix, M., Salmieri, S., Vu, K., and Follett, P. A. (2014). Basil oil fumigation increases radiation sensitivity in adult *sitophilus oryzae* (Coleoptera: Curculionidae). *J. Stored Prod. Res.* 59, 108–112. doi:10.1016/j.jspr.2014.06.003
- Hsiao, Y.-W., and Söderhjelm, P. (2014). Prediction of SAMPL4 host-guest binding affinities using funnel metadynamics. *J. Comput. Aided. Mol. Des.* 28, 443–454. doi:10.1007/s10822-014-9724-4
- Hu, X., and Shelper, W. H. (2003). Docking studies of matrix metalloproteinase inhibitors: Zinc parameter optimization to improve the binding free energy prediction. *J. Mol. Graph. Model.* 22, 115–126. doi:10.1016/s1093-3263(03)00153-0
- Hymavathi, A., Devanand, P., Suresh Babu, K., Sreelatha, T., Pathipati, U. R., and Madhusudana Rao, J. (2011). Vapor-phase toxicity of *Derris scandens* Benth.-derived constituents against four stored-product pests. *J. Agric. Food Chem.* 59, 1653–1657. doi:10.1021/jf104411h
- Jankowska, M., Rogalska, J., Wyszowska, J., and Stankiewicz, M. (2018). Molecular targets for components of essential oils in the insect nervous system-A review. *Molecules* 23, 34. doi:10.3390/molecules23010034
- Jo, S., Kim, T., and Im, W. (2007). Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS one* 2, e880. doi:10.1371/journal.pone.0000880
- Kiani, Y. S., Ranaghan, K. E., Jabeen, I., and Mulholland, A. J. (2019). Molecular dynamics simulation framework to probe the binding hypothesis of CYP3A4 inhibitors. *Ijms* 20, 4468. doi:10.3390/ijms20184468
- Kinoshita, M., and Okada, T. (2015). Structural conservation among the rhodopsin-like and other G protein-coupled receptors. *Sci. Rep.* 5, 9176–9179. doi:10.1038/srep09176
- Kostyukovsky, M., Rafaei, A., Gileadi, C., Demchenko, N., and Shaaya, E. (2002). Activation of octopaminergic receptors by essential oil constituents isolated from aromatic plants: Possible mode of action against insect pests. *Pest. Manag. Sci.* 58, 1101–1106. doi:10.1002/ps.548
- Kumar, D., and Kalita, P. (2017). Reducing postharvest losses during storage of grain crops to strengthen food security in developing countries. *Foods* 6, 8. doi:10.3390/foods6010008
- Laio, A., and Parrinello, M. (2002). Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* 99, 12562–12566. doi:10.1073/pnas.202427399
- Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* 8, 477–486. doi:10.1007/bf00228148
- Limongelli, V., Bonomi, M., and Parrinello, M. (2013). Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6358–6363. doi:10.1073/pnas.1303186110
- Lüthy, R., Bowie, J. U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* 356, 83–85. doi:10.1038/356083a0
- Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., Simmerling, C. J. J. o. c. t., et al. (2015). ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255
- Marzoug, H. N. B., Romdhane, M., Lebrihi, A., Mathieu, F., Couderc, F., Abderraba, M., et al. (2011). *Eucalyptus oleosa* essential oils: Chemical composition and antimicrobial and antioxidant activities of the oils from

different plant parts (stems, leaves, flowers and fruits). *Molecules* 16, 1695–1709. doi:10.3390/molecules16021695

Masson, F., Moné, Y., Vigneron, A., Vallier, A., Parisot, N., Vincent-Monégat, C., et al. (2015). Weevil endosymbiont dynamics is associated with a clamping of immunity. *BMC genomics* 16 (1), 1–13. doi:10.1186/s12864-015-2048-5

Miller, B. R., III, McGee, T. D., Jr, Swails, J. M., Homeyer, N., Gohlke, H., and Roitberg, A. E. (2012). MMPBSA.py: An efficient program for end-state free energy calculations. *J. Chem. Theory Comput.* 8, 3314–3321. doi:10.1021/ct300418h

Mirzadegan, T., Benkö, G., Filipek, S., and Palczewski, K. (2003). Sequence analyses of G-protein-coupled receptors: Similarities to rhodopsin. *Biochemistry* 42, 4310. doi:10.1021/bi033002f

Mondal, E., Majumdar, S., and Chakraborty, K. (2016). Report on sitophilus oryzae as a carrier of fungal pathogen of rice grain with a note on the nature of grain damage at upper Gangetic plains of West Bengal. *World J. Pharm. Med. Res.* 2, 139–145.

Nayak, M. K., Collins, P. J., and Pavic, H. (2007). Developing fumigation protocols to manage strongly phosphine-resistant rice weevils, *Sitophilus oryzae* (L.). *Proceedings of the international conference of controlled atmosphere and fumigation in stored products gold-Coast Australia*, 267–273.

Neethirajan, S., Karunakaran, C., Jayas, D., and White, N. (2007). Detection techniques for stored-product insects in grain. *Food control* 18, 157–162. doi:10.1016/j.foodcont.2005.09.008

Ocampo, A. B., Braza, M. K. E., and Nellas, R. B. (2020). The interaction and mechanism of monoterpenes with tyramine receptor (SoTyrR) of rice weevil (*Sitophilus oryzae*). *SN Appl. Sci.* 2, 1592. doi:10.1007/s42452-020-03395-6

Ohta, H., and Ozoe, Y. (2014). Molecular signalling, pharmacology, and physiology of octopamine and tyramine receptors as potential insect pest control targets. *Adv. Insect Physiology* 46, 73–166. doi:10.1016/b978-0-12-417010-0.00002-1

Ohta, H., Utsumi, T., and Ozoe, Y. (2004). Amino acid residues involved in interaction with tyramine in the *Bombyx mori* tyramine receptor. *Insect Mol. Biol.* 13, 531–538. doi:10.1111/j.0962-1075.2004.00511.x

Parrinello, M., and Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* 52, 7182–7190. doi:10.1063/1.328693

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., et al. (2013). Gromacs 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29, 845–854. doi:10.1093/bioinformatics/btt055

Raniolo, S., and Limongelli, V. (2020). Ligand binding free-energy calculations with funnel metadynamics. *Nat. Protoc.* 15, 2837–2866. doi:10.1038/s41596-020-0342-4

S, S., and Prakash, B. (2015). Assessment of toxicity, antifeedant activity, and biochemical responses in stored-grain insects exposed to lethal and sublethal doses of *Gaultheria procumbens* L. essential oil. *J. Agric. Food Chem.* 63, 10518–10524. doi:10.1021/acs.jafc.5b03797

Sadiq, S. K., Wright, D. W., Kenway, O. A., and Coveney, P. V. (2010). Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases. *J. Chem. Inf. Model.* 50, 890–905. doi:10.1021/ci100007w

Saleh, N., Hucke, O., Kramer, G., Schmidt, E., Montel, F., Lipinski, R., et al. (2018). Multiple binding sites contribute to the mechanism of mixed agonistic and

positive allosteric modulators of the cannabinoid CB1 receptor. *Angew. Chem.* 130, 2610–2615. doi:10.1002/ange.201708764

Saleh, N., Ibrahim, P., Saladino, G., Gervasio, F. L., and Clark, T. (2017). An efficient metadynamics-based protocol to model the binding affinity and the transition state ensemble of G-protein-coupled receptor ligands. *J. Chem. Inf. Model.* 57, 1210–1217. doi:10.1021/acs.jcim.6b00772

Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779–815. doi:10.1006/jmbi.1993.1626

Seada, M. A., Arab, R. A., Adel, I., and Seif, A. I. (2016). Bioactivity of essential oils of basil, fennel, and geranium against *Sitophilus oryzae* and *Callosobruchus maculatus*: Evaluation of repellency, progeny production and residual activity. *Egypt. J. Exp. Biol. (Zool.)* 12, 1–12.

Sharifi, S., and Mills, R. B. (1971). Radiographic studies of *Sitophilus zeamais* Mots. in wheat kernels. *J. Stored Prod. Res.* 7, 195–206. doi:10.1016/0022-474x(71)90007-5

Srinivasan, J., Trevathan, M. W., Beroza, P., and Case, D. A. (1999). Application of a pairwise generalized born model to proteins and nucleic acids: Inclusion of salt effects. *Theor. Chem. Accounts Theory, Comput. Model. Theor. Chimica Acta* 101, 426–434. doi:10.1007/s002140050460

Taylor, W. G., Fields, P. G., and Sutherland, D. H. (2007). Fractionation of lentil seeds (*Lens culinaris* Medik.) for insecticidal and flavonol tetraglycoside components. *J. Agric. Food Chem.* 55, 5491–5498. doi:10.1021/jf0705062

Teppen, B. J. (1992). HyperChem, release 2: Molecular modeling for the personal computer. *J. Chem. Inf. Comput. Sci.* 32, 757–759. doi:10.1021/ci00010a025

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids Res.* 25, 4876–4882. doi:10.1093/nar/25.24.4876

Tiway, P., and Parrinello, M. (2015). A time-independent free energy estimator for metadynamics. *J. Phys. Chem. B* 119, 736–742. doi:10.1021/jp504920s

Troussicot, L., Guillièrre, F., Limongelli, V., Walker, O., and Lancelin, J.-M. (2015). Funnel-metadynamics and solution NMR to estimate protein-ligand affinities. *J. Am. Chem. Soc.* 137, 1273–1281. doi:10.1021/ja511336z

Vanqualef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., et al. (2011). R.E.D. Server: A web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic acids Res.* 39, W511–W517. doi:10.1093/nar/gkr288

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J. Comput. Chem.* 25, 1157–1174. doi:10.1002/jcc.20035

Wang, S., Xu, Y., and Yu, X.-W. (2021). Propeptide in *Rhizopus chinensis* lipase: New insights into its mechanism of activity and substrate selectivity by computational design. *J. Agric. Food Chem.* 69, 4263–4275. doi:10.1021/acs.jafc.1c00721

Yuan, X., Raniolo, S., Limongelli, V., and Xu, Y. (2018). The molecular mechanism underlying ligand binding to the membrane-embedded site of a G-protein-coupled receptor. *J. Chem. Theory Comput.* 14, 2761–2770. doi:10.1021/acs.jctc.8b00046

Zhou, Y., Lin, X. W., Zhang, Y. R., Huang, Y. J., Zhang, C. H., Yang, Q., et al. (2013). Identification and biochemical characterization of *Laodelphax striatellus* neutral ceramidase. *Insect Mol. Biol.* 22, 366–375. doi:10.1111/imb.12028



OPEN ACCESS

EDITED BY

Adolfo Poma,
Institute of Fundamental Technological
Research, Polish Academy of Sciences,
Poland

REVIEWED BY

Julija Zavadlav,
Technical University of Munich,
Germany
Selim Sami,
University of California, Berkeley,
United States

*CORRESPONDENCE

Luigi Delle Site,
✉ luigi.dellesite@fu-berlin.de

SPECIALTY SECTION

This article was submitted to Theoretical
and Computational Chemistry,
a section of the journal *Frontiers in
Chemistry*

RECEIVED 17 October 2022

ACCEPTED 23 November 2022

PUBLISHED 15 December 2022

CITATION

Panahian Jand S, Nourbakhsh Z and
Delle Site L (2022), Nuclear quantum
effects in fullerene–fullerene
aggregation in water.
Front. Chem. 10:1072665.
doi: 10.3389/fchem.2022.1072665

COPYRIGHT

© 2022 Panahian Jand, Nourbakhsh and
Delle Site. 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.

Nuclear quantum effects in fullerene–fullerene aggregation in water

Sara Panahian Jand, Zahra Nourbakhsh and Luigi Delle Site*

Institute of Mathematics, Freie Universität Berlin, Berlin, Germany

We studied the effects of the quantum delocalization in space of the hydrogen atoms of water in the aggregation process of two fullerene molecules. We considered a case using a purely repulsive water–fullerene interaction, as such a situation has shown that water-mediated effects play a key role in the aggregation process. This study becomes feasible, at a reduced computational price, by combining the path integral (PI) molecular dynamics (MD) method with a recently developed open-system MD technique. Specifically, only the mandatory solvation shell of the two fullerene molecules was considered at full quantum resolution, while the rest of the system was represented as a mean-field macroscopic reservoir of particles and energy. Our results showed that the quantum nature of the hydrogen atoms leads to a sizable difference in the curve of the free energy of aggregation; that is, that nuclear quantum effects play a relevant role.

KEYWORDS

nuclear quantum effects, path integral molecular dynamics, PMF of aggregation of hydrophobic particles, fullerene, adaptive resolution simulation (AdResS) method

Introduction

The aggregation of large hydrophobic nanoparticles in water is a subject of interest for its technological and environmental relevance. In particular, the C_{60} fullerene, which is produced in a massive manner by, for example, the arc discharge of graphite electrodes (Montellano Lopez et al., 2011), is the most studied hydrophobic nanoparticle in water, both experimentally (Labille et al., 2009; Chae et al., 2010; Ma et al., 2010; Meng et al., 2010; Voronin et al., 2014) and theoretically (Li et al., 2005a; Li et al., 2005b; Maciel et al., 2011; Zangi, 2014; Makarucha et al., 2016). In this context, the potential of mean force as a function of the C_{60} fullerene–fullerene distance (PMF), that is, the ensemble-averaged fullerene–fullerene space-dependent force (Kirkwood, 1935; Darve, 2006), has been studied using several classical MD approaches (Makarucha et al., 2016). The PMF explains, in terms of (free) energy cost, the process of aggregation of the fullerene molecules, that is, how the two solutes reach aggregation by breaking the hydrogen bonding network of water and coming near each other. Simulation results based on classical models showed that aggregation eventually occurs without any significant energy barrier. However, the classical models used in previous work do not explicitly describe any quantum feature of

water and, thus, cannot account for its potential effects on the strength or flexibility of the hydrogen bonds. In this context, the question of interest is whether the use of a quantum molecular model leads to different results compared to a corresponding classical model in the aggregation process. When a long-range interaction between the carbon atoms of the fullerene and the oxygen atoms of water is used to model the system, water-mediated effects are not relevant in PMF determination (Li et al., 2005b); thus, one can conclude that nuclear quantum effects of water are not likely to play a key role. However, when a purely repulsive C-O interaction is used to model the system, the aggregation process is dominated by the water-mediated effects (Li et al., 2005b); therefore, nuclear quantum effects may become relevant. Experimental results promote the hypothesis that water-mediated effects actually regulate the aggregation (Voronin et al., 2014). The present study tested the relevance of the quantum nature of the hydrogen atoms in the C_{60} - C_{60} aggregation process at room conditions by modeling the C-O interaction as a purely repulsive interaction. This study applied the PIMD technique within the Adaptive Resolution approach (AdResS) (Praprotnik et al., 2005; Praprotnik et al., 2008; Wang et al., 2013; Agarwal et al., 2015; Delle Site and Praprotnik, 2017; Delle Site et al., 2019; Cortes-Huerto et al., 2021). The AdResS technique reduces simulation costs by requiring high (quantum) resolution only in the mandatory solvation region, while the rest of the system is treated at a lower resolution and a small computational cost. The size of the high-resolution region can be automatically and precisely defined by the AdResS method (Lambeth et al., 2010). Our results showed that, at the qualitative level, the PMF calculated with the quantum model did not differ from the PMF calculated with the various classical models; however, a one-to-one quantitative comparison with the TIP4P rigid model; i.e., the closest classical model to our quantum model, showed a sizable difference. Specifically, the depth of the minimum of the PMF curve differed such that one could see the classical model building a strong rigid cage around the aggregated fullerene molecules (deeper minimum), while in the quantum case, the H-bonding network was more flexible and easier to break (less deep minimum). These interesting results add to the methodological message of the paper demonstrating the utility of the open system MD approach to make possible tests of this kind with feasible computational resources. This report is organized as follows: we first provide a brief but essential review of the PIMD idea/technique, followed by the essential description of the AdResS/open system approach and its features. Although this method was previously validated for the quantum water model used here, we further validate the method by studying the solvation of a single fullerene in water and compare the results with simulations of reference. As anticipated, the case of a single fullerene also allowed the precise determination of the

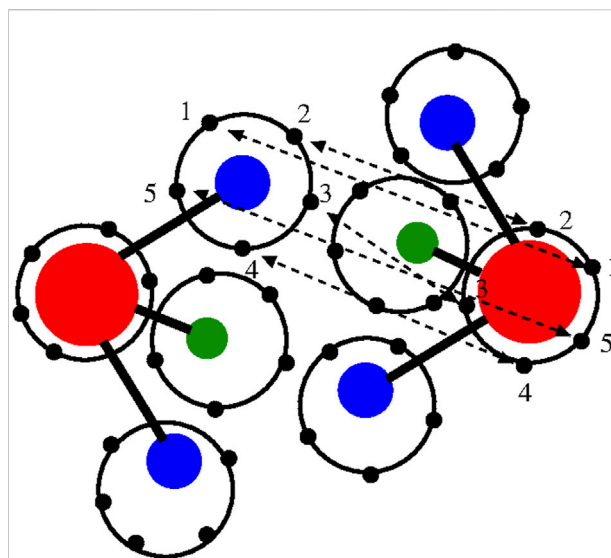


FIGURE 1

Graphical illustration of the path integral/polymer ring representation of two interacting water molecules of the TIP4P 4-site model used in this work (Habershon et al., 2009). Oxygen (red), hydrogen (blue), and additional site model (green). Each site is represented by a polymer ring; for graphical convenience, only five beads per atom/site are drawn although 30 beads per atom/site are used in the real simulation. Atoms of different molecules interact through bead-bead interactions. The beads involved in the interatomic/inter-site interactions are only the beads with the same label (here represented as 1, 2, 3, 4, and 5) of each atom/site. For simplicity, the oxygen-hydrogen interaction is illustrated. The interaction potential has a classical form as the potentials used in the atomistic simulation; however, in this case, the bead-bead interaction is scaled by the number of beads.

minimal solvation region of the two fullerene molecules and, thus, automatically fixed the minimum fullerene-fullerene distance in the PMF calculation. The discussion and conclusions close the paper, while the technical and computational details of the simulations are reported in the **Supplementary Appendix**.

The essentials of path integral molecular dynamics

Light atoms, such as the hydrogen atoms of water, are strongly characterized by quantum effects that lead to their delocalization in space. The path integral technique is a theoretical tool that satisfactorily describes such effects [see e.g., (Feynman and Hibbs, 1965) and references therein]. In particular, a practical method that approaches realistic systems with satisfactory results is the computational technique known as path integral (PI) molecular dynamics (MD) (Tuckerman, 2010; Tuckerman et al., 2014). In essence, one can use a classical potential and delocalize the interatomic interactions by

representing each atom as a polymer ring in which each bead represents an interaction site for the corresponding bead of another atom. The spatial deformation of the ring-polymer during an effectively classical simulation mimics the quantum delocalization of the atom in space (Figure 1); in principle, the larger the number of beads, the more accurate the description of the quantum effect of spatial delocalization.

However, in this representation, each bead counts as a degree of freedom; thus, the cost of simulation, compared to the equivalent classical representation, increases proportionally to the number of beads. This aspect implies a sizable increase in the overall simulation costs compared to classical systems. In general, an atom requires at least 16 beads for a first approximation of a realistic quantum representation. Thus, simulations of a system with 1,000 water molecules represented by a three-site water model with each atom represented by a ring-polymer of 16 beads (thus, 48 degrees of freedom per molecule) become essentially prohibitive, although in practice 30–32 beads are considered the standard for trustworthy simulations (Agarwal and Delle Site, 2015). However, such calculations are expensive and, in particular, for the case of the fullerene–fullerene PMF calculations in the present study, are prohibitive using standard computational resources. Overcoming this challenge requires the use of simulation tools that drastically reduce the mandatory degrees of freedom but provide reliable results. One such method is the recently developed open system MD technique (Delle Site et al., 2019) based on the AdResS technique which has been extensively tested regarding its merging to PIMD (Poma and Delle Site, 2010; Poma and Delle Site, 2011; Potestio and Delle Site, 2012; Agarwal and Delle Site, 2015; Agarwal and Delle Site, 2016; Evangelakis et al., 2021).

The basics of the adaptive resolution technique

AdResS treats an open subregion of the simulation domain at full quantum resolution and the rest as a thermodynamic reservoir of energy and particles, that is, as a large domain of non-interacting particles (tracers) thermalized by an external thermostat [the latest version is described in Delle Site et al. (2019) and Evangelakis et al. (2021)]. Figure 2 illustrates the concept, showing a high-resolution region (PI) embedded in a (usually) much larger region of tracers (TR) thermalized by an external reservoir that assures the correct thermodynamic conditions. Between the high-resolution and tracer regions is the so-called Δ (transition) region in which the molecules are at high resolution and experience the external (one-body) thermodynamic force. This force, together with the action of the thermostat, assures the physically consistent exchange of particles between the high-resolution and tracer regions. In essence, the additional force corrects from any difference in

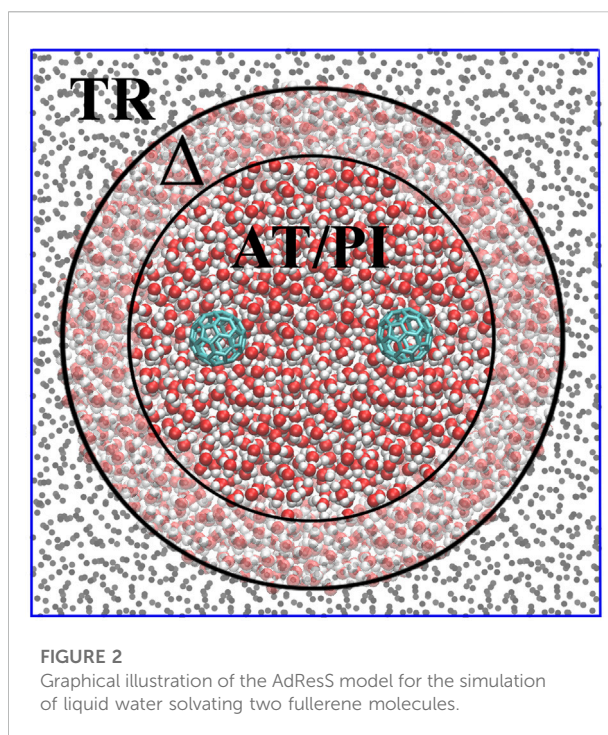


FIGURE 2
Graphical illustration of the AdResS model for the simulation of liquid water solvating two fullerene molecules.

the chemical potential between the different regions and ensures the exchange of particles at the chemical potential of a reference (full high-resolution) system. The calculation of the thermodynamic force is performed self-consistently during the equilibration run of the AdResS system (Poblete et al., 2010; Fritsch et al., 2012; Wang et al., 2013; Agarwal et al., 2014; Gholami et al., 2021a; Gholami et al., 2021b). Tracer particles entering the Δ region acquire the chemical structure of the water molecule and the corresponding path integral resolution; on the contrary, molecules leaving the Δ region for the TR region lose their high resolution and become non-interacting particles. Recent results have demonstrated the reliability of this technique for the four-site water model used here with 30 beads per atom, which means that molecules entering the TR region lose 120 degrees of freedom, while molecules entering the Δ region acquire 120 degrees of freedom (Evangelakis et al., 2021). The size of the Δ region is equal to the cut-off distance of the interaction potential such that there is no missing interaction between molecules in the PI and TR regions. The data on the PI region are used to calculate the properties of the open system, while the Δ region represents a sort of artificial region needed to implement the boundary conditions for the PI region so that molecules entering the PI region are automatically equilibrated with the PI environment at the thermodynamic conditions required by the study. The next section considers the solvation of a single fullerene in water and confirmed the reliability of the technique. We also define the maximal region of interest in the fullerene–fullerene aggregation.

Test of validity of the method: Solvation of a single fullerene in water

To define a physically meaningful open region for the PI resolution region of AdResS, the physical consistency was routinely checked in the AdResS: 1) the water density in the $AT + \Delta$ region should reproduce, within some numerical accuracy, the full reference PI simulation value. The thermodynamic force in Δ ensures that (1) is satisfied. 2) The radial distribution functions should reproduce, within some numerical accuracy, the reference full PI simulation value. These functions represent relevant structural properties that characterize a liquid and its solvation action at certain thermodynamic conditions. In addition, at the statistical mechanics level, their combination expresses the probability distribution function of the system in configuration space up to the two-body approximation (Wang et al., 2013; Agarwal et al., 2015; Evangelakis et al., 2021). 3) The probability distribution function of the particle number in PI, $P(N)$, must be consistent with $P(N)$ of an equivalent subregion in the full reference path integral simulation so that the exchange of particles between the PI region and the reservoir (TR) is physically consistent. The concurrent fulfillment of 1, 2, and 3 assures that the explicit quantum degrees of freedom of the PI region are sufficient to reproduce the key features of solvation, while the explicit quantum degrees of freedom outside this region are not relevant for characterizing its physical property and, thus, can be represented by a generic thermodynamic bath. The size of the PI region automatically defines the minimal extension of the mandatory solvation shell and the maximal fullerene–fullerene distance in the PMF calculation (Delle Site, 2022). The maximum fullerene–fullerene distance of interest in a PMF calculation can be accurately determined by the minimum size of the region around each fullerene. Here, water molecules, with their quantum degrees of freedom, directly influence the behavior of the fullerene; beyond this distance, water acts only as a thermodynamic bath and the corresponding hydrogen bonding structure has no direct effect on the fullerene. Regarding the PMF calculation, if the maximum fullerene–fullerene distance is equal to the sum of the radii of the smallest mandatory solvation shells of the single fullerenes, then automatically for larger distances, the two fullerenes do not experience the perturbation of the hydrogen bonding network caused by the other; thus, distances beyond these maximal values are of no interest in the PMF calculation. Figures 3–5 show the calculation of the water density, the various radial distribution functions, and the $P(N)$ for three different sizes of the PI region. The case of 1.22 nm agrees in a highly satisfactory manner with the results of the reference full path integral simulation; thus, it validates the technique as reliable to simulate a physically consistent open region. Moreover, 1.22 nm represents the mandatory

solvation region and implies that 2.44 nm is the largest fullerene–fullerene distance to be considered in the PMF calculation.

PMF of aggregation of two C_{60} molecules

As discussed previously, for the solvation of two fullerene molecules, the radius of the mandatory solvation shell is twice that of the single fullerene molecule, that is, 2.44 nm. This is also the maximal distance that must be considered for the calculation of the PMF. Figure 6 shows the PMF curve calculated for the quantum model with the PIMD-AdResS simulation, compared to the equivalent classical rigid model. Qualitatively, the aggregation process does not differ in the two cases and the aggregation eventually happens without any significant energy barrier. However, the aggregation in the classical model is energetically more favorable than in the quantum model as the two fullerene molecules approach a closer distance. Once the two fullerene molecules have come in contact, the system falls into a deeper minimum for the classical simulation compared to the quantum case. Thus, the aggregated fullerene molecules are more stable in the classical case compared to that in the quantum case, with a substantial difference in (free) energy of about 7 kcal/mol. At this point, the quantum model is the direct extension of the classical model, that is, its force field is enhanced by the intramolecular flexibility (OH bond stretching and HOH angular potential) together with the ring polymer representation of the atoms. The straightforward implication is that the molecular flexibility and the quantum delocalization of the H atoms can sizably influence the (re)organization hydrogen bonding network. For a purely repulsive C-O potential, as used in this study, the aggregation is driven by water-mediated effects; in other words, by the reorganization of the OH-bonding network as the two fullerenes approach each other. The curves in Figure 6 suggest that the degree of reorganization of the OH-bonding network passing from two cages localized around each fullerene, when the fullerenes are far apart, to a large cage that embeds both, once they aggregate, is higher in the classical case than in the quantum case. This idea was also hypothesized previously (Agarwal et al., 2017). Agarwal et al. (2017) also reported a less structured OH-bonding network in the quantum case compared to the classical case. The authors speculated, based on experimental data, that this result may imply a different characterization of aggregated C_{60} molecules when quantum effects are considered. In that study, calculations of the aggregation process were not yet possible using standard computational resources and were defined as “feasible in the near future.” The current results fill this gap and provide a quantitative argument for their hypothesis. A detailed analysis of the structure and dynamics of the bonding network would require the calculation of time correlation functions to explain in

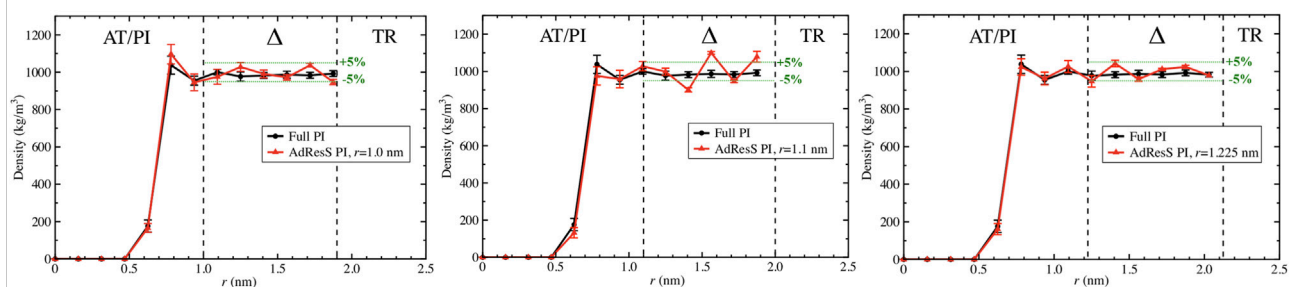


FIGURE 3

Particle number density calculated in the AdResS setup and compared to the density calculated in the reference simulations for three different radii of the PI region, namely, $r = 1\text{ nm}$, $r = 1.1\text{ nm}$, and $r = 1.22\text{ nm}$. All three figures show sufficient agreement with the reference density. For $r = 1\text{ nm}$, despite a satisfactory agreement in the Δ region, the AdResS density close to the fullerene shows a slight disagreement with the reference density. For $r = 1.1\text{ nm}$ in the Δ region, the accuracy of the density with respect to the density of reference is slightly beyond the 5% threshold. $r = 1.22\text{ nm}$ shows satisfactory agreement over the whole range and the accuracy of the density in the Δ region is within 5% compared to the reference value. 5% is usually considered a satisfactory threshold.

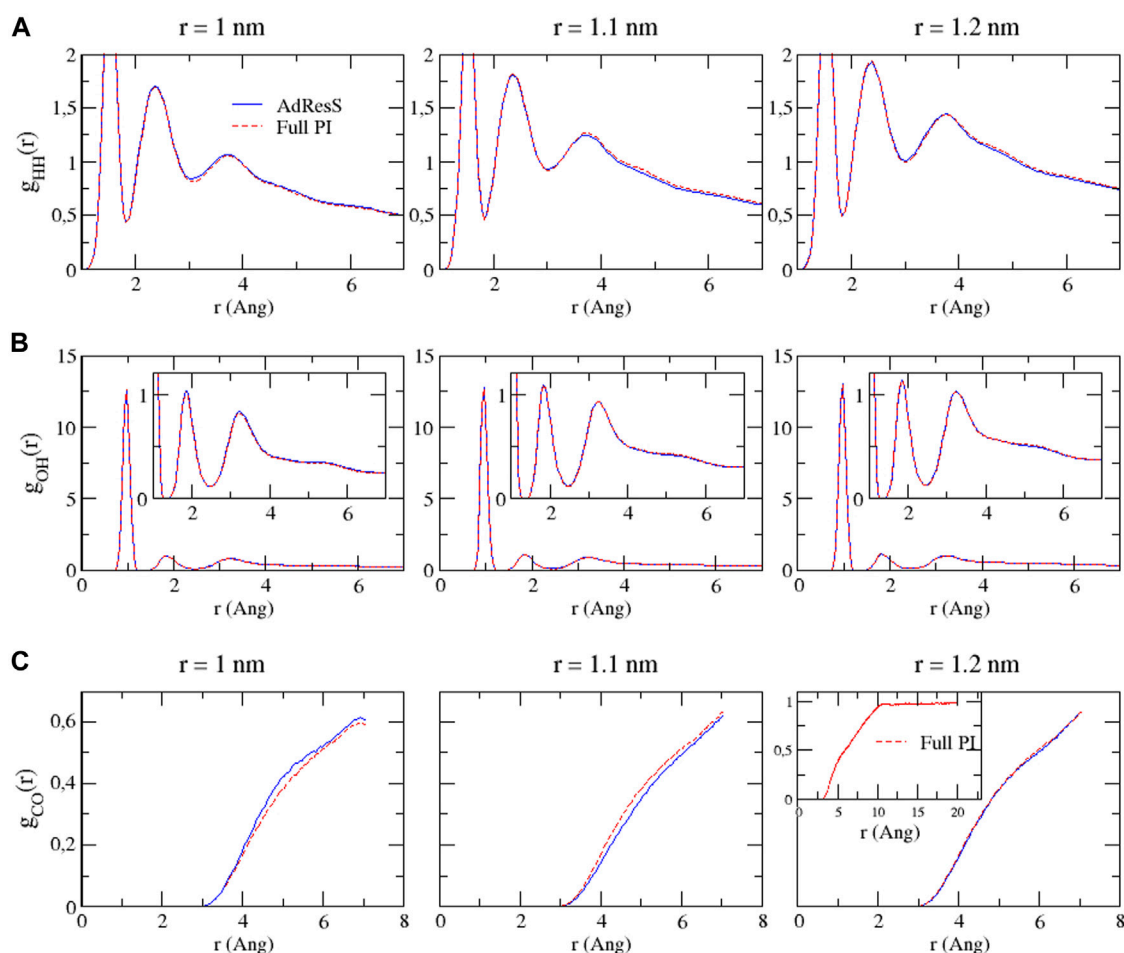


FIGURE 4

Bead-bead radial distribution functions for hydrogen-hydrogen (A), oxygen-hydrogen (B), and carbon-oxygen (C) calculated in the PI region of AdResS and the equivalent subregion of the reference simulation. Since these curves are calculated only in a subregion, they are not normalized.

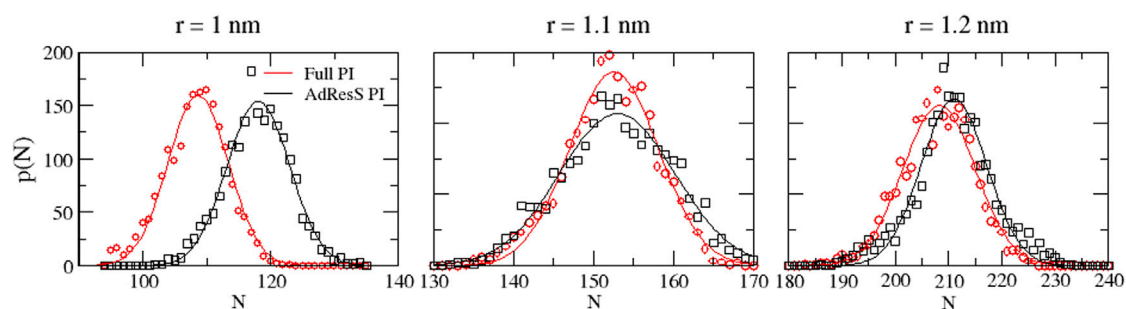


FIGURE 5

Particle number probability distributions calculated in the PI region and the equivalent subregion of the reference simulation.

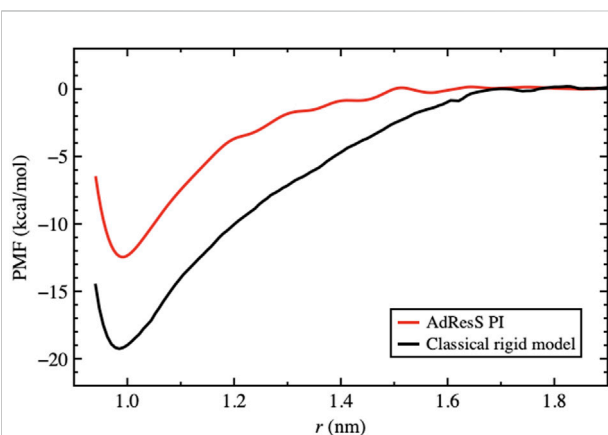


FIGURE 6

PMF for the path integral model using AdResS compared to the reference full atomistic classical simulation. The PMF is calculated as a function of the distance between the centers of mass of the C_{60} molecules. The zero of each curve was chosen to be the corresponding bulk solvation energy, that is, the value of the PMF at the plateau.

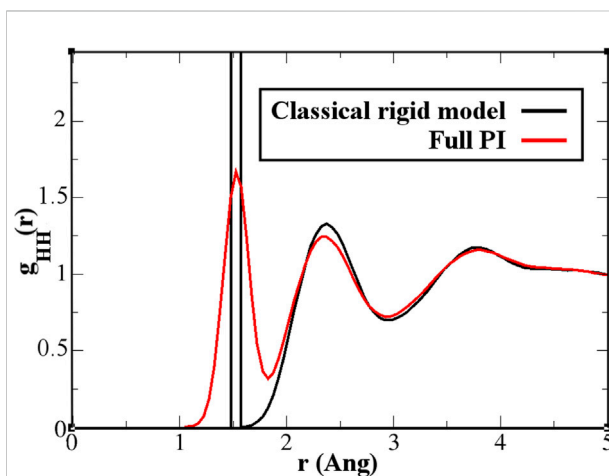


FIGURE 7

Hydrogen-hydrogen radial distribution function for a pure water system. The classical rigid model (black line) has a first sharply localized peak, while the quantum model (red line) spreads the probability over 1 Å. Further effects are visible, although in a light form, also beyond the intramolecular and first neighbor molecule environment.

detail the dynamics of the aggregation. Such a study, which requires much longer trajectories and the careful use of the thermostat only in regions where the dynamics is not investigated, goes beyond the scope of the present study, which aimed to characterize only the static structural properties of aggregation. In this context, the effect of the flexibility of the quantum model becomes evident in the hydrogen-hydrogen radial distribution function (Figure 7). The hydrogen atoms are the true quantum particles of the systems. In their spatial correlation, the quantum delocalization and the induced flexibility of the bonds are clearly expressed. Within the range of 1.0–2.5 Å, the well-structured classical model differs from the quantum model, in which the probability is spread across the whole range. Regarding the technical advantages of the AdResS, the explicit

computational gain is still modest compared to its full potential as the parallelization of the code is not yet optimized.

The straightforward comparison with full path integral simulations currently leads to a factor 3. Although not yet optimal, it is already a non-trivial gain as it reduces the requested computational resources to one-third. This difference becomes significant when a large number of calculations are required, as shown in the present case for the determination of the PMF. The additional advantages of this method include the possibility of determining the maximum distance required in a PMF by reducing the need to sample distances that are not relevant but that cannot be excluded *a priori*. Finally, the drastic reduction in the number of degrees of freedom requires a much lower allocation memory, while full

path integral simulations would require so much memory that would *a priori* prevent groups without significant computational resources from performing such simulations.

Conclusion

We applied the open system MD technique based on the AdResS protocol to study the aggregation of two C_{60} fullerene molecules in water considering quantum nuclear effects. After validating the simulation techniques and the corresponding technical set-up, we determined the PMF as a function of the centers of the mass distances of the two solutes. These calculations were performed for the quantum case and for the classical case where molecules are modeled as rigid objects. Only purely repulsive interactions between water and the C_{60} molecule were considered. In such cases, water-mediated effects have been shown to play a major role. In the case of a potential with an attractive part, this part would play a key role in the aggregation process; thus, the role of the H-bonding network becomes negligible. The difference in the PMF curve of aggregation was qualitatively similar, that is, aggregation occurs without barriers. However, quantitatively, the difference was sizable. This result can be interpreted as the combined effect of the molecular flexibility and the quantum delocalization of H atoms in the reorganization of the H-bonding network in the quantum case. Thus, nuclear quantum effects are very relevant in the aggregation process if a purely repulsive fullerene–water potential is used to model the interaction. From the methodological aspect, the results of this study demonstrated that the open system MD approach can significantly reduce the computational resource requirements, thus permitting studies to be performed that would otherwise be significantly more expensive.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors without undue reservation.

Author contributions

LD conceived the study. SP and ZN performed the calculations and analyzed the results. LD wrote the manuscript with input from all authors. SP and LD discussed

the results and contributed to the final manuscript. All authors read and approved the final manuscript.

Funding

This research was funded by Deutsche Forschungsgemeinschaft (DFG) (grant CRC 1114 “Scaling Cascades in Complex Systems”, Project Number 235221301, Projects C01 “Adaptive coupling of scales in molecular dynamics and beyond to fluid dynamics” to LD and ZN, grants DE 1140/7-3 to LD and SP, and DE 1140/11-1 to LD and SP). The simulations presented here were performed using HPC resources provided by the North-German Supercomputing Alliance (HLRN).

Acknowledgments

The authors would like to acknowledge open access funding provided by the Freie Universität Berlin.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fchem.2022.1072665/full#supplementary-material>

References

- Abascal, J. L., and Vega, C. (2005). A general purpose model for the condensed phases of water: Tip4p/2005. *J. Chem. Phys.* 123, 234505. doi:10.1063/1.2121687
- Abraham, M. J., Murtola, T., Schulz, R., Pall, S., Smith, J., Hess, B., et al. (2015). Gromacs: High performance molecular simulations through multi-level parallelism

from laptops to supercomputers. *SoftwareX* 1–2, 19–25. doi:10.1016/j.softx.2015.06.001

Agarwal, A., Clementi, C., and Site, L. D. (2017). Path integral-GC-AdResS simulation of a large hydrophobic solute in water: A tool to investigate the interplay

- between local microscopic structures and quantum delocalization of atoms in space. *Phys. Chem. Chem. Phys.* 19, 13030–13037. doi:10.1039/c7cp01629h
- Agarwal, A., and Delle Site, L. (2015). Path integral molecular dynamics within the grand canonical-like adaptive resolution technique: Simulation of liquid water. *J. Chem. Phys.* 143, 094102. doi:10.1063/1.4929738
- Agarwal, A., and Delle Site, L. (2016). Grand-canonical adaptive resolution centroid molecular dynamics: Implementation and application. *Comput. Phys. Commun.* 206, 26–34. doi:10.1016/j.cpc.2016.05.001
- Agarwal, A., Wang, H., Schütte, C., and Site, L. D. (2014). Chemical potential of liquids and mixtures via adaptive resolution simulation. *J. Chem. Phys.* 141, 034102. doi:10.1063/1.4886807
- Agarwal, A., Zhu, J., Hartmann, C., Wang, H., and Site, L. D. (2015). Molecular dynamics in a grand ensemble: Bergmann–Lebowitz model and adaptive resolution simulation. *New J. Phys.* 17, 083042. doi:10.1088/1367-2630/17/8/083042
- Chae, S.-R., Badireddy, A. R., Budarz, J. F., Lin, S., Xiao, Y., Therezien, M., et al. (2010). Heterogeneities in fullerene nanoparticle aggregates affecting reactivity, bioactivity, and transport. *ACS Nano* 4, 5011–5018. doi:10.1021/nn100620d
- Cortes-Huerto, R., Kremer, K., Praprotnik, M., and Delle Site, L. (2021). From adaptive resolution to molecular dynamics of open systems. *Eur. Phys. J. B* 94, 189. doi:10.1140/epjb/s10051-021-00193-w
- Darve, E. (2006). *Numerical methods for calculating the potential of mean force*. Berlin Heidelberg: Springer-Verlag.
- Delle Site, L., and Praprotnik, M. (2017). Molecular systems with open boundaries: Theory and simulation. *Phys. Rep.* 693, 1–56. doi:10.1016/j.physrep.2017.05.007
- Delle Site, L., Krekeler, C., Whittaker, J., Agarwal, A., Klein, R., and Höfling, F. (2019). Molecular dynamics of open systems: construction of a mean-field particle reservoir. *Adv. Theory Simul.* 2, 1900014. doi:10.1002/adts.201900014
- Delle Site, L. (2022). Investigation of water-mediated intermolecular interactions with the adaptive resolution simulation technique. *J. Phys. Condens. Matter* 34, 115101. doi:10.1088/1361-648x/ac29e2
- Evangelakis, A., Panahian Jand, S., and Delle Site, L. (2021). Path integral molecular dynamics of liquid water in a mean-field particle reservoir. *ChemistryOpen* 11, e202100286. doi:10.1002/open.202100286
- Feynman, R., and Hibbs, A. (1965). *Quantum mechanics and path integrals*. New York: McGraw-Hill.
- Fritsch, S., Poblete, S., Junghans, C., Ciccotti, G., Delle Site, L., and Kremer, K. (2012). Adaptive resolution molecular dynamics simulation through coupling to an internal particle reservoir. *Phys. Rev. Lett.* 108, 170602. doi:10.1103/physrevlett.108.170602
- Gholami, A., Höfling, F., Klein, R., and Delle Site, L. (2021). Thermodynamic relations at the coupling boundary in adaptive resolution simulations for open systems. *Adv. Theory Simul.* 4, 2000303. doi:10.1002/adts.202000303
- Gholami, A., Klein, R., and Delle Site, L. (2021). On the relation between pressure and coupling potential in adaptive resolution simulations of open systems in contact with a reservoir. *Adv. Theory Simul.* 4, 2100212. doi:10.1002/adts.202100212
- Girifalco, L. A. (1992). Molecular properties of fullerene in the gas and solid phases. *J. Phys. Chem.* 96, 858–861. doi:10.1021/j100181a061
- Habershon, S., Markland, T., and Manolopoulos, D. (2009). Competing quantum effects in the dynamics of a flexible water model. *J. Chem. Phys.* 131, 024501. doi:10.1063/1.3167790
- Kirkwood, J. G. (1935). Statistical mechanics of fluid mixtures. *J. Chem. Phys.* 3, 300–313. doi:10.1063/1.1749657
- Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comput. Chem.* 13, 1011–1021. doi:10.1002/jcc.540130812
- Labille, J., Masion, A., Ziarelli, F., Rose, J., Brant, J., Villieras, F., et al. (2009). Hydration and dispersion of c60 in aqueous systems: The nature of water-fullerene interactions. *Langmuir* 25, 11232–11235. doi:10.1021/la9022807
- Lambeth, B., Junghans, C., Kremer, K., Clementi, C., and Delle Site, L. (2010). Communication: On the locality of Hydrogen bond networks at hydrophobic interfaces. *J. Chem. Phys.* 133, 221101. doi:10.1063/1.3522773
- Li, L., Bedrov, D., and Smith, G. D. (2005). A molecular-dynamics simulation study of solvent-induced repulsion between C60 fullerenes in water. *J. Chem. Phys.* 123, 204504. doi:10.1063/1.2121647
- Li, L., Bedrov, D., and Smith, G. D. (2005). Repulsive solvent-induced interaction between C60 fullerenes in water. *Phys. Rev. E* 71, 011502. doi:10.1103/physreve.71.011502
- Lobaugh, J., and Voth, G. A. (1997). A quantum model for water: Equilibrium and dynamical properties. *J. Chem. Phys.* 106, 2400–2410. doi:10.1063/1.473151
- Ma, X., Wigington, B., and Bouchard, D. (2010). Fullerene c60: Surface energy and interfacial interactions in aqueous systems. *Langmuir* 26, 11886–11893. doi:10.1021/la101109h
- Maciel, C., Fileti, E. E., and Rivelino, R. (2011). Assessing the solvation mechanism of c60(oh)24 in aqueous solution. *Chem. Phys. Lett.* 507, 244–247. doi:10.1016/j.cplett.2011.03.080
- Makarucha, A., Baldauf, J. S., Downton, M. T., and Yiapanis, G. (2016). Size-dependent fullerene–fullerene interactions in water: A molecular dynamics study. *J. Phys. Chem. B* 120, 11018–11025. doi:10.1021/acs.jpcc.6b07471
- Meng, H., Xing, G., Sun, B., Zhao, F., Lei, H., Li, W., et al. (2010). Potent angiogenesis inhibition by the particulate form of fullerene derivatives. *ACS Nano* 4, 2773–2783. doi:10.1021/nn100448z
- Montellano Lopez, A., Mateo-Alonso, A., and Prato, M. (2011). Materials chemistry of fullerene c60 derivatives. *J. Mat. Chem.* 21, 1305–1318. doi:10.1039/c0jm02386h
- Monticelli, L. (2012). On atomistic and coarse-grained models for c60 fullerene. *J. Chem. Theory Comput.* 8, 1370–1378. doi:10.1021/ct3000102
- Poblete, S., Praprotnik, M., Kremer, K., and Delle Site, L. (2010). Coupling different levels of resolution in molecular simulations. *J. Chem. Phys.* 132, 114101. doi:10.1063/1.3357982
- Poma, A., and Delle Site, L. (2010). Classical to path-integral adaptive resolution in molecular simulation: Towards a smooth quantum-classical coupling. *Phys. Rev. Lett.* 104, 250201. doi:10.1103/physrevlett.104.250201
- Poma, A., and Delle Site, L. (2011). Adaptive resolution simulation of liquid parahydrogen: Testing the robustness of the quantum-classical adaptive coupling. *Phys. Chem. Chem. Phys.* 13, 10510. doi:10.1039/c0cp02865g
- Potestio, R., and Delle Site, L. (2012). Quantum locality and equilibrium properties in low-temperature parahydrogen: A multiscale simulation study. *J. Chem. Phys.* 136, 054101. doi:10.1063/1.3678587
- Praprotnik, M., Delle Site, L., and Kremer, K. (2005). Adaptive resolution molecular-dynamics simulation: Changing the degrees of freedom on the fly. *J. Chem. Phys.* 123, 224106. doi:10.1063/1.2132286
- Praprotnik, M., Delle Site, L., and Kremer, K. (2008). Multiscale simulation of soft matter: From scale bridging to adaptive resolution. *Annu. Rev. Phys. Chem.* 59, 545–571. doi:10.1146/annurev.physchem.59.032607.093707
- Tuckerman, M. E., Berne, B. J., Martyna, G. J., and Klein, M. L. (1993). Efficient molecular dynamics and hybrid Monte Carlo algorithms for path integrals. *J. Chem. Phys.* 99, 2796–2808. doi:10.1063/1.465188
- Tuckerman, M. (2014). “Path integration via molecular dynamics,” in *Quantum simulations of Complex many-body systems: From theory to algorithms*. Editors J. Grotenndorff, D. Marx, and A. Muramatsu (India: NIC), 10, 169–189.
- Tuckerman, M. E. (2010). *Statistical mechanics: Theory and molecular simulation*. New York: Oxford University Press.
- Voronin, D., Buchelnikov, A., Kostjukov, V., Khrapatiy, S., Wyrzykowski, D., Poisik, J., et al. (2014). Evidence of entropically driven c60 fullerene aggregation in aqueous solution. *J. Chem. Phys.* 140, 104909. doi:10.1063/1.4867902
- Wang, H., Hartmann, C., Schütte, C., and Delle Site, L. (2013). Grand-canonical-like molecular-dynamics simulations by using an adaptive-resolution technique. *Phys. Rev. X* 3, 011018. doi:10.1103/physrevx.3.011018
- Witt, A., Ivanov, S. D., Shiga, M., Forbert, H., and Marx, D. (2009). On the applicability of centroid and ring polymer path integral molecular dynamics for vibrational spectroscopy. *J. Chem. Phys.* 130, 194510. doi:10.1063/1.3125009
- Zangi, R. (2014). Are buckyballs hydrophobic. *J. Phys. Chem. B* 118, 12263–12270. doi:10.1021/jp508174a



OPEN ACCESS

EDITED BY

Sergio Pantano,
Institut Pasteur de Montevideo, Uruguay

REVIEWED BY

Paolo A. Calligari,
University of Rome Tor Vergata, Italy
Durba Sengupta,
National Chemical Laboratory (CSIR),
India

*CORRESPONDENCE

Paolo Ruggerone,
✉ paolo.ruggerone@dsf.unica.it

SPECIALTY SECTION

This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

RECEIVED 01 October 2022

ACCEPTED 06 December 2022

PUBLISHED 09 January 2023

CITATION

Oliva F, Musiani F, Giorgetti A, De Rubeis S, Sorokina O, Armstrong DJ, Carloni P and Ruggerone P (2023), Modelling eNvironment for Isoforms (MoNvlso): A general platform to predict structural determinants of protein isoforms in genetic diseases. *Front. Chem.* 10:1059593. doi: 10.3389/fchem.2022.1059593

COPYRIGHT

© 2023 Oliva, Musiani, Giorgetti, De Rubeis, Sorokina, Armstrong, Carloni and Ruggerone. 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.

Modelling eNvironment for Isoforms (MoNvlso): A general platform to predict structural determinants of protein isoforms in genetic diseases

Francesco Oliva^{1,2}, Francesco Musiani³, Alejandro Giorgetti^{2,4}, Silvia De Rubeis^{5,6,7,8}, Oksana Sorokina⁹, Douglas J. Armstrong^{2,9,10}, Paolo Carloni^{2,11,12} and Paolo Ruggerone^{1*}

¹Department of Physics, University of Cagliari, Monserrato (CA), Italy, ²Institute of Neuroscience and Medicine INM-9, Institute for Advanced Simulations IAS-5, Forschungszentrum Jülich, Jülich, Germany, ³Laboratory of Bioinorganic Chemistry, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, ⁴Department of Biotechnology, University of Verona, Verona, Italy, ⁵Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY, United States, ⁶Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, United States, ⁷The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, United States, ⁸Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, United States, ⁹The School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, ¹⁰Simons Initiative for the Developing Brain, University of Edinburgh, Edinburgh, United Kingdom, ¹¹Department of Physics, RWTH Aachen University, Aachen, Germany, ¹²JARA-Institute: Molecular Neuroscience and Neuroimaging, Institute for Neuroscience and Medicine INM-11/JARA-BRAIN Institute JBI-2, Forschungszentrum Jülich GmbH, Jülich, Germany

The seamless integration of human disease-related mutation data into protein structures is an essential component of any attempt to correctly assess the impact of the mutation. The key step preliminary to any structural modelling is the identification of the isoforms onto which mutations should be mapped due to there being several functionally different protein isoforms from the same gene. To handle large sets of data coming from omics techniques, this challenging task needs to be automatized. Here we present the MoNvlso (Modelling eNvironment for Isoforms) code, which identifies the most useful isoform for computational modelling, balancing the coverage of mutations of interest and the availability of templates to build a structural model of both the wild-type isoform and the related variants.

KEYWORDS

isoform identification, mutations, molecular modelling, proteins, diseases

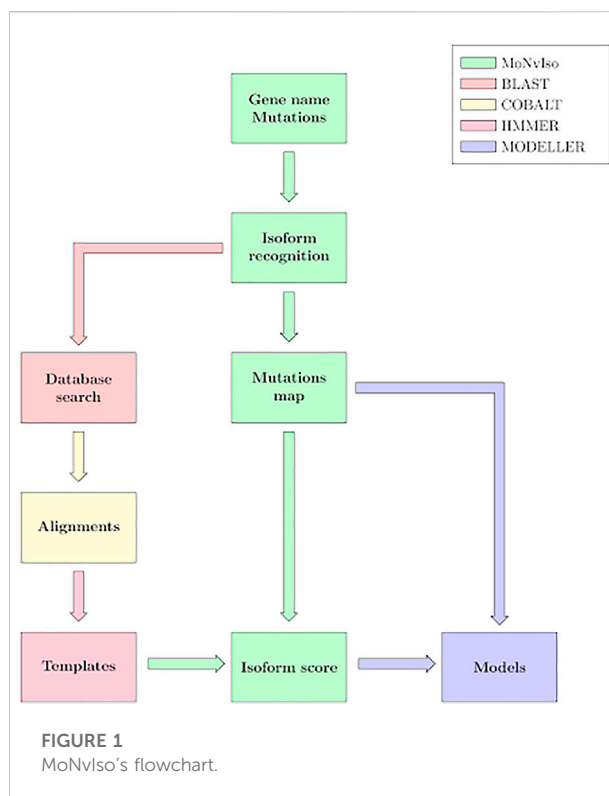
1 Introduction

The spatial and functional diversity of the 20,465 protein-coding genes (Howe et al., 2021) (<https://www.ensembl.org/>) in the human genome is dramatically augmented through alternative splicing that results in an enormous number of potential protein isoforms. Exact numbers are not fully known but common estimates for total isoforms are in the 10X range (245,000 transcripts in <https://www.ensembl.org/>). Alternative splicing can result in isoforms with relatively subtle changes through to those that vary enormously in their structure, function, and subcellular spatial expression (Park et al., 2018).

Indeed, most functional (and dysfunctional) biochemical processes are affected by the expressed isoforms, which feature distinct functional roles. Examples of this complexity include the neuroligin and neuroligin families, which perform synaptic regulatory functions that are surprisingly isoform specific (Markwick et al., 2007; Slabinski et al., 2007). This complexity may be increased by the addition of genetic variants, which can directly influence the protein structure and function of the isoform. Moreover, genetic variations can also affect the splice mechanisms and change the isoforms directly (Park et al., 2018), but this is not addressed in this study.

Further information, key to our understanding of genetic diseases, is the availability of three-dimensional structures of a protein. The structure of many human proteins is now available by accurate - yet time-consuming (Markwick et al., 2007; Slabinski et al., 2007) - experimental techniques (such as X-ray diffraction, NMR and electron microscopy (Murata and Wolf, 2018)). These accurate but demanding approaches are complemented by fast (and more approximate) computational predictions (Kuhlman and Bradley, 2019), including homology modelling (Kuhlman and Bradley, 2019) and deep learning techniques such as AlphaFold (AF) (Tunyasuvunakool et al., 2021), based on experimental structural information of evolutionarily related template protein(s) (Kuhlman and Bradley, 2019). Unfortunately, all these methods do not usually provide the isoforms most likely involved in the process of interest.

Here we present a computational platform that selects specifically the most useful isoform for molecular modelling and provides structural information, in the context of identified genetic variants. The presence of a variable number of protein isoforms makes it challenging to assign each mutation to a specific position in the protein sequence, which frequently hampers a reliable assessment of the impact of the genetic variations (including disease relevant mutations (Rees et al., 2010; Kato et al., 2018)) on an isoform suitable for molecular modelling. In other cases, a mutation is observed that is relevant to a specific isoform, but the databases reporting mutations related to a particular genetic disease usually lack a reference to the specific isoform.



Given a set of mutations at the protein expression level, our pipeline can correctly assign them to the corresponding isoforms at the protein level, providing important information that can be used for further investigations. The second key step of the determination of the isoform most useful for molecular modelling is achieved by combining the mutation-isoform map with the sequence coverage of available structural templates.

2 The MoNvIso (Modelling eNvironment for Isoforms) pipeline

The general workflow of MoNvIso is summarised in Figure 1 and proceeds according to three steps described in more details in the next subsections:

- 1) Step 1: check of the gene names provided in the input file, identification of canonical and additional isoforms extracted from the Uniprot database. In the input file a list of the mutations of interest is also present.
- 2) Step 2: check of the modelling propensity and how properly mutations are mapped on the available isoforms. The availability of templates is supervised by MoNvIso, as well as the association of the mutations to the appropriate isoforms. MoNvIso highlights failures in this mapping procedure, i.e., when mutations cannot be mapped on any available isoforms.

- 3) Step 3: Building of the structural model of the identified proteins. Model of the wild-type (WT) forms and of their variants (selected by MoNvIso according to Step 2) are built if experimental structures are not already available for the selected isoforms.

The selection procedure is based on a function, named **Selection**, (Step 2) that casts two contributions as follows:

$$\text{Selection} = w1 \cdot (\text{Structural function}) + w2 \cdot (\text{Mutation function})$$

The two terms, **Structural function** and **Mutation function** numerically translate the modelling propensity and the mapping of the mutations on the available isoforms to accomplish the two conditions. $w1$ and $w2$ are the weights of two terms. By default, $w1 = w2 = 10$ but they can be adjusted by the user. **Structural function** and **Mutation function** are described more in detail in the Subsection Step 2.

Collections of input and output files for the proteins KRAS and KDM5C are collected in example_p1.rar and example_p2.rar, which can be downloaded at <https://github.com/MoNvIsoModeling/MoNvIso>.

2.1 Step 1

MoNvIso checks the list of gene names and the set of point mutations provided by the user. The mutations can be indicated in the input file according to different formats: three-letters or single letter names for the amino acids. Additionally, spaces and tabs are also accepted to simplify the creation of the list by the user. Every gene name is searched against the Uniprot (Bateman et al., 2021) database, the results are extracted from two files, namely *uniprot_sprot.fasta*, which contains the aminoacidic sequence of the canonical isoforms according to the classification of Uniprot, and *uniprot_sprot_varsplic.fasta* collecting the sequences of the remaining isoforms obtained from Uniprot (see Supplementary Figure S1 for the list of folders and files created by MoNvIso).

2.2 Step 2

MoNvIso then performs an analysis on each isoform extracted from the Uniprot entry (see Step 1) based on two functions: 1) checking the modelling propensity and 2) mapping of the mutations. A score is associated with each function and the combination of the two is used to select the isoform most suitable to be modelled. Independently on the chosen isoform to be modelled, the information on the mapped mutations onto all the isoforms is provided by MoNvIso. In detail:

2.2.1 Checking the modelling propensity.

Each isoform is then processed according to a standard procedure: A search for homologous sequences is performed using BLAST API (Altschul et al., 1990), which allows users to submit BLAST searches for processing through cloud service provider(s) using HTTPS; and a multi sequence alignment (MSA) is generated using COBALT (Papadopoulos and Agarwala, 2007). Subsequently, based on the MSA, the *hmmsearch* function of HMMER (version 3.3.2 <http://hmmer.org/>) uses the HMM (Hidden Markov Model) (Baum and Petrie, 1966) to find relevant templates in the PDB. The 10 most similar sequences for the identified PDB structures are downloaded and the chains necessary for the homology modelling are extracted as separate files. The extracted structures are cleaned from water molecules, ligands, disordered atoms, and non-standard residues, then aligned to the MSA and are made available to the user in a folder (see Supplementary Figure S1).

The resulting structures are ranked by resolution and sequence identity to find the most appropriate templates, thus excluding crystals with poor resolution or with sequences that are very different from the original query (see Section Limitations). The default values of the sequence identity and resolution thresholds are 25% and 4.5 Å, respectively. However, the thresholds can be modified by the user. A further selection criterion is applied by calculating the coverage of the input sequence by the sequences of the templates. To this aim, MoNvIso identifies the minimum number of templates necessary to model the highest percentage of the target sequence. For a given target sequence (for example, Isoform 1 = ADRRVLTLY) and the set of templates identified as described above (for example, Template A: AD, Template B: AD, Template C: RRVLT, Template D: DRR), MoNvIso proceeds as follows:

- 1) Sorting of the templates according to the covered lengths, in our case Templates A, B, D, C;
- 2) Checking if the given sequence is covered by more than one template or by a combination of templates. In our case, Templates A and B cover the same portion;
- 3) If a single template covers the target, then this template is considered (which is not the case of our example);
- 4) If the target is covered either by a longer template or by a combination of other templates (with at least one covering extra portions of the protein), the proper selection is considered. In our example, this is accomplished by the combination of Templates A and C, being the choice between Templates A and B only dictated by the alphabetical order.

The described procedure is applied by MoNvIso to entire sequences or portions of them and to all the possible additional isoforms (our example deals with a second isoform, Isoform 2 = ADRKVLTY). Note that information about covered sections and

associated templates are stored in the *covered_intervals* file produced by MoNvIso.

Starting from the above description, the term **Structural function** in Eq. 1, accounts for the availability of crystallographic data defined as the number of amino acids (AAs) that are covered by a template (or a combination of templates) over the total number of AAs constituting the isoform

$$\text{Structural function} = \frac{(\text{Covered AA})}{(\text{Total AA})} \quad (2)$$

In the above example, for Isoform 1 we have **Total AA** = 8 and **Covered AA** = 7, resulting in a **Structural function** = 0.875, while for Isoform 2 the values of **Covered AA** and **Structural function** are 6 and 0.750, respectively.

2.2.2 Mapping of the mutations

The second term of Eq. 1, **Mutation Function**, considers the entire list of mutations provided for the considered gene, thus pinpointing to the isoform most suitable for homology modelling. Our program maps all mutations onto the appropriate isoform and increases by one the numerator, **Mutating AA that can be modelled**, if the mutated residue can be correctly located in the isoform sequence. The contribution of matched mutations to the selection function is evaluated as follows:

$$\text{Mutation function} = \frac{(\text{Mutating AA that can be modelled})}{(\text{Mutating AA found in at least 1 isoform})} \quad (3)$$

According to our example, for the three mutations T2A, R3A, R4L, MoNvIso highlights that the first mutation T2A is not mapped on the two present isoforms, while it evaluates **Mutating AA that can be modelled** equal to two and one for Isoforms 1 and 2, respectively. **Mutating AA found in at least one isoform** is two for both isoforms, **Mutation function** (Isoform 1) = 1, and **Mutation function** (Isoform 2) = 0.5.

For each gene and each isoform, the resulting **Selections** are reported in the *report.log* file. Moreover, this file contains a report on all mutations inserted in the input file, that is, i) the mapped mutations, ii) on which isoform they were mapped and iii) mutations not associated with any isoforms, together with iv) the isoform most suitable to be modelled (see **Supplementary Figure S2**). In our example, the selected isoform to be modelled is Isoform 1 with **Selection** = 18.75.

2.3 Step 3

Structural models for the selected isoform in its WT form and in all the variant(s) associated with the properly mapped mutation(s) are then created by using the MODELLER program (Webb and Sali, 2016) based on the sequence alignment obtained in the previous step. Regions not covered by the templates are not considered. The models are then ranked

by the DOPE score (Shen and Sali, 2006), and MoNvIso yields the top ranked one (the list of all the models with their DOPE score is in the file MYOUT.dat, see SI for the list of all the files generated by MoNvIso and their location). The modelling of the variants is then performed by taking the MODELLER input file containing the WT sequences of the templates and replacing the mutated AAs in the sequence. MODELLER is then run again to produce the model of the variant(s). This can be useful for mapping the position of mutations on a three-dimensional structure, allowing the study not only of the mutated residue but also of the amino acids in its vicinity and with which the mutated residue may be in contact.

3 Strengths

Our pipeline exploits a series of tools tailored to manage large sets of proteins. Useful information is provided at each step of the run so that decisions taken by the pipeline can be audited. In the case of a failure of the pipeline to provide a satisfactory structural model, the file *report.csv* traces the mutations on all the isoforms and provides an easy way to identify the isoform mapping the largest number of mutations. The previously mentioned *report.log* file is also important. This file contains all the data that would otherwise have to be manually collected such as the number of isoforms for a gene, the location of the mutations, which mutations cannot be mapped on any known isoform and finally the values of the selection functions. These data can provide a useful starting point if the user needs to manually model the protein. For example, the user, upon data retrieval, can also decide if another isoform should be prioritised because of a mutation of particular interest not present in the isoform selected by the program. Regarding the modelling part of the protocol, the final alignments, the used templates with detailed information on the selection process as well as the coverage are made available to the user, as specified thoroughly in Section 2. Although the process of building the variants can be time consuming if many of them need to be built, this part is fully automated. In most of the tested cases the models built showed a high quality and can be used for further studies (see Section Results). Thus, our pipeline reduces the time necessary to model a large number of proteins by automating the slowest parts of the process including the search for isoforms, the mapping of mutations, the search for crystallographic data to use as templates and the building of the alignments.

4 Limitations

As with any modelling study, also our method presents limitations. MoNvIso does not model the parts of the protein that are not covered by templates. The solution implemented in the program is the modelling of the single domains, although this

implies the uncertainty on reciprocal orientations of the domains. An additional drawback is the possible presence of several small portions that can be modelled but are interspersed by regions not covered by templates. In some cases, the search for templates with HMMER does not return any result (depends on HMMER's servers). When several successive searches for homologues are queued on BLAST, a slowdown of the runs may occur. Multiple point mutations coexisting on the same proteins are not modelled by MoNvIso concurrently. Rather, MoNvIso provides a series of structural models of single amino acid variants for pairwise comparison. Finally, MoNvIso selects the most useful isoform based on available structural data and mutation coverage but there is no guarantee this is the most functionally relevant one in every case.

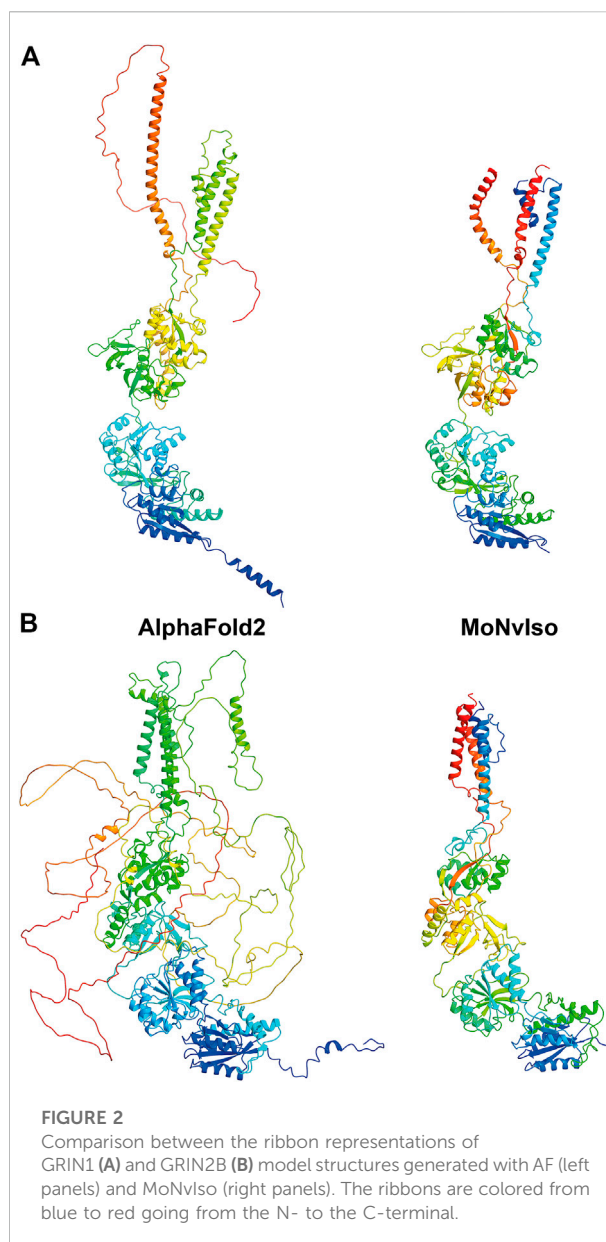
5 Case studies

We tested MoNvIso on a set of 70 proteins. A corresponding 257 human isoforms were extracted from the Uniprot database and relative mutations obtained from the relative Uniprot webpage, with a maximum cap of five mutations per protein. The genes and mutations considered are listed in the file *mutations.txt* provided in Supporting Materials. For all selected proteins MoNvIso was able to produce the alignments and to map the mutations onto the identified isoform. It successfully located, retrieved, and edited the templates to generate the WT structural models as well as the variants, when the identified mutations were in the modelled portions.

Out of the 70 proteins we modelled, 53 WT models could be compared against equivalent ones available in the AF database (DB) (<https://alphafold.ebi.ac.uk/>). This was done by extracting from the AF model the part of the sequence that we modelled and performing an RMSD analysis on the Ca.

For the remaining 17 proteins (BCL11A, CACNA1B, CAMKK1, CAMKK2, DNMT1, FMR1, GABRB3, GRIK2, GRM5, PLXNB1, SCN2A, SLC17A8, SNAP25, STX1A, SYN1, SYT1, TAF1), such comparison was not feasible because the isoform selected by MoNvIso was not the canonical one as considered by AF and was not sufficiently similar for direct comparison, i.e. the number of Ca was different. For a further 13 proteins out of 70 we modelled an isoform different from the canonical sequence but the RMSD comparison with the AF models was possible because the changes were localised in region not covered by templates.

Thus, for a total of 30 proteins out of 70 mutations are best modelled on non-canonical isoforms. The results of the comparison are presented in **Supplementary Table S1** together with the amount of residue for which AF has a high or very high confidence (pLDDT score >70) about their position. The genes are ordered from the one with lowest RMSD value to the highest. According to **Supplementary Table S1**, 44 out of 57 (77%) models present an RMSD below 20 Å, and a visual inspection reinforces



the validity of our results, since the larger RMSD values in this group are mainly due to small, disordered loops. In the group of models with RMSD above 20 there are subunits assuming different orientations in both MoNvIso and AF structures. When comparing the number of AA with a high or, very high, confidence score, we see that in most of our results (46 out of 57), the modelled portion retains at least 50% of these residues.

As an example, we show two structures in **Figure 2**: the proteins GRIN1 (Glutamate receptor ionotropic, NMDA one; also known as GluN1; Uniprot #Q05586) and GRIN2B (Glutamate receptor ionotropic, NMDA one; also known as GluN2B; Uniprot #Q13224). These two transmembrane

proteins are subunits of the N-methyl-D-aspartate (NMDA) glutamate receptor complex, which contribute to excitatory transmission in the brain. In the first case both AF and MoNvIso produce similar results that differ only in the domains for which no templates are available, but still modelled by AF. Examples of these domains are the C-terminal part, starting from K866 to S938 and the N-terminal helix (residues M1 to D23) that are modelled by AF and not by MoNvIso (see top left and bottom right in [Figure 2A](#), respectively). These two portions of the sequence are not considered by MoNvIso (see Step 3) since there are no available templates to correctly model them, but AF does attempt to model the whole chain. This leads to portions of the model with low or very low confidence scores (calculated by AF), and which corresponds to a pLDDT between 0 and 70, meaning that those parts of the model are generally unreliable.

The results for GRIN2B (see [Figure 2B](#)) demonstrate the differences between AF and MoNvIso predictions. AF successfully models the N-terminal part of the protein but fails to correctly build the trans and intra-membrane domains, which are then added as loops twisted around the correctly modelled section of the protein. Once again, the portions that are missing from the PDB database are poorly modelled. Since AF has been trained on the PDB dataset ([Tunyasuvunakool et al., 2021](#)), it still relies on available crystallographic data to correctly model structures. Thus, transmembrane domains such as those of GRIN2B, which are underrepresented in that training set because of the scarcity of experimentally determined structures of transmembrane proteins and their complexes ([Kermani, 2021](#)), may fail to be correctly built. In turn, MoNvIso automatically recognises the parts of the protein that can be modelled with confidence. As a result, MoNvIso cuts out of the sequence the extra AAs that cannot be modelled, producing a model ready to be used for further analysis.

6 Conclusion

Dissecting the impact of point mutations in the function of a protein are often hindered by a lack of an appropriate mapping of the mutation onto the correct isoform of a protein, of the identification of isoform(s) useful for molecular modelling, and of the associated building of a reliable structure. This knowledge is important because different isoforms of proteins can have widely differing functional roles and spatio-temporal expression profiles. As genomic variants associated with human traits and/or disease are being discovered at an increasing rate, approaches to link them to isoforms and find reliable structural models are urgently needed. MoNvIso addresses these two aspects: mapping a set of point mutations (provided by the user) on known isoforms, along with selecting the isoform most suitable to be modelled. The prediction of the structural models for the

WT isoforms and their variants is automated, making MoNvIso appropriate for high-throughput investigations. Although several platforms to provide accurate structures of a protein are available and routinely used ([Yang et al., 2014](#); [Webb and Sali, 2016](#); [Waterhouse et al., 2018](#)), surprisingly few of them can be implemented in a pipeline ([Webb and Sali, 2016](#)) to automate the modelling of multiple different proteins. Therefore, our protocol combines this final step with the key preliminary assessment of the isoform mapping correctly the mutation of interest. Importantly, all steps of our protocol yield results that can be used at different stages by the user: the identification of specific isoforms containing residues involved in selected mutations is *per se* a remarkable clue for genetic assessment of the impact of isoforms, especially by handling a large number of proteins and point mutations; the set of the templates eventually identified by MoNvIso with the section of the target protein covered by them are made available to the user; finally, the structural predictions represent a valuable starting point for additional refinements and investigations, such as molecular dynamics simulations ([Raval et al., 2012](#); [Hollingsworth and Dror, 2018](#); [Lazim et al., 2020](#); [Miller and Phillips, 2021](#); [Itoh and Okumura, 2022](#)), hot spots evaluation ([Murakami et al., 2017](#); [Liu et al., 2018](#); [Rosell and Fernández-Recio, 2018](#); [Rosensweig et al., 2018](#)), protein-protein docking ([Kangueane and Nilofer, 2018](#); [van Noort et al., 2021](#)) and more ([Poelwijk et al., 2016](#); [Rivoire et al., 2016](#); [Salinas and Ranganathan, 2018](#)). Finally, note that for isoforms without good quality-templates, users could choose to use predicted structures such as those provided by AF and RosettaFold ([Baek et al., 2021](#)) or other modelling packages and/or protocols to build their own structural models using the isoform(s) correctly associated with the selected point mutations.

The test of MoNvIso on a set of proteins and the comparison with the results of AF confirms the validity of our approach. Additionally, our computational protocol can be easily inserted in a pipeline suitable to perform extensive campaigns of investigation on protein-protein interactions. MoNvIso is particularly useful to evaluate the availability of templates for large sets of proteins and automatically selecting the isoform most suitable to be modelled containing the point mutations of interest. MoNvIso is freely available and can be downloaded from GitHub at the following link: <https://github.com/MoNvIsoModeling/MoNvIso>, implemented in Python 3.8 and tested on version 3.0, 3.7 and 3.9 and supported on Linux.

Key points

- 1) We have developed a computational protocol to map mutations on appropriate isoforms of protein.
- 2) The protocol identifies the available templates on which mutations can be located.

- 3) Ranking of the isoforms based on the number of located mutations and the template coverage.
- 4) Structural models are built for the WT and mutated isoforms if reliable templates are available.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://github.com/MoNvIsoModeling/MoNvIso>.

Author contributions

All authors provided contributions to study design, analysis and interpretation of data, drafting the article or revising it critically for important intellectual content. Here are the most important contributions of each author: PC, DA, OS, SR, and PR designed the study. FO, FM, AG, and PR developed the computational protocol. Data were collected by FO and PR. Analysis was carried out by FO, FM, AG, SR, PC, and PR.

Funding

SR received a Wilhelm Bessel Research Award from the Alexander von Humboldt Foundation. JA and OS received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3).

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/s0022-2836(05)80360-2
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876. doi:10.1126/science.abj8754
- Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., et al. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi:10.1093/nar/gkaa1100
- Baum, L. E., and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* 37, 1554–1563. doi:10.1214/aoms/1177699147
- Hollingsworth, S. A., and Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron* 99, 1129–1143. doi:10.1016/j.neuron.2018.08.011
- Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., et al. (2021). Ensembl 2021. *Nucleic Acids Res.* 49, D884–D891. doi:10.1093/nar/gkaa942
- Itoh, S. G., and Okumura, H. (2022). All-Atom molecular dynamics simulation methods for the aggregation of protein and peptides: Replica exchange/permutation and nonequilibrium simulations. *Methods Mol. Biol.* 2340, 197–220. doi:10.1007/978-1-0716-1546-1_10
- Kanguane, P., and Nilofer, C. (2018). Protein-protein docking: Methods and tools. *Protein-Protein Domain-Domain Interact.*, 161–168.
- Kato, G. J., Piel, F. B., Reid, C. D., Gaston, M. H., Ohene-Frempong, K., Krishnamurti, L., et al. (2018). Sickle cell disease. *Nat. Rev. Dis. Prim.* 4, 18010. doi:10.1038/nrdp.2018.10
- Kermani, A. A. (2021). A guide to membrane protein X-ray crystallography. *FEBS J.* 288, 5788–5804. doi:10.1111/febs.15676
- Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell. Biol.* 20, 681–697. doi:10.1038/s41580-019-0163-x
- Lazim, R., Suh, D., and Choi, S. (2020). Advances in molecular dynamics simulations and enhanced sampling methods for the study of protein systems. *Int. J. Mol. Sci.* 2121, 63396339. doi:10.3390/ijms21176339
- Liu, S., Liu, C., and Deng, L. (2018). Machine learning approaches for protein-protein interaction hot spot prediction: Progress and comparative assessment. *Molecules* 23, 2535. doi:10.3390/molecules23102535
- Markwick, P. R. L., Bouvignies, G., and Blackledge, M. (2007). Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. *J. Am. Chem. Soc.* 129, 4724–4730. doi:10.1021/ja0687668
- Miller, M. D., and Phillips, G. N. (2021). Moving beyond static snapshots: Protein dynamics and the protein data bank. *J. Biol. Chem.* 296, 100749. doi:10.1016/j.jbc.2021.100749

Acknowledgments

PC acknowledges the Deutsche Forschungsgemeinschaft (DFG) via the Research Training Group RTG2416 MultiSenses-MultiScales (368482240/GRK2416). We thank Emiliano Ippoliti (Jülich), Enrico Gandini (Milan), and Andrea Bosin (Cagliari) for technical support.

Conflict of interest

The 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.

The handling editor SP declared a past co-authorship with the author AG.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fchem.2022.1059593/full#supplementary-material>

- Murakami, Y., Tripathi, L. P., Prathipati, P., and Mizuguchi, K. (2017). Network analysis and *in silico* prediction of protein–protein interactions with applications in drug discovery. *Curr. Opin. Struct. Biol.* 44, 134–142. doi:10.1016/j.sbi.2017.02.005
- Murata, K., and Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica Biophysica Acta - General Subj.* 1862, 324–334. doi:10.1016/j.bbagen.2017.07.020
- Papadopoulos, J. S., and Agarwala, R. (2007). Cobalt: Constraint-based alignment tool for multiple protein sequences. *Bioinformatics* 23, 1073–1079. doi:10.1093/bioinformatics/btm076
- Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet.* 102, 11–26. doi:10.1016/j.ajhg.2017.11.002
- Poelwijk, F. J., Krishna, V., and Ranganathan, R. (2016). The context-dependence of mutations: A linkage of formalisms. *PLOS Comput. Biol.* 12, e1004771. doi:10.1371/journal.pcbi.1004771
- Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* 80, 2071–2079. doi:10.1002/prot.24098
- Rees, D. C., Williams, T. N., and Gladwin, M. T. (2010). Sickle-cell disease. *Lancet* 376, 2018–2031. doi:10.1016/s0140-6736(10)61029-x
- Rivoire, O., Reynolds, K. A., and Ranganathan, R. (2016). Evolution-based functional decomposition of proteins. *PLoS Comput. Biol.* 12, 1004817. doi:10.1371/journal.pcbi.1004817
- Rosell, M., and Fernández-Recio, J. (2018). Hot-spot analysis for drug discovery targeting protein-protein interactions. *Expert Opin. Drug Discov.* 13, 327–338. doi:10.1080/17460441.2018.1430763
- Rosensweig, C., Reynolds, K. A., Gao, P., Laothamatas, I., Shan, Y., Ranganathan, R., et al. (2018). An evolutionary hotspot defines functional differences between CRYPTOCHROMES. *Nat. Commun.* 9, 1138. doi:10.1038/s41467-018-03503-6
- Salinas, V. H., and Ranganathan, R. (2018). Coevolution-based inference of amino acid interactions underlying protein function. *Elife* 7, e34300. doi:10.7554/elife.34300
- Shen, M., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15, 2507–2524. doi:10.1110/ps.062416606
- Slabinski, L., Jaroszewski, L., Rodrigues, A. P. C., Rychlewski, L., Wilson, I. A., Lesley, S. A., et al. (2007). The challenge of protein structure determination—lessons from structural genomics. *Protein Sci.* 16, 2472–2482. doi:10.1110/ps.073037907
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Zidek, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* 596, 590–596. doi:10.1038/s41586-021-03828-1
- van Noort, C. W., Honorato, R. V., and Bonvin, A. M. J. J. (2021). Information-driven modeling of biomolecular complexes. *Curr. Opin. Struct. Biol.* 70, 70–77. doi:10.1016/j.sbi.2021.05.003
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi:10.1093/nar/gky427
- Webb, B., and Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.* 2016, 56–57.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2014). The I-TASSER suite: Protein structure and function prediction. *Nat. Methods* 12, 7–8. doi:10.1038/nmeth.3213



OPEN ACCESS

EDITED BY

Simón Poblete,
Universidad Austral de Chile, Chile

REVIEWED BY

Davit Potoyan,
Iowa State University, United States
Patricia Soto,
Creighton University, United States

*CORRESPONDENCE

Christine Peter,
✉ christine.peter@uni-konstanz.de

SPECIALTY SECTION

This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

RECEIVED 02 November 2022

ACCEPTED 23 December 2022

PUBLISHED 10 January 2023

CITATION

Hunkler S, Buhl T, Kukharensko O and Peter C (2023), Generating a conformational landscape of ubiquitin chains at atomistic resolution by back-mapping based sampling. *Front. Chem.* 10:1087963. doi: 10.3389/fchem.2022.1087963

COPYRIGHT

© 2023 Hunkler, Buhl, Kukharensko and Peter. 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.

Generating a conformational landscape of ubiquitin chains at atomistic resolution by back-mapping based sampling

Simon Hunkler¹, Teresa Buhl¹, Oleksandra Kukharensko² and Christine Peter^{1*}

¹Department of Chemistry, University of Konstanz, Konstanz, Germany, ²Max Planck Institute for Polymer Research, Mainz, Germany

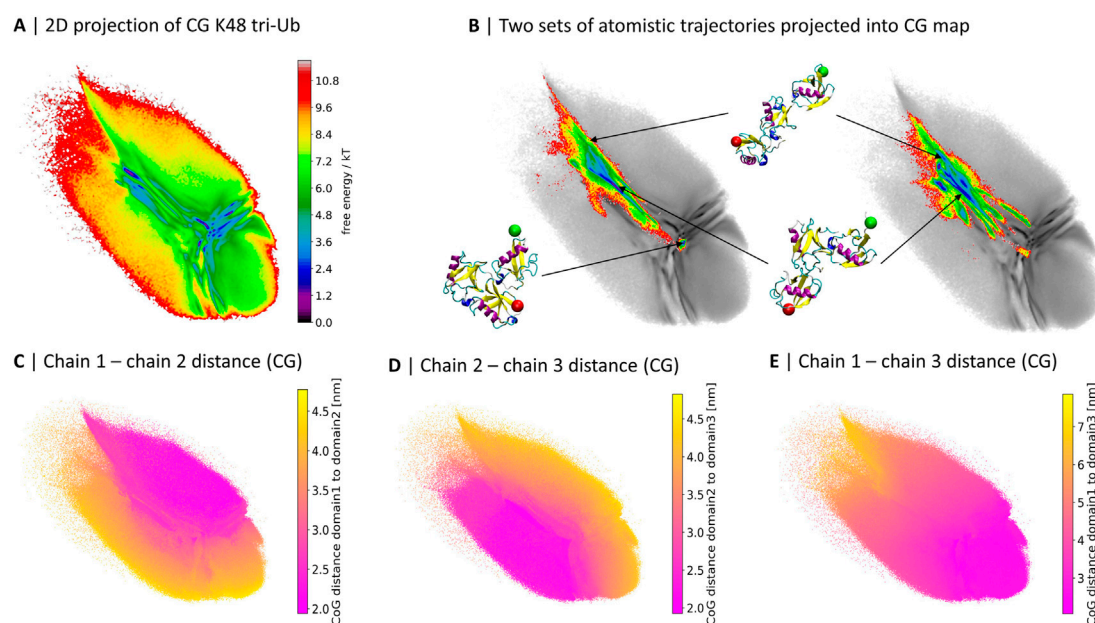
Ubiquitin chains are flexible multidomain proteins that have important biological functions in cellular signalling. Computational studies with all-atom molecular dynamics simulations of the conformational spaces of polyubiquitins can be challenging due to the system size and a multitude of long-lived meta-stable states. Coarse graining is an efficient approach to overcome this problem—at the cost of losing high-resolution details. Recently, we proposed the back-mapping based sampling (BMBS) approach that reintroduces atomistic information into a given coarse grained (CG) sampling based on a two-dimensional (2D) projection of the conformational landscape, produces an atomistic ensemble and allows to systematically compare the ensembles at the two levels of resolution. Here, we apply BMBS to K48-linked tri-ubiquitin, showing its applicability to larger systems than those it was originally introduced on and demonstrating that the algorithm scales very well with system size. In an extension of the original BMBS we test three different seeding strategies, i.e. different approaches from where in the CG landscape atomistic trajectories are initiated. Furthermore, we apply a recently introduced conformational clustering algorithm to the back-mapped atomistic ensemble. Thus, we obtain insight into the structural composition of the 2D landscape and illustrate that the dimensionality reduction algorithm separates different conformational characteristics very well into different regions of the map. This cluster analysis allows us to show how atomistic trajectories sample conformational states, move through the projection space and in sum converge to an atomistic conformational landscape that slightly differs from the original CG map, indicating a correction of flaws in the CG template.

KEYWORDS

molecular dynamics simulations, dimensionality reduction, back-mapping, coarse graining, clustering, ubiquitin, polyubiquitin

1 Introduction

Nowadays molecular dynamics (MD) simulation is a well established tool to investigate proteins and protein complexes at atomistic resolution. However it can still be computationally very expensive to obtain convergent MD trajectories for larger protein systems consisting of several thousand atoms. One typical way to overcome these limitations is to use coarse graining. Here, the number of degrees of freedom is significantly reduced by combining multiple atoms into one “super-atom” or “bead”.

**FIGURE 1**

2D projections of K48-linked tri-Ub trajectories from coarse grained (Berg et al. (2020)) (A) and two independent sets of atomistic simulations (B). (B) The atomistic simulations are colored based on free energy values, the CG map is gray and the same as in (A); three exemplary conformations from the atomistic simulations and their location in the map are illustrated. A red sphere is attached to the first residue, indicating the proximal unit, and a green sphere is attached to the last residue, indicating the distal moiety. (C–E) 2D map colored by the center of geometry (CoG) distance between two of the three Ub moieties in the CG simulations.

We used coarse grained (CG) MD simulations to study a chain of ubiquitin (Ub) proteins. Ub consists of 76 amino acids and plays an important role in cellular signaling. In a process called “ubiquitylation” an isopeptide bond is formed between a lysine group of a substrate protein and the C-terminal carboxylate group of an Ub molecule. Starting from this first Ub molecule other Ub moieties can be attached to form poly-ubiquitin chains (Ub-chains) of various lengths. The first attached ubiquitin offers eight potential linkage-sites: the N-terminal methionine (M1) and seven lysine residues (K6, K11, K27, K29, K33, K48, K63). Depending on the involved linkage-sites, chain length and topology, Ub-chains signal their substrate proteins for different functions, e.g., DNA damage tolerance or proteasomal degradation. (Pickart and Eddins, 2004; Komander and Rape, 2012).

To understand and explain differences in the physiological behavior of polyubiquitin chains one needs tools to characterize their conformational space. This is a challenging task due to a very dynamic behavior of Ub-conjugates and their conformational diversity. Thach et al. (2016) CG MD simulations in combination with dimensionality reduction and clustering techniques can be used to obtain a detailed description of the statistical ensemble of configurations populated by Ub-chains. Recently Berg et al. (2020) used a modified MARTINI v2.2 (Marrink et al., 2007; Monticelli et al., 2008; de Jong et al., 2013) CG force field and machine learning to describe and compare conformational spaces of di- and tri-Ub linked via all eight linkage-sites as well as free ubiquitins. Coarse graining massively speeds up the exploration of the phase space, but can potentially lead to inaccuracies. To assess the results of the CG sampling of tri-Ub we conducted extensive atomistic simulations (4 μ s of simulation time in total) of K48-linked tri-Ub-chains starting from an extended conformation. We compared the phase

spaces of CG and atomistic simulations by projecting all data to the same two-dimensional space (see Figures 1A,B, details on the projection method are given in Section 2).

Already at first sight, the comparison reveals that while the atomistic proteins quickly evolved from the extended starting conformation to more compact structures with contacts between the Ub-domains, large parts of the CG conformational space was not sampled during the 4 μ s of atomistic simulations.

Out of the 40 brute-force atomistic simulations only two sampled the area in the middle of the map, corresponding to a completely collapsed conformation (see Figure 1B). In order to get a better understanding of the meaning of the different regions of the map, in particular those visited by the CG model but not the atomistic one, we colored the projection of the CG simulations based on the pairwise distance between the centers of geometry (CoG) of the three Ub moieties (Figures 1C–E). The conformational landscape can roughly be divided into three parts, which are separated by a “T”-like shape of more frequently sampled areas: the upper-right part represents conformations where the first and second Ub moieties are in close contact; the lower-left side contains conformations with close contacts between the second and third moieties; and lastly there is a gradient in terms of the distance between the first and the third moiety from the upper-left hand side to the lower-right hand side.

Now the question arises whether the fact that the atomistic simulations do not visit substantial parts of the CG conformational space results from insufficient length of the atomistic simulations or unphysical conformations produced by the CG model. One method that is very well suited to address this question is back-mapping based sampling (BMBS) (Hunkler et al., 2019). We introduced this technique by analysing a rather drastically coarsened model of oligopeptides. The

application of BMBS allowed to reintroduce atomistic and dynamic information to the studied systems as well as to correct inaccuracies in the CG sampling. The core idea behind the method is the following: by navigating in two-dimensional free energy landscapes of very efficiently produced CG ensembles, selected conformations can be back-mapped to higher (e.g., atomistic) resolution to start new short explorative atomistic simulations in order to sample all of the accessible phase space as fast as possible. The convergence/divergence of the initial CG and obtained BMBS-guided atomistic landscapes is monitored quantitatively using a selected metric (earth mover's distance (EMD) (Applegate et al., 2011)). Details are given in Section 2.2 and (Hunkler et al., 2019).

In the following we show how the BMBS algorithm can be used to resolve the question whether the discrepancies between the CG and atomistic landscapes stem from insufficient atomistic sampling or from a major flaw in the CG model. Moreover, we demonstrate here that BMBS is applicable to much larger systems compared to the ones it was introduced on. We extend the originally introduced BMBS scheme with analysis of the influence of the initial weights/biases of the back-mapped configurations used to start the atomistic BMBS simulations. We also perform detailed analysis of the atomistic ensemble obtained with BMBS applying a newly introduced clustering scheme Hunkler et al. (2022).

2 Methods/Computational details

2.1 Simulation details

All atomistic simulations were performed using either the 2016.4 or the 2020.4 version of the GROMACS package (Bekker et al., 1993) with a modified GROMOS 54A7 force field (Schmid et al., 2011) and the SPC/E water model. The force field was altered by the introduction of an isopeptide bond, to be able to simulate the covalently linked Ub moieties. Furthermore the following settings were used: the time step was set to 2 fs, the temperature was set to 300 K using the velocity rescale thermostat and the pressure was set to 1 bar with the Parrinello-Rahman barostat. As an integrator algorithm, the leap-frog algorithm was used. Long range interactions were computed with the particle mesh Ewald method, where a Fourier grid spacing of .16 nm and a pme-order of 4 were used. For Coulomb and van-der-Waals interactions, a cutoff of 1.4 nm was used. In order to constrain all bonds, the LINCS algorithm was applied.

For the CG simulations a modified MARTINI force field was used (based on MARTINI v2.2) (Marrink et al., 2007; de Jong et al., 2013) where protein-water interactions were increased to avoid proteins being too sticky. The MARTINI non-polarizable CG water was used as the solvent. The temperature was set to 300 K using the velocity rescale thermostat, pressure was kept at 1 bar by the Parrinello-Rahman barostat. The Verlet cut-off scheme was applied, the LINCS algorithm was utilised for bond constraining and the leap-frog integrator was used. A 10 fs time step was used due to the soft elastic network potentials (IDEN) (Globisch et al., 2013). The cutoff distance for short-range van-der-Waals interactions was set to 1.1 nm, and electrostatics were treated by the reaction field method with a cutoff distance of 1.1 nm and a dielectric constant of 15. For more details on how the MARTINI force field was modified see Berg et al. (2018).

2.2 Back-mapping based sampling

The back-mapping based sampling (BMBS) algorithm (Hunkler et al., 2019) was used to efficiently reintroduce atomistic resolution to CG simulations and is shortly summarised here. BMBS uses a low-dimensional projection of CG free energy surfaces to initiate new atomistic simulations and consists of the following steps: 1) CG simulations are projected to a two-dimensional landscape; 2) a number of selected CG structures are back-mapped to full resolution atomistic level; 3) new short atomistic simulations are run from the selected structures to rapidly explore the phase space; 4) convergence or divergence is monitored by comparing CG and atomistic probability distributions in low-dimensional space. Those steps rely on five main components: high-dimensional collective variables (CVs) applicable to both CG and atomistic configurations, a dimensionality reduction scheme, a method to select starting configurations from the CG ensemble (seeding), a back-mapping strategy and a statistical metric to monitor convergence. All of them are described below.

2.2.1 Collective variables: Residue-wise minimal distances

In principle many different CVs/feature sets can be used in combination with the BMBS workflow. The specific choice of a CV is almost exclusively dependent on the given system. The only requirement regarding the CV is that it has to be able to describe the system in both resolutions (in the atomistic and the CG model). Therefore it must rely on coordinates that are present in both models. The CVs which we use here to describe and analyse the tri-Ub system are the residue-wise minimal distances (RMD). It has been shown before that the RMD are very well suited to describe the domain-domain configurations in ubiquitin chains since they are sensitive to the protein interfaces and to the distances and relative orientations of the domains (Berg et al., 2018; Berg and Peter, 2019; Berg et al., 2020). For one conformation of tri-Ub such a CV is a 432 dimensional vector, which contains the minimal distances of each of the 72 C_α atoms (the highly flexible residues 73–76 of ubiquitin were not considered) of each Ub domain to any C_α atom of each of the other moieties. This set of internal coordinates describes a distance as well as a relative orientation of individual ubiquitin moieties towards each other and can be applied to both atomistic as well as CG systems (if a backbone bead is present at any C_α location).

In order to describe the RMD vector of tri-Ub, the distal, middle and proximal moieties are abbreviated as A, B and C. In this notation “proximal” refers to the moiety with a free C-terminus with which the chain can be linked to the substrate and “distal” denotes the terminal moiety which is linked by its C-terminus to the middle Ub-unit. These three domains can be formulated as $A = (a_1, a_2, a_3, \dots, a_n)$, $B = (b_1, b_2, b_3, \dots, b_m)$ and $C = (c_1, c_2, c_3, \dots, c_o)$, where a_i , b_j and c_k are positions of the C_α or the backbone beads respectively. Then pairwise distance matrices $D_{A,B}$, $D_{B,C}$ and $D_{A,C}$ are computed. By taking the minimum values in each respective row and column the vectors of the residue-wise minimum distances between all three moieties (A_B , B_A , B_C , C_B , A_C , C_A) are calculated. Those vectors are then concatenated to one high-dimensional representation (432 dimensions) of the considered tri-Ub conformation, the RMD vector. All CG configurations are projected to two dimensions by using their RMD vectors as input features for the dimensionality reduction method encodermap (Lemke et al., 2019; Lemke and Peter, 2019).

TABLE 1 Encodermap parameters used to generate the 2D projection shown in this work.

Encodermap parameters	N_{steps}	N_{layers}	$N_{neurons}$	σ_{highD}	A	B	σ_{lowD}	a	b	k_a	k_s
Values	10,000	3	300	20	12	10	1	2	10	1	500

2.2.2 Dimensionality reduction: Encodermap

Encodermap (Lemke et al., 2019; Lemke and Peter, 2019) utilizes an autoencoder architecture but adjusts the autoencoder loss function by adding a multidimensional-scaling-like loss term [Equations 1 to (Eq. 3)]. This additional loss function transforms all distances by a sigmoid function (Eq. 4) and is termed as sketch-map loss due to its connection to the sketch-map dimensionality reduction method Ceriotti et al. (2011). The sketch-map loss function enables encodermap to reproduce the connectivity between high-dimensional data points in a 2D map, meaning that conformations with similar high-dimensional CVs are also located close to each other in the 2D projection. Furthermore, the autoencoder architecture enables the method to project huge amounts of data in a very short time.

$$L_{encodermap} = k_a L_{auto} + k_s L_{sketch} + Reg \quad (1)$$

$$L_{auto} = \frac{1}{N} \sum_{i=1}^N D(X_i, \tilde{X}_i) \quad (2)$$

$$L_{sketch} = \frac{1}{N} \sum_{i \neq j}^N [SIG_h(D(X_i, X_j)) - SIG_l(D(x_i, x_j))]^2 \quad (3)$$

Here, k_a , k_s are adjustable weights, Reg is a regularization used to prevent over-fitting; N denotes the number of data points to be projected; $D(\cdot, \cdot)$ is a distance between points, X is the high-dimensional input vector, x is the low-dimensional projection (the bottleneck layer); SIG_h and SIG_l are sigmoid functions of the form shown in Eq. 4,

$$SIG_{\sigma,a,b}(D) = 1 - \left(1 + \left(2^{\frac{b}{\sigma}} - 1 \right) \left(\frac{D}{\sigma} \right)^a \right)^{-\frac{b}{\sigma}}, \quad (4)$$

where a , b and σ are parameters defining the range of distances to preserve.

Once the network has been trained, the encoder works as a mathematical function that maps the high-dimensional inputs to the low-dimensional projection. In this mapping function lies one of the main advantages of the encodermap algorithm, namely the extremely efficient projection of additional high-dimensional input data points to the low-dimensional space.

Since the encodermap method is non-linear, the axes of the resulting 2D space do not necessarily allow a physical interpretation in terms of order parameters. Therefore we chose to omit the x- and y-axes for all 2D plots shown in this manuscript. Adding these axes would in our opinion rather mislead the reader than help in understanding the figures.

Similar to the choice of CVs, a different dimensionality reduction method can be chosen to be used with the BMBS workflow. However, such a method should fulfill a few requirements. First it has to be possible (and preferably fast) to project additional data points to the low-dimensional space. And secondly the method should be able to separate different structures reliably in the low-dimensional space (2D or 3D if one wants to visualize the projection). Encodermap performs remarkably well in both of these tasks and is extremely efficient in projecting data once it is trained.

The parameters for encodermap used in this work are given in Table 1. We used encodermap version 2.0.1 and its implementation from <https://github.com/AG-Peter/encodermap>.

2.2.3 Seeding

The obtained two-dimensional projection of the CG ensemble is used to seed new short atomistic MD simulations from back-mapped CG structures. If the starting conformations are chosen properly, it takes the BMBS simulations only a fraction of the simulation time compared to a standard MD to sample a comparable amount of the available phase space. In the original BMBS paper Hunkler et al. (2019) the starting configurations were chosen based on the minima in the two-dimensional CG landscape (Figure 2A). In this paper we want to explore in more details different seeding strategies and study their influence on the BMBS performance. In addition to the original seeding method, which we call here minima-focused, we test Boltzmann-weighted and uniform seeding (see Figure 2).

For the minima-focused seeding we chose the starting structures to replicate the deepest free-energy minima of the CG 2D distribution and their weighting as well as possible. To achieve this we applied a binning to the 2D CG space and created a list with the most populated bins. Then we randomly chose a data point from the highest populated bin and repeated this until the percentage of starting structures from this bin approximately matched the percentage of data points in this bin. This procedure was reiterated for all the most populated bins until a predefined number of starting conformations (50 in this paper) were obtained.

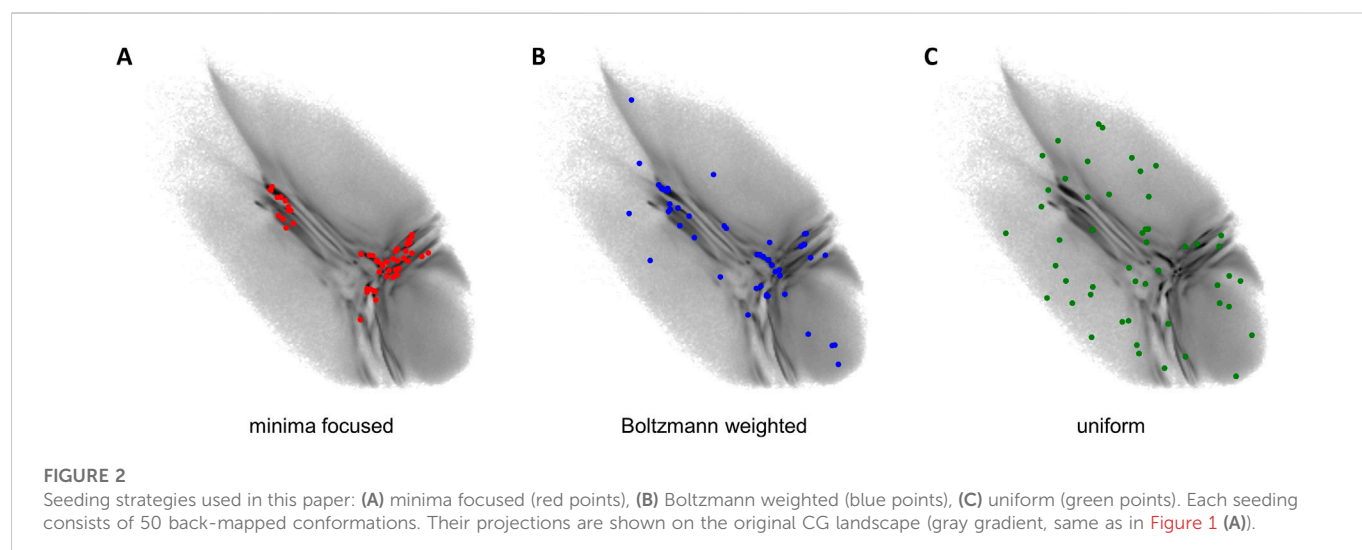
The Boltzmann-weighted seeding was chosen to also include rare conformations in the starting structures. We binned the 2D space as before but randomly picked one bin and accepted or rejected this bin with a Monte Carlo criterion (a probability proportional to the bin's population). A random data point from the accepted bin was chosen as a starting structure and the process was repeated until 50 data points were selected. Such a procedure allowed us to include rare conformations and retain as well as possible the original CG distribution given a very limited sample size (50 points). Theoretically with a much larger sample size this procedure would converge to a random selection of starting configurations from the full high-dimensional configuration space.

Lastly we chose a uniform seeding (with uniform referring to a uniform distribution in the 2D space). We again used the same binning as before and randomly chose one bin. From this bin one data point was randomly selected and the bin was then removed from the pool of available bins (the removal of a bin becomes important if the number of chosen data points approximates the number of available bins). This was again repeated until 50 starting points were selected.

The results of different seedings are compared in Section 3.1.

2.2.4 Back-mapping

In the main part of the original paper the back-mapping was done by taking an atomistic structure with CVs similar to a target CG structure. Then an external restrictive potential was applied to the



atomistic structure during an energy minimization step in order to force its conformation to retain the CVs of the CG target. In this work we used CG trajectories generated with the MARTINI model and thus applied the “backward” (Wassenaar et al., 2014) script to reintroduce an atomistic resolution into selected CG structures.

2.2.5 Statistical metric: Earth mover’s distance

To monitor a similarity between two conformational phase spaces, e.g., a CG and atomistic sampling, we use the earth mover’s distance (EMD) (also known as Wasserstein’s metric or Mallows distance). It is a metric that describes how similar or dissimilar two given multivariate distributions are. For a formal definition of the method see e.g., Applegate et al. (2011). In order to be able to quantitatively compare the EMD values we use unity-based normalized EMDs. This implementation of the EMD brings all values into the range (0,1) (Eq. 5).

$$EMD' = \frac{EMD - \min(EMD)}{\max(EMD) - \min(EMD)}, \quad (5)$$

with $\min(EMD) = 0$ and $\max(EMD) = 1.62$. The coefficient $\max(EMD)$ is hereby defined as the EMD for the comparison of the CG 2D projection with a uniform rectangular 2D distribution with the same amount of data points. The dimensions of this 2D rectangular area are given by the minimum and maximum x and y values of the CG projection. By implementing the EMD in such a way, a value of 0 means that two given distributions are exactly identical and a value of 1 means that two distributions are as dissimilar as the CG projection compared to a uniformly distributed data set. In order to compute the EMDs we used the python implementation *pyemd* v0.5.1 (Pele and Werman, 2009).

2.3 Clustering scheme

To analyse atomistic ensembles of such complex systems as tri-Ub we use a recently introduced clustering scheme which can effectively work with large amounts of high-dimensional data Hunkler et al. (2022). In this iterative clustering workflow we use HDBSCAN (Campello et al., 2015) as the clustering algorithm and combine it

with two different dimensionality reduction algorithms, namely *cc_analysis* (Diederichs, 2017) and *encodemap* (Section 2.2.2). HDBSCAN is a hierarchical density-based clustering algorithm which is able to find clusters of different shapes and densities requiring only a small number of input parameters (at least one). The *cc_analysis* is a multidimensional-scaling-like method that minimizes the differences between Pearson correlation coefficients of high-dimensional data points and the scalar product of low-dimensional vectors representing them.

In this clustering workflow the probability density in the *cc_analysis* projection is used as the clustering space (intermediate dimensionality; usually between 10 and 40 dimensions), while the 2D *encodemap* space is utilized to efficiently process large data sets and assign additional conformations to already identified clusters. The provided data set is clustered iteratively until a specified amount of conformations is assigned to clusters or until a specified amount of clustering iterations have been performed. In the process of assigning conformations to clusters a root-mean-square deviation (RMSD) cutoff of C_α atom positions is used to obtain conformationally very defined clusters.

For applying the clustering scheme to the tri-Ub system we set the HDBSCAN parameters “min_cluster_size” and “min_samples” to 80 and used an RMSD cutoff distance of 3 Å. The clustering scheme was run for three iterations.

3 Results and discussion

3.1 BMBS

We applied the BMBS method to the K48-linked trimer of ubiquitin with three different seeding algorithms: minima focused, Boltzmann weighted, and uniform (see Section 2.2.3 for detailed description). In each case we chose 50 starting points. For every starting structure we ran an atomistic MD simulation for 50 ns with a cumulative simulation time of 2.5 μs for each seeding. The location of the 50 starting points is shown in Figure 2. The BMBS simulation trajectories were projected to the original CG landscape and can be seen in Figures 3A–C. These three maps show that the choice of

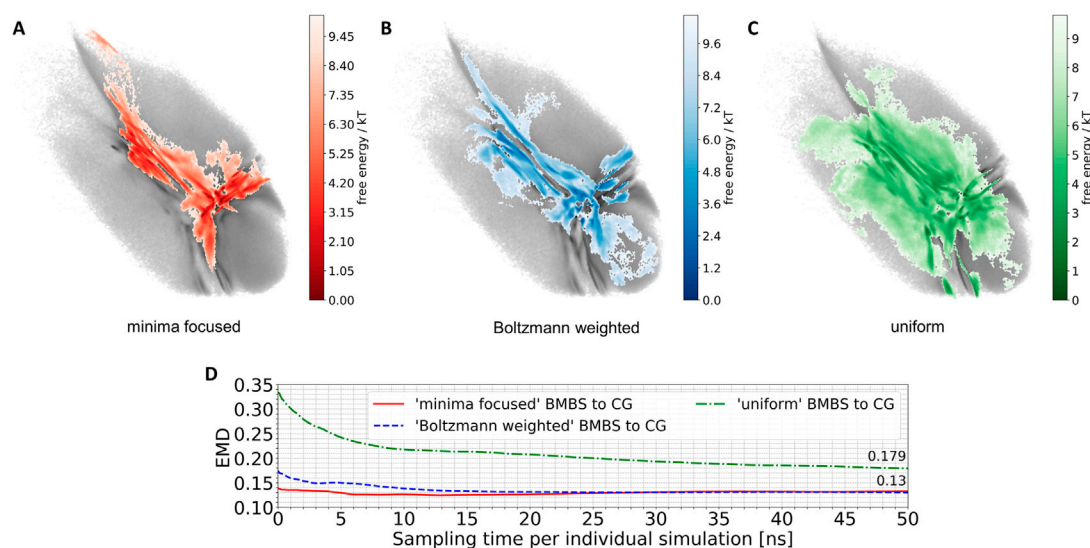


FIGURE 3

Projections of 50 atomistic simulations obtained using BMBS with minima-focused (A), Boltzmann-weighted (B) and uniform (C) seedings. (D) EMD values between CG and BMBS projections as a function of sampling time. Colors of the projections and EMDs lines correspond to the coloring in Figure 2.

starting points heavily influences the resulting conformational space (a detailed analysis of the obtained conformations and their spreading in the 2D projections is discussed in Section 3.3).

The BMBS with all three seedings visited the bottom part of the CG 2D map which was not sampled by the two initial 4 μ s atomistic simulations (compare to Figure 1B). Notably the uniformly seeded trajectories retain the “T” shaped arrangement of free-energy minima of the original distribution even though only few of the starting conformations were selected in those parts of the map. This indicates a rather quick progression of the trajectories that were seeded near the rims to the center part of the 2D projection.

A purely visual comparison of the obtained maps can be misleading as it is important to not only cover the CG phase space but to properly sample the free energy minima. For a quantitative comparison of such two-dimensional distributions we use the EMD, which fits perfectly into the BMBS workflow. The EMD is not sensitive to bin sizes (can be applied for comparing different histograms), is symmetric, and is more sensitive to similarities in highly populated regions than to the rarely populated ones. The EMD values comparing the original CG projection with the time evolution of the differently seeded BMBS projections are shown in Figure 3D. Contrary to visual perception, the EMD plot shows that both the minima-focused and Boltzmann-weighted seedings produce atomistic ensembles whose projections resemble the CG target map much better (an EMD value of .13 after 50 ns of simulation time of the individual runs) than the projection of the uniformly seeded trajectories (.179). On the other hand, the uniformly seeded BMBS approaches the CG distribution very quickly, especially in the first 10 ns of individual simulation time. To put these EMD values into perspective, the comparison of the projection of the initial 4 μ s atomistic simulations to the CG distribution gives an EMD of .815.

Therefore we can address the initial question on the reason of the discrepancy between the CG and atomistic ensembles. By applying the BMBS algorithm to the K48-linked tri-ubiquitin, we obtained 150 atomistic BMBS trajectories which provide enough evidence to

confidently say that the CG ensemble does not include a large amount of unphysical conformations. Given enough simulation time, the two initial atomistic trajectories would most likely also have sampled the conformations that reside in the lower parts of the 2D map.

The generation of these new atomistic trajectories is however only one aspect of the BMBS algorithm. Another part is the monitoring and comparison of the 2D histograms which develop over time. This analysis is provided in the next section.

3.2 EMD monitoring

In order to analyse the temporal/chronological development of the BMBS compared to the CG map we extracted 2D projections of BMBS trajectories for different sampling times. We chose to generate one histogram every 250 ps of individual simulation time for a good temporal resolution. This resulted in 200 projections for each seeding approach. For each of these histograms we computed the EMD to the CG 2D map and obtained EMD values shown in Figure 3D and Figure 4.

In addition to the time evolution of the minima-focused BMBS (red lines in both figures) provided in Figure 3D, Figure 4 shows the reversed timeline of the minima-focused BMBS histogram (orange line) to the CG map. By reversed we mean that the projection of the last frame of each minima-focused trajectory is the starting point from which the histogram grows contrary to the original timeline, meaning that each histogram starts from a point where the trajectory could sample for some time and therefore will most likely be in some meta stable state. The forward timeline (red line in Figure 4) has a non-monotonic behaviour with the initial decrease in EMD values (the two histograms become more similar to each other) until about 13 ns, followed by an increase and plateauing of the values at about .13. The same behaviour was found in the original (Hunkler et al., 2019) paper for a predictive CG model based on extrapolated data and could be explained as a correction of flaws in CG sampling. To reduce the

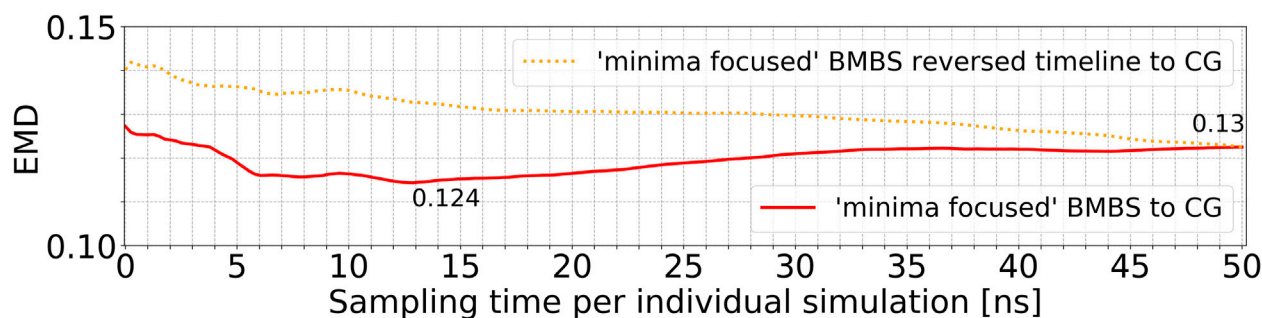


FIGURE 4

Time resolved EMDs of both forward (red solid line, same as in Figure 3D) and reversed (orange dotted line) timelines of minima-focused BMBS histograms to the CG map.

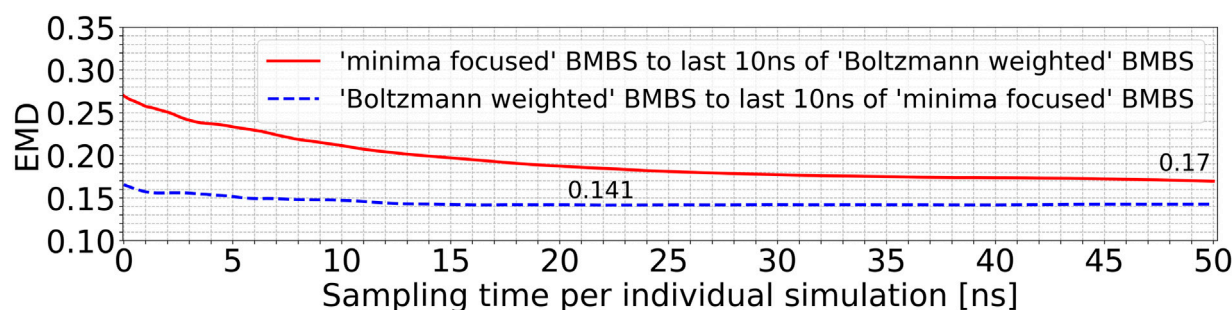


FIGURE 5

EMDs of the entire trajectories of the minima-focused seeding to the histogram of the last 10 ns of the Boltzmann-weighted seeding (red curve) and vice versa (blue curve).

influence of the seeding bias on the 200 time-resolved histograms we also included the reversed timeline (orange line in Figure 4). This timeline shows that the BMBS trajectories moved away from their initial seedings. With increasing simulation time the trajectories approach their original starting points, which leads to a decrease in the EMD values. This clearly shows that the BMBS trajectories move away from the most populated areas in the CG 2D map and indicates that the underlying CG distribution of conformations is not perfectly representing the conformational ensemble corresponding to the atomistic Hamiltonian.

Using EMDs we also monitored and compared the behaviour of different seeding approaches to each other. Figure 5 compares the minima-focused (red curve) and Boltzmann-weighted (blue curve) seedings to the histograms generated by the last 10 ns of the simulations from the respective other seeding. With this comparison we can identify if two sets of trajectories converge to sample a shared part of phase space or whether they diverge over time to different accessible areas of the conformational space. The blue curve in Figure 5 changes only slightly, while there is a much more significant decrease in the red curve. The minima-focused histograms are more similar to the histogram representing the last 10 ns of the Boltzmann-weighted trajectories than vice versa (reflected by the generally lower EMD values). These observations allow us to draw two conclusions. First, the

minima-focused trajectories initially move away from their seeding points but then do not change much in the remaining simulation time. And secondly, the Boltzmann-weighted trajectories significantly move away from their original seeding and approach the same areas in the 2D map as the minima-focused trajectories. This shows that the two systems evolve in the same general direction, even though they are partially sampling quite different areas of the 2D map at the end of the simulations.

Lastly we assess the question if the convergence of MD simulations can be monitored using EMDs. Generally, a continuous upwards or downwards trend in the EMD values indicates that the corresponding atomistic ensemble has not converged yet. However, even if the EMD curve has not changed significantly over a longer period of time, that does not imply that a convergence has been reached. As can be seen in Figure 3D the EMD plots from 25 to 50 ns of individual simulation time for all three seedings only show a very minimal change over time. But by comparing the three curves quantitatively, one observes higher EMD values for the uniform seeding compared to other two approaches, consequently the uniform simulations cannot be converged. Overall this means that none of the three BMBS ensembles can be considered converged and that an additional simulation time has to be invested to cover the full phase space and produce an ensemble that is representative of the actual atomistic free-energy landscape. However, the EMD of 2D

histograms can be an additional easily employed and efficient indicator of the current degree of non-convergence.

The general workflow which we propose in this manuscript is compatible with any atomistic force field, water model or CG model (as long as the CV of choice is available in both the atomistic and CG representations). In Hunkler et al. (2019) we demonstrated the use of the BMBS with different CG models, moreover it can be very informative in comparing the 2D probability distributions of various atomistic or coarse grained force fields with each other. As an example one could take the results of the comparison of the probability distributions generated by the two force fields used in this work (modified GROMOS 54A7 and modified Martini v2.2). We have shown that the resulting 2D distributions differ and have interpreted this difference as flaws in the CG model (i.e. due to the shape of the minima-focused EMD curve). Yet, it would be difficult to prove whether the discrepancies in the 2D projections actually stem from the CG or the atomistic model (or both). If however, we would now make the same comparison using a different atomistic force field (but the same back-mapped starting conformations), we could compare both the atomistic 2D distributions with the CG model, as well as the atomistic distributions with each other. This could lead to a much better understanding of the origin of the differences in the 2D projections and be useful for efforts to improve simulation models in either resolution.

To summarize, the EMD, especially if used in a time resolved fashion, is a very useful tool to analyse (2D) projections of the sampled phase space of MD trajectories. We showed that the EMD can be used to follow atomistic trajectories (that were specifically seeded based on the minima of a CG template map) evolution over time compared to the CG template. By first approaching the seeding template but then moving away from it, the EMD curve alludes to a correction of flaws in the CG map. This assessment of the quality of the CG model is one of the strongest features of a minima-focused back-mapping based sampling. The uniform seeding on the other hand is primarily useful in order to obtain atomistic conformations from all the CG space as fast as possible. However, if one wants to generate a (close to) converged atomistic ensemble that realistically represents the actual conformational landscape, the Boltzmann-weighted seeding is the best choice. It is on the one hand much faster in sampling of low energy conformations compared to the uniform seeding (assuming the CG model is somewhat viable) and on the other hand it includes less bias of the CG map compared to the minima-focused seeding.

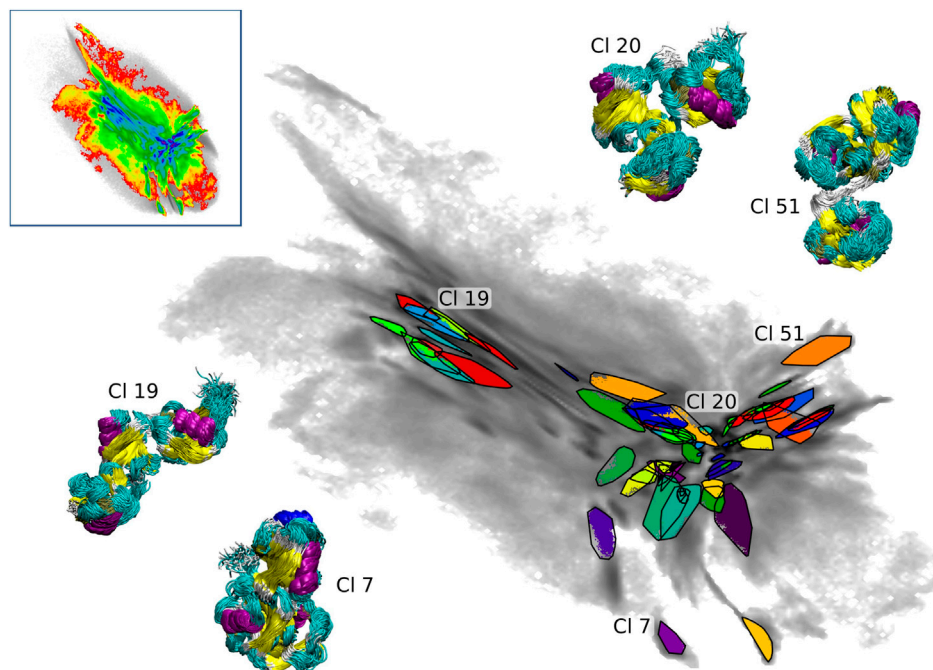
3.3 Cluster analysis

For the choice of starting configurations and the monitoring of the convergence, the BMBS scheme relies on the 2D projection of the CG configurational space. This is a radical reduction in dimensionality considering the size of tri-Ub. Thus we decided to assess a quality of this map by performing a clustering analysis in the high-dimensional space of the atomistic configurations sampled with BMBS. Such clustering can provide information on general conformational trends in the map (similar to the change in CoG distances between Ub moieties shown in Figures 1C–E) or show if the 2D projection is able to separate relatively similar conformations. Additionally it allows us to study the behaviour of individual short trajectories, e.g., whether

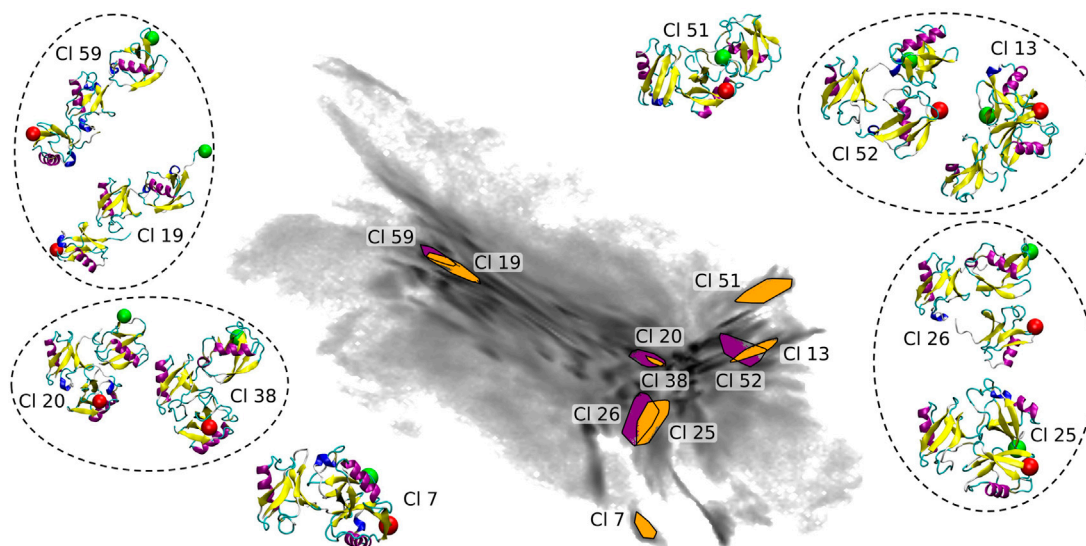
the same conformations were sampled by trajectories from different origins (i.e. different seeding schemes and different starting regions on the 2D map). This can complement the convergence analysis based on the EMDs discussed in Section 3.2. Considering the system sizes and complexity we used a recently developed clustering scheme which is specifically designed to efficiently cluster large MD trajectories Hunkler et al. (2022) (see Section 2.3).

We applied the clustering workflow to the combined atomistic data of all three seeding schemes (upper left inset in Figure 6). The data set contains 7.44 million conformations and 30% of these were assigned to 61 clusters after three iterations of the clustering process (the RMSD cutoff was set to 3 Å). As described in details in Section 2.3, the clustering was performed in the intermediate-dimensional space determined by cc_analysis and the resulting clusters were then projected into the 2D map. They are shown in Figure 6 including tri-Ub structures belonging to four example clusters (structure bundles in the insets) to demonstrate the structural consistency obtained by the clustering method (the shown cluster numbers are used as they are assigned during the clustering process and do not reflect any meaningful ordering e.g., by cluster size). The compact placement of the clusters on the map shows that the 2D map is a meaningful representation of the high-dimensional conformational landscape - a property that was important for the use of this projection for BMBS and for the comparison of the atomistic and CG sampling with EMD.

The coloring based on the CoG distances shown in Figures 1C–E provides a general understanding of the map. In order to get a more detailed insight we show 10 clusters (including representative tri-Ub configurations) from all parts of the 2D map (see Figure 7). These clusters were selected based on their location in the 2D projection. Conformations at the left hand side of the map (example clusters 19 and 59) are in general open chain conformations, meaning that the proximal and distal moieties extend to opposite directions from the middle moiety. The two clusters 20 (the largest cluster containing 3.5% of all conformations) and 38 in the center of the map adopt a collapsed conformation where each of the three moieties are roughly in equal distance to each other. Those are the most stable conformation in the system. One possible reason for this stability is that the hydrophobic patches on the distal and the middle moieties (primarily the part around the residues Ile 44 and Val 70) are orientated towards the other units and are thereby shielded from solvent. Cluster 38 intersects in the 2D projection with cluster 20. They are however still identified as two different clusters since they differ (mostly) in a small rotation of the distal moiety. This is a nice illustration of the precision and sensitivity of the proposed clustering workflow and its ability to pick up such minimal structural differences and separate the conformations into different clusters. Other examples of clusters overlapping in the 2D projection but having small structural differences identified by clustering in a higher-dimensional space are circled in Figure 7. In the clusters 51 and 52 (on the right hand side of the map) the middle and distal moieties (green sphere) are further apart than in the most populated cluster 20 (middle of the map). Especially in cluster 51, the proximal moiety is almost located between the other two. For cluster 7 the situation is exactly reversed, here the distal and middle chains are more distant and the proximal chain is located in between the two other units. So the clusters shown here confirm the general trends that we derived from the CoG distance distributions.

**FIGURE 6**

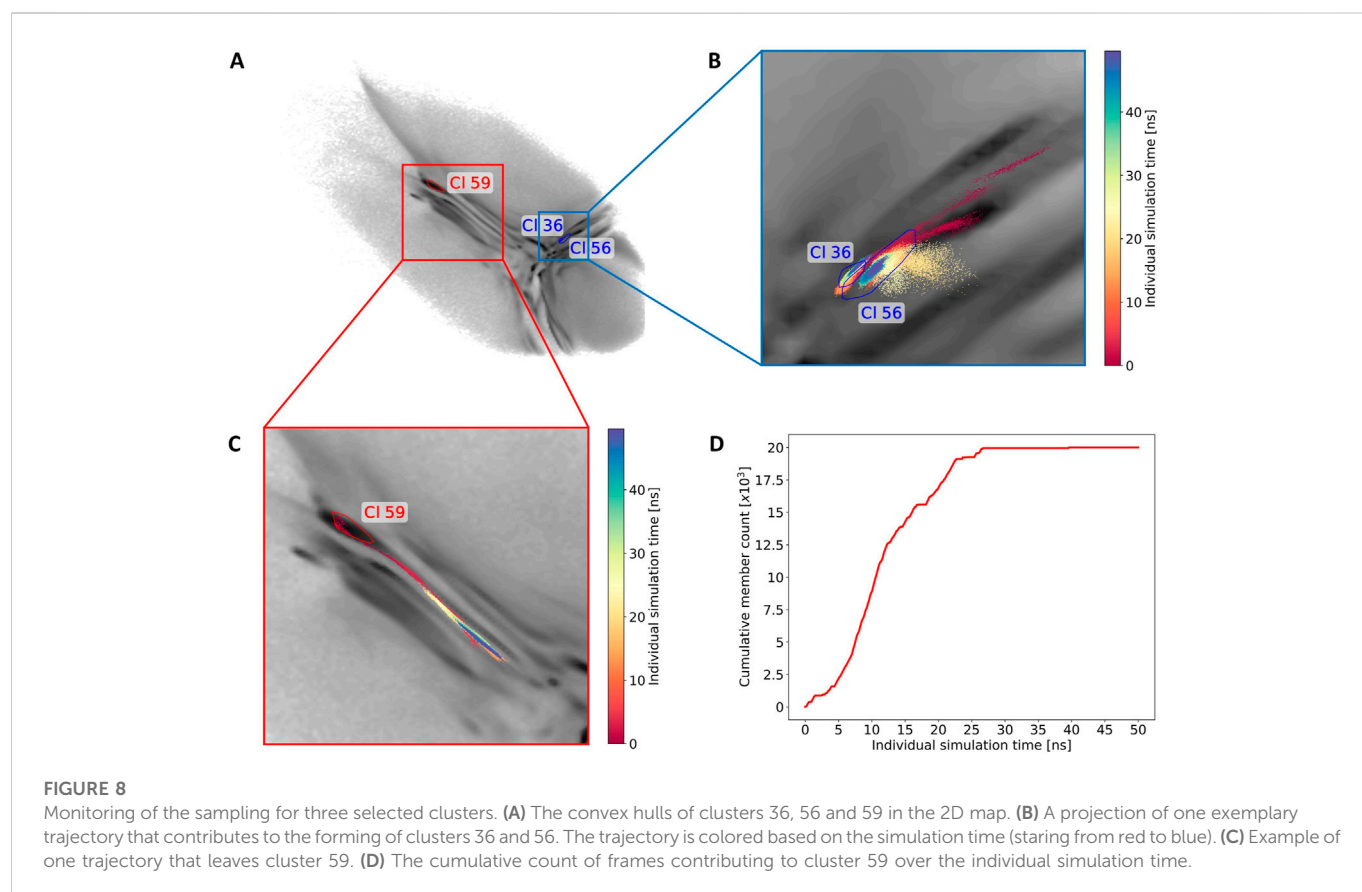
Projections of 61 clusters from the combined BMBS trajectories (gray gradient). Bundles of the structures (colored according to the secondary structure) from selected clusters are shown to visualize the homogeneity of the found clusters. The upper left inset shows the projection of the combined BMBS trajectories in the original CG landscape.

**FIGURE 7**

Selected clusters and their representative conformations in the BMBS projection (the same as in Figure 6). In all inset plots the middle moiety of the tri-Ub system is positioned in the same way. A red sphere is attached to the first residue, indicating the proximal unit, and a green sphere attached to the last residue, indicating the distal moiety. Conformations from clusters overlapping in 2D map are circled.

By using this clustering analysis we can also try to verify our statement about the ability of BMBS to correct flaws in the CG sampling using the minima-focused seeding. In Section 3.2 we argued (based on the minima-focused BMBS vs CG EMD plots) that the atomistic BMBS trajectories partially move away from the

area in the 2D projection they were seeded in and thereby generate an atomistic 2D distribution that slightly differs from the CG one. This process can be seen as a mending of inherent defects in the CG model. To verify this, we inspect a few clusters and follow individual trajectories in the 2D landscape (Figure 8A). We start



again with cluster 59 (left side of the map with extended conformations). Of the 150 independent trajectories 8 were initiated in or around that state but leave the cluster during the simulation time (a projection of one such trajectory is illustrated in **Figure 8C**). **Figure 8D** shows the cumulative number of members of cluster 59 *versus* the simulation time of the individual trajectories. This plot illustrates that the simulated trajectories indeed first sample cluster 59 and quickly populate it until around 11 ns of individual simulation time, but then the amount of conformations that are assigned to the cluster decreases. From around 25 ns onwards the cluster is not expanding. This means that after the first half of the simulated time all trajectories that have been initiated in this cluster (due to the high population of that specific area in the CG projection) have moved away from it. This example complements the correction trend observed in the EMD plots (**Figure 4**).

Next we show an example of two intersecting clusters 36 and 56 which are formed by several atomistic trajectories (**Figure 8A**). **Figure 8B** shows projections of two selected trajectories forming these clusters. In this case four BMBS trajectories that were initiated in and around a local minimum of the CG projection moved away from their seeding points and formed clusters in a less populated area of the CG map. This is another illustration where the 2D distribution of the atomistic BMBS trajectories slightly differs from the CG template distribution. This time, however, the BMBS trajectories do not collectively abandon one area of the map,

but rather collectively move towards one specific section that was not heavily populated by the CG model.

4 Conclusion

We have applied back-mapping based sampling to obtain a conformational free-energy landscape of a flexible multidomain protein—K48-linked tri-ubiquitin—at atomistic resolution. BMBS had been introduced for much smaller peptides, where we had shown that the method is able to very efficiently generate a correctly weighted atomistic ensemble based on a 2D projection of a coarse grained simulation ensemble. For tri-Ub we first generated a 2D projection of a set of extensive CG simulations with the help of the dimensionality reduction method encodermap. From projecting the structures from a long (4 μ s) atomistic simulation onto this 2D map, we found that these simulations had only visited a very limited part of the CG 2D landscape. By employing the BMBS algorithm, we found that the entire CG map is accessible to the atomistic trajectories, i.e. the CG simulations had in fact not sampled unphysical conformations. Rather, free energy barriers between different (metastable) conformational states are too high to be easily overcome on the timescales accessible to the atomistic model. This successful application of BMBS to tri-Ub illustrates that the method scales very well with system size. Furthermore we compared different

seeding methods to initiate the atomistic simulations in the 2D projection: minima focused, Boltzmann weighted and uniform. We argue that Boltzmann weighted seeding is more advantageous in its ability to retain a correct free energy profile on the one hand and, on the other hand, to explore bigger areas of conformational space. In this context we also illustrate and discuss the use of the EMD metric for the comparison of different (2D) distributions in a time-resolved fashion. Lastly, we employed a recently introduced conformational clustering workflow to the combined atomistic BMBS trajectories. In doing so we illustrate which parts of the 2D map represent which structural conformations. In this context we also show that the encodemap algorithm separates different conformational characteristics very well into different regions of the 2D map, which validates the whole BMBS approach. Finally, we show how individual atomistic BMBS trajectories sample conformational states, move through the 2D map and in sum converge to an atomistic 2D distribution that slightly differs from the CG one, indicating a correction of flaws in the CG template.

Data availability statement

The python notebooks used to analyze the data in this study, as well as a minimal example consisting of 28,000 random CG structures can be found in https://github.com/AG-Peter/BMBS_of_tri-ubiquitin. The CG trajectories, selected back-mapping points and encodemap projections of all used data can be found in <https://doi.org/10.48606/40>.

Author contributions

SH performed the simulations of all BMBS trajectories and the analysis of the data. TB contributed by finding suitable encodemap

parameters and by performing the initial two sets of atomistic tri-Ub simulations. SH, OK, and CP designed the research. SH, OK, and CP wrote the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by the DFG through CRC 969. Furthermore the authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1134-1 FUGG.

Acknowledgments

We would like to thank Andrej Berg for providing the CG data used in this work. Furthermore we thank Madlen Malcharek for helpful comments regarding the manuscript.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Applegate, D., Dasu, T., Krishnan, S., and Urbanek, S. (2011). "Unsupervised clustering of multidimensional distributions using Earth mover distance," in *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (New York, NY, USA: Association for Computing Machinery, KDD '11), 636–644. doi:10.1145/2020408.2020508
- Bekker, H., Berendsen, H., Dijkstra, E., Achterop, S., Vondrumen, R., Vanderspoel, D., et al. (1993). "Gromacs - a parallel computer for molecular-dynamics simulations," in 4th international conference on computational physics. Editors R. DeGroot and J. Nadrchal (Physics computing World Scientific Publishing), 252–256. (PC 92); Conference date: 24-08-1992 Through 28-08-1992.
- Berg, A., Franke, L., Scheffner, M., and Peter, C. (2020). Machine learning driven analysis of large scale simulations reveals conformational characteristics of ubiquitin chains. *J. Chem. Theory Comput.* 16, 3205–3220. doi:10.1021/acs.jctc.0c00045
- Berg, A., Kukharensko, O., Scheffner, M., and Peter, C. (2018). Towards a molecular basis of ubiquitin signaling: A dual-scale simulation study of ubiquitin dimers. *PLOS Comput. Biol.* 14, e1006589. doi:10.1371/journal.pcbi.1006589
- Berg, A., and Peter, C. (2019). Simulating and analysing configurational landscapes of protein-protein contact formation. *Interface Focus* 9, 20180062. doi:10.1098/rsfs.2018.0062
- Campello, J. G. B. R., Moulavi, D., Zimek, A., and Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data* 10, 1–51. doi:10.1145/2733381
- Ceriotti, M., Tribello, G. A., and Parrinello, M. (2011). Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci.* 108, 13023–13028. doi:10.1073/pnas.1108486108
- de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., et al. (2013). Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* 9, 687–697. doi:10.1021/ct300646g
- Diederichs, K. (2017). Dissecting random and systematic differences between noisy composite data sets. *Acta Crystallogr. Sect. D.* 73, 286–293. doi:10.1107/S2059798317000699
- Globisch, C., Krishnamani, V., Deserno, M., and Peter, C. (2013). Optimization of an elastic network augmented coarse grained model to study ccmv capsid deformation. *PLOS ONE* 8, e605822–e60618. doi:10.1371/journal.pone.0060582
- Hunkler, S., Diederichs, K., Kukharensko, O., and Peter, C. (2022). *Fast conformational clustering of extensive molecular dynamics simulation data. submitted.*
- Hunkler, S., Lemke, T., Peter, C., and Kukharensko, O. (2019). Back-mapping based sampling: Coarse grained free energy landscapes as a guideline for atomistic exploration. *J. Chem. Phys.* 151, 154102. doi:10.1063/1.5115398
- Komander, D., and Rape, M. (2012). The ubiquitin code. *Annu. Rev. Biochem.* 81, 203–229. PMID: 22524316. doi:10.1146/annurev-biochem-060310-170328
- Lemke, T., Berg, A., Jain, A., and Peter, C. (2019). EncoderMap(II): Visualizing important molecular motions with improved generation of protein conformations. *J. Chem. Inf. Model.* 59, 4550–4560. doi:10.1021/acs.jcim.9b00675
- Lemke, T., and Peter, C. (2019). EncoderMap: Dimensionality reduction and generation of molecule conformations. *J. Chem. Theory Comput.* 15, 1209–1215. doi:10.1021/acs.jctc.8b00975

- Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and de Vries, A. H. (2007). The martini force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B* 111, 7812–7824. doi:10.1021/jp071097f
- Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S. J. (2008). The martini coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.* 4, 819–834. doi:10.1021/ct700324x
- Pele, O., and Werman, M. (2009). “Fast and robust Earth mover’s distances,” in IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September 2009, 460–467. doi:10.1109/ICCV.2009.5459199
- Pickart, C. M., and Eddins, M. J. (2004). Ubiquitin: Structures, functions, mechanisms. *Biochimica Biophysica Acta (BBA) - Mol. Cell Res.* 1695, 55–72. doi:10.1016/j.bbamcr.2004.09.019
- Schmid, N., Eichenberger, A. P., Choutko, A., Riniker, S., Winger, M., Mark, A. E., et al. (2011). Definition and testing of the gromos force-field versions 54a7 and 54b7. *Eur. Biophysics J.* 40, 843–856. doi:10.1007/s00249-011-0700-9
- Thach, T. T., Shin, D., Han, S., and Lee, S. (2016). New conformations of linear polyubiquitin chains from crystallographic and solution-scattering studies expand the conformational space of polyubiquitin. *Acta Crystallogr. Sect. D.* 72, 524–535. doi:10.1107/S2059798316001510
- Wassenaar, T. A., Pluhackova, K., Böckmann, R. A., Marrink, S. J., and Tieleman, D. P. (2014). Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput.* 10, 676–690. doi:10.1021/ct400617g



OPEN ACCESS

EDITED BY

Adolfo Poma,
Institute of Fundamental Technological
Research, Polish Academy of Sciences,
Poland

REVIEWED BY

Rikhia Ghosh,
Icahn School of Medicine at Mount Sinai,
United States
Aykut Erbas,
Bilkent University, Türkiye
Miłosz Wierczór,
Institute for Research in Biomedicine,
Spain

*CORRESPONDENCE

Viviana Monje-Galvan,
✉ vmonje@buffalo.edu

SPECIALTY SECTION

This article was submitted to Theoretical
and Computational Chemistry,
a section of the journal
Frontiers in Chemistry

RECEIVED 03 November 2022

ACCEPTED 28 December 2022

PUBLISHED 12 January 2023

CITATION

Ramirez RX, Campbell O, Pradhan AJ,
Atilla-Gokcumen GE and Monje-Galvan V
(2023), Modeling the molecular fingerprint
of protein-lipid interactions of MLKL on
complex bilayers.
Front. Chem. 10:1088058.
doi: 10.3389/fchem.2022.1088058

COPYRIGHT

© 2023 Ramirez, Campbell, Pradhan,
Atilla-Gokcumen and Monje-Galvan. 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.

Modeling the molecular fingerprint of protein-lipid interactions of MLKL on complex bilayers

Ricardo X. Ramirez¹, Oluwatoyin Campbell¹, Apoorva J. Pradhan²,
G. Ekin Atilla-Gokcumen² and Viviana Monje-Galvan^{1*}

¹Department of Chemical and Biological Engineering, School of Engineering and Applied Sciences, University at Buffalo, Buffalo, NY, United States, ²Department of Chemistry, College of Arts and Sciences, University at Buffalo, Buffalo, NY, United States

Lipids, the structural part of membranes, play important roles in biological functions. However, our understanding of their implication in key cellular processes such as cell division and protein-lipid interaction is just emerging. This is the case for molecular interactions in mechanisms of cell death, where the role of lipids for protein localization and subsequent membrane permeabilization is key. For example, during the last stage of necroptosis, the mixed lineage kinase domain-like (MLKL) protein translocates and, eventually, permeabilizes the plasma membrane (PM). This process results in the leakage of cellular content, inducing an inflammatory response in the microenvironment that is conducive to oncogenesis and metastasis, among other pathologies that exhibit inflammatory activity. This work presents insights from long all-atom molecular dynamics (MD) simulations of complex membrane models for the PM of mammalian cells with an MLKL protein monomer. Our results show that the binding of the protein is initially driven by the electrostatic interactions of positively charged residues. The protein bound conformation modulates lipid recruitment to the binding site, which changes the local lipid environment recruiting PIP lipids and cholesterol, generating a unique fingerprint. These results increase our knowledge of protein-lipid interactions at the membrane interface in the context of molecular mechanisms of the necroptotic pathway, currently under investigation as a potential treatment target in cancer and inflammatory diseases.

KEYWORDS

protein-lipid interactions, lipid membrane modeling, local lipid fingerprint, MLKL protein, molecular dynamics simulations, mechanisms of cell death

Introduction

The plasma membrane (PM) is the natural barrier that encapsulates cells and cellular organelles; it is composed primarily of lipids arranged in a bilayer, proteins, and sugars (Corradi et al., 2018). Lipids constitute the structural backbone of the membrane, and their relative composition modulates membrane tension, rigidity, and shape (Casares and EscrivRossello, 2019). Furthermore, lipids have a dynamic interaction with transmembrane and peripheral membrane proteins (Kandt et al., 2008; Sapay and Tieleman, 2008; Monje-Galvan and Klauda, 2018) that is relevant to cell signaling cascades, ionic flux, cargo transport, mechanisms of cell death, and disease progression. The molecular-level understanding of these protein-lipid interactions at the membrane interface is relevant to understand mechanisms of membrane permeabilization and cell death. This knowledge can be potentially leveraged in the treatment of several diseases such as cancer.

For instance, necroptosis is a caspase-independent programmed cell death pathway under consideration as a potential cancer treatment (Wang et al., 2017; Gong et al., 2019; Qin et al.,

2019). This pathway initiates when the tumor necrosis factor TNF- α binds its receptor and ends with the permeabilization of the PM and the leakage of cellular content. The process responsible for PM permeabilization is the interaction of mixed lineage kinase-like (MLKL) protein with membrane lipids. MLKL is the final executor of necroptosis by translocating to the PM and causing membrane disruption (Galluzzi et al., 2017; Chen et al., 2019; Choi et al., 2019). However, the details of these protein-lipid interactions and the corresponding membrane permeabilization mechanism are unknown. Necroptosis is a relevant pathway in cancer, and also in neurodegenerative and inflammatory diseases (Choi et al., 2019). Similarly, there are other diseases where protein-lipid interactions alter normal function, such as in disruption of lipid metabolism in hepatitis C (Lee et al., 2020a); hence, there is an urgency to characterize their molecular mechanisms and understand the role of specific protein-lipid interactions in membrane remodeling as well as their relevance in the overall disease onset and progression.

MLKL is a pseudo-kinase with 469 residues distributed into three domains: the four helical bundle (4HB), residues 1-121; the brace, residues 133-175; and the pseudo-kinase domain (PsK), residues 193-459 (Zhang et al., 2016; Murphy, 2020; Petrie et al., 2020; Sethi et al., 2022a). MLKL is phosphorylated in preparation for the last step of necroptosis, currently considered a critical step in MLKL protein oligomerization. The 4HB domain of the oligomerized MLKL translocates to the PM and permeabilizes it; studies on MLKL lacking this domain show increase in cell viability (Zhang et al., 2021). The brace region consists of two helices and affects the interaction of the 4HB with the PM. Once the interaction between the brace and 4HB is disrupted (i.e., salt bridge between R30 and E136 breaks down), the 4HB interacts with and inserts in the PM (Su et al., 2014). Furthermore, decreasing the membrane binding of MLKL by inhibiting its S-acylation increases cell viability and restores membrane integrity (Parisi et al., 2019; Pradhan et al., 2021).

There is not yet a consensus on how the oligomerized MLKL permeabilizes the membrane. Some authors believe that it penetrates the membrane forming ion channels where cell content can leak (Zhang et al., 2021). However, other authors claim that, instead of forming ion channels, the 4HB forms cation channels or pores that allow cell content to flow (Xia et al., 2016). Furthermore, two additional models propose alternative mechanisms for membrane permeabilization, the carpet model and the toroidal pore model (Grage et al., 2016; Engelberg and Landau, 2020; Flores-Romero et al., 2020). Interestingly, the carpet model does not require the protein to cross the membrane. To increase our understanding of protein-lipid interactions in the context of mechanisms of cell death, we present an initial molecular dynamics study of a single MLKL protein with a complex lipid membrane model that mimics the environment of the PM. Our results suggest that binding of MLKL modulates lipid recruitment and can generate a unique lipid fingerprint enriched in phosphatidylinositol phosphates and cholesterol lipids at the protein binding site. These changes also affect the packing of lipids on the membrane surface of the binding leaflet, further modulating membrane surface topology and charge distribution. Proposing a final mechanism of membrane permeabilization is out of the scope of this work, which is intended as the first step in subsequent computational studies to characterize protein-lipid interactions in the context of MLKL-driven membrane remodeling and disruption.

Methods

Simulations setup

We used all-atom molecular dynamics simulations to model the interaction between a single MLKL protein (PDBID: 4BTF) and the PM as a starting point to characterize the molecular driving forces of late-stage necroptosis. The protein sequence corresponds to a murine model for MLKL, selected because its complete sequence of joint protein domains was available on the PDB server; on the contrary, the human MLKL tertiary structure is only available for separate domains on the PDB. **Supplementary Figure S1** shows the sequence alignment for the N-terminus of the protein, namely the 4HB and Brace domains, between the human (Uniprot: Q8NB16) and murine (PDBID: 4BTF) models showing excellent agreement between the structures.

The membrane model was based on the HT-29 cell line, built with a mixture of dioleoyl-phosphatidylcholine (DOPC): cholesterol (Chol): dioleoyl-phosphatidylethanolamine (DOPE): palmitoyl-oleoyl-phosphatidylinositol-4-phosphate (POPI-1,4): palmitoyl-oleoyl-phosphatidylinositol-(2,5)-bisphosphate (POPI-2,5) (40:32:20:4:4 mol%) to model the PM; hereon after, POPI-1,4 and POPI-2,5 are referred to as PIP and PIP₂. The membrane model was built using CHARMM-GUI *Membrane Builder* (Jo et al., 2007; Go and Jones, 2008; Jo et al., 2009; Cai et al., 2014; Lee et al., 2019), and the protein was solvated in a three-site water model using the *Solution Builder* (Lee et al., 2016; Lee et al., 2020b). The membrane model was built with 600 lipids per leaflet, fully hydrated with at least 50 water molecules per lipid. The default step-wise relaxation protocol from CHARMM-GUI was used for initial minimization and equilibration of the protein and membrane systems separately. Membrane-only systems were equilibrated for 200 ns, while the protein-water system was equilibrated for 50 ns before merging the equilibrated coordinates.

Upon equilibration, membrane and protein coordinates were merged and the simulation box rendered neutral using .15 mM KCl. The protein was positioned at different orientations above the lipid bilayer to ensure unbiased binding: (Rep1) vertical, with the pseudo-kinase domain facing membrane; (Rep2) vertical, with 4HB facing membrane; (Rep3) horizontal, with the brace facing away from the membrane; and (Rep4) horizontal, with the brace facing towards the membrane. **Supplementary Figure S2** illustrates these orientations relative to the membrane surface. **Table 1** summarizes the details of each system built for this study. The systems built for Rep1 and Rep2 are larger than Rep3 and Rep4 in terms of number of atoms and the *z* box vector. Rep1 and Rep2 started with the protein from a vertical conformation and had more water molecules to prevent the protein from interacting with image atoms from the bilayer during the simulation. Rep3 and Rep4 started with a horizontal protein, and were built in a smaller box to reduce the number of water molecules and reduce the computational cost. All systems were run with periodic boundary conditions and monitored to ensure no central atoms were interacting with its image atoms or with both membrane leaflets at the same time due to periodicity. The protein-membrane systems were run for 100ns to ensure they were equilibrated prior to transferring them to the Anton2 machine. The four replicas were run on this resource for at least 2,000 ns each, for a total of 8.76 μ s of simulated trajectory.

All systems were run using the CHARMM36m force field (Klauda et al., 2010; Vanommeslaeghe and MacKerell, 2012; Huang et al., 2017) and periodic boundary conditions. The Initial equilibration for

TABLE 1 Protein-membrane simulation systems.

System	# Water molecules	Total # atoms	Box cell size (x, y, z, in nm)	Sim. Time (ns)
Rep1	173,744	669,706	17.1 × 17.1 × 22.3	2,180
Rep2	173,750	669,724	17.2 × 17.2 × 21.9	2,180
Rep3	138,227	562,959	17.3 × 17.3 × 18.4	2,200
Rep4	135,548	554,910	17.3 × 17.3 × 18.2	2,200

the membrane-only and protein-only systems were performed using GROMACS (Abraham et al., 2015) with a timestep of 2 fs. Temperature was kept constant at 310.15 K using the Berendsen thermostat with a 1.0 ps coupling constant (Berendsen et al., 1984). Pressure was set at 1 bar and controlled semi-isotropically with the Berendsen barostat using a coupling time of 5.0 ps and compressibility of 4.5×10^{-5} (Berendsen et al., 1984). The merged protein-membrane systems were run with NPT dynamics, using the Nose-Hoover thermostat (Nosé, 1984; Hoover, 1985) and Parrinello-Rahman barostat (Parrinello and Rahman, 1981; Nosé and Klein, 1983); coupling and compressibility settings were kept as listed above. Non-bonded interactions were modeled using Verlet force-switch function with cutoffs set at 1.0 and 1.2 nm for Lennard-Jones interactions (Grubmüller et al., 1991). Particle Mesh Ewald was used for long-range electrostatics (Darden et al., 1993), and the LINCS algorithm (Hess et al., 1997) to constrain bonds with hydrogen atoms. The equilibration trajectories were run with resources available at the Center for Computational Research (CCR) at the University at Buffalo (Center for Computational Research UaB, 2019).

The production runs for each protein-membrane replica were computed on the Anton2 machine (Shaw et al., 2014a; Shaw et al., 2014b), hosted at the Pittsburgh Supercomputing Center (PSC). Simulation parameters were set by Anton2 internal guesser files, which are automated scripts designed to optimize the parameters for the integration algorithms of this machine. As such, the cut-off values to compute non-bonded interactions between neighboring atoms were set automatically during system preparation. Long-range electrostatics were computed using the Gaussian Split Ewald algorithm (Shan et al., 2005), and hydrogen bonds were constrained using the SHAKE algorithm (Ryckaert et al., 1977). Finally, the Nose-Hoover thermostat and MTK barostat (Martyyna et al., 1994) were used to control the temperature and pressure during NPT dynamics on Anton2 using optimized parameters set by the Multigrator integrator (Lippert et al., 2013) of the machine.

Trajectory analysis

We analyzed the trajectory primarily with VMD (William et al., 1996) and MDAnalysis (Michaud-Agrawal et al., 2011; Gowers et al., 2016). VMD was used to produce all the snapshots, and perform Hydrogen bond, RDFs, and packing defects analyses (Wildermuth et al., 2019). MDAnalysis was used to collect the raw data for the time series and histograms presented in this work, and in-house python scripts were used to further process the data and render all plots. Unless stated differently, all quantities are represented along with their standard error as computed from block averages during the listed time windows.

Cumulative plots were chosen to show lipid remodeling and recruitment by tracking the positions of the atoms in the lipid headgroup for a period of time. The size of this time window was determined to highlight differences between initial and final conditions on the membrane upon protein binding. The xyz coordinates were stored and accumulated for each lipid in one-nanosecond intervals for the determined time window. To show recruitment of inositol lipids, we rendered a scatter plot with a hue parameter of .5. To show height information, each z coordinate was selected and compared to the average z-coordinate of the first frame, z_o , to find the relative position, $z_f = z - z_o$. A scatter plot with z_f mapped to a color bar was plotted. Finally, to show the distribution of each lipid species per leaflet, we plotted a 2D density map in which the xy-plane was divided into a 2D grid and the number of points in each grid space was counted and plotted using a color bar.

Membrane lipids are free to move laterally, exposing regions of the hydrophobic core known as packing defects. These defects are enhanced when a protein binds and inserts into the membrane as it displaces lipids, making packing defects a good metric for protein insertion and membrane response. A method introduced by Wildermuth et al. measures the magnitude of the packing defects (Wildermuth et al., 2019). First, for a defined time window, images of the xy-plane are rendered with VMD. Hydrophobic atoms in the lipids are colored in yellow and the hydrophilic headgroups in blue. Supplementary Figure S3 shows a single-frame snapshot as an example; the yellow regions in the image correspond to packing defects. Multiple snapshots are taken from consecutive frames in the trajectory; an artificial intelligence algorithm for image analysis in the OpenCV python library identifies contours and measures their area. The code provided also allows for computation of the packing defect area underneath the protein, i.e. in the region delimited by the projection of the protein in the xy-plane (local packing defects). In this work, packing defects are used as a complementary measure for protein insertion in the binding leaflet.

Results

Protein binding conformation

MLKL binds the membrane within the first few hundred nanoseconds of simulation and remains bound the entire simulation. Figure 1A, shows the final bound conformation of the protein in each replica; despite the different initial orientations, common final bound states were found. Rep1 and Rep4 show a vertically bound conformation, with the PsK domain interacting with the membrane. Rep2 and Rep3 show the 4HB interacting with the membrane, which remains in contact with the bilayer until the end

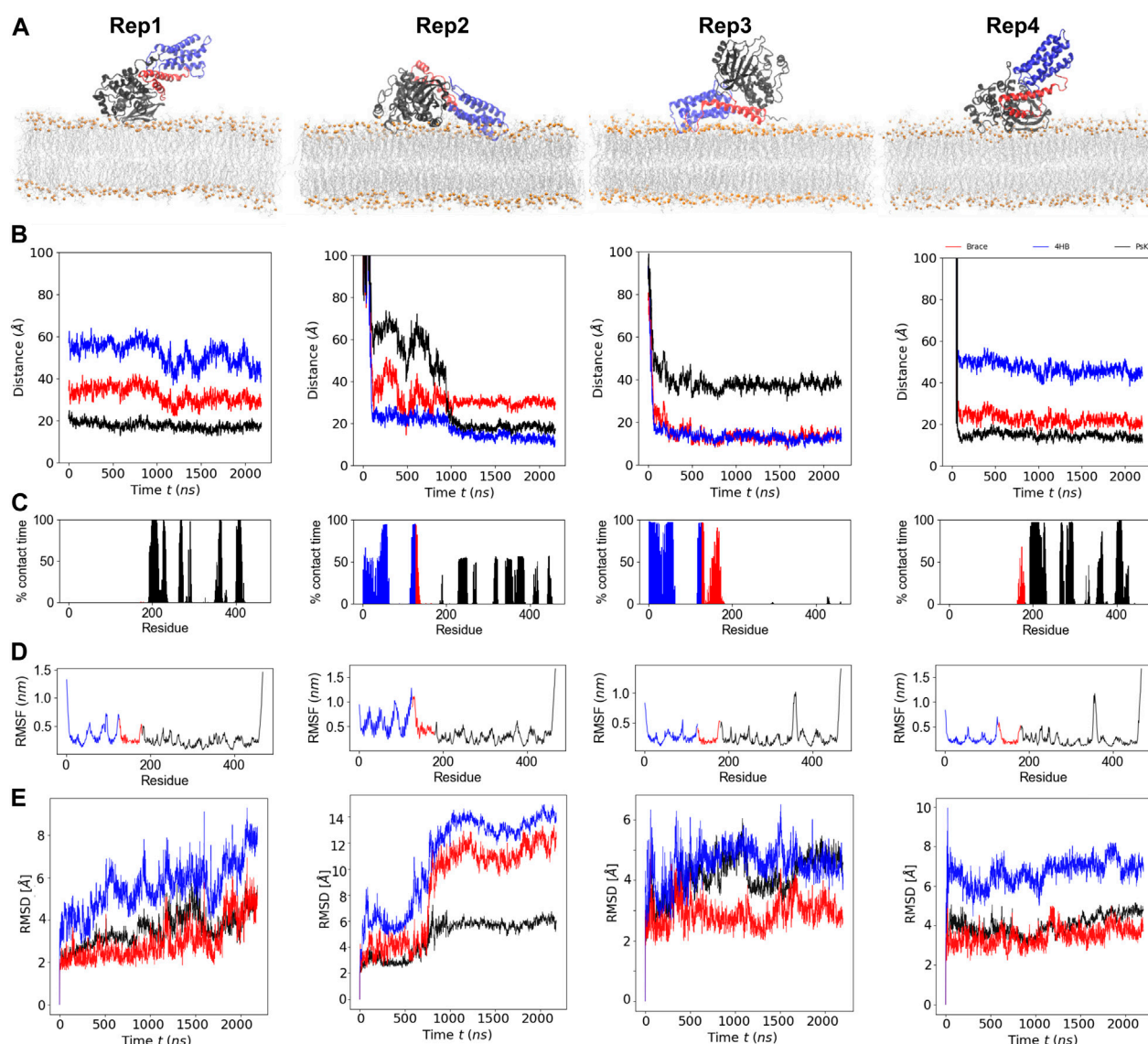


FIGURE 1

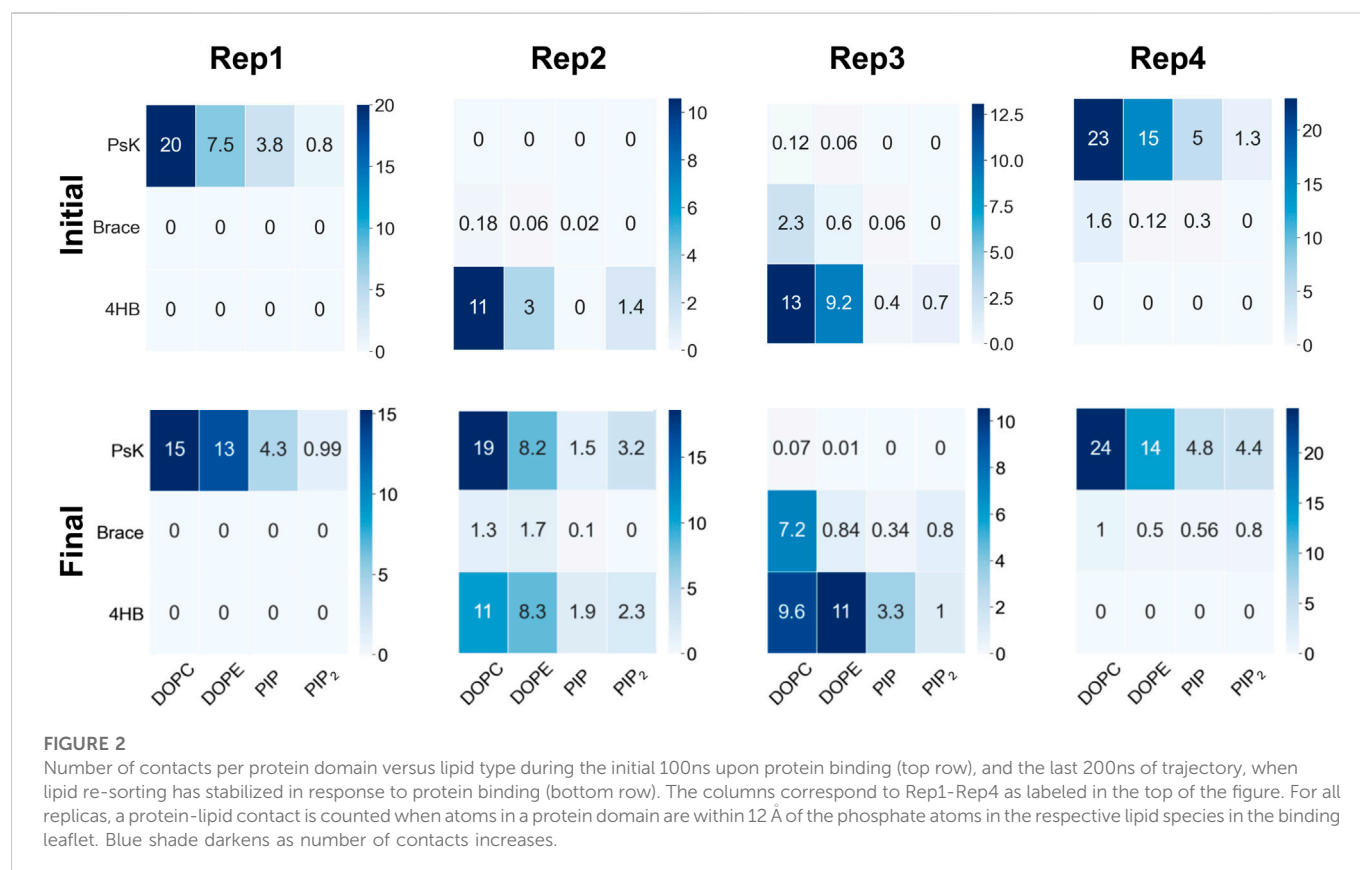
Binding conformation and dynamics of MLKL protein with a model membrane. (A) Final bound conformations of MLKL for each replica. The 4HB domain is shown in blue, the brace in red, and the PsK in black. Phosphate atoms of lipids are shown in orange, and fatty acid tails are shown in silver. (B) Center-of-mass distance of the C_{α} in each protein domain with respect to the phosphate groups in the binding leaflet. (C) Percent of time each residue in contact with any of the lipid species in the binding leaflet. (D) RMSF of protein residues. (E) Corresponding RMSD timeseries for each domain in the four replicas. Values corresponding to 4HB domain residues are indicated in blue, to the brace in red, and to the PsK in black in panels (B–E), matching the domain colors in panel (A).

of the simulation. However, the PsK domain comes in contact with the bilayer after the first microsecond of simulation in Rep2, for a final horizontal bound conformation.

Figure 1B, shows the time series of the distance between the center of mass (COM) of the individual protein domains and the phosphate groups (P atoms) of the lipids in the binding leaflet; the COM of each domain was computed from its C_{α} atoms. As expected from the final conformations in Figure 1A, the PsK is the closest to the membrane in Rep1 and Rep4, followed by the brace, and no interaction of the 4HB with the bilayer. On the other hand, the 4HB is the first to contact the membrane in Rep2, the brace interacts with the membrane in the first half of the simulation, but then remains pointing towards the water

when the PsK domain, shown in black, binds the membrane after the 1 μ s mark. Finally, Rep3 shows the 4HB and the brace both interact with the bilayer at the same plane, while the PsK positions towards the water, at nearly 180° with respect to its bound conformation on Rep1 and Rep4.

A residue is considered in contact with the membrane when its C_{α} is located within 12 Å of lipid phosphate groups in the binding leaflet; unless mentioned otherwise, this cutoff is used for all contact analyses. The frequency of contact analysis per protein residue during the entire trajectory is presented as % contact time in Figure 1C. Rep1 and Rep4 show similar trends for the % contact time, with corresponding residues in the PsK domain in contact with the bilayer in both replicas. Note the brace domain does interact with the bilayer intermittently in



Rep4, yet not permanently. Conversely, in Rep2 and Rep3, the frequency of contact is higher for residues in the 4HB domain. The main difference between these two replicas is that the PsK domain in Rep2 does interact with the membrane in the second half of the trajectory, adopting a fully horizontal position after 1 μ s of simulation. Some 4HB residues detach from the membrane as the PsK forms new contacts; **Figure 2**, Final stage, shows greater number of PsK-lipid contacts in Rep2 compared to PsK-lipid contacts in Rep3.

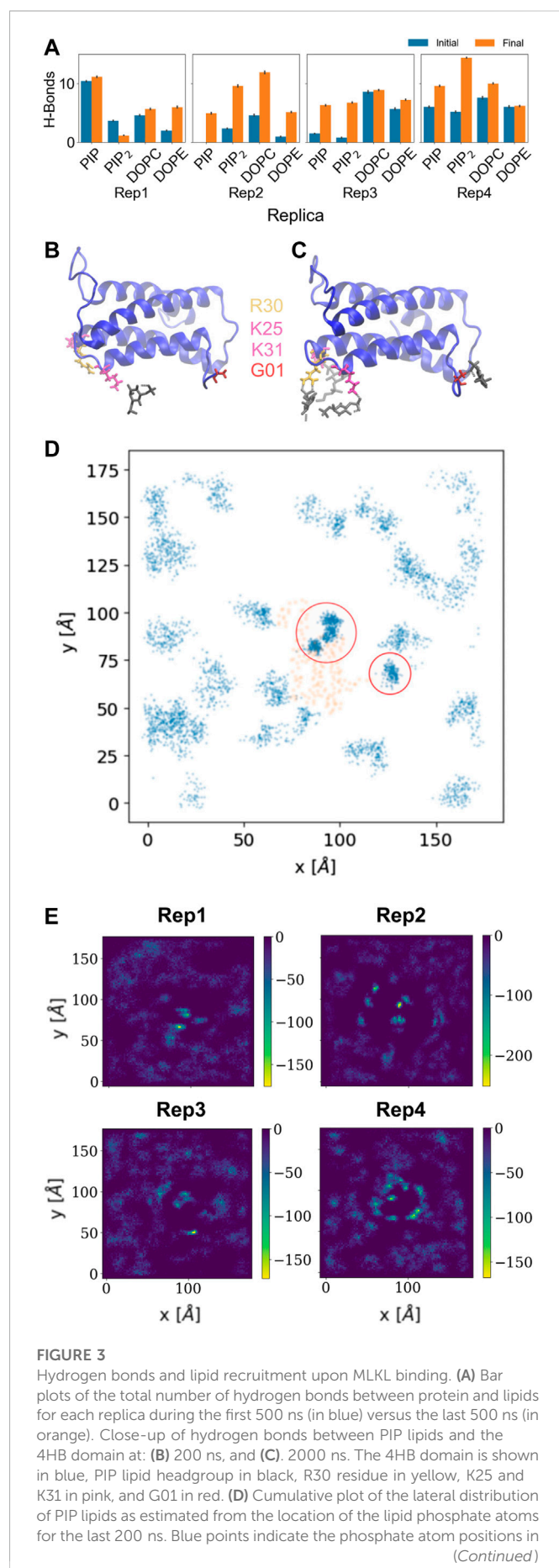
Figure 1D shows the root mean square fluctuations (RMSFs) of protein residues averaged over the total trajectory time. Rep3 and Rep4 exhibit a spike for residues surrounding residue 350 in the PsK domain (increased fluctuations in this region), while Rep2 exhibits larger fluctuations in the 4HB instead. Additionally, **Figure 1E** shows the root mean square displacement (RMSD) for each replica over the full trajectory. From this figure, it is evident there are no major conformational changes in the protein for Rep1, Rep3, and Rep4, which maintain their vertically bound conformation upon initial binding. Rep4 is the most stable of the four, as it barely changes over time with respect to its initial conformation. Rep1 experiences an increase in RMSD towards the end of the simulation, but the new conformation is stable for the last 200 ns of the trajectory. The PsK domain in Rep3 changes conformation after the first microsecond of the simulation, which was maintained for nearly 500 ns. There is a decrease in the RMSD of the black curve at this timepoint; however, the change is reverted and the RMSD returns to the value of the initial bound conformation for the rest of the trajectory.

The most interesting set of RMSD curves is that of Rep2 (**Figure 1E**), the only trajectory to exhibit both vertical and horizontal bound

conformations. Upon initial vertical binding by the 4HB, the protein is stable with no shifts in conformation for nearly 500 ns. There is a noticeable increase in the RMSD of all three protein domains between 700 and 1,000 ns timepoints in the trajectory. Of the three domains, the brace is the one to show the sharpest change in configuration as it interacts with the bilayer (see the red curve in this plot). Following the same trend, the PsK domain also has a large conformational change as the protein lays horizontally on the membrane surface. The next section discusses changes in the lateral organization of lipids in response to the bound protein.

Lipid contacts

As MLKL protein approaches the membrane, it interacts with specific lipid species in the binding leaflet, with a distinctive preference depending on the bound conformation. **Figure 2** shows contact heatmaps for each lipid species in our model upon initial protein binding, and during the last 200ns of trajectory: DOPC, DOPE, PIP, and PIP₂ with each protein domain (4HB, Brace, PsK). The overall number of contacts is higher with DOPC and DOPE in all cases, as expected, given their relative compositions in the membrane. PIP and PIP₂ have fewer total number of contacts with the protein due to their relative abundance compared to other lipid species. However, as summarized in **Figure 2**, PIP and PIP₂ co-localize to the protein binding site and increase their concentration rather notoriously. The following section expands on evidence of inositol recruitment to the protein binding site as evidenced by hydrogen bonding and 2D lipid density maps.



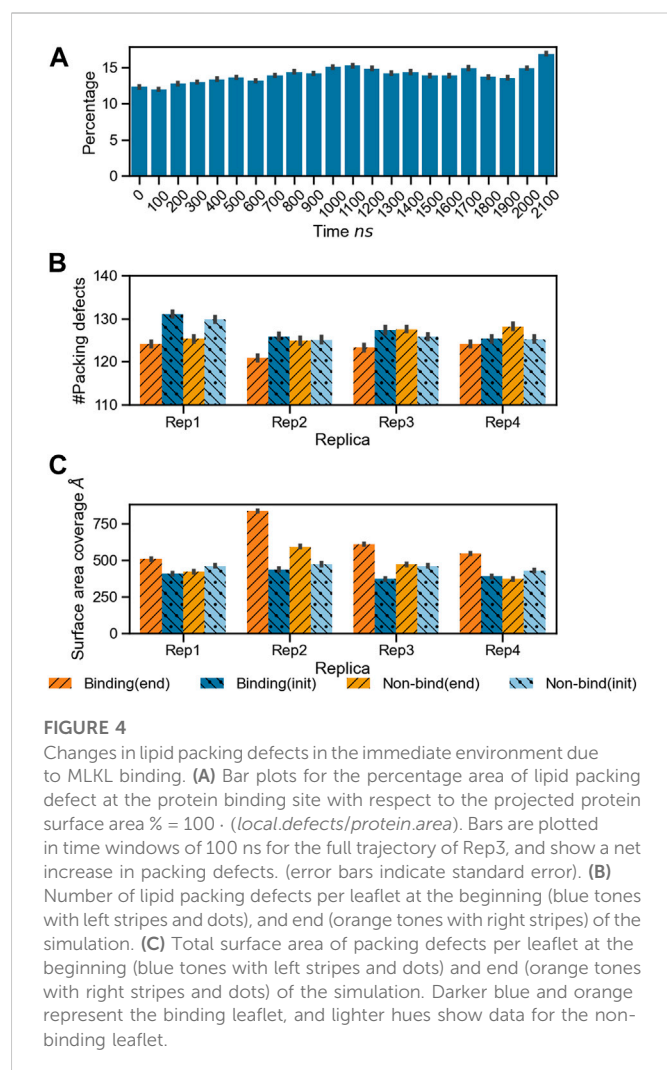
Protein-lipid interactions: Hydrogen bonding

The hydrogen bond analysis shown in [Figure 3A](#) was performed with a donor-acceptor distance of 3.2 Å and a cutoff angle of 30° on the initial and final 500 ns of the trajectory using VMD. [Figure 3A](#) and [Supplementary Table S1](#) summarize this analysis; the final number of hydrogen bonds between the protein and inositol lipids is highest for Rep1 and Rep4, compared to much lower net number of hydrogen bonds with DOPC or DOPE lipids. [Figure 3A](#) and [Supplementary Figure S4](#) show the number of hydrogen bonds increases consistently across replicas for inositol lipids, Rep1 being the one where PIP has the highest number of hydrogen bonds with the protein. Similarly, PIP₂ has larger number of hydrogen bonds as the simulation advances, except for Rep1; in all other cases, PIP₂ is the species with highest increase in hydrogen bonds as the simulation progresses. [Supplementary Figure S5A](#) shows final snapshots of all four replicas with PI lipids, shown in red and blue, and cholesterol, shown in yellow, underneath the protein. Taken together, these results suggest lipid resorting patterns upon protein binding leave a specific lipid fingerprint at the protein binding site.

[Figures 3B, C](#) show examples of residues that form hydrogen bonds with PIP lipids, most of which are positively charged arginine and lysine residues. Specifically, G01, K25, R30, and K31 are highlighted. [Figure 3D](#) further shows the 2D cumulative density of PIP lipids on the membrane plane over the last 200 ns. The red circles show regions with greater lipid density that match the location of the protein, shown as orange scatter. [Supplementary Figures S6D–F](#) show similar density maps for PIP lipids for the remaining replicas, where we find similar patterns. Furthermore, [Supplementary Figure S6G](#) shows the time progression of PI lipids under and around the protein binding site; enriched PI lipid regions are linked to charged regions as shown for all replicas in [Figure 3E](#). These plots show highly negative charged regions at and around the protein binding site, and correspond PIP and PIP₂ enriched zones. For instance, [Figure 3E](#) for Rep4 shows a charged ring that matches the binding site.

Membrane response: Lipid packing defects

As the protein interacts with the membrane, it influences the surface topology and lipid packing. We computed the lipid packing defects on the membrane prior to protein binding, and at the end of the simulation, when at least one microsecond of stable binding and subsequent lipid sorting around the protein has taken place. [Figure 4A](#) shows the percentage of surface area covered with lipid packing defects below the projected area of the protein during the trajectory for Rep3 (local packing defects, as described in the methods section). The area covered by



the packing defects underneath the protein increases over time, correlating with protein insertion as verified by depth of bound residues in the binding leaflet (see [Supplementary Figure S7](#)). [Figures 4B, C](#) show the number of lipid packing defects and respective surface area coverage per leaflet at the beginning and end of the simulation with their associated standard error. Interestingly, while Rep1 and Rep4 do not exhibit significant changes in the number of packing defects between leaflets at the beginning vs. the end of the trajectory, the surface area coverage does change, with larger values in the binding leaflet. This is accentuated in Rep2 and Rep3, which show the most interesting behavior in terms of bound conformation and insertion of the 4HB past the lipid headgroup region (see [Figures 1A, 7](#); [Supplementary Figure S7](#)). This fact is counterintuitive because the PsK domain, which binds the membrane in Rep1 and Rep4, is larger than the 4HB domain; yet, it does not insert as deep as 4HB ([Supplementary Figure S7](#)).

Membrane response: Surface topology

The protein bound conformation directly influences the local lipid environment and, consequently, the surface topology. [Figure 5](#) shows 2D histograms of the cumulative distribution of each lipid species per leaflet for the last 500 ns of the trajectory. Rep3 is shown as reference since it exhibits the deepest 4HB insertion across all replicas. This is in

agreement with multiple studies indicating the role of 4HB is for retention and insertion of MLKL into the PM ([Dondelinger et al., 2014](#); [Su et al., 2014](#); [Wang et al., 2014](#)). Each cumulative histogram was generated by mapping the membrane onto a grid and counting the number of phosphate groups in each zone. The DOPC map, for example, shows a more uniform distribution of these lipids in the non-binding leaflet; whereas, there is clear depletion of DOPC directly underneath the protein binding site (see corresponding plot in the bottom row). On the other hand, PIP and PIP₂, present at lower concentrations than DOPC, have a sharp increase around the protein site in the binding leaflet, shown in bright green/yellow in the map. This striking effect is also observed in cholesterol, which colocalizes underneath the protein binding site in the binding leaflet, and around the protein in the non-binding leaflet. Note that the cholesterol enrichment underneath the protein matches with the DOPC-depleted zone in the same leaflet. Similar 2D histograms for the other replicas are shown in [Supplementary Figure S8](#).

[Figure 6](#) shows changes in the lateral lipid organization through protein-lipid Radial Distribution Functions (RDF). These were calculated by using the C_α atom of the deepest inserted protein residue in each replica as a reference point (K268 for Rep1, Q53 for Rep2 and 3, and G406 for Rep4 – see [Supplementary Figure S7](#) for details on identifying these residues), and the phosphate or hydroxyl oxygen atoms of phospholipids and cholesterol, respectively. Each plot compares the RDF for the first 50 ns upon protein binding and the last 100 ns of the trajectory. Rep1 shows a slight increase in cholesterol and a noticeable decrease in PIP₂ near the bound protein at the end of the trajectory.

In most cases, the likelihood of observing cholesterol and PIP lipids under the protein or closer to the deepest inserted residue is higher at the end of the simulation compared to the beginning, which is also depicted in [Supplementary Figure S5A](#). The RDFs for DOPC and DOPE retain the location of the first solvation shell; DOPC experiences little to no change, but DOPE has higher relative abundance at the end in most of the cases. Rep2 exhibits the most interesting change for PIP₂ lipids, as the final RDF shows three distinctive shells. Note that Rep2 is the only replica that shows the protein interacting with the bilayer in a fully horizontal fashion, in which all the domains interact to some extent with the bilayer. The bottom row of [Figure 2](#) and [Supplementary Figure S5A](#) show PIP interaction with the protein. [Supplementary Figure S9](#) shows the corresponding RDF analysis for lipid-lipid interactions on the membrane plane of the binding leaflet. Results for all replicas show little to no change in the lateral distribution of DOPC, DOPE, and cholesterol species on the entire binding leaflet. However, the height and width of the solvation shells for PIP and PIP₂ species do change; in this case, Rep4 shows two distinct solvation shells for PIP and PIP₂ lipids at the end versus the beginning. Additionally, Rep2 shows an inward shift and higher probability for the first solvation shell of PI lipids at the simulation end versus the beginning.

Lipid sorting directly impacts the topology of the membrane surface; [Figure 7](#) shows the cumulative changes on the topology of the membrane surface during the last 500 ns of the trajectory for Rep2 and Rep3, respectively, as these replicas exhibit protein insertion past the lipid headgroup region. These changes are calculated using the first frame of the selected trajectory as the reference point. [Supplementary Figure S10](#) shows the corresponding plots for Rep1 and Rep4; the latter shows a small indentation of the PsK domain in the bilayer, but not as pronounced as the displacement of

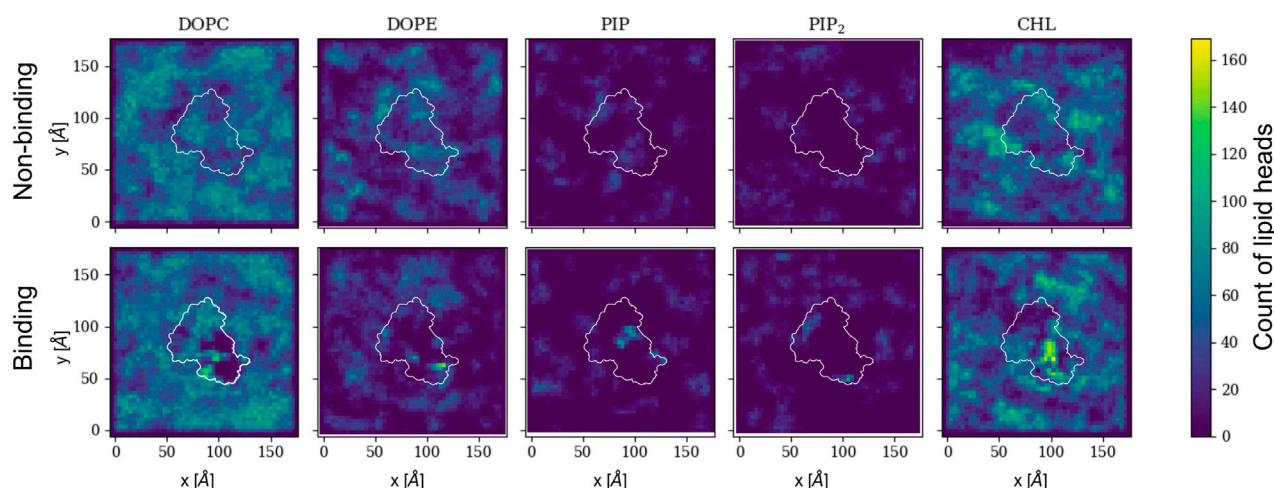


FIGURE 5

2D cumulative histograms for the last 500 ns of Rep3 trajectory. Top row corresponds to observations for non-binding leaflet, and the bottom row for the binding leaflet. White contours show a representative projection of the protein and the color bar is the cumulative number of lipids in each square of the 2D histogram, represented by either their P atom or O3 for cholesterol. Color intensity changes from dark blue to bright yellow as concentration of lipid head atoms increases.

lipids around the 4HB in Rep3. In Rep2 and Rep3, the 4HB domain inserts past the lipid headgroup region, and cholesterol can mitigate displacement of phospholipids by filling the space those lipids occupied without pushing the protein away.

Discussion and conclusion

Four replicas were run starting from different MLKL protein positions near a membrane model to characterize the protein binding mechanism and associated membrane response. Our results show MLKL is attracted to the membrane *via* electrostatic interactions. Then, as the protein binds the membrane, it remodels the local lipid environment by depleting DOPC, DOPE, and recruiting PIP, PIP₂, and cholesterol consistent with previously established experimental models (Dondelinger et al., 2014). Local remodeling of lipid composition depends, to a large extent, on the protein domain that binds the membrane. When the 4HB binds the membrane, it can insert past the phosphate region of the lipids, increasing the number of packing defects as it displaces the lipid headgroups and interacts with the hydrophobic core. Bound conformations with the 4HB interacting with the bilayer align to what has been proposed in the literature for MLKL during PM permeabilization.

Dondelinger et al. (2014) propose the 4HB as the executor of necroptosis, where the process is driven by interactions between highly conserved positive residues in the first two alpha helices in the 4HB and PIP lipids. More recently, experimental and simulation works suggest that the brace domain is an active player in the process of association of MLKL to lipid membranes (Yang et al., 2021; Sethi et al., 2022b). For example, it is reported that interactions between positively charged residues in MLKL and the membrane pull the brace away from 4HB for activation of this domain in human MLKL (Sethi et al., 2022b). Quarato et al propose a mechanism of initial recruitment of MLKL to the plasma membrane *via* low affinity interactions between positive residues on the helices of 4HB and membrane lipids, bringing the brace in closer proximity to the membrane – similar to what was observed in

Rep2 and Rep3 in this work. This unmasks further positive residues on the 4HB, leading to enhanced interaction with PIPs, also reproduced in our trajectories as shown in Figure 3, Supplementary Figures S4, S6. In our simulations, R30 exhibits hydrogen bonding with PIPs, this residue has also been identified as critical for binding to the membrane and stabilizing the interaction of the brace domain (Quarato et al., 2016). In line with these observations, Rep3 seems the most likely scenario to represent the interaction of MLKL with the plasma membrane in the cellular environment *via* both the 4HB and the brace.

The PsK domain is well known to interact with other proteins such as RIPK3 during necroptosis and act as a conformational-change switch that activates MLKL after undergoing phosphorylation (Petrie et al., 2017). Apart from this, not much is known about its interactions with membrane lipids, or if PsK-lipid interactions are relevant in the context of necroptosis and membrane disruption. Our results show Rep1 and Rep4 interact with the membrane through the PsK domain stably during the entire trajectory (see Figure 1E), and with Rep2 after 1 μ s. Our analyses examining the local lipid distribution, hydrogen bonding, and membrane response do not give direct strong evidence of a preferred binding domain. However, the number of contacts between each domain and specific lipid species does lead to a distinctive lipid fingerprint and local lipid distribution (see Figures 2, 3, 5; Supplementary Figures S9, S10). Based on the simulations presented here, it seems possible that a cooperative effect for MLKL binding and oligomerization could lead to membrane permeabilization; however, this event is not seen within the scope of our simulations in this study. Given that all 3 domains can interact with the membrane, it is possible that each domain contributes to specific lipid interactions that aid the process of membrane remodeling, and eventual bilayer disruption and permeabilization in a cooperative manner.

We observe the protein binds the membrane in all replicas well within the first 200 ns of simulation. Rep2 further exhibits a change in bound conformation after a microsecond of simulation and stable binding in a vertical conformation. The protein is able to turn and remain horizontally at the membrane interface with both the PsK and 4HB interacting with the lipids. In all replicas, the lipid distribution at the protein binding site

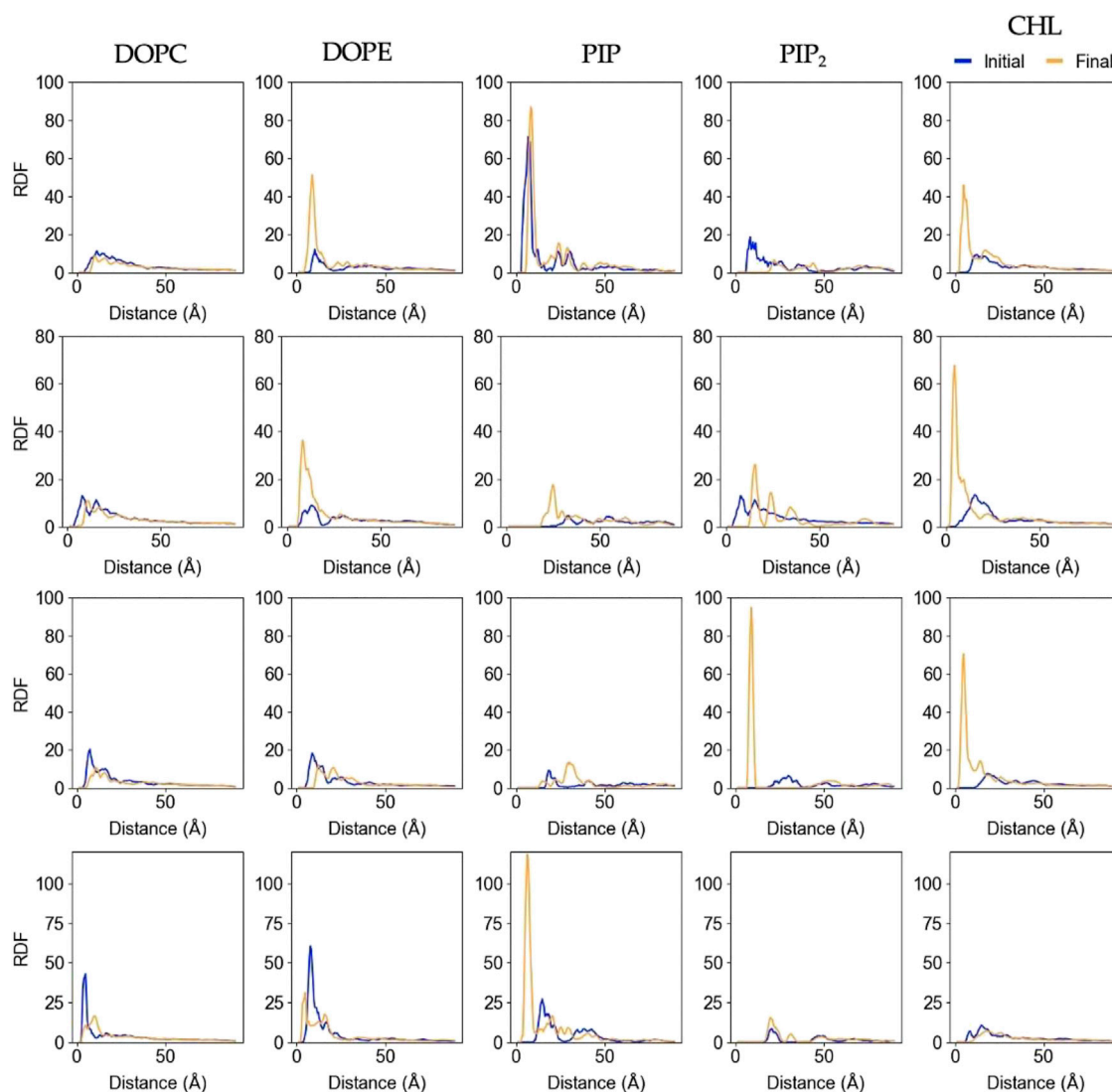


FIGURE 6

Lipid-Protein RDFs for each lipid species upon protein binding and at the end of the trajectory. The analysis was performed between the C_{α} atom of the deepest inserted protein residue in each replica (Rep1: K268, Rep2 and Rep3: Q53, and Rep4: G406), and the phosphorus (P) or hydroxyl oxygen (O3) of the lipids and cholesterol molecules, respectively. Initial curve is the averaged behavior during the first 50ns upon protein binding (shown in blue), and final is the average behavior over the last 100ns of simulation (shown in orange). Each row corresponds to Rep1-4, and each column corresponds to the lipid species listed at the top.

changed depending on the protein domain bound at the membrane. For example, Rep3 experiences a drastic change in local lipid composition when the brace domain contacts the membrane surface. The DOPE:PIP₂ ratio changes from 5:1 to ~1:1; additionally, at the 4HB-membrane contact site the DOPC:PIP₂ ratio changes from 10:1 to ~10:3. The initial ratios are based on the initial lipid composition, whereas the final ratios are extracted by counting the lipid species underneath the protein and determining their relative composition at the protein binding site. The change in lipid ratio is a clear indicator of local lipid redistribution directly modulated by the protein residues that bind the membrane. This distinct lipid fingerprint in the case of MLKL seems to result mainly from electrostatic interactions of positively charged residues and negatively charged lipid headgroups. The bottom panels in [Figure 3](#) further show a distinct distribution of charge around the protein, in the ring-like 2D density maps right at the edge of the protein binding site.

Some of the key charged residues that interact with the membrane are shown in [Figure 3](#). Notably, the positively charged residue K31, located in the second helix of the 4HB domain in the mouse model (4BTf) studied in this work, is conserved in human MLKL. Experiments with human MLKL have shown that the positively charged residues 22-35 are facilitators of MLKL oligomerization and recruitment to membranes as they interact with PIP lipids ([Dondelinger et al., 2014](#)). Our simulations with the murine protein model agree with a conserved behavior of these residues across both human and mice MLKL. There are differences in the report of relevant residues between the two orthologs; one work suggests that the mouse model associates with the membrane *via* residues found in the third and fourth helices of the 4HB, in contrast to the human model ([Sethi et al., 2022b](#)). While initial interactions of MLKL in our simulations are due to electrostatics, there is noticeable recruitment of PIP and PIP₂ lipids to the protein binding site, further stabilized by hydrogen bonding and

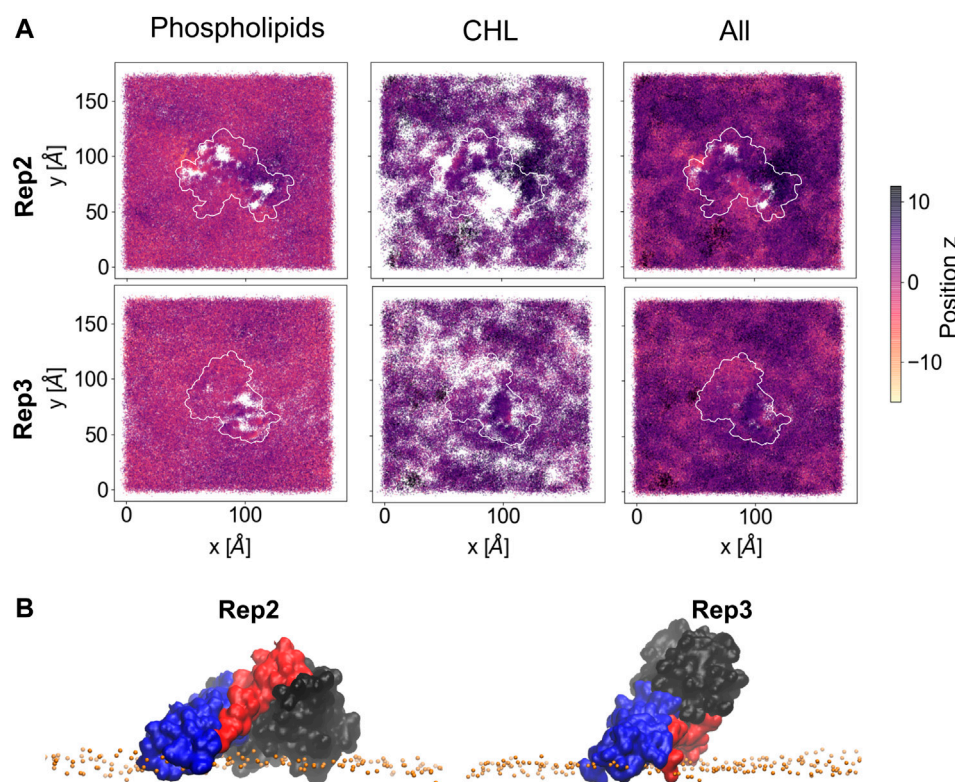


FIGURE 7

Membrane deformation due to protein binding. **(A)** Cumulative membrane height (z-coordinate) during the last 500 ns of simulation for Rep2 (top row) and Rep3 (bottom row). The color maps show the relative position of the lipid phosphate atoms (left), cholesterol hydroxyl group atoms (middle), and both types of atoms (right) on the binding leaflet. The relative positions are computed with respect to their initial coordinates in the analysis period (i.e., the first frame of the last 500ns of trajectory). Color intensity changes from pale yellow to dark purple as z position of atoms increase; white patches indicate absence of headgroup atoms. **(B)** Close up of the bound protein for Rep2 and Rep3, showing different domains inserted past the phosphate atoms in the membrane, shown as orange spheres. The 4HB is shown in blue, brace in red and PsK in black. Similar plots for Rep1 and Rep4 in [Supplementary Figure S10](#).

displacement of net neutral lipids like DOPC to the membrane bulk. This is shown in [Figure 5](#), the cumulative lipid density plots for different lipid species that show the regions where these are enriched or depleted on the membrane plane (Petrie et al., 2018; Murphy, 2020). The binding conformation of even a single protein is able to alter local lipid distribution, generating a distinctive lipid fingerprint and lateral organization patterns (see [Figure 7](#), RDFs). Cumulative plots of the lateral distribution of PIP lipids in [Figure 4D](#) and [Supplementary Figures S3D–F](#) further support this premise, showing stronger concentration of PIP lipids near the protein over time in agreement with experiment (Dondelinger et al., 2014).

The formation of a characteristic lipid fingerprint upon MLKL binding also impacts membrane lateral packing and surface topology, shown in [Figures 4, 7A](#), respectively. The relationship between protein binding and distribution of packing defects in the binding and non-binding leaflets is not trivial; we observe distinctive behavior for the different protein bound domains across our replicas. Packing defects underneath the protein were found to be significantly larger in Rep2 and Rep3 ([Figures 4B, C](#)), where the 4HB interacts with the membrane. This domain inserts past the phosphate region of the binding leaflet, resulting in a rearrangement of lipid packing. The overall number of lipid-packing defects decreases, but their overall surface coverage increases ([Figures 4B, C](#)); suggesting smaller packing defects merge into larger ones as lipid sorting and recruitment to the protein binding site progress. In contrast, the increase of packing defects surface area right below the protein in

Rep1 and Rep4 is rather subtle, and corresponds to a small or no insertion into the membrane. These results agree with previous observations in the literature that identified the 4HB as the killer domain (Dondelinger et al., 2014; Hildebrand et al., 2014).

From [Figures 5–7](#), it is evident that cholesterol is attracted to the protein in the binding leaflet, creating a distinctive fingerprint that differs from the non-binding leaflet. Accumulation of cholesterol is known to increase order in the membrane hydrophobic core and decrease membrane fluidity (Czub and Baginski, 2006). In the context of membrane permeabilization, accumulation of cholesterol under MLKL binding sites in the inner leaflet of the PM could potentially lead to a more fluid outer leaflet in the PM that allows easier permeation of small molecules around the protein or oligomers. Alternatively, the clustering of cholesterol near MLKL may be related to a necroptosis-independent role of the protein in lipid trafficking. Though there is little in literature that discusses the effects of cholesterol accumulation in the PM during necroptosis, it is relevant for intracellular membranes. Death of atherosclerotic lipid plaques is caused by cholesterol accumulation in the endoplasmic membrane, which triggers the unfolded protein response and in turn, apoptotic pathways (Tabas, 2004). Additional studies would help determine if lipid sorting due to MLKL binding follows a cooperative effect, in which more protein units are attracted to the initial protein binding site due to the local lipid composition remodeling. From our current studies, limited to a single MLKL near the membrane and time scales that did not show disruption of the

membrane, it seems plausible that oligomerization could be enhanced by lipid re-sorting caused by previous MLKL binding events (Flores-Romero et al., 2020).

Biochemical and lipidomic-based studies identified that phosphorylation of MLKL prior to plasma membrane association (Wang et al., 2014) or S-acylation of the protein can exacerbate membrane permeabilization (Parisi et al., 2017; Parisi et al., 2019; Pradhan et al., 2021); yet, how these modifications impact membrane permeability is not fully understood. The need of MLKL oligomers has been widely accepted in its mechanism to permeabilize the membrane during necroptosis; however, there are conflicting reports in the number of units that form the oligomer (Hildebrand et al., 2014). This work offers a basis for the study of membrane response and the specific lipid fingerprint that results upon binding of a peripheral membrane protein, specifically during initiation of MLKL driven mechanisms of cell death. The present work does not attempt to fully explain the process of protein-mediated permeabilization of the PM. Instead, it is geared towards characterizing the molecular mechanisms that may contribute to membrane remodeling and eventual disruption as a result of specific protein-lipid interactions. There is still much to explore in the context of MLKL-lipid interaction dynamics and how these shape the membrane surface topology, especially when multiple protein units are involved.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://www.rcsb.org/>- 4BTF and <https://www.uniprot.org/>- Q8NB16. The four simulation trajectories will be available upon publication on the Anton2 site, <https://antonweb.psc.edu/trajectories/Monje-Galvan/>. In house analysis scripts using Gromacs, VMD, and MDAnalysis are available upon request.

Author contributions

RR completed data collection and analysis, and prepare the manuscript. OC build the simulation systems, run the simulations on Anton2, and revise the manuscript. AP provided valuable contributions to comparing simulation data to experimental observations, and revised the final version of the manuscript.

References

- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1, 19–25. doi:10.1016/j.softx.2015.06.001
- Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81 (8), 3684–3690. doi:10.1063/1.448118
- Cai, Z., Jitkaew, S., Zhao, J., Chiang, H.-C., Choksi, S., Liu, J., et al. (2014). Plasma membrane translocation of trimerized MLKL protein is required for TNF-induced necroptosis. *Nat. Cell Biol.* 16 (1), 55–65. doi:10.1038/ncb2883
- Casares, D., Escibá, P. V., and Rosselló, C. A. (2019). Membrane lipid composition: Effect on membrane and organelle structure, function and compartmentalization and therapeutic avenues. *Int. J. Mol. Sci.* 20 (9), 2167. doi:10.3390/ijms20092167
- Center for Computational Research UaB (2019). CCR Facility Description 2019 [Available at: <https://ubir.buffalo.edu/xmlui/handle/10477/79221>].
- Chen, J., Kos, R., Garssen, J., and Redegeld, F. (2019). Molecular insights into the mechanism of necroptosis: The necrosome as a potential therapeutic target. *Cells* 8 (12), 1486. doi:10.3390/cells8121486
- Choi, M. E., Price, D. R., Ryter, S. W., and Choi, A. M. K. (2019). Necroptosis: A crucial pathogenic mediator of human disease. *JCI Insight* 4 (15), e128834. doi:10.1172/jci.insight.128834
- Corradi, V., Mendez-Villuendas, E., Inglfsson, H. I., Gu, R.-X., Siuda, I., Melo, M. N., et al. (2018). Lipid-protein interactions are unique fingerprints for membrane proteins. *ACS central Sci.* 4 (6), 709–717. doi:10.1021/acscentsci.8b00143
- Czub, J., and Baginski, M. (2006). Comparative molecular dynamics study of lipid membranes containing cholesterol and ergosterol. *Biophysical J.* 90 (7), 2368–2382. doi:10.1529/biophysj.105.072801
- Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98 (12), 10089–10092. doi:10.1063/1.464397
- Dondelinger, Y., Declercq, W., Montessuit, S., Roelandt, R., Goncalves, A., Bruggeman, I., et al. (2014). MLKL compromises plasma membrane integrity by binding to phosphatidylinositol phosphates. *Cell Rep.* 7 (4), 971–981. doi:10.1016/j.celrep.2014.04.026
- Engelberg, Y., and Landau, M. (2020). The Human LL-37 (17–29) antimicrobial peptide reveals a functional supramolecular structure. *Nat. Commun.* 11 (1), 3894. doi:10.1038/s41467-020-17736-x

GA-G and VM-G discussed the idea, guided the research, and edit the manuscript. All authors read and approved the final version.

Funding

This work was performed in part at the University at Buffalo's Center for Computational Research (Center for Computational Research UaB, 2019). Anton2 computer time was provided by the Pittsburgh Supercomputing Center (PSC) through Grant R01GM116961 from the National Institutes of Health, specific award MCB200093P. The Anton2 machine as PSC was generously made available by DE Shaw Research (Shaw et al., 2014b). OC was supported by the University at Buffalo Presidential Fellowship, and National Institute of Health's Initiative for Maximizing Student Development Training Grant 5T32GM144920-02 awarded to Dr. ML Dubocovich for the CLIMB program at the University at Buffalo.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fchem.2022.1088058/full#supplementary-material>

- D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, et al. (2014). *Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer*. SC '14: Proceedings of the international conference for high performance computing, networking, storage and analysis.
- Gowers, R. J., Linke, M., Barnoud, J., Reddy, T. J. E., Melo, M. N., Seyler, S. L., et al. (2016). "A Python package for the rapid analysis of molecular dynamics simulations," in *Proceedings of the 15th python in science conference*. Editors S. Benthall and S. Rostrup (Austin, TX: SciPy). doi:10.25080/majora-629e541a-00e
- Flores-Romero, H., Ros, U., and Garcia-Saez, A. J. (2020). Pore formation in regulated cell death. *EMBO J.* 39 (23), e105753. doi:10.15252/embj.2020105753
- Galluzzi, L., Kepp, O., Chan, F. K.-M., and Kroemer, G. (2017). Necroptosis: Mechanisms and relevance to disease. *Annu. Rev. Pathology Mech. Dis.* 12, 103–130. doi:10.1146/annurev-pathol-052016-100247
- Go, Y.-M., and Jones, D. P. (2008). Redox compartmentalization in eukaryotic cells. *Biochimica Biophysica Acta (BBA)-General Subj.* 1780 (11), 1273–1290. doi:10.1016/j.bbagen.2008.01.011
- Gong, Y., Fan, Z., Luo, G., Yang, C., Huang, Q., Fan, K., et al. (2019). The role of necroptosis in cancer biology and therapy. *Mol. cancer* 18 (1), 100–117. doi:10.1186/s12943-019-1029-8
- Grage, S. L., Afonin, S., Kara, S., Buth, G., and Ulrich, A. S. (2016). Membrane thinning and thickening induced by membrane-active amphipathic peptides. *Front. Cell Dev. Biol.* 4, 65. doi:10.3389/fcell.2016.00065
- Grubmüller, H., Heller, H., Windemuth, A., and Schulten, K. (1991). Generalized verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Mol. Simul.* 6 (1–3), 121–142. doi:10.1080/08927029108022142
- Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). Lincs: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18 (12), 1463–1472. doi:10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h
- Hildebrand, J. M., Tanzer, M. C., Lucet, I. S., Young, S. N., Spall, S. K., Sharma, P., et al. (2014). Activation of the pseudokinase MLKL unleashes the four-helix bundle domain to induce membrane localization and necroptotic cell death. *Proc. Natl. Acad. Sci.* 111 (42), 15072–15077. doi:10.1073/pnas.1408987111
- Hoover, W. G. (1985). Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* 31 (3), 1695–1697. doi:10.1103/physrev.31.1695
- Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., et al. (2017). CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* 14 (1), 71–73. doi:10.1038/nmeth.4067
- Jo, S., Kim, T., and Im, W. (2007). Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLOS ONE* 2 (9), e880. doi:10.1371/journal.pone.0000880
- Jo, S., Lim, J. B., Klauda, J. B., and Im, W. (2009). CHARMM-GUI membrane builder for mixed bilayers and its application to yeast membranes. *Biophysical J.* 97 (1), 41a–48a. doi:10.1016/j.bpj.2008.12.109
- Kandt, C., Mtyus, E., and Tieleman, D. P. (2008). Protein lipid interactions from a molecular dynamics simulation point of view. *Struct. Dyn. Membr. Interfaces*, 267–282. doi:10.1002/9780470388495.ch10
- Klauda, J. B., Venable, R. M., Freitas, J. A., O'Connor, J. W., Tobias, D. J., Mondragon-Ramirez, C., et al. (2010). Update of the CHARMM all-atom additive force field for lipids: Validation on six lipid types. *J. Phys. Chem. B* 114 (23), 7830–7843. doi:10.1021/jp101759q
- Lee, H.-R., Lee, G. Y., You, D.-G., Kim, H. K., and Yoo, Y. D. (2020). Hepatitis C virus p7 induces membrane permeabilization by interacting with phosphatidylserine. *Int. J. Mol. Sci.* 21 (3), 897. doi:10.3390/ijms21030897
- Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., et al. (2016). CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.* 12 (1), 405–413. doi:10.1021/acs.jctc.5b00935
- Lee, J., Hitznerberger, M., Rieger, M., Kern, N. R., Zacharias, M., and Im, W. (2020). CHARMM-GUI supports the Amber force fields. *J. Chem. Phys.* 153 (3), 035103. doi:10.1063/5.0012280
- Lee, J., Patel, D. S., Stähle, J., Park, S.-J., Kern, N. R., Kim, S., et al. (2019). CHARMM-GUI membrane builder for complex biological membrane simulations with glycolipids and lipoglycans. *J. Chem. Theory Comput.* 15 (1), 775–786. doi:10.1021/acs.jctc.8b01066
- Lippert, R. A., Predescu, C., Ierardi, D. J., Mackenzie, K. M., Eastwood, M. P., Dror, R. O., et al. (2013). Accurate and efficient integration for molecular dynamics simulations at constant temperature and pressure. *J. Chem. Phys.* 139 (16), 164106. doi:10.1063/1.4825247
- Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* 101 (5), 4177–4189. doi:10.1063/1.467468
- Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* 32 (10), 2319–2327. doi:10.1002/jcc.21787
- Monje-Galvan, V., and Klauda, J. B. (2018). Preferred binding mechanism of Osh4's amphipathic lipid-packing sensor motif, insights from molecular dynamics. *J. Phys. Chem. B* 122 (42), 9713–9723. doi:10.1021/acs.jpcc.8b07067
- Murphy, J. M. (2020). The killer pseudokinase mixed lineage kinase domain-like protein (MLKL). *Cold Spring Harb. Perspect. Biol.* 12 (8), a036376. doi:10.1101/cshperspect.a036376
- Nosé, S. (1984). A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* 52 (2), 255–268. doi:10.1080/00268978400101201
- Nosé, S., and Klein, M. L. (1983). Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* 50 (5), 1055–1076. doi:10.1080/00268978300102851
- Parisi, L. R., Li, N., and Atilla-Gökçumen, G. E. (2017). Very long chain fatty acids are functionally involved in necroptosis. *Cell Chem. Biol.* 24 (12), 1445–1454.e8. e8. doi:10.1016/j.chembiol.2017.08.026
- Parisi, L. R., Sowlati-Hashjin, S., Berhane, I. A., Galster, S. L., Carter, K. A., Lovell, J. F., et al. (2019). Membrane disruption by very long chain fatty acids during necroptosis. *ACS Chem. Biol.* 14 (10), 2286–2294. doi:10.1021/acscchembio.9b00616
- Parrinello, M., and Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* 52 (12), 7182–7190. doi:10.1063/1.328693
- Petrie, E. J., Birkinshaw, R. W., Koide, A., Denbaum, E., Hildebrand, J. M., Garnish, S. E., et al. (2020). Identification of MLKL membrane translocation as a checkpoint in necroptotic cell death using Monobodies. *Proc. Natl. Acad. Sci.* 117 (15), 8468–8475. doi:10.1073/pnas.1919960117
- Petrie, E. J., Hildebrand, J. M., and Murphy, J. M. (2017). Insane in the membrane: A structural perspective of MLKL function in necroptosis. *Immunol. Cell Biol.* 95 (2), 152–159. doi:10.1038/icb.2016.125
- Petrie, E. J., Sandow, J. J., Jacobsen, A. V., Smith, B. J., Griffin, M. D. W., Lucet, I. S., et al. (2018). Conformational switching of the pseudokinase domain promotes human MLKL tetramerization and cell death by necroptosis. *Nat. Commun.* 9 (1), 2422. doi:10.1038/s41467-018-04714-7
- Pradhan, A. J., Lu, D., Parisi, L. R., Shen, S., Berhane, I. A., Galster, S. L., et al. (2021). Protein acylation by saturated very long chain fatty acids and endocytosis are involved in necroptosis. *Cell Chem. Biol.* 28 (9), 1298–1309.e7. e7. doi:10.1016/j.chembiol.2021.03.012
- Qin, X., Ma, D., Tan, Y.-X., Wang, H.-Y., and Cai, Z. (2019). The role of necroptosis in cancer: A double-edged sword? *Biochimica Biophysica Acta (BBA)-Reviews Cancer* 1871 (2), 259–266. doi:10.1016/j.bbcan.2019.01.006
- Quarato, G., Guy Cliff, S., Grace Christy, R., Llambi, F., Nourse, A., Rodriguez Diego, A., et al. (2016). Sequential engagement of distinct MLKL phosphatidylinositol-binding sites executes necroptosis. *Mol. Cell* 61 (4), 589–601. doi:10.1016/j.molcel.2016.01.011
- Ryckaert, J.-P., Cicotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* 23 (3), 327–341. doi:10.1016/0021-9991(77)90098-5
- Sapay, N., and Tieleman, D. P. (2008). Molecular dynamics simulation of lipid-protein interactions. *Curr. Top. Membr.* 60, 111–130.
- Sethi, A., Horne, C. R., Fitzgibbon, C., Wilde, K., Davies, K. A., Garnish, S. E., et al. (2022). Membrane permeabilization is mediated by distinct epitopes in mouse and human orthologs of the necroptosis effector. *MLKL. Cell Death Differ.*, 1–12.
- Sethi, A., Horne, C. R., Fitzgibbon, C., Wilde, K., Davies, K. A., Garnish, S. E., et al. (2022). Membrane permeabilization is mediated by distinct epitopes in mouse and human orthologs of the necroptosis effector, MLKL. *MLKL. Cell Death Differ.* 29 (9), 1804–1815. doi:10.1038/s41418-022-00965-6
- Shan, Y., Klepeis, J. L., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2005). Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys.* 122 (5), 054101. doi:10.1063/1.1839571
- Shaw, D. E., Grossman, J. P., Bank, J. A., Batson, B., Butts, J. A., Chao, J. C., et al. (2014). *Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer*, 41–53.
- Su, L., Quade, B., Wang, H., Sun, L., Wang, X., and Rizo, J. (2014). A plug release mechanism for membrane permeation by MLKL. *Structure* 22 (10), 1489–1500. doi:10.1016/j.str.2014.07.014
- Tabas, I. (2004). Apoptosis and plaque destabilization in atherosclerosis: The role of macrophage apoptosis induced by cholesterol. *Cell Death Differ.* 11 (1), S12–S16. doi:10.1038/sj.cdd.4401444
- Vanommeslaeghe, K., and MacKerell, A. D. (2012). Automation of the CHARMM general force field (CGenFF) I: Bond perception and atom typing. *J. Chem. Inf. Model.* 52 (12), 3144–3154. doi:10.1021/ci300363c

- Wang, H., Sun, L., Su, L., Rizo, J., Liu, L., Wang, L-F., et al. (2014). Mixed lineage kinase domain-like protein MLKL causes necrotic membrane disruption upon phosphorylation by RIP3. *Mol. Cell* 54 (1), 133–146. doi:10.1016/j.molcel.2014.03.003
- Wang, T., Jin, Y., Yang, W., Zhang, L., Jin, X., Liu, X., et al. (2017). Necroptosis in cancer: An angel or a demon? *Tumor Biol.* 39 (6), 101042831771153. doi:10.1177/1010428317711539
- Wildermuth, K. D., Monje-Galvan, V., Warburton, L. M., and Klauda, J. B. (2019). Effect of membrane lipid packing on stable binding of the ALPS peptide. *J. Chem. Theory Comput.* 15 (2), 1418–1429. doi:10.1021/acs.jctc.8b00945
- William, H., Andrew, D., and Klaus, S. (1996). Vmd: Visual molecular dynamics. *J. Mol. Graph.* 14 (1), 33–38. doi:10.1016/0263-7855(96)00018-5
- Xia, B., Fang, S., Chen, X., Hu, H., Chen, P., Wang, H., et al. (2016). MLKL forms cation channels. *Cell Res.* 26 (5), 517–528. doi:10.1038/cr.2016.26
- Yang, Y., Xie, E., Du, L., Yang, Y., Wu, B., Sun, L., et al. (2021). Positive charges in the brace region facilitate the membrane disruption of MLKL-NTR in necroptosis. *Mol. [Internet]* 26 (17), 5194. doi:10.3390/molecules26175194
- Zhang, J., Yang, Y., He, W., and Sun, L. (2016). Necrosome core machinery: Mkl. *Cell. Mol. Life Sci.* 73 (11), 2153–2163. doi:10.1007/s00018-016-2190-5
- Zhang, Y., Liu, J., Yu, D., Zhu, X., Liu, X., Liao, J., et al. (2021). The MLKL kinase-like domain dimerization is an indispensable step of mammalian MLKL activation in necroptosis signaling. *Cell Death Dis.* 12 (7), 638. doi:10.1038/s41419-021-03859-6



OPEN ACCESS

EDITED BY
Simón Poblete,
Universidad Austral de Chile, Chile

REVIEWED BY
Marco Zoli,
University of Camerino, Italy
Antonio Suma,
University of Bari Aldo Moro, Italy

*CORRESPONDENCE
Peter Virnau,
✉ virnau@uni-mainz.de

SPECIALTY SECTION
This article was submitted to Theoretical
and Computational Chemistry,
a section of the journal
Frontiers in Chemistry

RECEIVED 11 November 2022
ACCEPTED 23 December 2022
PUBLISHED 17 January 2023

CITATION
Wettermann S, Datta R and Virnau P (2023),
Influence of ionic conditions on knotting in
a coarse-grained model for DNA.
Front. Chem. 10:1096014.
doi: 10.3389/fchem.2022.1096014

COPYRIGHT
© 2023 Wettermann, Datta and Virnau.
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.

Influence of ionic conditions on knotting in a coarse-grained model for DNA

Sarah Wettermann, Ranajay Datta and Peter Virnau*

Institut für Physik, Johannes Gutenberg-Universität, Mainz, Germany

We investigate knotting probabilities of long double-stranded DNA strands in a coarse-grained Kratky-Porod model for DNA with Monte Carlo simulations. Various ionic conditions are implemented by adjusting the effective diameter of monomers. We find that the occurrence of knots in DNA can be reinforced considerably by high salt conditions and confinement between plates. Likewise, knots can almost be dissolved completely in a low salt scenario. Comparisons with recent experiments confirm that the coarse-grained model is able to capture and quantitatively predict topological features of DNA and can be used for guiding future experiments on DNA knots.

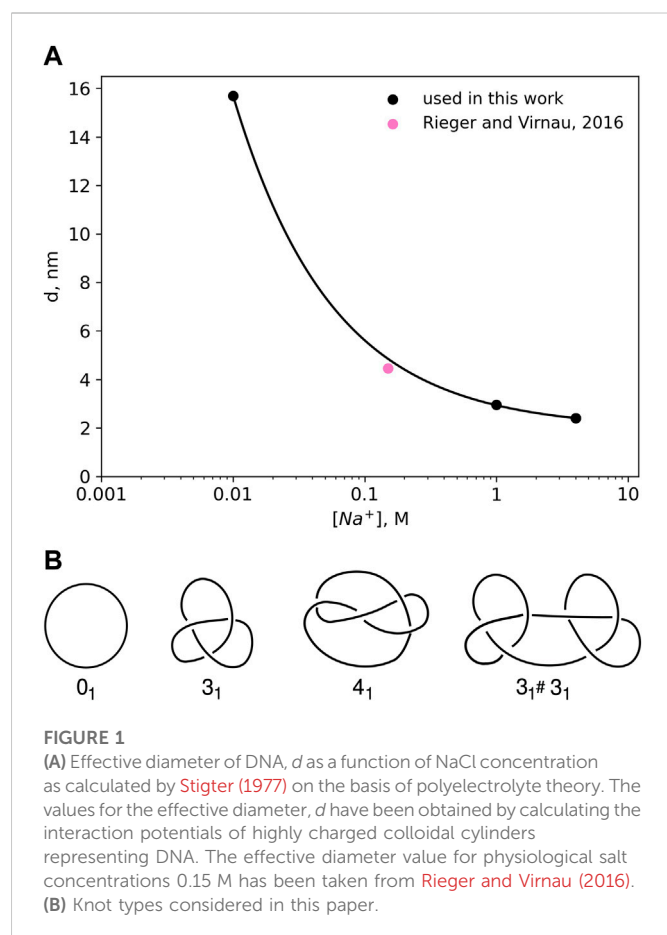
KEYWORDS

polymers, ionic conditions, knots, coarse-grained model, DNA

Introduction

The revelation of DNA packing and folding in the cell nucleus (Lieberman-Aiden et al., 2009; Siebert et al., 2017; Stevens et al., 2017; Ganji et al., 2018) and the emergence of commercially available nanopore techniques (Jain et al., 2016; Jain et al., 2018) has ushered in a new era of DNA research in the past decade. Knots, which emerge naturally as a byproduct in long macromolecules like DNA (Frisch and Wasserman, 1961; Delbrück, 1962), may however be detrimental to biological processes and technical applications. It is therefore of prime importance to study conditions and length scales at which they appear in equilibrium and develop strategies to enhance (Lua et al., 2004; Virnau et al., 2005; Tang et al., 2011; Amin et al., 2018) or suppress knotting (Di Stefano et al., 2014; Renner and Doyle, 2014). From a technical point of view, numerical simulations are a great tool for this task as structural and topological information are readily available. Coarse-grained models are particularly relevant as knots appear at scales beyond the Kuhn length and models with atomistic resolution are often poorly suited for efficient Monte Carlo algorithms required to scan configuration space. It is therefore crucial to test and improve coarse-grained models for DNA to quantitatively support and interpret experimental efforts.

A first link to double-stranded (ds) DNA was already established in the first simulation paper on polymer knots from 1974 (Vologodskii et al., 1974). In their seminal contribution, Vologodskii et al. determined knotting probabilities of random walks and associated single segments with the Kuhn length of DNA (100 nm)—a prediction which turns out to be surprisingly accurate as we will demonstrate later. This basic approach has been refined further in the early 1990s in conjunction with gel electrophoresis experiments on short DNA strands of up to 10 kbp (Rybenkov et al., 1993; Shaw and Wang, 1994). Ideal segments were replaced by cylinders with excluded volume interactions that depend on ionic conditions (Rybenkov et al., 1993), and it was also demonstrated that DNA knotting probabilities vary somewhat with solvent conditions (reaching about 4% in a high salt environment.) Higher resolution versions of this model in which one Kuhn length is represented by several segments have been used to study the effect of confinement on short strands in high salt conditions. Among other things, Orlandini, Micheletti and coworkers have



demonstrated with numerical simulations that confining DNA between plates or in nanopores increases knotting probabilities when typical distances between plates or nanopore diameters are in the order of the Kuhn length of DNA (Micheletti and Orlandini, 2012a; Micheletti and Orlandini, 2012b; Orlandini and Micheletti, 2013). Alternatively, coarse-grained bead-stick (Dai et al., 2012a; Dai et al., 2012b; Rieger and Virnau, 2016) or bead-spring (Trefz et al., 2014; Rothörl et al., 2022) representations for DNA can be used in which the effective diameter is adjusted to account for varying solvent conditions and which can be adapted for molecular dynamics simulations. Of particular relevance to our study is Dai et al. (2012b) in which the authors have studied knotting of closed DNA rings in bulk and plate geometry and to which our results for open strands can be compared. Variants of this model class have also been applied to investigate, e.g., statics (Dai et al., 2015; Jain and Dorfman, 2017) and dynamics of DNA knots in a nanochannel (Micheletti and Orlandini, 2014), packing of DNA in viral capsids (Marenduzzo et al., 2009; Reith et al., 2012) and recently for the reproduction of experimental knotting probabilities of λ phage DNA in high salt conditions (Kumar Sharma et al., 2019). Of course, there are also limits to this class of coarse-grained descriptions, and higher resolution models (Suma and Micheletti, 2017; Suma et al., 2018) may address questions which either require a detailed structural description or an explicit modelling of electrostatic interactions (Suma et al., 2018).

In this work we systematically extend previous analyses to DNA lengths relevant to modern experiments on λ (Plesa et al., 2016; Kumar Sharma et al., 2019) and T4 phages (Plesa et al., 2016). Our comprehensive study also covers the full range of ionic conditions

for free DNA and DNA confined between two plates, and comparisons with existing experimental data confirm the validity of the modelling approach. This enables us to show, amongst others, that for the considered strand sizes the dependence of knotting on salt concentrations (Rybenkov et al., 1993) can be used to effectively disentangle DNA prior to experiments where knots are undesired.

Methods

Implicit modelling of ionic solvent conditions

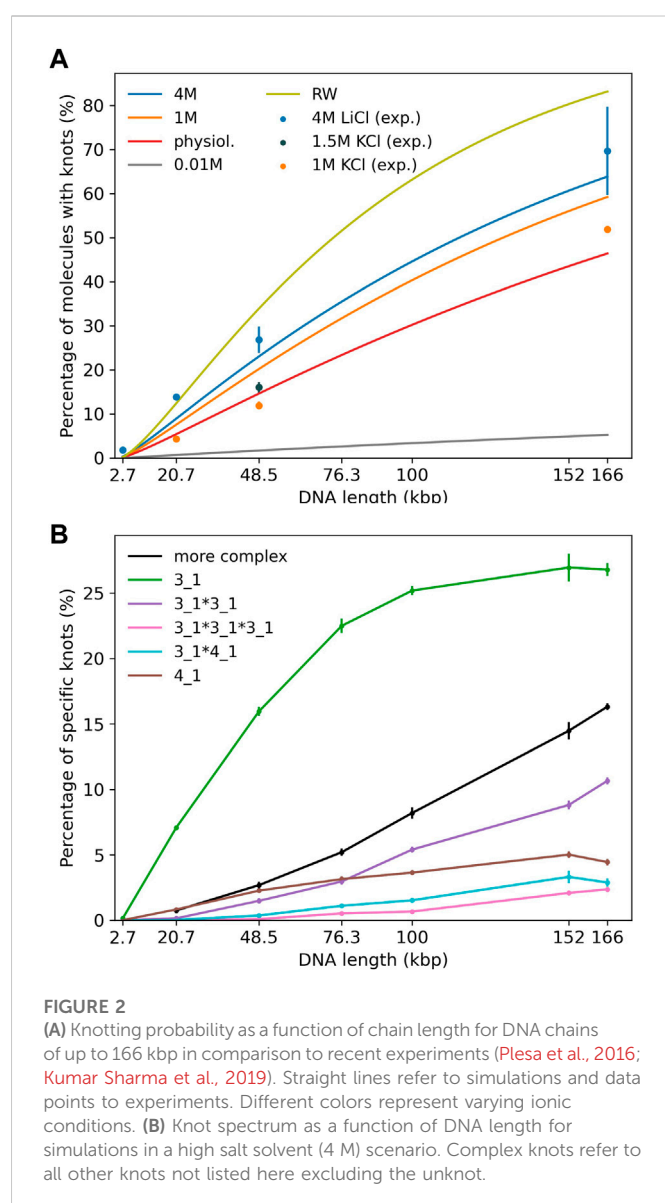
DNA is negatively charged, but long-range electrostatic interactions are partially or completely screened by counterions in solution. In this paper we follow an implicit solvent approach pioneered by Stigter (1977) and Rybenkov et al. (1993) in which screened charges are represented by effective excluded volume interactions. The diameter d of a DNA chain is a parameter that quantifies the latter and can be defined as the segment diameter of a representative chain which is uncharged, but has the same configurational and morphological properties as the original DNA with partial or completely screened charges. The magnitude of the electrostatic repulsion, and consequently, the numerical value of d , is a function of salt concentration. Stigter (1977) modeled DNA in sodium chloride solution as charged cylinders. Following the theory developed by McMillan and Mayer (1945) and the calculations of Hill (Hill, 1956; Hill, 1960), Stigter carried out analytic calculations to estimate the effective diameter of DNA as a function of sodium chloride concentration (see Figure 1A).

Already in 1993 Rybenkov et al. (1993) were able to confirm this approach (and Figure 1A) by representing DNA as a closed chain of cylinders of Kuhn length 100 nm and by matching experimentally determined knotting probabilities of a short P4 phage DNA strand (of around 10,000 base pairs) with those obtained from Monte Carlo simulations.

Here, we use a higher resolution variation of this ansatz which models DNA as a standard bead-stick chain and also resolves local structure at the scale of the persistence length (which according to Kratky-Porod theory is half of the Kuhn length). We keep, however, the same effective diameter to determine knotting probabilities in various ionic conditions for long, experimentally relevant DNA strands (like λ phage or T4). In a previous work (Rieger and Virnau, 2016), we have already validated this approach by determining simulation parameters for physiological conditions (0.15M) that reproduce experimental knotting spectra of short strands from Rybenkov et al. (1993) and Shaw and Wang (1994) even without making assumptions about the persistence or Kuhn length. Not only did these simulations confirm a value for d which is close to the value of Stigter (pink point in Figure 1A), they also confirmed the correct persistence length of DNA. While we use $d = 4.465$ nm for physiological conditions, values for other ionic conditions are directly taken from Figure 1A.

Bead-stick model. Simulations were performed using a discrete Kratky-Porod model (Kratky and Porod, 1949; Dai et al., 2012b; Marenz and Janke, 2016; Rieger and Virnau, 2016) with hard sphere interactions between monomers and a constant distance between adjacent beads. For simulations in slit confinement, walls are also hard and impenetrable. Chain stiffness is implemented *via* a bond-bending potential:

$$U = \kappa \sum_i (1 - \cos \theta_i) \quad (1)$$



where θ_i for $i = 1, \dots, N-1$ are the angles between adjacent bond vectors. Simulations were performed at various salt concentrations with values for d obtained from Figure 1A. We assume a persistence length l_p of 50 nm or 150 base pairs (bp) for all considered salt concentrations. For a Kratky-Porod chain the stiffness parameter κ for any given effective diameter d can be computed as (Fisher, 1964; Trefz et al., 2014; Rieger and Virnau, 2016)

$$\kappa \approx \frac{l_p \cdot k_B T}{d} \quad (2)$$

In dsDNA the distance between adjacent base pairs is 1/3 nm. By comparing the contour lengths, we conclude that a DNA strand with B base pairs is represented by a chain of

$$N \approx B \cdot 0.3333 \text{ nm}/d \quad (3)$$

beads.

Several simplifications are implied in this approach. The dependence of persistence length on ionic conditions was neglected as differences only amount to a few percent at least in the formalism of Odijk (1977);

Skolnick and Fixman (1977). Note, however, that for small DNA strands (up to several kilo bases) and high salt conditions the persistence length can be significantly smaller ($\approx 30\text{--}35 \text{ nm}$), (Kam et al., 1981; Manning, 1981; Post, 1983; Savelyev, 2012; Brunet et al., 2015; Rieger and Virnau, 2018) and also depends on the specific ions in the solvent (Brunet et al., 2015) (the influence of which we neglect as well). Nevertheless, for larger chains (such as those simulated in our paper) persistence length is expected to increase again and might actually be closer to 50 nm . In high salt conditions knotting probabilities also depend little on the actual value of the persistence length as demonstrated in Supplementary Information, which taken together justifies our simplified assumptions.

Experiments (Plesa et al., 2016; Kumar Sharma et al., 2019) displayed in Figure 2A use either KCl (1 and 1.5 M) or LiCl (4 M) as buffer. In our simulations we mainly study DNA strands of lengths 20,678 bp, 48,502 bp and 165,648 bp corresponding to a linearized plasmid, λ phage DNA and phage T4 GT7 DNA used in Plesa et al. (2016).

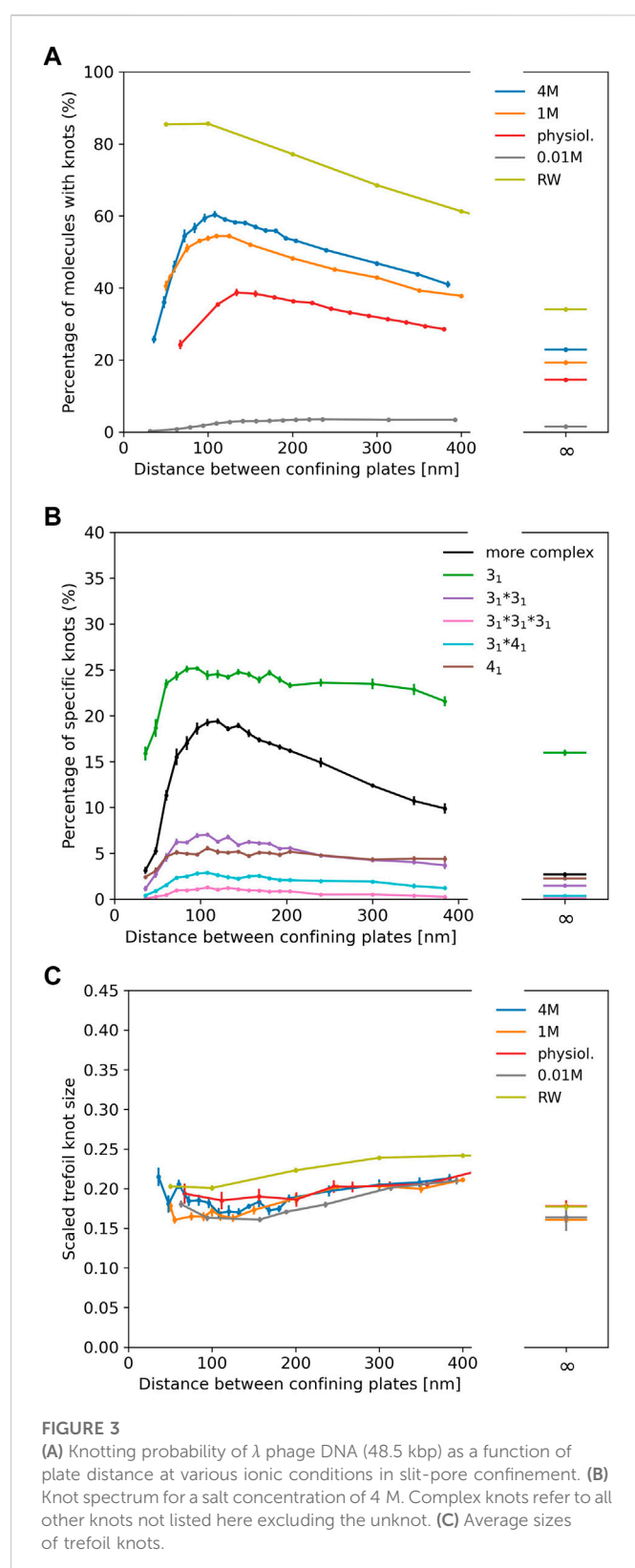
For comparison we have also implemented a simple random walk which can be mapped onto DNA by setting the Kuhn length to 100 nm, which takes over the role of d from Eq. 3. Interestingly, this simplistic model for DNA was already discussed in the first simulation paper on polymer knots from 1974 (Vologodskii et al., 1974) and yields, as we will see later, surprisingly reasonable results when compared with recent experiments on long DNA strands (Plesa et al., 2016). Of course, differences in knotting probabilities due to varying ionic conditions are not captured in this approach, but could in principle be included following Rybenkov et al. (1993). All chains were simulated with a pivot Monte Carlo algorithm (Madras and Sokal, 1988): After a pivot center is chosen at random, one arm of the polymer is rotated by a random angle around the pivot point and the move is accepted with the Metropolis criterion.

Knot analysis. Knots are defined only for closed chains and characterised by the minimum number of crossings when projected onto a two-dimensional plane (see Figure 1B) (Adams, 1994). The simplest knot, apart from an unknotted ring which is called the unknot (0), is the trefoil (3_1) with three essential crossings. Similarly, the next knot type to follow is the figure-eight knot (4_1) with four crossings. While there is only one knot with three and one knot with four crossings (as indicated by the index), eventually the number of different knots grows exponentially with the crossing number. In addition to prime knots, multiple knots can also be combined on a ring to form so-called composite knots as indicated in the right-most picture of Figure 1B.

Since we have simulated linear chains a closure to connect the two end points of our chains needs to be defined. For this we first connect the two termini with their centre of mass. Along these lines one can then define a closure which emerges from one end, follows the first line, connects to the second one far away from the polymer and ends at the second end of the chain (Virnau et al., 2006). Once the open chain has been closed the Alexander polynomial can be determined for which a detailed description can be found in Virnau (2010). The size of a knot can be determined by successively removing monomers from the two ends of the polymer (before closure) chain until the knot type changes.

Results

First, we investigate knotting probabilities as a function of DNA length and ionic conditions for unconfined DNA (Figure 2A). For better clarity, we only plot fitted curves according to Deguchi and



Tsurusaki (1997) for our simulated data. Experimental data from recent nanopore experiments (Plesa et al., 2016; Kumar Sharma et al., 2019) on 20,678 bp long linearized plasmids, λ phage (48,502 bp) and T4 GT7 DNA (165,648 bp) are displayed as data points. We notice that even in the range of 100 kbp, DNA already exhibits substantial knotting which strongly depends on ionic conditions. Knotting

probabilities are larger in high salt scenarios and can reach up to 70% for the largest strands. Intriguingly, our coarse-grained simulations also suggest that knotting can almost be avoided completely in a low salt scenario. For the same 166 kbp strand we only observe a knotting probability of 5%, which (if confirmed experimentally) would open up new possibilities for disentangling large DNA strands, e.g., in preparation of nanopore sequencing. The latter may, however, prove challenging experimentally as it becomes difficult to translocate at low ionic concentrations. These large discrepancies are indeed surprising as prior simulations of closed DNA rings with a similar model yielded significantly higher knotting probabilities, particularly for the low salt scenario (Dai et al., 2012b). Overall, agreement between predicted and experimentally determined knotting probabilities in medium to high salt conditions is quite good and differences only amount to a few percent. Surprisingly, comparisons with a simple random walk model still yield reasonable agreement even though occurrences of knots are overestimated systematically. At the length scales considered, the knot spectrum is still dominated by trefoil knots as is depicted for the high salt (4 M) scenario in Figure 2B. However, we already observe the emergence of composite knots as demonstrated before for even larger chains under physiological conditions in Rieger and Virnau (2016).

In Figure 3A we show results for λ phage DNA (48,502 bp) confined between two plates to study the interplay of ionic conditions with confinement. As no experimental data is available, Figure 3A only displays simulation results. The general shape of the curves follows results for shorter chains and high ionic conditions from Micheletti and Orlandini (2012a), Orlandini and Micheletti (2013) and for rings in Dai et al. (2012b): The knotting probability first increases with increasing plate distance, reaches a maximum at around 100–150 nm before falling off and approaching the value obtained for unconfined DNA. Here, we note again that knotting is suppressed substantially in low salt scenarios. For all salt concentrations, the number of knotted conformations in comparison to unconfined DNA is roughly increased by a factor of two at the maximum, and the position of the maximum shifts to lower plate distances with increasing salt concentrations as noted for closed rings in Dai et al. (2012b).

Figure 3B displays the knot spectrum as a function of plate distance for the 4 M high salt scenario. While the amount of complex knots decreases (and unknots thus increase) for distances beyond the maximum, the composition of trefoil, figure-eight and composite variants of the two only varies slightly.

In Figure 3C, we plot the scaled trefoil knot length (which we define as the ratio of the contour length of the trefoil knot to the contour length of the whole chain). For all concentrations and plate distances, a trefoil knot roughly occupies one-fifth of the chain and has a similar size as in the unconfined scenario. For a simple random walk, we roughly obtain the same result.

Discussion

We investigate with numerical simulations the influence of ionic conditions on knotting of free DNA and DNA confined between two

plates with a focus on long, experimentally relevant strands. From a technical point of view we test and confirm a coarse-grained bead-stick model by comparing simulations to recent nanopore experiments on DNA knotting. The model is not only susceptible to the influence of ionic conditions and reproduces the existing experimental knotting probabilities for unconfined DNA, but also resolves the structure of DNA below the persistence length. As such it is well-suited for the numerical description of recent (Plesa et al., 2016; Kumar Sharma et al., 2019) and ongoing DNA experiments in the range of tens to hundreds of kilo base pairs and could be easily adapted for molecular dynamics simulations. Extensions which account for smaller, varying persistence lengths in small strands could be implemented as well to study structural properties of DNA at these scales (Zoli, 2018). At large length scales we observe a strong dependence on solvent conditions: While knotting can be abundant in a high salt scenario in which negative charges on DNA are completely screened, it becomes almost negligible in low salt conditions. Experiments on DNA dynamics (Shusterman et al., 2004) also imply that characteristic time scales involved in these transitions may well be below typical times required, e.g., for nanopore sequencing even though further studies on this issue are certainly warranted. If this drastic change is confirmed experimentally in long strands, an adjustment of ionic conditions could indeed be used as a switch to effectively unknot DNA in scenarios where knots are undesired. Likewise, such experiments could further improve coarse-grained models by eliminating the need to assume effective excluded volume interactions, which could be fitted to knotting probabilities instead (Rieger and Virnau, 2016; Rieger and Virnau, 2018).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

SW: Simulation and analysis (lead); software (equal); writing of original draft (equal). RD: Simulation and topological analysis

References

- Adams, C. C. (1994). *The knot book*. Rhode Island: American Mathematical Soc.
- Amin, S., Khorshid, A., Zeng, L., Jimmy, P., and Reisner, W. (2018). A nanofluidic knot factory based on compression of single DNA in nanochannels. *Nat. Commun.* 9, 1506. doi:10.1038/s41467-018-03901-w
- Brunet, A., Tardin, C., Salomé, L., Rousseau, P., Destainville, N., and Manghi, M. (2015). Dependence of DNA persistence length on ionic strength of solutions with monovalent and divalent salts: A joint theory-experiment study. *Macromolecules* 48, 3641–3652. doi:10.1021/acs.macromol.5b00735
- Dai, L., Jones, J. J., van der Maarel, J. R. C., and Doyle, P. S. (2012a). A systematic study of DNA conformation in slitlike confinement. *Soft Matter* 8, 2972–2982. doi:10.1039/C2SM07322F
- Dai, L., Renner, C. B., and Doyle, P. S. (2015). Metastable knots in confined semiflexible chains. *Macromolecules* 48, 2812–2818. doi:10.1021/acs.macromol.5b00280
- Dai, L., van der Maarel, J. R. C., and Doyle, P. S. (2012b). Effect of nanoslit confinement on the knotting probability of circular DNA. *ACS Macro Lett.* 1, 732–736. doi:10.1021/mz3001622
- Deguchi, T., and Tsurusaki, K. (1997). Universality of random knotting. *Phys. Rev. E* 55, 6245–6248. doi:10.1103/PhysRevE.55.6245
- Delbrück, M. (1962). “Knotting problems in biology,” in *Mathematical problems in biological sciences*. Editor R. E. Bellman (Rhode Island: American Mathematical Society). vol. 14 of Proc. Symp. Appl. Math, 55.
- Di Stefano, M., Tubiana, L., Di Ventra, M., and Micheletti, C. (2014). Driving knots on DNA with ac/dc electric fields: Topological friction and memory effects. *Soft Matter* 10, 6491–6498. doi:10.1039/C4SM00160E
- Fisher, M. E. (1964). Magnetism in one-dimensional systems—The Heisenberg model for infinite spin. *Am. J. Phys.* 32, 343–346. doi:10.1119/1.1970340
- Frisch, H. L., and Wasserman, E. (1961). Chemical topology 1. *J. Am. Chem. Soc.* 83, 3789–3795. doi:10.1021/ja01479a015
- Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., et al. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science* 360, 102–105. doi:10.1126/science.aar7831
- Hill, T. (1960). *Introduction to statistical thermodynamics*. Boston: Adison-Wesley.
- Hill, T. (1956). *Statistical mechanics*. New York, USA: McGraw-Hill.
- Jain, A., and Dorfman, K. D. (2017). Simulations of knotting of DNA during genome mapping. *Biomechanics* 11, 024117. doi:10.1063/1.4979605

(supporting); software (supporting); writing of original draft (equal). PV: Conceptualization (lead); funding acquisition (lead); methodology (lead); supervision (lead); writing (lead); software (equal).

Acknowledgments

We are grateful to the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding this research: Project number 233630050-TRR 146 and 464588647-CRC 1551. The authors also acknowledge computing time granted on the supercomputer Mogon offered by the Johannes Gutenberg University Mainz (hpc.uni-mainz.de), which is a member of the AHRP (Alliance for High Performance Computing in Rhineland Palatinate, www.ahrp.info) and the Gauss Alliance e.V. P.V. would also like to acknowledge helpful discussions with Eugene Kim.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fchem.2022.1096014/full#supplementary-material>

- Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345. doi:10.1038/nbt.4060
- Jain, M., Olsen, H., Benedict, P., and Akerson, M. (2016). The oxford nanopore minION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 239. doi:10.1186/s13059-016-1103-0
- Kam, Z., Borochov, N., and Eisenberg, H. (1981). Dependence of laser light scattering of DNA on NaCl concentration. *Biopolymers* 20, 2671–2690. doi:10.1002/bip.1981.360201213
- Kratky, O., and Porod, G. (1949). Röntgenuntersuchung gelöster Fadenmoleküle. *Recl. Des. Trav. Chim. Des. Pays-Bas* 68, 1106–1122. doi:10.1002/recl.19490681203
- Kumar Sharma, R., Agrawal, I., Dai, L., Doyle, P. S., and Garaj, S. (2019). Complex DNA knots detected with a nanopore sensor. *Nat. Commun.* 10, 4473. doi:10.1038/s41467-019-12358-4
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi:10.1126/science.1181369
- Lua, R., Borovinskiy, A. L., and Grosberg, A. Y. (2004). Fractal and statistical properties of large compact polymers: A computational study. *Polymer* 45, 717–731. Conformational Protein Conformations. doi:10.1016/j.polymer.2003.10.073
- Madras, N., and Sokal, A. D. (1988). The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. *J. Stat. Phys.* 50, 109–186. doi:10.1007/BF01022990
- Manning, G. S. (1981). A procedure for extracting persistence lengths from light-scattering data on intermediate molecular weight DNA. *Biopolymers* 20, 1751–1755. doi:10.1002/bip.1981.360200815
- Marenduzzo, D., Orlandini, E., Stasiak, A., Sumners, D. W., Tubiana, L., and Micheletti, C. (2009). DNA–DNA interactions in bacteriophage capsids are responsible for the observed DNA knotting. *Proc. Natl. Acad. Sci.* 106, 22269–22274. doi:10.1073/pnas.0907524106
- Marenz, M., and Janke, W. (2016). Knots as a topological order parameter for semiflexible polymers. *Phys. Rev. Lett.* 116, 128301. doi:10.1103/PhysRevLett.116.128301
- McMillan, W. G., and Mayer, J. E. (1945). The statistical thermodynamics of multicomponent systems. *J. Chem. Phys.* 13, 276–305. doi:10.1063/1.1724036
- Micheletti, C., and Orlandini, E. (2012a). Knotting and metric scaling properties of DNA confined in nano-channels: A Monte Carlo study. *Soft Matter* 8, 10959–10968. doi:10.1039/C2SM26401C
- Micheletti, C., and Orlandini, E. (2014). Knotting and unknotting dynamics of DNA strands in nanochannels. *ACS Macro Lett.* 3, 876–880. doi:10.1021/mz500402s
- Micheletti, C., and Orlandini, E. (2012b). Numerical study of linear and circular model DNA chains confined in a slit: Metric and topological properties. *Macromolecules* 45, 2113–2121. doi:10.1021/ma202503k
- Odijk, T. (1977). Polyelectrolytes near the rod limit. *J. Polym. Sci. Polym. Phys. Ed.* 15, 477–483. doi:10.1002/pol.1977.180150307
- Orlandini, E., and Micheletti, C. (2013). Knotting of linear DNA in nano-slits and nano-channels: A numerical study. *J. Biol. Phys.* 39, 267–275. doi:10.1007/s10867-013-9305-0
- Plesa, C., Verschuere, D., Pud, S., van der Torre, J., Ruitenberg, J. W., Witteveen, M. J., et al. (2016). Direct observation of DNA knots using a solid-state nanopore. *Nat. Nanotechnol.* 11, 1093–1097. doi:10.1038/nnano.2016.153
- Post, C. B. (1983). Excluded volume of an intermediate-molecular-weight DNA. A Monte Carlo analysis. *Biopolymers* 22, 1087–1096. doi:10.1002/bip.360220406
- Reith, D., Cifra, P., Stasiak, A., and Virnau, P. (2012). Effective stiffening of DNA due to nematic ordering causes DNA molecules packed in phage capsids to preferentially form torus knots. *Nucleic Acids Res.* 40, 5129–5137. doi:10.1093/nar/gks157
- Renner, C. B., and Doyle, P. S. (2014). Untying knotted DNA with elongational flows. *ACS Macro Lett.* 3, 963–967. doi:10.1021/mz500464p
- Rieger, F. C., and Virnau, P. (2016). A Monte Carlo study of knots in long double-stranded DNA chains. *Plos Comput. Biol.* 12, e1005029. doi:10.1371/journal.pcbi.1005029
- Rieger, F. C., and Virnau, P. (2018). Coarse-grained models of double-stranded DNA based on experimentally determined knotting probabilities. *React. Funct. Polym.* 131, 243–250. doi:10.1016/j.reactfunctpolym.2018.08.002
- Rothörl, J., Wettermann, S., Virnau, P., and Bhattacharya, A. (2022). Knot formation of dsDNA pushed inside a nanochannel. *Sci. Rep.* 12, 5342. doi:10.1038/s41598-022-09242-5
- Rybenkov, V. V., Cozzarelli, N. R., and Vologodskii, A. V. (1993). Probability of DNA knotting and the effective diameter of the DNA double helix. *Proc. Natl. Acad. Sci. U. S. A.* 90, 5307–5311. doi:10.1073/pnas.90.11.5307
- Savelyev, A. (2012). Do monovalent mobile ions affect DNA's flexibility at high salt content? *Phys. Chem. Chem. Phys.* 14, 2250–2254. doi:10.1039/C2CP23499H
- Shaw, S., and Wang, J. (1994). Knotting of a DNA chain during ring closure. *Science* 260, 533–536. doi:10.1126/science.8475384
- Shusterman, R., Alon, S., Gavrinov, T., and Krichevsky, O. (2004). Monomer dynamics in double- and single-stranded DNA polymers. *Phys. Rev. Lett.* 92, 048303. doi:10.1103/PhysRevLett.92.048303
- Siebert, J. T., Kivel, A. N., Atkinson, L. P., Stevens, T. J., Laue, E. D., and Virnau, P. (2017). Are there knots in chromosomes? *Polymers* 9, 317. doi:10.3390/polym9080317
- Skolnick, J., and Fixman, M. (1977). Electrostatic persistence length of a wormlike polyelectrolyte. *Macromolecules* 10, 944–948. doi:10.1021/ma60059a011
- Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 544, 59–64. doi:10.1038/nature21429
- Stigter, D. (1977). Interactions of highly charged colloidal cylinders with applications to double-stranded DNA. *Biopolymers* 16, 1435–1448. doi:10.1002/bip.1977.360160705
- Suma, A., Di Stefano, M., and Micheletti, C. (2018). Electric-field-driven trapping of polyelectrolytes in needle-like backfolded states. *Macromolecules* 51, 4462–4470. doi:10.1021/acs.macromol.8b00019
- Suma, A., and Micheletti, C. (2017). Pore translocation of knotted DNA rings. *Proc. Natl. Acad. Sci.* 114, E2991–E2997. doi:10.1073/pnas.1701321114
- Tang, J., Du, N., and Doyle, P. S. (2011). Compression and self-entanglement of single DNA molecules under uniform electric field. *Proc. Natl. Acad. Sci.* 108, 16153–16158. doi:10.1073/pnas.1105547108
- Trefz, B., Siebert, J., and Virnau, P. (2014). How molecular knots can pass through each other. *Proc. Natl. Acad. Sci.* 111, 7948–7951. doi:10.1073/pnas.1319376111
- Virnau, P. (2010). Detection and visualization of physical knots in macromolecules. *Phys. Procedia* 6, 117–125. doi:10.1016/j.phpro.2010.09.036
- Virnau, P., Kantor, Y., and Kardar, M. (2005). Knots in globule and coil phases of a model polyethylene. *J. Am. Chem. Soc.* 127, 15102–15106. doi:10.1021/ja052438a
- Virnau, P., Mirny, L. A., and Kardar, M. (2006). Intricate knots in proteins: Function and evolution. *PLoS Comput. Biol.* 2, e122. doi:10.1371/journal.pcbi.0020122
- Vologodskii, A., Lukashin, A., and Frank-Kamenetskii, M. D. (1974). Topological interaction between polymer chains. *Sov. Phys.-JETP* 40, 932–936.
- Zoli, M. (2018). Short DNA persistence length in a mesoscopic helical model. *Europhys. Lett.* 123, 68003. doi:10.1209/0295-5075/123/68003



OPEN ACCESS

EDITED BY

Adolfo Poma,
Polish Academy of Sciences, Poland

REVIEWED BY

Rosana Collepardo,
University of Cambridge, United Kingdom
Roberto Covino,
Frankfurt Institute for Advanced Studies,
Germany

*CORRESPONDENCE

Siewert J. Marrink,
✉ s.j.marrink@rug.nl

SPECIALTY SECTION

This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

RECEIVED 23 November 2022

ACCEPTED 09 January 2023

PUBLISHED 18 January 2023

CITATION

Stevens JA, Grünewald F, van Tilburg PAM, König M, Gilbert BR, Brier TA, Thornburg ZR, Luthey-Schulten Z and Marrink SJ (2023), Molecular dynamics simulation of an entire cell. *Front. Chem.* 11:1106495. doi: 10.3389/fchem.2023.1106495

COPYRIGHT

© 2023 Stevens, Grünewald, van Tilburg, König, Gilbert, Brier, Thornburg, Luthey-Schulten and Marrink. 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.

Molecular dynamics simulation of an entire cell

Jan A. Stevens¹, Fabian Grünewald¹, P. A. Marco van Tilburg¹, Melanie König¹, Benjamin R. Gilbert², Troy A. Brier², Zane R. Thornburg², Zaida Luthey-Schulten² and Siewert J. Marrink^{1*}

¹Molecular Dynamics Group, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, Netherlands, ²Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Champaign, IL, United States

The ultimate microscope, directed at a cell, would reveal the dynamics of all the cell's components with atomic resolution. In contrast to their real-world counterparts, computational microscopes are currently on the brink of meeting this challenge. In this perspective, we show how an integrative approach can be employed to model an entire cell, the minimal cell, JCVI-syn3A, at full complexity. This step opens the way to interrogate the cell's spatio-temporal evolution with molecular dynamics simulations, an approach that can be extended to other cell types in the near future.

KEYWORDS

JCVI-syn3A, minimal cell, Martini force field, integrative modeling, coarse grain, polyply

Introduction

Biomolecular functions emerge from the molecular interactions taking place in cellular environments. Understanding each component's role in driving cell function poses an immense challenge. For a long time, experimental techniques have been our main window into the cellular environment. By resolving biomolecular structures and probing the dynamics of biomolecular processes, both *in vivo* and *in vitro*, researchers have developed a global understanding of how a cell functions.

A limitation of these experimental techniques is the spatio-temporal resolution that they can probe, particularly within the complex and crowded environment of the cell. Molecular dynamics (MD) simulations provide a suitable alternative approach, covering the relevant length and timescales at molecular resolution, albeit over short periods of a cell cycle. Over the past decades, MD has matured into a powerful tool that functions as a computational microscope (Lee et al., 2009; Dror et al., 2012). With the advances in available computer power, including the transition from using central processing units (CPUs) to graphical processing units (GPUs), the complexity and the spatio-temporal scales of MD simulations have increased remarkably. State-of-the-art simulations, containing hundreds of millions of atoms, include dynamic models of a photosynthetic chromatophore vesicle (Singharoy et al., 2019), the nuclear core complex (Moslaganti et al., 2022), the membranes of a mitochondrion (Pezeshkian et al., 2020), the bacterial cytoplasm (Yu et al., 2016), and a virus particle embedded in a nanoscale aerosol droplet (Dommer et al., 2022). The natural next step is, arguably, the scale of entire cells (Bhat and Balaji, 2020; Khalid and Rouse, 2020; Luthey-Schulten et al., 2022; Thornburg et al., 2022).

Creating a whole-cell model has long been a major goal in computational modeling. A computational cell will help us to resolve a more integral picture of how biomolecular interactions drive cell function since biomolecular processes operate on a hierarchy of interconnected scales. Thus, resolving the full cell function requires a holistic approach.

The current state-of-the-art uses static representations of heterogeneous cell-scale structures such as cellPACK (Johnson et al., 2015; Maritan et al., 2022), genome-scale well-stirred reaction models for metabolism (Karr et al., 2012; Macklin et al., 2014; Karr et al., 2015; Breuer et al., 2019; Macklin et al., 2020), or mesoscale kinetic models that attempt to include both structural and chemical states of the cell such as Lattice Microbes (Roberts et al., 2013). These computational techniques, despite granting impressive insights into the complexity of cellular processes, are limited by the spatial resolution they can achieve.

Constructing whole-cell models requires the integration of a large amount of experimental data, i.e., relies on an integrative modeling approach (Bonvin, 2021; Luthey-Schulten, 2021; Gupta et al., 2022). Obtaining such data with high spatial and dynamic detail, particularly in living cells, is very challenging, but exciting progress is being made in elucidating the architecture and stoichiometry of cellular components at unprecedented resolutions (Reading et al., 2017; Ando et al., 2018; Cheng, 2018; Chorev et al., 2018; Christie et al., 2020; Lorent et al., 2020; Narasimhan et al., 2020; Wietrzynski et al., 2020; Štefl et al., 2020), setting the stage for spatially detailed simulations of whole cells. To showcase this possibility, we consider one of the simplest cells known to date: the minimal cell created by the J. Craig Venter Institute (Hutchison et al., 2016), a stripped-down version of a *Mycoplasma* bacterium. The current strain, named JCVI-syn3A (Syn3A), contains only 493 genes and is still able to replicate independently (Breuer et al., 2019). This cell is particularly amenable to detailed computational modeling because it is of relatively small size (measuring 400 nm in diameter), and its precise composition has largely been resolved (Breuer et al., 2019).

Here we present our ongoing effort to simulate Syn3A using the Martini coarse-grained (CG) force field (Marrink et al., 2022). The Martini force field employs a four-to-one mapping scheme, where up to four heavy atoms and associated hydrogens are represented by one CG bead. This reduction in the number of particles in the system, together with a smoothening of the potential energy surface, speeds up the simulations by about three orders of magnitude, enabling simulations of systems that approach the size of entire cells. In the case of Syn3A, about 550 million CG particles are required, corresponding to more than six billion atoms.

In the remainder of this work, we first describe the set of tools needed to construct a system as complex as an entire cell at the Martini level (the Martini “ecosystem”), including a description of the stepwise procedure to construct a starting model for Syn3A, from building the chromosome and modeling the cytoplasm to the addition of the cell envelope. We end with the prospects of actually simulating this model and discuss the potential future avenues of simulating entire cells. The integrative modeling workflow is schematically depicted in Figure 1.

Building cells with the Martini ecosystem

Modeling cellular environments using a coarse-grained approach requires the use of a force field that incorporates enough detail to represent all biomolecules and their interactions explicitly. In this regard, the Martini force field is an excellent candidate, as demonstrated by the wide range of application studies using Martini over the past 2 decades (Marrink et al., 2022). Additionally, parameters for a large variety of biomolecules are already available, including proteins (de Jong et al., 2013), lipids

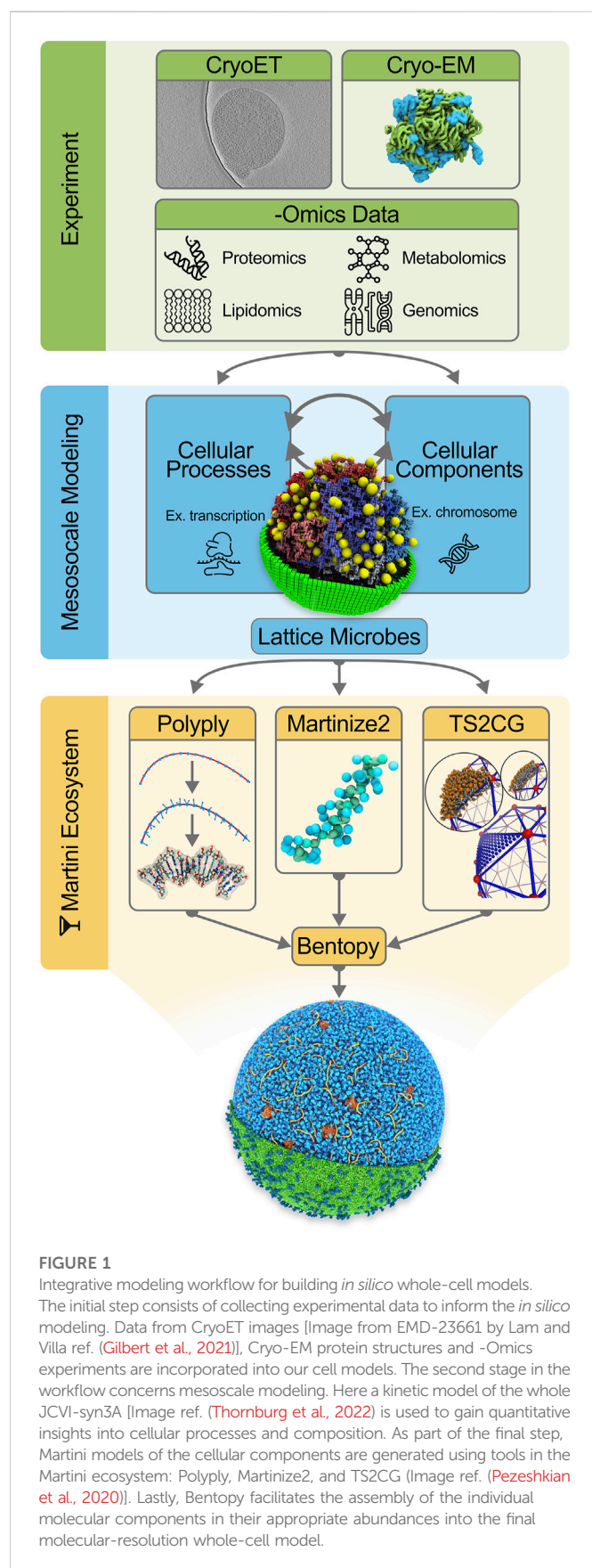
(Wassenaar et al., 2015), polynucleotides (Uusitalo et al., 2015; Uusitalo et al., 2017), carbohydrates (López et al., 2009; Grünewald and Punt, 2022) and metabolites (Sousa et al., 2021; Alessandri et al., 2022). A curated collection of all parameters is available from the Martini Database (Hilpert et al., 2022).

Accompanying the Martini force field is a collection of tools that compose the software side of the Martini ecosystem (Figure 1). The primary goal of this software suite is to facilitate the construction of topologies and initial coordinates for running CG Martini MD simulations. The Martini ecosystem is currently being extended around a central framework, named *Vermouth*. Making use of a graph-based description of molecules, *Vermouth* implements a unified handling of processes frequently used in Martini, such as topology and coordinate generation or resolution transformation, as a lightweight python library (Kroon et al., 2022).

Proteins comprise the bulk of a cell’s organic material, composing approximately 40% of the total intracellular volume (Ellis and Minton, 2003). The number of unique proteins expressed by the cell, i.e., the proteome, can range from a few hundred to several thousand. Consequently, describing realistic cellular environments requires generating topology files for proteins of varying shapes and sizes as well as packing these into a highly crowded solution. *Martinize2*, which is built on top of *Vermouth*, is designed for high-throughput generation of Martini topologies and coordinates for proteins from atomistic protein structures. The workflow used in *Martinize2* additionally performs quality checks on the atomistic protein structures and alerts to potential problems making it ideal for such a high-complexity use case (Kroon et al., 2022). To generate dense protein solutions on a cellular scale as required for whole-cell modeling, a new software tool, called *Bentopy*, is currently under development. It uses an efficient collision detection scheme (Howard et al., 2016) to generate random packings of proteins and protein complexes within volumetric constraints. Furthermore, functional annotations of proteins can be integrated into the algorithm, biasing their spatial distribution in the cytosol based on their known biochemical function.

Constructing coordinates and input files for chromosomal DNA presents another challenge for modeling of a whole cell. Even for a comparatively small genome as that of JCVI-syn3A, approaches that rely on reading input coordinates and forward mapping such as used in *Martinize2*, become too inefficient due to the sheer size of the molecule. Instead the *Polyply* software, which was originally developed to efficiently setup general polymeric systems, can be used (Grünewald et al., 2022). Within *Polyply*, a multiresolution graph-based approach is used to efficiently generate polymer topologies, in particular for DNA, only from the sequence. In addition to topologies, system coordinates can be generated using a specialized biased random walk protocol. This tool of the Martini ecosystem has successfully been applied to model dense polymer melts and simple ssDNA viral chromosomes. At the moment, the package is being extended to handle double-stranded nucleic acids, and generate more complex DNA structures such as bacterial chromosomes.

Lastly, modeling lipid membranes has historically been a leading application of the Martini force field (Marrink et al., 2019). Simulating arbitrarily complex membranes of various sizes, geometries, and lateral heterogeneities is facilitated by TS2CG (Pezeshkian et al., 2020). This tool implements a backmapping algorithm that converts triangulated surfaces into CG membrane models. As a result of the method’s high level of control, the curvature-



dependent lipid concentrations in both membrane leaflets can be precisely determined by the user. In addition, proteins can be inserted into the membrane together with their characteristic lipid shells

[i.e., lipid fingerprints (Corradi et al., 2018)], setting the stage for building cell envelopes.

In the following subsections, we describe the application of the aforementioned tools to construct a proof of principle whole-cell simulation of Syn3A, illustrating how the current Martini ecosystem enables users to study multi-component systems at the mesoscale.

Chromosome building

The minimal genome of JCVI-syn3A contains 493 genes and is encoded in a single circular chromosome of 543 kilobase pairs (kbps). Since the chromosome is contained inside the cell's cytosol, the structural organization is heavily influenced by the crowded intracellular environment. Due to the size and near-uniform distribution of ribosomes present in the cytosol, the excluded volume interactions of these protein-RNA complexes are known to have a significant influence on the nucleoid organization (Mondal et al., 2011).

The nucleoid structure of Syn3A was previously modeled by Gilbert et al. (2021) based on the ribosome distribution and cell boundary determined by cryo-electron tomography. A Monte Carlo (MC) method grew the chromosome, modeled by a self-avoiding polygon, on a lattice inside the cell boundary. Each MC step ensured that no model constraints were violated, resulting in a circular genome without steric clashes with the ribosomes or cell membrane. The algorithm was validated by comparing the chromosome conformation capture (3C) maps of ensembles of simulated nucleoid configurations with experimental 3C maps. 3C maps show spatial correlations between chromosomal regions, which are spatially close but can be distant in the nucleotide sequence. Based on the features in the 3C maps, we infer that the chromosome is organized more like a fractal globule with little persistent supercoiling.

Whilst the previous chromosome modeling approach with a lattice polymer was tailored to be highly compatible with the whole-cell simulations using Lattice Microbes, we have subsequently developed a new method to generate circular chromosomes organized as fractal globules in a continuum polymer model with 10 bp monomers. The generated chromosome model is relaxed using Brownian dynamics and an energy function for modeling dsDNA as a twistable worm-like chain from (Brackley et al., 2014). In order to connect the chromosome model to a Martini-level representation, the model is transformed to a one-bead-per-base-pair resolution by spline interpolation. Rotation minimizing frames are then constructed along the chromosomal contour, providing a consistent reference to which the Martini DNA model can be backmapped (Wang et al., 2008). After adding an equilibrium twist along the frame's tangent vector, Martini base pair templates matching the 543 kbp genome sequence are positioned along the chromosome following the local contour reference frame. By performing a short energy minimization the system is relaxed, resulting in a stable chromosome structure. The subsequent model consists of 543 kbps, which at a Martini resolution is equivalent to seven million beads. By implementing this backmapping procedure in *Polyply*, we are able to efficiently generate the coordinates for the chromosome in a force field agnostic manner. The overall chromosome building takes a matter of minutes, opening up the possibility of studying larger protein-DNA complexes like chromatin fibers and *Escherichia coli* chromosomes.

The required topology files were generated from the sequence using the default *Polyply* methods.

Cytosol modeling

In order to model the cytosol, it is essential to have a complete picture of the bacterial proteome, including both protein structures and proteomics counts. The genome reduction leading to Syn3A limits the number of different proteins that have to be taken into account by only retaining 452 protein-coding genes. This minimal genome has been extensively characterized, and only 91 genes remain without an annotated function. A recent study by Bianchi *et al.* (2022) uses computational analyses to further elucidate the function of uncharacterized genes and work toward complete functional characterization of the proteome. By gaining a better understanding of the function of encoded proteins, we will be able to inform the spatial distribution of proteins in our whole-cell model.

From the 452 different proteins expressed by Syn3A, 281 are characterized as cytosolic proteins, 63 as trans-membrane proteins, 42 as peripheral membrane proteins, and the remaining 66 still have an unknown localization. As part of the computational gene characterization workflow, Bianchi *et al.* modeled the protein structures of the entire proteome using AlphaFold2 (Jumper *et al.*, 2021). *Martinize2* successfully converted all but one of the predicted protein structures (451) to a corresponding Martini model. Using *Bentopy*, the cytosolic protein models are packed into the intracellular volume alongside the chromosome and ribosomes. The number of copies of each protein is based on available proteomics data (Breuer *et al.*, 2019; Thornburg *et al.*, 2022); in total, around 60,000 proteins were distributed within a spherical volume with a diameter of 400 nm. Concerning the ribosomes, we used bacterial homologs that we had already generated previously (Uusitalo *et al.*, 2017), placing 503 ribosomes in random orientations near the positions originally determined from the cryo-electron tomography map (Gilbert *et al.*, 2021). Single-stranded RNA fragments were not included at this stage.

The next major component of the cytosol are the small molecules that, together with enzymatic proteins, participate in the metabolic pathways. In the current model, we include only the metabolites for which Martini topologies were already available, primarily amino acids and nucleotide cofactors (Sousa *et al.*, 2021), and which are present at high concentrations inside Syn3A. The metabolite models were automatically generated from the topology files using *Polyply*. Based on available metabolomic data (Thornburg *et al.*, 2022), 1.7 million metabolites are distributed within the cytosol, approximately 55% of the metabolite count for the complete metabolome.

Constructing the envelope

Modeling the cell envelope of the Syn3A is a straightforward procedure since it is solely composed of a singular cytoplasmic membrane. Furthermore, experimental measurements indicate the absence of a cell capsule, drastically reducing the complexity of the cell boundary. The lipid membrane is constructed using *TS2CG* with a uniform lipid mixture across both membrane leaflets. It should be

noted that since the minimal cell acquires membrane components through lipid synthesis from fatty acids and direct incorporation of lipids from its environment, the lipid composition of the cellular membrane heavily depends on the growth medium. We base our model on the lipidomics data presented by (Thornburg *et al.*, 2022), indicating the presence of five main lipid types: cholesterol (59%), sphingomyelin (18%), cardiolipin (17%), phosphatidylcholines (4%), and phosphatidylglycerol (2%). In the absence of more detailed lipidomics data, all lipids are modeled with fully saturated palmitoyl tails. The total lipid count amounted to 1.3 million lipids.

Additionally, we randomly inserted membrane proteins into the cell membrane using *TS2CG*. From the available proteomics data, the number and types of membrane proteins are determined. While AlphaFold2 structure predictions can be used directly to model monomeric membrane proteins, experimental crystal structures are still required for the protein transport complexes. *Martinize2* is again used to generate the Martini models for the membrane proteins. For simplicity, we selected five abundant protein complexes and distributed these uniformly over the membrane. In total, 2,200 protein complexes were embedded in the cell envelope, corresponding with the expected number of membrane proteins present on the surface of Syn3A.

Solvating and simulating the cell

Having modeled all the cell components, the final step in constructing a starting structure for subsequent simulation is defining the periodic simulation box and solvating the system. Considering the whole-cell model's spherical shape, a logical choice for the periodic box is a rhombic dodecahedron. To solvate, a periodic water box is tiled across the cell model, removing the water beads that overlap with the model using a collision detection scheme. The system is neutralized by placing counter ions near the highly charged components in the cytosol, i.e., the chromosome and ribosomes; the overall negative charge is substantial, amounting to 3.2 million elementary charges. As part of the solvation procedure, we also replace an appropriate number of water beads with ion beads to establish an ion concentration of 135 mM NaCl across our system, mimicking the experimental buffer. Thus, we ended up with a system containing 447 million water beads (208 million inside, 239 outside of the cell), 8.5 million sodium, and 5.3 million chloride ions. Note that Martini CG water beads represent four real water molecules. The total bead count, including all biomolecules, adds up to 561 million beads. A snapshot of the full system is shown in Figure 2.

Having constructed a starting model for Syn3A, the current challenge is to perform an actual MD simulation. At the time being, this proved to be non-trivial. Gromacs (Abraham *et al.*, 2015), the main MD engine to run Martini-based simulations, is having difficulties handling systems comprising hundreds of millions of particles, in particular featuring large molecules such as the genome spread over multiple domains. The Gromacs developer team is aware of this problem and is dedicated to solving it. Possible other software engines to consider are ddcMD (Zhang *et al.*, 2020) and openMM (Eastman *et al.*, 2017), both of which are supporting Martini and offer simulation speeds comparable to those of Gromacs.

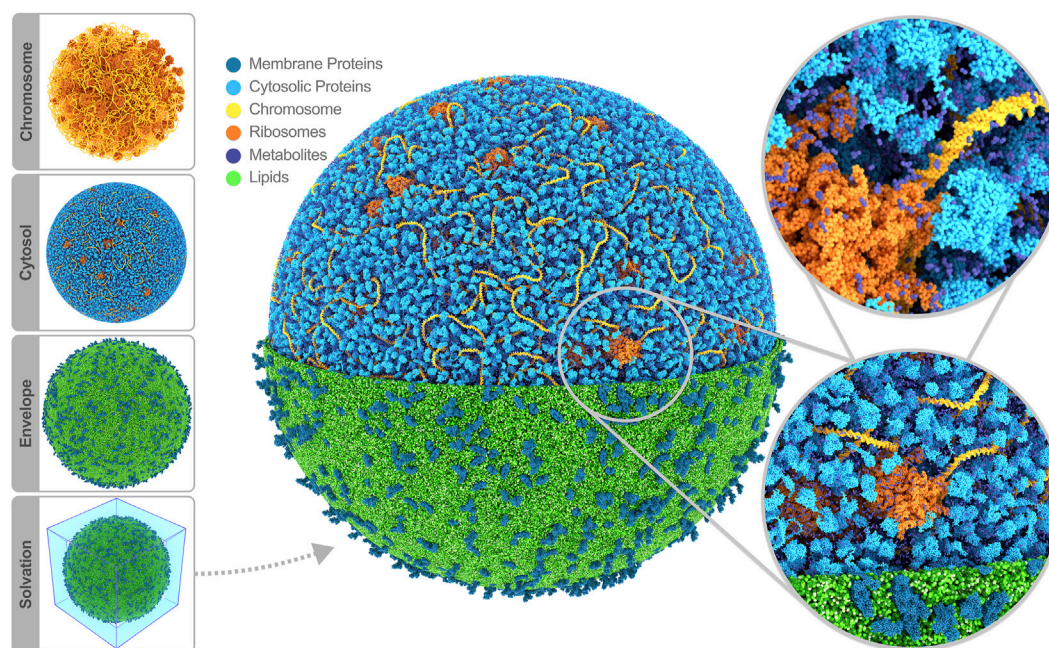


FIGURE 2

Whole-cell Martini model of JCVI-syn3A. The four stages of cell building are shown on the side. The final system contains 60,887 soluble proteins (light blue), 2,200 membrane proteins (blue), 503 ribosomes (orange), a single 500 kbp circular dsDNA (yellow), 1.3 million lipids (green), 1.7 million metabolites (dark blue), 14 million ions (not shown) and 447 million water beads (not shown) for a total of 561 million beads representing more than six billion atoms. Image rendered with Blender ([Blender Online Community, 2022](#)).

Discussion

In the wake of a continuous rise in computing power, MD simulations have transitioned from studying idealized representations of biomolecular systems to modeling their full complexity. The culmination of this development would be simulations at the level of entire cells. As a proof of principle that we are ready to meet this challenge, we presented a model of the complete minimal cell JCVI-syn3A, constructed using the Martini ecosystem. The final simulation box comprises more than 560 million CG beads, representing over six billion atoms in the cell ([Figure 2](#)).

Before looking at the broader prospects of this endeavor, it is important to discuss a number of limitations of our approach. The current model uses the Martini 2 version of the force field since Martini models for nucleic acids, and other essential cellular components are still under development for the latest Martini 3 release. However, the methods described in this paper can be straightforwardly transferred to the latest version of Martini when validated models become available. With over 800 different bead types and a recalibrated interaction matrix, Martini 3 offers an improved framework for CG MD simulations ([Souza et al., 2021](#)). Nevertheless, inherent limitations of Martini, such as an inability to sample protein secondary conformational changes, remain. We do not anticipate that such changes are of primary importance in determining the cellular organization, but details of protein-protein and protein-lipid interactions might be affected. This problem could perhaps be resolved by using Go potentials ([Poma et al., 2017](#); [Souza et al., 2019](#)), which are already integrated into *Martinize2* and *Bentopy*.

Even though our *in silico* cell contains more than 500 unique CG molecules and thereby presumably qualifies as the most complex

system simulated to date, it simplifies the composition of various cellular components of Syn3A. Firstly, limited by the availability of Martini models for the metabolites, only a small subset is currently included in the cytosol. Future iterations of our whole-cell model will include Martini models for the complete metabolome, which comprises about 188 different compounds, and are expected to benefit from the ongoing development of dedicated automatic topology builders ([Bereau and Kremer, 2015](#); [Potter et al., 2021](#)). Secondly, since AlphaFold2 was used to predict the protein structures of the whole proteome, only monomeric structures were initially available. Essential multimeric proteins like the ribosomes and membrane-embedded transport complexes are either left out or represented by homologous proteins with available experimental crystal structures. In the future, improved protein structure prediction algorithms will be used that also facilitate the modeling of multimeric protein structures. In addition, ongoing progress in the experimental characterization of the Syn3A proteome and lipidome, as well as the characterization of the spatial distributions of membrane proteins, will help further increase our model's realism. A "living" list of the complete composition of our *in silico* cell can be found in our GitHub repository ([marrink-lab, 2022](#)).

Another issue is the fine-tuning of the amount of interior solvent (both water and ions), together with the lipid balance between the inner and outer leaflet. Previous works on large-scale membrane-enveloped systems ([Pezeshkian et al., 2021](#); [Vermaas et al., 2022](#)) have shown that finding this balance is a non-trivial task. Unbalanced systems might experience strong osmotic pressures and membrane (curvature) stress, causing unwanted shape deformations all the way to membrane rupture. As a complicating factor, these effects may only appear after prolonged simulation times. Clearly, dedicated

computational resources are required for the simulation of whole cells or cell organelles. The forthcoming generation of supercomputers and simulation software is becoming increasingly efficient, and billion-particle simulations have already been achieved (Jung et al., 2019; Castagna et al., 2020).

An important challenge is reaching timescales long enough to allow meaningful analysis of such large systems. Assuming dedicated computer time on current infrastructure, we anticipate that we can reach timescales of the order of 10–100 μ s in the foreseeable future. Although this is typically considered a long enough simulation time for standard system sizes (e.g., a single membrane protein), it is clear that on the scale of an entire cell, we will not be able to equilibrate our system; the generated ensemble of configurations will remain dependent on our starting state. Equilibration will only happen locally, and multiple replicas will need to be generated to obtain statistically relevant data. Note that the 10–100 μ s range offers a nice overlap with state-of-the-art experimental techniques. For example, advanced MINFLUX microscopy from the Hell lab enables the tracking of particles as small as 1–2 nm for 100 s of microseconds (Eilers et al., 2018; Schmidt et al., 2021). Besides, Lattice Microbes simulations of the Luthey-Schulten group (Roberts et al., 2013) use time steps of the order of microseconds, which allows for a potential feedback loop between these computational approaches.

Another major limitation is the fact that real cells operate out-of-equilibrium, driven by the import and export of nutrients and an intricate metabolic network of chemical reactions. In our approach, which is based on classical MD, we do not take this into account. We are therefore limited to studying non-reactive processes, i.e., those arising from the physical interactions among the constituents. The current composition of our cell is based on average concentrations of proteins and metabolites and thus reflects a steady-state. Coupling our classic approach with approaches taking into account reactivity, such as the aforementioned Lattice Microbes simulations or other metabolic network models (see below), in principle, could capture the non-equilibrium aspect of real cells.

Keeping these limitations in mind, simulations of the minimal cell with a molecular resolution will make it possible to study a wide range of new aspects. Modelling cellular processes and chemical transformations involves a hierarchy of interconnected scales that cannot be separated without causing artefacts. Behaviour emerging from the interaction of millions of different compounds is easily missed when systems are simplified. One might question to what extent one part of the cell affects another, given the limited timescales likely to be reached. If the various cellular subsystems act independently, one might better simulate those in isolation. To find out, one needs to simulate the complete system in addition to the smaller-scale subsystems. Our whole-cell simulation is only a first step, which will benefit from imminent improvements in high-performance computing to extend these simulations to longer timescales, up to the point where all parts of the cell may influence each other. Currently, the internal organization of the cytosol of Syn3A is primarily a black box. Our model will allow us to observe how proteins inside the cytosol interact with macromolecular structures such as ribosomes and chromosomes. Viewing the cytosol from this perspective, we can observe emerging heterogeneities and viscosity gradients, following in the footsteps of other realistic models of the cytoplasm of various cell types (McGuffee and Elcock, 2010; Yu et al., 2016; Oliveira Bortot et al., 2020). We can expect arising interaction patterns between proteins and metabolites, and probe the possible appearance of biomolecular condensates (Guilhas et al., 2020; Rhine et al., 2020).

A simulation at the level of the entire cell allows us to characterize the extent to which the cell membrane affects (and is affected by) the cellular

interior. If we consider a membrane zone with a thickness of 30 nm (~20 nm of the membrane together with its embedded proteins, plus another 10 nm layer underneath), 40% of the total cell volume is part of this membrane zone. Our simulations will provide detailed insights into the nature and extent of depletion or crowding layers, and into the level of heterogeneity inside this membrane zone, providing information on the extent to which compounds are either enriched or depleted near the cell surface (Nawrocki et al., 2019). A full-cell membrane model might explain why the minimal cell grows on a diet of both saturated and unsaturated fatty acids, but not on a diet of just saturated ones as observed in lipidomics experiments from the Saenz lab (private communication). A related question is why the cell membrane contains such a high percentage of cholesterol (20%–60% dependent on growth medium); this is uncommon for bacterial membranes although generally *Mycoplasma* do contain some cholesterol for membrane stability.

Of special interest is the potential existence of dynamic highways, i.e., regions in the cell with greater mobility of the constituents, which may arise from crowding effects or liquid-liquid phase separation phenomena, or may be induced by proximity of the cell membrane. Such dynamic highways could be important in regulating transport in an otherwise glassy state of the cytoplasm. For regions of the cell showing particularly interesting behaviour, smaller systems can be extracted with the advanced TS2CG tool and simulated for extended timescales to increase the statistical relevance. Besides passively studying the cellular environment, holistic cell modeling poses the ideal computational sandbox in which we can introduce new components to the cellular environment. For instance, elucidating the non-specific interactions between the cytosol and drug candidates and showing how drug-receptor interactions affect the entire cell instead of just the receptor site.

Using a multiscale modeling approach, we could potentially explore cell dynamics at various stages in its life cycle. Compared to MD simulations, other low-resolution modeling approaches can more broadly explore timescales of several orders of magnitude longer. Integrating other computational models will make it possible to sprout MD simulations in interesting regimes observed with the lower-resolution models. The primary computational method we will focus on integrating into our framework is the whole-cell fully dynamical kinetic model developed by the Luthey-Schulten lab, which accounts for the metabolic pathways governing the cellular processes (Thornburg et al., 2022). By transferring structural information from the kinetic model into our high-resolution model, it will be possible to paint a more detailed picture of the cell's internal organization and dynamics at specific points of the cell's life cycle, including during cell fission.

Since most of the tools in the Martini ecosystem are force field agnostic, the workflow can also be applied to generate all-atom whole-cell models. Given the substantial increase in associated computational costs, it might be a wiser approach to only sample smaller subsystems at the all-atom level. These could be straightforwardly obtained from backmapping representative regions taken from the whole-cell CG model. A number of such backmapping tools, optimized for Martini, already exist (Louison et al., 2021; Vickery and Stansfeld, 2021; López et al., 2022).

A final challenge lies in the analysis and interpretation of the complex high-dimensional massive data that will be generated. Clearly, it will be impossible to perform a comprehensive analysis on a whole-cell trajectory, and one needs to focus on specific research questions. However, the trajectories can nowadays be easily shared with the broader community *via* dedicated open-access repositories

such as Zenodo (<https://zenodo.org/>), allowing others to perform whatever additional analysis they fancy. One can also envision the usage of data reduction schemes to efficiently analyse the whole-cell simulation. One possibility is storing only centers-of-mass movement of the non-aqueous components, which would facilitate the analysis of diffusional behavior, for instance. Another approach would be using a voxel-based method (Bruininks et al., 2021) to dynamically segment the whole-cell model into similarity regions, e.g., membrane periphery or chromosomal region. The system segmentation would allow for efficient quantitative comparison of the cytosolic properties within and between distinct regions of the cell. Furthermore, machine-learning can be invoked to extract interaction patterns and other emergent behavior that might be missed by standard analysis tools (Noé et al., 2020; Wang et al., 2020; Kaptan and Vattulainen, 2022). We foresee that our data sets will generate novel ways of dealing with this unprecedented level of complexity.

In conclusion, we presented a roadmap toward whole-cell MD simulations, illustrated with the construction of the first MD model of an entire cell using our Martini ecosystem. The model represents a next level realized with the computational microscope, providing a complete picture of the cell and making it possible to relate molecular structures and interactions to cellular function directly. In the long term, our computational framework will enable us to study a wide variety of mesoscopic systems, possibly informing the design of fully synthetic cells (Olivi et al., 2021) and modeling cells with more complex internal structures.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://github.com/marrink-lab/>.

References

- Abraham, M. J., Murtola, T., Schulz, R., Pall, S., Smith, J. C., Hess, B., et al. (2015). 'GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers'. *SoftwareX* 1–2, 19–25. doi:10.1016/j.softx.2015.06.001
- Alessandri, R., Barnoud, J., Gertsen, A. S., Patmanidis, I., de Vries, A. H., Souza, P. C. T., et al. (2022). Martini 3 coarse-grained force field: Small molecules. *Adv. Theory Simulations* 5 (1), 2100391. doi:10.1002/adts.202100391
- Ando, T., Bhamidimarri, S. P., Brending, N., Colin-York, H., Collinson, L., De Jonge, N., et al. (2018). The 2018 correlative microscopy techniques roadmap. *Correl. Microsc. Tech. roadmap*, *J. Phys. D Appl. Phys.* 51 (44), 443001. doi:10.1088/1361-6463/aad055
- Bereau, T., and Kremer, K. (2015). Automated parametrization of the coarse-grained Martini force field for small organic molecules. *J. Chem. Theory Comput.* 11 (6), 2783–2791. doi:10.1021/acs.jctc.5b00056
- Bhat, N. G., and Balaji, S. (2020). Whole-cell modeling and simulation: A brief survey. *New Gener. Comput.* 38 (1), 259–281. doi:10.1007/s00354-019-00066-y
- Bianchi, D. M., Pelletier, J. F., Hutchison, C. A., Glass, J. I., and Luthey-Schulten, Z. (2022). 'Toward the complete functional characterization of a minimal bacterial proteome'. *J. Phys. Chem. B* 126 (36), 6820–6834. doi:10.1021/acs.jpcc.2c04188
- Blender Online Community (2022). 'Blender - a 3D modelling and rendering package'. Amsterdam: Blender Foundation. Stichting Blender Foundation. Available at: <http://www.blender.org>
- Bonvin, A. M. J. J. (2021). 50 years of PDB: A catalyst in structural biology. *Nat. Methods* 18 (5), 448–449. doi:10.1038/s41592-021-01138-y
- Brackley, C. A., Morozov, A. N., and Marenduzzo, D. (2014). 'Models for twistable elastic polymers in Brownian dynamics, and their implementation for LAMMPS'. *J. Chem. Phys.* 140 (13), 135103. doi:10.1063/1.4870088
- Breuer, M., Earnest, T. M., Merryman, C., Wise, K. S., Sun, L., Lynott, M. R., et al. (2019). 'Essential metabolism for a minimal cell'. *eLife* 8, e36842. doi:10.7554/eLife.36842
- Bruininks, B. M. H., Thie, A. S., Souza, P. C. T., Wassenaar, T. A., Faraji, S., and Marrink, S. J. (2021). 'Sequential voxel-based leaflet segmentation of complex lipid morphologies'. *J. Chem. Theory Comput.* 17 (12), 7873–7885. doi:10.1021/acs.jctc.1c00446
- Castagna, J., Guo, X., Seaton, M., and O'Cais, A. (2020). Towards extreme scale dissipative particle dynamics simulations using multiple GPGPUs. *Comput. Phys. Commun.* 251, 107159. doi:10.1016/j.cpc.2020.107159
- Cheng, Y. (2018). 'Single-particle cryo-EM—how did it get here and where will it go'. *Science* 361 (6405), 876–880. doi:10.1126/science.aat4346
- Chorev, D. S., Baker, L. A., Wu, D., Beilsten-Edmands, V., Rouse, S. L., Zeev-Ben-Mordehai, T., et al. (2018). 'Protein assemblies ejected directly from native membranes yield complexes for mass spectrometry'. *Science* 362 (6416), 829–834. doi:10.1126/science.aau0976
- Christie, S., Shi, X., and Smith, A. W. (2020). 'Resolving membrane protein–protein interactions in live cells with pulsed interleaved excitation fluorescence cross-correlation spectroscopy'. *Accounts Chem. Res.* 53 (4), 792–799. doi:10.1021/acs.accounts.9b00625
- Corradi, V., Mendez-Villuendas, E., Ingolfsson, H. I., Gu, R. X., Siuda, I., Melo, M. N., et al. (2018). 'Lipid–Protein interactions are unique fingerprints for membrane proteins'. *ACS Central Sci.* 4 (6), 709–717. doi:10.1021/acscentsci.8b00143
- de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schafer, L. V., et al. (2013). 'Improved parameters for the Martini coarse-grained protein force field'. *J. Chem. Theory Comput.* 9 (1), 687–697. doi:10.1021/ct300646g
- Dommer, A., Casalino, L., Kearns, F., Rosenfeld, M., Wauer, N., Ahn, S. H., et al. (2022). 'COVIDisAirborne: AI-enabled multiscale computational microscopy of delta SARS-CoV-2 in a respiratory aerosol'. *Int. J. High Perform. Comput. Appl.*, 109434202211282. doi:10.1177/10943420221128233

Author contributions

SM and ZL-S conceived of the project. JS and FG constructed the model with help of PT, MK, BG, TB, and ZT. JS, FG, and SM wrote the manuscript with input from all authors.

Funding

SM acknowledges funding from the ERC with the Advanced grant 101053661 ("COMP-O-CELL"), and funding from NWO through the NWA grant "The limits to growth: The challenge to dissipate energy" and BaSys ("Building a Synthetic Cell") consortium. ZL-S acknowledges funding from NSF (MCB: 2221237 "Simulating a growing minimal cell: Integrating experiment and theory"; PHY: 1430124 "Center for the Physics of Living Cells"; PHY: 1505008 and 2014027 "Collaborative Research Network from Physics of Living Systems").

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Dror, R. O., Dirks, R. M., Grossman, J., Xu, H., and Shaw, D. E. (2012). Biomolecular simulation: A computational microscope for molecular biology. *Annu. Rev. Biophysics* 41 (1), 429–452. doi:10.1146/annurev-biophys-042910-155245
- Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., et al. (2017). 'OpenMM 7: Rapid development of high performance algorithms for molecular dynamics'. *PLOS Comput. Biol.* 13 (7), e1005659. doi:10.1371/journal.pcbi.1005659
- Eilers, Y., Ta, H., Gwosch, K. C., Balzarotti, F., and Hell, S. W. (2018). 'MINFLUX monitors rapid molecular jumps with superior spatiotemporal resolution'. *Proc. Natl. Acad. Sci.* 115 (24), 6117–6122. doi:10.1073/pnas.1801672115
- Ellis, R. J., and Minton, A. P. (2003). 'Join the crowd'. *Nature* 425 (6953), 27–28. doi:10.1038/425027a
- Gilbert, B. R., Thornburg, Z. R., Lam, V., Rashid, F. Z. M., Glass, J. I., Villa, E., et al. (2021). 'Generating chromosome geometries in a minimal cell from cryo-electron tomograms and chromosome conformation capture maps'. *Front. Mol. Biosci.* 8, 644133. doi:10.3389/fmolb.2021.644133
- Grünwald, F., Alessandri, R., Kroon, P. C., Monticelli, L., Souza, P. C. T., and Marrink, S. J. (2022). PolyPy: a python suite for facilitating simulations of macromolecules and nanomaterials. *Nat. Commun.* 13 (1), 68. doi:10.1038/s41467-021-27627-4
- Grünwald, F., Punt, M. H., Jefferys, E. E., Vainikka, P. A., König, M., Virtanen, V., et al. (2022). 'Martini 3 coarse-grained force field for carbohydrates'. *J. Chem. Theory Comput.* 18, 7555–7569. doi:10.1021/acs.jctc.2c00757
- Guilhas, B., Walter, J. C., Rech, J., David, G., Walliser, N. O., Palmeri, J., et al. (2020). 'ATP-Driven separation of liquid phase condensates in bacteria'. *Mol. Cell* 79 (2), 293–303.e4. doi:10.1016/j.molcel.2020.06.034
- Gupta, C., Sarkar, D., Tieleman, D. P., and Singharoy, A. (2022). 'The ugly, bad, and good stories of large-scale biomolecular simulations'. *Curr. Opin. Struct. Biol.* 73, 102338. doi:10.1016/j.sbi.2022.102338
- Hilpert, C., Beranger, L., Souza, P. C., Vainikka, P. A., Nieto, V., Marrink, S. J., et al. (2023). Facilitating CG simulations with MAD: The Martini Database server. *J. Chem. Inf. Model.* doi:10.1021/acs.jcim.2c01375
- Howard, M. P., Anderson, J. A., Nikoubashman, A., Glotzer, S. C., and Panagiotopoulos, A. Z. (2016). 'Efficient neighbor list calculation for molecular simulation of colloidal systems using graphics processing units'. *Comput. Phys. Commun.* 203, 45–52. doi:10.1016/j.cpc.2016.02.003
- Hutchison, C. A., Chuang, R. Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., et al. (2016). 'Design and synthesis of a minimal bacterial genome'. *Science* 351, aad6253. doi:10.1126/science.aad6253
- Johnson, G. T., Autin, L., Al-Alusi, M., Goodsell, D. S., Sanner, M. F., and Olson, A. J. (2015). cellPACK: A virtual mesoscope to model and visualize structural systems biology. *Nat. Methods* 12 (1), 85–91. doi:10.1038/nmeth.3204
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). 'Highly accurate protein structure prediction with AlphaFold'. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2
- Jung, J., Nishima, W., Daniels, M., Bascom, G., Kobayashi, C., Adedoyin, A., et al. (2019). 'Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations'. *J. Comput. Chem.* 40 (21), 1919–1930. doi:10.1002/jcc.25840
- Kaptan, S., and Vattulainen, I. (2022). 'Machine learning in the analysis of biomolecular simulations'. *Adv. Phys. X* 7 (1), 2006080. doi:10.1080/23746149.2021.2006080
- Karr, J. R., Sanghvi, J., Macklin, D., Gutschow, M., Jacobs, J., Bolival, B., et al. (2012). 'A whole-cell computational model predicts phenotype from genotype'. *Cell* 150 (2), 389–401. doi:10.1016/j.cell.2012.05.044
- Karr, J. R., Takahashi, K., and Funahashi, A. (2015). 'The principles of whole-cell modeling'. *Curr. Opin. Microbiol.* 27, 18–24. doi:10.1016/j.mib.2015.06.004
- Khalid, S., and Rouse, S. L. (2020). 'Simulation of subcellular structures'. *Curr. Opin. Struct. Biol.* 61, 167–172. doi:10.1016/j.sbi.2019.12.017
- Kroon, P. C., Grünwald, F., Barnoud, J., van Tilburg, M., Souza, P. C. T., and Marrink, S. J. (2022). 'Martinize2 and vermouth: Unified framework for topology generation'. arXiv. doi:10.48550/arXiv.2212.01191
- Lee, E. H., Hsin, J., Sotomayor, M., Comellas, G., and Schulten, K. (2009). 'Discovery through the computational microscope'. *Structure* 17 (10), 1295–1306. doi:10.1016/j.str.2009.09.001
- López, C. A., Rzepiela, A. J., de Vries, A. H., Dijkhuizen, L., Hunenberger, P. H., and Marrink, S. J. (2009). Martini coarse-grained force field: Extension to carbohydrates. *J. Chem. Theory Comput.* 5 (12), 3195–3210. doi:10.1021/ct900313w
- López, C. A., Zhang, X., Aydin, F., Shrestha, R., Van, Q. N., Stanley, C. B., et al. (2022). 'Asynchronous reciprocal coupling of Martini 2.2 coarse-grained and CHARMM36 all-atom simulations in an automated multiscale framework'. *J. Chem. Theory Comput.* 18 (8), 5025–5045. doi:10.1021/acs.jctc.2c00168
- Lorent, J. H., Levental, K. R., Ganesan, L., Rivera-Longworth, G., Sezgin, E., Doktorova, M., et al. (2020). 'Plasma membranes are asymmetric in lipid unsaturation, packing and protein shape'. *Nat. Chem. Biol.* 16 (6), 644–652. doi:10.1038/s41589-020-0529-6
- Louison, K. A., Dryden, I. L., and Laughton, C. A. (2021). Glimps: A machine learning approach to resolution transformation for multiscale modeling. *J. Chem. Theory Comput.* 17 (12), 7930–7937. doi:10.1021/acs.jctc.1c00735
- Luthey-Schulten, Z. (2021). 'Integrating experiments, theory and simulations into whole-cell models'. *Nat. Methods* 18 (5), 446–447. doi:10.1038/s41592-021-01150-2
- Luthey-Schulten, Z., Thornburg, Z. R., and Gilbert, B. R. (2022). 'Integrating cellular and molecular structures and dynamics into whole-cell models'. *Curr. Opin. Struct. Biol.* 75, 102392. doi:10.1016/j.sbi.2022.102392
- Macklin, D. N., Ahn-Horst, T. A., Choi, H., Ruggero, N. A., Carrera, J., Mason, J. C., et al. (2020). Simultaneous cross-evaluation of heterogeneous *E. coli* datasets via mechanistic simulation. *Science* 369, eaav3751. doi:10.1126/science.aav3751
- Macklin, D. N., Ruggero, N. A., and Covert, M. W. (2014). 'The future of whole-cell modeling'. *Curr. Opin. Biotechnol.* 28, 111–115. doi:10.1016/j.copbio.2014.01.012
- Maritan, M., Autin, L., Karr, J., Covert, M. W., Olson, A. J., and Goodsell, D. S. (2022). 'Building structural models of a whole Mycoplasma cell'. *J. Mol. Biol.* 434 (2), 167351. doi:10.1016/j.jmb.2021.167351
- Marrink, S. J., Monticelli, L., Melo, M. N., Alessandri, R., Tieleman, D. P., Souza, P. C. T., et al. (2022). Two decades of Martini: Better beads, broader scope. *WIREs Comput. Mol. Sci.*, e1620. doi:10.1002/wcms.1620
- Marrink, S. J., Corradi, V., Souza, P. C., Ingolfsson, H. I., Tieleman, D. P., and Sansom, M. S. (2019). 'Computational modeling of realistic cell membranes'. *Chem. Rev.* 119 (9), 6184–6226. doi:10.1021/acs.chemrev.8b00460
- marrink-lab (2022). 'Martini_Minimal_Cell'. Available at https://github.com/marrink-lab/Martini_Minimal_Cell.
- McGuffee, S. R., and Elcock, A. H. (2010). Diffusion, crowding and protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.* 6 (3), e1000694. doi:10.1371/journal.pcbi.1000694
- Mondal, J., Bratton, B. P., Li, Y., Yethiraj, A., and Weisshaar, J. (2011). 'Entropy-Based mechanism of ribosome-nucleoid segregation in *E. coli* cells'. *Biophysical J.* 100 (11), 2605–2613. doi:10.1016/j.bpj.2011.04.030
- Mosalaganti, S., Obarska-Kosinska, A., Siggel, M., Taniguchi, R., Turanova, B., Zimmerli, C. E., et al. (2022). 'AI-based structure prediction empowers integrative structural analysis of human nuclear pores'. *Science* 376, eabm9506. doi:10.1126/science.abm9506
- Narasimhan, S., Folkers, G. E., and Baldus, M. (2020). When small becomes too big: Expanding the use of in-cell solid-state NMR spectroscopy. *ChemPlusChem* 85 (4), 760–768. doi:10.1002/cplu.202000167
- Nawrocki, G., Im, W., Sugita, Y., and Feig, M. (2019). 'Clustering and dynamics of crowded proteins near membranes and their influence on membrane bending'. *Proc. Natl. Acad. Sci.* 116 (49), 24562–24567. doi:10.1073/pnas.1910771116
- Noé, F., Tkatchenko, A., Müller, K. R., and Clementi, C. (2020). 'Machine learning for molecular simulation'. *Annu. Rev. Phys. Chem.* 71 (1), 361–390. doi:10.1146/annurev-physchem-042018-052331
- Oliveira Bortot, L., Bashardaneh, Z., and van der Spoel, D. (2020). Making soup: Preparing and validating models of the bacterial cytoplasm for molecular simulation. *J. Chem. Inf. Model.* 60 (1), 322–331. doi:10.1021/acs.jcim.9b00971
- Olivi, L., Berger, M., Creighton, R. N. P., De Franceschi, N., Dekker, C., Mulder, B. M., et al. (2021). 'Towards a synthetic cell cycle'. *Nat. Commun.* 12 (1), 4531. doi:10.1038/s41467-021-24772-8
- Pezeshkian, W., Grünwald, F., Narykov, O., Lu, S., Arkhipova, V., Solodovnikov, A., et al. (2021). Molecular architecture and dynamics of SARS-CoV-2 envelope by integrative modeling. *Biophysics*. doi:10.1101/2021.09.15.459697
- Pezeshkian, W., König, M., Wassenaar, T. A., and Marrink, S. J. (2020). 'Backmapping triangulated surfaces to coarse-grained membrane models'. *Nat. Commun.* 11 (1), 2296. doi:10.1038/s41467-020-16094-y
- Poma, A. B., Cieplak, M., and Theodorakis, P. E. (2017). 'Combining the MARTINI and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins'. *J. Chem. Theory Comput.* 13 (3), 1366–1374. doi:10.1021/acs.jctc.6b00986
- Potter, T. D., Barrett, E. L., and Miller, M. A. (2021). 'Automated coarse-grained mapping algorithm for the Martini force field and benchmarks for membrane-water partitioning'. *J. Chem. Theory Comput.* 17 (9), 5777–5791. doi:10.1021/acs.jctc.1c00322
- Reading, E., Hall, Z., Martens, C., Haghighi, T., Findlay, H., Ahdash, Z., et al. (2017). 'Interrogating membrane protein conformational dynamics within native lipid compositions'. *Angew. Chem. Int. Ed.* 56 (49), 15654–15657. doi:10.1002/anie.201709657
- Rhine, K., Vidaurre, V., and Myong, S. (2020). 'RNA droplets'. *Annu. Rev. Biophysics* 49 (1), 247–265. doi:10.1146/annurev-biophys-052118-115508
- Roberts, E., Stone, J. E., and Luthey-Schulten, Z. (2013). Lattice microbes: High-performance stochastic simulation method for the reaction-diffusion master equation. *J. Comput. Chem.* 34 (3), 245–255. doi:10.1002/jcc.23130
- Schmidt, R., Weihs, T., Wurm, C. A., Jansen, I., Rehman, J., Sahl, S. J., et al. (2021). 'MINFLUX nanometer-scale 3D imaging and microsecond-range tracking on a common fluorescence microscope'. *Nat. Commun.* 12 (1), 1478. doi:10.1038/s41467-021-21652-z
- Singharoy, A., Maffeo, C., Delgado-Magnero, K. H., Swainsbury, D. J., Sener, M., Kleinekathofer, U., et al. (2019). Atoms to phenotypes: Molecular design principles of cellular energy metabolism. *Cell* 179, 1098–1111.e23. doi:10.1016/j.cell.2019.10.021
- Sousa, F. M., Lima, L. M. P., Arnarez, C., Pereira, M. M., and Melo, M. N. (2021). 'Coarse-Grained parameterization of nucleotide cofactors and metabolites: Coarse-

- grained parameterization of nucleotide cofactors and metabolites: Protonation constants, partition coefficients, and model topologies. *J. Chem. Inf. Model.* 61 (1), 335–346. doi:10.1021/acs.jcim.0c01077
- Souza, P. C. T., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grunewald, F., et al. (2021). Martini 3: A general purpose force field for coarse-grained molecular dynamics. *Nat. Methods* 18 (4), 382–388. doi:10.1038/s41592-021-01098-3
- Souza, P. C. T., Thallmair, S., Marrink, S. J., and Mera-Adasme, R. (2019). ‘An allosteric pathway in copper, an allosteric pathway in copper, zinc superoxide dismutase unravels the molecular mechanism of the G93A amyotrophic lateral sclerosis-linked mutation. *J. Phys. Chem. Lett.* 10 (24), 7740–7744. doi:10.1021/acs.jpclett.9b02868
- Štefl, M., Herbst, K., Rubsam, M., Benda, A., and Knop, M. (2020). ‘Single-Color fluorescence lifetime cross-correlation spectroscopy *in vivo*’. *Biophysical J.* 119 (7), 1359–1370. doi:10.1016/j.bpj.2020.06.039
- Thornburg, Z. R., Bianchi, D. M., Brier, T. A., Gilbert, B. R., Earnest, T. M., Melo, M. C., et al. (2022). ‘Fundamental behaviors emerge from simulations of a living minimal cell’. *Cell* 185 (2), 345–360.e28. doi:10.1016/j.cell.2021.12.025
- Uusitalo, J. J., Ingolfsson, H. I., Akhshi, P., Tieleman, D. P., and Marrink, S. J. (2015). Martini coarse-grained force field: Extension to DNA. *J. Chem. Theory Comput.* 11 (8), 3932–3945. doi:10.1021/acs.jctc.5b00286_FILE/CT5B00286_SI_001.PDF
- Uusitalo, J. J., Ingolfsson, H. I., Marrink, S. J., and Faustino, I. (2017). Martini coarse-grained force field: Extension to RNA. *Biophysical J.* 113 (2), 246–256. doi:10.1016/j.bpj.2017.05.043
- Vermaas, J. V., Mayne, C. G., Shinn, E., and Tajkhorshid, E. (2022). ‘Assembly and analysis of cell-scale membrane envelopes’. *J. Chem. Inf. Model.* 62 (3), 602–617. doi:10.1021/acs.jcim.1c01050
- Vickery, O. N., and Stansfeld, P. J. (2021). CG2AT2: An enhanced fragment-based approach for serial multi-scale molecular dynamics simulations. *J. Chem. Theory Comput.* 17 (10), 6472–6482. doi:10.1021/acs.jctc.1c00295
- Wang, W., Juttler, B., Zheng, D., and Liu, Y. (2008). ‘Computation of rotation minimizing frames’. *ACM Trans. Graph. (TOG)* 18 (1), 1. doi:10.1145/1330511.1330513
- Wang, Y., Lamim Ribeiro, J. M., and Tiwary, P. (2020). ‘Machine learning approaches for analyzing and enhancing molecular dynamics simulations’. *Curr. Opin. Struct. Biol.* 61, 139–145. doi:10.1016/j.sbi.2019.12.016
- Wassenaar, T. A., Ingolfsson, H. I., Bockmann, R. A., Tieleman, D. P., and Marrink, S. J. (2015). Computational lipidomics with insane: A versatile tool for generating custom membranes for molecular simulations. *J. Chem. Theory Comput.* 11 (5), 2144–2155. doi:10.1021/acs.jctc.5b00209
- Wietrzynski, W., Schaffer, M., Tegunov, D., Albert, S., Kanazawa, A., Plitzko, J. M., et al. (2020). ‘Charting the native architecture of Chlamydomonas thylakoid membranes with single-molecule precision’. *eLife* 9, e53740. doi:10.7554/eLife.53740
- Yu, I., Mori, T., Ando, T., Harada, R., Jung, J., Sugita, Y., et al. (2016). ‘Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm’. *eLife* 5, e19274. doi:10.7554/ELIFE.19274
- Zhang, X., Sundram, S., Oppelstrup, T., Kokkila-Schumacher, S. I. L., Carpenter, T. S., Ingolfsson, H. I., et al. (2020). ddcMD: A fully GPU-accelerated molecular dynamics program for the Martini force field. *J. Chem. Phys.* 153 (4), 045103. doi:10.1063/5.0014500



OPEN ACCESS

EDITED BY

Adolfo Poma,
Institute of Fundamental Technological
Research, Polish Academy of Sciences,
Poland

REVIEWED BY

Adam Liwo,
University of Gdansk, Poland
Thu Tran,
Vietnam National University, Vietnam

*CORRESPONDENCE

Mateusz Chwastyk,
✉ chwastyk@ifpan.edu.pl

[†]Deceased

SPECIALTY SECTION

This article was submitted to Recent
Advances in Computational Modelling of
Biomolecular Complexes,
a section of the journal
Frontiers in Chemistry

RECEIVED 23 November 2022

ACCEPTED 29 December 2022

PUBLISHED 25 January 2023

CITATION

Pham DQH, Chwastyk M and Cieplak M
(2023), The coexistence region in the Van
der Waals fluid and the liquid-liquid
phase transitions.
Front. Chem. 10:1106599.
doi: 10.3389/fchem.2022.1106599

COPYRIGHT

© 2023 Pham, Chwastyk and Cieplak. 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.

The coexistence region in the Van der Waals fluid and the liquid-liquid phase transitions

Dinh Quoc Huy Pham, Mateusz Chwastyk* and Marek Cieplak[†]

Institute of Physics, Polish Academy of Sciences, Warsaw, Poland

Cellular membraneless organelles are thought to be droplets formed within the two-phase region corresponding to proteinaceous systems endowed with the liquid-liquid transition. However, their metastability requires an additional constraint—they arise in a certain region of density and temperature between the spinodal and binodal lines. Here, we consider the well-studied van der Waals fluid as a test model to work out criteria to determine the location of the spinodal line for situations in which the equation of state is not known. Our molecular dynamics studies indicate that this task can be accomplished by considering the specific heat, the surface tension and characteristics of the molecular clusters, such as the number of component chains and radius of gyration.

KEYWORDS

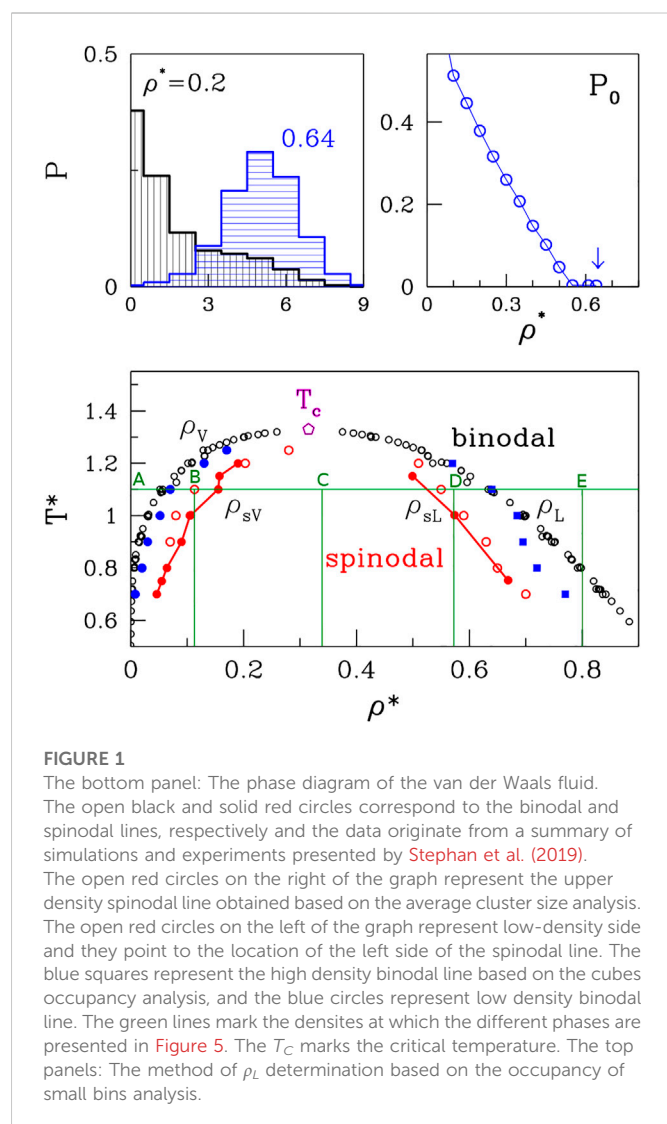
Van der Waals fluid, phase diagram, liquid-liquid phase transitions, intrinsically disordered proteins (IDPs), molecular dynamics simulations (MD)

1 Introduction

Cellular organelles can be either membraneless or membrane-bound. The membranes arise as droplets during a liquid-liquid phase transition (Brangwynne et al., 2009; Brangwynne et al., 2011; Shin and Brangwynne, 2017; Boeynaems et al., 2018; Elbaum-Garfinkle, 2019) as a result of thermal fluctuations. These biological droplets can be micrometers in size and exhibit hydrodynamical characteristics such as fusion (Caragine et al., 2018; Caragine and Haley, 2019). Proteinaceous liquids involved in the phase transition have been found to be composed primarily of intrinsically disordered proteins (IDPs) (Uversky, 2002; Dyson and Wright, 2005; Fink, 2005; Dunker et al., 2008; Ferreon et al., 2010; Uversky and Dunker, 2010; Babu et al., 2011; Wright and Dyson, 2015; Banani et al., 2017; Chwastyk and Cieplak, 2020; de Aquino et al., 2020) that allow for a multitude of ways to bind and aggregate.

The droplets may form only within the coexistence region of the phase diagram of the two fluids but their functionality requires that they are metastable. The paradigm model that yields such a coexistence region is the van der Waals (vdW) fluid as described by the well-known equation of state that generalizes the perfect gas law. In the density (ρ)—temperature (T) plane, the phase diagram of the vdW fluid includes the coexistence region of gas and liquid that is bounded by the inverted parabola, as shown in the bottom panel of Figure 1. Its vertex corresponds to the critical temperature (T_c) above which one cannot distinguish between the two phases. Such a phase diagram can be obtained for the system of n_m monatomic particles that interact through the 6–12 Lennard-Jones (LJ) potential (Hensen and McDonald, 1973) given by:

$$\Phi_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right], \quad (1)$$



where ε and σ are the uniform energy and length parameters. A significant increase in ρ , at any T , results in solidification. A sufficient decrease in T yields a similar effect. The solids may have several kinds of symmetry and the more complete schematic phase diagram can be found in ref. (Schultz and Kofke, 2018).

It should be noted that the separation into the two phases in the coexistence region can occur either through nucleation or by spinodal decomposition, depending on ρ and T . Both processes can be triggered by quenching from a temperature above T_c , but have a different physical mechanism. Nucleation arises as a result of a rare but large energy fluctuation and is associated with metastability (Frenkel, 1955; Feder et al., 1966; Abraham, 1975; Binder and Stauffer, 1976). On the other hand, phase separation through spinodal decomposition takes place in an initially unstable system in which all fluctuations grow because there is no energy barrier (Cahn and Hilliard, 1958; Cahn and Hilliard, 1971; Langer, 1971; Huang et al., 1974; Binder et al., 1978). Thus, the generation of metastable droplets can take place only between the binodal and spinodal lines. The spinodal line for the vdW system is also an inverted parabola that is placed within the coexistence region (cf. The bottom panel in Figure 1). The region within the spinodal line is chaotic, unstable, and beyond a

thermodynamic description. Any short-lived clusters of atoms there cannot be analogues of the “organelles.” Thus, the determination of the proper conditions for the droplet formation involves figuring out not only the position of the binodal line but also of the spinodal boundary. It should be noted that droplets of a higher (lower) density than the environment arise in the region that borders with the gas (liquid) phase. The biophysical context assumes the higher density situation.

In the absence of theoretically validated equations of state for the protein solutions, we resort to considering a simpler system: a homogeneous vdW fluid. This will allow us to test the novel concepts related to the determination of the phase diagram, giving us much-needed insight on how to deal with more complicated situations. In principle, for the vdW fluid, one can derive the free energy of the system, consistent with the equation of state, and analyze its stability. Our purpose, however, is to find alternative ways to locate the spinodal and binodal lines that could be used in molecular dynamics simulations of proteins.

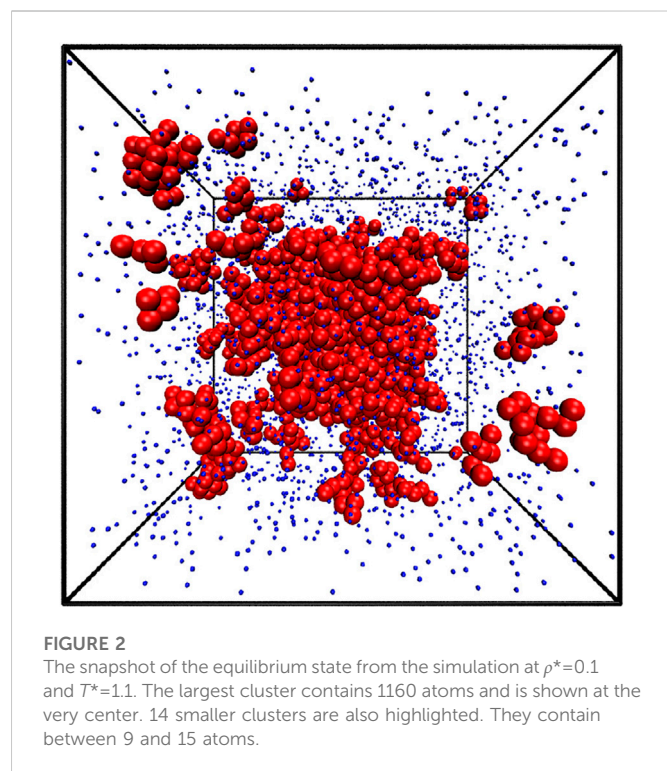
2 The phase diagram construction

A series of simulations and experimental studies (Nicolas et al., 1979; Panagiotopoulos, 1994; Baidakov et al., 2000) of the van der Waals fluids has been reviewed by Stephan et al. (Stephan et al., 2019) and the data shown in the bottom panel of Figure 1 is based on this reference. The results are presented in reduced units (the symbols are denoted by an asterisk) that involve the length parameter, σ , and the depth of the energy well, ε . Density is given in units of the number of monomers per σ^3 . For the cutoff value of 6.85σ , the critical point is at temperature T_c of 1.31 and density $\rho_c = 0.316$ that is consistent with a direct analysis of the equation of state.

The theoretical and experimental data (Stephan et al., 2019) were the references for our simulations. We performed molecular dynamics simulations for two systems of 4,000 particles: one for 4,000 non-bonded particles and the other one for 200 20-bead chains. During our simulations we monitored the cluster sizes, appearance of cavities and their volumes (Chwastyk et al., 2014a; Chwastyk et al., 2016), specific heat (Chwastyk et al., 2015; Chwastyk et al., 2017) as well as a number of other parameters to find a way to determine the phase diagram.

2.1 Details of the simulations

Our simulations were conducted by using the LAMMPS software package (Plimpton, 1995). The cut-off for LJ potential was at a distance of $r_c = 6.85\sigma$. We used the Verlet algorithm to integrate the equations of motion. The time was measured in units of $\tau_{LJ} \equiv \sqrt{m\sigma^2/\varepsilon}$, where m is the mass of each particle. This time unit corresponds to the characteristic period of undamped oscillations at the bottom of a 6–12 potential (Chwastyk et al., 2014b; Zhao et al., 2017a; Zhao et al., 2017b). We used the integration step of $\Delta t = 0.005\tau_{LJ}$ for the simulations of monomers and $\Delta t = 0.001\tau_{LJ}$ for chains. The length of our simulations was 1 000 000 and 5 000 000 steps for monomers and chains, respectively, which corresponds to 5000 τ_{LJ} of total simulation time. The trajectory analysis was done based on the last 1666 τ_{LJ} of the simulation in each case and the other part of the simulation was the equilibration. Two atoms were considered to be in contact if the distance between them does not exceed 1.3σ . Two chains



are treated as belonging to the same cluster when they have at least one inter-chain contact. As a consequence, we define a cluster as a group of beads (in the case of the simulation of monomers) or chains connected by at least one contact. The chain is defined as a line of monomers connected by harmonic potential:

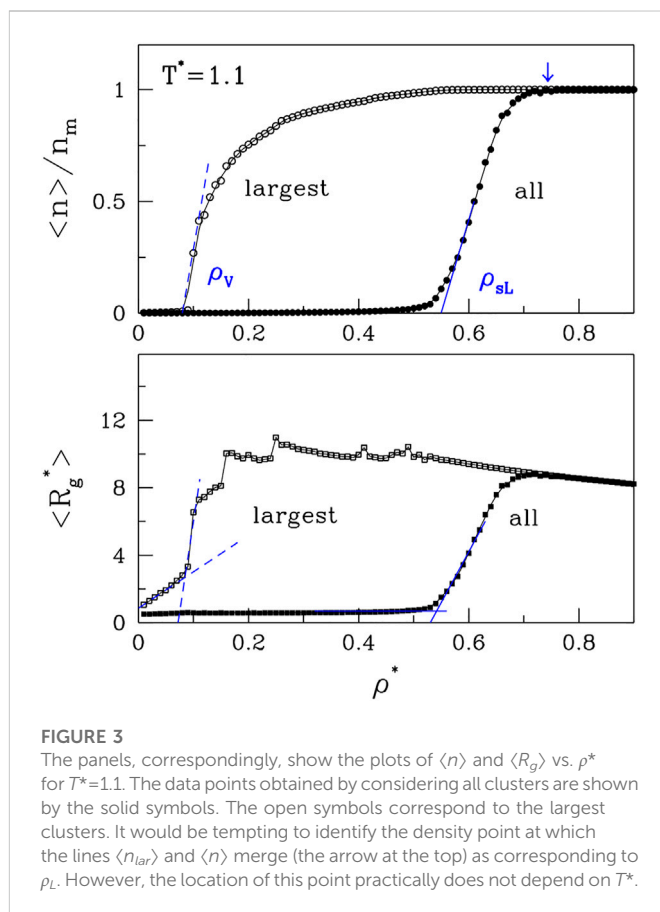
$$U_{\text{bond}}(r) = k_b(r - \sigma)^2, \quad (2)$$

where $k_b = 75000\epsilon/\sigma^2$ is a force constant, strong enough to keep two atoms at distance of σ . This assumption was established according to the results of Kevin S. Silmore *et al.* (Silmore *et al.*, 2017). We used the canonical ensemble (NVT) and the temperature was controlled by Nose-Hoover thermostat with damping parameter of $1.0\tau_{LJ}$ for monomers and $10.0\tau_{LJ}$ for chains.

2.2 The simulations results

Our simulations were conducted for 90 different densities from $\rho^* = 0.01$ to 0.90 , at nine different temperatures for monomers: $T^* \in \{0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.25, 1.3\}$ and eight temperatures: $T^* \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ for chains. The example atomic configuration is presented in Figure 2 for $\rho^* = 0.1$ and $T^* = 1.1$. The largest cluster composed of 1160 atoms is positioned in the center and 14 smaller clusters are highlighted by red larger balls. They contain from 5 to 15 atoms.

The top panels of Figure 1 pertain to our method of determination of the high-density branch, ρ_L , of the binodal line. The idea is to divide the volume into small cubes with a side of order of the size of the molecule. For polymers or proteins, the radius of gyration, r_g would be an appropriate length scale. For monatomic molecules 2σ was found to be optimal. The top left panel shows the probability distributions of the cubic bins to have 0, 1, 2, 3, etc. Atoms. At the low density ($\rho^* = 0.2$)



there is a substantial probability, P_0 , of having an empty bin. The top-right panel shows that P_0 decreases with ρ^* and at around 0.64 it approaches zero. We take this value as defining ρ_L . By considering several other temperatures we get the data points indicated as blue squares. They agree fairly well with the literature results except at the very low T^* s which appear to require longer averaging.

Let us now consider the average cluster (or droplet) sizes. These can be characterized by either the radii of gyration, R_g , or the number of molecules, n , that a droplet contains. The largest possible value of n is n_m . We find it useful to either consider averages over all clusters or only over the largest clusters. In the latter case, the corresponding average size will be denoted by n_{lar} . The results for $\langle n \rangle$ and $\langle n_{lar} \rangle$ are shown in the top panel of Figure 3. We observe that $\langle n_{lar} \rangle$ undergoes a rapid growth at the low density branch of the binodal line, ρ_v , which delineates the vapor phase. The average cluster size of all clusters also undergoes a rapid growth, but at a higher density, ρ_{sl} . The growth coincides with the upper spinodal line (the open red circles in Figure 1). The lower panel shows the corresponding plots for $\langle R_g \rangle$ and $\langle R_{g,lar} \rangle$. They basically mimic the curves related to n except that the growth of R_g for the largest cluster is affected by the fluctuating morphology of the cluster, which affects R_g while not affecting n .

In order to determine ρ_{sv} , the low density branch of the spinodal line, we study the specific heat, C_v . Since C_v is a measure of the energy fluctuations, we would expect volatile energy changes upon entering the non-thermodynamic spinodal region. Indeed, we observe sudden spikes in C_v as a function of ρ , as illustrated in the bottom panel of Figure 4 for $T^* = 1.1$. The tallest of them is on the low-density side and its location is marked by open red circles on the left side of Figure 1.

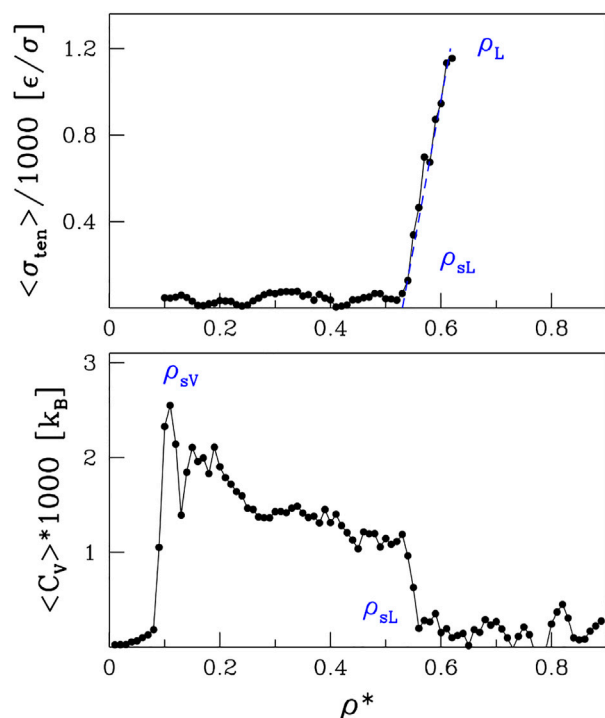


FIGURE 4
The top panel shows the coefficient of the surface tension as a function of ρ^* . The lower panel shows the specific heat. Both panels are for $T = 1.1^*$.

When going to higher densities, there is a sudden drop in C_v that coincides with ρ_{sL} obtained from $\langle n \rangle$.

Yet another way to assess the boundary of the spinodal region is through the surface tension, σ_{ten} . It can be derived, both theoretically and experimentally, by invoking the energy equipartition theorem (Caragine et al., 2018; Caragine and Haley, 2019) leading to $\sigma_{ten} = k_B T / u^2$ where u^2 is the fluctuation in the droplet linear size and k_B is the Boltzmann constant. In molecular dynamics, we take u^2 to be a fluctuation of R_g and average it over time. We perform this procedure for sufficiently large droplets, as they are better defined. However, to avoid the finite-size effects, we do not consider droplets that span the whole system. The surface tension, σ_{ten} , calculated in this manner is shown in the upper panel of Figure 4 for $T^* = 1.1$. The behavior of σ_{ten} as a function of density is rather irregular in the spinodal region but then there a nearly monotonic increase is observed between ρ_{sL} and ρ_L . The phase separation induced by the density changes at $T^* = 1.1$ is presented schematically in Figure 5. The nucleation process can be observed in panels B and D for light and dense phases, respectively. Panels A and E represent one-phase regimes.

3 The phase diagram for chains of monomers

Proteins differ from the Lennard-Jones atoms discussed so far in two major ways: first, their molecules are in the form of chains, and second, the monomers in the chains are of a heterogeneous nature as

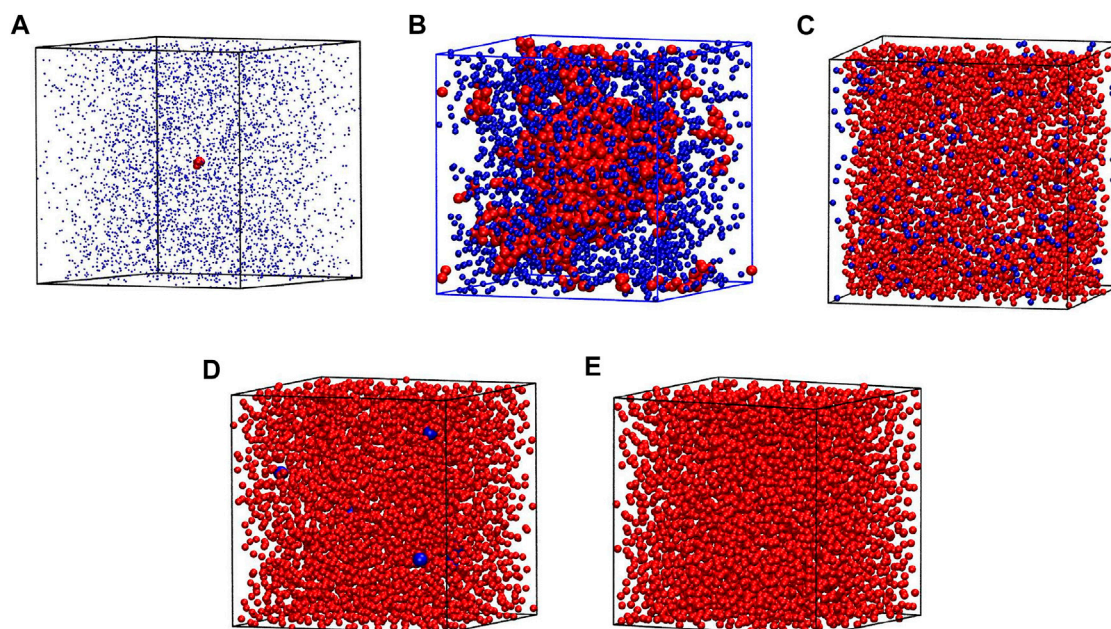
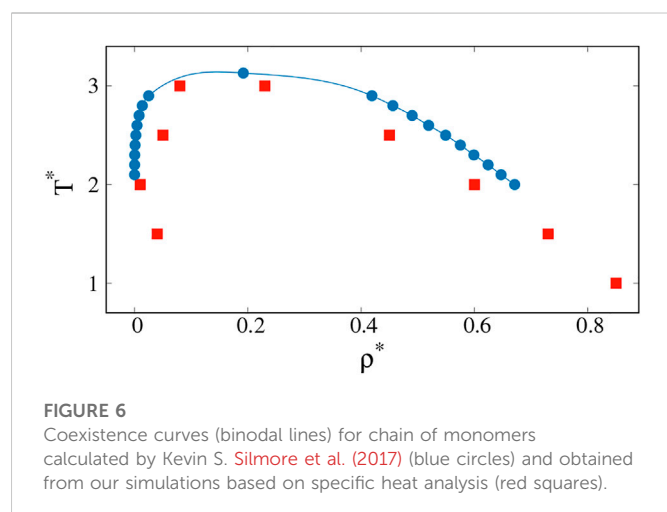


FIGURE 5
The phase separation during the density changes at $T^* = 1.1$ for system composed of 4,000 atoms. The densities of each box are $\rho^* = 0.01, 0.11, 0.35, 0.57$ and 0.8 for boxes (A), (B), (C), (D), and (E), respectively. The clusters are marked by red in panels (A,B). At the dense phase (C,D) the clusters are marked by blue. The positions on the phase diagram of particular cases are marked by green letters in Figure 1.



they represent 20 types of amino acid. The second aspect requires special studies along the lines of Dignon et al. (Dignon et al., 2018a; Dignon et al., 2018b; Dignon et al., 2019) or Mioduszeewski et al. (Mioduszeewski and Cieplak, 2018; Mioduszeewski and Cieplak, 2020; Mioduszeewski and Cieplak, 2021) who use different coarse-grained models to analyze the protein dynamics. In addition, proteins may have inverted binodal lines when hydrophobic effects intervene (Li et al., 2002; Urry et al., 2002; Dignon et al., 2019). We now consider the first of these aspects by performing molecular dynamics simulations for $n_m = 400$ chain molecules of length 20 each. The atoms in the chains are connected at a distance of σ . The binodal lines for this system have been derived by Silmore et al. (Silmore et al., 2017) by using the procedure of Rowlinson and Widom (Rowlinson and Widom, 1982) in which one starts with a dense blob of molecules in the center of an elongated periodic box and reaches a heterogeneous equilibrium. These results are presented by blue points in Figure 6. The chains exhibit more cohesion, and therefore the critical point is moved up in temperature in comparison to the monomeric system.

The clusters that are analogues of the biological droplets are those that should be present immediately to the left of the left branch of the spinodal line, i.e. close to the gas phase. To the right of the right branch of the spinodal line, there are droplets of the low density regions that are essentially like cavities in the liquid phase. The cavities disappear on crossing the binodal line towards the single-component liquid phase. In numerical practice, finding the left branch of the spinodal line can be achieved by considering C_v . This works also for the right-hand side spinodal line but monitoring the surface tension offers an additional tool. Our results for the determination of the spinodal line, based on the C_v analysis are presented by red squares in Figure 6.

4 Conclusion

In principle, a precise determination of both the binodal and spinodal line requires procedures of finite-size scaling. Our purpose here, however, was to determine quantities to accomplish the task

of determining the region in which the metastable droplets could be studied theoretically. In previous theoretical studies (Dignon et al., 2018a; Dignon et al., 2018b; Mioduszeewski and Cieplak, 2018; Dignon et al., 2019; Mioduszeewski and Cieplak, 2020) analyzing proteinaceous droplets, no attempt was made to locate spinodal lines within the two-phase region. The proposed approach should help in such cases, as it allows for the determination of binodal and spinodal line positions for fluids of complex composition.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

All of the authors developed the model, analysed results and wrote the paper. PQ performed the simulations. The manuscript was submitted after the death of MC.

Funding

This research has received support from the National Science Centre (NCN), Poland, under grant No. 2018/31/B/NZ1/00047 and the European H2020 FETOPEN-RIA-2019-01 grant PathoGelTrap No. 899616. The computer resources were supported by the PL-GRID infrastructure. This project is also a part of the European COST Action EUTOPIA.

Acknowledgments

Discussions with P. R. F. de Carvalho and his technical help are appreciated. We are grateful for comments of Piotr Szymczak about the manuscript.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Abraham, F. F. (1975). *Homogeneous nucleation theory*. New York: Academic Press.
- Babu, M. M., van der Lee, R., de Groot, N. S., and Gsponer, J. (2011). Intrinsically disordered proteins: Regulation and disease. *Curr. Opin. Struct. Biol.* 21, 432–440. doi:10.1016/j.sbi.2011.03.011
- Baidakov, V. G., Chernykh, G. G., and Protchenko, S. P. (2000). Effect of the cut-off radius of the intermolecular potential on phase equilibrium and surface tension in Lennard-Jones systems. *Chem. Phys. Lett.* 321, 315–320. doi:10.1016/S0009-2614(00)00217-7
- Banani, S. F., Lee, H. O., Hyman, A. A., and Rosen, M. K. (2017). Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* 18, 285–298. doi:10.1038/nrm.2017.7
- Binder, K., Billotet, C., and Mirol, P. (1978). On the theory of spinodal decomposition in solid and liquid binary mixtures. *Z. Phys. B* 30, 183–195. doi:10.1007/bf01320985
- Binder, K., and Stauffer, D. (1976). Statistical theory of nucleation, condensation and coagulation. *Adv. Phys.* 25, 343–396. doi:10.1080/00018737600101402
- Boeynaems, S., Alberti, S., Fawzi, N. L., Mittag, T., Polymenidou, M., Rousseau, F., et al. (2018). Protein phase separation: A new phase in cell biology. *Trends Cell Biol.* 28, 420–435. doi:10.1016/j.tcb.2018.02.004
- Brangwynne, C. P., Mitchison, T. J., and Hyman, A. A. (2011). Active liquid-like behavior of nucleoli determines their size and shape in *Xenopus laevis* oocytes. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4334–4339. doi:10.1073/pnas.1017150108
- Brangwynne, C. P., Eckmann, C. R., Courson, D. S., Rybarska, A., Hoege, C., Gharakhani, J., et al. (2009). Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* 324, 1729–1732. doi:10.1126/science.1172046
- Cahn, J. W., and Hilliard, J. E. (1958). Free energy of a nonuniform system. I. Interfacial free energy. *J. Chem. Phys.* 28, 258–267. doi:10.1063/1.1744102
- Cahn, J. W., and Hilliard, J. E. (1971). Spinodal decomposition: A reprise. *Acta Met.* 19, 151–161. doi:10.1016/0001-6160(71)90127-1
- Caragine, C. M., and Haley, S. C. (2019). Nucleolar dynamics and interactions with nucleoplasm in living cells. *eLife* 8, e47533. doi:10.7554/eLife.47533
- Caragine, C. M., Haley, S. C., and Zidovska, A. (2018). Surface fluctuations and coalescence of nucleolar droplets in the human cell nucleus. *Phys. Rev. Lett.* 121, 148101. doi:10.1103/physrevlett.121.148101
- Chwastyk, M., and Cieplak, M. (2020). Conformational biases of α -synuclein and formation of transient knots. *J. Phys. Chem. B* 124, 11–19. doi:10.1021/acs.jpcc.9b08481
- Chwastyk, M., Galera-Prat, A., Sikora, M., Gómez-Sicilia, Á., Carrión-Vázquez, M., and Cieplak, M. (2014). Theoretical tests of the mechanical protection strategy in protein nanomechanics. *Proteins* 82, 717–726. doi:10.1002/prot.24436
- Chwastyk, M., Jaskolski, M., and Cieplak, M. (2014). Structure-based analysis of thermodynamic and mechanical properties of cavity-containing proteins – case study of plant pathogenesis-related proteins of class 10. *FEBS J.* 281, 416–429. doi:10.1111/febs.12611
- Chwastyk, M., Jaskolski, M., and Cieplak, M. (2016). The volume of cavities in proteins and virus capsids. *Proteins* 84, 1275–1286. doi:10.1002/prot.25076
- Chwastyk, M., Poma Bernal, A., and Cieplak, M. (2015). Statistical radii associated with amino acids to determine the contact map: Fixing the structure of a type I cohesin domain in the Clostridium thermocellum cellulosome. *Phys. Biol.* 12, 046002. doi:10.1088/1478-3975/12/4/046002
- Chwastyk, M., Vera, A. M., Galera-Prat, A., Gunnoo, M., Thompson, D., Carrión-Vázquez, M., et al. (2017). Non-local effects of point mutations on the stability of a protein module. *J. Chem. Phys.* 147, 105101. doi:10.1063/1.4999703
- de Aquino, B. R. H., Chwastyk, M., Mioduszecki, L., and Cieplak, M. (2020). Networks of interbasin traffic in intrinsically disordered proteins. *Phys. Rev. Res.* 2, 013242. doi:10.1103/physrevresearch.2.013242
- Dignon, G. L., Zheng, W., Best, R. B., Kim, Y. C., and Mittal, J. (2018). Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* 115, 9929–9934. doi:10.1073/pnas.1804177115
- Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B., and Mittal, J. (2018). Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* 14, e1005941. doi:10.1371/journal.pcbi.1005941
- Dignon, G. L., Zheng, W., Kim, Y. C., and Mittal, J. (2019). Temperature-controlled liquid-liquid phase separation of disordered proteins. *ACS Cent. Sci.* 5, 821–830. doi:10.1021/acscentsci.9b00102
- Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, V. L. (2008). Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* 18, 756–764. doi:10.1016/j.sbi.2008.10.002
- Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6, 197–208. doi:10.1038/nrm1589
- Elbaum-Garfinkle, S. (2019). Matter over mind: Liquid phase separation and neurodegeneration. *JBC Rev.* 294, 7160–7168. doi:10.1074/jbc.rev118.001188
- Feder, J., Russell, K. C., Lothe, J., and Pound, G. M. (1966). Homogeneous nucleation and growth of droplets in vapours. *Adv. Phys.* 15, 111–178. doi:10.1080/00018736600101264
- Ferreon, A. C. M., Moran, C. R., Gambin, Y., and Deniz, A. A. (2010). Single-molecule fluorescence studies of intrinsically disordered proteins. *Methods Enzymol.* 472, 179–204. doi:10.1016/S0076-6879(10)72010-3
- Fink, A. L. (2005). Natively unfolded proteins. *Curr. Opin. Struct. Biol.* 15, 35–41. doi:10.1016/j.sbi.2005.01.002
- Frenkel, J. (1955). *Kinetic theory of nucleation*. New York: Dover.
- Hensen, J. P., and McDonald, I. R. (1973). *Theory of simple liquids*. New York: Academic Press.
- Huang, J. S., Goldburg, W. I., and Bjerkaas, A. W. (1974). Study of phase separation in a critical binary liquid mixture: Spinodal decomposition. *Phys. Rev. Lett.* 32, 921–923. doi:10.1103/physrevlett.32.921
- Langer, J. S. (1971). Theory of spinodal decomposition in alloys. *Ann. Phys.* 65, 53–86. doi:10.1016/0003-4916(71)90162-x
- Li, B., Alonso, D. O. V., and Daggett, V. (2002). Stabilization of globular proteins via introduction of temperature-activated elastin-based switches. *Structure* 10, 989–998. doi:10.1016/S0969-2126(02)00792-x
- Mioduszecki, L., and Cieplak, M. (2018). Disordered peptide chains in an α -C-based coarse-grained model. *Phys. Chem. Chem. Phys.* 20, 19057–19070. doi:10.1039/c8cp03309a
- Mioduszecki, L., and Cieplak, M. (2020). Protein droplets in systems of disordered homeopeptides and the amyloid glass phase. *Phys. Chem. Chem. Phys.* 22, 15592–15599. doi:10.1039/d0cp01635g
- Mioduszecki, L., and Cieplak, M. (2021). Viscoelastic properties of wheat gluten in a molecular dynamics study. *PLoS Comput. Biol.* 17, e1008840. (in press). doi:10.1371/journal.pcbi.1008840
- Nicolas, J. J., Gubbins, K. E., Street, B., and Tildesley, D. J. (1979). Equation of state for the Lennard-Jones fluid. *Mol. Phys.* 37, 1429–1454. doi:10.1080/00268977900101051
- Panagiotopoulos, A. Z. (1994). Molecular simulation of phase coexistence: Finite-size effects and determination of critical parameters for two- and three-dimensional Lennard-Jones fluids. *Int. J. Thermophys.* 15, 1057–1072. doi:10.1007/bf01458815
- Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *J. Comp. Phys.* 117, 1–19. doi:10.1006/jcph.1995.1039
- Rowlinson, J. S., and Widom, B. (1982). *Molecular theory of capillarity*. Oxford: Clarendon Press.
- Schultz, A. J., and Kofke, D. A. (2018). Comprehensive high-precision high-accuracy equation of state and coexistence properties for classical Lennard-Jones crystals and low-temperature fluid phases. *J. Chem. Phys.* 149, 204508. doi:10.1063/1.5053714
- Shin, Y., and Brangwynne, C. P. (2017). Liquid phase condensation in cell physiology and disease. *Science* 357, eaaf4382. doi:10.1126/science.aaf4382
- Silmore, K. S., Howard, M. P., and Panagiotopoulos, A. Z. (2017). Vapour-liquid phase equilibrium and surface tension of fully flexible Lennard-Jones chains. *Molec. Phys.* 115, 320–327. doi:10.1080/00268976.2016.1262075
- Stephan, S., Thol, M., Vrabec, J., and Hasse, H. (2019). Thermophysical properties of the Lennard-Jones fluid: Database and data assessment. *J. Chem. Info. Model.* 59, 4248–4265. doi:10.1021/acs.jcim.9b00620
- Urry, D. W., Hugel, T., Seitz, M., Gaub, H. E., Scheiba, L., Dea, J., et al. (2002). Elastin: A representative ideal protein elastomer. *Phil. Trans. R. Soc. Lond. B* 357, 169–184. doi:10.1098/rstb.2001.1023
- Uversky, V. N., and Dunker, A. K. (2010). Understanding protein non-folding. *Biochem. Biophys. Acta* 1804, 1231–1264. doi:10.1016/j.bbapap.2010.01.017
- Uversky, V. N. (2002). Natively unfolded proteins: A point where biology waits for physics. *Prot. Sci.* 11, 739–756. doi:10.1110/ps.4210102
- Wright, P. E., and Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 6, 18–29. doi:10.1038/nrm3920
- Zhao, Y., Chwastyk, M., and Cieplak, M. (2017). Structural entanglements in protein complexes. *J. Chem. Phys.* 146, 225102. doi:10.1063/1.4985221
- Zhao, Y., Chwastyk, M., and Cieplak, M. (2017). Topological transformations in proteins: Effects of heating and proximity of an interface. *Sci. Rep.* 7, 39851. doi:10.1038/srep39851



OPEN ACCESS

EDITED BY

Adolfo Poma,
Institute of Fundamental Technological
Research, Polish Academy of Sciences,
Poland

REVIEWED BY

Wojciech Plazinski,
Jerzy Haber Institute of Catalysis and
Surface Chemistry, Polish Academy of
Sciences, Poland
Amit Das,
Indian Institute of Technology Delhi, India

*CORRESPONDENCE

Rafael C. Bernardi,
✉ rcbernardi@auburn.edu

SPECIALTY SECTION

This article was submitted to Theoretical
and Computational Chemistry,
a section of the journal
Frontiers in Chemistry

RECEIVED 24 November 2022

ACCEPTED 25 January 2023

PUBLISHED 08 February 2023

CITATION

Gomes PSFC, Forrester M, Pace M,
Gomes DEB and Bernardi RC (2023), May
the force be with you: The role of hyper-
mechanostability of the bone sialoprotein
binding protein during early stages of
Staphylococci infections.
Front. Chem. 11:1107427.
doi: 10.3389/fchem.2023.1107427

COPYRIGHT

© 2023 Gomes, Forrester, Pace, Gomes
and Bernardi. 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.

May the force be with you: The role of hyper-mechanostability of the bone sialoprotein binding protein during early stages of *Staphylococci* infections

Priscila S. F. C. Gomes, Meredith Forrester, Margaret Pace,
Diego E. B. Gomes and Rafael C. Bernardi*

Department of Physics, College of Sciences and Mathematics, Auburn University, Auburn, AL, United States

The bone sialoprotein-binding protein (Bbp) is a mechanoactive MSCRAMM protein expressed on the surface of *Staphylococcus aureus* that mediates adherence of the bacterium to fibrinogen- α (Fg α), a component of the bone and dentine extracellular matrix of the host cell. Mechanoactive proteins like Bbp have key roles in several physiological and pathological processes. Particularly, the Bbp: Fg α interaction is important in the formation of biofilms, an important virulence factor of pathogenic bacteria. Here, we investigated the mechanostability of the Bbp: Fg α complex using *in silico* single-molecule force spectroscopy (SMFS), in an approach that combines results from all-atom and coarse-grained steered molecular dynamics (SMD) simulations. Our results show that Bbp is the most mechanostable MSCRAMM investigated thus far, reaching rupture forces beyond the 2 nN range in typical experimental SMFS pulling rates. Our results show that high force-loads, which are common during initial stages of bacterial infection, stabilize the interconnection between the protein's amino acids, making the protein more "rigid". Our data offer new insights that are crucial on the development of novel anti-adhesion strategies.

KEYWORDS

mechanobiology, *Staphylococcus* infection, biofilm, adhesins, molecular dynamics

1 Introduction

Staphylococcus aureus infections have a high clinical and communal impact with an estimated mortality rate that can reach 30.2% [Bai et al. \(2022\)](#). The persistence of these infections lies on the *Staphylococcus aureus*' ability to form biofilms [Costerton et al. \(1999\)](#); [Archer et al. \(2011\)](#); [Suresh et al. \(2019\)](#), and the eventual dissemination of these pathogenic bacteria throughout the body [Kwieceński and Horswill \(2020\)](#). Despite the increase in sterilization and hygienic measures, modern medical devices play a key role in the transfer of these bacterial colonies through device-associated biofilm infections [Wertheim et al. \(2004\)](#); [Otto \(2009\)](#); [Lister and Horswill \(2014\)](#). The contamination of patients during medical and dental procedures is of increasing relevance, particularly with the emergence of drug-resistant bacteria. In the dental field, it has been estimated that the carrier prevalence of *S. aureus* in healthy adults varies from 24% to 84% [Donkor and Kotey \(2020\)](#). Additionally, the oral cavity is a source for cross infection and dissemination of the infection directly into the bloodstream, increasing the likelihood of septicemia and possibly death [McCormack et al. \(2015\)](#); [Garbacz et al. \(2021\)](#); [Jevon et al. \(2021\)](#).

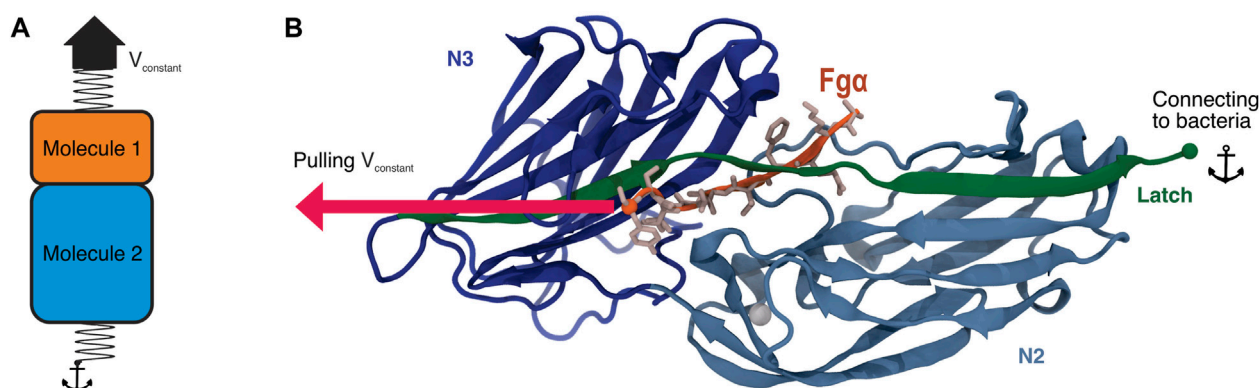


FIGURE 1

Bbp's adhesion domain. (A) Scheme illustrating the SMD protocol applying force at the interface between two molecules of interest. In this protocol, a spring is attached to one of the termini of each molecule, in our case, the C-terminal end of both Bbp and Fga peptide. While the end of Molecule two is fixed, the end of Molecule one is then pulled at constant velocity. (B) Tridimensional structure of Bbp. The protein is represented in cartoon, colored by its different domains. The latch is highlighted in green. Fga is colored in orange and its aminoacids represented as sticks colored in light pink. The SMD pulling and anchor points are indicated in the image as spheres.

Biofilms shelter the bacteria and enhance the persistence of infection by eluding innate and adaptive host defenses [González et al. \(2018\)](#); [Versey et al. \(2021\)](#). Biofilms also form a barrier, protecting colonies from biocides and antibiotic chemotherapies [Sharma et al. \(2019\)](#). Adhesins play critical roles during infection, especially during the early step of adhesion when bacterial cells are exposed to mechanical stress [Latasa et al. \(2006\)](#). Adhesins bind to their target ligands, holding it tight to them even at extreme force loadings that largely outperform classical binding forces [Gomes P. S. F. C. et al. \(2022\)](#). The resilience to mechanical forces provides the pathogen with a means to withstand high levels of mechanical stress during biofilm formation, thus yielding these pathogens highly resistant to breaking these cell adhesion bonds. These unusual stress-dependent molecular interactions play an integral role during bacterial colonization and dissemination and when studied, reveal critical information about pathosis [Dufrêne and Viljoen \(2020\)](#).

Among *S. aureus* adhesins, the bone sialoprotein binding protein (Bbp) is a bifunctional Microbial Surface Component Recognizing Adhesive Matrix Molecule (MSCRAMM) [Gillaspie et al. \(1998\)](#). Bbp is part of the MSCRAMM serine-aspartate repeat (Sdr) family that also includes SdrF and SdrG in *Staphylococcus epidermidis*, and clumping factor A (ClfA), B (ClfB), SdrC, and SdrE in *S. aureus* [Josefsson et al. \(1998\)](#); [McDevitt et al. \(1994\)](#); [Ni Eidhin et al. \(1998\)](#); [Tung et al. \(2000\)](#). Ligand-binding for Bbp occurs generally in the N-terminal region, from residues 273 to 598, where Bbp binds to fibrinogen- α (Fga), a glycopeptide on bone and dentine extracellular matrix (ECM). Bbp's binding region is subdivided into domains N2 and N3, which are made up of two layers of β -sheets with an open groove at the C-terminus where primary ligand binding occurs [Zhang et al. \(2015\)](#) (Figure 1B). The binding of Fga follows a “dock, lock, and latch” mechanism [O'Connell \(2003\)](#); [Ponnuraj et al. \(2003\)](#); [Bowden et al. \(2008\)](#); [Foster et al. \(2014\)](#); [Zhang et al. \(2017\)](#), that has been previously investigated by a myriad of techniques [Herman et al. \(2014\)](#); [Vanzielegheem et al. \(2015\)](#); [Vitry et al. \(2017\)](#); [Herman-Bausier et al. \(2018\)](#); [Milles et al. \(2018\)](#). Thus, the pathogenic bacteria does not invade a host cell, but rather adheres to the ECM via Bbp: Fga interactions [Patti et al. \(1994\)](#).

Using a combination of *in silico* and *in vitro* single-molecule force spectroscopy (SMFS), we have previously reported that *S. epidermidis*'

adhesin SdrG, when in complex with Fg β , was able to withstand extreme mechanical loads [Milles et al. \(2018\)](#). The necessary force applied to rupture the SdrG: Fg β complex was shown to be an order of magnitude stronger than that needed to rupture the widely employed Streptavidin:biotin complex [Sedlak et al. \(2018\)](#), and more than twice of that of cellulosomal cohesin:dockerin interactions [Schoeler et al. \(2014\)](#); [Bernardi et al. \(2019\)](#). Most biological complexes rupture at a relatively low force range [Seppälä et al. \(2017\)](#); [Haataja et al. \(2019\)](#); [Hoelz et al. \(2011; 2012\)](#); [Mendes et al. \(2012\)](#); [Bernardi and Pascutti \(2012\)](#), including other host-pathogen interactions [Bauer et al. \(2022\)](#). A molecular mechanism for a catch-bond behavior of the SdrG: Fg β was then revealed by investigating the system in a “force-clamp” regime [Melo et al. \(2022\)](#), with magnetic tweezers based SMFS revealing that the SdrG: Fg β bond can live for hours under force loads [Huang et al. \(2022\)](#). Here, taking advantage of a powerful *in silico* SMFS approach, we describe how Bbp plays a key role in bacterial adhesion during nosocomial infections, by investigating the Bbp: Fga complex at different pulling velocities combining all-atom (aa) and coarse-grained (CG) steered molecular dynamics (SMD) simulations (Figures 1A, B). Building on *in vitro* SMFS data, our results point to Bbp's interaction with the extracellular matrix fibrinopeptide as the most mechanostable so far investigated, independent of the loading rate. Our findings reveal that a few key interactions are responsible for the outstanding force resilience of the complex. Furthermore, our results offer insights into the development of anti-adhesion strategies.

2 Results

2.1 Bbp is highly mechanostable under stress

To probe the mechanics of the interaction between Bbp and Fga, and to characterize the atomic details of the complex under force load, we performed aa-SMD simulations with Bbp anchored by its C-terminal while Fga was pulled at different velocities (Supplementary Table S1). The simulations resulted in Force vs extension curves that reveal a clear one-step rupture event, as

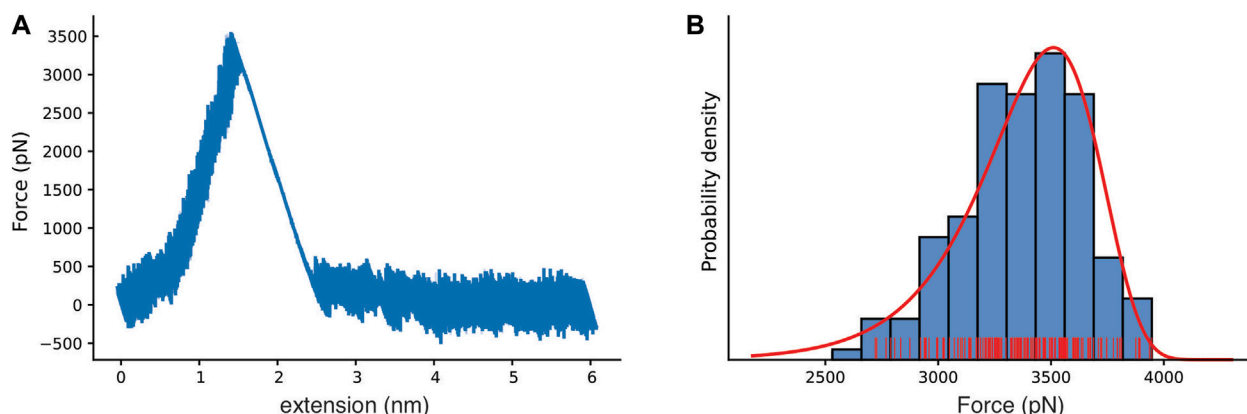


FIGURE 2

Bbp mechanostability under high mechanical load. (A) Force versus extension curve as an exemplary trace, with rupture peak force at 3,510 pN. (B) Histogram for the most probable rupture force (blue, rugged plot in red) with the Bell-Evans (BE) model for the first rupture peak (red), based on the all-atom steered molecular dynamics simulation replicas with the slowest simulated pulling velocity (2.5×10^{-4} nm/ps).

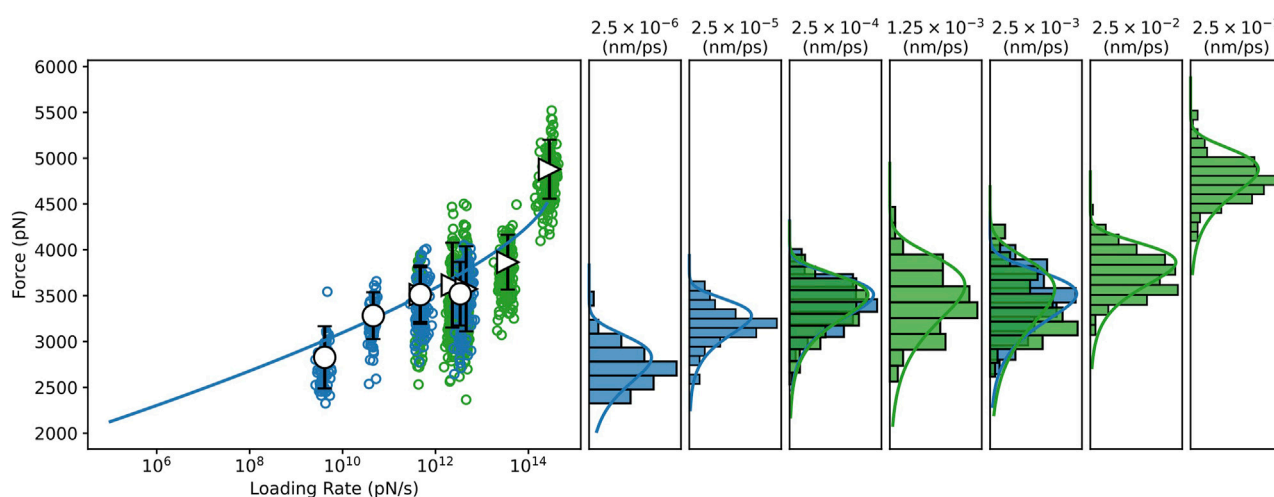


FIGURE 3

Dynamic Force spectrum for the Bbp: Fg α complex combining data from all-atom and coarse-grained SMD simulations. All-atom, and Coarse-grained steered molecular dynamics simulations (CG-SMD and aa-SMD) were performed at different velocities: 2.5×10^{-6} to 2.5×10^{-3} nm/ps (blue) and 2.5×10^{-4} to 2.5×10^{-1} nm/ps (green), respectively. A Dudko-Hummer-Szabo [Dudko et al. \(2006\)](#) (DHS) fit was performed through the SMD dataset predicting $\Delta x = 7.489 \times 10^{-2}$ nm, $k_{off}^0 = 2.596 \times 10^{-12}$ s $^{-1}$, $\Delta G = 2.293 \times 10^2$ $k_B T$.

represented in [Figure 2A](#). For the slowest pulling velocity, 160 replicas were performed following a wide-sampling paradigm previously developed in our group [Sedlak et al. \(2020\)](#). At the pulling velocity of 2.5×10^{-4} nm/ps, we observed that the most probable rupture force for the complex was 3,510 pN, as described by the Bell-Evans (BE) [Bell \(1978\)](#); [Evans and Ritchie \(1997\)](#) fit of the peak forces at that pulling speed (see [Figure 2B](#)). Our results reveal that Bbp: Fg α is the most mechanostable complex investigated thus far, which is in agreement with previous experimental data where we showed that SdrG: Fg β complex can withstand forces on the 2 nN range, equivalent to breaking of covalent bonds [Milles et al. \(2018\)](#).

To investigate the dependence of the mechanostability of Bbp: Fg α on the force loading rate, we performed CG-SMD simulations at

several, much lower, pulling speeds ([Supplementary Table S1](#)). We have recently shown that aa-SMD and CG-SMD can be combined to in an *in silico* SMFS approach [Gomes D. E. et al. \(2022\)](#); [Melo et al. \(2022\)](#). Here, the combination of the two levels of molecular details is capable of rendering predictions that are consistent with theory and experimentation with the advantage of being 10 to approximately 100 times faster than aa-SMD simulations, depending on the pulling speed [Gomes D. E. et al. \(2022\)](#); [Melo et al. \(2022\)](#). A Dudko-Hummer-Szabo [Dudko et al. \(2006\)](#) (DHS) fit was performed through the SMD data, including both the aa-SMD, and the CG-SMD (see [Figure 3](#)). The DHS fit suggests that the system should rupture at forces higher than 2 nN at 10^5 pN/s force loading rate, in agreement with experimental data [Milles et al. \(2018\)](#). It is interesting

TABLE 1 Hydrogen bonds occupancy between Bbp and Fg α residues calculated and averaged before the main rupture event.

Bbp	Fg α	Occupancy (%)	Nature
Asp334	Thr565	54.84	Side-chain
Asp556	Lys562	45.39	Salt-bridge
Leu584	Thr565	36.14	Backbone
Thr582	Ser567	35.62	Backbone
Thr586	Gln563	35.03	Backbone:Side-chain
Ser333	Phe564	18.54	Side-chain
Asp588	Gln563	12.77	Side-chain
Thr587	Ser561	12.66	Side-chain

to note that the BE model is able to fit well all the simulation results, at both aa and CG level, as evidenced by the density plots in [Figure 3](#).

The influence of the peptide size on the rupture force was also investigated. We have shown previously that SdrG complexed with shortened Fg β peptides had lower unbinding forces [Milles et al. \(2018\)](#). Here, we simulated a model of Bbp complexed with Fg α elongated by nine residues (See Methods section) by aa and CG-SMD simulations ([Supplementary Table S1](#)). Our results show that the force loading rate was not significantly impacted by the size of the peptide ([Supplementary Figure S1](#)), indicating that the original complex formed at the crystal structure has the minimal length to keep the important contacts with the protein latch to hold the DLL configuration.

2.2 Key hydrogen bonds are responsible for Bbp: Fg α high mechanostability

After confirming that Bbp: Fg α complex presents a hyperstable interaction under shear mechanical load, we used the approximately 3 μ s of aa-SMD simulation data to investigate the molecular origin of the mechanostability of the complex. Previously, simulations of the SdrG: Fg β revealed the presence of frequent and persistent hydrogen bonds (H-bonds) between the peptide and the protein backbone, showing that the high-force resilience of the complex was largely independent of the peptide side-chains interactions, and therefore the peptide's sequence [Milles et al. \(2018\)](#). Here, we computed the occupancy of the H-bonds between the Bbp and Fg α before the complex rupture. We identified the key amino acid interactions responsible for keeping the complex together at high force loads ([Table 1](#)). Different than SdrG: Fg β , Bbp: Fg α interactions are not dominated by backbone-backbone interactions, with a significant amount of side-chain interaction of the peptide playing an important role in the complex mechanostability. The backbone interactions between Bbp^{Leu584, Thr582, Thr586} and Fg α ^{Thr565, Ser567, Thr586} have been previously described as important for Fg α binding at the crystal structure [Zhang et al. \(2015\)](#). However, we noticed that the side-chain H-bonds are rearranged upon application of mechanical stress on the complex. On the crystal, Bbp^{Asp334} forms a side-chain H-bond with Fg α ^{Ser566}, and during the SMD simulations, this interaction shifts to Fg α ^{Thr565}, being the H-bond with the highest occupancy over the trajectories. Another shift occurs between

Bbp^{Asp334, Ile335} interacting with Fg α ^{Phe564}, on the crystal, to Bbp^{Ser333} interacting with Fg α ^{Phe564} in our simulations. The H-bond between Bbp^{Asp588} and Fg α ^{Gln563} is described as important to lock the peptide N-terminus and is still present before the rupture of the complex, although with lower occupancy. Instead, a charged side-chain interaction arises with significant occupancy values: Bbp^{Asp556}: Fg α ^{Lys562}. These data corroborates the importance of backbone interactions to maintain the high mechanostability and also highlights important side chain H-bonds plasticity that occurs when Bbp: Fg α is exposed to mechanical stress.

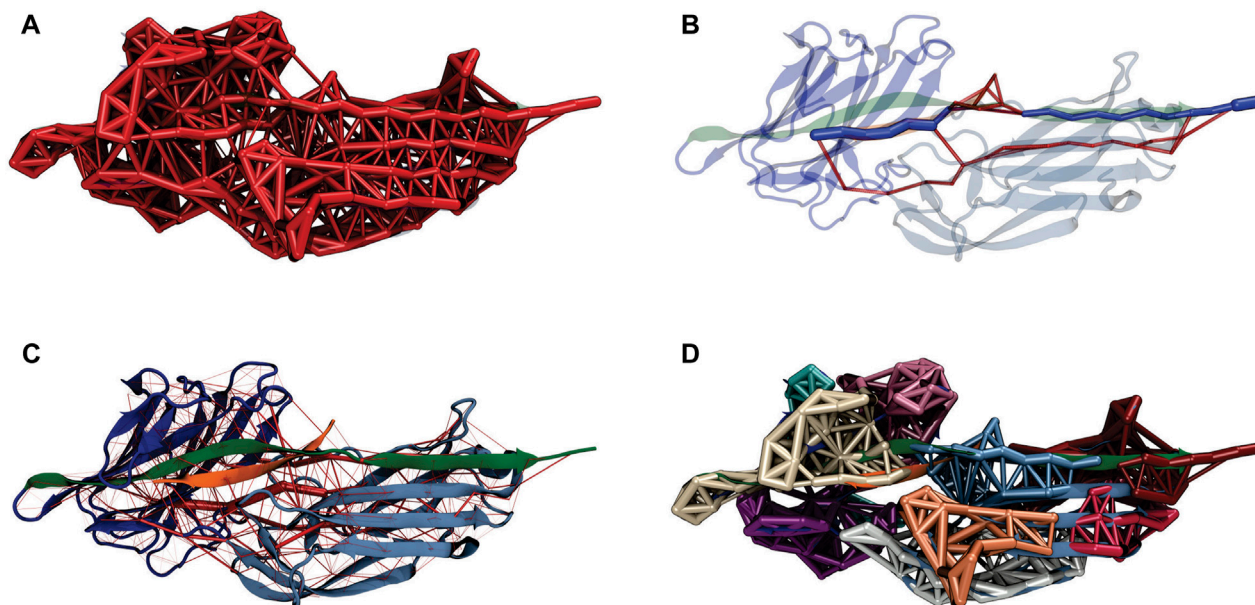
2.3 The force propagates indirectly from the latch to the peptide

How a shear force load “activates” the hyperstability of the complex can be investigated by analysing the evolution of pairwise interactions during the force-loading event. Such analysis can be used to investigate how a catch-bond may be formed in the Bbp: Fg α complex [Liu et al. \(2020\)](#). Previously, it has been shown that SdrG: Fg β presents a catch-bond behavior [Huang et al. \(2022\)](#), which is expected also for Bbp: Fg α . To analyse the pairwise interactions during the SMD, we employed the generalized correlation-based dynamical network analysis method [Melo et al. \(2020\)](#), which can also be used to calculate force propagation pathways [Schoeler et al. \(2015\)](#). [Figure 4A](#) shows the pairwise interactions obtained from the network analysis. The thickness of the connections between nodes (amino acid residues) represents how well correlated the motion of these nodes are, and therefore how well connected are these amino acid residues.

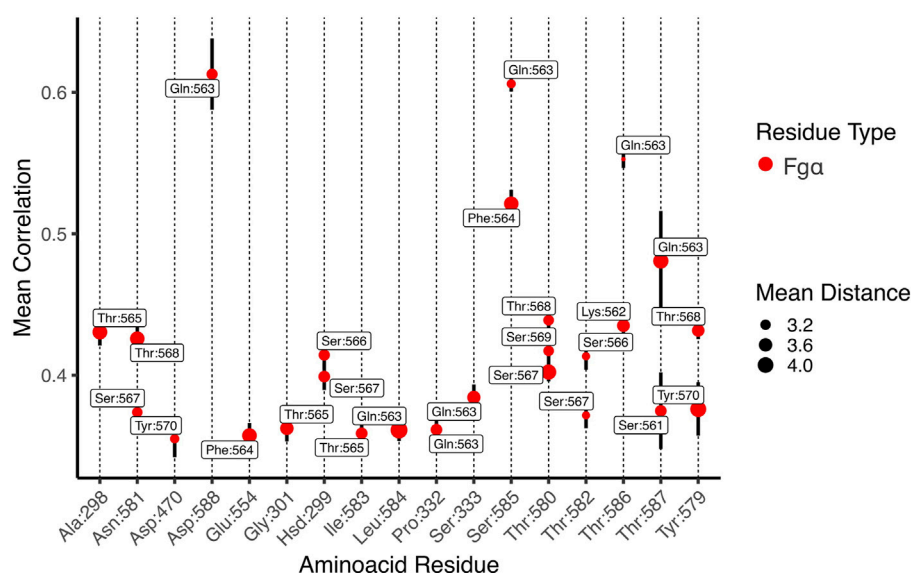
The force propagation pathway that connects the pulling and the anchoring residues shows that most of the force is propagating from the protein latch directly to the peptide, passing by the center of Bbp's N2 domain ([Figure 4B](#)). These results are slightly different than the ones obtained for the SdrG: Fg β complex upon high mechanical stress [Milles et al. \(2018\)](#). However, in a previous study, it was observed that changes in the pulling velocities can lead to different force propagation pathways, suggesting different unbinding mechanisms at different pulling rates [Melo et al. \(2022\)](#).

The rigidity of the protein under high-force load can also be studied using the betweenness map from the dynamical network analysis (see [Figure 4C](#)). The betweenness is defined as the number of shortest paths from all vertices to all others that pass through that node, in this case, an amino acid residue. If an amino acid residue has high betweenness, it tends to be important for controlling inter-domain communication within a protein [Melo et al. \(2020\)](#). High betweenness values (thicker red tubes) are seen on the latch that is in direct contact with Fg α , highlighting the strong correlation between the motif and the peptide. Interestingly, high betweenness is also found at connections intra N2 domain, pointing that Bbp: Fg α complex becomes more rigid under high force loads, particularly in the region interconnecting the latch, the peptide and the N2 domains. Such behavior helps the stabilization of the interactions under high forces.

A representation of the network in subgroups, or communities, is shown at [Figure 4D](#). The communities group the amino acid residues that are most inter-connected in relation to the rest of the network. We can see that Bbp: Fg α is subdivided in a handful of communities. The latch, most of Fg α , and part of the N2 domains are united in the same community in light blue, showing that these amino acids are highly

**FIGURE 4**

Bbp:Fga dynamical network under high mechanical load. **(A)** Representation of the dynamical network. The thickness of the links between the nodes (amino acid residues) represents the correlation of motion between these residues. **(B)** The force propagates from the latch indirectly to the peptide, passing by the N2 domain of the protein. The color scheme of the complex is the same from Figure 1. The network's optimal path is colored in dark blue while the sub-optimal paths are colored in red. **(C)** Full dynamical network revealing the most correlated regions of the complex. The weight of the network edges (represented by the thickness of red tubes) is given by the betweenness values. **(D)** Generalized correlation-based communities represented by different colors of the nodes and edges in the network.

**FIGURE 5**

Mean generalized coefficients for contacts along Bbp:Fga interface. The x-axis is labeled by Bbp amino acid residues and the y-axis indicates the averaged generalized correlation values (vertical bars indicate the standard error of the mean), labeled by Fga amino acid residues. The circle sizes indicates the average Cartesian distance. Only amino acid residues with a mean correlation higher than 0.35 are shown.

connected. We also measured the correlation between motions on the interface residues to determine how cooperative their motion is and the essential contacts that are keeping the complex stable under high mechanical load. Essentially, the higher the correlation between

residues, the more relevant is their interaction for the stability of the protein complex. We noticed that two Fga residues are highly correlated (values equal or superior to 0.5) to Bbp at the interface, namely: Fga^{Gln:563}:Bbp^{Asp:588,Ser:585,Thr:586,Thr:587} and Fga^{Phe:564}:Bbp^{Ser:585}

(Figure 5). The importance of Fgα^{Gln563} described as a persistent H-bond contact with Bbp^{Asp588,Thr586} and important locking contact is reinforced by its high correlation values. The same analysis was performed for the simulations of Bbp complexed with the elongated Fgα peptide (Supplementary Figure S2). The pattern of contacts is very similar, reinforcing the importance of Fgα^{Gln563}, and we observe the absence of new contacts made by the extra residues, corroborating that the short peptide contains the key residues responsible for holding the complex tight at the DLL configuration.

3 Discussion

During infection, Gram-positive bacteria are frequently exposed to high mechanical stress. These bacteria have evolved an intricate host-binding mechanism to efficiently form colonies under the most inhospitable conditions. Key for the maintenance of the colonies, biofilms are an important virulence factor developed by *S. aureus* among other bacteria. In the initial steps of infection and biofilm formation, MSCRAMMS adhesins have an important role in clinging the bacteria to their human hosts Otto (2009); Latasa et al. (2006). *Staphylococcus aureus* isolated from patients suffering from septic arthritis and osteomyelitis specifically interacts with bone sialoprotein, present at bone and dentine extracellular matrix. This interaction is mediated by an specific adhesin protein, namely Bbp Ryden et al. (1987); Ganss et al. (1999); Tung et al. (2000).

Here we have explored the interaction of Bbp with Fgα by using an *in silico* SMFS approach that relies on aa- and CG-SMD simulations. CG-SMD simulations have proven to bridge the force-loading gap between *in vitro* SMFS data with *in silico* data obtained from aa-SMD simulations, distanced by orders of magnitude Gomes D. E. et al. (2022). In addition, CG-SMD simulations require much less computational power Liu et al. (2021); Poma et al. (2019), enabling us to explore pulling speeds unfeasible to simulate *via* aa-SMD Gomes D. E. et al. (2022). Using an approach previously described Souza et al. (2019), we combined GōMartini approach Poma et al. (2017) with Martini 3 Souza et al. (2021) obtaining sensible results. The higher spread of rupture force at faster pulling rates suggests that force-induced extensions may result in lost of relevant interactions between CG-bead pairs, indicating that further optimization of the contact map or redefinition of the native contacts is necessary to improve the results Mahmood et al. (2021).

Here, we showed that Bbp: Fgα complex can withstand forces even higher than the previously investigated SdrG: Fgβ complex Milles et al. (2018), overcoming the 2 nN force range for rupture forces, equivalent to breaking covalent bonds, demonstrating the high mechanostability of the Bbp: Fgα complex. We revealed that the force propagation pathway between the anchoring and pulling points of the Bbp: Fgα complex goes beyond the interactions between the latch and the peptide, passing through an intricate network involving several amino acids of the Bbp N2 domain (Figure 4). We were also able to point the key residues H-bonds responsible for keeping the complex stable at such high mechanical stress, highlighting important backbone-backbone interactions between Bbp^{Leu584, Thr582, Thr586} and Fgα^{Thr565, Ser567,Thr586} but also side-chain connections, such as Bbp^{Asp334}:Fgα^{Thr565}, Bbp^{Ser333}:Fgα^{Phe564} and Bbp^{Asp588}:Fgα^{Gln563} (Table 1). The latter being an important contact to lock the peptide N-terminus Zhang et al. (2015). Fgα^{Gln563} has also revealed to be a key network hub, being highly correlated with several residues on the complex interface

such as Bbp^{Asp588,Ser585,Thr586,Thr587} (Figure 5). We also showed that the short Fgα peptide is able to hold the key interactions responsible for its mechanostability by probing an elongated Fgα in complex with Bbp (Supplementary Figures S1 and S2).

By probing the Bbp: Fgα complex under high mechanical load, we discovered the molecular mechanism that triggers Bbp's unique resilience to shear forces. The high force-loads that can be found during initial stages of bacterial infection stabilize the interconnection between the protein's amino acids, particularly along the β-sheets that, due to their force-loading geometry, cannot be "peeled" like other β-sheet-rich proteins, such as green fluorescent protein (GFP) Hughes and Dougan (2016); Dietz et al. (2006) and human filamins Seppälä et al. (2017); Haataja et al. (2019). Our results build on previous knowledge of host-microbial interactions, supporting the idea that anti-adhesion therapies might be fundamental in our fight against nosocomial bacteria infections.

Antiadhesion therapies are attractive since they would not target essential processes and have the potential advantage of eliciting less and slower resistance acquisition. Some of the approaches using peptides have been reviewed elsewhere Dufrêne and Viljoen (2020). Our findings support that a short peptide is capable of holding the essential interactions to keep the protein locked in the DLL configuration. This could be explored on the design of small peptidomimetic compounds that can mimic these interactions. Moreover, peptidomimetics overcome the poor pharmacokinetic profile and low selectivity associated with peptide therapies, the main drawback for this kind of approach Li Petri et al. (2022). Another possible strategy would be to replace the peptide backbone for a small drug-like molecule with substituents that could mimic the bioactive conformation of the native peptide Spiegel et al. (2012).

Due to the good agreement between our *in silico* SMFS protocol and experiments, we could use our simulations as a platform to study structure-activity relationships and not only screen the early potential drug candidates, but also decipher their mechanisms of action. The best candidates can be later probed by SMFS experiments. In summary, our work presents a key step in creating a intelligent design for a new class of antibiotics that act on the initial stages of bacterial infection.

4 Methods

4.1 Structure preparation

The structure of Bbp in complex with Fgα has been previously solved by means of X-ray crystallography at 1.45 Å resolution Zhang et al. (2015) and deposited at the Protein Data Bank (PDB ID: 5CFA). Here we retrieved this structure and prepared it for molecular dynamics (MD) simulations using VMD Humphrey et al. (1996) and its plugin QwikMD Ribeiro et al. (2016). To investigate the loading rate dependency on the size of the peptide, we used Modeller v.10.1 Webb and Sali (2016) to create an additional structure of the complex where the Fgα was elongated by nine residues at its C-terminal end, in respect of the crystal structure, following the sequence of Fgα from *Homo sapiens* (Uniprot ID: P02671). The model followed the same preparation as described for the crystal structure.

4.2 All-atom molecular dynamics simulations

The complexes between Bbp and Fgα in its short or longer configuration were solvated using the TIP3P water model

Jorgensen et al. (1983), with the net charge of the protein neutralized using a 150 mM concentration of sodium chloride. Steered molecular dynamics (SMD) simulations were carried out using NAMD 3 Phillips et al. (2020), with the CHARMM36 force field Best et al. (2012). The simulations were performed assuming periodic boundary conditions in the isothermal-isobaric ensemble (NPT) with temperature maintained at 300 K using Langevin dynamics for temperature and pressure coupling, the latter kept at 1 bar. A distance cut-off of 11.0 Å was applied to short-range non-bonded interactions, whereas long-range electrostatic interactions were treated using the particle-mesh Ewald (PME) Darden et al. (1993) method. Taking advantage of a hydrogen-mass repartitioning method implemented in VMD's autopsfgen, the time step of integration was chosen to be 4 fs for all production aa-MD simulations performed. Before the SMD simulations, the system was submitted to an energy minimization protocol for 1,000 steps. An MD simulation with position restraints in the protein backbone atoms was performed for 1 ns, with temperature ramping from 0 K to 300 K in the first 0.5 ns at a timestep of 2.0 fs in the NVT ensemble, which served to pre-equilibrate the system. In an *in silico* single-molecule force spectroscopy (SMFS) strategy Verdorfer et al. (2017); Bernardi et al. (2019), SMD simulations were carried out in several replicas, using a constant velocity stretching protocol at three different pulling speeds (Supplementary Table S1). SMD was employed by harmonically restraining the position of the amino acid at the C-ter of Bbp and moving a second restraint point at the C-ter of Fgα peptide with a 5 kcal/mol Å² spring constant, with constant velocity in the z-axis. The force applied to the harmonic spring is then monitored during the time of the SMD. The pulling point was moved with constant velocity along the z-axis and due to the single anchoring point and the single pulling point the system is quickly aligned along the z-axis. The number of replicas for each velocity is indicated at Supplementary Table S1.

4.3 Coarse-grained molecular dynamics simulations

The atomistic model of Bbp: Fgα was modeled onto the Martini 3.0 Coarse-grained (CG) force field (v.3.0.b.3.2) Souza et al. (2021) using martinize2 v0.7.3 Kroon (2020). A set of native contacts, based on the rCSU + OV contact map protocol, was computed from the equilibrated all-atom structure using the rCSU server Wolek et al. (2015) and used to determine Gō-MARTINI interactions Poma et al. (2017) used to restrain the secondary and tertiary structures with the effective depth ϵ of Lennard-Jones potential set to 9.414 kJ.mol⁻¹. All CG-MD simulations were performed using GROMACS version 2021.5 Abraham et al. (2015). The Bbp: Fgα complex was centered in a rectangular box measuring with 10.0, 10.0, 25.0 nm to the x, y, and z directions. The anchor (Bbp C-terminal) and pulling (peptide C-terminal) backbone (BB) atoms were used to align the protein to the Z-axis. The box was then solvated with Martini3 water molecules. Systems were minimized for 10,000 steps with steepest descent, followed by a 10 ns equilibration at the NPT ensemble using the Berendsen thermostat at 298K, while pressure was kept at 1 bar with compressibility set to 3e⁻⁴bar⁻¹, using the Berendsen barostat. A time step of 10 fs was used to integrate the equations of motion. Pulling simulations were subsequently done at the NVT ensemble with a time step of 20 fs the temperature was controlled using the v-rescale thermostat Bussi et al. (2007) with a coupling time of 1 ps for all

CG-MD simulations, the cutoff distance for Coulombic and Lennard-Jones interactions was set to 1.1 nm De Jong et al. (2016), with the long-range Coulomb interactions treated by a reaction field (RF) Tironi et al. (1995) with $\epsilon_r = 15$. The Verlet neighbor search Verlet (1967) was used in combination with the neighbor list, updated every 20 steps. The LINCS Hess et al. (1997) algorithm was used to constrain the bonds and the leapfrog integration algorithm for the solution of the equations of motion. Several replicas of CG-SMD simulations were performed at a range of speeds described at Supplementary Table S1.

4.4 Simulation data analysis

All analysis presented at the main text correspond to the Bbp: Fgα original complex. Force loading rate and mean correlation values for Bbp complexed with the elongated Fgα peptide are found on the Supplementary Information material. H-bonds occupancy between Bbp and Fgα were calculated and averaged for aa-MD simulations 1 ns before the main rupture event, using VMD Humphrey et al. (1996) with standard parameters for the calculation: residue pairs; donor-acceptor distance of 3.0 Å; angle cutoff of 20°. Mean correlation and dynamical network pathways were calculated using the generalized dynamical network analysis Melo et al. (2020) and VMD for aa-SMD at pulling velocity of 2.5 × 10⁻⁴ nm/ps. In this analysis, a network is defined as a set of nodes that represent amino acid residues, and the node's position is mapped to that of the residue's α -carbon. Edges connect pairs of nodes if their corresponding residues are in contact, and two non-consecutive residues are said to be in contact if they are within 4.5 Å of each other for at least 75% of analyzed frames. The interface residues between Bbp: Fgα were defined in a radius of 10 Å between nodes in each molecule. A representative for the full-network, optimal and suboptimal paths and communities was rendered using one of the SMD trajectory replicas. The mean correlation analysis was carried out 1 ns before the first rupture event using a cutoff of 0.35 for the mean correlation coefficients. All charts were generated using in-house python scripts. The protein image was rendered using VMD.

Data availability statement

The raw data supporting the conclusion of this article will be made available upon request to the corresponding author.

Author contributions

PG, MF and MP contributed to performing simulations, analysing data, and writing of the manuscript. DG contributed to analysing data and discussion. RB contributed to writing and discussion on *in silico* force spectroscopy, proof-reading, manuscript revision and approval of the submitted version. PG and RB coordinated the project.

Funding

This work was supported by the National Science Foundation under Grant MCB-2143787 (CAREER: *In Silico* Single-Molecule

Force Spectroscopy) and ACCESS Allocations (Project: BIO220009).

Acknowledgments

We thank Auburn University and the College of Sciences and Mathematics for the computational resources provided by Dr. Bernardi faculty startup funds. We thank Dr. Marcelo Melo for the fruitful discussions.

Conflict of interest

The 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.

References

- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2, 19–25. doi:10.1016/j.softx.2015.06.001
- Archer, N. K., Mazaitis, M. J., Costerton, J. W., Leid, J. G., Powers, M. E., and Shirtliff, M. E. (2011). *Staphylococcus aureus* biofilms: Properties, regulation, and roles in human disease. *Virulence* 2, 445–459. doi:10.4161/viru.2.5.17724
- Bai, A. D., Lo, C. K., Komorowski, A. S., Suresh, M., Guo, K., Garg, A., et al. (2022). *Staphylococcus aureus* bacteraemia mortality: A systematic review and meta-analysis. *Clin. Microbiol. Infect.* 28, 1076–1084. doi:10.1016/j.cmi.2022.03.015
- Bauer, M. S., Gruber, S., Hausch, A., Gomes, P. S., Milles, L. F., Nicolaus, T., et al. (2022). A tethered ligand assay to probe SARS-CoV-2:ACE2 interactions. *Proc. Natl. Acad. Sci.* 119, e2114397119. doi:10.1073/pnas.2114397119
- Bell, G. I. (1978). Models for the specific adhesion of cells to cells. *Science* 200, 618–627. doi:10.1126/science.347575
- Bernardi, R. C., Durner, E., Schoeler, C., Malinowska, K. H., Carvalho, B. G., Bayer, E. A., et al. (2019). Mechanisms of nanonewton mechanostability in a protein complex revealed by molecular dynamics simulations and single-molecule force spectroscopy. *J. Am. Chem. Soc.* 141, 14752–14763. doi:10.1021/jacs.9b06776
- Bernardi, R. C., and Pascutti, P. G. (2012). Hybrid qm/mm molecular dynamics study of benzocaine in a membrane environment: How does a quantum mechanical treatment of both anesthetic and lipids affect their interaction. *J. Chem. theory Comput.* 8, 2197–2203. doi:10.1021/ct300213u
- Best, R. B., Zhu, X., Shim, J., Lopes, P. E., Mittal, J., Feig, M., et al. (2012). Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. theory Comput.* 8, 3257–3273. doi:10.1021/ct300400x
- Bowden, M. G., Heuck, A. P., Ponnuraj, K., Kolosova, E., Choe, D., Gurusiddappa, S., et al. (2008). Evidence for the “dock, lock, and latch” ligand binding mechanism of the staphylococcal microbial surface component recognizing adhesive matrix molecules (mscramm) sdr. *J. Biol. Chem.* 283, 638–647. doi:10.1074/jbc.m706252200
- Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126, 014101. doi:10.1063/1.2408420
- Costerton, J. W., Stewart, P. S., and Greenberg, E. P. (1999). Bacterial biofilms: A common cause of persistent infections. *Science* 284, 1318–1322. doi:10.1126/science.284.5418.1318
- Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald: An $n \log(n)$ method for ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397
- De Jong, D. H., Baoukina, S., Ingólfsson, H. I., and Marrink, S. J. (2016). Martini straight: Boosting performance using a shorter cutoff and gpus. *Comput. Phys. Commun.* 199, 1–7. doi:10.1016/j.cpc.2015.09.014
- Dietz, H., Berkemeier, F., Bertz, M., and Rief, M. (2006). Anisotropic deformation response of single protein molecules. *Proc. Natl. Acad. Sci.* 103, 12724–12728. doi:10.1073/pnas.0602995103
- Donkor, E. S., and Kotey, F. C. (2020). Methicillin-resistant staphylococcus aureus in the oral cavity: Implications for antibiotic prophylaxis and surveillance. *Infect. Dis. Res. Treat.* 13, 117863372097658. doi:10.1177/1178633720976581
- Dudko, O. K., Hummer, G., and Szabo, A. (2006). Intrinsic rates and activation free energies from single-molecule pulling experiments. *Phys. Rev. Lett.* 96, 108101. doi:10.1103/PhysRevLett.96.108101
- Dufrène, Y. F., and Viljoen, A. (2020). Binding strength of gram-positive bacterial adhesins. *Front. Microbiol.* 11, 1457. doi:10.3389/fmicb.2020.01457
- Evans, E., and Ritchie, K. (1997). Dynamic strength of molecular adhesion bonds. *Biophysical J.* 72, 1541–1555. doi:10.1016/s0006-3495(97)78802-7
- Foster, T. J., Geoghegan, J. A., Ganesh, V. K., and Höök, M. (2014). Adhesion, invasion and evasion: The many functions of the surface proteins of staphylococcus aureus. *Nat. Rev. Microbiol.* 12, 49–62. doi:10.1038/nrmicro3161
- Ganss, B., Kim, R. H., and Sodek, J. (1999). Bone sialoprotein. *Crit. Rev. Oral Biol. Med.* 10, 79–98. doi:10.1177/10454411990100010401
- Garbacz, K., Kwapisz, E., Piechowicz, L., and Wierzbowska, M. (2021). *Staphylococcus aureus* isolated from the oral cavity: Phage susceptibility in relation to antibiotic resistance. *Antibiotics* 10, 1329. doi:10.3390/antibiotics10111329
- Gillaspay, A. F., Lee, C. Y., Sau, S., Cheung, A. L., and Smeltzer, M. S. (1998). Factors affecting the collagen binding capacity of staphylococcus aureus. *Infect. Immun.* 66, 3170–3178. doi:10.1128/iai.66.7.3170-3178.1998
- Gomes, D. E., Melo, M. C., Gomes, P. S., and Bernardi, R. C. (2022a). Bridging the gap between *in vitro* and *in silico* single-molecule force spectroscopy. *bioRxiv*. doi:10.1101/2022.07.14.500151
- Gomes, P. S. F. C., Gomes, D. E. B., and Bernardi, R. C. (2022b). Protein structure prediction in the era of ai: Challenges and limitations when applying to *in silico* force spectroscopy. *Front. Bioinforma.* 2, 983306. doi:10.3389/fbinf.2022.983306
- González, J. F., Hahn, M. M., and Gunn, J. S. (2018). Chronic biofilm-based infections: Skewing of the immune response. *Pathogens Dis.* 76, fty023. doi:10.1093/femspd/fty023
- Haataja, T. J., Bernardi, R. C., Lecoate, S., Capoulade, R., Merot, J., and Pentikäinen, U. (2019). Non-syndromic mitral valve dysplasia mutation changes the force resilience and interaction of human filamin a. *Structure* 27, 102–112.e4. doi:10.1016/j.str.2018.09.007
- Herman, P., El-Kirat-Chatel, S., Beaussart, A., Geoghegan, J. A., Foster, T. J., and Dufrène, Y. F. (2014). The binding force of the staphylococcal adhesin sdr is remarkably strong. *Mol. Microbiol.* 93, 356–368. doi:10.1111/mmi.12663
- Herman-Bausier, P., Labate, C., Towell, A. M., Derclaye, S., Geoghegan, J. A., and Dufrène, Y. F. (2018). *Staphylococcus aureus* clumping factor a is a force-sensitive molecular switch that activates bacterial adhesion. *Proc. Natl. Acad. Sci.* 115, 5564–5569. doi:10.1073/pnas.1718104115
- Hess, B., Bekker, H., Berendsen, H. J., and Fraaije, J. G. (1997). Lincs: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463–1472. doi:10.1002/(sici)1096-987x(199709)18:12<1463:aid-jcc4>3.0.co;2-h
- Hoelz, L. V., Bernardi, R. C., Horta, B. A., Araújo, J. Q., Albuquerque, M. G., da Silva, J. F., et al. (2011). Dynamical behaviour of the human β_1 -adrenoceptor under agonist binding. *Mol. Simul.* 37, 907–913. doi:10.1080/08927022.2011.572167
- Hoelz, L. V., Ribeiro, A. A., Bernardi, R. C., Horta, B. A., Albuquerque, M. G., da Silva, J. F., et al. (2012). The role of helices 5 and 6 on the human β_1 -adrenoceptor activation mechanism. *Mol. Simul.* 38, 236–240. doi:10.1080/08927022.2011.616501
- Huang, W., Le, S., Sun, Y., Lin, D. J., Yao, M., Shi, Y., et al. (2022). Mechanical stabilization of a bacterial adhesion complex. *J. Am. Chem. Soc.* 144, 16808–16818. doi:10.1021/jacs.2c03961
- Hughes, M. L., and Dougan, L. (2016). The physics of pulling polypeptides: A review of single molecule force spectroscopy using the afm to study protein unfolding. *Rep. Prog. Phys.* 79, 076601. doi:10.1088/0034-4885/79/7/076601
- Humphrey, W., Dalke, A., and Schulten, K. (1996). Vmd: Visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi:10.1016/0263-7855(96)00018-5
- Jevon, P., Abdelrahman, A., and Pigadas, N. (2021). Management of odontogenic infections and sepsis: An update. *Adj Team* 8, 24–31. doi:10.1038/s41407-021-0520-4

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fchem.2023.1107427/full#supplementary-material>

- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi:10.1063/1.445869
- Josefsson, E., McCrea, K. W., Eidhin, D. N., O'Connell, D., Cox, J., Hook, M., et al. (1998). Three new members of the serine-aspartate repeat protein multigene family of staphylococcus aureus. *Microbiology* 144, 3387–3395. doi:10.1099/0021287-144-12-3387
- Kroon, P. (2020). *Aggregate, automate, assemble*. Groningen: Ph.D. thesis, University of Groningen.
- Kwiecinski, J. M., and Horswill, A. R. (2020). *Staphylococcus aureus* bloodstream infections: Pathogenesis and regulatory mechanisms. *Curr. Opin. Microbiol.* 53, 51–60. doi:10.1016/j.mib.2020.02.005
- Latasa, C., Solano, C., Penadés, J. R., and Lasa, I. (2006). Biofilm-associated proteins. *Comptes rendus Biol.* 329, 849–857. doi:10.1016/j.crv.2006.07.008
- Li Petri, G., Di Martino, S., and De Rosa, M. (2022). Peptidomimetics: An overview of recent medicinal chemistry efforts toward the discovery of novel small molecule inhibitors. *J. Med. Chem.* 65, 7438–7475. doi:10.1021/acs.jmedchem.2c00123
- Lister, J. L., and Horswill, A. R. (2014). *Staphylococcus aureus* biofilms: Recent developments in biofilm dispersal. *Front. Cell. Infect. Microbiol.* 4, 178. doi:10.3389/fcimb.2014.00178
- Liu, Z., Liu, H., Vera, A. M., Bernardi, R. C., Tinnfeld, P., and Nash, M. A. (2020). High force catch bond mechanism of bacterial adhesion in the human gut. *Nat. Commun.* 11, 4321. doi:10.1038/s41467-020-18063-x
- Liu, Z., Moreira, R. A., Dujmovic, A., Liu, H., Yang, B., Poma, A. B., et al. (2021). Mapping mechanostable pulling geometries of a therapeutic anticalin/ctla-4 protein complex. *Nano Lett.* 22, 179–187. doi:10.1021/acs.nanolett.1c03584
- Mahmood, M. I., Poma, A. B., and Okazaki, K.-i. (2021). Optimizing Gō-MARTINI Coarse-Grained Model for F-BAR Protein on Lipid Membrane-martini coarse-grained model for f-bar protein on lipid membrane. *Front. Mol. Biosci.* 8, 619381. doi:10.3389/fmolb.2021.619381
- McCormack, M., Smith, A., Akram, A., Jackson, M., Robertson, D., and Edwards, G. (2015). *Staphylococcus aureus* and the oral cavity: An overlooked source of carriage and infection? *Am. J. Infect. Control* 43, 35–37. doi:10.1016/j.ajic.2014.09.015
- McDevitt, D., Francois, P., Vaudaux, P., and Foster, T. (1994). Molecular characterization of the clumping factor (fibrinogen receptor) of staphylococcus aureus. *Mol. Microbiol.* 11, 237–248. doi:10.1111/j.1365-2958.1994.tb00304.x
- Melo, M. C., Bernardi, R. C., De La Fuente-Nunez, C., and Luthy-Schulten, Z. (2020). Generalized correlation-based dynamical network analysis: A new high-performance approach for identifying allosteric communications in molecular dynamics trajectories. *J. Chem. Phys.* 153, 134104. doi:10.1063/5.0018980
- Melo, M. C., Gomes, D. E., and Bernardi, R. C. (2022). Molecular origins of force-dependent protein complex stabilization during bacterial infections. *J. Am. Chem. Soc.* 145, 70–77. doi:10.1021/jacs.2c07674
- Mendes, Y. S., Alves, N. S., Souza, T. L., Sousa, I. P., Jr, Bianconi, M. L., Bernardi, R. C., et al. (2012). *The structural dynamics of the flavivirus fusion peptide-membrane interaction*.
- Milles, L. F., Schulten, K., Gaub, H. E., and Bernardi, R. C. (2018). Molecular mechanism of extreme mechanostability in a pathogen adhesin. *Science* 359, 1527–1533. doi:10.1126/science.aar2094
- Ní Eidhin, D., Perkins, S., Francois, P., Vaudaux, P., Höök, M., and Foster, T. J. (1998). Clumping factor b (clfb), a new surface-located fibrinogen-binding adhesin of staphylococcus aureus. *Mol. Microbiol.* 30, 245–257. doi:10.1046/j.1365-2958.1998.01050.x
- O'Connell, D. (2003). Dock, lock and latch. *Nat. Rev. Microbiol.* 1, 171. doi:10.1038/nrmicro788
- Otto, M. (2009). Staphylococcus epidermidis—the ‘accidental’ pathogen. *Nat. Rev. Microbiol.* 7, 555–567. doi:10.1038/nrmicro2182
- Patti, J. M., Allen, B. L., McGavin, M. J., and Höök, M. (1994). Mscramm-mediated adherence of microorganisms to host tissues. *Annu. Rev. Microbiol.* 48, 585–617. doi:10.1146/annurev.mi.48.100194.003101
- Phillips, J. C., Hardy, D. J., Maia, J. D., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., et al. (2020). Scalable molecular dynamics on cpu and gpu architectures with namd. *J. Chem. Phys.* 153, 044130. doi:10.1063/5.0014475
- Poma, A. B., Cieplak, M., and Theodorakis, P. E. (2017). Combining the martini and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins. *J. Chem. Theory Comput.* 13, 1366–1374. doi:10.1021/acs.jctc.6b00986
- Poma, A. B., Guzman, H. V., Li, M. S., and Theodorakis, P. E. (2019). Mechanical and thermodynamic properties of aβ42, aβ40, and α-synuclein fibrils: A coarse-grained method to complement experimental studies. *Beilstein J. Nanotechnol.* 10, 500–513. doi:10.3762/bjnano.10.51
- Ponnuraj, K., Bowden, M. G., Davis, S., Gurusiddappa, S., Moore, D., Choe, D., et al. (2003). A “dock, lock, and latch” structural model for a staphylococcal adhesin binding to fibrinogen. *Cell* 115, 217–228. doi:10.1016/s0092-8674(03)00809-2
- Ribeiro, J. V., Bernardi, R. C., Rudack, T., Stone, J. E., Phillips, J. C., Freddolino, P. L., et al. (2016). Qwikmd—Integrative molecular dynamics toolkit for novices and experts. *Sci. Rep.* 6, 1–14. doi:10.1038/srep26536
- Ryden, C., Maxe, I., Franzén, A., Ljungh, A., Heinegård, D., and Rubin, K. (1987). Selective binding of bone matrix sialoprotein to staphylococcus aureus in osteomyelitis. *Lancet* 330, 515. doi:10.1016/s0140-6736(87)91830-7
- Schoeler, C., Bernardi, R. C., Malinowska, K. H., Durner, E., Ott, W., Bayer, E. A., et al. (2015). Mapping mechanical force propagation through biomolecular complexes. *Nano Lett.* 15, 7370–7376. doi:10.1021/acs.nanolett.5b02727
- Schoeler, C., Malinowska, K. H., Bernardi, R. C., Milles, L. F., Jobst, M. A., Durner, E., et al. (2014). Ultrastable cellulosome-adhesion complex tightens under load. *Nat. Commun.* 5, 5635–5638. doi:10.1038/ncomms6635
- Sedlak, S. M., Schendel, L. C., Gaub, H. E., and Bernardi, R. C. (2020). Streptavidin/biotin: Tethering geometry defines unbinding mechanics. *Sci. Adv.* 6, eaay5999. doi:10.1126/sciadv.aay5999
- Sedlak, S. M., Schendel, L. C., Melo, M. C., Pippig, D. A., Luthy-Schulten, Z., Gaub, H. E., et al. (2018). Direction matters: Monovalent streptavidin/biotin complex under load. *Nano Lett.* 19, 3415–3421. doi:10.1021/acs.nanolett.8b04045
- Seppälä, J., Bernardi, R. C., Haataja, T. J., Hellman, M., Pentikäinen, O. T., Schulten, K., et al. (2017). Skeletal dysplasia mutations effect on human filamins' structure and mechanosensing. *Sci. Rep.* 7, 4218. doi:10.1038/s41598-017-04441-x
- Sharma, D., Misba, L., and Khan, A. U. (2019). Antibiotics versus biofilm: An emerging battleground in microbial communities. *Antimicrob. Resist. Infect. Control* 8, 76–10. doi:10.1186/s13756-019-0533-3
- Souza, P. C., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grünewald, F., et al. (2021). Martini 3: A general purpose force field for coarse-grained molecular dynamics. *Nat. methods* 18, 382–388. doi:10.1038/s41592-021-01098-3
- Souza, P. C. T., Thallmair, S., Marrink, S. J., and Mera-Adasme, R. (2019). An allosteric pathway in copper, zinc superoxide dismutase unravels the molecular mechanism of the g93a amyotrophic lateral sclerosis-linked mutation. *J. Phys. Chem. Lett.* 10, 7740–7744. doi:10.1021/acs.jpclett.9b02868
- Spiegel, J., Mas-Moruno, C., Kessler, H., and Lubell, W. D. (2012). Cyclic aza-peptide integrin ligand synthesis and biological activity. *J. Org. Chem.* 77, 5271–5278. doi:10.1021/jo300311q
- Suresh, M. K., Biswas, R., and Biswas, L. (2019). An update on recent developments in the prevention and treatment of staphylococcus aureus biofilms. *Int. J. Med. Microbiol.* 309, 1–12. doi:10.1016/j.ijmm.2018.11.002
- Tironi, I. G., Sperb, R., Smith, P. E., and van Gunsteren, W. F. (1995). A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* 102, 5451–5459. doi:10.1063/1.469273
- Tung, H.-s., Guss, B., Hellman, U., Persson, L., Rubin, K., and Rydén, C. (2000). A bone sialoprotein-binding protein from staphylococcus aureus: A member of the staphylococcal sdr family. *Biochem. J.* 345, 611–619. doi:10.1042/bj3450611
- Vanzielegheem, T., Herman-Bausier, P., Dufrene, Y. F., and Mahillon, J. (2015). Staphylococcus epidermidis affinity for fibrinogen-coated surfaces correlates with the abundance of the sdrg adhesin on the cell surface. *Langmuir* 31, 4713–4721. doi:10.1021/acs.langmuir.5b00360
- Verdorfer, T., Bernardi, R. C., Meinhold, A., Ott, W., Luthy-Schulten, Z., Nash, M. A., et al. (2017). Combining in vitro and in silico single-molecule force spectroscopy to characterize and tune cellulosomal scaffoldin mechanics. *J. Am. Chem. Soc.* 139, 17841–17852. doi:10.1021/jacs.7b07574
- Verlet, L. (1967). Computer “experiments” on classical fluids. i. thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* 159, 98–103. doi:10.1103/physrev.159.98
- Versey, Z., da Cruz Nizer, W. S., Russell, E., Zigic, S., DeZeeuw, K. G., Marek, J. E., et al. (2021). Biofilm-innate immune interface: Contribution to chronic wound formation. *Front. Immunol.* 12, 648554. doi:10.3389/fimmu.2021.648554
- Vitry, P., Valotteau, C., Feuillie, C., Bernard, S., Alsteens, D., Geoghegan, J. A., et al. (2017). Force-induced strengthening of the interaction between staphylococcus aureus clumping factor b and loricrin. *MBio* 8, e01748-17. doi:10.1128/mbio.01748-17
- Webb, B., and Salí, A. (2016). Comparative protein structure modeling using modeller. *Curr. Protoc. Bioinforma.* 54, 5.6.1–5.6.37. doi:10.1002/cpbi.3
- Wertheim, H. F., Vos, M. C., Ott, A., van Belkum, A., Voss, A., Kluytmans, J. A., et al. (2004). Risk and outcome of nosocomial staphylococcus aureus bacteraemia in nasal carriers versus non-carriers. *Lancet* 364, 703–705. doi:10.1016/s0140-6736(04)16897-9
- Wolek, K., Gómez-Sicilia, A., and Cieplak, M. (2015). Determination of contact maps in proteins: A combination of structural and chemical approaches. *J. Chem. Phys.* 143, 243105. doi:10.1063/1.4929599
- Zhang, X., Wu, M., Zhuo, W., Gu, J., Zhang, S., Ge, J., et al. (2015). Crystal structures of bbp from staphylococcus aureus reveal the ligand binding mechanism with fibrinogen α. *Protein & Cell* 6, 757–766. doi:10.1007/s13238-015-0205-x
- Zhang, Y., Wu, M., Hang, T., Wang, C., Yang, Y., Pan, W., et al. (2017). Staphylococcus aureus sdr captures complement factor h's c-terminus via a novel ‘close, dock, lock and latch’ mechanism for complement evasion. *Biochem. J.* 474, 1619–1631. doi:10.1042/bj20170085



OPEN ACCESS

EDITED BY

Sergio Pantano,
Pasteur Institute of Montevideo, Uruguay

REVIEWED BY

Simón Poblete,
Universidad Austral de Chile, Chile

*CORRESPONDENCE

Alberto Perez,
✉ perez@chem.ufl.edu

SPECIALTY SECTION

This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

RECEIVED 24 November 2022

ACCEPTED 27 January 2023

PUBLISHED 13 February 2023

CITATION

Liu Q and Perez A (2023), Assessing a computational pipeline to identify binding motifs to the $\alpha 2\beta 1$ integrin. *Front. Chem.* 11:1107400. doi: 10.3389/fchem.2023.1107400

COPYRIGHT

© 2023 Liu and Perez. 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.

Assessing a computational pipeline to identify binding motifs to the $\alpha 2\beta 1$ integrin

Qianchen Liu and Alberto Perez*

Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, FL, United States

Integrins in the cell surface interact with functional motifs found in the extracellular matrix (ECM) that queue the cell for biological actions such as migration, adhesion, or growth. Multiple fibrous proteins such as collagen or fibronectin compose the ECM. The field of biomechanical engineering often deals with the design of biomaterials compatible with the ECM that will trigger cellular response (e.g., in tissue regeneration). However, there are a relative few number of known integrin binding motifs compared to all the possible peptide epitope sequences available. Computational tools could help identify novel motifs, but have been limited by the challenges in modeling the binding to integrin domains. We revisit a series of traditional and novel computational tools to assess their performance in identifying novel binding motifs for the I-domain of the $\alpha 2\beta 1$ integrin.

KEYWORDS

molecular recognition, integrin, AlphaFold, molecular modeling, binding

1 Introduction

The integrin superfamily (Hynes, 1987) encompasses 24 different integrins in humans responsible for communication and signaling between cells and with the extracellular matrix (ECM). Structurally, they are $\alpha\beta$ heterodimers with two non-covalent subunits (arising from 18 α and 8 β subunits) located on the cell's membrane (Hynes, 2002; Takada et al., 2007). Their normal behavior controls cellular processes such as cell adhesion, migration and differentiation [(Critchley et al., 1999); (Mizuno et al., 2000); (Mercurio et al., 2001)]. Usually, these integrins recognize specific peptide epitope motifs present in large fibrous proteins that form the extracellular matrix such as collagen or fibronectins (see Figure 1). Hence, designing molecules that disrupt or enhance these interactions has long been a potential therapeutic target. A recent study (Slack et al., 2022) shows over 60 integrin-target therapies have been recorded (<https://www.clinical-trials.gov> and <https://www.clinical-trialsregister.eu/ctrsearch/search> using the search term “integrin”) targeting diseases like Multiple Sclerosis (Kawamoto et al., 2012) or Crohn's disease (Hutchinson, 2007). Most binding occurs through an “I-like domain” in the β subunit which contains a “metal ion-dependent adhesion site” (MIDAS). Some peptide epitope binding motifs like RGD (Arginine-Glycine-Aspartic) are present in many ECM fibers and bind many integrins (Hatley et al., 2018). However, there is selectivity and specificity among their ligands—and even for the RGD motif there is an interplay between the conformation it adopts and the specificity to a particular integrin (Aumailley et al., 1991; Kapp et al., 2017). In the field of biomaterial engineering, there is growing interest to develop computational pipelines that can identify functional motifs to incorporate into engineered ECMs that trigger cellular response (Perez et al., 2021).

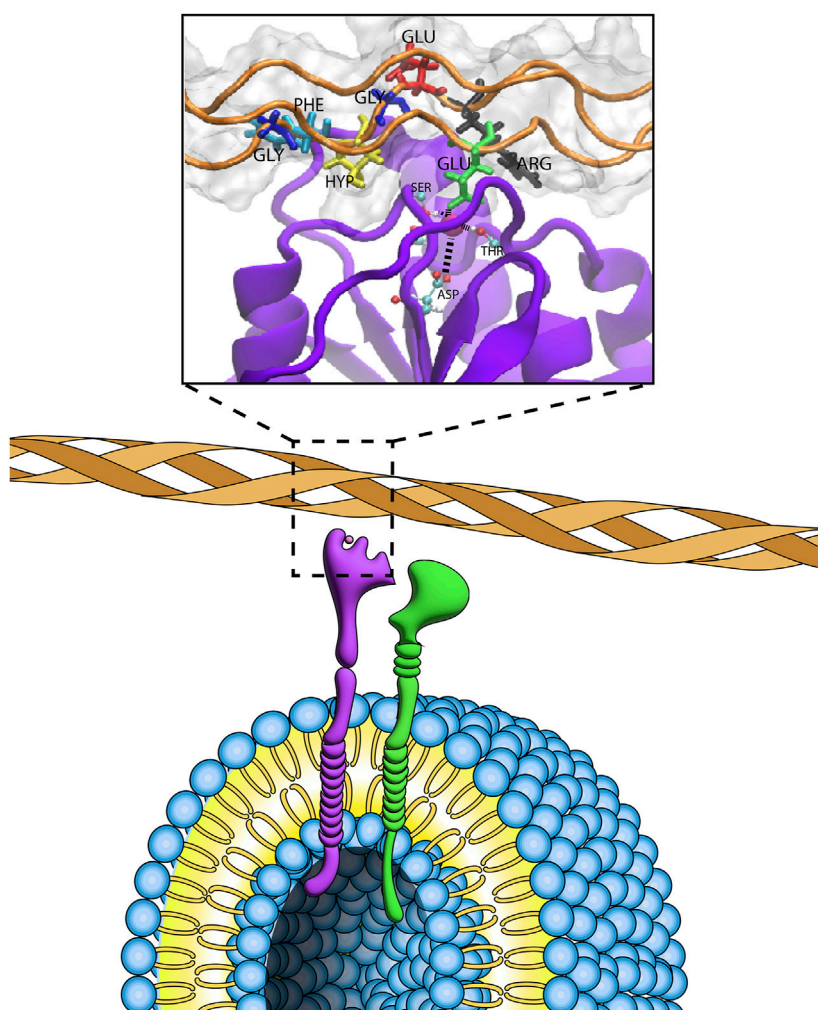


FIGURE 1

System of study. Artistic representation of the $\alpha 2 \beta 1$ binding collagen. The inset corresponds to the PDB structure 1dzi, focusing on the specific motif area "GFOGER" on the collagen fiber (orange).

Our existing understanding of integrin-ligand recognition has mainly been driven from experimental observations including affinity chromatography [Otey et al. \(1993\)](#), antibodies against cell epitopes [Ley et al. \(2016\)](#), and the use of NMR experiments [Siebert et al. \(2010\)](#). Computational tools on the other hand have been challenged by the complexity of modeling integrin-ECM interactions as well as the diversity of function/structure relationships arising from the multidomain architecture that merit attention such as the origin of selectivity, mechanism for signal transduction ([Kalli et al., 2011](#)), effect of the lipid environment ([Kalli et al., 2017](#)), or interaction between the different domains and their role in active/inactive conformations to name a few ([Chen et al., 2011](#)). Although the number of computational studies for integrin systems is limited, there is a wide range of approaches that have been used including physics based approaches such as docking ([Guzzetti et al., 2017](#)), atomistic and coarse grained molecular dynamics (MD) ([Craig et al., 2004](#); [Murcia et al., 2008](#); [Choi et al., 2010](#); [Zhu et al., 2010](#); [Wang et al., 2015](#); [Farina et al., 2016](#); [Fravev and Sirimulla, 2019](#)), QM/MM

approaches ([Freindorf et al., 2012](#)), and machine learning ([Mehdi et al., 2013](#); [Prytuliak et al., 2017](#); [Asgari et al., 2019](#)). Typically, ligand docking calculations are applied to filter ligands with high affinity, MD approaches are used to either predict free energy differences with thermodynamic integration (TI) or conformational changes *via* enhanced sampling, while ML approaches have been traditionally used to discover new binding motifs in protein-peptide complexes such as the well-known RGD, GPR (the recognition site for $\alpha \times \beta 2$), or DLLEL (the binding site for $\alpha \times \beta 6$) for integrins.

We focus on the I-domain of the $\alpha 2 \beta 1$ integrin, which contains a binding motif and has been shown to retain the binding activity of the whole integrin in recombinant studies expressing only the I motif (PDB code 1dzi) ([Emsley et al., 2000](#)). The binding domain undergoes a conformational change between the unbound and bound forms in which three loops participate in coordinating a central metal ion, with a glutamic acid from the collagen completing the coordination of the metal ([Emsley et al., 2000](#)). The collagen used here introduces a six aminoacid peptide motif (GFOGER, where O

stands for hydroxyproline), that forms triple helices analogous to canonical collagen. Even though the three strands are homologous for triple-helix formation, during binding each strand becomes distinct, with one containing a critical Glutamic acid residue (E) for binding (“leading strand”). By comparison, the other two strands have been previously named “middle” and “trailing” strands) (Emsley et al., 2000). Given the 20^6 possible peptide sequences covering the length of the *GFOGER* motif, we expect there are many other sequences that might bind this integrin. Indeed, amongst integrins that bind collagen, there are differences amongst canonical motifs (*GxOGER*, where $x = F, L, M, A$) and non-canonical motifs (Hamaia et al., 2012). Hence we ask the question of whether computational pipelines can suggest new motifs and if they are capable of assessing which of those suggested motifs are actually better binders.

We seek to assess the advantages/disadvantages of using traditional and novel pipelines combining multiple computational techniques readily available. We divide the pipelines in three stages: 1) predicting new motifs, 2) predicting their ability to bind, and 3) predicting their stability. Overall, finding new interacting motifs against integrins remains challenging regardless of the pipeline used.

2 Computational methods

2.1 Identification of novel motifs

We started from the X-ray crystal structure of the $\alpha 2$ I domain from $\alpha 2\beta 1$ in complex with collagen [PDBid: 1dzi (Emsley et al., 2000)] and performed a scan of all possible mutations (for the 20 common amino acids) at each position along the “*GFOGER*” motif, collecting the expected free energy changes ($\Delta\Delta G$) these programs predict. Integrin complexes were first optimized in the FoldX suite. Next, a position scan was conducted with the command “Position Scan” on the *GFOGER* motif and the output results showed the difference of binding energy for each mutation per amino acid on collagen. $\Delta\Delta G_{bind}$ was also calculated using *RosettaDDG* predictions, with the backrub trajectory stride set to 35,000 and making three trials for each $\Delta\Delta G$ calculation.

The ProteinMPNN (message passing neural network) (Dauparas et al., 2022) has recently been developed as a way to identify the ideal sequence that will adopt a certain 3D structure. In this model, we provided the PDB structure of the complex and asked the model to design new motifs to replace the native *GFOGER* motif.

2.2 Stability MD simulations

We used standard minimization and equilibration protocols (Braun et al., 2018) followed by production runs using Langevin dynamics for 500 ns in the NPT ensemble using a Monte Carlo barostat (Åqvist et al., 2004). Simulations used AMBER’s (Case et al., 2020) *pmemd* module (Salomon-Ferrer et al., 2013). We simulated the top 20 FoldX and Rosetta predictions using ff14SB (Maier et al., 2015) solvated in a truncated octahedron box [OPC water model (Izadi et al., 2014)], and 150 mM concentration of Na^+ and Cl^- ions (Joung and Cheatham, 2008). As a control, we simulated the I-domain in the presence and absence of the wild

type (WT) collagen (PDBid: 1dzi). All simulations were carried out with a Co^{2+} ion in the MIDAS binding site. We simulated 10,000 steps of energy minimization, switching from steepest descent to conjugate gradient after 5,000 cycles. The resulting minimized system was heated from 0 to 100 K in NVT condition for 50 ps with Langevin dynamics, and 100–300 K in NPT for 500 ps using Langevin dynamics, followed by a short (5 ns) equilibration process at constant pressure (1 atm) and temperature (300 K). Finally, unbiased and unrestrained system went through production in a periodic boundary condition for 500 ns in NPT by Langevin thermostat and Monte Carlo barostat conditions. Bonds involving hydrogen were constrained by the SHAKE algorithm. Cpptraj (Roe and Cheatham, 2013) was used to analyze the root mean square deviation (RMSD) and Dynamical Cross Correlation (Kamberaj and Vaart, 2009) within the ensembles comparing them to the wild type complex.

2.3 Structure predictions with AlphaFold

We used Alphafold Multimer (Evans et al., 2021b) to predict the structure of the complex using either sequence data or templates (containing the collagen and integrin domain far from each other). Results were analyzed in terms of the predicted local distance difference test (pLDDT) score as is standard in the field (Jumper et al., 2021). In short, the pLDDT score gives a per residue and global value to show how confident the Alpha Fold prediction results are. Results above 80 typically reflect high confidence in the prediction.

2.4 Thermodynamic integration (TI) calculations

TI was used to calculate the relative binding affinity ($\Delta\Delta G_{bind}^{mutant-WT}$) between collagen and $\alpha 2\beta 1$ upon mutation of certain residues in collagen. Here, we applied “One-step” transformations (Steinbrecher et al., 2011) to decrease the simulation time, in which electrostatic and van der Waals forces are varied synchronously (Shirts et al., 2003). The initial system was prepared using AMBER’s *tiMerge* to eliminate redundant bonding terms and increase calculation efficiency. We ran TI simulations with *pmemd*. The complex and mutant ligand were solvated separately in a cubic box with explicit OPC (Izadi et al., 2014) water and a 10 Å clearance. We employed ff14SB (Maier et al., 2015) for the protein parameters and general AMBER force field (He et al., 2020) for general atom and bonds parameters. Minimization, heating and equilibration process was performed in the NVT ensemble with a Monte Carlo barostat. The TI production phase was done in the NPT ensemble (300 K and 1 atm), running for 500 ns Softcore potentials were applied to reduce issues with the integration step at the endpoints (Steinbrecher et al., 2011). Eleven independent MD simulations were performed spaced evenly between the end-points ($\lambda \in [0, 1]$). We performed six replicates for each simulated system. The average and standard deviation for $\Delta\Delta G_{bind}$ were calculated from the differences amongst replicates.

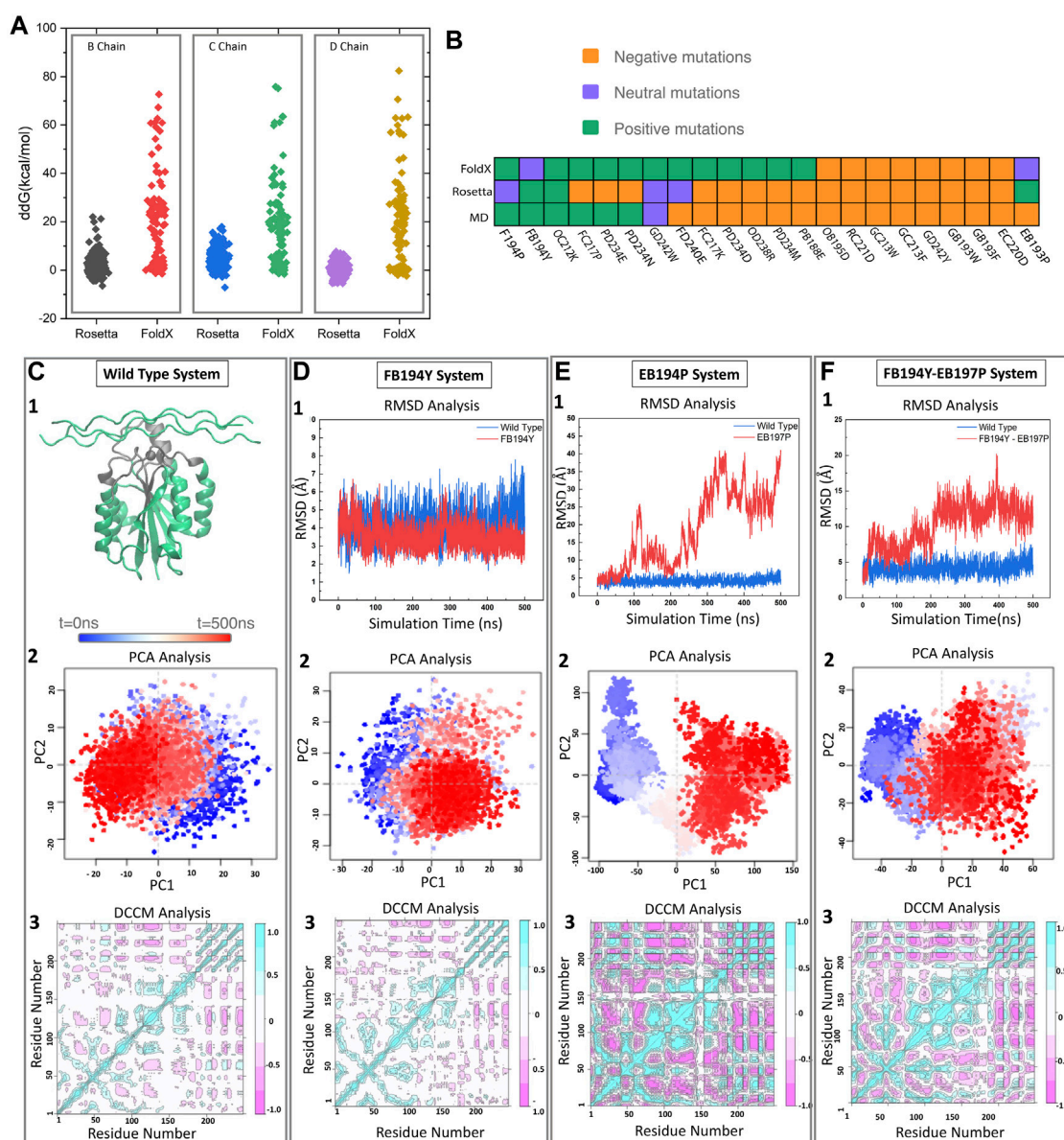


FIGURE 2

Pipeline for selecting new motifs. (A) FoldX and Rosetta are used to estimate relative free energy changes upon mutating each residue in the binding motif to all possible amino acids. (B) Predicted effect of mutations by three different methods (FoldX, Rosetta, and MD) for a set of 22 mutations. (C) The wild type samples a single state throughout the trajectory as identified by projecting onto the two first principal components. Reference Dynamic cross correlation matrix (DCCM) for the wild type state. (D–F) Examples of a stable (D) and unstable (E,F) mutations as identified from RMSD, PCA, and DCCM.

2.5 Sampling collagen binding modes with MELD

The Modeling Employing Limited Data (MELD) approach uses H,T-REMD (Sugita and Okamoto, 1999) to sample rare events. The method changes the Hamiltonian by enforcing information that guides to different conformations that might be compatible with the end state. The caveat is that the data is framed as ambiguous and noisy—thus MELD relies on Bayesian inference to identify the best interpretation of the data compatible with the forcefield. In this

process, analyzing the resulting ensemble (e.g., through clustering) identifies the states (conformations) most compatible with the information and force field.

To guide the binding process we first placed harmonic distance restraints amongst native contacts in the integrin (so it would not unfold), and also between the three collagen strands, so it would not dissociate. We then selected residues in the active site of the integrin and in those of the collagen binding motif. Based on those two lists of residues, we generated a list of twenty five possible contacts (some of which were present in the native state and some of which were not).

We found that when enforcing 15 or more restraints, replica exchanges were inefficient, leading to poor sampling. At the other extreme, satisfying less than four restraints sampling was not restrictive enough to sample native-like bound conformations. We thus required that only eight restraints out of the 25 possible ones be satisfied. Satisfying different subsets of eight restraints give rise to different binding modes.

MELD simulations used the ff14SB force field (Maier et al., 2015) for side chains and ff99SB (Hornak et al., 2006) for backbone, together with the GBneck2 (Nguyen et al., 2013) implicit solvent model. The collagen fiber was placed over 30 Å away from the integrin. The temperature range was set between 300 and 500 K, with 30 replicas. Ensembles were analyzed using hierarchical clustering as implemented in CPPTRAJ (Roe and Cheatham, 2013) with an $\epsilon = 2$ value, including heavy atoms at the interface of the complex in the native state.

3 Results

3.1 Local search for new interaction motifs

Traditional design strategies start with a known binding motif and search for single amino acid mutants that increase binding affinity ($\Delta\Delta G_{bind}$). Such strategies lead to local sequence optimization, with designs similar to the original motif. Here we used FoldX (Schymkowitz et al., 2005) and Rosetta (Barlow et al., 2018) (see methods), two traditional approaches with varying computational cost and success rate. We observed that FoldX single point mutations have a wider $\Delta\Delta G_{bind}$ distribution, and are generally shifted towards higher energies (see Figure 2A). While there is good agreement on the failed mutations, the more computationally demanding Rosetta is better at discriminating mutations that FoldX finds favorable.

To further assess the predicted motifs with an independent methodology, we performed MD simulations of a selected group of 15 mutants. We expect that monitoring standard structural and dynamical properties like RMSD and dynamical cross correlation functions would be enough to distinguish those mutations that remain stable in the 500 ns timescale vs. those that are unlikely to bind (see Figures 2C–F). We monitored the RMSD of the interface region, defined as heavy atom contacts to collagen in the native structure (using a 10 Å cutoff). In this timescale, the integrin oscillates around 1 Å from the initial structures, with few deviations to higher RMSD values (2.5 Å). In the presence of collagen we observe a similar behavior, where there are no deviations to larger RMSD states in the 500 ns timescale. The RMSD of the whole complex oscillates at around 4 Å. Figure 2 showcases the behavior of the wild type, neutral and negative mutation on sampling [RMSD and projection on the top two principal components using the Bio3d package (Grant et al., 2006)]. Figure 2 exemplifies a negative control mutation (which rapidly dissociates) and a neutral mutation that remains close to the starting conformation.

We find that the more computationally efficient FoldX is capable of filtering out mutations that are likely detrimental to the binding affinity. While the ones predicted to be beneficial do not always agree with MD and Rosetta results (see Figure 2B). We notice several disagreements with Rosetta and MD—this is not surprising as Rosetta has been designed to predict free energy differences while

short conventional MD trajectories do not contain enough sampling to assess the free energy. We thus decided to perform thermodynamic integration calculations to further identify the agreement between Rosetta and MD-based approaches.

Thermodynamic integration increases the complexity in system setup and analysis with respect a conventional MD trajectory—but the computational costs (considering replicates needed, see methods) remains relatively small compared to other MD approaches. We selected 15 mutations and compared results using Rosetta and TI (see Supplementary Figure S1). For most residue mutations, both programs agree in sign if not in magnitude. Previous work points to systems including multiple binding modes or systems that are sensitive to local conformational changes (such as the MIDAS binding site) (Armacost et al., 2020) as problematic for TI. For example, Guest and coworkers performed free energy perturbation studies on a series of small molecule inhibitors to the $\beta 6$ integrin with an average error of 1.5 kcal/mol with respect to the experimental results (Guest et al., 2020).

We searched for alternative binding modes by using the MELD approach, which can simulate multiple binding/unbinding events. MELD combines ambiguous/noisy information with molecular simulations through Bayesian inference and has been routinely used for predict the binding of macromolecules [protein-protein (Brini et al., 2019), protein-peptide (Morrone et al., 2017; Mondal et al., 2022), protein-DNA (Bauzá and Pérez, 2021), and protein-small molecule (Liu et al., 2020)]. We derived ambiguous information based on native contacts present in the crystal structure in such a way that different interpretations of the data is compatible with different binding modes. We expected, that the force field would be able to recognize the most native-like amongst the binding modes for those sequences that have a high affinity (clusters with high population) (Lang and Perez, 2021). Unfortunately, due to the small interface region between collagen and the integrin, the different binding modes found give rise to large deviations in binding angles between the collagen in MELD simulations with respect to the native structure (see Supplementary Figure S2). On the other hand, satisfying more information overrides the force field preferences and yields native-like binding modes regardless of the sequence. Similarly, competitive binding simulations (Morrone et al., 2017) with MELD also failed to distinguish which collagen mutations were more likely to lead to more stable complexes. Presumably, these limitations arise from the use of an implicit solvent (Nguyen et al., 2013) needed for the MELD binding simulations.

Similarly, the recent successes of the AlphaFold (AF) (Evans et al., 2021a) machine learning approach did not translate to this system. We used a local installation of AlphaFold and performed predictions in the presence/absence of structural templates. In our hands, AlphaFold multimer predictions were confident about the $\alpha 2$ I-domain structure (high pLDDT scores), but failed to predict the structure of the collagen triple helix structure—and hence of the complex (see Supplementary Figure S2).

3.2 Recent machine learning approaches can suggest novel sequences based on the structure

Whereas we used FoldX and Rosetta to predict local changes in the sequence (single mutants), the recent protein MPNN (Dauparas

et al., 2022) machine learning approach can in principle find an optimal sequence given the structure of the complex. Contrary to the other two methods, this approach does not provide a relative binding affinity. We first generated two predictions in which we allowed any residue in the motif along the tree collagen strands to change (see “Prediction 1” and “Prediction 2” in [Supplementary Figure S3](#)). This gave rise to four different binding motifs. We next generated four more sequences by creating homo-trimer collagen strands with each of the four predicted motifs (see the latter four motifs in [Supplementary Figure S3A](#)). We assessed the viability of these motifs by running conventional MD. All sequences in which the leading strand had an E to P mutation were unstable. Whereas if this mutation occurred in other strands, the system remained stable. This is expected as the Glutamic acid coordinates with a divalent site when interacting with the integrin.

4 Discussion

In this work we focused on identifying collagen-like motifs that bind the I-domain of the $\alpha2\beta1$ integrin. Despite their biological relevance and some successes ([Craig et al., 2004](#); [Murcia et al., 2008](#); [Choi et al., 2010](#); [Zhu et al., 2010](#); [Wang et al., 2015](#); [Farina et al., 2016](#); [Fratev and Sirimulla, 2019](#)), integrins remain challenging systems to study through molecular modeling. The collagen fiber with the *GFOGER* motif that we study was initially suggested based on docking calculations ([Emsley et al., 2000](#)), which led to the crystallization of the complex (pdb code 1dzi). Our use of local (single mutant) and global (proteinMPNN) approaches shows that current methodologies are better at discerning unfavorable mutations than at providing reliable predictions. However, consensus between different methods increases the likelihood of success. Our use of MD stability analysis showed that it can be a helpful tool to distinguish unfavorable mutations, but stable simulations are not a guarantee of favorable mutations as timescales remain limited. This becomes an issue even when using thermodynamic integration, as multiple binding modes are possible. While this is an actively developed field for small molecule binders [Gill et al. \(2018\)](#), it remains more challenging for flexible molecules such as collagen. For such flexible systems, we have previously found the MELD Bayesian inference approach can typically identify differences amongst different binder sequences. Due to the small interface area, our standard protocol results in binding modes where the collagen binds in the right region, but with orientations that can deviate up to 90° from their experimental binding mode. The caveat of increasing the number of restraints in MELD to solve this issue leads to the inability to distinguish motif sequence preferences.

Molecular modeling pipelines are undergoing rapid and drastic changes thanks to the eruption of machine learning approaches. The CASP event served as the perfect scenario for the first iteration of AlphaFold to show the potential of machine learning in protein structure prediction ([Senior et al., 2020](#)). Their initial approach relied on following the leading strategies in the field: determine pairwise distance distributions between residues to impose as restraints to predict structures. Two years later, AlphaFold presented a novel strategy based on attention networks with an impressive performance in CASP ([Jumper et al., 2021](#)). Making the network

available to the community and the appearance of collaborative notebooks ([Mirdita et al., 2022](#)) rapidly allowed groups to apply it to a myriad of problems: for molecular recognition (protein-protein and protein-peptide) ([Humphreys et al., 2021](#); [Tsaban et al., 2022](#)), for predicting multiple biological states ([Wayment-Steele et al., 2022](#)), relative binding affinities ([Chang and Perez, 2022](#)), or even for designing new proteins *via* deep network hallucination ([Anishchenko et al., 2021](#)). As these networks learn from data deposited in the protein data Bank, they also implicitly learn about the position of ions or ligands in active sites. However, AF multimer was not able to predict the structures of the 1dzi complex. Recent work showed that partial retraining of AF weights for specific targets could lead to an improved ability to correctly identify bound or unbound peptides binding to the Major Histocompatibility Complex (MHC) ([Motmaen et al., 2022](#)). This was possible thanks to a large database of peptides known to be either binders/non-binders to MHC. Such type of initiatives could soon provide accurate results for predicting complexes involving integrins, which combined with competitive binding strategies ([Chang and Perez, 2022](#)) could lead to rapid identification of functional motifs.

During the writing of this paper, several new machine learning approaches appeared in the literature which make us optimistic about the future: we highlight three that are relevant to the discussion above. The first one is RosettaFoldNA ([Baek et al., 2022](#)), which predicts the folding of RNA as well as nucleic acid-protein complexes. The approach draws on the AF principles but incorporates an additional physics-inspired term (Lennard Jones potentials taken from Rosetta) to better reproduce geometries (e.g., reduce the overlap between protein and nucleic acids). In this process, the algorithm has learned to assemble double-stranded DNA, much like we hope the collagen triple helix can be predicted. The second development is the OpenFold ([Ahdriz et al., 2022](#)) initiative—a pyTorch-based implementation trainable to reproduce AlphaFold levels of accuracy at a lower computational cost. The authors also report the OpenProteinSet used to train the model. In the last few months, the field used AF beyond what it was originally designed to do. OpenFold will now give users the possibility to retrain a tool equivalent to AF for new purposes. Finally, a recent study ([Akdel et al., 2022](#)) highlights the potentially transformative role of AF in structural biology, its accuracy matching experiments for many applications, as well as the role of potential biases, and its ability to identify features that are not typically present in databases.

5 Conclusion

In this work we assessed the role of different computational tools to identify novel collagen-integrin binding motifs. FoldX serves as a fast mutant screen, to filter out mutations that do not improve binding affinity. A combination of Rosetta and MD (TI) serves to further identify those mutations most likely to lead to improved binding affinities. Although we were very enthusiastic about the possibility of using AlphaFold to differentiate amongst binding motifs, we found no evidence that it could predict the native state. However, in light of recent work it seems like partial retraining of the weights against known binders/non-binders might lead to a feasible pipeline. Finally, proteinMPNN was able

to correctly identify that mutations to the glutamic acid involved in binding would be deleterious only in the leading strand. Although further assessment is needed, proteinMPNN paves the way to identifying functional motifs far from the starting sequence motif.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors upon request.

Author contributions

QL performed computational simulations and analysis. QL and AP discussed the results and wrote the paper.

Funding

Research was sponsored by the Army Research Office under Grant Number W911NF-22-1-0142. The views and conclusion contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ARO or the U.S. Government. The authors acknowledge University of Florida Research Computing for providing computational resources and support that have contributed to the research results reported in this publication.

References

- Ahdritz, G., Bouatta, N., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T. J., et al. (2022). *OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization*. doi:10.1101/2022.11.20.517210
- Akdal, M., Pires, D. E. V., Pardo, E. P., Jänes, J., Zalevsky, A. O., Mészáros, B., et al. (2022). A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* 29, 1056–1067. doi:10.1038/s41594-022-00849-w
- Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., et al. (2021). De novo protein design by deep network hallucination. *Nature* 1, 547–552. doi:10.1038/s41586-021-04184-w
- Åqvist, J., Wennerström, P., Nervall, M., Bjelic, S., and Brandsdal, B. O. (2004). Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm. *Chem. Phys. Lett.* 384, 288–294. doi:10.1016/j.cplett.2003.12.039
- Armocost, K. A., Riniker, S., and Cournia, Z. (2020). Exploring novel directions in free energy calculations. *J. Chem. Inf. Model.* 60, 5283–5286. doi:10.1021/acs.jcim.0c01266
- Asgari, E., McHardy, A. C., and Mofrad, M. R. K. (2019). Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). *Sci. Rep.* 9, 3577. doi:10.1038/s41598-019-38746-w
- Aumailley, M., Gurrath, M., Müller, G., Calvete, J., Timpl, R., and Kessler, H. (1991). Arg-gly-asp constrained within cyclic pentapeptides strong and selective inhibitors of cell adhesion to vitronectin and laminin fragment p1. *FEBS Lett.* 291, 50–54. doi:10.1016/0014-5793(91)81101-d
- Baek, M., McHugh, R., Anishchenko, I., Baker, D., and DiMaio, F. (2022). *Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA*. doi:10.1101/2022.09.09.507333
- Barlow, K. A., Ó Conchúir, S., Thompson, S., Suresh, P., Lucas, J. E., Heinonen, M., et al. (2018). Flex ddp: Rosetta Ensemble-Based estimation of changes in Protein-Protein binding affinity upon mutation. *J. Phys. Chem. B* 122, 5389–5399. doi:10.1021/acs.jpcc.7b11367
- Bauzá, A., and Pérez, A. (2021). MELD-DNA: A new tool for capturing protein-DNA binding. *bioRxiv*. 2021.06.24.449809. doi:10.1101/2021.06.24.449809
- Braun, E., Gilmer, J., Mayes, H. B., Mobley, D. L., and Monroe, J. I. (2018). *Best practices for foundations in molecular simulations*. Article v1. 0.
- Brini, E., Kozakov, D., and Dill, K. (2019). Predicting protein dimer structures using MELD × MD. *J. Chem. theory Comput.* 15, 3381–3389. doi:10.1021/acs.jctc.8b01208
- Case, D., Belfon, K., Ben-Shalom, I., Brozell, S., Cerutti, D., Cheatham, T., et al. (2020). *Amber 2020*.
- Chang, L., and Perez, A. (2022). Ranking peptide binders by affinity with AlphaFold. *Angew. Chem. Int. Ed.*, e202213362. doi:10.1002/anie.202213362
- Chen, W., Lou, J., Hsin, J., Schulten, K., Harvey, S. C., and Zhu, C. (2011). Molecular dynamics simulations of forced unbending of integrin V3. *PLoS Comput. Biol.* 7, e1001086. doi:10.1371/journal.pcbi.1001086
- Choi, Y., Kim, E., Lee, Y., Han, M. H., and Kang, I.-C. (2010). Site-specific inhibition of integrin $\alpha_5\beta_1$ -vitronectin association by a ser-asp-val sequence through an Arg-Gly-Asp-binding site of the integrin. *Proteomics* 10, 72–80. doi:10.1002/pmic.200900146
- Craig, D., Gao, M., Schulten, K., and Vogel, V. (2004). Structural insights into how the MIDAS ion stabilizes integrin binding to an RGD peptide under force. *Structure* 12, 2049–2058. doi:10.1016/j.str.2004.09.009
- Critchley, D. R., Holt, M. R., Barry, S. T., Priddle, H., Hemmings, L., and Norman, J. (1999). Integrin-mediated cell adhesion: The cytoskeletal connection. *Biochem. Soc. Symp.* 65, 79–99.
- Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., et al. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. *Science* 378, 49–56. doi:10.1126/science.add2187
- Emsley, J., Knight, C., Farnale, R. W., Barnes, M. J., and Liddington, R. C. (2000). Structural basis of collagen recognition by integrin α_1 . *Cell* 101, 47–56. doi:10.1016/s0092-8674(00)80622-4
- Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. (2021a). Protein complex prediction with alphafold-multimer. *bioRxiv*. doi:10.1101/2021.10.04.463034
- Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al. (2021b). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. 2021.10.04.463034. doi:10.1101/2021.10.04.463034

Acknowledgments

We are thankful for startup resources to run simulations on UF research computing resources.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fchem.2023.1107400/full#supplementary-material>

- Farina, B., de Paola, I., Russo, L., Capasso, D., Liguoro, A., Gatto, A. D., et al. (2016). A combined NMR and computational approach to determine the RGD α 5 β 1 integrin recognition mode in isolated cell membranes. *Chemistry* 22, 681–693. doi:10.1002/chem.201503126
- Frave, F., and Sirimulla, S. (2019). An improved free energy perturbation FEP+ sampling protocol for flexible ligand-binding domains. *Sci. Rep.* 9, 16829. doi:10.1038/s41598-019-53133-1
- Freindorf, M., Furlani, T. R., Kong, J., Cody, V., Davis, F. B., and Davis, P. J. (2012). Combined QM/MM study of thyroid and steroid hormone analogue interactions with integrin. *J. Biomed. Biotechnol.* 2012, 1–12. doi:10.1155/2012/959057
- Gill, S. C., Lim, N. M., Grinaway, P. B., Rustenburg, A. S., Fass, J., Ross, G. A., et al. (2018). Binding modes of ligands using enhanced sampling (BLUES): Rapid decorrelation of ligand binding modes via nonequilibrium candidate Monte Carlo. *J. Phys. Chem. B* 122, 5579–5598. doi:10.1021/acs.jpcc.7b11820
- Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A., and Caves, L. S. (2006). Bio3d: An r package for the comparative analysis of protein structures. *Bioinformatics* 22, 2695–2696. doi:10.1093/bioinformatics/btl461
- Guest, E. E., Oatley, S. A., Macdonald, S. J. F., and Hirst, J. D. (2020). Molecular simulation of v6 integrin inhibitors. *J. Chem. Inf. Model.* 60, 5487–5498. doi:10.1021/acs.jcim.0c00254
- Guzzetti, I., Civera, M., Vasile, F., Arosio, D., Tringali, C., Piarulli, U., et al. (2017). Insights into the binding of cyclic RGD peptidomimetics to α 5 β 1 integrin by using Live-Cell NMR and computational studies. *ChemistryOpen* 6, 128–136. doi:10.1002/open.201600112
- Hamaia, S. W., Pugh, N., Raynal, N., Némaz, B., Stone, R., Gullberg, D., et al. (2012). Mapping of potent and specific binding motifs, GLOGEN and GVOGEA, for integrin 11 using collagen toolkits II and III. *J. Biol. Chem.* 287, 26019–26028. doi:10.1074/jbc.m112.353144
- Hatley, R. J. D., Macdonald, S. J. F., Slack, R. J., Le, J., Ludbrook, S. B., and Lukey, P. T. (2018). Anv-RGD integrin inhibitor toolbox: Drug discovery insight, challenges and opportunities. *Angew. Chem. Int. Ed.* 57, 3298–3321. doi:10.1002/anie.201707948
- He, X., Liu, S., Lee, T.-S., Ji, B., Man, V. H., York, D. M., et al. (2020). Fast, accurate, and reliable protocols for routine calculations of protein–ligand binding affinities in drug design projects using AMBER GPU-TI with ff14SB/GAFF. *ACS Omega* 5, 4611–4619. doi:10.1021/acsomega.9b04233
- Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65, 712–725. doi:10.1002/prot.21123
- Humphreys, I. R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., et al. (2021). Computed structures of core eukaryotic protein complexes. *Science* 374, ea6m4805. doi:10.1126/science.abm4805
- Hutchinson, M. (2007). Natalizumab: A new treatment for relapsing remitting multiple sclerosis. *Ther. Clin. Risk Manag.* 3, 259–268. doi:10.2147/tcrm.2007.3.2.259
- Hynes, R. O. (1987). Integrins: A family of cell surface receptors. *Cell* 48, 549–554. doi:10.1016/0092-8674(87)90233-9
- Hynes, R. O. (2002). Integrins: Bidirectional, allosteric signaling machines. *Cell* 110, 673–687. doi:10.1016/s0092-8674(02)00971-6
- Izadi, S., Anandakrishnan, R., and Onufriev, A. V. (2014). Building water models: A different approach. *J. Phys. Chem. Lett.* 5, 3863–3871. doi:10.1021/jz501780a
- Joung, I. S., and Cheatham, T. E. (2008). Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* 112, 9020–9041. doi:10.1021/jp8001614
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2
- Kalli, A. C., Campbell, I. D., and Sansom, M. S. P. (2011). Multiscale simulations suggest a mechanism for integrin inside-out activation. *Proc. Natl. Acad. Sci.* 108, 11890–11895. doi:10.1073/pnas.1104505108
- Kalli, A. C., Rog, T., Vattulainen, I., Campbell, I. D., and Sansom, M. S. P. (2017). The integrin receptor in biologically relevant bilayers: Insights from molecular dynamics simulations. *J. Membr. Biol.* 250, 337–351. doi:10.1007/s00232-016-9908-z
- Kamberaj, H., and Vaart, A. v. d. (2009). Correlated motions and interactions at the onset of the DNA-induced partial unfolding of ets-1. *Biophysical J.* 96, 1307–1317. doi:10.1016/j.bpj.2008.11.019
- Kapp, T. G., Rechenmacher, F., Neubauer, S., Maltsev, O. V., Cavalcanti-Adam, E. A., Zarka, R., et al. (2017). A comprehensive evaluation of the activity and selectivity profile of ligands for RGD-binding integrins. *Sci. Rep.* 7, 39805. doi:10.1038/srep39805
- Kawamoto, E., Nakahashi, S., Okamoto, T., Imai, H., and Shimaoka, M. (2012). Anti-integrin therapy for multiple sclerosis. *Autoimmune Dis.* 2012, 1–6. doi:10.1155/2012/357101
- Lang, L., and Perez, A. (2021). Binding ensembles of p53-MDM2 peptide inhibitors by combining bayesian inference and atomistic simulations. *Molecules* 26, 198. doi:10.3390/molecules26010198
- Ley, K., Rivera-Nieves, J., Sandborn, W. J., and Shattil, S. (2016). Integrin-based therapeutics: Biological basis, clinical use and new drugs. *Nat. Rev. Drug Discov.* 15, 173–183. doi:10.1038/nrd.2015.10
- Liu, C., Brini, E., Perez, A., and Dill, K. A. (2020). Computing ligands bound to proteins using MELD-accelerated MD. *J. Chem. Theory Comput.* 16, 6377–6382. doi:10.1021/acs.jctc.0c00543
- Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. theory Comput.* 11, 3696–3713. doi:10.1021/acs.jctc.5b00255
- Mehdi, A. M., Sehgal, M. S. B., Kobe, B., Bailey, T. L., and Bodén, M. (2013). Dlocalmotif: A discriminative approach for discovering local motifs in protein sequences. *Bioinformatics* 29, 39–46. doi:10.1093/bioinformatics/bts654
- Mercurio, A. M., Rabinovitz, I., and Shaw, L. M. (2001). The α 6 β 4 integrin and epithelial cell migration. *Curr. Opin. cell Biol.* 13, 541–545. doi:10.1016/s0955-0674(00)00249-0
- Mirdita, M., Schütze, K., Moriawaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nat. Methods* 19, 679–682. doi:10.1038/s41592-022-01488-1
- Mizuno, M., Fujisawa, R., and Kuboki, Y. (2000). Type I collagen-induced osteoblastic differentiation of bone-marrow cells mediated by collagen- α 2 β 1 integrin interaction. *J. Cell. physiology* 184, 207–213. doi:10.1002/1097-4652(200008)184:2<207::aid-jcp8>3.0.co;2-u
- Mondal, A., Swapna, G., Hao, J., Ma, L., Roth, M. J., Montelione, G. T., et al. (2022). Structure determination of protein-peptide complexes from NMR chemical shift data using MELD. *bioRxiv*. 2021.12.31.474671. doi:10.1101/2021.12.31.474671
- Morrone, J. A., Perez, A., MacCallum, J., and Dill, K. A. (2017). Computed binding of peptides to proteins with MELD-accelerated molecular dynamics. *J. Chem. Theory Comput.* 13, 870–876. doi:10.1021/acs.jctc.6b00977
- Motmaen, A., Dauparas, J., Baek, M., Abedi, M. H., Baker, D., and Bradley, P. (2022). Peptide binding specificity prediction using fine-tuned protein structure prediction networks. *bioRxiv*. 2022.07.12.499365. doi:10.1101/2022.07.12.499365
- Murcia, M., Jiroukova, M., Li, J., Collier, B. S., and Filizola, M. (2008). Functional and computational studies of the ligand-associated metal binding site of 3 integrins. *Proteins Struct. Funct. Bioinforma.* 71, 1779–1791. doi:10.1002/prot.21859
- Nguyen, H., Roe, D. R., and Simmerling, C. (2013). Improved generalized born solvent model parameters for protein simulations. *J. Chem. theory Comput.* 9, 2020–2034. doi:10.1021/ct3010485
- Otey, C. A., Vasquez, G. B., Burrige, K., and Erickson, B. W. (1993). Mapping of the alpha-actinin binding site within the beta 1 integrin cytoplasmic domain. *J. Biol. Chem.* 268, 21193–21197. doi:10.1016/s0021-9258(19)36909-1
- Perez, J. J., Perez, R. A., and Perez, A. (2021). Computational modeling as a tool to investigate PPI: From drug design to tissue engineering. *Front. Mol. Biosci.* 8, 681617–681637. doi:10.3389/fmolb.2021.681617
- Prytulak, R., Volkmer, M., Meier, M., and Habermann, B. H. (2017). HH-MOTIF: De novo detection of short linear motifs in proteins by hidden markov model comparisons. *Nucleic Acids Res.* 45, 10921–W477. doi:10.1093/nar/gkx810
- Roe, D. R., and Cheatham, T. E. (2013). PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. theory Comput.* 9, 3084–3095. doi:10.1021/ct400341p
- Salomon-Ferrer, R., Götz, A. W., Poole, D., Grand, S. L., and Walker, R. C. (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *J. Chem. theory Comput.* 9, 3878–3888. doi:10.1021/ct400314y
- Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Res.* 33, W382–W388. doi:10.1093/nar/gki387
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710. doi:10.1038/s41586-019-1923-7
- Shirts, M. R., Pitera, J. W., Swope, W. C., and Pande, V. S. (2003). Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.* 119, 5740–5761. doi:10.1063/1.1587119
- Siebert, H.-C., Burg-Roderfeld, M., Eckert, T., Stötz, S., Kirch, U., Diercks, T., et al. (2010). Interaction of the α 2A domain of integrin with small collagen fragments. *Protein Cell* 1, 393–405. doi:10.1007/s13238-010-0038-6
- Slack, R. J., Macdonald, S. J. F., Roper, J. A., Jenkins, R. G., and Hatley, R. J. D. (2022). Emerging therapeutic opportunities for integrin inhibitors. *Nat. Rev. Drug Discov.* 21, 60–78. doi:10.1038/s41573-021-00284-4

- Steinbrecher, T., Joung, I., and Case, D. A. (2011). Soft-core potentials in thermodynamic integration: Comparing one-and two-step transformations. *J. Comput. Chem.* 32, 3253–3263. doi:10.1002/jcc.21909
- Sugita, Y., and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314, 141–151. doi:10.1016/s0009-2614(99)01123-9
- Takada, Y., Ye, X., and Simon, S. (2007). The integrins. *Genome Biol.* 8, 215. doi:10.1186/gb-2007-8-5-215
- Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A., and Schueler-Furman, O. (2022). Harnessing protein folding neural networks for peptide–protein docking. *Nat. Commun.* 13, 176. doi:10.1038/s41467-021-27838-9
- Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., et al. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* 137, 2695–2703. doi:10.1021/ja512751q
- Wayment-Steele, H. K., Ovchinnikov, S., Colwell, L., and Kern, D. (2022). Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. *bioRxiv.* 2022.10.17.512570. doi:10.1101/2022.10.17.512570
- Zhu, J., Zhu, J., Negri, A., Provasi, D., Filizola, M., Collier, B. S., et al. (2010). Closed headpiece of integrin IIb3 and its complex with an IIb3-specific antagonist that does not induce opening. *Blood* 116, 5050–5059. doi:10.1182/blood-2010-04-281154

Frontiers in Chemistry

Explores all fields of chemical science across the periodic table

Advances our understanding of how atoms, ions, and molecules come together and come apart. It explores the role of chemistry in our everyday lives - from electronic devices to health and wellbeing.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

