INTEGRATIVE STRUCTURAL BIOLOGY OF PROTEINS AND MACROMOLECULAR ASSEMBLIES: BRIDGING EXPERIMENTS AND SIMULATIONS

EDITED BY: Paulo Ricardo Batista, Mario Oliveira Neto and David Perahia

PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-610-9
DOI 10.3389/978-2-88976-610-9

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

INTEGRATIVE STRUCTURAL BIOLOGY OF PROTEINS AND MACROMOLECULAR ASSEMBLIES: BRIDGING EXPERIMENTS AND SIMULATIONS

Topic Editors:
Paulo Ricardo Batista, Oswaldo Cruz Foundation (Fiocruz), Brazil
Mario Oliveira Neto, São Paulo State University, Brazil
David Perahia, UMR8113 Laboratoire de biologie et pharmacologie appliquée (LBPA), France

Citation: Batista, P. R., Oliveira Neto, M., Perahia, D., eds. (2022). Integrative Structural Biology of Proteins and Macromolecular Assemblies: Bridging Experiments and Simulations. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-610-9

Table of Contents

frontiers | Frontiers in Molecular Biosciences

EDITORIAL
published: 27 June 2022
doi: 10.3389/fmolb.2022.947582

Editorial: Integrative Structural Biology of Proteins and Macromolecular Assemblies: Bridging Experiments and Simulations

Paulo Ricardo Batista[1]*, *Mario Oliveira Neto*[2] *and David Perahia*[3]

[1]Programa de Computação Científica, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, [2]Departamento de Biofísica e Farmacologia, Instituto de Biociências de Botucatu, Universidade Estadual Paulista Júlio de Mesquita Filho, Botucatu, Brazil, [3]Laboratoire de Biologie et Pharmacologie Appliquée, École Normale Supérieure Paris-Saclay, Centre National de la Recherche Scientifique, Gif-sur-Yvette, France

Keywords: integrative modeling, enhanced sampling, hybrid simulations, protein structure, dynamics, CryoEM, SAXS, AFM

Editorial on the Research Topic

Integrative Structural Biology of Proteins and Macromolecular Assemblies: Bridging Experiments and Simulations

OPEN ACCESS

Edited and reviewed by:
Andrea Mozzarelli,
University of Parma, Italy

****Correspondence:***
Paulo Ricardo Batista
pbatista@fiocruz.br

Specialty section:
This article was submitted to Integrative Physiology, a section of the journal Frontiers in Physiology

Received: *18 May 2022*
Accepted: *10 June 2022*
Published: *27 June 2022*

Citation:
Batista PR, Oliveira Neto M and Perahia D (2022) Editorial: Integrative Structural Biology of Proteins and Macromolecular Assemblies: Bridging Experiments and Simulations. Front. Mol. Biosci. 9:947582. doi: 10.3389/fmolb.2022.947582

Despite continued advances in canonical experimental methods for protein structure determination, they all still possess significant limitations, which have become more evident when applied to challenge systems (Seffernick and Lindert, 2020). For example, this is the case with large transient macromolecular complexes, biomolecular machines, very flexible and intrinsically disordered proteins, or domains (Černý, 2017). Nevertheless, experimental structural information (from various sources and resolutions) can be joined through an integrative modeling approach, resulting in a more accurate model - or set of models. For this, experiments and predictions on how the subunits interact allow obtaining spatial restraints, which help determine the molecular architecture of proteins (Webb et al., 2018).

The success rate of resolving a protein structure is often inversely related to protein size and flexibility. Thus, predictions and measurements of protein dynamics are essential for effectively characterizing structural ensembles (Kuhlman and Bradley, 2019). Hence, hybrid simulations guided by experimental data efficiently sample an ensemble of function-relevant states (Seffernick and Lindert, 2020).

The goal of this Research Topic is to present applications of experimental and computational strategies regarding integrative structural approaches. This collection of nine articles explores experiments of high-speed atomic force microscopy (HS-AFM), small-angle X-ray scattering (SAXS), hydrogen-deuterium exchange mass spectrometry (HDX-MS), crosslinking MS (XL-MS), time-lapse microscopy, circular dichroism (CD), applied with complementary modeling tools, such as molecular docking, normal modes, molecular dynamics (MD), and hybrid enhanced sampling methods.

Soares et al. determined the molecular architecture of the complex formed between native DM64 glycoprotein and myotoxin II from *Bothrops asper* venom using an integrative structural modeling approach. Distance constraints from XL-MS guided the docking of DM64 domains to the myotoxin II crystal structure. SAXS and molecular dynamics simulations indicated that the complex is flexible and structured with an anisotropic shape. In addition, interprotein cross-links and limited hydrolysis analyses revealed inhibitory regions involved in the interaction with toxins.

Molza et al. have developed a complete integrative modeling pipeline that incorporates theoretical knowledge with advanced interactive and immersive modeling tools. As a result, it is possible to extract biologically relevant information by combining automated and human-driven interactive steps.

Proteins are polymers composed of simple units–amino acid residues. However, the dynamic folding process is highly complex (Scheraga et al., 2007). In the review published by Almeida et al. there is a theoretical discussion on the subject, presenting the relative contribution of the hydrophobic effect versus the stabilization of proteins by surface forces, which can sometimes go unnoticed.

Most proteins undergo conformational changes to perform their functions (Orellana, 2019). The article published by Pattanayak et al. monitored conformational changes of an outer membrane efflux protein resulting from the interaction with kanamycin. The bound structure was predicted by molecular docking, and the conformational changes occurred during the MD simulations. CD data suggests that the protein is less flexible and more compact in the presence of kanamycin.

The prediction/determination of relevant conformational states of proteins can be obtained by different methods. Dasgupta et al. modeled multiple conformations of a bacterial ClpB through HS-AFM images. The function of this protein relies on the interconversion between these different conformations. Therefore, even at low resolution, the knowledge of such states can help interpret dynamical-related properties. Still, on the subject, Fagnen et al. studied the dynamical properties of an engineered bacterial channel specifically designed to rest in an open conformation. Using an *in-silico* hybrid method that combines MD simulations and normal modes (Costa et al., 2015), global and local motions were captured and compared to HDX-MS experiments. The simulations also provided an estimation of the probability of the different opening states, agreeing with the electrophysiological experiments.

The major limitation of predicting protein dynamics using MD simulations is to sample rare long-time-scale events. Several enhanced sampling MD methods have been developed to address the problem (Yang et al., 2019). The article of Kaynak et al. presents a systematic comparison of the usefulness and limitations of four normal modes-based hybrid methods to explore the conformational space of proteins. They were applied to four well-studied proteins: triosephosphate isomerase, 3-phosphoglycerate kinase, HIV-1 protease, and HIV-1 reverse transcriptase.

Urazbaev et al. demonstrated that the microtubule growth in living cells could be approximated by a constant velocity with large stochastic fluctuations through time-lapse microscopy and mathematical modeling.

Ahmed et al. described an accurate prediction of residue burial aside from a quantitative prediction of specific residue contributions to protein stability and activity. However, this is a major challenge, especially in the absence of experimental structural information. Therefore, the authors resorted to the yeast surface display of a CcdB bacterial toxin saturation mutagenesis library to investigate the relationship between ligand binding and the level of expression of the displayed protein, with both *in vivo* solubility and *in vitro* thermal stability.

AUTHOR CONTRIBUTIONS

PRB and MON wrote, and DP reviewed the editorial. All authors contributed to the article and approved the submitted version.

REFERENCES

Černý, R. (2017). Crystal Structures from Powder Diffraction: Principles, Difficulties and Progress. *Crystals* 7, 142. doi:10.3390/cryst7050142

Costa, M. G. S., Batista, P. R., Bisch, P. M., and Perahia, D. (2015). Exploring Free Energy Landscapes of Large Conformational Changes: Molecular Dynamics with Excited Normal Modes. *J. Chem. Theory Comput.* 11, 2755–2767. doi:10.1021/acs.jctc.5b00003

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

Orellana, L. (2019). Large-Scale Conformational Changes and Protein Function: Breaking the In Silico Barrier. *Front. Mol. Biosci.* 6, 117. doi:10.3389/fmolb.2019.00117

Scheraga, H. A., Khalili, M., and Liwo, A. (2007). Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annu. Rev. Phys. Chem.* 58, 57–83. doi:10.1146/annurev.physchem.58.032806.104614

Seffernick, J. T., and Lindert, S. (2020). Hybrid Methods for Combined Experimental and Computational Determination of Protein Structure. *J. Chem. Phys.* 153, 240901. doi:10.1063/5.0026025

Webb, B., Viswanath, S., Bonomi, M., Pellarin, R., Greenberg, C. H., Saltzberg, D., et al. (2018). Integrative Structure Modeling with the Integrative Modeling Platform. *Protein Sci.* 27, 245–258. doi:10.1002/pro.3311

Yang, Y. I., Shao, Q., Zhang, J., Yang, L., and Gao, Y. Q. (2019). Enhanced Sampling in Molecular Dynamics. *J. Chem. Phys.* 151, 070902. doi:10.1063/1.5109531

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.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 15 April 2021
doi: 10.3389/fmolb.2021.636286

Kanamycin-Mediated Conformational Dynamics of *Escherichia coli* Outer Membrane Protein TolC

Biraja S. Pattanayak*[1], *Budheswar Dehury*[2], *Mamali Priyadarshinee*[3], *Suman Jha*[4], *Tushar K. Beuria*[5], *Dhananjay Soren*[1] and *Bairagi C. Mallick*[6]

[1]Department of Zoology, Ravenshaw University, Cuttack, India, [2]Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, [3]School of Life Sciences, Ravenshaw University, Cuttack, India, [4]Department of Life Sciences, National Institute of Technology, Rourkela, India, [5]Institute of Life Sciences, Bhubaneswar, India, [6]Department of Chemistry, Ravenshaw University, Cuttack, India

TolC is a member of the outer membrane efflux proteins (OEPs) family and acts as an exit duct to export proteins, antibiotics, and substrate molecules across the *Escherichia coli* cell membrane. Export of these molecules is evidenced to be brought about through the reversible interactions and binding of substrate-specific drug molecules or antibiotics with TolC and by being open for transport, which afterward leads to cross-resistance. Hence, the binding of kanamycin with TolC was monitored through molecular docking (MD), the structural fluctuations and conformational changes to the atomic level. The results were further supported from the steady-state fluorescence binding and isothermal titration calorimetry (ITC) studies. Binding of kanamycin with TolC resulted in a concentration dependent fluorescence intensity quenching with 7 nm blue shift. ITC binding data maintains a single binding site endothermic energetic curve with binding parameters indicating an entropy driven binding process. The confirmational changes resulting from this binding were monitored by a circular dichroism (CD) study, and the results showed insignificant changes in the α-helix and β-sheets secondary structure contents, but the tertiary structure shows inclusive changes in the presence of kanamycin. The experimental data substaintially correlates the RMSD, $R_{g,}$ and RMSF results. The resulting conformational changes of the TolC-kanamycin complexation was stabilized through H-bonding and other interactions.

Keywords: *Escherichia coli*, efflux proteins, TolC, kanamycin, molecular docking, molecular dynamic simulation

OPEN ACCESS

Edited by:
Mario Oliveira Neto,
São Paulo State University, Brazil

Reviewed by:
Elena B Tikhonova,
Texas Tech University Health Sciences Center, United States
Gulam Rabbani,
Yeungnam University, South Korea

****Correspondence:***
Bairagi C. Mallick
bcmallick@ravenshawuniversity.ac.in

Received: *01 December 2020*
Accepted: *17 February 2021*
Published: *15 April 2021*

Citation:
Pattanayak BS, Dehury B, Priyadarshinee M, Jha S, Beuria TK, Soren D and Mallick BC (2021) Kanamycin-Mediated Conformational Dynamics of Escherichia coli *Outer Membrane Protein TolC. Front. Mol. Biosci. 8:636286. doi: 10.3389/fmolb.2021.636286*

INTRODUCTION

The Gram-negative bacteria *Escherichia coli* (*E.coli*) is a well-known human pathogen that causes urinary tract infections (UTIs), bacteremia, neonatal meningitis, and serious food-borne infections worldwide (Money et al., 2010; Pennington, 2014). The treatment of *E. coli* infection depends on its early diagnosis and scheduled administration of antibiotics. The extensive, prolonged, and unethical use of antibiotics has meant antibacterial resistance has been developed in *E. coli* strains (Cagnacci et al., 2008; Michael et al., 2014). The reasons for the emergence of antibiotic resistance has been evidenced to be associated with the overexpression of efflux pumps (Wang et al., 2001). These efflux pumps are classified into five different families: the ATP-binding cassette (ABC) superfamily, the major facilitator superfamily (MFS), the multidrug and toxic compound extrusion (MATE) family, the small multidrug resistance (SMR) family and the resistance nodulation division (RND) family (Poole, 2007; Li and Nikaido, 2009; Delmar et al., 2014). The RND family proteins, functioning as a tripartite system in *E. coli*, are composed of an inner membrane protein (IMP), a periplasmic adaptor membrane fused protein (MFP) family, and an outer membrane factor (OMF) family protein (Nikaido and Zgurskaya, 2001). The outer membrane protein (OMP) TolC works with a combination of other RND, ABC, and MFS efflux pumps to excrude drugs molecues across the lipid bilayer. Any irregularities or the absence of any of these components makes the triparty system non-functional.

In *E. coli*, the outer membrane protein TolC has long been known as a transporter of diverse small molecules and antibacterial drugs to large protein toxins (Morona and Reeves, 1981; Morona and Reeves, 1982; Morona et al., 1983). The TolC protein belongs to the outer membrane efflux proteins (OEP) family or outer membrane factor (OMF) family. The members of this family functions in conjunction with three-types of transport systems; ATP-binding cassette (ABC-type), resistance nodulation division (RND-type), and major facilitator super family (MFS-type) (Paulsen et al., 1997). The association between transporters and outer membrane efflux protein is mediated by periplasmic proteins, named membrane fusion proteins (MFPs) (Dinh et al., 1994; Zgurskaya, 2009). The inner membrane transporter AcrB and the periplasmic membrane fused protein AcrA together forms the translocase complex as AcrAB and recognizes specific substrates for efflux (Tikhonova et al., 2011) by recruiting TolC. TolC spans in the outer membrane and projects deep into the cell periplasm to export substrates out of the cell with the help of membrane-fused translocase complex and AcrAB proteins that forms the AcrAB-TolC efflux system (Fralick et al., 1996; Zgurskaya and Nikaido, 1999; Du et al., 2015). The AcrAB-TolC complex is a $\Delta\mu H^+$-dependent pump responsible for resistance to many compounds (Paulsen et al., 1996). Studies have shown that the interaction of these proteins to form the multi-protein efflux channel to export antibiotics is mediated through the conformational dynamic (Zhao et al., 2017). The recent crystal structure of TolC has confirmed the trimeric assembly of polypeptide chains with an entrance just like a cannon and a long axis of 140 Å (Koronakis et al., 2000). Each monomer is composed of 471 amino acids and 12-beta sheets (four from each monomer) which together constitute the portion of the tunnel, which is located in the bacterial outer membrane. Similarly, 12-alpha helices forming the body part of the tunnel reside in the bacterial periplasmic space and a mixed α/β-barrel as an equatorial domain. The opening and closing of the coiled-coil domain through the interactions with MFPs and IMPs initiates the efflux process. TolC is recruited to form the efflux system when the translocate complex is bound to export substrates (Sulavik et al., 2001; Koronakis et al., 2004; Li and Nikaido, 2009). The AcrAB-TolC triparty system creates a direct efflux pathway from the cytoplasm to the extracellular space and allows the efflux of substrate molecules across the cell membrane. This efflux pump has been found to contribute to the inherent resistance for ß-lactams, fluoroquinolones, aminoglycosides (Nidaido, 1994; Wang et al., 2001; Nishino et al., 2003), and other toxic compounds such as antiseptics, detergents, and dyes (Sulavik et al., 2001). Overexpression of these efflux-pump proteins have resulted in the development of antibiotic resistance bacteria (Nikaido, 1998; Li and Nikaido, 2009).

Aminoglycosides are broad-spectrum bactericide antibiotics used for the treatment of short-term infections caused by infectious bacteria such as *E. coli, Proteus* spp., *Enterobactor areogenes, Klabsiella pneumoniae, Serratia marcescens,* and *Acinetobacter* spp (Ristuccia and Cunha, 1985; Aggen et al., 2010; Landman et al., 2010). Kanamycin is a promising aminoglycoside antibiotic that has been extensively used in the treatment of *E. coli* infections worldwide. It acts on bacteria by inhibiting protein synthesis, which is essential for its survival (Zhang et al., 2015). The extensive use of kanamycin for *E. coli* treatment has developed its resistance to a greater extent and the same is considered to be brought about through the involvement of outer membrane efflux pump proteins. However, the exact mechanism of efflux and the involvement of these proteins still remains largely unknown (CarolBaker et al., 1974). It is also known that TolC-dependent efflux of antibiotics and other molecules involves the interaction of TolC with the translocase/efflux pump of two inner membrane proteins, AcrAB (Aono et al., 1998; Fralick,1996). Despite these encouraging results, the binding of antibiotics to the triparty system and their active efflux remains incomplete and needs comprehensive analysis. Hence, we aimed to explore the binding of kanamycin with TolC by molecular docking and biophysical methods, such as fluorescence spectroscopy, isothermal titration calorimetry (ITC), and conformational stability by circular dichroism (CD) studies. The conformational changes resulting from TolC-kanamycin interaction binding was established through the molecular dynamic simulation study.

MATERIALS AND METHODS

Materials

Oligonucleotides were purchased from IDT, Singapore and the DNA polymerase, dNTPs (dATP, dGTP, dTTP, dCTP) Phusion buffer, restriction endonuclease (Sac I and Hind III), and T4 DNA ligase were purchased from NEB, USA. *Pfu* polymerase was purchased from Promega, USA. L-arabinose, protease inhibitor, lysozyme, Octyl β-D-glucopyranoside, nitrocellulose membrane, and primary antibody (anti-mouse monoclonal antibody) were purchased from Sigma-Aldrich, USA. Ni-NTA resin was procured from Qiagen. Ampicillin and kanamycin were purchased from HiMedia. Tris and phosphate buffer were purchased from MP Biomedical, LLC (France).

Expression and Purification of TolC

TolC was effectively overexpressed in *E. coli* BL21 (DE3) competent cells and purified using a Ni-NTA affinity column. The purified recombinant TolC protein was re-purified using FPLC (Akta, Ge-Healthcare) to remove impurities (**Supplementary Figure S1**). BL21 (DE3) cells harboring the plasmid pBAD-*tolC* construct were allowed to grow in an orbital shaker, and the protein expression was induced with 0.2% (w/v) of L-arabinose at 37°C for 4 h. The cell pellet obtained was resuspended in lysis buffer (20 mM sodium phosphate buffer, 150 mM NaCl, 10 μl protease inhibitor solution and 1 mg/ml Lysozyme, pH 7.5) and ruptured by passing in a French press. The unruptured cells were separated through centrifugation at 9,000 g for 30 min at 4°C, and then the membrane fraction was separated and collected at 120,000 g by using an ultracentrifuge. The pellet obtained was again resuspended in 20 ml of binding buffer (5% Triton-X-100 in Lysis buffer, pH 7.5). The solution with membrane fraction was sonicated for 20 min on ice and then centrifuged at 20,000 g for an hour at 4°C. The supernatant was mixed with Ni-NTA resin and incubated for 1 h at 4°C. Before the incubation, the Ni-NTA resin was equilibrated with 10 column volumes of binding buffer containing 2% Triton-X-100. Then the column was thoroughly washed with 10 column volume of washing buffer (20 mM sodium phosphate buffer pH 7.5, 150 mM NaCl, 20 mM Imidazole and 2% Triton-X-100). The elution was completed with elution buffer (20 mM sodium phosphate buffer pH 7.5, 150 mM NaCl, 300 mM imidazole and 2% of Triton-X-100) at a flow rate of 1 ml/min. Finally, the elute was passed through a Superdex-200 gel-filtration column pre-equilibrated with the running buffer (20 mM sodium phosphate buffer at pH 7.5, 150 mM NaCl with 1% Octyl β-D-glucopyranoside). The fractions collected were concentrated using Amicon ultrafiltration membrane and quantified using UV-visible spectrophotometer. The molecular size of the TolC protein was confirmed through the MALDI mass spectroscopy analysis (**Supplementary Table S1**).

Fluorescence Study

Fluorescence quenching experiments of TolC with kanamycin complexation was carried out using a Carry Eclipse spectrofluorometer (Agilent Technologies) equipped with an external Peltier temperature controller (EC-50) to maintain the cell temperature at 25 ± 0.1°C. Prior to the stock solution preparation, 20 mM Tris-HCl buffer pH 7.2 and 1% Octyl β-D-glucopyranoside detergent were filter-sterilized and degassed to avoid scattering and the interference of dissolved oxygen. The emission spectra of TolC and kanamycin binding experiments were recorded with successive additions of increasing concentrations of kanamycin from 5 mM to 50 mM, and the spectra were recorded between 300–450 nm by keeping the excitation wavelength fixed at 280 nm. Each protein and kanamycin scan was subtracted with corresponding kanamycin concentration in buffer to avoid the dilution effect. A constant protein concentration of 5 μM was maintained throughout all the experiments to avoid data irregularities. The binding experiments were carried out with the same excitation and emission slit widths (5 nm) in an auto response time mode. Each scan was an average of five accumulations with least smoothening.

Isothermal Titration Calorimeter Measurements

Binding of kanamycin with TolC was measured by using MicroCal PEAQ-ITC (Malvern Panalytical, United States) equipped with a temperature-controlled cell of volume 200 μl and a 40 μl micro syringe. The stock protein sample was prepared by dialyzing against 20 mM Tris-HCl buffer pH 7.2 added with 1% Octyl β-D-glucopyranoside detergent and stored at 4°C for further use. Before experiments, each solution was thoroughly degassed under vacuum (140 mbar for 10 min) to remove the dissolved air bubbles. The sample cell was loaded with 1 ml of 20 μM TolC solution and the calorimetry syringe with 1.2 mM of kanamycin in the same buffer to avoid baseline error. Both the sample and syringe solutions were added with 1% Octyl β-D-glucopyranoside detergent to avoid the dilution effects and prevent signal instability during measurements. The temperature of the sample cell and syringe for each experiment was isothermally maintained at 25°C and experiments were programmed for 20 injections with 2 μL per injection each at 300 s interval to obtain a saturated binding curve. The heat of dilution experiments of kanamycin and buffer added with 1% Octyl β-D-glucopyranoside detergent were performed in the same conditions and subtracted from the protein binding titration data (**Supplementary Figure S2**). Each binding isotherm was performed in triplicate and the data obtained was analyzed with MicroCal PEAQ-ITC analysis software using a single-site-binding model to determine the thermodynamic parameters viz., stoichiometry of binding (N), binding constant (K_a), enthalpy change (ΔH^o), and entropy change ($T\Delta S^o$) at temperature T. The change in Gibb's free energy (ΔG^o) was calculated using the Gibb's-Helmholtz equation:

$$\Delta G^o = \Delta H^o - T\Delta S^o$$

Circular Dichroism Measurements

The far UV-CD spectra of TolC and kanamycin binding interaction were recorded in a Jasco-815 spectropolarimer (Jasco, Japan) using a quartz cell of 0.1 cm path length at 298.15 K. All the CD scans were performed in 20 mM Tris-HCl buffer at pH 7.2, added with 1% Octyl β-D-glucopyranoside. The blank scans of kanamycin in buffer without protein were subtracted from the protein and kanamycin in buffer scans to obtain the final spectra. For all conditions, TolC concentration was kept constant at 10 μM and the kanamycin concentration was varied from 10 to 30 μM. CD spectra were collected with a scanned speed of 50 nm/min with a fixed bandwidth of 1 nm in the range of 190–250 nm. Nitrogen gas with 99% purity was purged through the sample compartment to create an oxygen free environment. The results obtained were expressed as the mean molar ellipticity (θ):

$$[\theta] = 100 \times [\theta_{\mathrm{obsd}}/lc]$$

Where θ_{obsd} is the observed ellipticity in degree, c is the concentration of the residues in mol/cm^3, and l is the path length of the cell (Johnson, 1990).

Molecular Docking

The sequence information and atomic coordinates of TolC protein in SDF format was obtained from Protein Data Base (PDB) of NCBI and Protein Data Bank (PDB ID:1EK9). The structural information of kanamycin was collected from

PubChem database (NCBI) and Drug Bank in SDF format. Both the SDF format of TolC and kanamycin were converted to PDB format for docking purposes. Flexible docking of both the molecules were carried out through Schrodinger software (Friesner et al., 2004; Dinesh et al., 2010; Ahmad et al., 2011), and the different parameters, like bond orders, the addition of hydrogen, proper ionization state of residues, capping and termini, and so forth, were taken into consideration for the preparation of receptors. Then the refining of receptors with the H-bond assignment (water orientation, at neutral pH) and minimization of energy with OPLS 2005 force field was performed. Generation of the grid for the protein was conducted using the site surrounding the selected residues' centroids. The OPLS 2005 force filed parameters, ionization at pH 7.0 ± 2.0, stereoisomers, and generated tautomer were utilized to prepare the ligands in LigPrep (Abdullah et al., 2016). The flexible nitrogen inversion and ring conformation ligands were then docked by the Extra Precision [EP] method.

Molecular Dynamic Simulation

All simulations were performed using GROMACS v2018.4 with the CHARMM36m force field in TIP3 water model. There was a total of 238 POPC for the top (head) and bottom (tail) bilayer with a water thickness of 17.5 Å (36,969 water molecules) from the top and bottom of the lipid head group. The systems were neutralized by adding counter ions to each of the systems. These were 115 positive sodium and 101 negative chloride ions at a 0.15 M concentration. Then the system underwent 50,000 steps of steepest descent energy minimization to remove close van der Waals force contacts. Afterward, the system was subjected to a two-step equilibration phase, namely NVT (constant number of particles, Volume, and Temperature) for 1,000 ps to stabilize the temperature of the system and a short position restraint NPT (constant number of particles, Pressure, and Temperature) for 1,000 ps to stabilize the pressure of the system by relaxing the system and keeping the protein restrained. The V-rescale temperature-coupling method was used for the NVT ensemble, with constant coupling of 0.1 ps at 303.15 K under a random sampling seed. The temperature was maintained at 303.15 K using a Nosé-Hoover thermostat with a coupling time constant of 1.0 ps. For NPT, Parrinello-Rahman pressure coupling was turned on with constant coupling of 0.1 ps at 303.15 K under conditions of position restraints (all H-bonds). Electrostatic forces were calculated for both NVT and NPT using Particle Mesh Ewald method. After equilibration, the simulation was carried out for 50 ns under the NPT ensemble without any position restraints.

All trajectory analyses were performed using the analysis tools in GROMACS package. Intrinsic dynamics stability parameters, including root-mean-square deviation (RMSD), solvent accessible surface area (SASA), radius of gyration (Rg), and root-mean-square fluctuation (RMSF) calculations, were computed using GROMACS analysis tools. Principal component analysis (PCA) or essential dynamics (ED) were performed to understand the global motion of TolC in the presence of Kanamycin in a lipid bilayer considering the main-chain atoms. Using the MD trajectory, the translational and rotational movements were removed from the complex system using *gmx cover* toolkit to construct a covariance matrix. Then, eigenvectors and eigenvalues were calculated by diagonalizing the covariance matrix. The eigenvectors that correspond to the largest eigenvalues are called "principal components (PCs)" or "eigenvectors (EVs)," as they represent the largest-amplitude collective motions. In this study we considered the first two PCs for further exploration as they represent ~89.3% of the total motion of the complex. Clustering analysis was performed using the GROMACS tool *gmx cluster* to identify different conformational states of the complex. Further, structural superposition of the docked conformation and the representative of top ranked cluster were performed to see the variations in the ligand binding.

TABLE 1 | List of six antibiotics and their Docking scores, XP G score, and Glide G score (kcal/mol) with TolC.

Sl. No	Antibiotics	Docking score	Xp G score	Glide G score
1	Kanamycin	−6.32	−6.53	−6.53
2	Tetracycline	−5.05	−5.44	−5.44
3	Erythromycin	−4.64	−4.65	−4.65
4	Chloramphenicol	−2.88	−2.88	−2.88
5	Norfloxacin	−2.70	−2.83	−2.83
6	Rifampicin	5.92	3.63	3.63

RESULTS

Screening of Antibiotics

Antibiotics like kanamycin, tetracycline, erythromycin, chloramphenicol, norfloxacin, and rifampicin were screened for their feasible interaction with TolC using a molecular docking method to correlate their involvement in the efflux process that causes *E. coli* resistance. Molecular docking has provided valuable information and has helped to evaluate the binding patterns of six antibiotics and calculate their binding affinities toward residues in the active site of TolC. The comparative docking score and binding energies of six antibiotics estimated from the docking results of TolC are listed in **Table 1**.

It clearly indicates that among all chosen antibiotics, kanamycin showed a comparative strong binding affinity toward TolC with an estimated high value of free energy change, i.e., $\Delta G = -6.5$ kcal/mol. Based on the screening results, kanamycin was chosen as the antibiotics of interest to understand its interaction with TolC and correlate the process with the efflux mechanism that causes *E. coli* resistance.

Purification of Recombinant TolC

The recombinant his-tagged TolC was purified using a modified method, described previously (Koronakis et al., 1997; Koronakis et al., 2000). Purified protein concentration was determined using the DC protein assay (Bio-Rad) with bovine serum albumin as standard and the purity was checked by 12% SDS-PAGE (**Figure 1**), which showed a band at ~ 54 kDa. The protein was then further purified using a FPLC, Akta, GE Healthcare

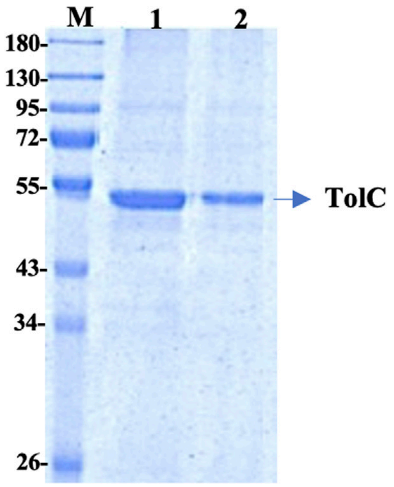


FIGURE 1 | 12% SDS PAGE of the purified TolC protein showing the molecular mass of ~54 kDa. First Lane; Marker, Lane-1 and Lane-2 are purified recombinant TolC protein.

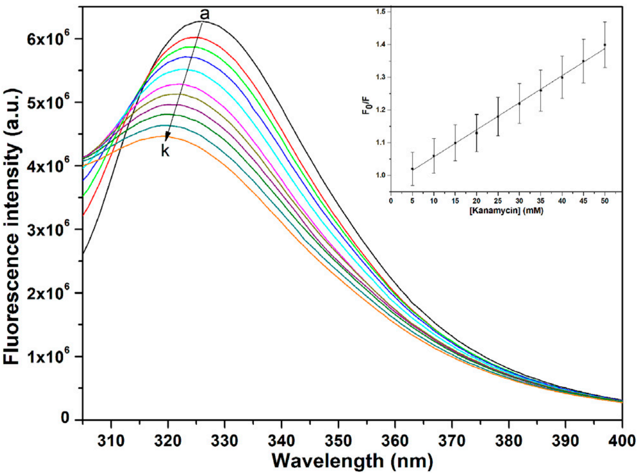


FIGURE 2 | Fluorescence spectra of TolC in the presence of increasing concentrations of kanamycin monitored at 25°C. Protein in buffer as control **(A)**, and with increasing kanamycin concentrations: 5 mM **(B)**; 10 mM **(C)**; 15 mM **(D)**; 20 mM **(E)**; 25 mM **(F)**; 30 mM **(G)**; 35 mM **(H)**; 40 mM **(I)**; 45 mM **(J)**; and 50 mM **(K)** respectively. Protein concentration was kept constant at 5 μM and the experiments were performed in triplicate to obtain the best spectra. The samples were excited at 280 nm and the emission spectra were recorded in the range of 300–400 nm.

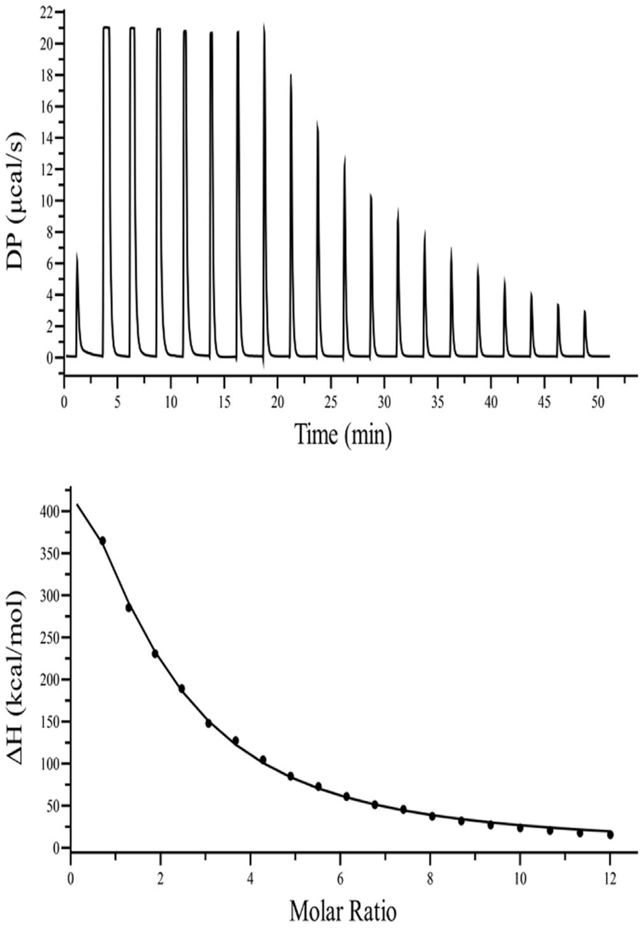


FIGURE 3 | ITC binding isotherm of TolC with kanamycin at 25°C. The upper panel indicates the raw data points for heat produced with time for each titration of 1.2 mM of kanamycin with 20 μM TolC at 25°C. The lower panel shows the binding isotherm obtained after subtracting the heat of dilution with integration of peak areas and normalization to produce a plot of molecular heat change against the kanamycin/TolC molar ratio. The fitting line in the curve is shown using a black solid line.

to eliminate impurities. The accurate protein molecular mass was determined from the MALDI mass spectroscopy (**Supplementary Table S1**).

Fluorescence Measurements

The binding affinity parameters of kanamycin with TolC was calculated using fluorescence measurements. Molecular interactions, such as formation of excited state charge-transfer complex, intersystem crossing, molecular rearrangement, and ground-state complexation, between the fluorophore and quencher can lead to fluorescence quenching (Lakowicz, 2006). TolC contains 1-tryptophan (Trp), 21-tyrosine (Try) and 10-phenylalanine (Phe) residues. All these residues can act as fluorophore to exploit the conformational changes in the neighbourhood environment of ligand binding sites in the TolC through fluorescence quenching experiments (Rabbani et al., 2011; Rabbani et al., 2013). However, as the fluorescence of tryptophan residue dominates over all others, we exclusively excite it at 295 nm to monitor its quenching effects on kanamycin binding (Ahmad et al., 2012). In **Figure 2**, the fluorescence emission spectra of TolC shows some change in shape with increasing kanamycin concentrations. Binding titration curves of kanamycin with TolC indicates significant decreases in the tryptophan fluorescence intensities with a blue shift of 7 nm. The inset shows the linear dependence of kanamycin on the quenching of TolC fluorescence intensities. The resulted

TABLE 2 | Thermodynamic parameters obtained from the ITC binding of kanamycin with TolC measured at 298.15 K. The values are means of triplicates ±SD.

N	K_a (M^{-1})	ΔH (kcal/mol)	$-T\Delta S$ (kcal/mol)	ΔG (kcal/mol)
1.46	2.34 ± 0.06 × 10^4	80.0 ± 0.9	−88.3 ± 0.8	−8.32 ± 0.56

spectra were obtained by subtracting the corresponding blanks, and the data obtained was analyzed using Stern-Volmer equation by plotting the intrinsic fluorescence intensities at λ_{max} against the [kanamycin].

$$\text{Stern} - \text{Volmer equation} : \frac{F_o}{F} = 1 + K_{sv}\ [Q]$$

Here, F_o and F are the fluorescence intensities in the absence and presence of kanamycin. K_{sv} is the Stern-Volmer quenching constant and [Q] represents the molar concentration of quencher [kanamycin]. At 298.15 K, the value of $K_{sv} = 6.07 \pm 0.05 \times 10^4\ M^{-1}$ with R-value of 0.989 justifies an effective quenching process of TolC protein fluorescence intensities by kanamycin binding.

ITC Measurements

Fluorescence quenching studies indicate that kanamycin has a high binding affinity toward TolC. And to support this binding interaction result, we performed ITC binding titration of TolC with kanamycin and calculated the thermodynamic parameters as binding constant (K_a), enthalpy change (ΔH), entropy change (ΔS), binding stoichiometry (N), and free energy change (ΔG). **Figure 3** shows the ITC isotherm produced from the titration of kanamycin with TolC at 298.15 K. The upper panel of binding isotherm with positive heat change indicates that an endothermic thermogram produced unfavourable binding standard enthalpy of 80.0 ± 0.9 kcal/mol and calculated favourable value of binding entropy change at T (K) is −88.3 ± 0.8 kcal/mol. The lower panel shows the amount of heat released with each successive injection as a function of molar ratio of TolC and kanamycin. The heat change on dilution of kanamycin into Tris-buffer was measured in the same conditions and subtracted from the heat changes by the titration of kanamycin with TolC. The overall heat changes due to the interaction of kanamycin with TolC was plotted against the molar ratio of TolC and kanamycin. The raw data obtained for kanamycin and TolC interaction was fitted to the binding models using the inbuilt origin software. The negative values of Gibb's free energy indicates the spontaneous formation of the TolC-kanamycin complex (Rahman et al., 2019). The best fitted ITC isotherm shows a single binding site model and the thermodynamic parameters obtained are listed below in **Table 2**.

The calculated positive enthalpy change indicates an endothermic reaction counterbalanced with a high value of positive entropy change, making the spontaneous binding of kanamycin with TolC an entropy driven process.

CD Measurements

Circular dichroism (CD) is a relatively easy and highly applicable spectroscopic technique to study protein drug binding and the conformational changes (Rabbani et al., 2012; Varshney et al., 2010). It is also useful to understand the mode of interaction and determine the change in protein secondary structure (Rabbani and Choi, 2018; Rabbani et al., 2015). The intermolecular and intramolecular interactions involved in the protein secondary structure stability are affected upon binding with ligands or drug molecules (Varshney et al., 2014; Thakur et al., 2017). CD spectra is established with one positive spectra on 195 nm and two significant negative peaks centered on 209 nm and 222 nm in the far-UV region due to the n → π* transition that shifts the peptide bond, which is characteristic of α-helix structure (Varlan and Hillebrand, 2010; Rogozea, 2012; Suryawashi, 2016; Zhang et at., 2016). To corroborate the fluorescence and ITC studies for kanamycin binding to TolC protein, CD measurements of TolC protein were carried out to observed the conformational behavior in the presence and absence of kanamycin. For all the experimental conditions, the protein concentration was kept constant at 10 μM. The change in ellipticity in the far-UV region from 200–240 nm provides information about the change in the regular secondary structure contents, such as α-helix and β-sheets of TolC, in the presence of kanamycin, which was used for deconvolution of regular secondary structure using the K2D2 online server (Perez-Iratxeta and Andrade-Navarro, 2008) and presented in **Table 3**. In **Figure 4A**, the binding of kanamycin with TolC indicates insignificant change in the secondary structure contents with a near decrease in the molar ellipticity in the range of 30 μM. However, the near-UV CD spectra in **Figure 4B** shows the change in the micro environment of the tryptophan residues around the binding site. The CD data indicates less flexible secondary structures with a more compact and stabilized complex in the presence of 30 μM kanamycin.

Molecular Docking of Kanamycin with TolC

Hence, to analyze the interaction of kanamycin with TolC, we performed molecular docking (MD) to understand the nature and the mechanism of interaction at the molecular level. We identified that kanamycin interacts with TolC through a set of functionally active residues in the active site of the protein. The surface potential view in **Figure 5A** and the binding pocket view in **Figure 5B** illustrates the involvement of GLU16, LYS19, SER20, ASP23, THR97, ASP101, ASN108, and GLN189 amino acids in the active site of TolC, which interact with kanamycin; the residual atom involved and their hydrogen bonding lengths are presented in **Table 4**.

During the MD, it was observed that the complex was stabilized by 12-hydrogen bonds, which involve five amino acid residues in the interactions. The hydrogen bonds are

TABLE 3 | Estimation of secondary structure content of TolC on addition of kanamycin.

Protein and antibiotics	α-Helix (%)	β-Sheets (%)
TolC	43.29	10.79
TolC +10 μM kanamycin	45.81	10.79
TolC +20 μM kanamycin	43.01	10.79
TolC +30 μM kanamycin	40.01	10.08

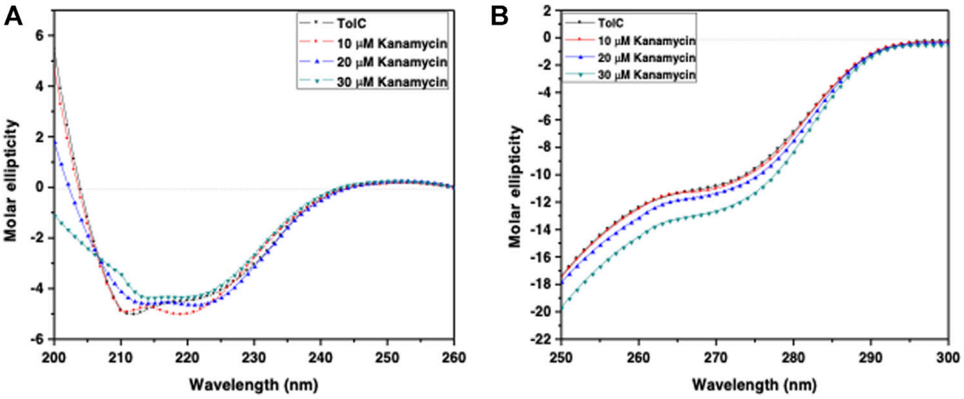


FIGURE 4 | CD spectra of the purified outer membrane protein TolC without and with different concentrations of kanamycin measured in 20 mM phosphate buffer, added with 1% Octyl β-D-glucopyranoside, pH 7.2 at 25°C **(A)** Far UV-CD spectra of Tol C protein (■), and in the presence of increasing concentrations of kanamycin 10 μM (●), 20 μM (▲) and 30 μM (▼) **(B)** Near UV-CD spectra of TolC protein (■), and in the presence of increasing concentrations of kanamycin 10 μM (●), 20 μM (▲) and 30 μM (▼) respectively.

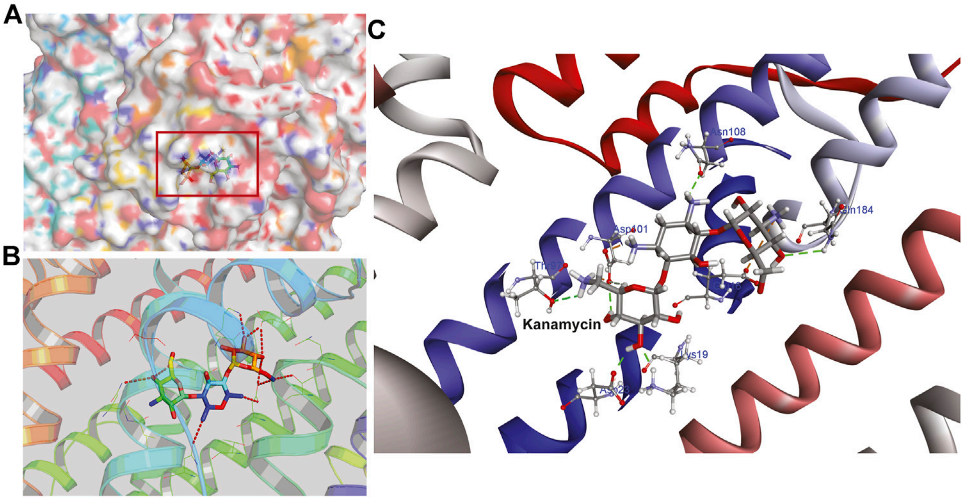


FIGURE 5 | The structural representation of TolC residues interacting with kanamycin **(A)** surface potential view of TolC binding pocket (marked in square) occupied by kanamycin **(B)** ribbon diagram of TolC showing the interaction of docked kanamycin in the active site, and **(C)** amino acids residues with different positions (shown in gray sticks) involved in the binding with kanamycin (shown in gray and red sticks).

strong enough to keep hold tightly of kanamycin inside the active pocket of the protein. Similar salt bridge interactions were retained in the post docked structures and demonstrated that protein and kanamycin have high affinity and interact strongly. In addition, the system was also stabilized by non-polar interactions, such as pi-pi interactions, which were formed by residues with kanamycin [**Figures 6A,B**]. Interestingly, as compared to the docked conformation, the top ranked cluster representative complex showed loss of hydrogen bonding in the simulated complex which is due the structural reorientation of ligand and few residues of the protein in the binding site, thereby affecting the binding process.

System Preparation

The best ranked pose i.e., 1EK9-kanamycin complex structure obtained from the docking simulations was subjected to 50 ns MD in lipid bilayer (**Figure 7**). The orientation of the 1EK9 structure with respect to the membrane was determined using the Positioning of Proteins in Membrane [PPM] server. The membrane-oriented protein-ligand was then inserted in the 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine [POPC] lipid bilayer using the CHARMM-GUI membrane builder. The protein-membrane system was solvated with TIP3P water model and 0.15 M NaCl and the ligand force field parameters were obtained using ParamChem with CHARMM General Force Field [CGenFF].

TABLE 4 | The amino acid residues in the active site of TolC that are involved in the interaction with kanamycin.

Sl. No	Residue involved	Residues atom involved	H-bond length (Å)
1	GLU16	O	2.9
2	GLU16	O	2.7
3	LYS19	O	2.8
4	SER20	O	3.5
5	ASP23	O	2.7
6	ASP23	O	3.6
7	THR97	N	3.1
8	ASP101	O	3.2
9	ASP101	N	2.8
10	ASP101	N	2.8
11	ASN108	N	3.0
12	GLN184	O	3.5

Structural Changes in TolC on Kanamycin Binding

The structural motion and internal fluctuation of kanamycin bound form of TolC was analyzed through MD trajectory by essential dynamics. Initially, a diagonal covariance matrix from the main-chain of all atoms of the protein that captures the strenuous motion of the atom through eigenvectors and eigen values were inferred. The eigen values present the atomic contribution of motion (dynamics behavior and degree of fluctuations) while eigenvectors describe the overall direction of motion of the atoms. The first two eigenvectors (EV1 and EV2) obtained from ED analysis capture more than ~74% of the total motion, indicating that these vectors define the essential subspace of the system [**Figures 8A,B**]. The trace value for the complex was computed to be 52.11 nm^2. To quantitatively understand the movement directions captured by the eigenvectors, a porcupine plot was generated using the extreme projections on principal component PC1 and PC2. The direction of the arrow represents the direction of motion, while the length of the arrow characterizes the movement strength. The obtained plot suggests that rotational concerted movements are observed in two EVs, where the periplasmic and the extracellular ends of the protein display a high degree of inward movement, i.e. toward the pore.

To understand the structural diversity of the ensembles during MD, GROMOS method was employed to perform the clustering with a 0.2 nm cut off. The RMSD-based clustering generated a total of three clusters where the top ranked cluster harbors 2,293 structures (91.72% with an average RMSD of 0.175 nm), the second one occupies 199 structures (~8% with RMSD of 0.148 nm), and the third has the least 9 structures with an average RMSD value of 0.157 nm.

Dynamics of TolC-Kanamycin Complex

The binding of small ligands or drugs to a protein channel is a microscopic event that happens within fractions of a second. So, understanding molecular interactions and energetics of this binding is difficult to accomplish in detail. However, such complex processes can be investigated through molecular dynamics (MD) simulation studies (Dehury et al., 2017; Dehury et al., 2019), and the calculated root-mean-square deviation (rmsd) values can be used to estimate the structural

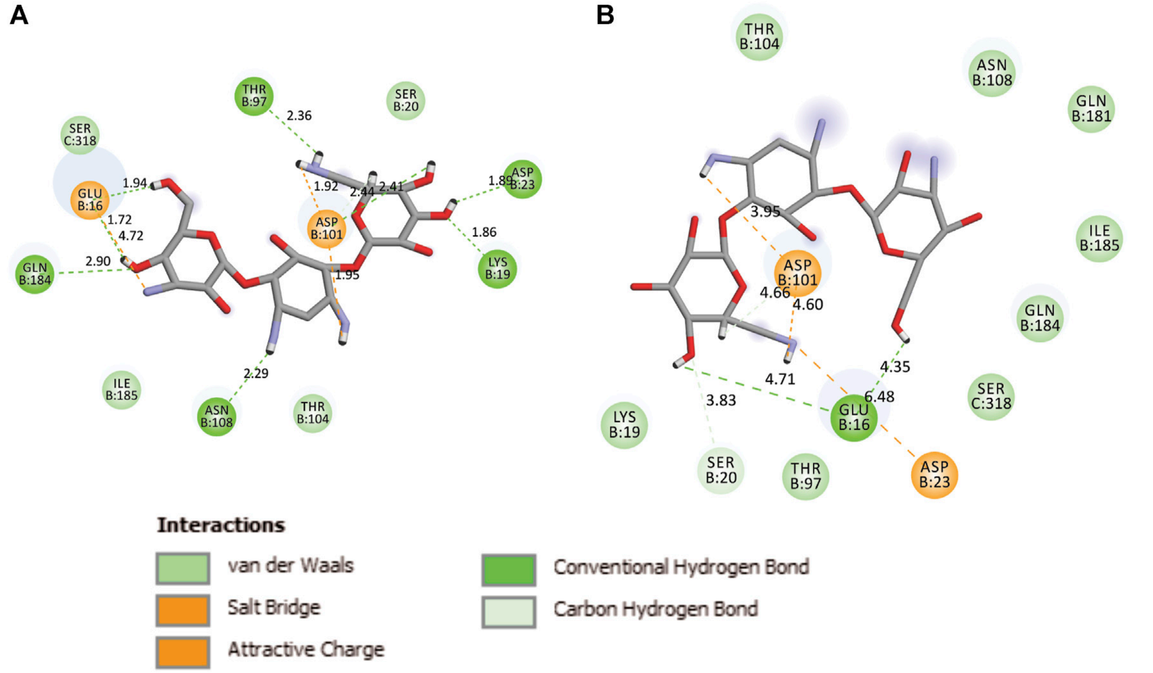


FIGURE 6 | Two-dimensional representation of TolC-kanamycin interaction complex analysis. **(A)** TolC docked with kanamycin and stabilized through different interactions, and **(B)** the stabilizing interaction of residues through H-bonding (green color) and salt-bridging (orange color) in the active site of TolC protein.

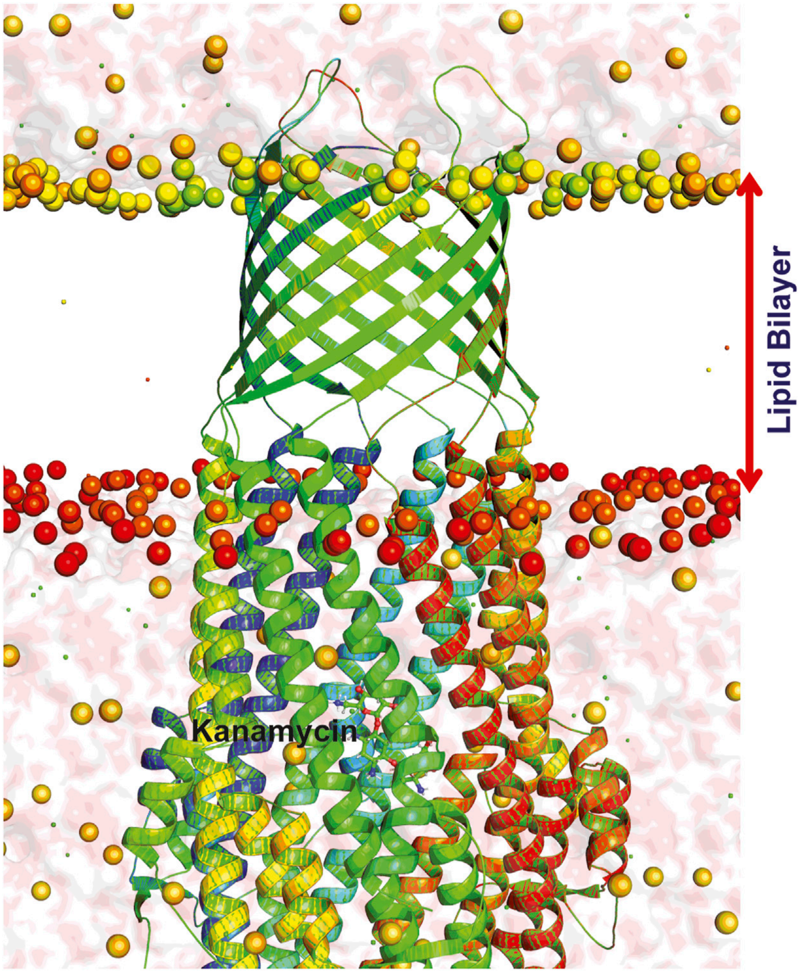


FIGURE 7 | Dynamics of TolC-kanamycin complex in POPC lipid bilayer. The protomers are colored as blue, red, and green with the lipid bilayer of the bacterial outer membrane. The outer membrane embedded ß-barrel is open to the extracellular medium whereas the coiled coils taper close the periplasmic entrance of the α-helical barrel.

dynamics of proteins at the atomic level (Gupta et al., 2017; Beg et al., 2018; Gulzar et al., 2019). MD simulation was performed with the top ranked conformation of kanamycin complex obtained from the molecular docking in POPC homogeneous lipid bilayer for 50 ns to assess all the structural dynamics through RMSDs calculations of the system. The system was comprised of 1,62,774 atoms with 238 POPC lipids, 36969TIP3 water molecules, 115 sodium, and 101 chlorine molecules at 303.15 K. The backbone RMSDs relative to the initial structure computed for the system have been depicted in **Figure 5**.

As evidenced from **Figure 9A**, the RMSD of ligand bound conformation of TolC shown in blue is mostly distributed within the range of 0.287 to 0.312 nm (average of ~0.3 nm) with a constant trend indicating that the system reached equilibrium after 20 ns and maintained a stable trend till 50 ns. Like the protein-ligand complex, the RMSD of the ligand (magenta) also displayed a stable trend, which signifies that ligand within binding pocket remains stable throughout the MD. The radius of gyration (R_g), which portrays the structure compactness and overall dimension of protein, is shown in **Figure 9B**. It represents the mass weighted root mean square distance of a collection of atoms from their common center of mass (Pathak et al., 2018). Here, the average value of R_g ~ 4.06 nm achieved a stable trend after 20 ns, which indicates that the binding of kanamycin with TolC increases the compactness of the secondary structures packed into the 3D structure of protein.

The root-mean-square fluctuation (RMSF) of C_α atoms provides insights into the residual fluctuation and flexibility of the complex. The RMSF values of C_α atoms of each residue of three chains of the protein have been shown in **Figure 10**. The most flexible regions were located at the periplasmic and the extracellular ends of the protein with higher RMSF values. Except for these regions, including loops, most of the residues of the complex displayed a constant trend in RMSF. The ligand binding regions also displayed higher RMSF values with large peaks which reflect the flexibilities of binding site residue and their constant participation in the ligand recognition. The list of amino acids participating in ligand binding and their respective C_α-RMSF values are represented in **Table 5**.

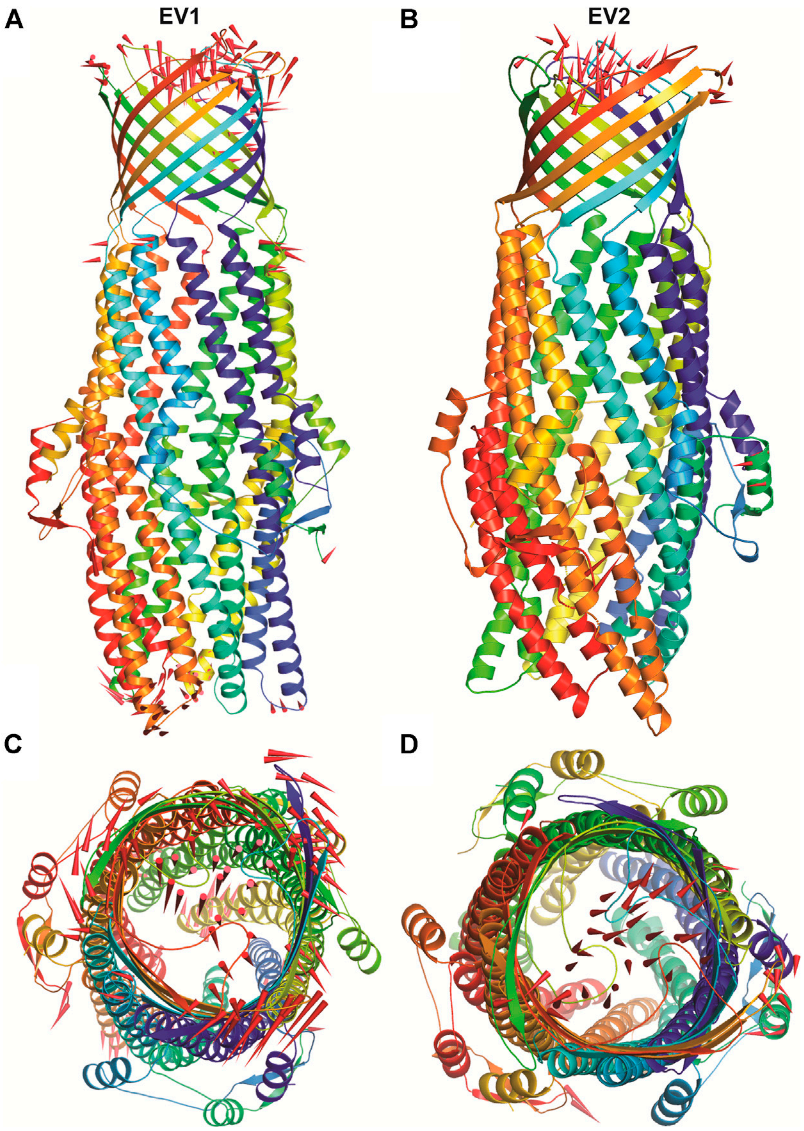


FIGURE 8 | The overall architecture of TolC protein generated with RIBBONS (*p*). **(A)** and **(B)** are the C_α trace of TolC in closed and open form with individual promoters colored differently. The molecular threefold axis is aligned vertically, normal to the plane of the outer membrane. The β-barrel is at the top (distal) end and the *a*-helical domain is at the bottom (proximal) end. **(C)** and **(D)** are the exterior views of the proximal end of the α-helical barrel showing the coiled coils closing the proximal end of the tunnel and open state channel respectively. As PCA only describe the motion of main chain atoms of the complex the ligand (kanamycin) these structures could not be seen.

Intermolecular H-Bonding Analysis

The stability of a biomolecular complex depends upon the intermolecular force of attraction, such as hydrogen bonding. The H-bonding formed in the protein-ligand complex and its stability can be explored to understand the molecular recognition and specificity of the interacting partners (Hubbard and Kamran Haider, 2001). Thus, the intermolecular hydrogen-bonding pattern in TolC protein and kanamycin complex was calculated using the *gmxh bond* tool by measuring the donor-acceptor distances during the MD simulations. Over the 50 ns time scale, TolC and kanamycin complex system exhibited differential intermolecular H-bond pattern with an average 2.86 hydrogen bonds per frame. The observed significant reduction in H-bonds in the complex after MD as compared to docked conformation were compensated by new H-bonds and electrostatic contacts, shown in **Figure 11**. Slight reorientation of the ligand within the binding pocket of TolC might be indicative of the foremost structural changes which might induce a significant decrease in the occupancy of the most H-bonds.

Principal Component Analysis and Free Energy Landscape

Principal component analysis (PCA) is a robust statistical method that eases the convolution of a data set to extract

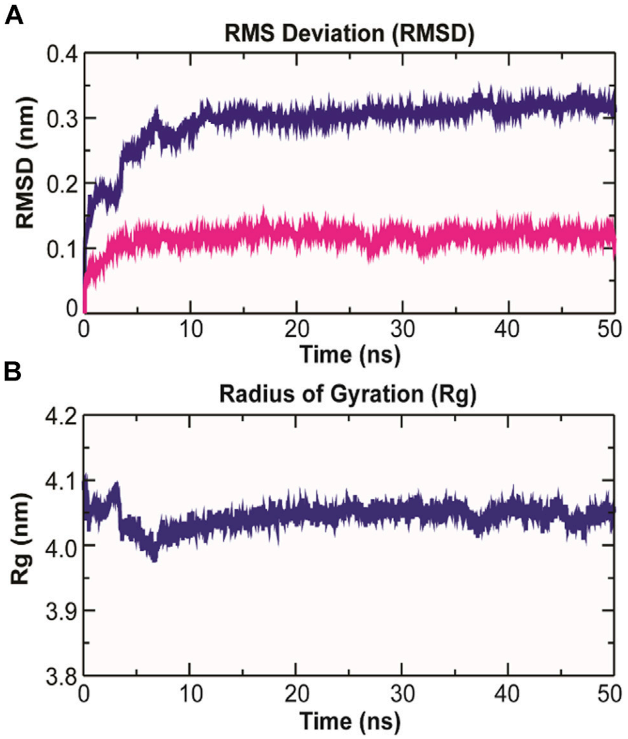


FIGURE 9 | Structural dynamics of TolC on kanamycin binding. **(A)** The RMSD of the C_α atoms from their initial coordinator as function of time (ns). The blue line shows the RMSD for all the C_α atoms and the magenta line indicates the RMSD of the extracellular loops. **(B)** The radius of gyration (R_g) of the C_α atoms from their initial coordinator as function of time (ns).

biologically relevant movements of protein from irrelevant localized motions of atoms. Here, we visualized the sampled conformations in the subspace along the first two PCs using gmx anaeig in a two-dimensional projection, and porcupine plots were plotted (**Figure 12**) to visualize the motions. To illuminate the possible different conformations of the complex adopted during a simulation, free energy landscape (FEL) analysis was conducted using gmx sham module of GROMACS along the first two PCs. Here, we observed that TolC and kanamycin complex was flexible at both EVs and is also quite stable during the simulation after binding to kanamycin.

TABLE 5 | RMSF values of C_α atoms of TolC residues binding with kanamycin.

Residues	RMSF (nm)
Asp101	0.0772
Glu16	0.0941
Lys19	0.0867
Gln184	0.0676
Thr97	0.08
Asn108	0.0624
Asp23	0.0818
Ser20	0.0819

DISCUSSION

Antimicrobial export through the membrane bound efflux pumps is a frequent event that causes microbial resistance (Puzhao et al., 2016). In *E. coli,* the tripartite AcrAB-TolC efflux pump exports antibiotics and is the major cause of developing cross resistance (Nikaido and Zgurskaya., 2001). Studies on *E. coli* AcrAB-TolC efflux-protein complexed system have revealed that the apo-TolC mostly remain in the closed conformation, whereas the holo-TolC switches to interact with other partners, as seen in both the crystal and cryo-EM structures (Koronakis et al., 2000; Zhao et al., 2017). Hence, to get a detailed insight of the kanamycin binding with the outer membrane efflux protein TolC at a molecular level, we performed docking and molecular dynamic (MD) simulation studies. Analysis of the initial drug screened docking results revealed that kanamycin has a competitive docking score with a strong binding affinity with free energy of −6.5 kcal/mol towards TolC protein than other antibiotics.

The molecular weight of TolC obtained from the mascot MALDI mass spectroscopy results was found to be ~54 kDa, which is the same as that earlier reported (Koronakis et al., 2000). The high affinity binding of kanamycin with TolC causes fluorescence intensities quenching. ITC data strengthens the TolC-kanamycin binding process through a single side binding endotherm, and the thermodynamics data obtained reveals an entropy driven stabilized binding process.

The intermolecular and intramolecular forces that stabilize the protein secondary and tertiary structures get affected when exposed to co-solutes or co-solvents (Kelly et al., 2005). The far-UV CD spectra of TolC on kanamycin binding showing the insignificant change in the α−helix and β−sheets contents of TolC protein indicates a high stable secondary structure. Whereas the near-UV CD spectra shows some changes in the tertiary structure that can be correlated with the

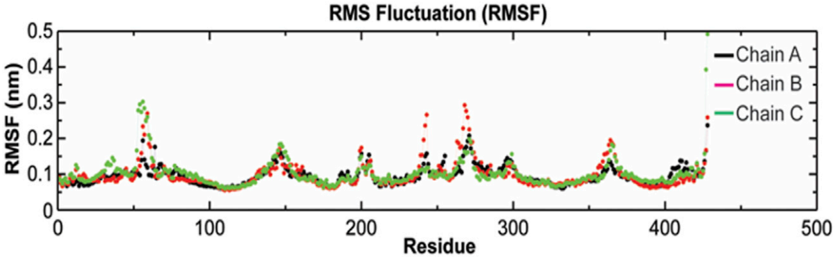


FIGURE 10 | Root-mean-square fluctuations (RMSF) with respect to the residual dynamics of TolC protein on complexation with kanamycin.

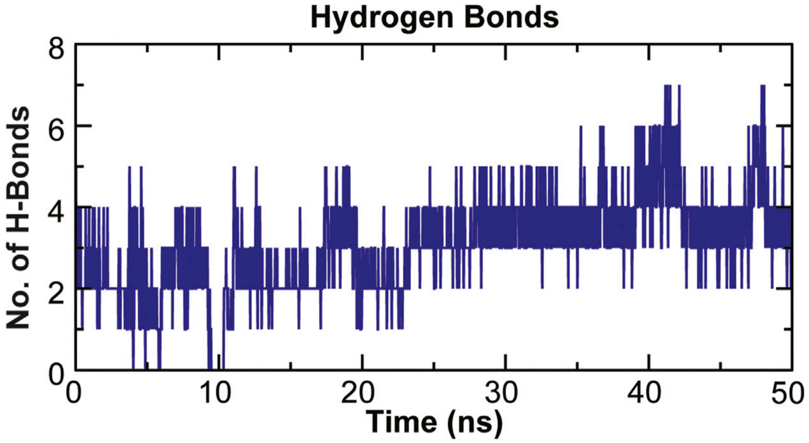


FIGURE 11 | Stability of the TolC-kanamycin complex formed through intermolecular hydrogen bonding and their probable distributions in and around. The *y*-axis represents the number of hydrogen bonds forming with time (ns).

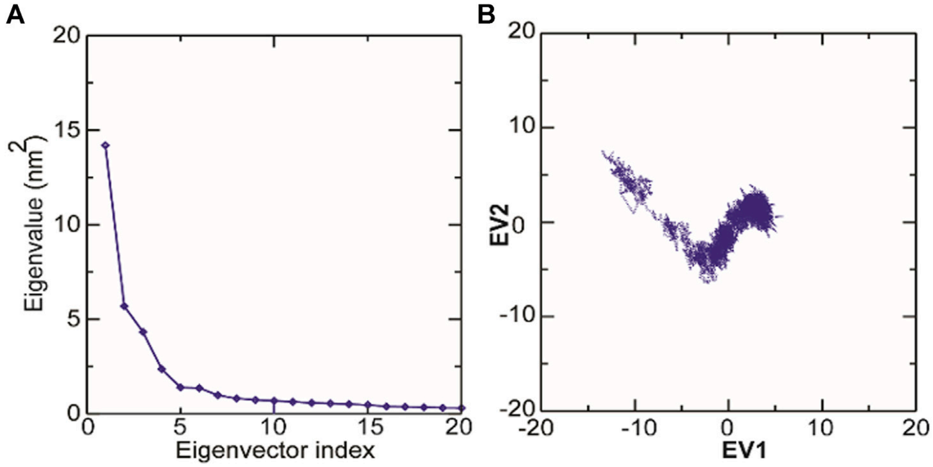


FIGURE 12 | Molecular dynamics analysis of conformers of TolC. **(A)** The plot of eigenvalues corresponding to eigenvector index for the kanamycin complexed with TolC. **(B)** 2D projection of TolC along the EV1 and EV2 template for first two principal components.

structural fluctuation resulting from kanamycin binding and that facilitate TolC to interact with other binding partners in the triparty efflux pump complex.

The docking results have validated the molecular interaction and binding of kanamycin in the active pocket of TolC through the hydrogen bonding, van der Waal forces, and other expected interactions to form a stable TolC-kanamycin complex. The protein TolC was quite stable in the POPC lipid bilayer, and the conformational conversion in the MD study confirms the closed and open conformation as apo- and kanamycin-TolC complexes (**Figures 8A,B**), respectively. This conformational conversion has been reported to be mediated through an iris-like expansion of the periplasmic end, which is in fact necessary to maintain TolC-AcrA contact to permit the drug molecules to pass through the pump (Koronakis et al., 1997).

Further, we observed that the kanamycin binding with TolC is promoted through the inter-promoter hydrogen-bonding network involving five amino acid residues that are bound through twelve H-bonds and are mostly from the α-helices in the protein. The MD simulation data indicates the stability of these conformational changes occurring due to the kanamycin binding with TolC, which was further assessed by 50 nm MD simulation studies. The RMSD, R_g, and RMSF average values all together indicate that the binding of kanamycin with TolC stabilizes the structural complex with insignificant fluctuations. However, within 10 ns of initial simulations, some random fluctuation was observed that was stabilized up to 50 ns Thus, the kanamycin-TolC complex is quite stable after 10 ns of equilibration. The data altogether indicate that the entropy driven kanamycin binding to TolC results in the conformation fluctuation in the protein that favors the efflux. However, there are still limited experimental data to support the antibiotics binding to TolC and efflux across the *E. coli* membrane.

CONCLUSION

We studied the binding of kanamycin with the outer membrane protein TolC and correlated its role in efflux. The results obtained using fluorescence, ITC and bioinformatics techniques indicated the preferential binding of kanamycin with TolC in the active-site and brings conformational changes in the protein. The far-UV CD results also supported the binding conformational changes in both the α-helix and β-sheets, which are part of the tunnel domain and the channel domain, respectively. This change in TolC conformation possibly regulates its interaction with other efflux pump partners (Koronakis et al., 2004). The binding of kanamycin with TolC and its efflux through the AcrAB-TolC system defines the cause of kanamycin resistance *E. coli* strains. Finally, the data altogether provides necessary information to establish the binding of kanamycin with TolC and thereby achieve efflux through the AcrAB-TolC channel. Thus, knowing the antibiotic binding and its efflux will help to discover potential antibacterial agents for the treatment of drug-resistant bacterial infections. However, this study needs further detailed analysis through different techniques to understand the mechanism of drug efflux in a more complete way.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

BP, TB and BM designed the experiments and BP, MP, BD performed it. At the end, BP, BD, SJ and BM analyzed the results and wrote the paper.

FUNDING

This work was supported by UGC, Govt. of India, under the MRP grant for the period 2012–2015 (File No.: 41-283/2012(SR)). Part of this work was also supported through the DBT Project (File No.: BT/PR5614/MED29/687/2013) and HED, Govt. of Odisha to the Center of Excellence in Environmental Sciences and Public Health, Ravenshaw University, Cuttack (File No.:26913/HE-PTC-WB-02-17).

ACKNOWLEDGMENTS

BP thanks UGC Govt. of India for the award of JRF and SRF fellowship. BCM sincerely thanks UGC, Govt. of India for the financial support through the MRP grant which allowed the researchers to carry out this work. We also thank CSIR, Govt. of India for the financial support to Ms. MP. We are grateful to the Center of Excellence in Environment Science and Public Health for financial support. We sincerely acknowledge the facility that we used in the Institute of Life Sciences, BBSR, Department of Chemistry and Center of Excellence, Ravenshaw University, Cuttack to complete this work. We are thankful to Dr Amol R. Suryawanshi, Scientist-D, ILS Bhubaneswar for helping in MALDI mass spectroscopy results. We thank Mr. Ashish Yadav for helping in performing the ITC experiments, ILS, BBSR. We thank Dr Neeraj K. Mishra University of Minnesota, USA for valuable suggestions and criticism about the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.636286/full#supplementary-material.

REFERENCES

Abdullah, S. M. S., Fatma, S., Rabbani, G., and Ashraf, J. M. (2017). A spectroscopic and molecular docking approach on the binding of tinzaparin sodium with human serum albumin. *J. Mol. Struct.* 1127, 283–288. doi:10.1016/j.molstruc.2016.07.108

Aggen, J. B., Armstrong, E. S., Goldblum, A. A., Dozzo, P., Linsell, M. S., Gliedt, M. J., et al. (2010). Synthesis and spectrum of the neoglycoside ACHN-490. *Antimicrob. Agents Chemother.* 54 (11), 4636–4642. doi:10.1128/AAC.00572-10

Ahmad, E., Rabbani, G., Zaidi, N., Ahmad, B., and Khan, R. H. (2012). Pollutant-induced modulation in conformation and β-lactamase activity of human serum albumin. *PLoS one* 7 (6), e38372. doi:10.1371/journal.pone.0038372

Ahmad, E., Rabbani, G., Zaidi, N., Singh, S., Rehan, M., Khan, M. M., et al. (2011). Stereo-selectivity of human serum albumin to enantiomeric and isoelectronic pollutants dissected by spectroscopy, calorimetry and bioinformatics. *PLoS one* 6 (11), e26186. doi:10.1371/journal.pone.0026186

Aono, R., Tsukagoshi, N., and Yamamoto, M. (1998). Involvement of outer membrane protein TolC, a possible member of the mar-sox regulon, in maintenance and improvement of organic solvent tolerance of *Escherichia coli* K-12. *J. Bacteriol.* 180 (4), 938–944. doi:10.1128/jb.180.4.938-944.1998

Baker, C. J. M. D., Barrett, F. F. M. D., and Clark, D. J. (1974). Incidence of kanamycin resistance among *Escherichia coli* isolates from neonates. *J. Pediatr.* 84, 126–130. doi:10.1016/s0022-3476(74)80573-1

Beg, M. A., Thakur, C., and Meena, L. S. (2018). Structural prediction and mutational analysis of Rv3906c gene of Mycobacterium tuberculosis H37Rv to determine its essentiality in survival. *Adv. Bioinform.* 2018, 1–12. doi:10.1155/2018/6152014

Cagnacci, S., Gualco, L., Debbia, E., Schito, G. C., and Marchese, A. (2008). European emergence of ciprofloxacin-resistant *Escherichia coli* clonal groups O25:H4-ST 131 and O15:K52:H1 causing community-acquired uncomplicated cystitis. *J. Clin. Microbiol.* 46 (8), 2605–2612. doi:10.1128/JCM.00640-08

Dehury, B., Behera, S. K., and Mahapatra, N. (2017). Structural dynamics of Casein Kinase I (CKI) from malarial parasite Plasmodium falciparum (Isolate 3D7): insights from theoretical modelling and molecular simulationsPlasmodium falciparum (Isolate 3D7): insights from theoretical modelling and molecular simulations. *J. Mol. Graphics Model.* 71, 154–166. doi:10.1016/j.jmgm.2016.11.012

Dehury, B., Tang, N., and Kepp, K. P. (2019). Molecular dynamics of C99-bound γ-secretase reveal two binding modes with distinct compactness, stability, and active-site retention: implications for Aβ production. *Biochem. J.* 476, 1173–1189. doi:10.1042/bcj20190023

Delmar, J. A., Su, C.-C., and YU, E. W. (2014). Bacterial multidrug efflux transporters. *Ann. Rev. Biophys.* 43, 93–117. doi:10.1146/annurev-biophys-051013-022855

Dinesh, K. B., Vignesh, K. P., Bhubaneswar, S. P., and Mitra, A. (2010). Drugs rules and regulations: different countries (India, China, Russia and United States) - a review. *Int. J. Pharm. Pharma Sci.* 2, 16.

Dinh, T., Paulsen, I. T., and Saier, M. H., Jr (1994). A family of extracytoplasmic proteins that allow transport of large molecules across the outer membranes of gram-negative bacteria. *J. Bacteriol.* 176, 3825–3831. doi:10.1128/jb.176.13.3825-3831.1994

Du, D., van Veen, H. W., and Luisi, B. F. (2015). Assembly and operation of bacterial tripartite multidrug efflux pumps. *Trends Microbiol.* 23, 311–319. doi:10.1016/j.tim.2015.01.010

Fralick, J. A. (1996). Evidence that TolC is required for functioning of the mar/AcrAB efflux pump of *Escherichia coli*. *J. Bacteriol.* 178, 5803–5805. doi:10.1128/jb.178.19.5803-5805.1996

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* 47, 1739–1749. doi:10.1021/jm0306430

Gulzar, A., Borau, L. V., Buchenberg, S., Wolf, S., and Stock, D. (2019). Energy transport pathways in proteins: a non-equilibrium molecular dynamics simulation study. *J. Chem. Theory Comput.* 15 (10), 5750–5757. doi:10.1021/acs.jctc.9b00598

Gupta, A., Mishra, S., Singh, S., and Mishra, S. (2019). (2017). Prevention of IcaA regulated poly N-acetyl glucosamine formation in Staphylococcus aureus biofilm through new-drug like inhibitors: in silico approach and MD simulation study. *Microb. Pathog.* 110, 659–669. doi:10.1016/j.micpath.2017.05.025

Hubbard, R. E., and Kamran Haider, M. (2001). *Hydrogen Bonds in Proteins: role and Strength. eLS.* John Wiley & Sons. doi:10.1002/9780470015902.a0003011.pub2

Johnson, W. C., Jr. (1990). Protein secondary structure and circular dichroism: a practical guide. *Proteins* 7, 205–214. doi:10.1002/prot.340070302

Kelly, S. M., Jess, T. J., and Price, N. C. (2005). How to study proteins by circular dichroism. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1751, 119–139. doi:10.1016/j.bbapap.2005.06.005

Koronakis, V., Eswaran, J., and Hughes, C. (2004). Structure and function OF tolc: the bacterial exit duct for proteins and drugs. *Annu. Rev. Biochem.* 73, 467–489. doi:10.1146/annurev.biochem.73.011303.074104

Koronakis, V., Li, J., Koronakis, E., and Stauffer, K. (1997). Structure of TolC, the outer membrane component of the bacterial type I efflux system, derived from two-dimensional crystals. *Mol. Microbiol.* 23 (3), 617–626. doi:10.1046/j.1365-2958.1997.d01-1880.x

Koronakis, V., Sharff, A., Koronakis, E., Luisi, B., and Hughes, C. (2000). Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature* 405, 914–919. doi:10.1038/35016007

Lakowicz, Joseph. R. (2006). *Principle of fluorescence spectroscopy*. Boston, MA: Springer, 277–330.

Landman, D., Babu, E., Shah, N., Kelly, P., Bäcker, M., and Bratu, S. (2010). Activity of a novel aminoglycoside, ACHN-490, against clinical isolates of Escherichia coli and Klebsiella pneumoniae from New York City. *J. Antimicrob. Chemother.* 65 (10), 2123–2127. doi:10.1093/jac/dkq278

Li, X. Z., and Nikaido, H. (2009). Efflux-mediated drug resistance in bacteria: an update. *J. Bacteriol.* 69, 1555–1623. doi:10.2165/11317030-000000000-00000

Michael, C. A., Dominey-Howes, D., and Labbate, M. (2014). The antimicrobial resistance crisis: causes, consequences, and management. *Front. Pub. Health* 2 (145), 1–8. doi:10.3389/fpubh.2014.00145

Money, P., Kelly, A. F., Gould, W. J., Denholm-Price, J., Threlfall, E. J., and Fielder, M. D. (2010). Cattle, weather and water: mapping Escherichia coli O157:H7 infections in humans in England and Scotland. *Env. Microbiol.* 12 (10), 2633–2644. doi:10.1111/j.1462-2920.2010.02293.x

Morona, R., Manning, P. A., and Reeves, P. (1983). Identification and characterization of the ToIC protein, an outer membrane protein from *Escherichia coli*. *J. Bacteriol.* 153, 693–699. doi:10.1128/JB.150.3.1016-1023.1982

Morona, R., and Reeves, P. (1981). Molecular cloning of the *tolC* locus of *Escherichia coli* K-12 with the use of transposon Tn*10*. *Mol. Gen. Genet.* 184, 430–433.doi:10.1007/BF00352517

Morona, R., and Reeves, P. (1982). The *tolC* locus of *Escherichia coli* affects the expression of three major outer membrane proteins. *J. Bacteriol.* 150, 1016–1023. doi:10.1128/JB.150.3.1016-1023.1982

Nidaido, H. (1994). Prevention of drug access to bacteria targets: permeability barriers and active efflux. *Science* 264, 382–388. doi:10.1126/science.8153625

Nikaido, H. (1998). Multiple antibiotic resistance and efflux. *Curr. Opin. Microbiol.* 1, 516–523. doi:10.1016/s1369-5274(98)80083-0

Nikaido, H., and Zgurskaya, H. I. (2001). AcrAB and related multidrug efflux pumps of *Escherichia coli*. *J. Mol. Microbiol. Biotechnol.* 3, 215–218. doi:10.1159/000103594

Nishino, K., Yamada, J., Hirakawa, H., Hirata, T., and Yamaguchi, A. (2003). Roles of TolC-dependent multidrug transporters of *Escherichia coli* in resistance to β-lactams. *Antimicrobe Agent Chemother.* 47, 3030–3033. doi:10.1128/AAC.47.9.3030-3033.2003

Pathak, R. K., Gupta, A., Shukla, R., and Baunthiyal, M. (2018). Identification of new drug-like compounds from millets as Xanthine oxidoreductase inhibitors for treatment of Hyperuricemia: a molecular docking and simulation study. *Comput. Biol. Chem.* 76, 32–41. doi:10.1016/.j.compbiolchem.2018.05.015

Paulsen, I. T., Brown, M. H., and Skurray, R. A. (1996). Proton-dependent multidrug efflux systems. *Microbiol. Rev.* 60 (4), 575–608.

Paulsen, I. T., Park, J. H., Choi, P. S., and Saier, M. H., Jr. (1997). A family of Gram-negative bacterial outer membrane factors function in the export of proteins, carbohydrates, drugs and heavy metals from Gram-negative bacteria. *FEMS Microbiol. Lett.* 156, 1–8. doi:10.1111/j.1574-6968.1997.tb12697.x

Pennington, T. H. (2014). *E. coli* O157 outbreaks in the United Kingdom: past, present, and future. *Infect. Drug Resist.* 7, 211–222. doi:10.2147/IDR.S49081

Perez-Iratxeta, C., and Andrade-Navarro, M. A. (2008). K2D2: estimation of protein secondary structure from circular dichroism spectra. *BMC Struct. Biol.* 8 (25), 1–5. doi:10.1186/1472-6807-8-25

Poole, K. (2007). Efflux pumps as antimicrobial resistance mechanisms. *Ann. Med.* 39 (3), 162–176. doi:10.1080/07853890701195262

PuZhao, Y. Z., Li, Y., Zou, J., Ma, Q., Zhao, Y., Ke, Y., et al. (2016). Enhanced efflux activity facilitates drug tolerance in dormant bacterial cells. *Mol. Cel* 62, 737–775. doi:10.101/j.molcel.2016.03.035

Rabbani, G., Ahmad, E., Khan, M. V., Ashraf, M. T., Bhat, R., and Khan, R. H. (2015). Impact of structural stability of cold adapted *Candida antartica* lipase B (CaLB): in relation to pH, chemical and thermal denaturation. *RSC Adv.* 5 (26), 20115–20131.doi:10.1039/C4RA17093H

Rabbani, G., Ahmad, E., Zaid, N., and Khan, R. H. (2011). pH-dependent conformational transitions in conalbumin (ovotransferrin), a metalloproteinase from hen egg white. *Cell Biochem Biophys* 61, 551–560. doi:10.1007/s12013-011-9237-x

Rabbani, G., Ahmad, E., Zaidi, N., Fatima, S., and Khan, R. H. (2012). pH-induced motel globule state of *Rhizopus niveus* lipase is more resistance against thermal and chemical denaturation than its native state. *Cel Biochem Biophys* 62, 487–499. doi:10.1007/s12013-011-9335-9

Rabbani, G., and Choi, I. (2018). Roles of osmolytes in protein folding and aggregation in cells and their biotechnology applications. *Int. J. Biol. Macromol* 109, 483–491. doi:10.1016/j.ijbiomac.2017.12.100

Rabbani, G., Kaur, J., Ahmad, E., Khan, R., and Jain, S. K. (2013). Structural characteristics of the thermostable immunogenic outer membrane protein from *Salmonella enterica* serovar Typhi. *Appl. Microbiol. Biotechnol.* 98, 2533–2543. doi:10.1007/s00253-013-5123-3

Rahman, S., Rehman, M. T., Rabbani, G., Khan, P., AlAjmi, M. F., Hassan, M. I., et al. (2019). *Int. J. Mol. Sci.* 20, 2727. doi:10.3390/ijms20112727

Ristuccia, A. M., and Cunha, B. A. (1985). An overview of Amikacin. *Therap. Drug Monitor.* 7 (1), 12–25. doi:10.1097/00007691-198503000-00003

Rogozea, A. (2012). EPR and circular dichroism solution studies on the interactions of bovine serum albumin with ionic surfactants and β-cyclodextrin. *J. Phys. Chem. B* 116, 14245–14253.

Sulavik, M. C., Houseweart, C., Cramer, C., Jiwani, N., MurgoloGreen, N. J., DiDomenico, B., et al. (2001). *Antimicrobe Agents Chemother*. 45, 1126–1136. doi:10.1128/AAC.45.4.1126-1136.2001

Suryawanshi, V. D., Walekar, L. S., Gore, A. H., Anbhule, P. V., and Kolekar, G. B. (2016). Spectroscopic analysis on the binding interaction of biologically active pyrimidine derivative with bovine serum albumin. *J. Pharm. Anal.* 6, 56–63. doi:10.1016/j.jpha.2015.07.001

Thakur, K., Kaur, T., Singh, J., Rabbani, G., Khan, R. H., Hora, R., et al. (2017). *Sauromatum guttatum* lectin: spectral studies, lectin-carbohydrate interaction, molecular cloning and *in silico* analysis. *Int. J. Biol. Macromol* 104, 1267–1279. doi:10.1016/j.ijbiomac.2017.06.123

Tikhonova, E. B., Yamada, Y., and Zgurskaya, H. I. (2011). Sequential mechanism of assembly of multidrug efflux pump. *Acrab-tolc*. Chem. Biol. 18, 454–463.

Varlan, A., and Hillebrand, M. (2010). Bovine and human serum albumin interactions with 3-carboxyphenoxathiin studied by fluorescence and circular dichroism spectroscopy. *Molecules* 15, 3905–3919.doi:10.3390/molecules15063905

Varshney, A., Ahmad, B., Rabbani, G., Kumar, V., Yadav, S., and Khan, R. H. (2010). Acid-induced unfolding of didecameric keyhole limpet hemocyanin: detection and characterizations of decameric and tetrameric intermediate states. *Amino Acids* 39, 899–910. doi:10.1007/s00726-010-0524-4

Varshney, A., Rabbani, G., Badr, Gamal., and Khan, R. H. (2014). Cosolvents induced unfolding and aggregation of keyhole limpet hemocyanin. *Cel Biochem Biophys* 69, 103–113. doi:10.1007/s1203-013-97764-4

Wang, H., Dzink-Fox, J. L., Chen, M., and Levy, S. B. (2001). Genetic characterization of highly fluoroquinolone-resistant clinical *Escherichia coli* strains from China: role ofacrR mutations. *Antimicrob. Agents Chemother*. 45, 1515–1521. doi:10.1128/AAC.45.5.1515-1521.2001

Zgurskaya, H. I. (2009). Multicomponent drug efflux complexes: architecture and mechanism of assembly. *Future Microbiol.* 4, 919–932. doi:10.2217/fmb.09.62

Zgurskaya, H. I., and Nikaido, H. (1999). Bypassing the periplasm: reconstitution of the AcrAB multidrug efflux pump of *Escherichia coli*. *Proc. Natl. Accad. Sci. U.S.A.* 96, 7190–7195. doi:10.1073/pnas.96.13.7190

Zhang, D., Li, H., Lin, X., and Peng, X. (2015). Outer membrane proteomics of kanamycin-resistant *Escherichia coli* identified MipA as a novel antibiotic resistance-related protein. *FEMS Microbiol. Lett.* 362, 1–8. doi:10.1093/femsle/fnv074

Zhang, L., Sahu, I. D., Xu, M., Wang, Y., and Hu, X. (2016). Data for β-lactoglobulin conformational analysis after (-)-epigallocatechingallate and metal ions binding. *Data Brief* 10, 474–477. doi:10.1016/j.dib.2016.12.021

Zhao, W., Guizhen, F., Corey, F. H., James, N. B., Irina, I. S., Michael, F. S., et al. (2017). An allosteric transport mechanism for the AcrAB-TolC multidrug efflux pump. *eLife* 29 (6), e24905. doi:10.7554/eLife.24905

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.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 10 June 2021
doi: 10.3389/fmolb.2021.691901

Unexpected Gating Behaviour of an Engineered Potassium Channel Kir

Charline Fagnen[1,2], *Ludovic Bannwarth*[1], *Dania Zuniga*[1], *Iman Oubella*[1], *Rita De Zorzi*[3], *Eric Forest*[4], *Rosa Scala*[5], *Samuel Guilbault*[5], *Saïd Bendahhou*[5], *David Perahia*[2] and *Catherine Vénien-Bryan*[1]*

[1]UMR 7590, CNRS, Muséum National d'Histoire Naturelle, Institut de Minéralogie, Physique des Matériaux et Cosmochimie, IMPMC, Sorbonne Université, Paris, France, [2]Laboratoire de Biologie et de Pharmacologie Appliquée, Ecole Normale Supérieure Paris-Saclay, Centre National de la Recherche Scientifique, Gif-sur-Yvette, France, [3]Department of Chemical and Pharmaceutical Sciences, University of Trieste, Trieste, Italy, [4]IBS University Grenoble Alpes, CNRS, CEA, Grenoble, France, [5]Faculté de Médecine, CNRS UMR7370, LP2M, Labex ICST, University Côte d'Azur, Nice, France

In this study, we investigated the dynamics and functional characteristics of the KirBac3.1 S129R, a mutated bacterial potassium channel for which the inner pore-lining helix (TM2) was engineered so that the bundle crossing is trapped in an open conformation. The structure of this channel has been previously determined at high atomic resolution. We explored the dynamical characteristics of this open state channel using an *in silico* method MDeNM that combines molecular dynamics simulations and normal modes. We captured the global and local motions at the mutation level and compared these data with HDX-MS experiments. MDeNM provided also an estimation of the probability of the different opening states that are in agreement with our electrophysiological experiments. In the S129R mutant, the Arg129 mutation releases the two constriction points in the channel that existed in the wild type but interestingly creates another restriction point.

Keywords: molecular dynamics and normal modes, HDX-mass spectrometry, single channel recording, potassium channel KirBac3.1, mutation effect

OPEN ACCESS

Edited by:
Aljaz Godec,
Max Planck Institute for Biophysical Chemistry, Germany

Reviewed by:
Harley Takatsuna Kurata,
University of Alberta, Canada
Wojciech Kopec,
Max Planck Institute for Biophysical Chemistry, Germany

Correspondence:
Catherine Vénien-Bryan
catherine.venien-bryan@upmc.fr

Specialty section:
This article was submitted to Biophysics,
a section of the journal Frontiers in Molecular Biosciences

Received: *07 April 2021*
Accepted: *26 May 2021*
Published: *10 June 2021*

Citation:
Fagnen C, Bannwarth L, Zuniga D, Oubella I, De Zorzi R, Forest E, Scala R, Guilbault S, Bendahhou S, Perahia D and Vénien-Bryan C (2021) Unexpected Gating Behaviour of an Engineered Potassium Channel Kir. Front. Mol. Biosci. 8:691901. doi: 10.3389/fmolb.2021.691901

INTRODUCTION

A detailed study of function requires careful dissection of the mechanistic steps. Protein engineering can provide a powerful tool for studying the relationships between structure and function. The design of various potassium channels with carefully chosen replacement residues has helped describe the gating mechanism of these channels. For instance, we can mention mutations close to the selectivity filter (Capener et al., 2003), on the wall of the cytoplasmic pore (Fujiwara and Kubo, 2006), on the cytoplasmic domain (Inanobe et al., 2013), at the end of the cytoplasmic pore (Pegan et al., 2005), at the extracellular domain of Kir2.2 (Li et al., 2014), at the bottom of the bundle crossing (Linder et al., 2015), or at the level of the cytoplasmic domain subunit interfaces (Wang et al., 2017). All these investigations, either *in silico* or experimental (NMR, FRET, etc.) provided valuable information.

A few years ago, the open state kir channel's crystal structure was revealed by KirBac3.1 S129R (Bavro et al., 2012; Zubcevic et al., 2014), which was designed so that the channel was trapped in an open conformation. Indeed, before the publication of this structure, most structures were known in the closed state, with the conduction pathway occluded. The use of an engineered protein made it possible to observe for the first time at high resolution the KirBac channel with the bundle-crossing gate in an open conformation, where the constriction points (Leu124 and Tyr132) are released (Bavro et al., 2012) as shown in **Figure 1**. This structure allowed proposing a mechanism for opening the channel. In this structure, we noticed that the mutated residue Arg was facing the channel's center

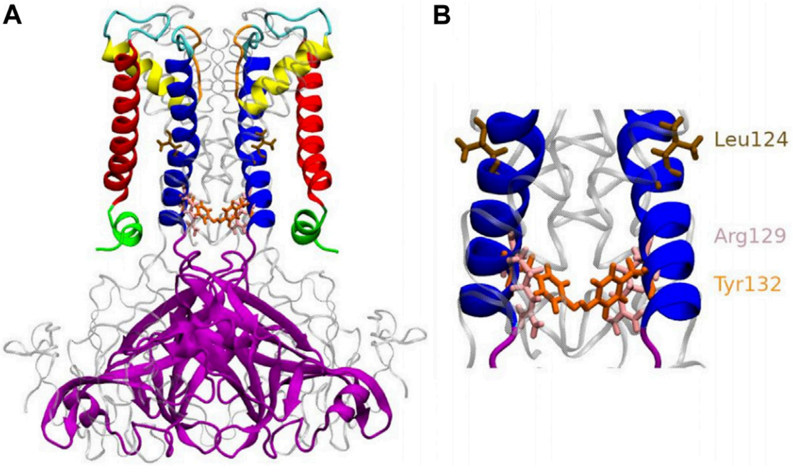


FIGURE 1 | Anatomy of KirBac3.1 S129R from the modeled KirBac3.1 S129R **(A)** The transmembrane portion of each monomer of KirBac3.1 is composed of four helices: slide helix (green), transmembrane helix 1 (red), pore helix (yellow), and transmembrane helix 2 (blue). The mutation S129R is situated at the level of the helix bundle (pink), at the bottom of the inner helix **(B)** Leu124 and Tyr132 are shown in color brown and orange respectively. Arg129 is shown in pink.

and therefore could create the condition for another unexpected constriction point. However, this channel is functionally open (Paynter et al., 2010) and we did not notice any toxic effect of this engineered protein on the host cell (De Zorzi et al., 2013). We then decided to investigate further this mutated channel's function and dynamics using an experimental and *in silico* study, allowing us to explain its particular behavior, which, despite unexpected characteristics, provided valuable information on the open state structure.

RESULTS AND DISCUSSION

Hydrogen/Deuterium Exchange Coupled to Mass Spectrometry

We investigated the protein conformational flexibility of the S129R mutant protein (open state) using HDX-MS. This technique is based on the exchange of deuterium atoms at the amide backbone of a protein, reflecting its conformational dynamics, followed by proteolytic digestion and spectrometry analysis. HDX has been widely used on soluble and membrane proteins (Forest et al., 2016). HDX was performed on the purified KirBac3.1 S129R mutant protein in the presence of detergent (**Figure 2B**). We have established in previous work that the presence of detergent does not affect conformational changes of the KirBac channel (Gupta et al., 2010; Fagnen et al., 2020). The results were compared to those of WT for which the same detergent was used [(Fagnen et al., 2020) and **Figure 2A**], the comparison is shown in **Figure 2C** where red shows a S129R segment more flexible compared with the same segment in WT. The Optimized conditions resulted in sequence coverage of 86% with nepenthesin (Fagnen et al., 2020). However, this enzyme did not allow covering the regions 57–87 (top half of the TM1 and the beginning of the pore helix), 143–147 (β3) and 195–203 (second half of the β7), for nomenclature see **Figure 2C**. Deuterium incorporation was monitored as a function of time for each peptide generated from the S129R mutant (**Supplementary Figure S1**).

Comparison Between the KirBac3.1 WT (Closed State) and S129R Mutant (Open State)

Our data shows that the most flexible regions for KirBac3.1 S129R are the loops extending outside the CTD (aa 271–285 between β14 and β15, in red **Figure 2B**; see also **Figure 2C** for the nomenclature). This external loop is subjected to the swinging movement during the gating (Fagnen et al., 2020). If we compare the closed state of KirBac3.1 and the open state of KirBac3.1 S129R, the latter shows slightly more pronounced flexibility with a maximum value for the HDX of 68% ± 2.2 (Mean ± S.D $n = 3$) against 59.1% ± 2.5 for the closed state as shown in **Figure 2C**.

Structural Flexibility of the Transmembrane Domain During the Gating

The largest change in the deuteration exchange percentage in the S129R mutant (35 vs. 21%, compared with WT) is observed for the Thr93-Leu112 peptides [end of α3 (pore helix), selectivity filter and top of α4 (TM2)], a feature described previously (Gupta et al., 2010). This includes the Met94 which is located towards the base of the pore-loop helix and packs closely with the Gly120 in TM2. Based on deuterium exchange percentages, the inner helix in the mutant S129R appears more flexible than in WT. At the TM1 level, there is an increase in deuteration, particularly at the bottom of the external helix, of 53.7% ± 2.2 against 45.2% ± 1.3 (**Figure 2C**).

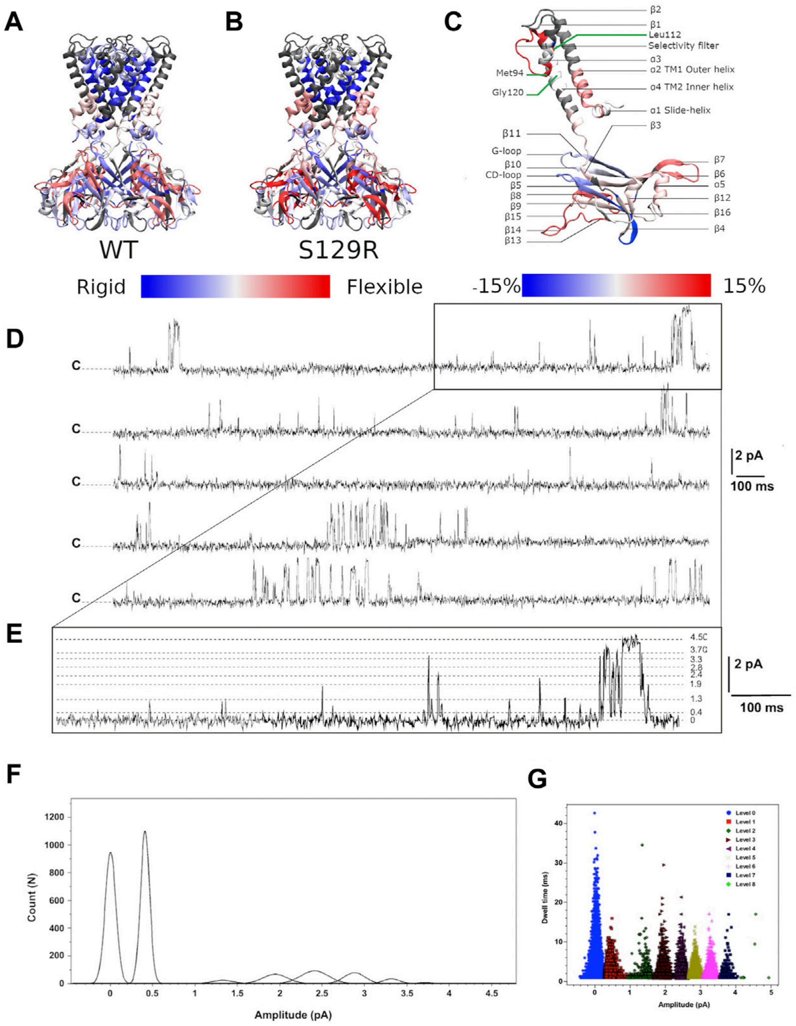


FIGURE 2 | HDX-MS and single channel recordings HDX-MS rates of peptides reported for **(A)** KirBac3.1 WT (Forest et al., 2016) and **(B)** the mutant S129R. Identified peptides are drawn with blue (low exchange and low flexibility) to red (high exchange and high flexibility) color according to their percentage of deuterium exchange after 1,200 s (scale of exchange shown at the bottom of the figures). **(C)** Nomenclature of secondary structures and rate of exchanges between the mutant S129R and KirBac3.1 WT. Scale of deuterium exchange rate is shown at the bottom of the figures, same color code as for A and B. Red appears when S129R is more flexible than WT and blue when S129R is more rigid than WT **(D)** Single channel recordings from KirBac3.1 S129R channels, traces obtained from 6 min consecutive recordings at +150 mV holding potential. Closed state is labeled c **(E)**, an enhancement of the upper trace to show the multiple sub states induced by the S129R mutant **(F)** amplitude histogram fits (Gaussian) of all events for the selected levels gating between closed and open states during 6 min **(G)** S129R channels gate with multiple subconductance states. Dwell time for all events at 150 mV test potential is plotted for all current levels.

Cytoplasmic Domain

These domains do not remain static during gating, and conformational changes should occur as the channel opens and closes. For both closed and open states, the greatest flexibility is found at the external loop. The KirBac3.1 cytoplasmic domain consists of two major β-sheets, one (which we refer to as βI, includes the large β6, β10, and β11), that is tilted about 45° to the membrane plane, and a second referred to as βII (which includes the shorter β3, β5, and β9) is approximately parallel to the pore axis as described in (Wang et al., 2012). Our flexibility measure shows that the main βI sheet is more rigid than the other major βII sheet for both the WT (closed) and S129R mutant (open). The cytoplasmic domain's interior is more rigid than the exterior, with the highest values of flexibility for the open state.

On the contrary, the G-loop, located next to the CD loop is slightly less flexible in the S129R mutant (open state) (9% ± 1.2 against 12% ± 0.8 in the close state). The G-loop has been described as very flexible (Bichet et al., 2003; Pegan et al., 2005; Nishida et al., 2007; De Zorzi et al., 2013). The five amino acids, 162 to 174 of the CD-loop, are also more rigid in the S129R mutant (open state) than in the WT (closed state) (49% ± 2.3 vs. 58% ± 2.5). The decrease in deuteration in these two loops shows that they are involved in a network of interactions with neighboring amino acids, which are therefore less flexible. This was also found in (Gupta et al., 2010).

To Further Assess the Structural Deviations of These Two States

we compared the root mean square displacements (rmsd) of the residues (equation given in *Materials and Methods*) calculated over the MDeNM relaxed structures of KirBac3.1 WT (closed) and KirBacS129R (open). The results are shown in **Supplementary Figure S2**. Minor rmsd values were observed in the closed state, particularly in the transmembrane region. For both structures, the smallest rmsd values (less than 2.8 Å) are in the transmembrane region. The open state shows slightly higher values, particularly in the region of TM1, the pore helix, and the bottom of TM2 with a high value at the position of the S129R mutation. The cytoplasmic domain exhibits higher rmsd values particularly the external loop, which reaches 5.8 and 5.2 Å, for the open and closed states, respectively. This is in agreement with HDX-MS data, for which the highest flexibility is in this region.

Current Recordings of KirBac3.1 in Planar Lipid Bilayers

When reconstituted into black lipid membranes, the KirBac3.1 S129R channels exhibit significant gating activity as shown by current recordings for 6 min at +150 mV (**Figure 2D**). As reported for the KirBac3.1 WT and KirBac 1.1, the KirBac 3.1 S129R gates with multiple subconductance states (Cheng et al., 2009; Clarke et al., 2010). The amplitude values for these levels are 0.40 ± 0.001, 1.32 ± 0.002, 1.94 ± 0.003, 2.41 ± 0.001, 2.88 ± 0.002, 3.31 ± 0.001, 3.71 ± 0.002, and 4.50 ± 0.007 pA (**Figure 2E**). Fits of Gaussian distributions of amplitude histograms resulted in a single channel current level of 4.5 ± 0.007 pA, corresponding to a conductance value of 30 pS (**Figure 2F**). A value similar to that has been obtained for KirBac 3.1 WT (47 pS) (Fagnen et al., 2020). KirBac S129R gates with subconductance level activity that increases the Po to levels as high as 44.05 ± 2.6% (Mean ± S.D. *n* = 16,107, number of events). Plotting dwell time vs. current amplitude of all the events shows that the S129R mutant gates with more subconductance levels, than the WT, that are contributing to the overall Po as follow: 7, 3, 11, 9, 5, 6, 4, and 1%, respectively, (**Figure 2G**). We have already shown that WT channels gate with only two subconductance levels (1 and 2 pA current amplitude) (Fagnen et al., 2020).

Theoretical Results

The theoretical results presented in this section are based on MDeNM (Molecular Dynamics with excited Normal Modes) simulations in which different linear combinations of a selected set of normal modes (NMs) related to the opening/closing of the channel are excited in molecular dynamics (MD) simulations (Costa et al., 2015). Through such a combined use of both methods, MDeNM allows a realistic exploration of the normal mode space relevant for the opening/closing mechanism taking into account the full environment (membrane, water, ions) of the protein at the ambient temperature. Standard MD simulations were thereafter carried out on a uniformly distributed set of structures obtained from MDeNM, which provided a reasonably good estimate of the populations of open/closed (and partially open) states. MDeNM is based on covering uniformly without any bias the whole normal mode space defined by a selected set of low frequency modes. In our study the normal modes that were chosen are all those that are involved in the opening/closing motion of the channel. Therefore, the open and closed states were equally and uniformly sampled. The tests that were done in the original article (Costa et al., 2015) has shown an extensive not biased sampling, giving a good estimation of the probabilities of different states.

Constriction Points Along the Channel in the KirBac3.1 WT and the S129R Mutant

In this work, a "closed" state is defined by a conduction pathway which is sterically occluded and an "open" state in which the pathway is sufficiently wide to accommodate at least a non hydrated potassium ion. The channel encompasses the region between residues 121 and 133. The constriction points in this region for the WT are located at the levels of Leu124 and Tyr132 (**Figure 1B**). To have a dynamical view of the relaxed structures obtained in the MD simulations (that follow the MDeNM), we calculated the shortest atom-atom distance (including hydrogens) between the same residues that are at the opposite chains at a given constriction level (Tyr132 or S/R129). The distances were calculated on KirBac3.1WT (closed state) and the KirBac3.1 S129R (open state) (**Figure 3A,B**, respectively). The average shortest distance at the level of Ser129 in KirBac3.1 WT is 9.53 Å (SD = 0.47 Å) (**Figure 3A** in red) and that at the level of Arg129 in S129R mutant is narrower with a value of 4.46 Å (SD = 1.28 Å) (**Figure 3A** in blue). This is a marked decrease in the channel diameter at the level of the mutation. Interestingly, the mutation at residue 129 introduces another constriction point in the channel which should not allow the K^+ ion to pass easily. **Figure 3** shows a set of distances situated in the gray zone which is indicative of a closed state.

Tyr132 has been described as a constriction point in KirBac3.1 WT confirmed with an average shortest distance obtained of 3.72 Å (SD = 0.87 Å) (Fagnen et al., 2020). In the KirBac3.1 S129R mutant, the average shortest distance, at this constriction point is larger with a value of 5.75 Å (SD 0.87 Å) (**Figure 3B** points in blue and **Figure 1B** in orange). This point of constriction is therefore released. The Leu124 constriction point is also in an open configuration in the KirBac3.1 S129R mutant (**Figure 1B** in brown). We provided in **Supplementary Table S1** the radius values of the pore (computed with the HOLE program) at the constriction levels, as well as the kink angle values of the TM1 for the KirBac3.1 WT and KirBac3.1 S129R crystallographic structures, and KirBac3.1 S129R simulated structures in one of their closed or open conformations. It is worth to notice that the pore at the level of the mutation can adopt a narrower radius in a representative MDeNM structure comparatively to its X-ray structure. We also notice that at the level of the residue 132 we have a larger pore radius for the simulated closed structure than in the X-ray structure. The kink angles of TM1 in three

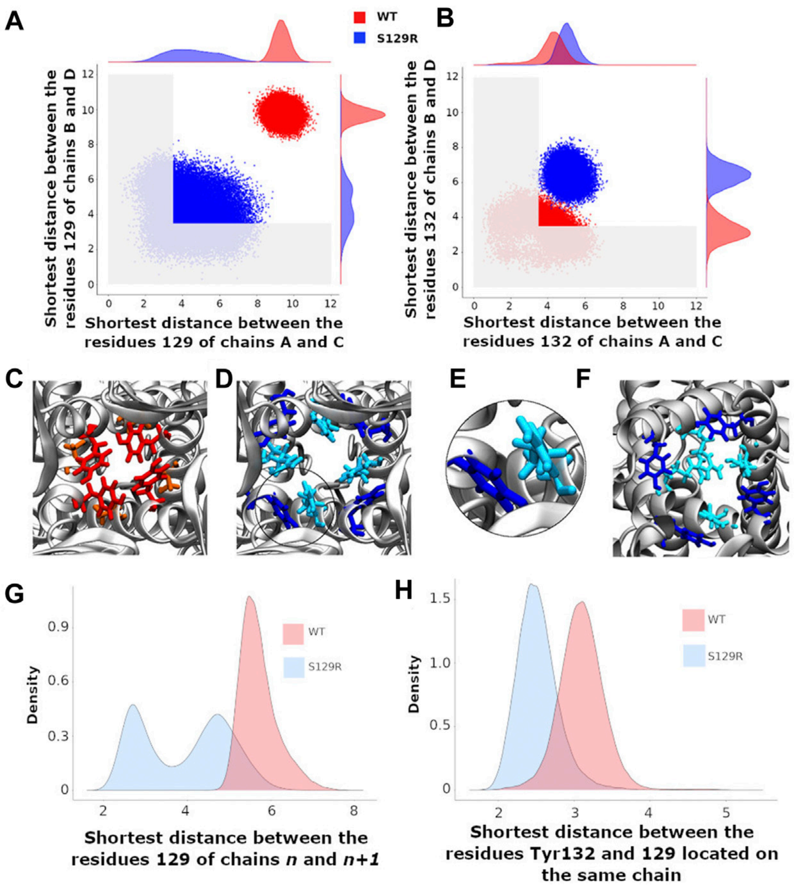


FIGURE 3 | Gating at the mutation S129R and at the constriction point Tyr132 **(A)** Scatter plots of the shortest distances between the chains B and D and between A and C at the level of the residue 129. The shortest distance between two residues is that between their respective atoms including the hydrogens; **(B)** Scatter plots of the shortest distances between the chains B and D and between A and C at the level of the residue Tyr132. A,B) Red and blue points correspond to the KirBac3.1 WT and KirBac3.1 S129R, respectively. The gray area delimits the region where the channel is closed **(C–F)**: representative relaxed structures; **(C)** Locations of the residues Tyr132 (red) and Ser129 (orange) in KirBac3.1 WT; **(D)** Locations of the residues Tyr132 (dark blue) and Arg129 (cyan) in KirBac3.1 S129R; **(E)** Zoom on the Van der Waals contacts between Tyr132 and Arg129 in KirBac3.1 S129R; **(F)** Locations of the residues Tyr132 (dark blue) and Arg129 (cyan) pointing toward the center of the channel in KirBac3.1 S129R; **(G)** Histograms of the shortest distance between residue 129 of the chain *n* and the residue 129 of the chain *n*+1; **(H)** Histograms of the values of the shortest distances between the residues 129 and Tyr132 from the same chain. The histograms in panels **(G,H)** were established by taking into account all the chains.

chains out of four are also larger in the chosen simulated open structure than in the closed one, but their mean values (**Figure 4**) are all larger in the mutant than in the WT, a trend similar to what is observed when comparing both X-ray structures.

Interactions of the Residue 129

We investigated the interactions of residue 129 with its neighboring residues and the changes in the interaction network caused by the mutation in all the relaxed MDeNM structures (Fagnen et al., 2020).

The probability density of the distances between two adjacent Ser129 in WT showed a single peak around 5.9 Å, the two residues being quite distant (**Figure 3G**, in red). Note that higher density of probability means more favorable interactions. For the mutant, the distances between two adjacent Arg129 are distributed into two populations centered around 2.7 and 4.8 Å (**Figure 3G**, in blue). The first peak corresponds to a close interaction between the two arginine heads which point towards the center of the channel and therefore obstructing it (**Figure 3F** in cyan), the second representing more distant residues similar to KirBac3.1 WT. The interaction energy computed between pairs of Arg shows that those belonging to opposite chains can interact favorably as shown in the scatter plot of the interaction energy *vs.* the shortest distance in Figure S3 computed for the ensemble of the MDeNM relaxed structures. It is seen that interaction energies can reach values close to −4 kcal/mol. The representative structure corresponding to the lowest interaction energy is also displayed in this figure. Although Arg has a positive charge it was shown that they can interact favorably between them adopting different orientations (Pednekar et al., 2009).

Note that the distribution of shortest distances depend on which chains are considered. Indeed, the crystallographic structures of KirBac can exhibit four-fold symmetry, but more

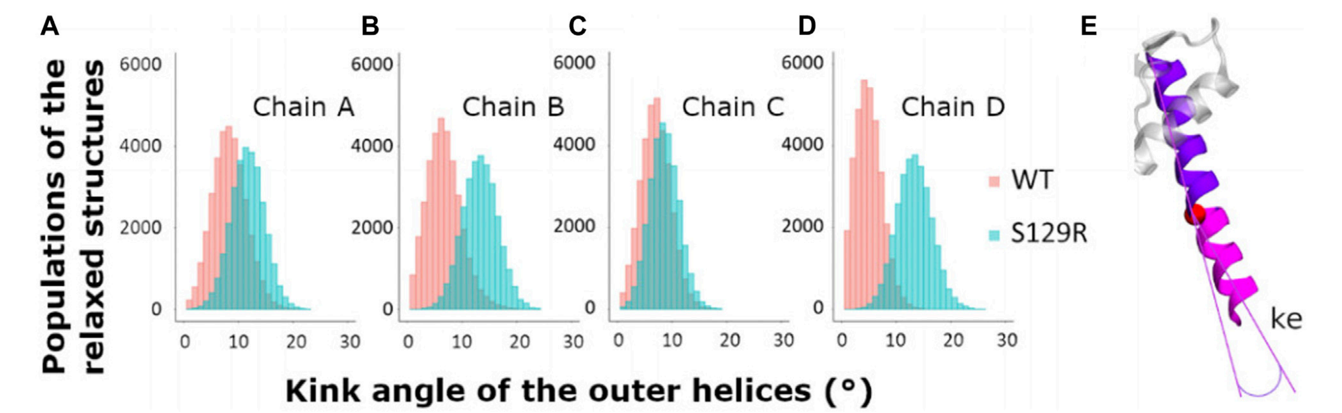


FIGURE 4 | Kink of the outer helices, TM1. Histograms of the values of the kink of the external helices during the relaxed simulations of molecular dynamics after MDeNM. The structures populations of KirBac3.1 WT are represented in red and the structures of KirBac3.1 S129R in blue **(A)** Kink of the outer helix of the chain A **(B)** Kink of the outer helix of the chain B **(C)** Kink of the outer helix of the chain C **(D)** Kink of the outer helix of the chain D **(E)** Representation of the outer kink angle (ke).

often two-fold symmetry or even no symmetry (Clarke et al., 2010; Bavro et al., 2012; De Zorzi et al., 2013). Moreover, the side chains in these crystallographic structures are slightly asymmetric.

We investigated the interaction between the mutated residue and Tyr132 which constitutes a region of constriction along the channel in WT (**Figure 3H**). The most populated shortest distance is greater for WT (3.3 Å) than for mutant (2.5 Å). The contact between the two residues for the mutant is shown in **Figures 3D,E**.

In the closed state, residue 132 points towards the center of the channel and obstructs the passage of the K^+ ion (**Figure 3C** in red). In the mutant, Tyr132 is displaced from the channel's center coming in contact with Arg129, as shown in **Figure 3D** (See also **Supplementary Figure S4** for details).

The interaction energies between the Arg129 and Tyr132 as a scatter plot are given in **Supplementary Figures S5, S6** with the molecular graphics of the most favorable structures. They show that they can interact strongly with an interaction energy around −10 kcal/mol in the case of Arg129B and Tyr132C forming a hydrogen bond between them (**Supplementary Figure S6**), and around −5 kcal/mol between Arg129B and Tyr132B. Interestingly, the Arg129 residues in all four chains point either to the center of the channel and thus block the passage of K^+ ion (**Figure 3F**) or interact with the aromatic ring of Tyr132 residues (**Figure 3D**).

Open and Closed State Populations

Four channel-gating states can be defined based on the open or closed conformation of the two main constriction points (Leu124 and Tyr132) as observed in relaxed structures [for more details, see (Fagnen et al., 2020) **Supplementary Figure S7**]: 1) Fully open state, sufficiently wide to accommodate at least a non-hydrated K^+ ion, that is, when the shortest diametrically opposed inter-chain atomic distances at the two constriction points are greater than the ionic diameter of K^+ [diameter of K^+ ion considered 3.5275 Å (Huang and MacKerell, 2013)]; 2) fully closed state, when both distances are less than the ionic diameter of the potassium ion; 3) partially open state 1, when the gate at residue Leu124 is open, and the gate at residue Tyr132 is closed; 4) partially open state 2 when the gate at residue Tyr132 is open, and the gate at residue Leu124 is closed. Considering the previous observation showing that the side chain of the mutant Arg129 can point towards the center of the channel and constitutes another constriction point, we therefore added another partially closed configuration at the level of Arg129.

We analyzed 34,086 relaxed structures issued from MDeNM simulations for KirBac3.1 WT and 29,600 for KirBac3.1 S129R to have important information on these states' populations. The populations of the different WT and mutant states are shown in **Table 1**, indicating that the fully open state in KirBac3.1 WT is only populated by about 6.8%. Such a low value is consistent with the population obtained by previous electrophysiological experiments (Fagnen et al., 2020). In contrast, the S129R mutant is mostly open by about 53%, the Arg129 keeping the two Leu124 and Tyr132 restriction points, always open. Interestingly, the Arg can adopt two conformations: one in which it interacts directly with Tyr132, and the other where it points to the channel's center obstructing it. Therefore, the closed state of the mutant depends only on the conformation of Arg. Ironically, the mutated residue, which was introduced in the protein to force the channel to open, by trapping the bundle crossing in an open conformation, causes some obstruction to the potassium ion's passage.

Structural Modifications Between the Closed (KirBac3.1 WT) and Open (KirBac3.1 S129R) States

Kink of the Outer Helix TM1

The TM1 outer helices' kink angles were calculated to determine the extent to which their bending is involved in the channel's opening. We compared the kink angles for the TM1s for KirBac3.1 WT and S129R, which are given in **Figure 4**.

We calculated the kink of the outer helix on all the relaxed structures. The mean values of the kink of TM1 for the chains A,

TABLE 1 | Populations (in percentage) of different opening states in the relaxed structures of KirBac S129R obtained through MDeNM simulations and single channel recording; comparison with KirBac3.1 WT is shown.

Opening types	KirBac3.1 WT (%)	KirBac3.1 S129R (%)
Fully open	6.8	52.8
Fully closed	50.2	0.0
Gating 124 open, gating 132 closed	28.8	0.0
Gating 132 open, gating 124 closed	14.2	0.0
Gating 129 closed	0.0	47.2
Current recordings of KirBac3.1 in planar lipid bilayers	9.9 (±1.3, *n* = 1803) (ref 14)	44.05 (± 2.6% mean ± S.D. *n*= 16,107)

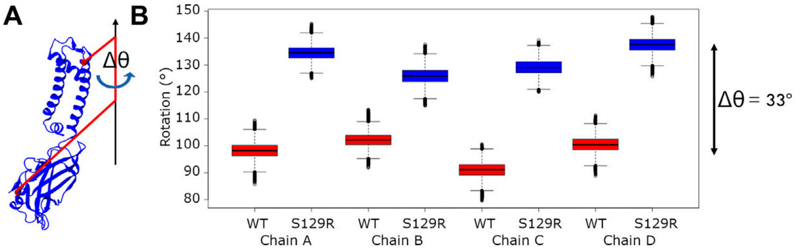


FIGURE 5 | Rotation of the cytoplasmic domain **(A)** Definition of the cytoplasmic domain rotation angle of a given chain; **(B)** Boxplots of the rotation angle's values of the cytoplasmic domain for each chain for KirBac3.1 WT and KirBac3.1 S129R. The angles were computed on all the relaxed structures of KirBac3.1 WT (red) and KirBac3.1 S129R (blue). The middle line of each box corresponds on the median values. The variation of the 33° between the WT and mutant were obtained from these average values.

B, C, and D of KirBac3.1 WT are respectively of 8.39° (SD = 3.09°), 6.51° (SD = 3.12°), 7.04° (SD = 2.70°), and 4.97° (SD = 2.39°) while the mean values for KirBac3.1 S129R are 11.70° (SD = 3.03°), 13.33° (SD = 3.18°), 8.59° (SD = 2.66°), and 13.308° (SD = 3.18°). These results highlight that the presence of the mutation S129R on TM2 has a knock-on effect on TM1, triggering a greater kink of this outer helix as noticed in the cryo-EM analysis (Paynter et al., 2010).

Rotation of the Cytoplasmic Domain

Motions of the cytoplasmic domain that couple ligand binding to the gating of the channel have been thoroughly investigated and various models have been proposed. From the KirBac3.1 S129R mutant's crystallographic data, a model described as "twist to open" has been proposed, on which a rotation of about 25° of the CTD around the central axis of the channel perpendicular to the membrane is crucial to allow gating. This is why we performed a thorough examination of the angle of rotation of each chain's cytoplasmic domain around the central axis on all the relaxed structures (See the definition of the rotation angle **Figure 5A**).

We compared the rotation angles of the cytoplasmic domain of the KirBac3.1 WT and KirBac3.1 S129R, they are given in **Figure 5B**. The mean cytoplasmic domain rotation values are 97.95° (SD = 5.13°) and 131.78° (SD = 5.41°) for WT and mutant, respectively. An average difference of 33° is observed between the two systems. This is to be compared with the data obtained from crystallographic structures, which show a difference of the cytoplasmic domain rotation angle of 30° between the KirBac3.1 WT (closed state) and KirBac3.1 S129R (open state) (Bavro et al., 2012). It can be observed that both experimental and *in silico* studies are very comparable. The structures obtained by MDeNM appear very stable with very small variations.

To study the S129R mutation effect on the interaction between the cytoplasmic domain and the membrane interface, we calculated two distances: 1) between His35 on the slide helices and Arg167 (CD-loop on the CTD); 2) between Pro138 (the linker between the transmembrane and the cytoplasmic domain) and Phe250 (G-loop on the CTD). The G-loop has been described as being mobile during gating in molecular dynamics studies (Bernsteiner et al., 2019; Li et al., 2019) and X-ray data (Pegan et al., 2005). The location of residues is indicated in **Figure 6C**.

The structures resulting from the dynamics highlight the differences between WT and the mutant. **Figure 6A** shows that the interaction between Pro138 and Phe250 (chain B) is stronger in S129R compared to WT. The distances between the amino acids are from 1.91 to 5.06 Å in the S129R mutant (in blue) and from 1.82 to 7.70 Å in KirBac3.1 WT (in red) predominantly closed (93.2% of the structures). The same trend is observed for the chain C, shown in **Figure 6C**, as they do not exceed 2.34 Å while the range of values for WT extends over 9.58 Å (Figure 6B). The distances between the residues His35 Chain B and Phe167 Chain C are also lower in S129R

The S129R mutant shows a stronger interaction between the cytoplasmic domain (G-loop and C loop) and the membrane interface (slide helix and linker) and a greater stability.

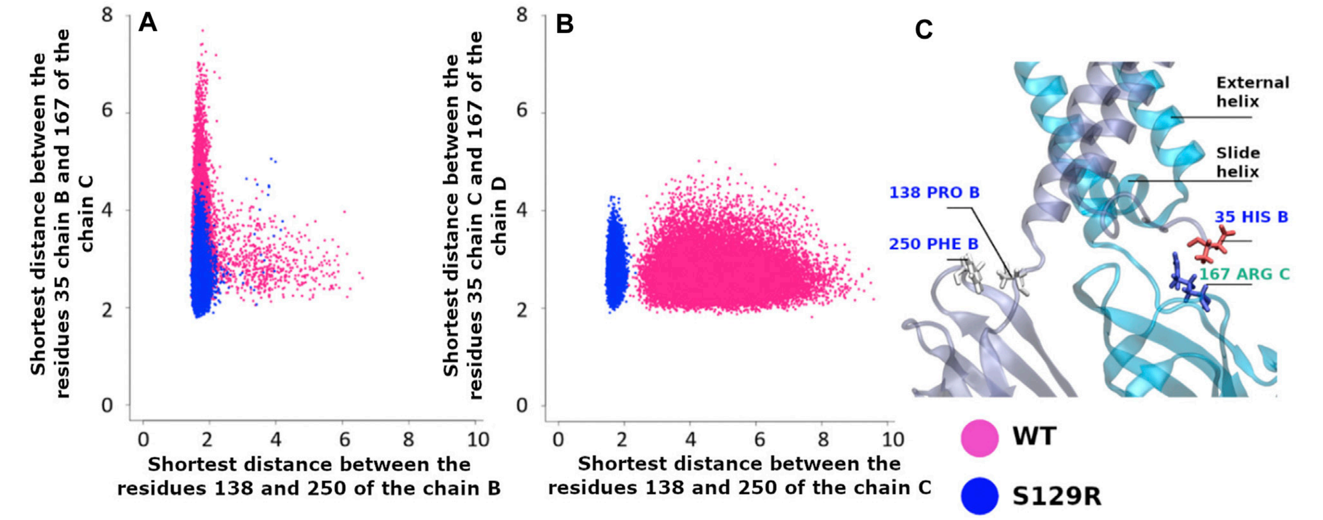


FIGURE 6 | Interaction between the cytoplasmic and the membrane interface. Pink points are for KirBac3.1 WT, blue points for the S129R mutant **(A)** Shortest distance between the slide-helix (residue Asp35 Chain B and the CD-loop (residue Arg167 Chain C) vs. the shortest distance between the linker (residue Pro138 Chain B) and the G-Loop (residue Phe250 Chain B); **(B)** Shortest distance between the slide-helix (residue Asp35 Chain C) and the CD-loop (residue Arg167 Chain D) vs. the distances between the linker (residue Pro138 Chain C) and the G-Loop (residue Phe250 Chain C); **(C)** Location of the interactions.

CONCLUSION

Our MDeNM simulations on S129R show clearly that the mutation leads to a greater opening probability of about 52.8% compared to 6.8% in the WT, which is corroborated by functional data from single channel recordings 44.05 (±2.6% Mean ± S.D. *n* = 16,107) compared with 9.9 (±1.3, *n* = 1803) in the WT. In this study, we observed the opening of the constriction points in the channel and the inherent motions of KirBac3.1 S129R associated with the gating in the absence of K^+ inside the pore. Our *in silico* results, in the estimation of open state population, carried out in the absence of K^+ ions, are very similar with experimental electrophysiological accounting of these ions. The gating is therefore mainly linked to the intrinsic dynamical properties of the channel and not dependent only on the presence of the K^+ ions. The presence of the Arg mutation triggers the release of the two constriction points that existed in WT protein, but at the same time, this residue can block the passage of the K^+ ion through the channel. Indeed, Arg can adopt two conformations, pointing either towards the channel's center or standing parallels to the channel and interacting with Tyr132. This explains why the opening probability is only 52.8% and not 100%. Also, the transmembrane external helices (TM1) show a more pronounced kink and flexibility in the case of the S129R mutant in agreement with our HDX-MS experiments. In addition, the contacts between Pro138 and Phe250 (G-loop), and the inter-chain contact between His35 (slide-helix) and Arg167 (CD-loop) are stabilized, and the mutant shows a greater stability.

One of the conclusions of this study is that care must also be taken when selecting replacement residues since this could affect importantly the structural and dynamical behaviour of the system under consideration as it is the case here.

MATERIALS AND METHODS

This article focuses on understanding the structural behavior of KirBac3.1 S129R, making the comparison with the KirBac3.1 WT necessary to detect its specific aspects. The data and details concerning the latter were described in our previous article (Fagnen et al., 2020).

Protein Expression and Purification

Same construct as for the protein used for the structure determination was used (Bavro et al., 2012). Protein expression and purification of this mutant channel were performed as outlined before (De Zorzi et al., 2013). Briefly, after cell disruption by French press, the protein was solubilized with 45 mM DM (Decyl *β*-D-maltopyranoside), centrifuged, and the supernatant was loaded onto a Co2+ affinity column. The protein was promptly purified on a Superdex 200 column pre-equilibrated with 2 mM TriDM buffer. Concentrated preparations (1–2.5 mg/ml) of purified proteins (>95% purity, judged by SDS-PAGE) were stored at −80°C in a buffer containing 20 mM Tris, pH 7.4, 150 mM KCl and 0.2 mM TriDM.

Pepsin Digestion, Hydrogen/Deuterium Exchange Coupled to Mass Spectrometry and HPLC Peptide Separation and Mass Spectrometry of Peptides.

These experiments were performed as outlined before (Fagnen et al., 2020). Briefly, all protein digestions in solution were performed in an ice bath at 0°C. Protease solutions were prepared in 500 mM glycine (pH 2.2). KirBac3.1 S129R protein samples were digested in the same buffer for 2–5 min using a protease/

substrate ratio of 1:1 or 1:10 (wt/wt) for pepsin and nepenthesin, respectively, either in solution or immobilized on a resin. The increase in digestion time did not affect the proteolysis. HDX-MS reactions were carried out on KirBac3.1 S129R at a protein concentration of about 10 μM. The reaction was initiated by a 10x dilution of the protein samples (10 μl) into a deuterated buffer containing 50 mM KCl and 0.2 mM TriDM. The time course of the HDX was followed over a 20-min period by sequential withdrawing 120 μl of deuterated samples, which were immediately added to 26 μl of quenching buffer (8 M guanidium chloride, 500 mM glycine HCl, pH 2.2), rapidly mixed, and flash-frozen in liquid nitrogen. After protease digestion in solution or on column in an ice bath at 0°C, the peptides were loaded onto a peptideMicroTrap (Michrom Bioresources) column and washed with 0.03% TFA in water (HPLC). They were then separated on a reversed-phase C12 column (1 × 50 mm, Jupiter; Phenomenex) using a linear gradient of 15–45% (vol/vol) of solution B (CH3CN 95% and TFA 0.03%) during 26 min. The tandem MS (mapping) analyzes were performed on an ion trap mass spectrometer (Esquire 3000+; Bruker Daltonics) to identify the peptides after their separation on HPLC. Accurate mass measurements and local deuteration kinetics analysis were performed on a time-of-flight (TOF) mass spectrometer (6210; Agilent Technologies) equipped with an electrospray source. Each deuteration experiment was performed in triplicate. Data were processed as described in (Fagnen et al., 2020).

Electrophysiology

An Orbit mini was used (Nanion, Germany, horizontal planar lipid bilayer system), where two aqueous chambers (150 μl) are separated by a partition with a 150-μm hole where the lipid bilayer is formed by 1,2-diphytanoyl-sn-glycero-3-phosphocholine (DPhPC,15–50 pF). The lower chamber contained 150 mM KCl, 10 mM MOPS, pH 7.4. After membrane bilayer formation, the upper chamber solution was changed to 150 mM KCl, 10 mM MOPS pH 8.1 μl of purified KirBac3.1 S129R (90 μg/ml) in DDM (n-DoDecyl-β-D-maltoside) detergent (0.015%) was added to the upper chamber to a preformed bilayer. Currents were recorded using Elements Data Reader (Nanion, Germany) and analyzed using Clampfit (Axon Instrument Inc., United States) software, sampled at 100 μs and filtered at 1.25 kHz. Recordings were performed at 24°C.

Molecular Modeling

KirBac3.1 was modeled in two different states, the closed one, modeled from the PDB structure of the wild type, 2WLJ (at 2.60 Å atomic resolution) described also previously (Fagnen et al., 2020), and the open one, modeled from the PDB structure 3ZRS (KirBac3.1 S129R at 3.05 Å resolution). KirBac3.1 S129R had missing C and N-terminal fragments, so the modeled structure describes the protein from residue 35 to 295; other missing atoms were rebuilt using the CHARMM program. We applied the $P42_12$ symmetry to the 3ZRS PDB structure using the Protein Interface Surface and Assemblies software (Krissinel and Henrick, 2007) to build the tetramer. We applied the "Orientation of Proteins in Membranes" (OPM) software (Lomize et al., 2012) and the protein was then included into a 114.566 × 114.566 × 128.572 $Å^3$ water box containing 36,629 water molecules with 150 mM KCl, and an 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC) membrane using CHARMM-GUI (Jo et al., 2008; Lee et al., 2016).

Normal Modes

Structures were minimized using the steepest descent (SD) and conjugate-gradient (CG) methods followed by Adopted Basis Newton Raphson (ABNR) algorithm. Harmonic restraints were applied during SD minimization, which were gradually reduced. Then the system was subjected to 50,000 CG steps without restraints followed by ABNR minimization until a convergence of 10^{-5} kcal mol^{-1} $Å^{-1}$ RMS energy gradient was reached. The first 200 lowest frequency normal modes, ranged in ascending order of frequencies, with all the atoms taken into account were computed using the iterative DIMB method (Mouawad and Perahia, 1993; Perahia and Mouawad, 1995) in CHARMM.

It was necessary to select the modes that contribute the most to the channel's conformational changes to proceed with the MDeNM calculations. The channel was defined from residues 121 to 133. For each normal mode, the minimized energy structure was first globally displaced by 2 Å of RMSD. To select them, a process similar to the one used for the KirBac3.1 WT was applied (Fagnen et al., 2020). Firstly, the ten normal modes showing the highest variations at the level of the channel were retained. To evaluate these variations, the distances between the residues 125 of opposite chains were considered. Secondly, only the normal modes describing spherical and elliptical opening of the channel were taken into account. Redundant modes were excluded keeping only the lowest frequency normal modes. This protocol allowed us to select four modes that affect the most the gating of the S129R.

Molecular Dynamics Using Excited Normal Modes

The Molecular Dynamics using excited Normal Modes (MDeNM) method (Costa et al., 2015), promotes large conformational changes while taking into account the coupling with local fluctuations. This method allows a larger exploration of the conformational space and the generation of a wide variety of different structures at a lower computation time, which would not have been possible using only the standard MD. The method consists first to achieve different linear combinations of the selected normal mode vectors such that the combined vectors describe different movements of a given region (here the channel region). They are chosen such that the displaced structures along them up to a given distance (1 Å) display a uniform distribution of local RMSDs between them (see ref (Costa et al., 2015) for more details).

In a second stage these directions are used in MD simulations for defining additional velocities oriented along these very directions, and corresponding to a given kinetic energy, that are added to the current MD velocities. Such a kinetic excitation was periodically repeated at a given period of time (called the relaxation time) for propagating the movement to larger distances and allowing the coupling with local motions. A sufficient number of excitations were applied to reach an energetically acceptable large displacement. The simulations were carried out independently for every NM combined vector, each of these simulations being called a replica simulation. For each replica 10 successive excitations were applied, each one corresponding to a 4 K rise of the overall temperature; the relaxation time between two excitations was 1 ps The numbers of replica used are 62 and 66 for KirBac3.1 WT and KirBac3.1 S129R, respectively.

Relaxation of the MDeNM Structures

Free MD simulations were carried out on MDeNM structures to relax them further energetically and release the excess kinetic energy that would have been accumulated during the excitations. These relaxation simulations were carried out on a limited number of representative structures obtained by clustering the MDeNM structures to save simulation time. The VMD clustering tool (Humphrey et al., 1996) was used to find at least 100 different clusters separated by a distance greater than an RMS threshold of 0.9 Å on the channel (from Met121 to Ala133). A representative structure was chosen for each cluster, which was the closest to the cluster's average structure. Unique structures not belonging to any cluster were also selected. Overall, 114 and 99 clusters were obtained for KirBac3.1 WT and KirBac3.1 S129R, respectively. Each representative structure was subjected to a free MD simulation of 0.4 ns to release the excess kinetic energy and allow local movements to occur, amounting to a total of 39.6 ns for all the KirBac3.1 S129R's structures considered. The simulations were carried out with NAMD v2.10 (Phillips et al., 2005) at constant temperature of 300 K and constant pressure of 1 atm using Langevin piston. Periodic Boundary Conditions and the Particle Mesh Ewald method were used for the electrostatic interactions. The motion propagation is driven by the Leapfrog Verlet algorithm. Concerning the non-bonded interactions, the cut-on and the cut-off were 10 and 12 Å, respectively. Charmm36 force field was used for the simulations. The parameters used were the same for the MDeNM simulations.

Analysis of Molecular Dynamics Simulations.

The results presented in this article are based on the relaxed structures from free MD simulations in which only the last three-quarters of the trajectories were kept, that represents 29,600 structures.

The shortest distance between two residues was calculated considering the distances between all their respective atoms including the hydrogens with the CHARMM software. The shortest distances calculated at the different levels of the channel and different pairs of residues were used to define the various open/closed states. Six shortest distances were considered: 1) between Leu124 of chain A and Leu124 of chain C, 2) between Leu 124 of chain B and Leu124 of chain D, 3) between Tyr132 of chain A and Tyr132 of chain C, 4) between Tyr132 of chain B and Tyr132 of chain D, 5) between Arg129 of chain A and Arg of chain C, 4) between Arg129 of chain B and Arg129 of chain D.

The kink of each of the outer helices is defined by the angle between the axis of the first part of the helix going from Trp46 to Leu56 and that of the second part going from Leu56 to Asp80.

The cytoplasmic domain rotation for each of the chains was calculated as the pseudo bonds' dihedral angle defined by four successive points defined by Leu108, its projection on the central Z-axis passing through the channel, the projection of Ile266 on the same axis, and Ile 266 itself.

The kink and the dihedral angles were calculated using the CHARMM software.

The Root-Mean-Square Deviation of a given atom i ($\mathrm{RMSD_i}$) was computed over the ensemble of all the MDeNM relaxed structures for KirBac3.1 WT (34,086 structures) and KirBac3.1 S129R (29,600 structures), respectively. The $\mathrm{RMSD_i}$ is defined by the equation

$$RMSD_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left|r_i(n) - r_i^{ref}\right|^2}$$

where i is the atom number, N the total number of structures considered, $r_i(n)$ the position of the atom i in the structure n, r_i^{ref} the position of the atom i in the reference structure. The RMSD of a given residue was calculated by averaging the $\mathrm{RMSD_i}$ of atoms belonging to the given residue.

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 authors.

AUTHOR CONTRIBUTIONS

DP, EF, and CV-B conceived and designed the research. CF performed the MDeNM calculations under the supervision of DP; RZ, LB, and IO expressed the protein. LB, IO, and EF performed the Hydrogen Deuterium Exchange Mass spectrometry (HDX-MS) experiments. SB, RS and SG performed single channel investigations. CF, DP, and CV-B wrote the manuscript with the help of comments from all authors.

FUNDING

SERVIER and ANR CACSICE (#11-EQPX-0008) (CV-B); AFM grant (#23207) (SB and CV-B) for consumable for experimental data. PhD research funding: MRT (CF). This work was performed using HPC resources from the "Mésocentre" computing center of CentraleSupélec and ENS Paris-Saclay supported by CNRS and Région Île-de-France (http://mesocentre.centralesupelec.fr/). It was granted access to the HPC resources of IDRIS under the allocation 100512 made by GENCI/IDRIS (DP).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.691901/full#supplementary-material

REFERENCES

Bavro, V. N., De Zorzi, R., Schmidt, M. R., Muniz, J. R. C., Zubcevic, L., Sansom, M. S. P., et al. (2012). Structure of a KirBac Potassium Channel with an Open Bundle Crossing Indicates a Mechanism of Channel Gating. *Nat. Struct. Mol. Biol.* 19, 158–163. doi:10.1038/nsmb.2208

Bernsteiner, H., Zangerl-Plessl, E.-M., Chen, X., and Stary-Weinzinger, A. (2019). Conduction through a Narrow Inward-Rectifier K+ Channel Pore. *J. Gen. Physiol.* 151, 1231–1246. doi:10.1085/jgp.201912359

Bichet, D., Haass, F. A., and Jan, L. Y. (2003). Merging Functional Studies with Structures of Inward-Rectifier K+ Channels. *Nat. Rev. Neurosci.* 4, 957–967. doi:10.1038/nrn1244

Capener, C. E., Proks, P., Ashcroft, F. M., and Sansom, M. S. P. (2003). Filter Flexibility in a Mammalian K Channel: Models and Simulations of Kir6.2 Mutants. *Biophysical J.* 84, 2345–2356. doi:10.1016/S0006-3495(03)75040-1

Cheng, W. W. L., Enkvetchakul, D., and Nichols, C. G. (2009). KirBac1.1: It's an Inward Rectifying Potassium Channel. *J. Gen. Physiol.* 133, 295–305. doi:10.1085/jgp.200810125

Clarke, O. B., Caputo, A. T., Hill, A. P., Vandenberg, J. I., Smith, B. J., and Gulbis, J. M. (2010). Domain Reorientation and Rotation of an Intracellular Assembly Regulate Conduction in Kir Potassium Channels. *Cell.* 141, 1018–1029. doi:10.1016/j.cell.2010.05.003

Costa, M. G. S., Batista, P. R., Bisch, P. M., and Perahia, D. (2015). Exploring Free Energy Landscapes of Large Conformational Changes: Molecular Dynamics with Excited Normal Modes. *J. Chem. Theor. Comput.* 11, 2755–2767. doi:10.1021/acs.jctc.5b00003

De Zorzi, R., Nicholson, W. V., Guigner, J. M., Erne-Brand, F., and Vénien-Bryan, C. (2013). Growth of Large and Highly Ordered 2D Crystals of a K$^+$ Channel, Structural Role of Lipidic Environment. *Biophys. J.* 105, 398–408. doi:10.1016/j.bpj.2013.05.054

Fagnen, C., Bannwarth, L., Oubella, I., Forest, E., De Zorzi, R., de Araujo, A., et al. (2020). New Structural Insights into Kir Channel Gating from Molecular Simulations, HDX-MS and Functional Studies. *Sci. Rep.* 10, 8392. doi:10.1038/s41598-020-65246-z

Forest, E., and Man, P. (2016). "Conformational Dynamics and Interactions of Membrane Proteins by Hydrogen/Deuterium Mass Spectrometry," in *Heterologous Expression of Membrane Proteins: Methods and Protocoles*. Editor I MusVeteau. 2nd edition, 269–279. doi:10.1007/978-1-4939-3637-3_17

Fujiwara, Y., and Kubo, Y. (2006). Functional Roles of Charged Amino Acid Residues on the wall of the Cytoplasmic Pore of Kir2.1. *J. Gen. Physiol.* 127, 401–419. doi:10.1085/jgp.200509434

Gupta, S., Bavro, V. N., D'Mello, R., Tucker, S. J., Vénien-Bryan, C., and Chance, M. R. (2010). Conformational Changes during the Gating of a Potassium Channel Revealed by Structural Mass Spectrometry. *Structure.* 18, 839–846. doi:10.1016/j.str.2010.04.012

Huang, J., and MacKerell, A. D., Jr (2013). CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* 34, 2135–2145. doi:10.1002/jcc.23354

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graphics.* 14, 33–38. doi:10.1016/0263-7855(96)00018-5

Inanobe, A., Nakagawa, A., and Kurachi, Y. (2013). Conformational Changes Underlying Pore Dilation in the Cytoplasmic Domain of Mammalian Inward Rectifier K+ Channels. *PLoS One.* 8, e79844. doi:10.1371/journal.pone.0079844

Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008). CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* 29, 1859–1865. doi:10.1002/jcc.20945

Krissinel, E., and Henrick, K. (2007). Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* 372, 774–797. doi:10.1016/j.jmb.2007.05.022

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. Theor. Comput.* 12, 405–413. doi:10.1021/acs.jctc.5b00935

Li, D., Jin, T., Gazgalis, D., Cui, M., and Logothetis, D. E. (2019). On the Mechanism of GIRK2 Channel Gating by Phosphatidylinositol Bisphosphate, Sodium, and the Gβγ Dimer. *J. Biol. Chem.* 294, 18934–18948. doi:10.1074/jbc.RA119.010047

Li, J., Xie, X., Liu, J., Yu, H., Zhang, S., Zhan, Y., et al. (2014). Lack of Negatively Charged Residues at the External Mouth of Kir2.2 Channels Enable the Voltage-dependent Block by External Mg2+. *PLoS One.* 9, e111372. doi:10.1371/journal.pone.0111372

Linder, T., Wang, S., Zangerl-Plessl, E.-M., Nichols, C. G., and Stary-Weinzinger, A. (2015). Molecular Dynamics Simulations of KirBac1.1 Mutants Reveal Global Gating Changes of Kir Channels. *J. Chem. Inf. Model.* 55, 814–822. doi:10.1021/acs.jcim.5b00010

Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I., and Lomize, A. L. (2012). OPM Database and PPM Web Server: Resources for Positioning of Proteins in Membranes. *Nucleic Acids Res.* 40, D370–D376. doi:10.1093/nar/gkr703

Mouawad, L., and Perahia, D. (1993). Diagonalization in a Mixed Basis: A Method to Compute Low-Frequency normal Modes for Large Macromolecules. *Biopolymers.* 33, 599–611. doi:10.1002/bip.360330409

Nishida, M., Cadene, M., Chait, B. T., and Mackinnon, R. (2007). Crystal Structure of a Kir3.1-prokaryotic Kir Channel Chimera. *Embo J.* 26, 4005–4015. doi:10.1038/sj.emboj.7601828

Paynter, J. J., Andres-Enguix, I., Fowler, P. W., Tottey, S., Cheng, W., Enkvetchakul, D., et al. (2010). Functional Complementation and Genetic Deletion Studies of KirBac Channels. *J. Biol. Chem.* 285, 40754–40761. doi:10.1074/jbc.M110.175687

Pednekar, D., Tendulkar, A., and Durani, S. (2009). Electrostatics-defying Interaction between Arginine Termini as a Thermodynamic Driving Force in Protein-Protein Interaction. *Proteins.* 74, 155–163. doi:10.1002/prot.22142

Pegan, S., Arrabit, C., Zhou, W., Kwiatkowski, W., Collins, A., Slesinger, P. A., et al. (2005). Cytoplasmic Domain Structures of Kir2.1 and Kir3.1 Show Sites for Modulating Gating and Rectification. *Nat. Neurosci.* 8, 279–287. doi:10.1038/nn1411

Perahia, D., and Mouawad, L. (1995). Computation of Low-Frequency normal Modes in Macromolecules: Improvements to the Method of Diagonalization in a Mixed Basis and Application to Hemoglobin. *Comput. Chem.* 19, 241–246. doi:10.1016/0097-8485(95)00011-g

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* 26, 1781–1802. doi:10.1002/jcc.20289

Wang, S., Borschel, W. F., Heyman, S., Hsu, P., and Nichols, C. G. (2017). Conformational Changes at Cytoplasmic Intersubunit Interactions Control Kir Channel Gating. *J. Biol. Chem.* 292, 10087–10096. doi:10.1074/jbc.M117.785154

Wang, S., Lee, S.-J., Heyman, S., Enkvetchakul, D., and Nichols, C. G. (2012). Structural Rearrangements Underlying Ligand-Gating in Kir Channels. *Nat. Commun.* 3, 617. doi:10.1038/ncomms1625

Zubcevic, L., Bavro, V. N., Muniz, J. R. C., Schmidt, M. R., Wang, S., De Zorzi, R., et al. (2014). Control of KirBac3.1 Potassium Channel Gating at the Interface between Cytoplasmic Domains. *J. Biol. Chem.* 289, 143–151. doi:10.1074/jbc.M113.501833

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.

frontiers in Molecular Biosciences

REVIEW
published: 09 July 2021
doi: 10.3389/fmolb.2021.706002

Protein Surface Interactions—Theoretical and Experimental Studies

Fabio C. L. Almeida [1,2]*, *Karoline Sanches* [1,2,3], *Ramon Pinheiro-Aguiar* [1,2], *Vitor S. Almeida* [1,2] *and Icaro P. Caruso* [1,2,3]*

[1] *Institute of Medical Biochemistry—IBqM, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil,* [2] *National Center for Structural Biology and Bioimaging (CENABIO)/National Center for Nuclear Magnetic Resonance (CNRMN), Federal University of Rio de Janeiro, Rio de Janeiro, Brazil,* [3] *Multiuser Center for Biomolecular Innovation (CMIB), Institute of Biosciences, Letters and Exact Sciences (IBILCE), São Paulo State University "Júlio de Mesquita Filho" (UNESP), São Paulo, Brazil*

In this review, we briefly describe a theoretical discussion of protein folding, presenting the relative contribution of the hydrophobic effect versus the stabilization of proteins *via* direct surface forces that sometimes may be overlooked. We present NMR-based studies showing the stability of proteins lacking a hydrophobic core which in turn present hydrophobic surface clusters, such as plant defensins. Protein dynamics measurements by NMR are the key feature to understand these dynamic surface clusters. We contextualize the measurement of protein dynamics by nuclear relaxation and the information available at protein surfaces and water cavities. We also discuss the presence of hydrophobic surface clusters in multidomain proteins and their participation in transient interactions which may regulate the function of these proteins. In the end, we discuss how surface interaction regulates the reactivity of certain protein post-translational modifications, such as S-nitrosation.

Keywords: surface, solvation, clusters, interdomain, NMR, dynamics, hydrophobic surface clusters

OPEN ACCESS

Edited by:
David Perahia,
UMR8113 Laboratoire de biologie et pharmacologie appliquée (LBPA), France

Reviewed by:
Maximo Sanz-Hernandez,
Imperial College London, United Kingdom
Carmelo Corsaro,
University of Messina, Italy

****Correspondence:***
Fabio C. L. Almeida
falmeida@bioqmed.ufrj.br
Icaro P. Caruso
icaro.caruso@unesp.br

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *06 May 2021*
Accepted: *29 June 2021*
Published: *09 July 2021*

Citation:
Almeida FCL, Sanches K, Pinheiro-Aguiar R, Almeida VS and Caruso IP (2021) Protein Surface Interactions—Theoretical and Experimental Studies. Front. Mol. Biosci. 8:706002. doi: 10.3389/fmolb.2021.706002

INTRODUCTION

Since the Anfinsen (Anfinsen, 1972) discovery that after chemical denaturation proteins can be refolded *in vitro*, the thermodynamic hypothesis became the most accepted model for protein folding. The question of what are the major forces that stabilize the protein and control the folding pathway arose as fundamental. The balance between hydrophobicity and hydrophilicity is crucial for protein folding and also for its structural property. This balance is known to be essential for the folding of globular proteins. While in intrinsically disordered proteins (IDPs) the balance is shifted toward the hydrophilicity, displaying an excess of hydrophilic residues, for globular proteins, there is a hydrophobic collapse mediated by a solvent entropic effect. The entropy increases significantly when a hydrophobic solute is transferred from an aqueous solution to a nonpolar environment, such as the protein core. The explanation (Kauzmann, 1959; Baldwin and Rose, 2016) for the entropy penalty is in the Gibbs free energy (ΔG°), enthalpy (ΔH°), and entropy (ΔS°) change for the hydration of hydrophobic solutes. For aqueous alkane, the enthalpic contribution is negative and favorable, while the entropic contribution is largely negative and unfavorable. The original explanation was on the rigidity of hydrocarbon hydration shells made of clathrate water (Kauzmann, 1959). New experiments suggest that the hydration shell made of clathrate water does not explain completely the entropic penalty. The best description is a hydration shell formed by Van der Waals (VDW) attraction, which also leads to increased rigidity of the water molecules (Baldwin, 2014; Baldwin and

Rose, 2016). Until now we poorly understand the solvation of hydrophobic residues. Baldwin and Rose state that this limitation precludes our ability to predict the free-energy values that can drive the folding process (Baldwin and Rose, 2016). In this review, we do not intend to deeply discuss the physical chemistry of protein solvation. We start analyzing the importance of solvation in proteins without a typical hydrophobic core (named core-less proteins) (Machado et al., 2018a; Pinheiro-Aguiar et al., 2020). These proteins form hydrophobic clusters at the surface that need to be better understood. They are an unexplored avenue to the understanding of the solvation of exposed hydrophobic residues. We extend the analysis showing that the surface hydrophobic clusters are present in globular proteins and have an important impact on transient interactions in multi-domain proteins (Pinheiro et al., 2019) and may have an impact on the reactivity of residues participating in post-translational modifications.

Proteins are large molecules, imposing an intrinsic large number of degrees of freedom on the polypeptide chain and an astronomical number of possible conformations. If a protein had to sequentially sample among all possible conformations, it would take a time longer than the age of the Universe to access its native folded conformation. The Levinthal paradox (Levin, 2004) is based on the fact that, despite all that, proteins fold spontaneously on milliseconds to seconds time-scale. Computational calculations show that polypeptide chains evolved to reach the minimum energy conformational state by shaping kinetic favorable pathways in the energy landscape (Bryngelson et al., 1995; Ferreiro et al., 2014). The evolutionary shaping of the energy funnel is mediated by the balance of forces at each conformational state, in the local and global minima and its transition states (Wolynes, 2015).

As described earlier, water plays an essential role in determining protein stability and the kinetic folding pathways. It is believed that hydrophobic collapse is the dominant effect that drives protein folding. Nevertheless, hydrophilic residues have also an important contribution to protein stability and folding pathways (Durell and Ben-Naim, 2017). It is well known that waters interact with proteins in different regimes. While bulk water has high degrees of freedom when compared to the freedom of protein dihedrals, there are tightly bound waters, which bind to the protein as if it were part of it, having similar freedom as the protein chains. There is an intermediate regime, with water molecules that are part of the solvation shell, which interacts transiently with the protein surface contributing to the protein stability and modulating the interaction of protein itself or other ligands (Fernández and Scheraga, 2003; Papoian et al., 2003). Water molecules have been shown to participate in protein-protein interaction, bridging the protein interfaces.

The solvation of the hydrophilic and hydrophobic residues involved in the protein surface forces is the main topic of this review. (Ben Naim, 2013; Durell and Ben-Naim, 2017) used computational simulations to predict the attractive and repulsive potential of mean force (PMF) of the hydrophilic and hydrophobic interactions that contribute to protein folding. Their data suggest that hydrophilic surface forces may have an underestimated contribution to the overall structure stabilization. First, they considered the forces and potentials in nonpolar solutes, demonstrating that the interactions between hydrophobic solutes in water are mediated by direct and solvent-induced forces. Repulsive forces occur because of the steric reorganization and disruption of the water molecules around the solute. Attractive direct and solvent induced PMF contribute to the interaction of two hydrophobic molecules in water. The simulations suggest that the solvent-induced attractive force exerts a higher contribution than the direct forces for methane-methane stabilization. For larger alkanes, the solute induced PMF contributes approximately equally to the direct force. In all simulated situations, the water solvation of hydrophobic molecules contributes significantly to the total attractive force.

On the surface of proteins and other biomolecules, the role of the solvent-induced attractive force may be underestimated. Feng and colleagues (Feng et al., 2019) demonstrated the importance of this effect by investigating the effect of the hydrophobic interaction on the stability of the DNA duplex. They measured the effect of polyethylene glycol 400 (PEG400) in the base-pairing stability. Using atomic force microscopy, they verified a decrease in the DNA stretching force due to a decrease in base-stacking energy caused by transient fluctuations of the bases. The fluctuations, referred to as longitudinal breathing, involve the disruption of the hydration shell of the DNA, leading to increased hydration in the interior of the DNA and a consequent decrease in hydrogen bond energy. In the presence of abundant water, the intact hydration shell promotes a solvent-induced hydrophobic interaction among the nitrogenous bases, the DNA interior gets dryer, allowing hydrogen bonds to occur in an ideal geometry. This important finding reveals that, instead of the base-pairing hydrogen bonds being the main responsible for the DNA stabilization, it involves the coin-pile stacking of base pairs due to the hydrophobic effect. The hydration shell shields the DNA, making its interior dry. The disruption of the hydration shell decreases the solvent-induced hydrophobic forces among the bases, causing fluctuation of the bases, promoting the formation of holes in the DNA interior and a consequent weakening of the Watson-Crick hydrogen bonds.

For hydrophilic groups interacting in an aqueous solution, the solvent-induced forces contribute even more significantly. For the simulation with two hydroxyls, the direct force is repulsive while the solvent-induced force is attractive by the formation of a water molecule bridging the two hydroxyls. For the simulations with 3 and 4 hydroxyls, the solvent-induced attractive force is even bigger, and the direct forces are also attractive. In the surface of a globular protein, the probability of finding clusters of hydrophilic side chains is high, and thus there is a good chance of having water molecules playing the role of bridging side-chains, contributing significantly to the stabilization of the overall fold. In bulk water, the exposed hydrophilic side-chains are likely to be hydrated, making it more likely to have bridging water than direct hydrogen bonds (Durell and Ben-Naim, 2017).

The contribution of hydrogen bonds is pivotal for the stabilization of the secondary structure elements. However, it is seen as small for the tertiary and quaternary structure stabilization, due to the presence in similar amounts in the

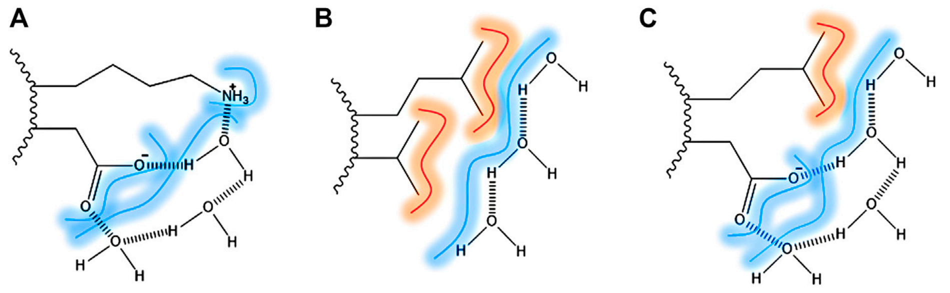


FIGURE 1 | Surface direct forces: amino acid residues exposed to the surface bridged by hydrogen bonds and VDW interaction with water (dash). **(A)** Lysine and aspartic acid as an example of two polar amino-acid side chains bridged by water molecules; **(B)** Leucine and alanine side-chains as an example of two apolar amino-acid side chains and the VDW interactions; **(C)** Leucine and aspartic acid as an example of apolar and polar amino-acid side chains.

folded and unfolded state. The contribution of water in protein stabilization by forming bridges between protein surface and hydrophilic residues has been raised as significant by many authors. In **Figure 1**, it is shown the water bridging at the protein surface (Ben-Naim, 2013; Durell and Ben-Naim, 2017). In **Figure 1A**, the water is bridging two polar side-chains through a hydrogen bond. In **Figure 1B**, water is shielding two hydrophobic residues exposed by the solvent, through hydrogen bonds among water molecules and VDW interactions between water and the aliphatic chains. In **Figure 1C**, the water is bridging hydrophobic and hydrophilic residues.

ARE SOLVENT-INDUCED SURFACE FORCES UNDERESTIMATED?

There is not a clear answer to this question. As briefly described earlier, there is a vast literature, using mainly computational prediction, which suggests that the answer is yes. It is not our intention to give a final answer but rather discuss experimental results in which these hydrophilic surface forces play important roles. We will present experimental results of i) protein dynamics in water cavities, ii) the structure and dynamics of core-less proteins, such as plant defensins and toxins, iii) the importance of surface forces in transient interactions in the modulation of multi-domain proteins dynamics and iv) how these forces may modulate reactivity at the surface.

Methods Available to Study Protein Solvation

One of the methods to study protein solvation is x-ray crystallography, where immobile and symmetric crystal waters can be directly observed in the electron density map. This is important to give information on tightly bound waters, which frequently are involved in catalysis and biological function.

The Nuclear Magnetic Resonance (NMR) spectroscopy can also be used, allowing the measurements of protein hydration in solution. The residence time obtained for the interior water molecules is in the range of 10^{-8} to 10^{-2} s, while surface water molecules are in the sub nanoseconds timescale. For this, NMR is a powerful technique to obtain information on the location and the residence time of individual hydration water molecules, being able to distinguish between tightly bound and unbound waters from the solvation layer. As mentioned before, it is an important mechanism involved in protein stability.

The method is based on the dipolar coupling between the hydrogen nuclei spin of water and hydrogens of the protein. This is achieved by measuring the nuclear Overhauser effect at the laboratory frame (NOE) and the rotating frame (ROE). Otting and collaborators (Otting et al., 1991b) have shown the difference in the residence time between water molecules in the protein interior as well as in the protein surface and compared with the water molecules observed in the corresponding crystal structure. When short mixing times are used in the NOESY/ROESY experiments, the contributions from auto relaxation and spin diffusion are minimal. As a result, the NOE/ROE intensities are almost exclusively from the contribution of cross-relaxation between water-protein nuclear spin. The authors constructed protein hydration models to verify how the sign and value of the NOE/ROE intensity ratio inform about the residence time and location. A negative NOE/ROE intensity ratio reveals dipolar interaction between the water and the protein residue and the presence of motional retarded water molecules.

It was verified later that non-local water molecules in fast exchange with the interacting water molecule also contribute to the negative ratio (Modig et al., 2004). Therefore, the presence of a negative ratio is not in itself conclusive regarding the presence of local tightly bound water. Thousands of non-local water molecules contribute to the negative NOE/ROE intensity ratio. The non-local effect of this method can be overcome by restricting the water mobility and exchange or the amount of water. This is the case of the water cavity of thioredoxins, where tightly bound water is found in the water cavity, hydrogen-bonded to the buried aspartic acid and these internal water molecules are motionally retarded (Cruzeiro-Silva et al., 2014; Iqbal et al., 2015).

Solvation Studies in Reverse Micelles

The measurement of protein hydration is not an easy task, once the quantity of water molecules in the solvent is extensive, and because of its fast motion and exchange. NMR experiments may provide the residence time and locality of water interactions through dipolar magnetization exchange of the hydrogens from protein-water interactions. By comparing the nuclear Overhauser effect at the NOE and ROE it can be verified the contributions of dipolar and chemical exchanges (Otting et al., 1991b). However, when the protein is embedded in bulk water there is a non-local effect that contributes to the intensities, leading to misleading data (Modig et al., 2004). Also, hydrogens exchange with the hydrogens from water solvent can make the analyses confusing to be evaluated.

To minimize the difficulty in measure protein hydration by NOE and ROE *via* NMR experiments, Wang has proposed the idea of protein encapsulation through a reverse micelle. It provides a reduction of water molecules in the system, which consequently leads to a decrease in hydrogens exchange and the non-locality artifact. It is also possible to regulate the protein tumbling time using the solvent solution with low viscosity to improve the relaxation parameter and the signal/noise ratio.

Wang et al. used ^{15}N, ^{2}H-ubiquitin encapsulated in reverse micelles in propane solution. It is remarkable the difference between the two spectra, ubiquitin in aqueous solution and the reverse micelle. The non-local effect and the long-range coupling to water were not present in the spectra, but just NOEs and ROEs with NOE distance from the surface were in the ^{1}H planes of ^{13}C-resolved NOESY and ROESY, which validating the method. The molecules of water in the protein interior have a residence time from about 10^{-8} to 10^{-2} s while the water molecules from the protein surface hydration in solution present residence time of sub-nanosecond (Otting et al., 1991a). By the $\sigma_{\mathrm{NOE}}/\sigma_{\mathrm{ROE}}$ ratio, it was verified the interaction between the water and the protein, where the ratio goes from 0 to −0.5 (small and high residence time, respectively). The ratio near -0.5 refers to protein-water interaction through the rotational correlation time and the ratio of 0 to interactions with times smaller than the rotational correlation time of the protein.

The NOE and ROE measurements showed to be a powerful method to verify protein hydration through NMR. Protein hydration has been studied because of its role in macromolecule functions, as in the case of ubiquitin which is an important protein involved in interactions that regulated the protein degradation inside the cell. It has also been reported the involvement of water molecules in protein-protein interactions and the association of protein subunits.

Core-Less Proteins–A Particular View at the Dynamics of Plant Defensins

There are a vast number of proteins that are stabilized by an extensive number of disulfide bridges. Plant defensins are a good example of these proteins and they share the same cysteine stabilized *αβ* fold (CSαβ) (Thomma et al., 2003; Thevissen et al., 2004). Their primary sequence is diverse, where the cysteines are the only conserved residues required to maintain the CSαβ fold (Gachomo et al., 2012). Similar cysteine stabilized folds are also known such as toxins, channel blockers, disintegrins, and many enzyme inhibitors (Yount and Yeaman, 2004). Different from typical globular proteins, they are core-less globular proteins. They lack a typical hydrophobic core, and instead, they are stabilized by disulfide bonds and the contacts between surface-exposed hydrophobic and hydrophilic residues. Most of the hydrophobic residues are exposed to the protein surface, and yet plant defensins are soluble and monomeric. Surface forces, along with the covalent disulfide bonds are the main forces that stabilize the CSαβ.

Because defensins have a high number of disulfide bonds, it was implicit that they would possess restricted backbone mobility. Remarkably, measurements of nuclear spin relaxation by NMR revealed extensive conformational dynamics in regions of the protein that are dominated by exposed hydrophobic residues. For instance, the Psd1 plant defensin exhibits millisecond time scale dynamics in the $\beta1/\alpha1$ loop (Medeiros et al., 2010), which forms the membrane recognition site. Almost all hydrophobic side chains in Psd1 are exposed at this surface patch (**Figures 2A,B**). Similarly, Sd5 defensin undergoes millisecond dynamics involving all secondary structure elements (de Paula et al., 2011; Machado et al., 2018b). For this defensin, hydrophobic residues are exposed all over the protein surface (**Figures 2E,F**). Another example of this feature is found in Psd2, where the presence of conformational exchange is correlated with the exposure of hydrophobic residues at the protein surface (**Figures 2C,D**).

A closer look at the structure in the solution of plant defensins, along with the analysis of the contact map provided by the nuclear spin dipolar interactions (NOEs) used in the structure calculation, points toward the formation of surface clusters. Interestingly, the exposed hydrophobic residues form dynamic hydrophobic/hydrophilic surface clusters containing long and linear extended polar side chains, such as arginine, lysine, and glutamate. Prolines are also present in such clusters, conferring some hydrophilic/hydrophobic balance to the surface-clusters. These hydrophobic surface clusters showed in **Figure 2** for the plant defensin Psd1, Psd2, and Sd5 (Machado et al., 2018c), align with the idea that water solvation can bridge surface polar and apolar side chains (Ben-Naim, 2013, 2014), acting as a surface direct stabilizing force (**Figure 1**).

For plant defensins, long polar side chains interact with the surface hydrophobic amino acid side chains, minimizing their exposure to the solvent. It protects the hydrophobic amino acids from complete exposure to the solvent, making cysteine stabilized folds water-soluble and monomeric.

Measurements of protein dynamics by NMR, verifying the presence of conformational exchange at the hydrophobic surface clusters, are a powerful tool to study these clusters. The measurement of ^{15}N relaxation parameters (R1, R2, and hetero-nuclear NOE) enables the mapping of the residues in thermal conformational flexibility (pico- to nanosecond timescale dynamics) and in conformational exchange (micro- to millisecond dynamics) at the core-less proteins, leading to the description of the backbone dynamics. The measurement of CPMG relaxation dispersion profiles in several temperatures enables the quantification of the micro- to millisecond conformational equilibrium and to understand the structural property of the first thermally accessible high energy conformational state. For Sd5 (Machado et al., 2018c), it

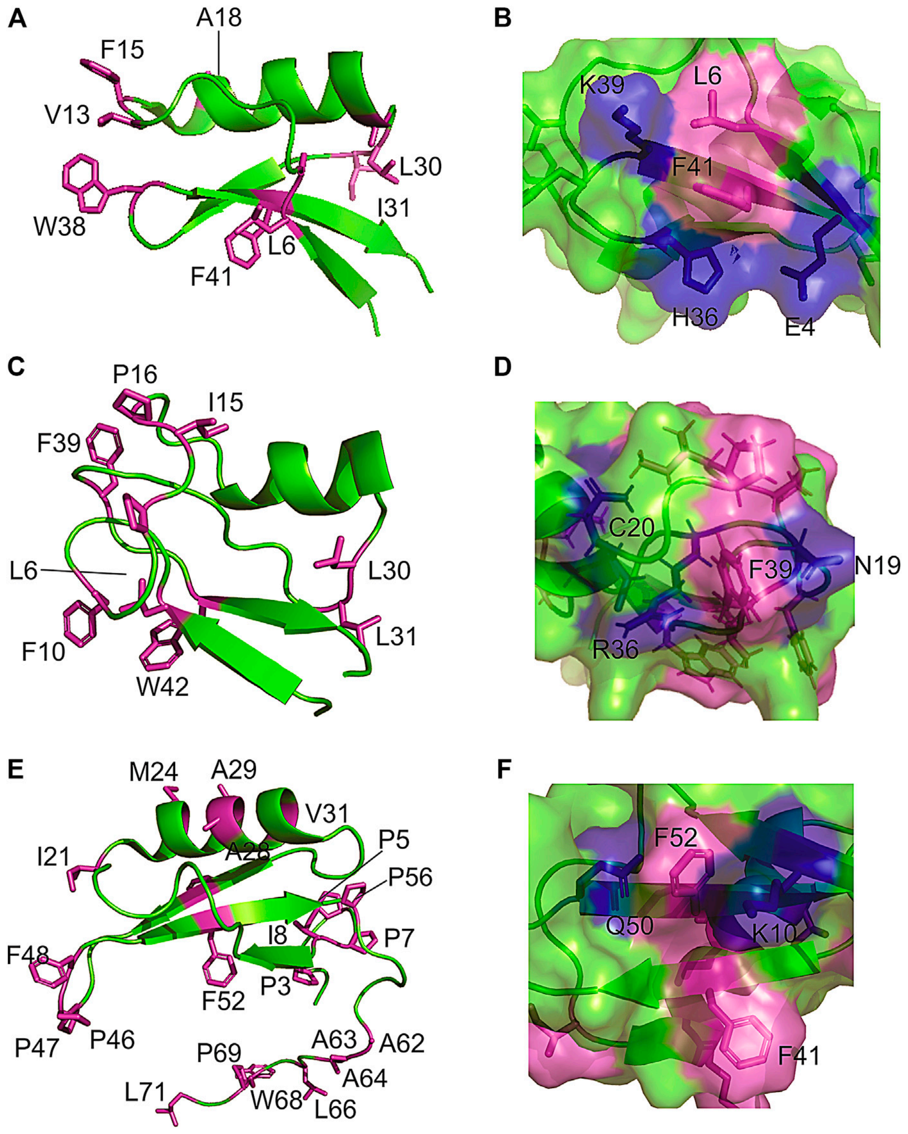


FIGURE 2 | Examples of hydrophobic residues exposing to solvent forming a hydrophobic/hydrophilic surface cluster in plant defensins. The extended and hydrophilic side chains of some amino acids are showing to protect the surface patch in **(A,B)** plant defensins Psd1; **(C,D)** Psd2; and **(E,F)** Sd5.

suggested that the most important changes of the high-energy conformational state are within the α-helix, dismantling some of the hydrophobic surface clusters. Remarkably, this high-energy state is more compact. The protein dynamic is the key factor to fully understand the complexity of the intrinsic structural behavior of the hydrophobic surface clusters and to measure their contribution to the stability of cysteine stabilized fold. Further studies are necessary.

Transient Surface Interactions in Multidomain Proteins

Surface clusters are not only present in core-less proteins, they are also present in globular proteins and may have a fundamental role in regulating transient interaction in multi-domain proteins (Pinheiro et al., 2019). These transient interactions are difficult to measure and are important to regulate inter-domain motion, which ultimately controls the activity of multi-domain proteins (**Figure 3A**). Hydration has an important role in mediating these interactions. The force of hydrogen bonds is regulated by the access to water. Fernández and Scheraga (2003) described that the wrapping of the backbone within the protein structure dehydrates and strengthen the hydrogen bonds, while insufficiently dehydrated hydrogen bonds created by unwrapping cause by packing defects make a "sticky" surface and prompt to bind to other sites (Fernández and Scheraga, 2003). Corroborating this idea, binding sites are often involved in conformational exchange (Valente et al., 2006) possibly due to the packing defects, promoting an increase in water access and exposure of hydrophobic surfaces (**Figure 3A**). Water also mediates

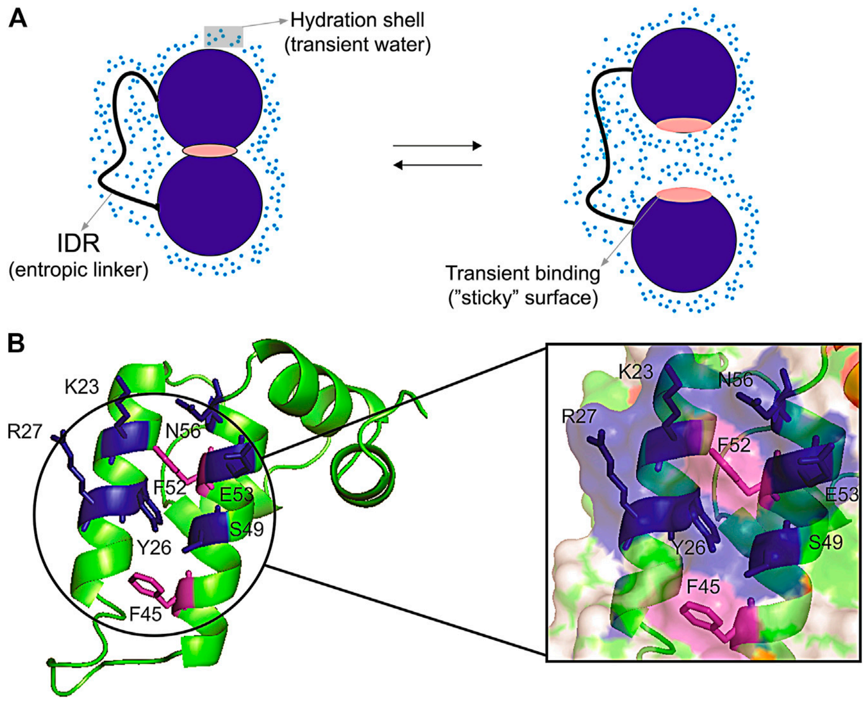


FIGURE 3 | Representation of transient interdomain interactions and surface cluster patch. **(A)** Scheme of a hypothetic open/close equilibrium being regulated through inter-domain transient interaction. The two domains (blue circumference) linked by an intrinsically disordered region (IDR) are interacting in a close conformation through the transient binding of a "sticky" surface (pink), and the solvent-induced interaction in an open conformation. The transient binding sites ("sticky surfaces") can be formed by defective packing, which may lead to insufficiently dehydrated hydrogen bonds and exposed hydrophobic residues. The blue dots denote the hydration shell formed by transient water molecules; **(B)** The Sis1 J-domain showing the surface patch with exposed hydrophobic residues (purple) protected by polar side chains (blue). This surface patch presents transient inter-domain interactions which are pivotal for protein recognition (Pinheiro et al., 2019).

protein-protein interfaces, bridging the protein interfaces (Papoian et al., 2003). It has also been proposed that small organic molecular (osmolytes) that can accumulate in the cells modulate thermodynamic stability or proteins, enzyme activity, and protein oligomerization (Rumjanek, 2018). The addition of cosolvents (osmolytes) changes the hydration shell of proteins, modulating the hydrophobic effect (Van Der Vegt and Nayar, 2017). Changes in hydrations also promote allosteric effects, which as measured for hemoglobin (Colombo et al., 1992). Consequently, the hydrophobic surface clusters observed in the core-less protein have the potential to modulate intermolecular and inter-domain interactions.

Multi-domain proteins are abundant in eukaryotic genomes and are advantageous to accelerate the search for cellular targets (Madan et al., 2011; Zmasek and Godzik, 2012). The domains are defined as an evolutionary and independent unit that can be part of a multi-domain protein or even be an independent single protein (Vogel et al., 2004). The domain function can vary, displaying different catalytic activity, cofactor binding, carry protein-protein recognition motifs, and more (Vogel et al., 2004). In evolution, novel domain arrangements are formed, increasing diversity and performance. The creation of new domain architectures has high adaptive potential through evolution and a relevant functional role (GUON and CHUNG, 2016). The inter-domain dynamics in multi-domain and multicomponent proteins is a key feature that determines protein functionality (Valente et al., 2006). The interdomain dynamics are still poorly known, mainly because it involves the understanding, among other features, of the thermodynamic role of the flexible linkers (intrinsically disordered region–IDR), known to contribute to the entropy of binding and allosteric events (entropic linkers) (Wright and Dyson, 2014; Li et al., 2018). It also involves the understanding of the structural features of the linkers and inter-domain surface interaction patches (Zhu et al., 2000; Wriggers et al., 2005). There are only a few examples in the literature that shows that the inter-domain transient interactions, along with the flexible linkers are the key elements that regulate the flexibility of a multi-domain and ultimately the protein function. In this review, we will show some examples of proteins where the flexibility of the domains has been recognized and where the surface interaction patches were described.

A reported example is the inter-domain contacts and stability of Serralysin protease from *Serratia marcescens* (Zhang et al., 2015). This protease family is known to be involved in pneumonia, empyema, urinary tract infection, and more. It presents high similarity, being composed of an N-terminal helix, a protease domain with an active-site motif (HEXXHXXGXXH), and an Asp/Gly-rich Repeats-in-ToXin (RTX) domain in the C-terminal portion. The folding of serralysin protease is regulated through the binding of Ca^{2+} at the C-terminal domain, and interestingly the folding hyper-stabilization is due to domain-domain interactions between the N and C-terminal of RTX protease (Zhang et al., 2015).

Another interesting example is in the Hsp40 co-chaperones family, for which the surface inter-domain interaction patch was recently described and mapped. It regulates the transient surface interactions of the J-domain and shows to be important for protein functions. The importance of the co-chaperones family is widely known in the literature for its activity in protein quality control in the proteostasis system (Summers et al., 2009; Kampinga and Craig, 2010; Cyr and Ramos, 2015). The Sis1 protein, an Hsp40, is composed of an N-terminal J-domain followed by a glycine-rich flexible linker containing G/F and G/M, and a C-terminal domain that contains the dimerization interface. Recently, Pinheiro et al. (2019) showed the pivotal role of transient inter-domain interactions of Sis1 protein. The comparison of the solution structure of the Sis1 J-domain with the full-length protein and its interaction with Hsp70, essential for the delivery of the client protein, revealed a surface interaction patch composed of hydrophobic and positive residues in the helix II and III (**Figure 3B**). The patch mediates internal transient interactions in the full-length Sis1 and the interaction with Hsp70. The patch is formed by residues V2, T39, F52, D9, R27, and R73 and resembles the hydrophobic surface clusters described for the plant defensins Sd5, Psd1, and Psd2 (Machado et al., 2018c), in which the exposed hydrophobic residues are protected by polar side chains (**Figure 3B**). These transient inter-domain interactions mediated by the hydrophobic surface cluster are favorable for the Hsp40 and Hsp70 interaction, being pivotal for the delivery of the client protein (Pinheiro et al., 2019).

An important feature in multi-domain architecture is the presence of IDRs, which work as flexible linkers, modulating allosteric events and the kinetics of target recognition (Vuzman and Levy, 2012; Li et al., 2018). The growth factor receptor-bound protein 2 (Grb2) presents a high quantity of loops which are pivotal for the protein dynamics and enable its necessary plasticity to bind to multiple cellular targets (Yuzawa et al., 2001; Sanches et al., 2019, 2021). Grb2 presents an equilibrium between monomeric and dimeric states, which is fundamental for the activation and regulation of signaling pathways (Yuzawa et al., 2001). The flexibility of the monomer of Grb2 is pivotal for the protein function in the recognition of cellular partners (Yuzawa et al., 2001). The SH2 domain of Grb2 comprises a unique dynamic behavior involving two independent subdomains. Subdomain I, responsible for the direct recognition of the phosphotyrosine, is in the fast-exchange regime and subdomain II, is the phosphotyrosine +2 residues specificity pocket is in the intermediate exchange regime (Sanches et al., 2020). This fascinating dancing behavior found in each protein is fundamental for molecular recognition. Further studies are necessary to fully understand the correlation between the inter-domain dynamics and protein function.

Nowadays, the interest in studying and understanding the way of working intrinsically disordered proteins (IDPs) has been increasing. Due to the intrinsic hydrophilicity, IDPs and IDRs are highly hydrated, but they do not behave as a random coil. They are intrinsically disordered, but the remaining order is important for the protein functions, especially when acting as entropic linkers in multi-domain proteins (Wright and Dyson, 2014). The IDPs and multidomain proteins are present in protein-protein interactions in many important biological systems. For example, it is reported that the intrinsically disordered region adopt a folded structure upon binding with its respective partner, or even with high concentrations of osmolytes in cellular stress condition (Rumjanek, 2018). Recently, Borgia and colleagues showed a disordered protein-protein interaction with physiological importance (Borgia et al., 2018). They investigated an interaction between the histone H1 linker, a largely unstructured and positively charged protein, known to be present in chromatin condensation, and the ProTα nuclear protein. The ProTα is fully disordered, negatively charged, and has an important participation in chromatin remodeling, transcription, cellular proliferation, and apoptosis. When interacting, H1 and ProTα remain unstructured, and the binding is driven by the large opposite net charge of the proteins. The interaction between H1 and ProTα represents significant evidence that a high-affinity interaction can occur even in the absence of a well-defined binding site. The presence of electrostatic interactions was demonstrated to be enough for the complex stabilization. It is an excellent example of the importance of expanding our studying and understands of the IDP's functionality.

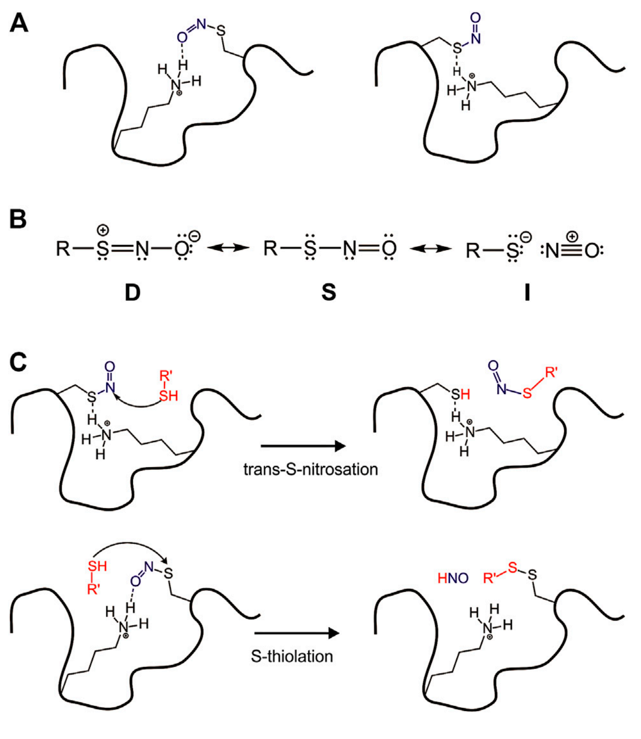


FIGURE 4 | Surface effect on the reactivity of S-nitrosation site. **(A)** Surface interactions modulating–SNO group in the protein; **(B)** RSNO as three resonance structures; **(C)** The protein environment driving nucleophilic attack in different positions in the RSNO.

Surface Effect on the Reactivity of Certain Protein Post-Translational Modifications, Such as S-Nitrosation

Nitric oxide (NO) is a free radical known as an important second messenger for signal transduction in cells. In mammals, three nitric oxide synthase isoforms are responsible for NO synthesis: neuronal (nNOS), inducible (iNOS), and endothelial (eNOS) (Smith and Marletta, 2012)· Once formed, NO reacts with the cysteine side chain as well as peptides and proteins that possess cysteine residues in their sequence, forming a posttranslational modification named S-nitrosation. A range of diseases including Parkinson's, Alzheimer's, heart failure, arrhythmia, diabetes (type I and type II), asthma, and cancer correlates with dysregulated S-nitrosation of proteins (Anand and Stamler, 2012). Due to dynamics and reversible features, S-nitrosation of protein is also implicated in regulating enzymatic activity, protein stability, subcellular localization, and protein-protein interaction (Morris et al., 2016). Than, we want to discuss how S-nitrosation is a function of the surface interactions (**Figure 4A**).

In the cell, S-nitrosothiols (RSNOs), the products of S-nitrosation, play an important role due to their capacity to store, transport, and transfer NO to different targets, a biological event known as *trans*-S-nitrosation. The pathway of *trans*-S-nitrosation consists of a nucleophilic attack of a thiolate or thiol at the nitrogen atom from RSNO. Despite the recognized biological importance of RSNOs, the reactivity of these species deserves more investigation to understand how or what exactly controls it *in vivo*. Besides the mechanism of *trans*-S-nitrosation, RSNOs can perform an alternative mechanism of S-thiolation, which consists of a nucleophilic attack of the thiol at the S atom from the–SNO group. That second possible reaction leads to a disulfide formation and HNO release, an nitrogen species with great pharmacological potential (Miranda, 2005).

The mechanisms of *trans*-S-nitrosation and S-thiolation are possible due to RSNOs distinct electronic structure. Timerghazin et al. (2007) have proposed three resonance forms to RSNOs to predict and understand their overall chemical reactivity, structural, and conformational properties. According to this idea, the electronic structure of RSNOs consists of a combination among the conventional RSNOs structure (S), which has a single bond between S and N atoms, a zwitterionic structure (D) with a double bond between S and N atoms, and an RS^{-}/NO^{+} ion pair (I) (**Figure 4B**).

Despite the great number of S-nitrosated proteins identified *in vivo*, the reactivity control of those biological RSNOs is not well understood. To investigate this phenomenon, Talipov et al. have used a range of models to computationally demonstrate that specific interactions of RSNOs with charged and polar residues in proteins can result in significant modification of RSNO characteristics, including their reactivity (Talipov and Timerghazin, 2013; Timerghazin and Talipov, 2013). These reported interactions stabilize and modulate formal charges in RSNOs nitrogen and sulfur atoms, so the protein environment tightly drives a nucleophilic attack in the RSNO (**Figure 4C**).

The reactivity of cysteine residues to S-nitrosation is dependent on the protonation state of the thiol group. Thiolates are nucleophiles prompt to attack the RSNOs. When the nucleophilic attack is at the sulfur the product is a disulfide (thiolate reaction), conversely, when the attack is at the nitrogen the reaction is S-trans-nitrosation (Talipov and Timerghazin, 2013; Timerghazin and Talipov, 2013). The content of thiolate depends on the pKa of the cysteine residue, which varies according to the amino acid residues in the microenvironment and to the access to water. Turan and Meuwly (2021) showed that the S-nitrosation of myoglobin can increase the density of water molecules closer to the nitrosation site due to the polar NO group, suggesting that the hydration can be modulated by S-nitrosation (Turan and Meuwly, 2021). Protein surfaces have many nuances that may impact the S-nitrosation/thiolation reaction (**Figure 4C**). The surface forces acting on a cysteine due to the protein vicinity and access of transient water may be key to the regulation of the cysteine thiol/thiolate equilibrium (free cysteine) or to the stabilization of the resonance form (A, B or C, **Figure 4B**) in an S-nitrosated cysteine. The role of water is unknown and further studies are needed.

CONCLUSION

Hydrophobic collapse is considered the dominant driving force in protein folding in globular proteins. There is an increasing investigation on the contribution of surface direct forces to protein stability and folding. The balance between hydrophobicity and hydrophilicity is crucial for protein folding and also for its structural properties. There is no clear answer if the direct surface forces are underestimated. We discussed experimental results on the structure and dynamics of core-less proteins. They lack a hydrophobic core and are stabilized by disulfide bonds and the contacts between surface-exposed hydrophobic and hydrophilic residues. These proteins have many hydrophobic residues exposed to the solvent and yet, they are water-soluble and monomeric. They form locally stabilized hydrophobic surface clusters, in which the hydrophobic side chain is protected by the long linear side chains of the hydrophilic residues. Possibly, the hydration contributes to the cluster stabilization, by forming bridges between hydrophilic and hydrophobic side chains, as illustrated in **Figure 1**. The measurement of nuclear spin relaxation by NMR was important to describe the hydrophobic surface clusters (Machado et al., 2018c). There is no description in the literature of the stability of these clusters. NMR relaxation and direct solvation studies (NOE/ROE intensity ratio) may contribute significantly to a better understanding of these clusters' properties and their importance in protein functions.

Surface clusters are also present in globular proteins and may have an important role in regulating transient interaction in multi-domain proteins. CSP analysis through NMR revealed a hydrophobic surface patch in the co-chaperone Hsp40, which modulates inter-domain dynamics, regulates internal transient interaction and the interaction with Hsp70, being pivotal for the protein function.

We also discussed the importance of the protein surface properties to modulate the reactivity of cysteines to the post-translational modification mediated by nitric oxide, forming S-nitrosated species. The methods revised here may be of extreme importance to fully understand the surface effect.

AUTHOR CONTRIBUTIONS

FA—writing, and supervision of the writing. Overall scientific coordination. KS—writing and literature search of the theoretical simulations and multi-domain proteins. RP-A—writing and literature search of the defensins. VA—writing and literature search S-nitrosation. IC—writing, and supervision of the writing. Overall scientific coordination.

FUNDING

Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro—FAPERJ, Brazil: Grant 255.940/2020, 202.279/2018, 239.229/2018, 210.361/2015, and 204.432/2014. Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq, Brazil: 309564/2017-4.

ACKNOWLEDGMENTS

The author IPC gratefully acknowledges the financial support by postdoctoral fellowship from FAPERJ and the PROPe UNESP.

REFERENCES

Anand, P., and Stamler, J. S. (2012). Enzymatic Mechanisms Regulating Protein S-Nitrosylation: Implications in Health and Disease. *J. Mol. Med.* 90, 233–244. doi:10.1007/s00109-012-0878-z

Anfinsen, C. B. (1972). The Formation and Stabilization of Protein Structure. *Biochem. J.* 128, 737–749. doi:10.1042/bj1280737

Baldwin, R. L. (2014). Dynamic Hydration Shell Restores Kauzmann's 1959 Explanation of How the Hydrophobic Factor Drives Protein Folding. *Proc. Natl. Acad. Sci.* 111, 13052–13056. doi:10.1073/pnas.1414556111

Baldwin, R. L., and Rose, G. D. (2016). How the Hydrophobic Factor Drives Protein Folding. *Proc. Natl. Acad. Sci. USA* 113, 12462–12466. doi:10.1073/pnas.1610541113

Ben-Naim, A. (2013). Water's Contribution in Providing strong Solvent-Induced Forces in Protein Folding. *Eur. Phys. J. Spec. Top.* 223, 927–946. doi:10.1140/epjst/e2013-01981-1

Ben-Naim, A. (2014). Water's Contribution in Providing strong Solvent-Induced Forces in Protein Folding. *Eur. Phys. J. Spec. Top.* 223, 927–946. doi:10.1140/epjst/e2013-01981-1

Borgia, A., Borgia, M. B., Bugge, K., Kissling, V. M., Heidarsson, P. O., Fernandes, C. B., et al. (2018). Extreme Disorder in an Ultrahigh-Affinity Protein Complex. *Nature* 555, 61–66. doi:10.1038/nature25762

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995). Funnels, Pathways, and the Energy Landscape of Protein Folding: a Synthesis. *Proteins* 21, 167–195. doi:10.1002/prot.340210302

Colombo, M., Rau, D., and Parsegian, V. (1992). Protein Solvation in Allosteric Regulation: A Water Effect on Hemoglobin. *Science* 256, 655–659. doi:10.1126/science.1585178

Cruzeiro-Silva, C., Gomes-Neto, F., Machado, L. E. S. F., Miyamoto, C. A., Pinheiro, A. S., Correa-Pereira, N., et al. (2014). Hydration and Conformational Equilibrium in Yeast Thioredoxin 1: Implication for H+Exchange. *Biochemistry* 53, 2890–2902. doi:10.1021/bi401542v

Cyr, D. M., and Ramos, C. H. (2015). Specification of Hsp70 Function by Type I and Type II Hsp40. *Sub-cellular Biochem.* 78, 91–102. doi:10.1007/978-3-319-11731-7_4

de Medeiros, L. N., Angeli, R., Sarzedas, C. G., Barreto-Bergter, E., Valente, A. P., Kurtenbach, E., et al. (2010). Backbone Dynamics of the Antifungal Psd1 Pea Defensin and its Correlation with Membrane Interaction by NMR Spectroscopy. *Biochim. Biophys. Acta* 1798, 105–113. doi:10.1016/j.bbamem.2009.07.013

de Paula, V. S., Razzera, G., Barreto-Bergter, E., Almeida, F. C. L., and Valente, A. P. (2011). Portrayal of Complex Dynamic Properties of Sugarcane Defensin 5 by NMR: Multiple Motions Associated with Membrane Interaction. *Structure* 19, 26–36. doi:10.1016/j.str.2010.11.011

Durell, S. R., and Ben-Naim, A. (2017). Hydrophobic-hydrophilic Forces in Protein Folding. *Biopolymers* 107, e23020. doi:10.1002/bip.23020

Feng, B., Sosa, R. P., Mårtensson, A. K. F., Jiang, K., Tong, A., Dorfman, K. D., et al. (2019). Hydrophobic Catalysis and a Potential Biological Role of DNA Unstacking Induced by Environment Effects. *Proc. Natl. Acad. Sci. USA* 116, 17169–17174. doi:10.1073/pnas.1909122116

Fernández, A., and Scheraga, H. A. (2003). Insufficiently Dehydrated Hydrogen Bonds as Determinants of Protein Interactions. *Proc. Natl. Acad. Sci.* 100, 113–118. doi:10.1073/pnas.0136888100

Ferreiro, D. U., Komives, E. A., and Wolynes, P. G. (2014). Frustration in Biomolecules. *Quart. Rev. Biophys.* 47, 285–363. doi:10.1017/S0033583514000092

Gachomo, E. W., Jimenez-Lopez, J. C., Kayodé, A. P. P., Baba-Moussa, L., and Kotchoni, S. O. (2012). Structural Characterization of Plant Defensin Protein Superfamily. *Mol. Biol. Rep.* 39, 4461–4469. doi:10.1007/s11033-011-1235-y

Guon, T. E., and Chung, H. S. (2016). Hyperoside and Rutin of *Nelumbo nucifera* Induce Mitochondrial Apoptosis through a Caspase-dependent Mechanism in HT-29 Human colon Cancer Cells. *Oncol. Lett.* 11, 2463–2470. doi:10.3892/ol.2016.4247

Iqbal, A., Gomes-Neto, F., Myiamoto, C. A., Valente, A. P., and Almeida, F. C. L. (2015). Dissection of the Water Cavity of Yeast Thioredoxin 1: The Effect of a Hydrophobic Residue in the Cavity. *Biochemistry* 54, 2429–2442. doi:10.1021/acs.biochem.5b00082

Kampinga, H. H., and Craig, E. A. (2010). The HSP70 Chaperone Machinery: J Proteins as Drivers of Functional Specificity. *Nat. Rev. Mol. Cel Biol.* 11, 579–592. doi:10.1038/nrm2941

Kauzmann, W. (1959). Some Factors in the Interpretation of Protein Denaturation. *Adv. Protein Chem.* 14, 1–63. Academic Press. doi:10.1016/S0065-3233(08)60608-7

Levin, D. (2004). Biohydrogen Production: Prospects and Limitations to Practical Application. *Int. J. Hydrogen Energ.* 29, 173–185. doi:10.1016/S0360-3199(03)00094-6

Li, M., Cao, H., Lai, L., and Liu, Z. (2018). Disordered Linkers in Multidomain Allosteric Proteins: Entropic Effect to Favor the Open State or Enhanced Local Concentration to Favor the Closed State? *Protein Sci.* 27, 1600–1610. doi:10.1002/pro.3475

Machado, L. E. S. F., De Paula, V. S., Pustovalova, Y., Bezsonova, I., Valente, A. P., Korzhnev, D. M., et al. (2018a). Conformational Dynamics of a Cysteine-Stabilized Plant Defensin Reveals an Evolutionary Mechanism to Expose Hydrophobic Residues. *Biochemistry* 57, 5797–5806. doi:10.1021/acs.biochem.8b00753

Machado, L. E. S. F., De Paula, V. S., Pustovalova, Y., Bezsonova, I., Valente, A. P., Korzhnev, D. M., et al. (2018b). Conformational Dynamics of a Cysteine-Stabilized Plant Defensin Reveals an Evolutionary Mechanism to Expose Hydrophobic Residues. *Biochemistry* 57, 5797–5806. doi:10.1021/acs.biochem.8b00753

Machado, L. E. S. F., De Paula, V. S., Pustovalova, Y., Bezsonova, I., Valente, A. P., Korzhnev, D. M., et al. (2018c). Conformational Dynamics of a Cysteine-Stabilized Plant Defensin Reveals an Evolutionary Mechanism to Expose Hydrophobic Residues. *Biochemistry* 57, 5797–5806. doi:10.1021/acs.biochem.8b00753

Madan, L. L., Veeranna, S., Shameer, K., Reddy, C. C. S., Sowdhamini, R., and Gopal, B. (2011). Modulation of Catalytic Activity in Multi-Domain Protein Tyrosine Phosphatases. *PLoS One* 6, e24766. doi:10.1371/journal.pone.0024766

Miranda, K. M. (2005). The Chemistry of Nitroxyl (HNO) and Implications in Biology. *Coord. Chem. Rev.* 249, 433–455. doi:10.1016/j.ccr.2004.08.010

Modig, K., Liepinsh, E., Otting, G., and Halle, B. (2004). Dynamics of Protein and Peptide Hydration. *J. Am. Chem. Soc.* 126, 102–114. doi:10.1021/ja038325d

Morris, G., Berk, M., Klein, H., Walder, K., Galecki, P., and Maes, M. (2016). Nitrosative Stress, Hypernitrosylation, and Autoimmune Responses to Nitrosylated Proteins: New Pathways in Neuroprogressive Disorders Including Depression and Chronic Fatigue Syndrome. *Mol. Neurobiol.* 54, 4271–4291. doi:10.1007/s12035-016-9975-2

Otting, G., Liepinsh, E., and Wüthrich, K. (1991a). Protein Hydration in Aqueous Solution. *Science* 254, 974–980. doi:10.1126/science.1948083

Otting, G., Liepinsh, E., and Wüthrich, K. (1991b). Protein Hydration in Aqueous Solution. *Science* 254, 974–980. doi:10.1126/science.1948083

Papoian, G. A., Ulander, J., and Wolynes, P. G. (2003). Role of Water Mediated Interactions in Protein−Protein Recognition Landscapes. *J. Am. Chem. Soc.* 125, 9170–9178. doi:10.1021/ja034729u

Pinheiro, G. M. S., Amorim, G. C., Iqbal, A., Almeida, F. C. L., and Ramos, C. H. I. (2019). Solution NMR Investigation on the Structure and Function of the Isolated J-Domain from Sis1: Evidence of Transient Inter-domain Interactions in the Full-Length Protein. *Arch. Biochem. Biophys.* 669, 71–79. doi:10.1016/j.abb.2019.05.020

Pinheiro-Aguiar, R., do Amaral, V. S. G., Pereira, I. B., Kurtenbach, E., and Almeida, F. C. L. (2020). Nuclear Magnetic Resonance Solution Structure of Pisum Sativum Defensin 2 Provides Evidence for the Presence of Hydrophobic Surface-Clusters. *Proteins Struct. Funct. Bioinforma.* 88, 242–246. doi:10.1002/prot.25783

Rumjanek, F. D. (2018). Osmolyte Induced Tumorigenesis and Metastasis: Interactions with Intrinsically Disordered Proteins. *Front. Oncol.* 8, 1–7. doi:10.3389/fonc.2018.00353

Sanches, K., Caruso, I. P., Almeida, F. C. L., and Melo, F. A. (2020). The Dynamics of Free and Phosphopeptide-Bound Grb2-SH2 Reveals Two Dynamically Independent Subdomains and an Encounter Complex with Fuzzy Interactions. *Sci. Rep.* 10, 1–13. doi:10.1038/s41598-020-70034-w

Sanches, K., Dias, R. V. R., da Silva, P. H., Caruso, I. P., Fossey, M. A., de Souza, F. P., et al. (2021). Thermodynamic Profile and Molecular Modeling of the Interaction between Grb2 Dimer and Flavonoids Rutin and Morin. *J. Mol. Struct.* 1234, 130164. doi:10.1016/j.molstruc.2021.130164

Sanches, K., Dias, R. V. R., da Silva, P. H., Fossey, M. A., Caruso, Í. P., de Souza, F. P., et al. (2019). Grb2 Dimer Interacts with Coumarin through SH2 Domains: A Combined Experimental and Molecular Modeling Study. *Heliyon* 5, e02869. doi:10.1016/j.heliyon.2019.e02869

Smith, B. C., and Marletta, M. A. (2012). Mechanisms of S-Nitrosothiol Formation and Selectivity in Nitric Oxide Signaling. *Curr. Opin. Chem. Biol.* 16, 498–506. doi:10.1016/J.CBPA.2012.10.016

Summers, D. W., Douglas, P. M., Ramos, C. H. I., and Cyr, D. M. (2009). Polypeptide Transfer from Hsp40 to Hsp70 Molecular Chaperones. *Trends Biochem. Sci.* 34, 230–233. doi:10.1016/j.tibs.2008.12.009

Talipov, M. R., and Timerghazin, Q. K. (2013). Protein Control ofS-Nitrosothiol Reactivity: Interplay of Antagonistic Resonance Structures. *J. Phys. Chem. B* 117, 1827–1837. doi:10.1021/jp310664z

Thevissen, K., Warnecke, D. C., François, I. E. J. A., Leipelt, M., Heinz, E., Ott, C., et al. (2004). Defensins from Insects and Plants Interact with Fungal Glucosylceramides. *J. Biol. Chem.* 279, 3900–3905. doi:10.1074/jbc.M311165200

Thomma, B., Cammue, B., and Thevissen, K. (2003). Mode of Action of Plant Defensins Suggests Therapeutic Potential. *Cdtid* 3, 1–8. doi:10.2174/1568005033342000

Timerghazin, Q. K., Peslherbe, G. H., and English, A. M. (2007). Resonance Description of S-Nitrosothiols: Insights into Reactivity. *Org. Lett.* 9, 3049–3052. doi:10.1021/ol0711016

Timerghazin, Q. K., and Talipov, M. R. (2013). Unprecedented External Electric Field Effects on S-Nitrosothiols: Possible Mechanism of Biological Regulation?. *J. Phys. Chem. Lett.* 4, 1034–1038. doi:10.1021/jz400354m

Turan, H. T., and Meuwly, M. (2021). Spectroscopy, Dynamics, and Hydration of S-Nitrosylated Myoglobin. *J. Phys. Chem. B* 125, 4262–4273. doi:10.1021/acs.jpcb.0c10353

Valente, A., Miyamoto, C., and Almeida, F. C. L. (2006). Implications of Protein Conformational Diversity for Binding and Development of New Biological Active Compounds. *Cmc* 13, 3697–3703. doi:10.2174/092986706779026147

Van Der Vegt, N. F. A., and Nayar, D. (2017). The Hydrophobic Effect and the Role of Cosolvents. *J. Phys. Chem. B* 121, 9986–9998. doi:10.1021/acs.jpcb.7b06453

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. (2004). Structure, Function and Evolution of Multidomain Proteins. *Curr. Opin. Struct. Biol.* 14, 208–216. doi:10.1016/j.sbi.2004.03.011

Vuzman, D., and Levy, Y. (2012). Intrinsically Disordered Regions as Affinity Tuners in Protein-DNA Interactions. *Mol. Biosyst.* 8, 47–57. doi:10.1039/c1mb05273j

Wolynes, P. G. (2015). Evolution, Energy Landscapes and the Paradoxes of Protein Folding. *Biochimie* 119, 218–230. doi:10.1016/j.biochi.2014.12.007

Wriggers, W., Chakravarty, S., and Jennings, P. A. (2005). Control of Protein Functional Dynamics by Peptide Linkers. *Biopolymers* 80, 736–746. doi:10.1002/bip.20291

Wright, P. E., and Dyson, H. J. (2014). Intrinsically Disordered Proteins in Cellular Signalling and Regulation. *Nat. Rev. Mol. Cel Biol.* 16, 18–29. doi:10.1038/nrm3920

Yount, N. Y., and Yeaman, M. R. (2004). Multidimensional Signatures in Antimicrobial Peptides. *Proc. Natl. Acad. Sci.* 101, 7363–7368. doi:10.1073/pnas.0401567101

Yuzawa, S., Yokochi, M., Hatanaka, H., Ogura, K., Kataoka, M., Miura, K.-i., et al. (2001). Solution Structure of Grb2 Reveals Extensive Flexibility Necessary for Target recognition11Edited by P. E. Wright. *J. Mol. Biol.* 306, 527–537. doi:10.1006/jmbi.2000.4396

Zhang, L., Morrison, A. J., and Thibodeau, P. H. (2015). Interdomain Contacts and the Stability of Serralysin Protease from *Serratia marcescens*. *PLoS One* 10, e0138419. doi:10.1371/journal.pone.0138419

Zhu, G., Xia, Y., Nicholson, L. K., and Sze, K. H. (2000). Protein Dynamics Measurements by TROSY-Based NMR Experiments. *J. Magn. Reson.* 143, 423–426. doi:10.1006/jmre.2000.2022

Zmasek, C. M., and Godzik, A. (2012). This Déjà Vu Feeling-Analysis of Multidomain Protein Evolution in Eukaryotic Genomes. *Plos Comput. Biol.* 8, e1002701. doi:10.1371/journal.pcbi.1002701

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.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fmolb.2021.704274

Reconstruction of Three-Dimensional Conformations of Bacterial ClpB from High-Speed Atomic-Force-Microscopy Images

Bhaskar Dasgupta[1], Osamu Miyashita[1], Takayuki Uchihashi[2,3,4] and Florence Tama[1,4,5]*

[1]Computational Structural Biology Research Team, RIKEN-Center for Computational Science, Kobe, Japan, [2]Institute for Glyco-core Research (iGCORE), Nagoya University, Nagoya, Japan, [3]Exploratory Research Center on Life and Living Systems (ExCELLS), National Institutes of Natural Sciences, Okazaki, Japan, [4]Department of Physics, Graduate School of Science, Nagoya University, Nagoya, Japan, [5]Institute of Transformative Bio-Molecules, Nagoya University, Nagoya, Japan

ClpB belongs to the cellular disaggretase machinery involved in rescuing misfolded or aggregated proteins during heat or other cellular shocks. The function of this protein relies on the interconversion between different conformations in its native condition. A recent high-speed-atomic-force-microscopy (HS-AFM) experiment on ClpB from *Thermus thermophilus* shows four predominant conformational classes, namely, open, closed, spiral, and half-spiral. Analyses of AFM images provide only partial structural information regarding the molecular surface, and thus computational modeling of three-dimensional (3D) structures of these conformations should help interpret dynamical events related to ClpB functions. In this study, we reconstruct 3D models of ClpB from HS-AFM images in different conformational classes. We have applied our recently developed computational method based on a low-resolution representation of 3D structure using a Gaussian mixture model, combined with a Monte-Carlo sampling algorithm to optimize the agreement with target AFM images. After conformational sampling, we obtained models that reflect conformational variety embedded within the AFM images. From these reconstructed 3D models, we described, in terms of relative domain arrangement, the different types of ClpB oligomeric conformations observed by HS-AFM experiments. In particular, we highlighted the slippage of the monomeric components around the seam. This study demonstrates that such details of information, necessary for annotating the different conformational states involved in the ClpB function, can be obtained by combining HS-AFM images, even with limited resolution, and computational modeling.

Keywords: ClpB, 3D modeling, Monte-Carlo sampling, Gaussian mixture model, atomic-force-microscopy image analysis

OPEN ACCESS

Edited by:
David Perahia,
UMR8113 Laboratoire de biologie et pharmacologie appliquée (LBPA), France

Reviewed by:
Sophie Sacquin-Mora,
UPR9080 Laboratoire de Biochimie Théorique (LBT), France
Elodie Laine,
Université Pierre et Marie Curie, France

****Correspondence:***
Florence Tama
florence.tama@nagoya-u.jp

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *02 May 2021*
Accepted: *13 July 2021*
Published: *05 August 2021*

Citation:
Dasgupta B, Miyashita O, Uchihashi T and Tama F (2021) Reconstruction of Three-Dimensional Conformations of Bacterial ClpB from High-Speed Atomic-Force-Microscopy Images. Front. Mol. Biosci. 8:704274. doi: 10.3389/fmolb.2021.704274

INTRODUCTION

For a healthy cell, specific machinery relieves the effect of stress and disease on the cell. One such biomolecular machine that helps to recover cells from the deposition of aggregated proteins due to heat and proteotoxic stresses is the Hsp100 chaperon in cooperation with Hsp70 (Glover and Lindquist, 1998; Haslberger et al., 2010; Doyle et al., 2013; Mogk et al., 2018). The Hsp100 proteins are prevalent in bacteria (known as ClpB), or Yeast (known as Hsp104), and belongs to the AAA+

superfamily of ATPase proteins, hosting two ATPase domains per monomer in its hexameric structure (Deville et al., 2017; Deville et al., 2019). The disaggregation function of Hsp100 proteins takes place when the substrate proteins pass through its central pore which involves large-scale conformational changes of Hsp100 (Weibezahn et al., 2004; Gates et al., 2017; Rizo et al., 2019). Although the mechanism of such conformational changes has gained recent attention, the characterization of the dynamics including many underlying conformational states is still limited (Uchihashi et al., 2018).

In *E. coli* ClpB and Yeast Hsp104, in addition to the two ATPase domains (AAA1+ and AAA2+), each monomer includes an N-terminal domain associated with substrate binding (Barnett et al., 2005; Lee et al., 2007; Mizuno et al., 2012) and a long coiled-coil domain—acting as a "propeller" to bind Hsp70 (Carroni et al., 2014; Mogk et al., 2015). The AAA1+ and AAA2+ domains constitute the hexameric core structure with a pore in the middle, which has been shown to bind a casein substrate from cryo-electron microscopy (cryo-EM) reconstruction (Gates et al., 2017; Rizo et al., 2019). The AAA1+ and AAA2+ domains incorporate Walker A and B motifs that are responsible for ATP binding and hydrolysis, and cooperative ATP binding is associated with the structural changes in the hexamer (Mogk et al., 2003). The structural studies of ClpB/Hsp104 revealed its hexameric form, however, high-speed-atomic-force-microscopy (HS-AFM) imaging also indicates that the hexamer is fragile and breaks frequently as required in the disaggregation mechanism (Uchihashi et al., 2018). The non-rigid nature of the hexamer is also observed in a recent cryo-EM analysis revealing a spiral two-tier AAA+ ring of interaction (Yokom et al., 2016).

In the HS-AFM experiments, the structural dynamics of ClpB from *Thermus thermophilus* was investigated (Uchihashi et al., 2018). The HS-AFM images clearly indicated that the hexamer ring is fragile to form not only the round closed structure but also open or spiral conformations. The HS-AFM images include four main conformational classes, *open*, *closed*, *spiral,* and *half-spiral*. In the closed or spiral structure, a common feature is a seam between two monomers, along which the monomers separate to form the open conformations. The half-spiral architecture resembles a dimer of trimer, forming an additional seam in the opposite end. However, it should be noted that such conformational classes were inferred from the HS-AFM images, in which only a partial structure of ClpB viewed from the top was observed. Therefore, we aim to model the three-dimensional ClpB structures to visualize and interpret the salient feature of the hexameric structures and further help relate the structure of ClpB to its function.

Hybrid modeling approaches, combining computation and experiment have been developed to generate 3D models from low-resolution data (Rout and Sali, 2019; Srivastava et al., 2020). In such approaches, data from multiple sources are combined through the lens of computational sampling, aiming to better interpret experimental data. Even for low-resolution structural data, the usage of computational modeling enables us to discuss the function of biomolecules in terms of 3D models. Such hybrid or integrative modeling techniques have been widely used, where they are applied to recover structural details from small-angle X-ray scattering profile (Gorba and Tama, 2010; Derevyanko and Grudinin, 2014; Schindler et al., 2016; Ekimoto and Ikeguchi, 2018; Chen et al., 2019), cross-linking mass spectrometry (Faini et al., 2016; Degiacomi et al., 2017), cryo-EM (Trabuco et al., 2008; Grubisic et al., 2010; Miyashita et al., 2017; Kim et al., 2019; Malhotra et al., 2019), X-ray free-electron laser (XFEL) (Tokuhisa et al., 2016; Nagai et al., 2018; Nakano et al., 2018) and AFM (Amyot and Flechsig, 2020; Dasgupta et al., 2020; Niina et al., 2020; Fuchigami et al., 2021) studies. Some of these methods aim to recover structural details from experimental data by simulating conformational changes from a known conformational state. Recently, we have developed such an approach to relate 3D conformational changes embedded in theoretical AFM images (Dasgupta et al., 2020). In this study, we apply our algorithm to experimental ClpB AFM images.

Our algorithm uses Monte-Carlo (MC) sampling to fit an initial low-resolution 3D structural model to a target AFM image. Structural models are represented at low-resolution using 3D Gaussian density distributions since the target AFM data is a low-resolution image usually from a large system. A 3D Gaussian mixture model (3D-GMM), derived from an atomically detailed structure, is used as the initial low-resolution model (Kawabata, 2008). It should be noted that an atomic structure is not necessary to generate the initial 3D-GMM. The MC sampling algorithm is based on three crucial factors. First, during the optimization, we need to generate a pseudo-AFM image from our 3D models. 3D-GMM can be used to rasterize over the set of kernels generating a low-resolution pseudo-AFM image. Second, we need to compare this generated pseudo-AFM image to the target AFM image of a given protein. Third, candidate models generated during the Monte Carlo sampling need to be evaluated to keep the model structurally compact.

The HS-AFM experiments on ClpB dynamics were performed under near-physiological conditions revealing a variety of ClpB conformations, which were significantly different from either of the known conformations (Uchihashi et al., 2018). In our current study, we started from two atomically detailed conformations obtained from cryo-EM experiments, an asymmetric non-rigid two-tier spiral structure of Hsp104 from Yeast (Yokom et al., 2016) and a symmetric closed ring conformation of ClpB from *E. coli* (Deville et al., 2017). We modified our algorithm to consider both conformations as initial models and perform sampling based on a mechanical potential defined by combining the initial spiral and closed ring conformations. We performed 3D modeling on four different conformational classes observed in the HS-AFM experiments (**Figure 1**, **Supplementary Table 1**). The reconstructed 3D models from our sampling can be used to detect salient features within ClpB conformations. Moreover, we decoded some finer details of the ClpB hexameric architecture, which cannot be clearly observed from HS-AFM images. Lastly, we could also interpolate between different conformational classes to compare novel ClpB structures. These results demonstrate that our 3D structure modeling approach from AFM images can be applied to experimental data, providing a new approach to study conformational transitions in macromolecular complexes through AFM-computation hybrid modeling.

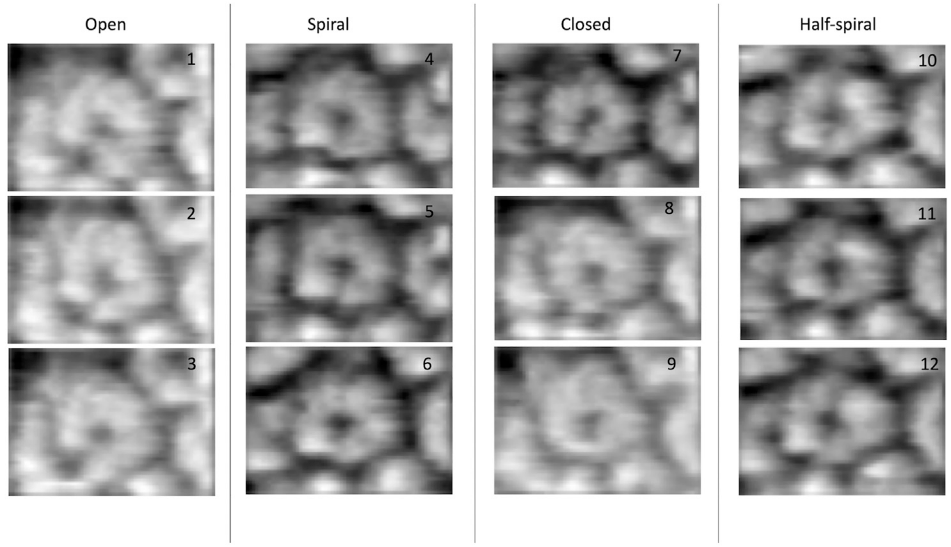


FIGURE 1 | Experimental AFM images used for 3D structure modeling (Uchihashi et al., 2018).

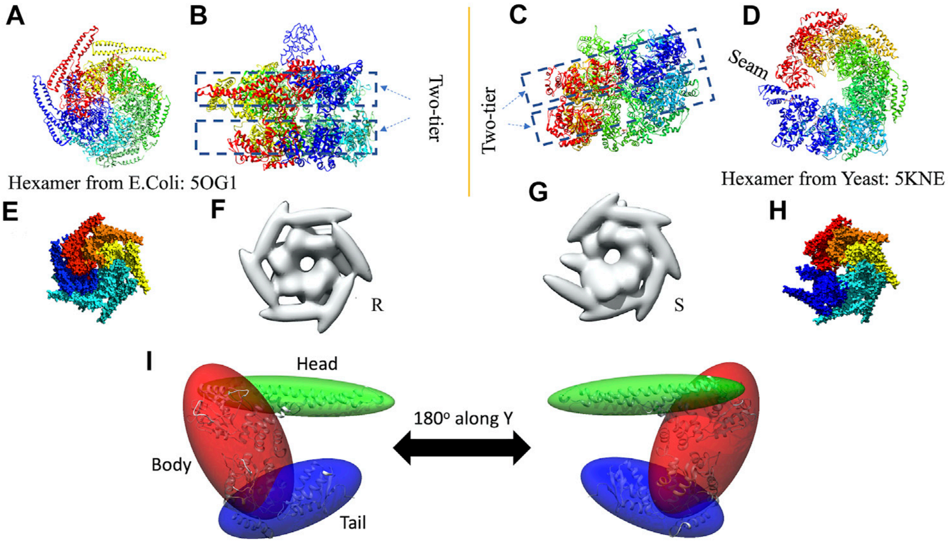


FIGURE 2 | Hexameric structures selected as initial models from *E. coli* **(A,B)** and Yeast **(C,D)** from 5OG1 and 5KNE, respectively. In **(A)** and **(D)** the structures are shown from the top and in **(B,C)**, a side-view is shown describing their two-tier structure. In **(D)** a seam is shown between two neighboring chains denoted by chains "A" (in blue) and "F" (in red). The corresponding seam is less prominent in 5OG1 (see **Figure 2A**). In the middle row **(E–H)**, the models with *T. thermophilus* monomer (from 1QVR, chain A) are shown, and in **(I)** the 3-kernel model for 1QVR, chain A is shown (with the atomic structure embedded); the head, body, and tail kernels are annotated and shown in different colors. In **(E,H)**, the atomic hexamer models after superimposition to 5OG1 and 5KNE are shown from a top view, respectively. In **(F,G)**, the 18-kernel representations based on 5OG1 and 5KNE which are referred to as R and S initial models in the text.

MATERIALS AND METHODS

Preparation of Initial Models

In 3D modeling against the AFM images, we have used two initial models based on structures originally obtained from cryo-EM reconstructions of Yeast ClpB, which adopts a two-tier spiral conformation (PDB ID: 5KNE) (Yokom et al., 2016) and *E. coli* ClpB, whose conformation is a symmetric closed ring (PDB ID: 5OG1) (**Figures 2A–D**) (Deville et al., 2017).

These two structures are sequentially different from the *T. thermophilus* construct used in the HS-AFM experiment (Uchihashi et al., 2018). To mimic the experimental construct, the missing residues in 1QVR (a trimeric structure of ClpB from *T. thermophilus* with sequence identity to chain A of 5KNE is

44.8% and to chain A of 5OG1 is 56.1%), chain A were modeled with Modeller using the automodel class (residues 1 to 3, 851 to 854 in the terminals and three loop regions from residues 235 to 245, 272 to 290, and 637 to 650) (**Supplementary Figure 1**) (Martí-Renom et al., 2000; Webb and Sali, 2014). To prepare the initial structures based on 5KNE and 5OG1, we first superposed 1QVR chain A to each of the chains of 5KNE and 5OG1 (chains A to F) by using Chimera matchmaker (see **Supplementary Section 2A**, **Supplementary Table 2**) (Pettersen et al., 2004; Meng et al., 2006). This is followed by deletion of the N-terminal regions (residue 1–165) generating two hexameric ClpB arrangements, i.e., a closed ring and a spiral conformation. Finally, the whole complex was oriented so that the C-terminal regions face towards the bottom as observed in the HS-AFM experiments (Uchihashi et al., 2018) (**Figures 2E,H**).

The AFM images are of nanometer resolution, therefore conformational transitions cannot be discussed with atomic-level details and thus we employ a coarse-grained three-dimensional Gaussian mixture model. In this technique a polypeptide chain is described by a weighted sum of three-dimensional Gaussian kernels. A Gaussian kernel is parameterized by its center and a covariance matrix. The volume of the kernel within a certain threshold is geometrically represented by an ellipsoid (see **Supplementary Section 2B**). The parameters of the mixture model can be obtained by expectation-maximization optimizing algorithm. In the current study, we used "gmconvert" software (Kawabata, 2008; Kawabata, 2018) to obtain such Gaussian mixture models.

We employ a Gaussian mixture model defined by an 18 Gaussian kernel arrangement, where a chain in the hexamer is described by 3 kernels (**Supplementary Figure 1**, **Figure 2I**). Each of the domains of the ClpB monomer is included in different Gaussian kernels. The coiled-coil domain of ClpB is included in a long elliptical kernel which we refer to *head* kernel. The C-terminal domain of ClpB is included in the *tail* kernel. The nucleotide-binding domains AAA1+ and AAA2+ together with the linker domain are included in the largest kernel that we refer to as *body* kernel. Looking from the top, as imaged in the HS-AFM experiment, the upper part of the body kernel hosts the AAA1+ domain while the bottom part hosts the AAA2+ domain. The coiled-coil domain is found at the top and the C-terminal region at the bottom. These 18-kernel Gaussian mixture models starting from 5OG1 and 5KNE are used as initial conformations to model AFM images and are hereafter are referred to as R (closed symmetric Ring conformation) and S (Spiral asymmetric conformation), respectively (**Figures 2F,G**). We call each of the six chains in the models A to F, where F and A are across the seam.

In the closed ring or spiral conformation, a common feature is a seam between two monomers. Therefore, we manually rotated around the *z*-axis of the S model, for which the seam is more prominent, to align it with the seam observed in the AFM images (see **Supplementary Sections 2C,D**, **Supplementary Tables 1, 3**). For the R model, we applied a z-rotation identical to that performed on the S model. Such rotated models were used to start the MC sampling. This protocol works well (in terms of final converged models) for the AFM images with the annotations spiral, open and close, however, poor convergence was observed for half-spiral cases. In such a case, the R model also needed to be rotated, independently of the S model. The z-rotations applied to different models are given in **Supplementary Table 3**.

Kernel Position Restraints for Sampling

In the MC algorithm, a random move is applied to one of the kernels and the restraint score for this new model is calculated to determine whether the model should be kept. This score is based on an empirical scoring function to model attraction and repulsion between kernels to ensure that the kernels would not overlap or move too far apart. The restraint scores are defined for each of the kernel pairs using their overlap values. Each term has the lowest score when the overlap value is the average of those for the two initial models (see **Supplementary Section 4**). The restraint score is given by an asymmetric harmonic function, with different curvature parameters for attraction and repulsion. These parameters are selected based on the types of kernel pairs. The restraints between the three kernels within each chain are set to be strongest (see **Figure 2I** for the kernel definitions) so that three domains keep the overall shape of the monomer while maintaining enough flexibility to allow conformational transitions. The interchain restrains between the neighboring "body" kernels are set to be less restrictive than intrachain pairs. The tail-kernels repulsion parameters between neighboring chains are set to be identical to the one between neighboring body kernels since the "tail" kernels closely follow "body" kernels. The "head" kernels comprising coiled-coil domain are least restricted owing to their flexible nature. One important aspect of the parameters is how we define the interaction between the chains A and F, which completes the circular arrangement. AFM results, as well as previous atomic structures of ClpB, revealed the possible presence of a seam between these chains. In addition, for the open-class conformation described in AFM images, those two chains are separated. Therefore, we assume that the interaction between the chains A and F is special in that the attraction is weaker. The details of those parameters are given in **Supplementary Section 4**, **Supplementary Table 4**.

Apart from the above attraction-repulsion scoring, to incorporate the connectivity of the underlying molecular structure into modeling, we also ensure that the intersections between kernels in the initial model remained. Such intersections can be viewed as hinges between kernels. To do so, a pair of phantom particles is assigned to the intersection site, one from each of the overlapping kernels (**Supplementary Figure 4**). Those phantom particles are initially on top of each other, but due to sampling, they may go far and thereby breaks the connection. We ensure, via a connectivity restraint, that the distance between two such phantom particles should be within 5 Å and each of them always falls within the overlapping region.

Sampling against AFM Image

During MC sampling we compare a pseudo-AFM image from the current candidate model (aka candidate image) to the target AFM input image. In this study, we used a more robust image similarity measure—the structural similarity metric (SSIM) (Wang et al., 2004; Wang and Bovik, 2009) instead of L1-norm in our previous study (Dasgupta et al., 2020). The SSIM takes its maximum value of 1.0 when two images are identical. The SSIM between the candidate image and target AFM image is first calculated. Then

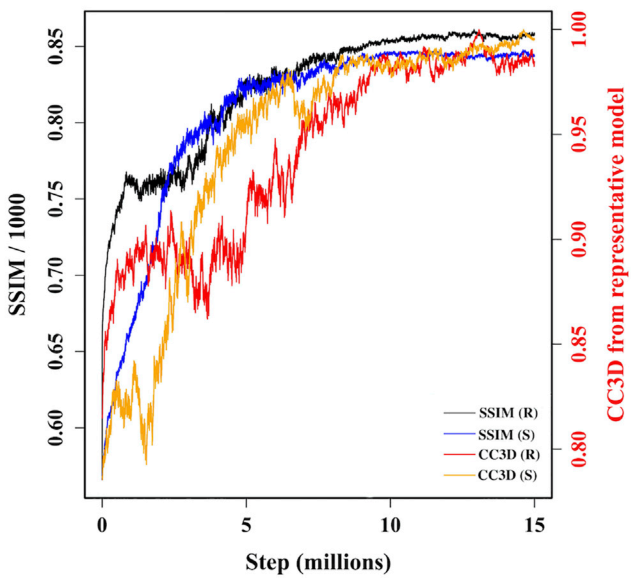


FIGURE 3 | A plot of variation of similarities between candidate models along two MC trajectories, initiated from R and S models, and target AFM image 7. The left vertical axis shows variation of SSIM between candidate models and target AFM image. The right vertical axis shows variation of the correlation coefficient after rigid-body fitting of kernel centers between intermediate candidate models and final converged model.

the candidate model (in 3D) is updated by applying a random move to one of the 18 Gaussian kernels. After applying such a move, we either accept or reject the updated model based on the change in the restraint score. The change in the score is converted to a probability associated with a temperature parameter. If the updated model is accepted in this first step, then we measure the similarity between the pseudo-AFM image corresponding to the updated model to the target AFM image and compare it to the previously calculated SSIM between candidate and target AFM images. The change in SSIM is again converted to a probability value to determine whether the updated model (in 3D) is accepted or rejected. The flowchart of our MC algorithm is given in **Supplementary Figure 2**, and the details of parameters used in Monte-Carlo algorithms are discussed in **Supplementary Section 3**. As shown in **Figure 3**, sampling runs started from both R and S models converge similarly with SSIM reaching a high value. In addition, the models accepted during the sampling become similar in 3D to our final reconstructed model.

In the above two-step MC algorithm, the scoring function for the first step is defined from the change in empirical restraints while the scoring function for the second step is defined from the change in SSIM. Therefore, in principle, the temperature parameters associated with each step should be different. To avoid this, we multiplied SSIM by 10000 and use identical temperature parameters for both steps. With this simplification, we could use one annealing scheme (**Supplementary Figure 3A**) modifying the temperature between 0.2 and 1.0. In **Supplementary Figure 3B**, we show how scores change in first and second MC steps with accepted steps for one of the trajectories, demonstrating our first MC step is more restrictive in rejecting new model than the second MC step, however, the ranges of change in score are comparable.

Selecting Representative Models for an AFM Image

The structure optimization process is run in two phases. In the first phase, for a given AFM image we ran twenty trajectories in parallel each for 1.5 × 10^6 steps, where ten trajectories randomly seeded were initiated from model R and ten trajectories were initiated from model S. At the end of the sampling, we select a few best models (based on a quantile value of 0.2 for SSIM over all the models) from each set of ten trajectories. The models from R and the models from S are compared pair-wise by calculating 3D correlation coefficients after rigid-body fitting of the kernel centers to identify which of the two models initiated from R or S are most similar, hence the most converged solutions starting from two different initial models.

In the second phase, we start from these converged solutions and continue for additional 1.5 × 10^6 steps for twenty trajectories, ten from the R converged model obtained previously and ten from S converged model, all randomly seeded. At the end, we repeat the above analysis to determine final converged models, providing us with two solutions for a given AFM image. In total, one optimization process for one AFM image with two phases takes about 166 h.

Comparison of Models in 3D

Two measures were used to compare the Gaussian mixture 3D models (Kawabata, 2008). In one approach, we performed a rigid-body fitting using the kernel centers between the two models, followed by computation of correlation coefficient. We use an in-house python script to perform rigid-body fitting (Diamond, 1988), and then the "gmconvert" program is used to compute the correlation coefficient (Kawabata, 2008). We also perform rigid-body fitting with UCSF Chimera using a full density distribution, where densities are computed according to a Gaussian mixture model (Cheng et al., 2015). The resulting correlation coefficients are referred to as $\text{CC3D}_{\text{density}}$.

We also clustered potential candidate solutions to understand their conformational landscape (**Supplementary Table 5**). Gaussian mixture models are converted into voxel-based density maps and principal component analysis of the set of maps was performed. The voxel-based density maps are then projected into a lower 2-dimensional space defined by the first two principal components. Then we performed clustering in the lower dimensional space by DBSCAN algorithm using scikit-learn. For clustering, we have used "min_sample" value of 4, and the "eps" parameter is optimized by visually checking the clustering solution. For each cluster, we determined the median conformation from the PCA projection. The clusters are ranked in terms of their size (**Table 1**, **Supplementary Table 5**).

TABLE 1 | Details of representative models for AFM images of different class.

AFM image annotation	Image index (see Figure 1)	Top rank pairwise correlation-coefficient between two sets of candidates ($CC3D_{density}$)	Details of representative model from R			Details of representative model from S		
			SSIM	Cluster	Similarity from median	SSIM	Cluster	Similarity from median
Open	1	0.9578	8,316.7	6	0.9885	8,315.2	6	0.9883
Open	2	0.9708	8,000.2	1	0.9702	8,038.6	3	0.9941
Open	3	0.8824	8,571.0	4	0.9802	8,584.2	2	0.9869
Spiral	4	0.9113	8,467.6	5	0.9939	8,770.2	1	0.9662
Spiral	5	0.8567	8,644.5	4	0.9824	8,669.8	1	0.9758
Spiral	6	0.9118	8,641.7	1	0.9714	8,467.6	4	0.99
Closed	7	0.978	8,789.9	1	0.9603	8,797.4	2	0.978
Closed	8	0.9327	8,644.4	2	0.9773	8,597.2	6	0.9848
Closed	9	0.9011	8,578.2	3	0.9777	8,528.8	1	0.9987
Half-spiral	10	0.8441	8,495.5	1	0.9726	8,447.5	2	0.9616
Half-spiral	11	0.8294	8,752.8	1	0.9871	8,384.8	5	0.9971
Half-spiral	12	0.8742	8,545.3	2	0.9948	8,566.4	2	0.9882

RESULTS

Dataset of AFM Images and Conformations of ClpB Observed in AFM Experiment

The dataset of AFM images of ClpB molecules used in this work was previously described (Uchihashi et al., 2018). The conformations observed in the HS-AFM experiments were classified into four categories—*open, spiral, closed, and half-spiral*, with spiral being the major conformation. The detail method of defining such conformations from 2D AFM analysis is explained in the method section of the reference (Uchihashi et al., 2018). Briefly, the open conformations were identified from a histogram of "circularity" (defined from the ratio of the perimeter and area outlining each molecule). From the rest of the conformations, closed, spiral, and half-spiral conformations were identified by analyzing the height profile along the top surface of the ring. In current study we have used the annotations of conformational classes used in the above reference. The full dataset of the above AFM images consisted of 340 images. We randomly select twelve AFM images, taking three AFM images for each type of conformation (**Figure 1**). The AFM image dimensions are 66 × 42 (width x height) pixels, where each pixel along *x*-direction (width) is 4.545 Å and that along *y*-direction (height) is 5.477 Å. By manual inspection, we define masks for the selected AFM images over a region of interest that includes the ClpB oligomer. The preprocessing of the AFM images is discussed in detail in the **Supplementary Section 2**. Background corrected masked AFM images (**Supplementary Table 1**) were used as input for the 3D modelling problem.

In the ClpB hexameric structure, a prominent feature, seam, between two neighboring monomers is observed in spiral and close conformations (**Figure 1**). In open conformation, two neighboring monomers along the seam are separated. The half-spiral conformation is one of the unique conformations detected in the AFM experiment, for which the hexamer can be understood as a dimer of trimer or hexamer with two farthest seams (Uchihashi et al., 2018). Because such classes are based on 2D height data from AFM images, we aim to recover the corresponding full 3D information (**Figure 1**). For all AFM image classes, we begin MC sampling against a given AFM image from one of the two initial models—R (ring closed symmetric conformation based on 5OG1) or S (spiral asymmetric conformation based on 5KNE). The correlation coefficient between the two initial models, $CC3D_{density}$, is 0.72.

Reconstruction of *Open* Class AFM Images

After MC sampling against open class images (AFM image 1, 2, and 3), the $CC3D_{density}$ between models derived from R and S increase up to 0.97 for AFM image 2, 0.96 for AFM image 1, and 0.88 from AFM image 3 (see **Table 1**). The similarity between the AFM image and pseudo-AFM image generated from a representative model is also converged as the values from R and S solutions are similar. In addition, the converged representative models are almost identical (~0.97 $CC3D_{density}$ or better) to the medoids of the cluster they belong to (**Table 1**; **Supplementary Figure 5B**). Even though models are converged, due to the low-resolution nature of AFM images, some conformational differences are apparent. For example, both candidate models obtained either from R or S against AFM images 1 and 2 show a clearer separation around the original seam than in AFM image 3 (**Figure 4**). The body kernel arrangements for chains A and F are also different. Finally, for AFM images 1 and 3, while the body kernels orientations are similar, the coiled-coil domain orientations are different (**Supplementary Figure 6**). These results indicate that conformational diversity can exist even within the open state.

Reconstruction of *Spiral* Class AFM Images

The best-converged models against the spiral class AFM images are shown in **Figure 5** and details of their similarities are given in **Table 1**. The convergence of the models is high for AFM images 4 and 6 ($CC3D_{density}$ >0.91). For AFM images 4 and 5, models derived from R show a hexameric arrangement (**Figures 5A,B**, in the left), which is less prominent in models derived from S. In addition, the S derived models show a clear separation around the seam between chains A and F (**Figures 5A,B**, in the middle). The arrangement of head kernels in all the models derived from R

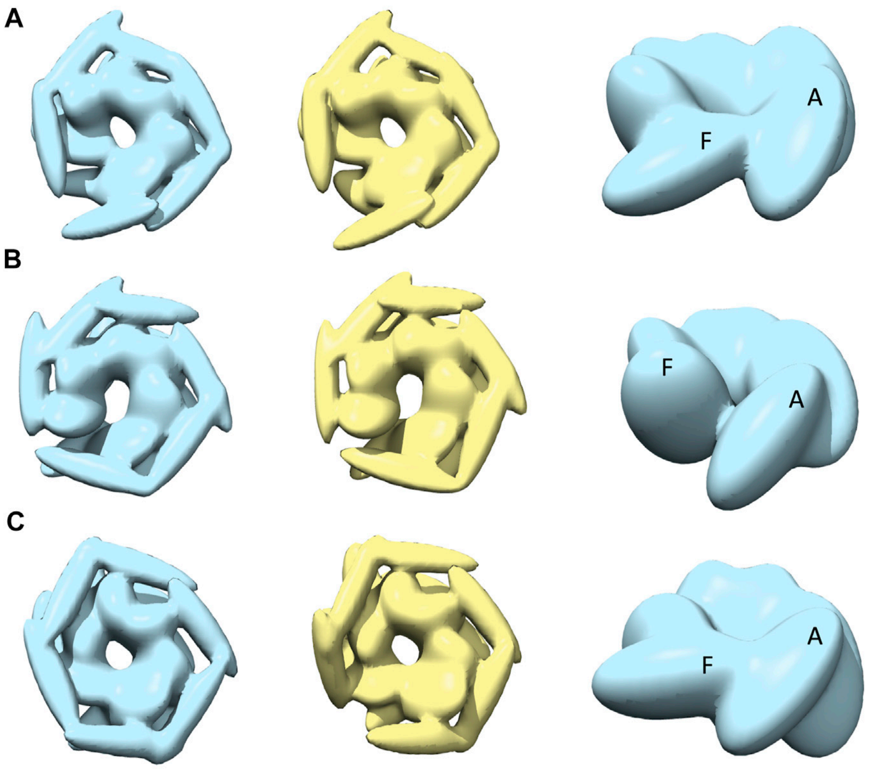


FIGURE 4 | The most converged model between structural modelling starting from R and S models for open class AFM images. The rows **(A–C)** are for AFM images 1, 2 and 3, respectively. The left column shows candidate solutions from R (light blue) and middle column shows candidate solutions from S (yellow). In the right column, only the body kernel is shown from R candidate solutions given in the left column. Chains A, F are indicated on top of the kernels.

is similar to the blades of a propeller (Carroni et al., 2014), where one end is connected to the AAA1+ domain and the distal end is pointing towards a neighboring AAA1+ domain. Regarding the body kernels, their orientations around the seam show some diversity (**Figure 5**, right). More specifically, for the R model of AFM image 6, chains A and F are interacting through AAA1+ domain residues (residues 160 and 330) (**Figure 5C**, left and right), whereas for AFM images 4 and 5, chain A AAA2+ domain residues (residues 560–750) are interacting with chain F AAA1+ domain (**Figures 5A,B**, more evident from S models shown in the middle). Therefore, the spiral nature of the conformations is clearly exhibited in the models obtained against AFM images 4 and 5; a twist in the orientation of body kernels resulted in a vertical shift in an upward direction for chain F from chain A. Such a twist does not appear in the model obtained against AFM image 6. However, in this case, a spiral feature can be seen from head kernels, showing one end of the head kernel from the chain F is more upward than the head kernel from chain A.

Reconstruction of *Closed* Class AFM Images

The best-converged models against closed class AFM images are shown in **Figure 6**. In particular, for AFM image 7, we obtained the best-converged results out of all 12 AFM images (0.978) (**Table 1**). The convergence is also high for AFM images 8 and 9 (>0.90). Note that for such AFM images a pre-rotation of the initial model along the z-axis to match the seam observed in the AFM image was needed and as a result, the whole system is rotated (compared to the result for AFM image 7) (**Figure 6**).

The closed nature of the conformation as observed in AFM images (**Figure 1**) is clearer in the case of R-derived models (**Figure 6**, leftmost column), with chains A and F tightly closed and more parallel to each other. However, the interaction between chains A and F, in R-derived models, is different for AFM image 7, compared to AFM images 8 and 9, indicating conformational heterogeneity among the closed state. For AFM image 7, the chain A body kernel region hosting the AAA2+ domain is interacting with the chain F body kernel region hosting the AAA1+ domain (**Figure 6**, rightmost column), while the orientations of chain A and F body kernels for AFM images 8 and 9 are more slanted and parallel (**Figures 6B,C**, in the right). In addition, chains A, E and F head kernels are laying horizontally in the XY-plane and no interaction is observed between the distal end of chains B, C, and D head kernels as they are oriented downward keeping a parallel placement (**Figure 6A**, left). For AFM image 8, the placement of the chain F head kernel is different (distal end facing downward) from the other head

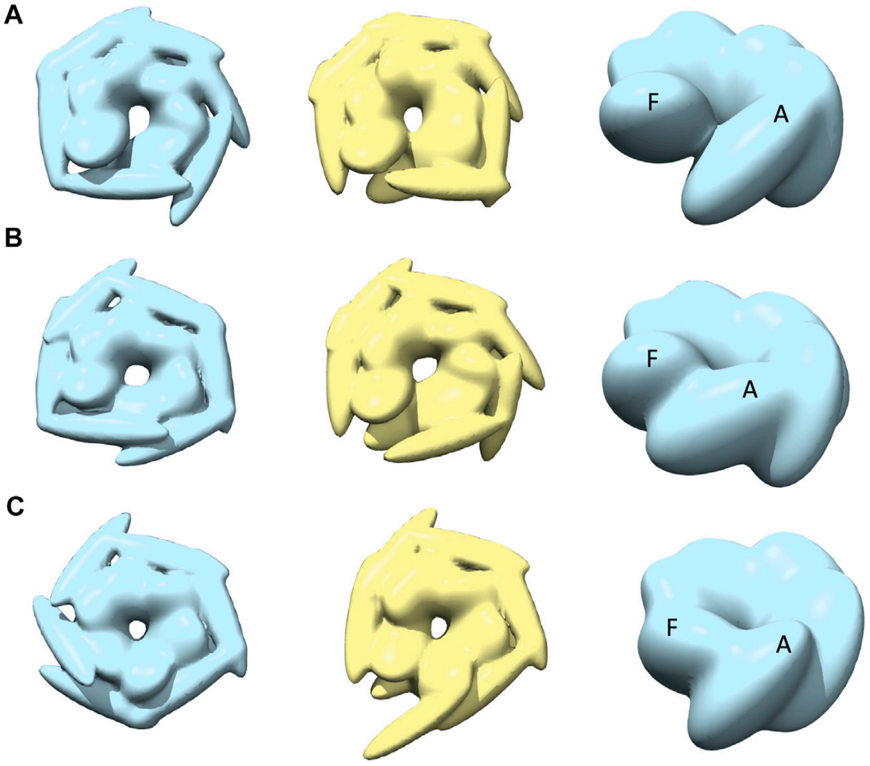


FIGURE 5 | The most converged model between structural modelling starting from R (light blue) and S (yellow) for spiral class AFM images. The rows **(A–C)** are for AFM images 4, 5 and 6, respectively.

kernels (**Figure 6B**, left), and for AFM image 9, head kernels are arranged following a hexameric symmetry, similar to the blades of a fan (**Figure 6C**, left).

While the representative models derived from S are less compact than those from R, their 2D similarities to the target AFM image are similar. A more open-like structure is seen for models derived from S (AFM images 8 and 9), where chains are separated at the seam (**Figures 6B,C** in the middle). In this case, the chain A tail kernel is pointing towards the neighboring AAA2+ domain from chain F. Such an interaction, between the C-terminal domain and AAA2+ domain indicates an alternative way of forming the closed hexamer. The diversity of resulting models indicates a possible conformational heterogeneity within closed state structures, which is difficult to solely characterize from the top view experiments by AFM.

Finally, we should also note that the arrangement of the head kernels differs from the models obtained in other classes. For spiral class AFM images, head kernels are oriented to form a hexameric arrangement (**Figure 5**), which is quite different from the model for the closed AFM images where coiled-coiled domains are oriented further away from each other (**Figure 6**).

Reconstruction of *Half-Spiral* Class AFM Images

Modeling against the half-spiral class AFM images was also performed (**Figure 7** and **Table 1**). In all the cases, two opened seams are observed with additional seams between chains C and D (see **Figures 7A–C**). However, the maximum similarities between best models are lower than for other target AFM images (>0.8), indicating that the modeling of the half-spiral state is more complex. More specifically, S derived models show more open-like conformations than R derived models. The best convergence among half-spiral cases is obtained for AFM image 12 (**Figure 7C**), in which the seam between C and D is less prominent than between chain A and F. These models provide insights into the possible overall arrangements of the hexamer chains in the half-spiral conformational state.

Conformational Variations within Models Other than Converged Models

To assess conformational variations of the models for a particular target AFM image, we also examined other candidate models using clustering (Materials and Methods *Comparison of Models in 3D* and **Supplementary Table 5**) in addition to the most converged model. For example, we compared a converged representative model to the median of the cluster to which it belongs. In most cases, the converged model is very similar to the cluster median (>0.95 $\mathrm{CC3D_{density}}$, **Supplementary Figure 5B**). We also checked the similarity of the converged representative models to the median of the largest cluster which may not include the representative model

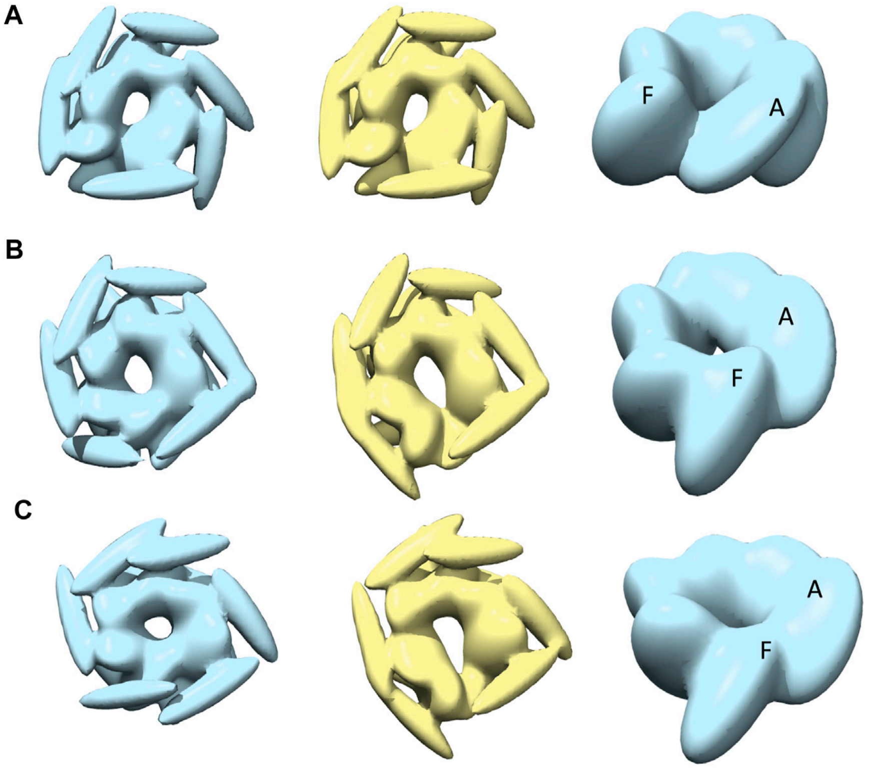


FIGURE 6 | The most converged model between structural modelling starting from R (light blue) and S (yellow) for closed class AFM images. The rows **(A–C)** are for AFM images 7, 8 and 9, respectively.

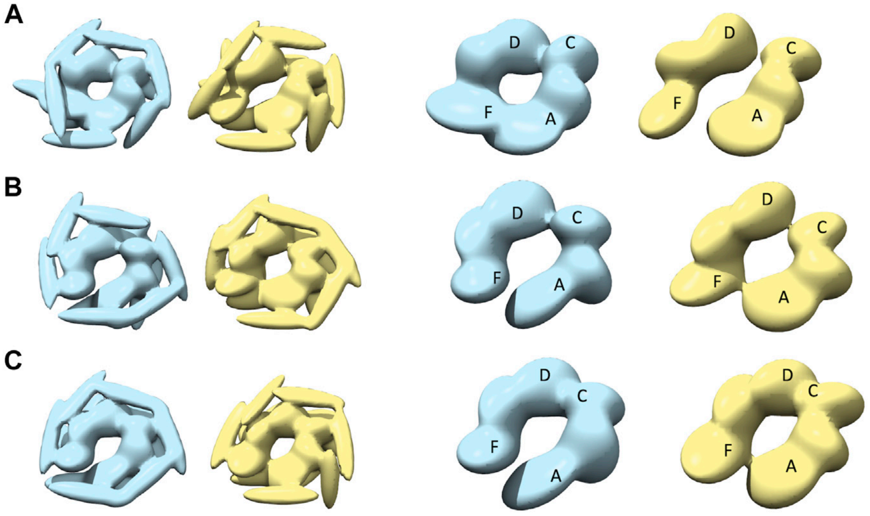


FIGURE 7 | The most converged model between structural modeling starting from R (light blue) and S (yellow) for half-spiral class AFM images. The rows **(A–C)** are for AFM images 10, 11, and 12, respectively. In the third and fourth columns, only the body kernel are shown from R and S candidate solutions given in the first and second columns, respectively. The chains A, F, C, and D are shown on top of the kernels.

(**Supplementary Figure 5A**) and found high similarity in most of the cases. We observed that the conformational variation is mostly due to the orientation of the flexible coiled-coil domain and the positioning of the tail-kernels in relation to the body kernel around the seam (**Supplementary Table 6**). This confirms that our converged models capture the hexameric arrangement robustly with finer differences arising due to the flexible coiled-coil region. In addition, we analyzed the similarity between initial conformations and final reconstructed models to characterize the class of such conformations (**Supplementary Figure 7**). The MC sampling generates models significantly different from the initial model which indicates that our sampling generates novel models based on the respective AFM images. This is particularly true for some half-spiral models that show the presence of two seams (**Figure 7**).

Interpolated Conformational Transition between Different Converged Models

The HS-AFM imaging clearly demonstrated that ClpB undergoes large conformational changes. In the current study, we did not aim to recover the mechanism of such conformational change. However, it is possible to compare two reconstructed models from different AFM image classes and visualize such transitions by interpolation between volume models using Chimera "morph map" feature (Pettersen et al., 2004; Meng et al., 2006). We have used the body kernel representation (from the R model) of the four most converged reconstructions, namely AFM images 1 (open), 2 (open), 6 (spiral), and 7 (closed). For the spiral class, we observe chains A and F along the seam are oriented in a slanted way such that the AAA1+ domain in the body kernel from chain F is interacting with the AAA1+ domain from chain A. However, for the closed class AFM image, the orientation of the body kernel from chain A slips relative to chain F, and its AAA2+ domain interacts with the AAA1+ domain of chain F. Such a movement is clearly visualized in the **Supplementary Movie 1**. Similarly, we compare the orientations of body kernels for the closed class AFM image against the open class model. In the open class AFM image 2, chains A and F are separated while keeping a similar orientation as observed in AFM image 7 (closed). This also can be seen in the interpolated visualization (**Supplementary Movie 2**). AFM image 1 is also an open class image, however, in this case, the body kernel from chain F is more slanted than in AFM image 2; the AAA1+ domain from chain F is pointing towards the AAA2+ domain of chain A. A comparison of spiral class AFM image and open class AFM image 1 indicates that the body kernel from chain F is getting separated while keeping a slanted orientation (**Supplementary Movie 3**). Lastly, morphing of the model generated from AFM image 2 (open) into the model generated from AFM image 6 (spiral) shows that the orientations of chain A and F remain rather similar (**Supplementary Movie 4**).

DISCUSSION

HS-AFM images of biomolecules can include a wealth of information on various conformational states, yet 3D reconstructed models can provide further functional implications. To this aim, we propose a novel method for reconstructing 3D models using MC sampling based on a coarse-grained representation—the Gaussian mixture model. We applied our method to reconstruct the hexameric structure of *T. thermophilus* ClpB protein as observed in the HS-AFM experiment conducted in presence of ATP (Uchihashi et al., 2018).

The basic algorithm applied to a few theoretical systems was previously published (Dasgupta et al., 2020). Here we further extend our algorithm, in particular, we adopted a state-of-the-art 2D image comparison technique—SSIM for quantifying the agreement between the 3D model and AFM images. Our previous restraint scheme was also improved by introducing a new harmonic restraint scheme built from the interpolation of the correlation coefficient from two different initial models. This extension was necessary because the HS-AFM experiments observed ClpB conformational dynamics for a much wider timescale, and as a result, some conformations could not be characterized by any of the known conformational states (Uchihashi et al., 2018). For example, while spiral arrangements are frequently observed in the HS-AFM experiment, the overall conformation is different from the known asymmetric spiral conformation (Deville et al., 2017). Therefore, 3D modeling against AFM images was performed from the two known conformational states that capture some of ClpB conformational diversity in terms of symmetric/asymmetric systems. In addition, in the harmonic restraining scheme, we treat attraction and repulsion in a separate way (**Supplementary Section 4**), with stronger repulsive interactions to eliminate any severe steric clash between the 3D ellipsoidal kernels. We also used a connectivity restraint between kernels from the same monomer (**Supplementary Figure 4**) to ensure that these kernels, within a monomer, remain connected behaving as hinges between a pair of kernels during random moves.

Models with good convergence were obtained for three AFM image classes—open, closed and spiral. The open and spiral classes show highly conserved 3D models (**Table 1**), and for the closed class we obtained fairly conserved reconstructions. However, for the half-spiral class, modeling was more challenging due to a characteristic feature in this class—an additional seam between chains C and D. One structural feature of ClpB important during model constructions is the presence of a seam between chains A and F. The initial conformation S has a prominent seam and important for the modeling for the spiral models from the AFM images. However, for closed class of AFM images, the models derived from R (without a prominent seam) show more consistent results. In the case of half-spiral states, there are two seams, and therefore such conformations are more distinct from either of the initial conformations. This is likely to be a reason for the difficulty to model half-spiral forms. Nonetheless, the initial correlation between two models (based on 5OG1 and 5KNE) was 0.72 which significantly increased during the modeling to more than 0.82 even for half-spiral AFM images. Moreover, the similarity between initial models and the final conformations are low for most of the cases (**Supplementary Figure 7**). These results demonstrate that

TABLE 2 | Cross-correlation ($CC3D_{density}$) between converged representative models for each AFM image derived from R and S models. The entries above the diagonal show similarity between reconstructed models derived from the R model, and those below the diagonal show similarity between reconstructed models derived from the S model.

S model		R model	AFM Image annotation											
			Open			Spiral			Closed			Half-spiral		
			1	2	3	4	5	6	7	8	9	10	11	12
AFM Image annotation	Open	1		0.8220	0.7921	0.7859	0.7989	0.8222	0.7019	0.7144	0.7345	0.7267	0.7839	0.8175
		2	0.8152		0.7896	0.8034	0.7793	0.8046	0.7594	0.7173	0.7353	0.6824	0.7805	0.8130
		3	0.7465	0.8015		0.7742	0.7721	0.7619	0.7785	0.7292	0.7838	0.6839	0.7556	0.7873
	Spiral	4	0.7637	0.8017	0.7791		0.8815	0.8392	0.8004	0.7769	0.7856	0.7448	0.8448	0.8306
		5	0.7300	0.7436	0.7689	0.8876		0.8850	0.7965	0.7810	0.7713	0.7646	0.9278	0.8947
		6	0.8164	0.7973	0.7704	0.8317	0.8232		0.7715	0.7724	0.7874	0.7967	0.8417	0.8714
	Closed	7	0.7189	0.7836	0.8099	0.8017	0.7905	0.8151		0.8184	0.8674	0.7269	0.7751	0.7953
		8	0.7055	0.7489	0.7824	0.8190	0.7767	0.7853	0.8613		0.8611	0.7344	0.7559	0.7911
		9	0.7113	0.7437	0.7951	0.8211	0.7795	0.7996	0.8843	0.8714		0.7351	0.7485	0.7700
	Half-spiral	10	0.6506	0.6552	0.6951	0.6913	0.7620	0.6985	0.7142	0.7206	0.7050		0.7363	0.8043
		11	0.7862	0.7749	0.7524	0.7943	0.8310	0.7700	0.7426	0.7743	0.7330	0.7673		0.8678
		12	0.7407	0.7549	0.7498	0.7950	0.8499	0.8091	0.8127	0.8091	0.7667	0.8758	0.8291	

TABLE 3 | Summary of **Table 2** to average cross-correlation between different types of AFM images (by averaging over each 3 × 3 block in **Table 2**).

$<CC3D_{density}>$	Open	Spiral	Closed	Half-spiral
Open	0.7945	0.7891	0.7393	0.7590
Spiral	0.7745	0.8580	0.7825	0.8352
Closed	0.7555	0.7987	0.8606	0.7591
Half-spiral	0.7289	0.7779	0.7531	0.8134

our algorithm with coarse-grained model can sample models far from the initial models.

To examine the overall distributions and relations between reconstructed models predicted against different AFM images, the cross-correlation between the 3D models were calculated (**Tables 2**, **3**). These data show that the models reconstructed against the same class of AFM images are more similar to each other than for other classes, i.e., the models are capturing the essential details of AFM experimental observation. On the other hand, it can be observed that cross-correlations between models for closed and spiral AFM image classes are moderately high (>0.81), indicating that the conformations of these two classes are similar. The open conformation can have considerable different hexameric arrangements, which is illustrated from lower cross-correlations within the open class AFM images.

While we focused on the analysis of conformational dynamics of ClpB in this study, we developed our algorithms aiming to apply to other AFM studies. In the current implantation we need at least some knowledge of the protein under investigation and how many domains we should expect. Also, we need to know rough size and shape of those domains. Since our approach is based on coarse-grained model, information from homologous structures can be utilized as we used the homologous structures (5OG1 and 5KNE) in this study. The algorithm can be theoretically applied in a completely *ab-initio* manner; however, such a detailed evaluation is yet to be done.

To conclude, in the present work, we applied a hybrid modeling approach to provide further interpretation of experimental HS-AFM images of bacterial ClpB. The structural dynamics of ClpB is crucial to understand how a healthy cell maintains its proper functioning. Previously, the HS-AFM experiment was performed without substrate in presence of ATP, and analysis of 2D images revealed novel conformational classes of the ClpB oligomer. Our current study using hybrid modeling from AFM data enabled us to reconstruct 3D models for each of the conformational classes. Such models show the full hexameric architecture of ClpB including all domains related to disaggregation and capture specific features of the four conformational states observed in the AFM images. In particular, some novel conformational classes were suggested, such that the open class could be divided into sub-classes. In addition, conformational transitions between 3D models representing the different classes were obtained. More specifically, we observed that a slipping motion between two monomers around the seam in spiral conformation might be necessary to reach the closed conformation. We have used a coarse-grained representation for ClpB structure in line with the low-resolution nature of the AFM data and thus a detailed atomic-level picture is not directly obtained from our analysis. However, our study provides domain-level 3D structural information with structural insights into the different types of ClpB hexameric arrangements observed in the HS-AFM experiments. While the present work aimed to address growing attention on the function of ClpB disaggretase, it also demonstrated the usability of our algorithm, hybrid modeling based 2D AFM images, on HS-AFM experimental data. In the future, such applications could help to address questions regarding the structure, dynamics, and function of biomolecules from HS-AFM experiments.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

BD, OM, and FT conceived and designed the research. BD implemented the method and performed the research, wrote first draft of the manuscript. BD, OM, TU, and FT analyzed the results and wrote the final manuscript. FT supervised the project.

FUNDING

This work was supported by JSPS KAKENHI Grant Number 20H05453, 19H05389 and 18H04512, RIKEN Dynamic Structural Biology Project, and FOCUS for Establishing Supercomputing Center of Excellence.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.704274/full#supplementary-material

REFERENCES

Amyot, R., and Flechsig, H. (2020). BioAFMviewer: An Interactive Interface for Simulated AFM Scanning of Biomolecular Structures and Dynamics. *Plos Comput. Biol.* 16, e1008444. doi:10.1371/journal.pcbi.1008444

Barnett, M. E., Nagy, M., Kedzierska, S., and Zolkiewski, M. (2005). The Amino-Terminal Domain of ClpB Supports Binding to Strongly Aggregated Proteins. *J. Biol. Chem.* 280, 34940–34945. doi:10.1074/jbc.M505653200

Carroni, M., Kummer, E., Oguchi, Y., Wendler, P., Clare, D. K., Sinning, I., et al. (2014). Head-to-tail Interactions of the Coiled-Coil Domains Regulate ClpB Activity and Cooperation with Hsp70 in Protein Disaggregation. *Elife* 3, e02481. doi:10.7554/eLife.02481

Chen, P.-c., Shevchuk, R., Strnad, F. M., Lorenz, C., Karge, L., Gilles, R., et al. (2019). Combined Small-Angle X-ray and Neutron Scattering Restraints in Molecular Dynamics Simulations. *J. Chem. Theor. Comput.* 15, 4687–4698. doi:10.1021/acs.jctc.9b00292

Cheng, A., Henderson, R., Mastronarde, D., Ludtke, S. J., Schoenmakers, R. H. M., Short, J., et al. (2015). MRC2014: Extensions to the MRC Format Header for Electron Cryo-Microscopy and Tomography. *J. Struct. Biol.* 192, 146–150. doi:10.1016/j.jsb.2015.04.002

Dasgupta, B., Miyashita, O., and Tama, F. (2020). Reconstruction of Low-Resolution Molecular Structures from Simulated Atomic Force Microscopy Images. *Biochim. Biophys. Acta Gen. Subj.* 1864, 129420. doi:10.1016/j.bbagen.2019.129420

Degiacomi, M. T., Schmidt, C., Baldwin, A. J., and Benesch, J. L. P. (2017). Accommodating Protein Dynamics in the Modeling of Chemical Crosslinks. *Structure* 25, 1751–1757. doi:10.1016/j.str.2017.08.015

Derevyanko, G., and Grudinin, S. (2014). HermiteFit: Fast-Fitting Atomic Structures into a Low-Resolution Density Map Using Three-Dimensional Orthogonal Hermite Functions. *Acta Cryst. D Biol. Crystallogr.* 70, 2069–2084. doi:10.1107/S1399004714011493

Deville, C., Carroni, M., Franke, K. B., Topf, M., Bukau, B., Mogk, A., et al. (2017). Structural Pathway of Regulated Substrate Transfer and Threading through an Hsp100 Disaggregase. *Sci. Adv.* 3, e1701726. doi:10.1126/sciadv.1701726

Deville, C., Franke, K., Mogk, A., Bukau, B., and Saibil, H. R. (2019). Two-Step Activation Mechanism of the ClpB Disaggregase for Sequential Substrate Threading by the Main ATPase Motor. *Cell Rep.* 27, 3433–3446. doi:10.1016/j.celrep.2019.05.075

Diamond, R. (1988). A Note on the Rotational Superposition Problem. *Acta Cryst. Sect. A.* 44, 211–216. doi:10.1107/s0108767387010535

Doyle, S. M., Genest, O., and Wickner, S. (2013). Protein rescue from Aggregates by Powerful Molecular Chaperone Machines. *Nat. Rev. Mol. Cel. Biol.* 14, 617–629. doi:10.1038/nrm3660

Ekimoto, T., and Ikeguchi, M. (2018). Hybrid Methods for Modeling Protein Structures Using Molecular Dynamics Simulations and Small-Angle X-Ray Scattering Data. *Adv. Exp. Med. Biol.* 1105, 237–258. doi:10.1007/978-981-13-2200-6_15

Faini, M., Stengel, F., and Aebersold, R. (2016). The Evolving Contribution of Mass Spectrometry to Integrative Structural Biology. *J. Am. Soc. Mass. Spectrom.* 27, 966–974. doi:10.1007/s13361-016-1382-4

Fuchigami, S., Niina, T., and Takada, S. (2021). Case Report: Bayesian Statistical Inference of Experimental Parameters via Biomolecular Simulations: Atomic Force Microscopy. *Front. Mol. Biosci.* 8, 636940. doi:10.3389/fmolb.2021.636940

Gates, S. N., Yokom, A. L., Lin, J., Jackrel, M. E., Rizo, A. N., Kendsersky, N. M., et al. (2017). Ratchet-like Polypeptide Translocation Mechanism of the AAA+ Disaggregase Hsp104. *Science* 357, 273–279. doi:10.1126/science.aan1052

Glover, J. R., and Lindquist, S. (1998). Hsp104, Hsp70, and Hsp40: A Novel Chaperone System that Rescues Previously Aggregated Proteins. *Cell* 94, 73–82. doi:10.1016/s0092-8674(00)81223-4

Gorba, C., and Tama, F. (2010). Normal Mode Flexible Fitting of High-Resolution Structures of Biological Molecules Toward SAXS Data. *Bioinform. Biol. Insights* 4, 43–54. doi:10.4137/bbi.s4551

Grubisic, I., Shokhirev, M. N., Orzechowski, M., Miyashita, O., and Tama, F. (2010). Biased Coarse-Grained Molecular Dynamics Simulation Approach for Flexible Fitting of X-ray Structure into Cryo Electron Microscopy Maps. *J. Struct. Biol.* 169, 95–105. doi:10.1016/j.jsb.2009.09.010

Haslberger, T., Bukau, B., and Mogk, A. (2010). Towards a Unifying Mechanism for ClpB/Hsp104-Mediated Protein Disaggregation and Prion propagation. *Biochem. Cell Biol.* 88, 63–75. doi:10.1139/o09-118

Kawabata, T. (2018). Gaussian-input Gaussian Mixture Model for Representing Density Maps and Atomic Models. *J. Struct. Biol.* 203, 1–16. doi:10.1016/j.jsb.2018.03.002

Kawabata, T. (2008). Multiple Subunit Fitting into a Low-Resolution Density Map of a Macromolecular Complex Using a Gaussian Mixture Model. *Biophys. J.* 95, 4643–4658. doi:10.1529/biophysj.108.137125

Kim, D. N., Moriarty, N. W., Kirmizialtin, S., Afonine, P. V., Poon, B., Sobolev, O. V., et al. (2019). Cryo_fit: Democratization of Flexible Fitting for Cryo-EM. *J. Struct. Biol.* 208, 1–6. doi:10.1016/j.jsb.2019.05.012

Lee, S., Choi, J.-M., and Tsai, F. T. F. (2007). Visualizing the ATPase Cycle in a Protein Disaggregating Machine: Structural Basis for Substrate Binding by ClpB. *Mol. Cell* 25, 261–271. doi:10.1016/j.molcel.2007.01.002

Malhotra, S., Träger, S., Dal Peraro, M., and Topf, M. (2019). Modelling Structures in Cryo-EM Maps. *Curr. Opin. Struct. Biol.* 58, 105–114. doi:10.1016/j.sbi.2019.05.024

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Sali, A. (2000). Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29, 291–325. doi:10.1146/annurev.biophys.29.1.291

Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C., and Ferrin, T. E. (2006). Tools for Integrated Sequence-Structure Analysis with UCSF Chimera. *BMC Bioinform.* 7, 339. doi:10.1186/1471-2105-7-339

Miyashita, O., Kobayashi, C., Mori, T., Sugita, Y., and Tama, F. (2017). Flexible Fitting to Cryo-EM Density Map Using Ensemble Molecular Dynamics Simulations. *J. Comput. Chem.* 38, 1447–1461. doi:10.1002/jcc.24785

Mizuno, S., Nakazaki, Y., Yoshida, M., and Watanabe, Y.-h. (2012). Orientation of the Amino-Terminal Domain of ClpB Affects the Disaggregation of the Protein. *FEBS J.* 279, 1474–1484. doi:10.1111/j.1742-4658.2012.08540.x

Mogk, A., Bukau, B., and Kampinga, H. H. (2018). Cellular Handling of Protein Aggregates by Disaggregation Machines. *Mol. Cell* 69, 214–226. doi:10.1016/j.molcel.2018.01.004

Mogk, A., Kummer, E., and Bukau, B. (2015). Cooperation of Hsp70 and Hsp100 Chaperone Machines in Protein Disaggregation. *Front. Mol. Biosci.* 2, 22. doi:10.3389/fmolb.2015.00022

Mogk, A., Schlieker, C., Strub, C., Rist, W., Weibezahn, J., and Bukau, B. (2003). Roles of Individual Domains and Conserved Motifs of the AAA+ Chaperone ClpB in Oligomerization, ATP Hydrolysis, and Chaperone Activity. *J. Biol. Chem.* 278, 17615–17624. doi:10.1074/jbc.M209686200

Nagai, T., Mochizuki, Y., Joti, Y., Tama, F., and Miyashita, O. (2018). Gaussian Mixture Model for Coarse-Grained Modeling from XFEL. *Opt. Express* 26, 26734. doi:10.1364/OE.26.026734

Nakano, M., Miyashita, O., Jonic, S., Tokuhisa, A., and Tama, F. (2018). Single-particle XFEL 3D Reconstruction of Ribosome-Size Particles Based on Fourier Slice Matching: Requirements to Reach Subnanometer Resolution. *J. Synchrotron Radiat.* 25, 1010–1021. doi:10.1107/S1600577518005568

Niina, T., Fuchigami, S., and Takada, S. (2020). Flexible Fitting of Biomolecular Structures to Atomic Force Microscopy Images via Biased Molecular Simulations. *J. Chem. Theor. Comput.* 16, 1349–1358. doi:10.1021/acs.jctc.9b00991

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera? A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* 25, 1605–1612. doi:10.1002/jcc.20084

Rizo, A. N., Lin, J., Gates, S. N., Tse, E., Bart, S. M., Castellano, L. M., et al. (2019). Structural Basis for Substrate Gripping and Translocation by the ClpB AAA+ Disaggregase. *Nat. Commun.* 10, 2393. doi:10.1038/s41467-019-10150-y

Rout, M. P., and Sali, A. (2019). Principles for Integrative Structural Biology Studies. *Cell* 177, 1384–1403. doi:10.1016/j.cell.2019.05.016

Schindler, C. E. M., de Vries, S. J., Sasse, A., and Zacharias, M. (2016). SAXS Data Alone Can Generate High-Quality Models of Protein-Protein Complexes. *Structure* 24, 1387–1397. doi:10.1016/j.str.2016.06.007

Srivastava, A., Tiwari, S. P., Miyashita, O., and Tama, F. (2020). Integrative/Hybrid Modeling Approaches for Studying Biomolecules. *J. Mol. Biol.* 432, 2846–2860. doi:10.1016/j.jmb.2020.01.039

Tokuhisa, A., Jonic, S., Tama, F., and Miyashita, O. (2016). Hybrid Approach for Structural Modeling of Biological Systems from X-ray Free Electron Laser Diffraction Patterns. *J. Struct. Biol.* 194, 325–336. doi:10.1016/j.jsb.2016.03.009

Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible Fitting of Atomic Structures into Electron Microscopy Maps Using Molecular Dynamics. *Structure* 16, 673–683. doi:10.1016/j.str.2008.03.005

Uchihashi, T., Watanabe, Y.-H., Nakazaki, Y., Yamasaki, T., Watanabe, H., Maruno, T., et al. (2018). Dynamic Structural States of ClpB Involved in its Disaggregation Function. *Nat. Commun.* 9, 2147. doi:10.1038/s41467-018-04587-w

Wang, Z., and Bovik, A. C. (2009). Mean Squared Error: Love it or Leave it? A New Look at Signal Fidelity Measures. *IEEE Signal. Process. Mag.* 26, 98–117. doi:10.1109/MSP.2008.930649

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image Quality Assessment: from Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* 13, 600–612. doi:10.1109/tip.2003.819861

Webb, B., and Sali, A. (2014). Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinform.* 47, 5.6.1–5.6.32. doi:10.1002/0471250953.bi0506s47

Weibezahn, J., Tessarz, P., Schlieker, C., Zahn, R., Maglica, Z., Lee, S., et al. (2004). Thermotolerance Requires Refolding of Aggregated Proteins by Substrate Translocation through the central Pore of ClpB. *Cell* 119, 653–665. doi:10.1016/j.cell.2004.11.027

Yokom, A. L., Gates, S. N., Jackrel, M. E., Mack, K. L., Su, M., Shorter, J., et al. (2016). Spiral Architecture of the Hsp104 Disaggregase Reveals the Basis for Polypeptide Translocation. *Nat. Struct. Mol. Biol.* 23, 830–837. doi:10.1038/nsmb.3277

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.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 08 November 2021
doi: 10.3389/fmolb.2021.745089

On the Relationship Between EB-3 Profiles and Microtubules Growth in Cultured Cells

Arshat Urazbaev[1], *Anara Serikbaeva*[2,3], *Anna Tvorogova*[4], *Azamat Dusenbayev*[1], *Sholpan Kauanova*[2,5] and *Ivan Vorobjev*[2,4,5]*

[1]*National Laboratory Astana, Nazarbayev University, Nur-Sultan, Kazakhstan,* [2]*Laboratory of Biophotonics and Imaging, National Laboratory Astana, Nazarbayev University, Nur-Sultan, Kazakhstan,* [3]*Department of Physiology and Biophysics (M/C 901), University of Illinois at Chicago, Chicago, IL, United States,* [4]*A.N.Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia,* [5]*Department of Biology, School of Sciences and Humanities, Nazarbayev University, Nur-Sultan, Kazakhstan*

OPEN ACCESS

Edited by:
Paulo Ricardo Batista,
Oswaldo Cruz Foundation (Fiocruz),
Brazil

Reviewed by:
Gavin Buxton,
Robert Morris University,
United States
Francisco Gomes Neto,
Oswaldo Cruz Institute (Fiocruz), Brazil

****Correspondence:***
Ivan Vorobjev
ivan.vorobyev@nu.edu.kz

Specialty section:
This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences

Received: *21 July 2021*
Accepted: *27 September 2021*
Published: *08 November 2021*

Citation:
Urazbaev A, Serikbaeva A, Tvorogova A, Dusenbayev A, Kauanova S and Vorobjev I (2021) On the Relationship Between EB-3 Profiles and Microtubules Growth in Cultured Cells.
Front. Mol. Biosci. 8:745089.
doi: 10.3389/fmolb.2021.745089

Microtubules are dynamic structures undergoing rapid growth and shrinkage in living cells and *in vitro*. The growth of microtubules *in vitro* was analyzed with subpixel precision (Maurer et al., Current Biology, 2014, 24 (4), 372–384); however, to what extent these results could be applied for microtubules growing *in vivo* remains largely unknown. Particularly, the question is whether microtubule growth velocity in cells could be sufficiently approximated by a Gaussian distribution or its variability requires a more sophisticated description? Addressing this question, we used time-lapse microscopy and mathematical modeling, and we analyzed EB-3 comets forming on microtubules of cultured cells with subpixel precision. Parameters of comets (shape, form, and velocity) were used as topological characteristics of 3D voxel objects. Using regression analysis, we determined the real positions of the microtubule tips in time-lapse sequences. By exponential decay fitting of the restored comet intensity profile, we found that *in vivo* EB-3 rapidly exchanges on growing microtubule ends with a decoration time ~ 2 s. We next developed the model showing that the best correlation between comet length and microtubule end growth velocity is at time intervals close to the decoration time. In the cells, EB comet length positively correlates with microtubule growth velocity in preceding time intervals, while demonstrating no correlation in subsequent time intervals. Correlation between comet length and instantaneous growth velocity of microtubules remains under nocodazole treatment when mean values of both parameters decrease. Our data show that the growth of microtubules in living cells is well-approximated by a constant velocity with large stochastic fluctuations.

Keywords: microtubules, dynamic instability, EB proteins, fluorescent microscopy, cell culture, live cell imaging

INTRODUCTION

Microtubules (MTs) are polymers of α/β-tubulin dimers that exhibit dynamic instability behavior (Mitchison and Kirschner, 1984), with their plus ends frequently switching between growth and shrinkage phases. MT dynamics *in vivo* are generally reported by six parameters: the rates of growth and shortening, duration of the attenuated state (pauses), and the frequencies of switching between these three phases. Tracking of growing MTs in cells using fluorescently labeled tubulin is a

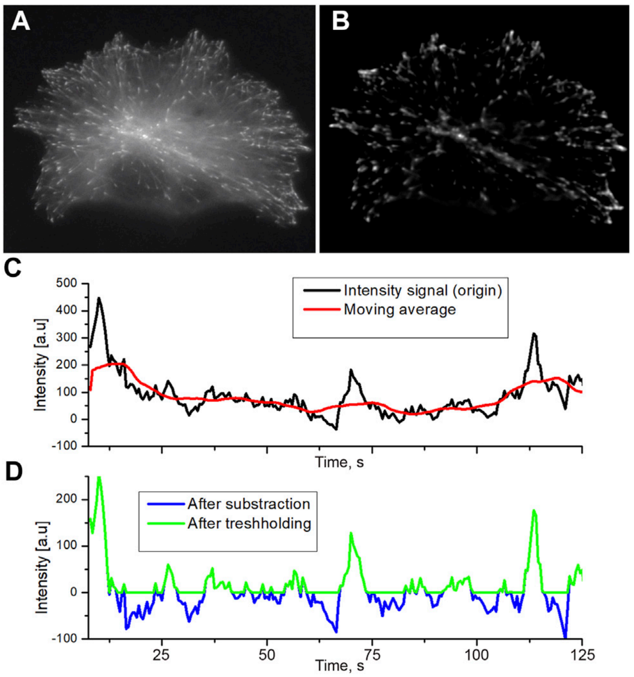


FIGURE 1 | Raw initial image of EB comets in the transfected HT-1080 cell **(A)** and normalized image by the running averaging and thresholding applied **(B)**. The background in the normalized image is set to 0 intensity. Original profiles and moving average filtering **(C)** black line: original intensity, red line: the result of moving average. The thresholding of intensity after subtraction of the moving average **(D)** is marked by a blue line. Green line: intensity after thresholding. All negative values were assigned to zero.

challenging task because of the relatively high density of these structures in the cytoplasm (Odde, 1997; Vorobjev et al., 1999). The alternative strategy was developed by using fluorescently labeled plus end–tracking proteins (+TIPs; Schuyler and Pellman, 2001). Fluorescently tagged EB proteins bound to the ends of growing MTs and appear as comet-like structures that could be traced in time-lapse experiments with high precision (Bieling et al., 2007; Seetapun et al., 2012; Maurer et al., 2014).

EB proteins turn over on growing MT ends and forms characteristic cap there during the MT growth period (Dragestein et al., 2008; Duellberg et al., 2016; Guesdon et al., 2016). These proteins bind exclusively to the growing MT plus ends and are best suitable to examine MT dynamics *in vivo* (Galjart, 2010; Matov et al., 2010). EB intensity profile on the MT tips represents a cap, made of GTP-bound tubulin subunits (Bowne-Anderson et al., 2013; Brouhard and Rice, 2018). This profile is approximated with an exponential decay to extract the characteristic comet length–L (Bieling et al., 2008; Seetapun et al., 2012; Maurer et al., 2014). *In vitro* EB cap size is proportional to the rate of MT elongation (Rickman et al., 2017) and thus could be used for indirect measurements of the MT growth rate. However, analysis of MT dynamics *in vivo* does not consider the precise structure of the plus end comet (Matov et al., 2010), and the results of measurements are mean velocity. When measurements of the plus end displacement are made at short time intervals (to determine instantaneous growth rate), the accuracy of such measurements is doubtful, unless the position of the MT tip is determined with subpixel precision. Taking this into account, we recently conducted a detailed analysis of the comet length measurement and show that EB comets could be described by a piecewise exponential/Gaussian function approximation (Mustyatsa et al., 2019). The exponential/Gaussian function approximation or Gausspont method allows achieving the sub-pixel precision obtained at *in vitro* measurements. Thus, the use of the comet length might be beneficial for MT dynamics analysis, but to use it, one needs to minimize the limitations of such analysis. In the current study, we have undertaken the analysis of the correlation between the size of EB-3 comets and MT growth velocity in the cancer cells. We developed the model allowing precise determination of the comet length and head position considering the shift of the start (zero points) of exponential decay from the brightest point on the profile. Our data show that EB-3 rapidly accumulates on the growing MT end with a short decoration time. The growth of a microtubule in a cell can be described as a process going on at a constant rate with random fluctuations.

RESULTS

We performed a live cell imaging experiment to collect images of the growing microtubule ends (the MT "comets") labeled by EB-3-RFP fluorescent probes with a time resolution of 500 ms. The collected sequences of the comet's images were used to describe the fine structure of EB-3 comets and to determine the dependence of the length of the comet tail on the velocity of displacement of the growing MT. (**Figure 1**). To further measure the comet size, velocity, and shape, we developed an automated analytical routine.

Processing software consists of two parts:

1) A comet tracker for the detection of the growing MT tip based on a topology algorithm. A tracker code was developed in MATLAB, and the function of the flood fill algorithm was developed in c++. This part of the program tracks the comets, builds trajectories for comets, and filters comets from other objects in image sequences.
2) A comet tail analyzer to extract the tail length from the data of the tracked comets. This part was developed on MATLAB. The comet tail analyzer allows extraction of the 2D profile of comet intensity (from head to tail) and making regression with the convoluted function of brightness. It allows getting the tail length of a comet.

Image Preprocessing for Automated Analysis of the "Comets"

The original raw image of MT comets contains noises, uneven illumination, and unwanted autofluorescent particles that need to be removed (**Figure 1A**). The preprocessing protocol we developed exploits spatial and time properties of growing comet objects to remove unwanted background pollution.

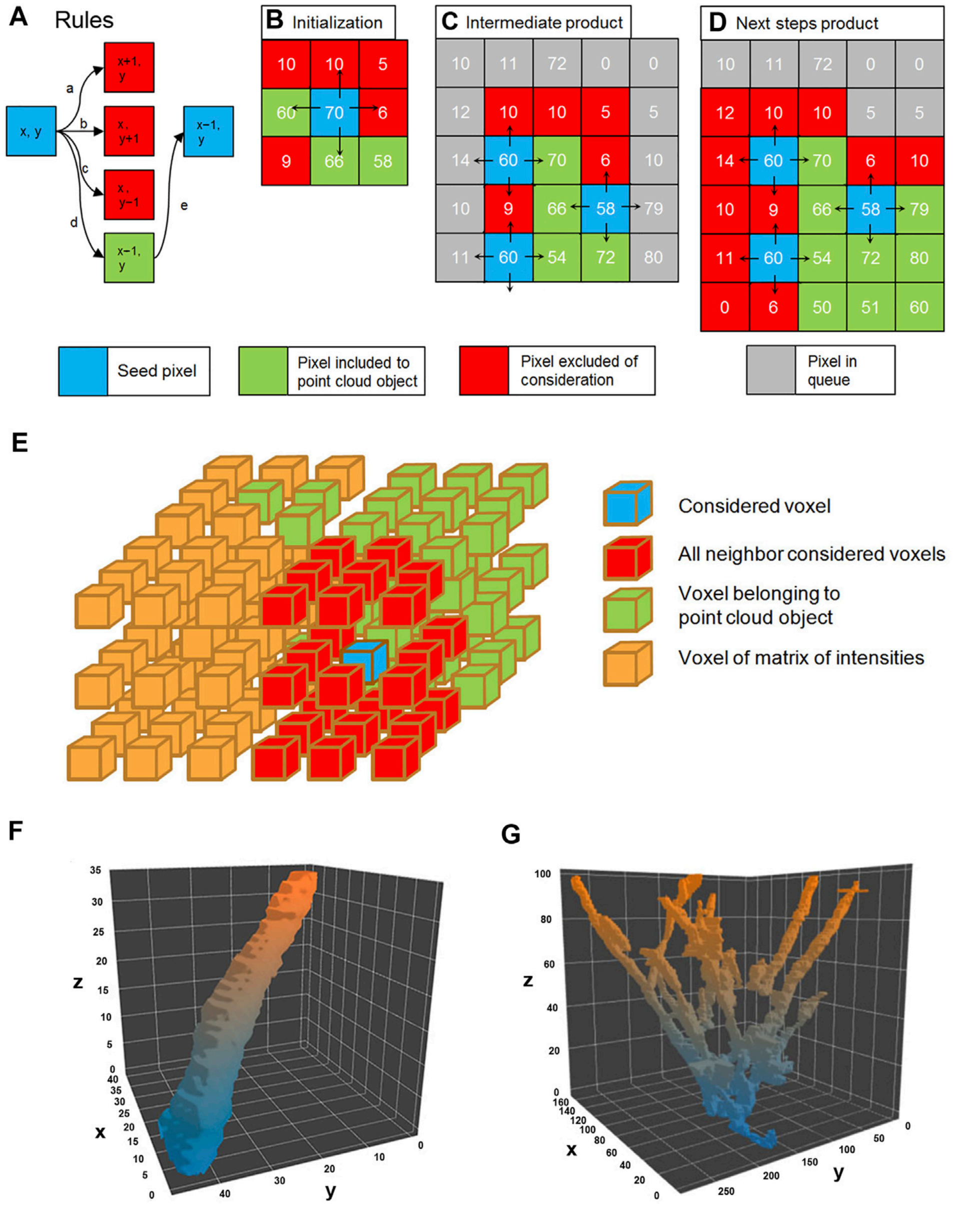


FIGURE 2 | Flood fill algorithm procedure rules **(A)**. The only pixel which satisfies the rule (marked green) is chosen as seed to the next step marked green. The initialization of flood fill by seed pixel **(B)**; in the consecutive steps **(C)** and **(D)** flood fill a search for other pixels satisfying the rules. The 3D representation of the flood fill algorithm **(E)**. The result is 3D visualization of the plus end comet in the spatial–temporal domain **(F)**. A pillar represents the trajectory of the segmented growing single microtubule, and the Z domain is time. The multitude of growing microtubules forms an object point cloud where comets form a tree shape, where comets' tracks may overlap **(G)**.

At the initial step, images were filtered by Fourier transformation (both high pass and low pass) to reduce noises. High-pass filtering reduces illumination unevenness. The low-pass filtering suppresses the high-frequency noises generated by the detector matrix (**Figure 1B**). Denoised images obtained could be described as 2D matrix, where columns are y coordinate, rows are x coordinate, and the intensity of a pixel will be the value of matrix cell I (x, y). These images were further used to render a 3D data matrix I (x, y, t), where I (x, y) is the intensity matrix of microscope images for the time point t as a 4th semi-dimension. Thus, we get the sequence I (t) for each point with spatial coordinates x, y. The running average with 30 time points was used to produce an array I_s (t) of local n-point mean values, where each means is calculated over a sliding window of length *n* across neighboring elements of I(t).

When *n* is odd, the window is centered about the element in the current position. When *n* is even, the window is centered on the current and previous elements. The window is automatically truncated at the endpoints when there are not enough elements to fill it. When the window is truncated, the average is taken over only for the elements that fill the window. I_s (t) has the same size as I (t).

The final operation was the subtraction of I_s (t) from an original sequence I (t). This subtraction keeps only objects changing in the time domain such as comets, while static objects are wiped off. After subtraction, the average background intensity becomes about zero or negative, and background pixels have negative values (**Figure 1C**, blue profile). To remove residual noise from the image without affecting comet intensity profiles, the threshold bringing all negative values to zero was applied (**Figure 1D**).

When the intensity of the comet against the background is high, the level of the threshold will be less than 10% of the average level of the comets. Comets with moderate and low intensity, where thresholding cut a significant part of the profile, were not analyzed further.

Topological Comet Tracker Based on Gradient and Flood Fill Method

To find a correlation between tail length and velocity, our growing MT end tracker treats each comet as a 4D object, where the third spatial dimension is time and intensity is in a hidden 4th dimension, as described in the previous section. We next applied a flood fill algorithm to detect all bright objects in the stacks based on the gradient analysis. The comet pixel profile consists of pixels with brightness above the background and using the differences of intensities between adjacent pixels this algorithm determines a margin of comet head. The initial "seed" pixels were taken as having significant brightness, and a pixel cloud was formed around them (**Figure 2A–D**) for each comet. Object's edge was determined as forming a large gradient with the local pixel neighborhood (**Supplementary Material S1**).

Clouds of positive voxels were determined, and 3D profiles of the comets were generated (**Figure 2E**). Each comet is presented as a 4D object: spatial g (x, y, z), where z represents time (t) and I (intensity) is a "hidden" dimension. Since the comet is moving in time and space domains in the 3D representation (x, y, t), it appears as an elongated pillar (**Figure 2F**). The length of the pillar reflects the life span of the comet.

However, not all pillars represent individual comets—some comet's tracks collide with each other because Airy disks of the closely located MT tips overlap, and residual non-comet objects might be present (**Figure 2G**).

Since the distance between growing MT ends might be well below the resolution of the light microscope (Airy disk radius), we need decoupling to make sure that automated analysis is always assigned to individual tracks without jumping from one track onto another. To remove apparently colliding comets from further analysis, we decomposed obtained point cloud into single comet tracks by the "decoupling" method based on the topological structure of the cloud and determining features of comet tracks (lifespan, size, and intensity) (**Figure 3**). We removed all planes where pillar fusion (comet apparent collision) happened from the further analysis and dissected each object tree into several pieces, where each piece reflects a single comet track or part of the track (**Figure 3B–C**).

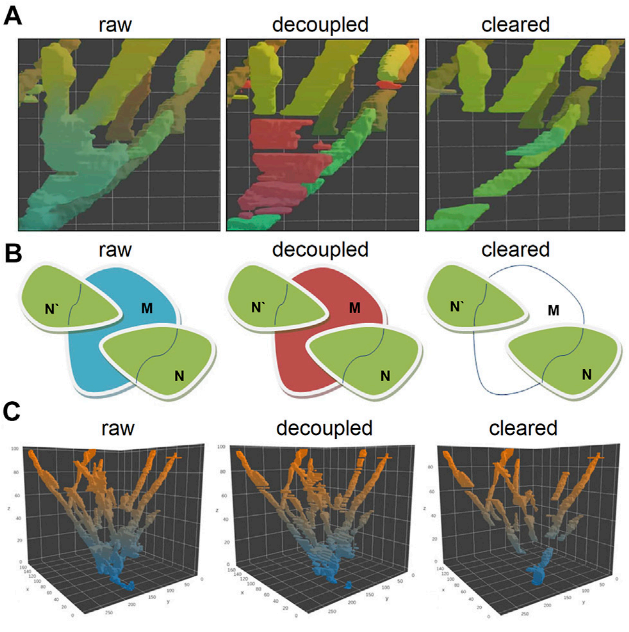


FIGURE 3 | Visualization of the plus end comet collision in a spatial–temporal **(A)** and 2D projection of the collision point **(B)**, before decoupling and after removing all non-comet objects. The decoupling algorithm compares the topological sets of XY coordinates with each other in T_n–T_{n+1} for a given tree object of objects N and N′ shift toward each other along the *Z*-axis form merged intersection object M at a certain point of time. The merged tracks of comets M belong to the T_{n+1} plane, track N, and track N′ belong to T_n. This object can be represented as a set M in T_{n+1} and as the sets N and N′ in T_n for the same X-Y coordinates. So, if M ∩ N′ and M ∩ N are non-empty sets and if the number of these non-empty sets are 2 or greater (for a greater number of comets), this means collision of comets occurs. For such cases, M set is removed from the voxel cloud and splits into several pieces. The 3D visualization of the decoupling process of the topological object point cloud **(C)**: after decoupling of the comet collision, event the solid "tree" shape object converts into a set of separate comet tracks.

The other non-comet objects were removed by analyzing the binary point cloud object mask in the 3D space. Since the comet moves in the (x, y, t) domain (where t means time), it forms the pillar structure with a certain "tilt" to the *z*-axis in a 3D coordinate cube. The objects with a small angle (nearly vertical) or occupying not more than three consecutive slices along the time axis were excluded.

Besides, we considered the presence of the noise and removed from further processing small objects with the integral size, below than 500 voxels, or the cross-section in the T_n smaller than 50 pixels (in one X–Y plane). Only pillar objects with relatively large cross sections and going through at least four consecutive X–Y planes were left for further analysis (**Figure 3C**, cleared). These binary masks are the tracks of individual comets allowing estimation of the velocity of MT growth. The overall number of individual comets used for further analysis was 31,623 in control HT1080 cells and 6,640 among the cells treated with nocodazole. A detailed analysis of the comet length and velocity is given further.

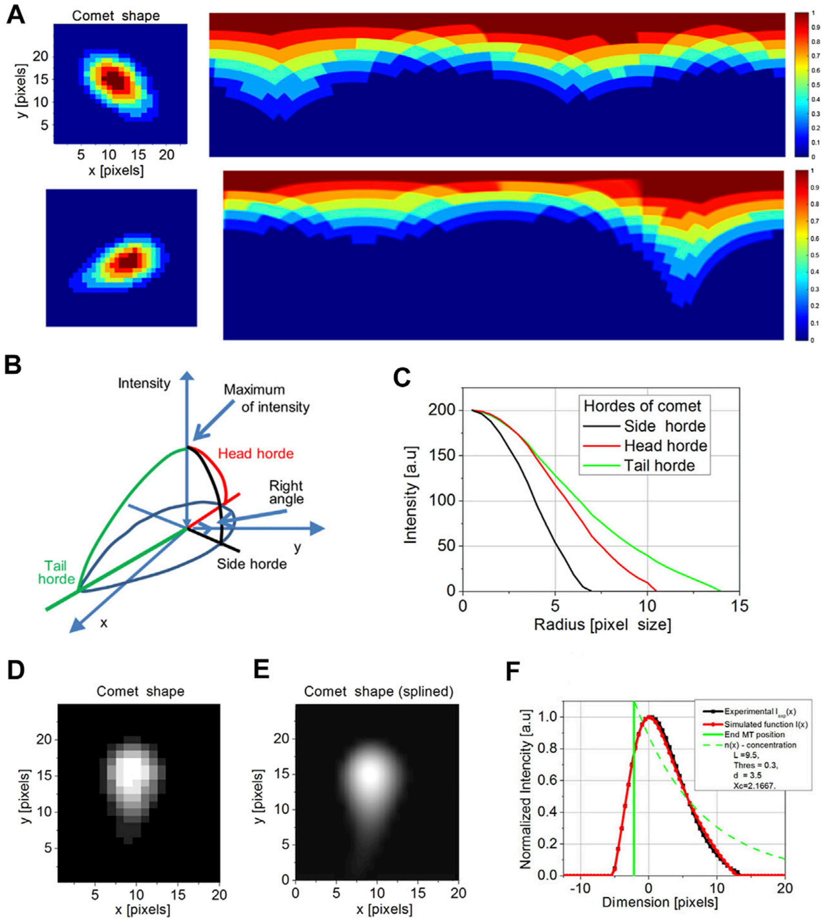


FIGURE 4 | Cartesian to polar transformation of a comet in a single time point for the comet with a short and long-tail **(A)**. The initial coordinates for transformation are set by maximum intensity pixel. The rainbow scale bar represents the intensity of each pixel. The model of comet intensities in XY and I (intensity) representation **(B)** was used to determine side (black) head (red) and tail (green) hordes. The model hordes of side (black), head (red), and tail (green) **(C)** were then compared with comet profiles. The original comet image is pixelized **(D)**. To compare it with a model, we perform a spline of comet images **(E)**. After determining the tail (red) and the real position of the tip of the microtubule (maximum in green), the comparison of exponential regression obtained by the current model (green), experimental profile (black), and simulated profile (red) was performed **(F)**. There is a significant shift between the MT end and an apparent maximum of intensity (set to 0).

Calculation of the Comet Length

Comet in a digital image is a 3D object with the XY coordinate and I intensity dimension. We assume all "true" comets having pixel profiles close to a cabochon shape with a roundish "head" and elongated "tail." The growing MT itself is a quasi-1D object. Therefore, there is no need to consider a 3D object. It is enough to consider a 2D object, which is a slice of the original object by a plane passing along the tube itself. It will also simplify calculations. The first step is to convert an image of this 2D object from Cartesian (x, y) to polar coordinates (r, ϕ), where the pole of new coordinates is located at the maximum of the comet intensity (**Figure 4A**). We assume that the center of intensity lies on the microtubule. Then we applied 64 angular profiles ("hordes," with a step Δϕ = 5.625) to determine the comet long axis direction (**Figures 4B,C**). To smooth the transition between coordinate systems, we used 10x rescaling by cubic splining (**Figure 4D**). As we mentioned before, the high- and low-pass Fourier filters were used. Therefore, our profiles are already "smooth," and using cubic spline routing does not cause a problem with additional mode spawning. Profiles of intensity were determined along with each horde. The comet "tail" horde is taken as the function with a maximal integral value. Comet "head" horde is determined as opposed to the tail horde, and the comet "side" hordes are determined as perpendicular to the direction of comet "tail".

We detect the length of the tail (comet length) from the entire profile of each comet formed by detecting the intensity signal produced by EB-3–RFP probe proteins (**Figure 4F**). According to the existing model of EB accumulation, the distribution of the EB-3 protein on the tip of the growing microtubule should have exponential decay due to a decrease in the number of EB-3–binding sites along growing MT (Bieling et al., 2008; Duellberg et al., 2016). Thus, the comet's tail intensity profile is as follows:

$$n(x,t) = N(t) * e^{-\frac{x}{L}},$$

where n (x, t) is a function of light intensity, N (t) is time dependent on the function that is considered at a certain time point, N (t) remains constant (A), and L is the required comet tail length:

$$\boldsymbol{n(x) = A * e^{-\frac{x}{L}}}.$$

We model the comet intensity profile acquired by the sCMOS detector array as the sum of Airy disks generated by individual EB-3-RFP molecules. It is denoted as I_{exp} (x_k).

The intensity of light I in the comet profile is a convolution of n(x) with PSF (point spread function) of the microscope (Maurer et al., 2014) that is well-approximated by Gaussian function (Zhang et al., 2007):

$$\boldsymbol{PSF(x) = e^{-\frac{x^2}{d^2}}},$$

where d of Gaussian is the width of PSF.

In our model, *d* is determined for each comet from the further regression analysis. The variable value of *d* represents slight defocusing of the comet images obtained by wide-field microscopy.

The real position of the comet tip (x_c) is shifted from the maximum of intensity in the microscopic image (Maurer et al., 2014). The shift between maximum intensity position and concentration function n(x) denoted as x_c is described in simulation function:

$$\begin{cases} \boldsymbol{n(x) = A*e^{-\frac{x - x_c}{L}}}; \; \boldsymbol{x > x_c} \\ \boldsymbol{n(x) = 0; x < x_c} \end{cases}.$$

Total comet profile intensity will be the product of convolution:

$$\boldsymbol{I(x) = (n(x) * PSF(x))(x)}.$$

And, since n(x) and PSF depends on the intensity function,

$$\boldsymbol{I(x) \sim \sum_{x_k} n(x_k) * e^{-\frac{(x-x_k)^2}{d^2}}}.$$

Another important parameter in the model is T (threshold) used to cut off pixels with low intensity and assigning their values to zero (**Figure 1D**). Thus, we added a condition related to the thresholding operation: ***if*** $\boldsymbol{I(x)} < 0$ ***then*** $\boldsymbol{I(x)} = 0$; as mentioned earlier, we used time averaging and thresholding in preprocessing of images. In our way of preprocessing, the threshold is a slowly changing function and for different comets is not constant. Threshold affects the shape of the intensity curve, and we considered it. Thus, to increase fitness accuracy, we added a condition related to the thresholding operation: ***if*** $\boldsymbol{I(x)} < 0$ ***then*** $\boldsymbol{I(x)} = 0$.

Using $\boldsymbol{T}$, $\boldsymbol{x_c}$, $\boldsymbol{d}$, and $\boldsymbol{L}$ as variables, we perform fitting of the real comet profile with the simulated one. The simulated profile was constructed as follows:

$$\begin{cases} \boldsymbol{I(x) = \sum_{x_k} n(x_k) * e^{-\frac{(x - x_k)^2}{d}} - T} \\ \boldsymbol{and\ if\ I(x) < 0\ than\ I(x) = 0} \end{cases}.$$

To find the best fit values of $\boldsymbol{T}$, $\boldsymbol{x_c}$, $\boldsymbol{d}$, and $\boldsymbol{L}$ for a comet, the regression with the experimental function I_{exp} (x_k) was performed:

$$\boldsymbol{(T,\ x_c,\ d,\ L) = arg \min_{T,\ x_c,\ d,\ L} \left(\sum_k \left(I(x_k) - I_{exp}(x_k) \right)^2 \right)}. \tag{4}$$

By the regression analysis, we obtained the following comet features: the threshold weight T, x_c that is the coordinate of the MT end in the local system of coordinates, *d* is the blur of PSF, and finally, L is the length of comet tail (distance where intensity drops down e times from the maximum). Reconstruction of the intensity profile of a comet is shown in **Figure 4F**. Two parameters are taken for further consideration: position of the MT end (x_c) and comet length (L).

The intensity maxima of comets had been obtained from local polar coordinates (x_c) was then returned to the global Cartesian coordinate system:

$$\begin{cases} \boldsymbol{x_{end} = x_{max} + x_c * cos(\varphi_{head})} \\ \boldsymbol{y_{end} = y_{max} + x_c * sin(\varphi_{head})} \end{cases},$$

where $(\boldsymbol{x_{max}}, \boldsymbol{y_{max}})$ are coordinates of the maximum of the intensity of comet in a global system of coordinates and $\boldsymbol{\varphi_{head}}$ is the angle of the head horde.

Measuring the Velocity of the MT Growth

For the determination of the velocity of MT growth, we next analyze comet head displacement between successive frames. The glowing comet appears trembling on the sequence of images while passing through the interior of the cell. It might represent ambiguity of our measurements due to "fishtailing" (lateral displacement) of the MT ends. In manual measurements, by connecting the consecutive comet head positions with a straight line when time discrete is short (<2–3 s) frequently gives relatively high angles between comet long axis and comet displacement vector (**Figures 5A,B**). This non-axial displacement might cause an overestimation of true MT growth events if the angle is too big (**Figures 5C,D**). The tracking algorithm keeps the only comet within a 35°angle between the displacement vector and comet axis (**Figure 5E**).

The "instantaneous" velocity $\boldsymbol{\vartheta}$ is measured as a distance between MT ends of the tracked comet between the z plane and z+1 plane:

$$\vartheta = \frac{\sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^2 + (\boldsymbol{y}_1 - \boldsymbol{y}_2)^2}}{\boldsymbol{t}} = $$

$$\frac{\sqrt{\left(\boldsymbol{x}_{\mathbf{max1}} - \boldsymbol{x}_{\mathbf{xmax2}} + (\mathrm{x}_{c1} - \mathrm{x}_{c2} * \boldsymbol{cos}(\varphi_{\mathbf{head}}))^2 + (y_{max1} - y_{xmax2} + (\mathrm{x}_{c1} - \mathrm{x}_{c2}) * \boldsymbol{sin}(\varphi_{head}))^2\right)}}{\boldsymbol{t}}.$$

For different time intervals, we obtained growth velocity distributions for both cell lines close to the Gaussian one (**Figure 6C**), which means the elongation of MTs in these cells could be approximated as growth with constant velocity (with random normal noise). A MT velocity histogram for HT 1080 cells is approximated by Gaussian distribution, with μ describing the Gaussian maximum coordinates and σ the function width at the decrease in e time. Growth velocity for HT-1080 is

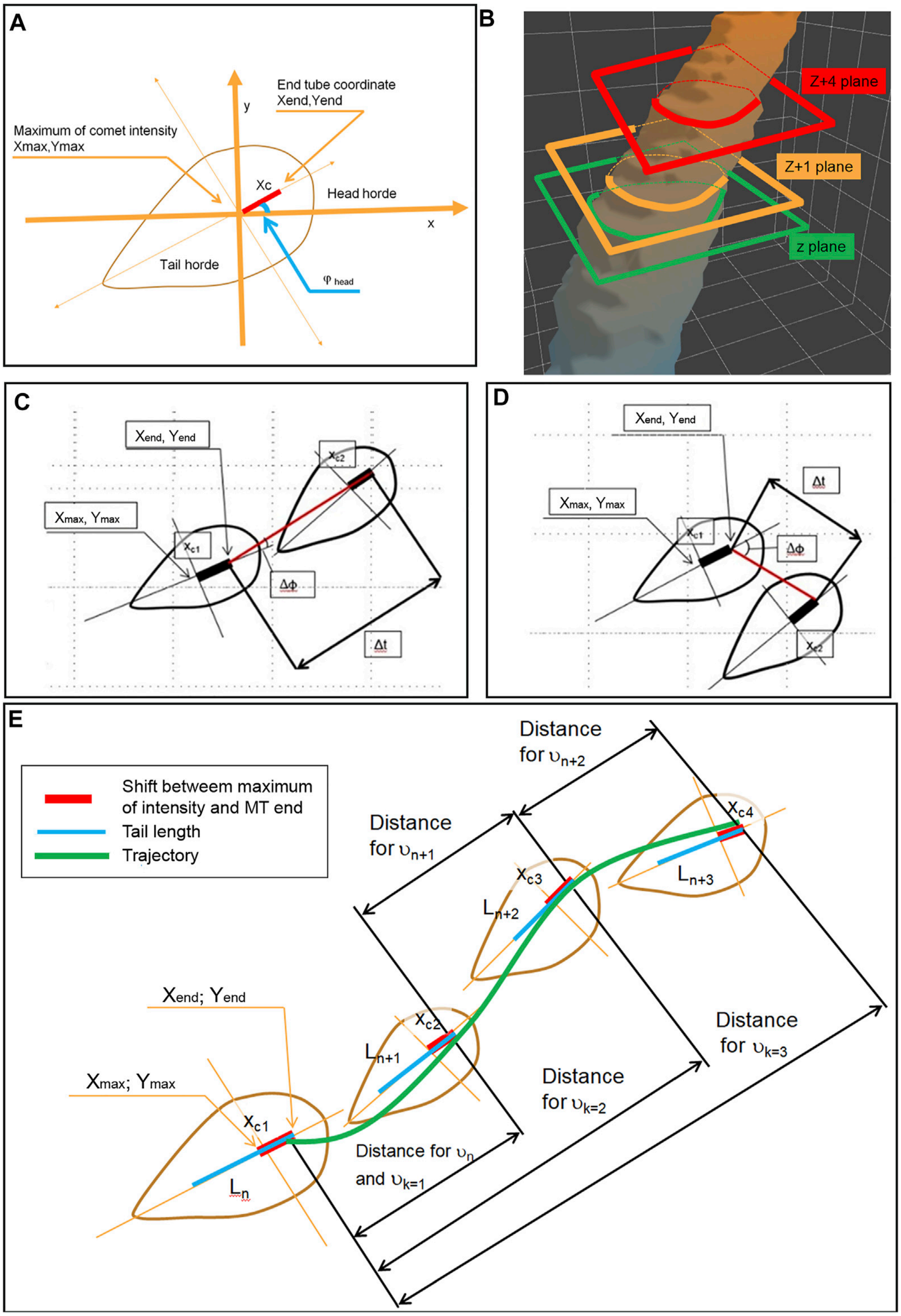


FIGURE 5 | Determination of the MT growth trajectory and measurements of the velocity by determination of the head horde and scheme of MT end coordinate measurements **(A)** and by neighbor planes, z and z+1, relates to time t and time t+1 on the topological object **(B)**. The plane z+4 relates to t+4 in case when the direction of a comet (head-to-tail axis) is close to MT growing direction (red line) **(C)** and in case where the direction of a comet is different from MT growing direction **(D)**. **(E)** The velocity measurements performed on the comets with both directions coinciding (with L_n, L_{n+1}, L_{n+2}, and L_{n+3} were measured sequentially). The distances for the measurement of v_n, v_{n+1}, and v_{n+2} are shown. Also, distances for calculation $v_{k=1}$ $v_{k=2}$ $v_{k=3}$ are shown. The green line shows the splined trajectory of the MT growth over few frames.

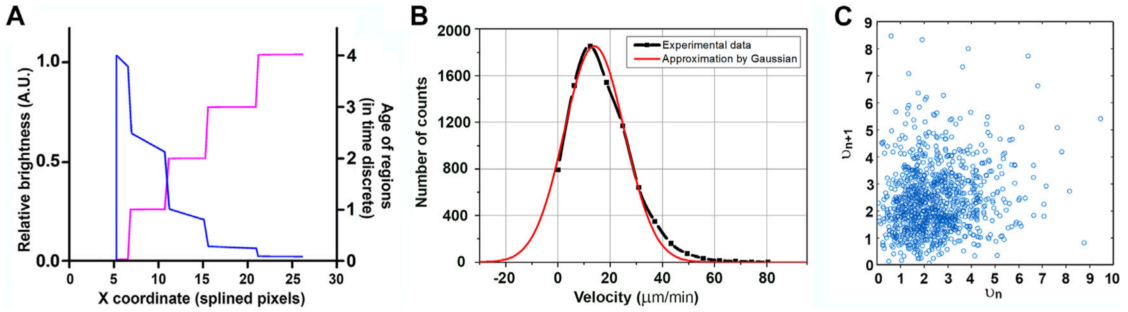


FIGURE 6 | EB-3 intensity by piecewise approximation. The older the ROI, the lower is the signal in it. A signal is recorded intermittently, and piecewise approximation represents the expected intensity distribution within the comet if the affinity of the MT to EB-3 decreases exponentially with time (blue line) **(A)**. The "age" of regions depended on the x coordinate for the microtubule (purple line). **(B)**. MT velocity histogram for HT 1080 cells approximated by Gaussian distribution. *Y*-axis: comet counts. μ = 13.9 μm/min; σ = 16.3 μm/min; R^2>0.98. **(C)** Lack of correlation between growth velocities (given in A.U.) in subsequent time periods (v_n versus v_{n+1}).

μ = 13.9 μm/min; σ = 16.3 μm/min (large σ reflects the presence of negative velocities (see Maurer et al., 2014 for more details). It is important to notice that instantaneous velocities *in vivo* rapidly change, and taken in subsequent time intervals do not correlate with each other (**Figure 6D**).

Modeling MT Growth

The analysis of the correlation between the comet length and microtubule growth rate is not straightforward. We obtain the comet tail length directly from one snapshot, while the velocity comes from a measurement of displacement between two consecutive snapshots. It means the velocity is measured between the moments when the tail length is determined, generating uncertainty when both values are stochastically changing. Thus, a more accurate result will be obtained when determining a correlation between averaged values of velocities and/or tail lengths. Addressing this question, we developed the computer model of microtubule growing. We use a model with parameters that can be found experimentally by developed code, which simplifies it compared to the one proposed earlier for *in vitro* measurements (Rickman et al., 2017).

In our model, the MT tip is considered as a 1D object which grows along the axis x. The velocity of growth is a random variable with normal (Gaussian) distribution. This statement has good agreement with experimental data (**Figure 6B**) and *in vitro* data (Vasquez et al., 1997; Dragestein et al., 2008; Rickman et al., 2017).

The MT comet image could be decomposed into several "regions of interest" (ROI on **Figure 6A**). Each ROI is generated by a single snapshot taken by the camera and represents elongation of the MT tip during a time interval between successive snapshots. In the proposed model, every region has its length depending on the growth rate and intensity depending on its age.

We set the time discreet in our model equal to one. At zero time point, the length of the MT comet is zero, and the regions are built subsequently one by one following time discrete. The age of the first region in a sequence is assigned to zero, its length is t*V_{rand}, where t = 1 is time discrete and V_{rand} is velocity equal to some random value distributed normally. The coordinate of the MT end becomes $X_1 = X_0 + 1 * V_{rand}$. During the next period, another region is added, and the previous region will fade. This new region has zero age, but for the previous region, I will now have age 1. This process will continue, and in the time period, n coordinates of the MT end will be $X_{n+1} = X_n + V_{rand}$. Because of the gradual disassembly of EB-3 from the tubulin (MT wall) behind the MT tip (Galjart, 2010), the intensity of each region will decrease with time until it drops below the threshold.

At some moment n, we will have X_n sequence of all regions that are above the threshold (**Figure 6A**). The age of the first region is always zero. The number of regions might be less than n because of negative velocity (MT shortening). For every region, the concentration of EB-3 decreases along with its age by exponential law:

$$\boldsymbol{n}(\mathrm{T}) = \boldsymbol{e}^{-\frac{T}{\mathrm{DT}}}.$$

The half-time of the existence of EB-3-binding sites at microtubule plus ends is termed "decoration time" or DT (Bieling et al., 2008). Time (T) in our model is determined as a discrete function as the age of ROI. So, the dependence of the concentration of EB on the x coordinate as continuous function could be described as follows (**Figure 6B**, orange line):

$$\boldsymbol{n}(\mathrm{X}) = \boldsymbol{e}^{-\frac{T(X)}{\mathrm{DT}}}.$$

Analysis of the comet length and displacement requires precise determination of the real position of the MT tip.

Assuming the growth velocity is constant, the intensity of a given region of the comet is equal to maximal intensity divided by DT (decoration time). However, the real value deviates from it because of stochastic variations. The question is how to determine the position and real intensity of the comet maximum. The comet intensity profiles were previously approximated by Gaussponent (Mustyatsa et al., 2019). Because of the image convolution by the microscope, the position of the exponent maximum in this model is localized within the "Gaussian" part. Our model takes into account that the MT tip is shifted from the brightest point on the intensity profile because of the image blurring by the microscope. The amount of shift depends on the comet brightness (Maurer

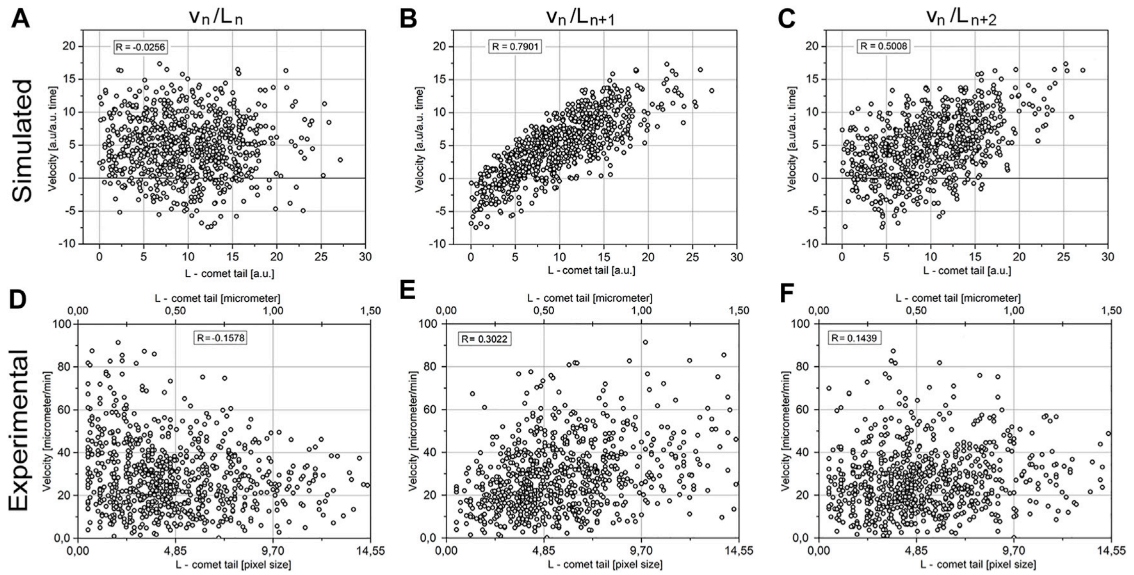


FIGURE 7 | Correlation between the velocity of MT elongation and tail length. The comparison of the simulated model data **(A–C)** and the experimental data **(D–F)**. Images **(A, D)** represent the correlation between v_n and L_n, image **(B, E)** represents v_n/L_{n+1}, and **(C, F)** represents v_n/L_{n+2}.

et al., 2014) and is relatively large (319 ± 164 nm for HT1080 cells).

Another ambiguity comes from measurements of the comet's velocity. The comet profile is obtained by summing the signals from all fluorescent molecules for the exposure time. When velocity is nearly constant for a long time, the intensity profile could be determined by using frame averaging (Rickman et al., 2017). When apparent velocity fluctuates significantly, averaging might give biased results. To overcome the possible problem caused by superaveraging, we developed the model, allowing determination of the full length of an individual comet and position of its maximum with subpixel precision. Analysis of MT growth velocity performed in a standard way and taking into account the real position of the MT tip (according to **Figure 5E**) show that velocity in HT1080 cells could be well-approximated by Gaussian distribution (**Figure 6C**). It is important to notice that velocities determined in subsequent short time intervals (0.5 s) do not correlate with each other (**Figure 6C**).

On the Correlation Between Comet Tail Length and MT Growth Velocity

Another ambiguity comes from measurements of the comet's velocity. The MT growth velocity is calculated from the distance between two consecutive measurements, while comet length is determined only at the beginning and the end of this period. It causes uncertainty in determining a correlation between the velocity (v_n) and comet length (L_n) (**Figure 7**). The average velocity of MT growth measured *in vitro* remains nearly constant over long periods—more than 30–40 s (Rickman et al., 2017), while *in vivo* usually only relatively short tracks of MT growth could be visualized and instantaneous growth velocity seemingly varies (Vorobjev et al., 1999; Matov et al., 2010).

Addressing this question, we analyzed the correlation between v_n and L_{n+t}, where t is the time interval against v_n. Taking into account a relatively large dataset (more than 6,000 events), we assumed that all values of the Pearson coefficient R > 0.1 will be significant. The model predicts no correlation between v_n and L_n, the highest correlation between v_n and L_{n+1}, and decreased correlations between v_n and L_{n+k}, where k > 1 (**Figures 7A–C**).

The experimental data are in good agreement with the model. There is maximal R = 0.3 for v_n/L_{n+1} for experiment (**Figure 7E**) and R = 0.79 for model (**Figure 7B**). For v_n/L_{n+2} both for experiment and model there is reducing correlation: experiment R = 0.14 (**Figure 7F**) and model R = 0.5 (**Figure 7C**). There is good agreement between experiment and model.

The modeled data have higher coefficients of correlation than experimental ones since the model is ignoring errors in measurements. Also, in the model, MT is assumed to grow along the straight line, that is, in one dimension, while in the experiment, MT growth deviates from the straight line and occurs in the 2D space.

The only significant discrepancy between the model and experimental data is correlation v_n/L_n, where simulation of the model data gives R = −0.02 that is below the level of confidence (**Figure 8A**), while in experimental data, the correlation between v_n/L_n in an experiment has a negative value of R = −0.15 (**Figure 8D**) that is by a module larger than the confidence level. This means an inverse correlation between the comet length and growth velocity of MTs in the following period of time exists in the cells.

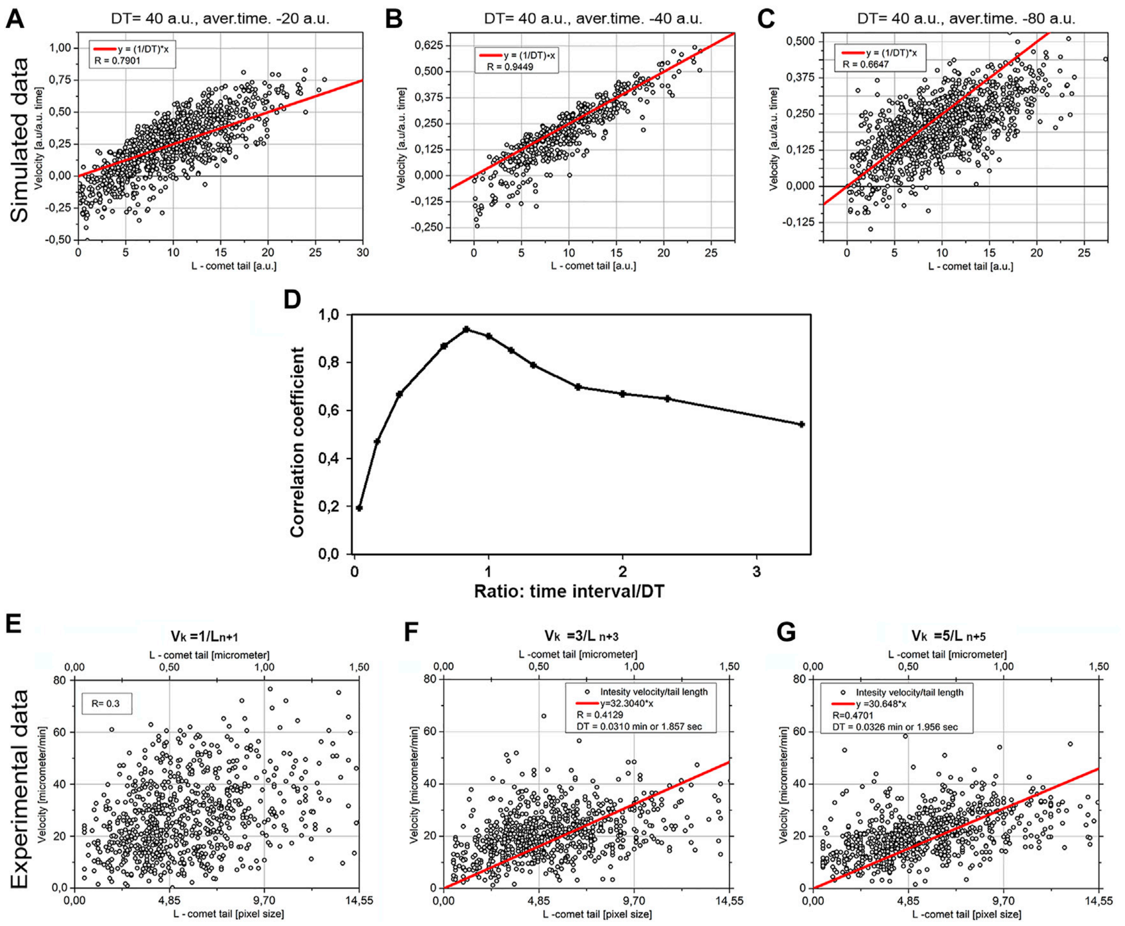


FIGURE 8 | Correlation between tail length and averaged velocity of MT elongation. The comparison of the simulated model data **(A–C)** and the experimental data **(D–F)** retrospectively. The model **(A)** DT = 40 a.u., averaging time = 20 a.u., **(B)** DT = 40 a.u., averaging time = 40 a.u., **(C)** DT = 40 a.u., averaging time = 80 a.u. The correlation of tail length/velocity for $v_k = 1/L_{n+1}$ in experimental data **(E)** the correlation (R) for $v_k = 3/L_{n+3}$ **(F)**, and correlation for $v_k = 5/L_{n+5}$ **(G)**. For **(F, G)** the calculated regression line y = ax also shown.

Finding Decoration Time for EB Protein on MTs *in vivo*

For the MTs, growing *in vitro* decoration time was found by dividing the comet tail length by the MT growth speed, that is, as a coefficient of the slope of line L/v (Bieling et al., 2008). However, for the *in vivo* data we have obtained, both parameters (L and v) are fluctuating, and the maximal correlation coefficient R is equal to 0.3 (**Figure 8E**), precluding finding the coefficient of the slope with reliable precision (i.e., with large S.D.-to-mean ratio). Thus, another approach was needed. Our model predicts that maximal length/velocity correlation occurs when the time interval for the velocity measurement is close to the DT value (**Figure 8D**), and comet length is taken from the frame after the interval where velocity was determined.

The averaged velocity of MT growth was obtained by skipping intermediate planes: we calculated velocities from the shifts of MT tip between planes (time points)—z and z+2, z and z+3, etc. (**Figure 8**). Denoting k as a distance between sequential planes, the average velocity will be as follows:

$$\vartheta_k = \frac{\sqrt{(x_n - x_{n+k})^2 + (y_n - y_{n+k})^2}}{k * t},$$

where $\sqrt{(x_n - x_{n+k})^2 + (y_n - y_{n+k})^2}$ are distances obtained for different v_k (**Figure 8D**), t defines the time interval between successive frames (500 ms), and v_k is velocity averaged for the time interval k*t. Should note that higher k required a long trajectory for a comet. For k = 5, for example, the comet must have no collision with any other comet or any other object for 2.5 s. As a consequence, the number of these comets is not very large. For example, the number of points for k = 5 is about 800, for k = 6 about 300. In this case, we need to keep the balance between a good statistic and good correlation coefficients.

The result for modeling is shown in **Figure 8**. Three cases are shown: when time-averaged is less than DT (**Figure 8A**), when averaged time is equal to DT (**Figure 8B**), and averaged time exceeds DT (**Figure 8C**). The maximal correlation value (R = 0.9449), shown in **Figure 8B**, means that averaging time represents DT. The overall dependence of correlation

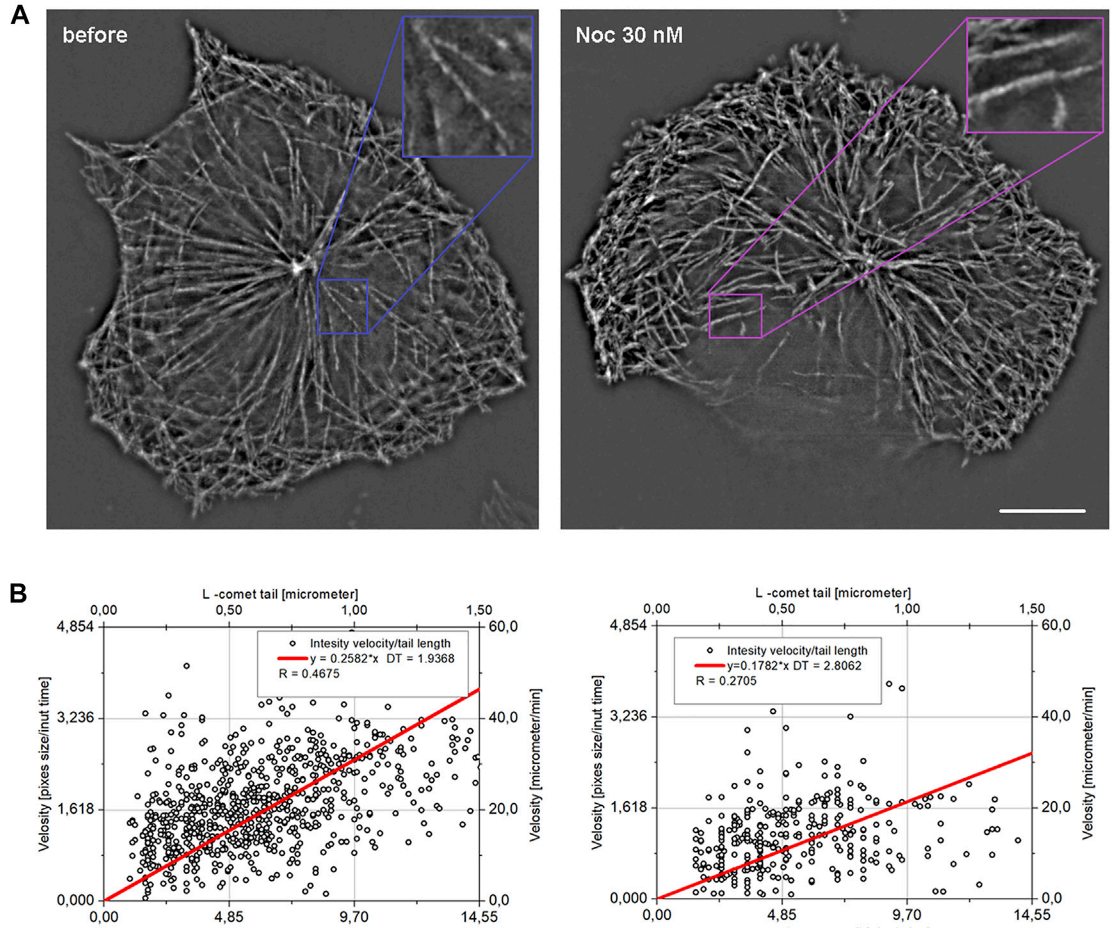


FIGURE 9 | Projections showing MT growth in HT 1080 **(A)** cells before and after treatment with nocodazole. Images were taken for 2-s intervals (each 4th frame) and merged as maximal intensity projection. In the internal cytoplasm, faster translocation of the EB comets in the control cell is evident (enlargement). The length–velocity correlation **(B)** for k = 4 ($v_{k=4}$ versus L_{n+4}, 2 s interval) for control cells and cells after nocodazole treatment are shown.

coefficient versus time/DT ratio in the model is given in **Figure 8D**. Thus, the model confirms the applicability of the method for estimating DT as the coefficient of a slope when R reaches its maximum.

We next plotted experimental data v_k versus L_{n+k} for different k values (**Figures 8E–G**). The maximal R coefficient (R = 0.47) was obtained for k = 5, which means the averaging time is 2.5 s. Linear regression y = a×x was made fitting the data on **Figures 8F,G**. This regression gives the coefficient of the slope a = 0.51–0.53. DT in this case will be DT = 1/a so DT = 1.86–1.96 s.

The increased coefficient of correlation with k is in good agreement with the model. In this case, the model predicts the maximum correlation for k equal or slightly greater than DT (**Figure 8D**). The found DT is about 2 s, which corresponds to k = 4, that is, k = 4~5 will be more or less optimal for our case. A further increase in k leads to a decrease in statistics—for k = 6, we will have only 300 points; for k = 7, only 170; and so on. Such small statistics will reduce the reliability of the data.

MT Growth Under Nocodazole Treatment

We further analyzed the effect of nocodazole applied in nanomolar concentrations. In the range of concentrations, 30–50 nM, the velocity of MT growth estimated by manual tracking did not significantly change (Serikbaeva et al., 2018). However, in some cases, smaller displacements of individual MT ends were evident (**Figure 9A**). Automated analysis using a much larger sample demonstrated a decrease in the mean velocity –236 ± 0.150 nm/s compared to 329 ± 171 nm/s (19.74 µm/min) in control cells (p < 0.0001) and an even more profound decrease in the comet length –525 ± 245 nm compared to 1,252 ± 624 nm in control (p < 0.0001).

Since the maximum correlation between comet tail length and growth velocity will be when velocity is averaged on time ~ DT, we used averaged velocity on k = 4 that is relate to 2 s (**Figure 9B,C**). The R coefficient = 0.27 obtained after treatment is higher than the confidence level (for 300 cases it is about 0.2), but rather low compared to untreated cells. Based on these data, DT is 2.8 s (determined from the linear approximation slope). Thus, we conclude that DT for EB-3 is the same for reference and nocodazole-treated cells. EB comets were also found at higher nocodazole concentration, but MTs were nearly stationary under such treatment, and it was not possible to determine the L value (data not shown).

DISCUSSION

The developed model for MT comet approximation from individual images of plus end comets and new MT tip tracker shows good capabilities to extract the spatial and temporal characteristics of MT comets in living cells. The advantage of our model is that it gives the real position of the MT tip for a given comet with the accuracy not achieved before for the *in vivo* imaging. Our algorithm of comet length determination is similar to the Gaussponent approximation developed previously (Mustyatsa et al., 2019) but takes into account the shift of the real comet tip from its apparent position in the microscopic image (Maurer et al., 2014, and **Figure 4F**).

Theoretically, the position of the MT tip could be calculated precisely from the deconvolution of the comet image using the microscope point spread function (PSF) (Maurer et al., 2014). However, iterative deconvolution requires 3D imaging (Dias-Zamboni et al., 2007) or could be applied only to relatively homogenous datasets (Maurer et al., 2014; Rickman et al., 2017). Data for successful deconvolution of MT comet profiles were obtained by superaveraging of hundreds to thousands of individual images recorded *in vitro* (Maurer et al., 2014; Rickman et al., 2017). This approach cannot be used *in vivo* because the MT elongation velocity at the population level in the cultured cells is not constant, instead varies significantly even within one cell (Serikbaeva et al., 2018).

Our observations show that approximation of the comet profile with subpixel resolution could be achieved without averaging. Thus, it was possible to determine the velocity of MT growth with higher precision than the previous analysis when comet displacement was determined as the distance between maximal intensity pixels in the comet profile in sequential images (Matov et al., 2010; Applegate et al., 2011; Serikbaeva et al., 2018).

Measurements of MT growth velocity in the current study show that despite the heterogeneity of the actual velocity of MTs, correlation between velocity and length of an EB comet determined for individual growing MTs *in vivo* is significant. The maximum observed coefficient of correlation is equal to R = 0.47 when the time of averaging was about or slightly longer than decoration time (DT). In the HT1080 and U-118 cells, the decoration time is about 2 s, and the best correlation between v_k and L_{n+k} is observed for k = 4~5 (k is a time discrete, which in our case is 0.5 s). This coefficient is highly significant. Good agreement between model and experimental data, in general, confirms the mechanism of MT growth that was earlier proposed for *in vitro* conditions (Galjart, 2010).

This suggestion is confirmed by observations of cells treated with a low dose of nocodazole (30 or 50 nM). Under nocodazole treatment, significant reduction of both growth velocity and comet length occurs, velocity drops down to 236 ± 150 nm/s compared to 329 ± 171 nm/s (19.74 um/min) in untreated cells ($p < 0.0001$), and the comet length shrinks to 525 ± 245 nm against 1,252±624 nm in untreated cells ($p < 0.0001$). The correlation of velocity and comet length after treatment with nocodazole exists, and DT seems to be the same as without treatment.

MT instantaneous growth velocity distribution determined *in vivo* fits in the normal distribution (**Figure 6B**), and velocities determined in subsequent time intervals do not correlate with each other (**Figure 6C**). The decoration time for EB-3 assembly on the growing MT tip for HT1080 cells is 2–2.5 s, which is significantly shorter than *in vitro* (Bieling et al., 2008). We suggest that faster accumulation of EB proteins on the growing MT tips *in vivo* than the *in vitro* measurements is achieved due to the action of additional plus TIPs, such as CLASPs facilitating assembly of EB proteins with polymerized tubulin (Watanabe et al., 2009; Grimaldi et al., 2014).

CONCLUSION

Automated unbiased tracking of MT plus ends decorated with EB-3 shows that elongation of individual MTs in cultured cells could be approximated by the constant velocity with random fluctuations and thus do not differ from the dynamics of MTs observed *in vitro*. Large variations of the instantaneous growth velocity *in vivo* compared to the *in vitro* measurements could be explained by the shorter decoration time. Shorter decoration time *in vivo* could be explained by the presence of other decorating proteins like chTOG, assembling at MT tip in living cells before EB (Nakamura et al., 2012). Growth velocity measured at time intervals close to the decoration time and comet length taken at the end of this period correlate with each other, and this correlation remains under low-dose nocodazole treatment when both velocity and comet length is significantly reduced.

MATERIALS AND METHODS

Cell Culture, Transfection, and the Microtubule Drugs

Experiments were performed on HT1080 and U-118 cultured cells. Cells were maintained at DMEM supplemented by 10% fetal bovine serum in the presence of 4–8 μm of L-glutamine and 100 U/ml penicillin/streptomycin mix. The transfection was performed with Evrogen plasmid (Cat. No. FP142) carrying EB-3-TagRFP fusion protein (EX.555 nm/EM.584 nm) with Xtreme ROCHE transfection reagent (Sigma, Cat. No. 6365787001) according to the manufacturer's protocol. The cells were maintained in the standard growth media. Nocodazole (Sigma, Cat. No. M1404-10MG) in the concentrations of 30 and 50 nM was used as a microtubule growth inhibition drug. For imaging, cells were subcultured onto 8-well glass-bottom plates (Nunc Lab-Tek II, Thermo Cat. No. 155409), and transfection was performed on the following day. The culture medium was replaced 24 h post-transfection, and microscopy recording was performed 36–48 h after transfection.

Live Cell Imaging

Fluorescent microscopy was performed on a Cell Observer SD microscope (Carl Zeiss GmbH), equipped with a heating incubator chamber. Imaging was performed on cells with an appropriate level of transfection without overexpression patterns. Comets chosen were resolvable from the background, and measured signal intensity was at least 4–5 times higher than the background of the cytoplasm in an area of 10–50 pixels of an image. These cells were recorded as a control in imaging CO_2-independent DMEM growth medium, and then coordinates of each cell were stored for further steps. After the media was carefully replaced with a drug-containing culture medium and cells were exposed for 60 min on the microscope stage, the temperature of 37°C was maintained by heating the incubator chamber. Imaging of treated cells was performed according to obtained coordinates to have valid pair of control (cell before treatment) and experiment (same cell after addition of nocodazole).

To obtain maximal intensity without photobleaching of comet during time-lapse, image acquisition was performed at a single focal plane in the wide-field fluorescence mode or using spinning disk confocal microscopy. Image sequences were recorded with 500-ms step, with an exposure time of 300 ms using an HXP120 metal halide lamp and Rhodamine filter cube (Filter set 43 HE) and Hamamatsu ORCA V2 sCMOS camera (2,048 × 2,048). Some image samples were collected using a Yokogawa CSU-X1 spinning disk confocal unit using a 561-nm laser and Rolera EM-CCD detector (1,024 × 1,002). Imaging in both setups was performed with Plan Apochromat 63x/1.46 oil immersion objective, and the image pixel size was 103 nm. On five separate experimental days, we recorded time-lapse videos of HT-1080 and U-118 transfected cells. Images were obtained using Zen software and exported as 16-bit TIFF uncompressed image sequences for processing. The original tiff image stacks were cropped manually to the margins of the individual cell which contained clearly visible microtubule comets. Each image stack contained 100 images. The number of individual cells (recorded before treatment and recorded the second time after nocodazole treatment for >30 min) for each experimental day was about 50–100. The image sets obtained after treatment with nocodazole at concentrations of 30 and 50 nM were further merged, while images obtained under nocodazole 100 nM treatment were excluded from analysis due to significant shrinkage of EB comets precluding measurement of their length. In total, we collected 486 image sequences of individual HT1080 cells before treatment and 328 cells after nocodazole treatment (30 nM + 50 nM). Further analysis was performed in MATLAB and C++ environments, as described in the Results section. Cells, where photobleaching was significant, were discarded from automated analysis. The tool and commentary uploaded in https://github.com/astrv-103/tracker.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

AU created the image processing tool and the model and performed the modeling in silica and data processing and data interpretation. AS performed the cell transfection, cell imaging, and manual image processing. AT performed the cell transfection, cell imaging, and manual image processing. AD performed the 3D visualization of growing comets and data processing. SK performed imaging condition optimization, cell transfection, cell imaging, and image preparation for image processing. IV suggested the model, the experimental design, and data collection and performed mentoring of the data collection and discussion on result interpretation. AU, AS, AD, AT, and SK drafted the manuscript results and method. AU and IV drafted the discussion. IV performed the peer-review and composing of a final version of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants AP08857554 (MES, Kazakhstan) and ORAU 240919FD3937 (Nazarbayev University) to IV. The research grant from MES Kazakhstan AP08053387 provided support for AU

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.745089/full#supplementary-material

Supplementary Video S1 | Related to **Figure 1**. Time-lapse sequence of fluorescence images of EB-3 comets translocation.

REFERENCES

Applegate, K. T., Besson, S., Matov, A., Bagonis, M. H., Jaqaman, K., and Danuser, G. (2011). PlusTipTracker: Quantitative Image Analysis Software for the Measurement of Microtubule Dynamics. *J. Struct. Biol.* 176 (2), 168–184. doi:10.1016/j.jsb.2011.07.009

Bieling, P., Kandels-Lewis, S., Telley, I. A., van Dijk, J., Janke, C., and Surrey, T. (2008). CLIP-170 Tracks Growing Microtubule Ends by Dynamically Recognizing Composite EB1/tubulin-Binding Sites. *J. Cel. Biol.* 183 (7), 1223–1233. doi:10.1083/jcb.200809190

Bieling, P., Laan, L., Schek, H., Munteanu, E. L., Sandblad, L., Dogterom, M., et al. (2007). Reconstitution of a Microtubule Plus-End Tracking System *In Vitro*. *Nature* 450 (7172), 1100–1105. doi:10.1038/nature06386

Bowne-Anderson, H., Zanic, M., Kauer, M., and Howard, J. (2013). Microtubule Dynamic Instability: a New Model with Coupled GTP Hydrolysis and Multistep Catastrophe. *Bioessays* 35 (5), 452–461. doi:10.1002/bies.201200131

Brouhard, G. J., and Rice, L. M. (2018). Microtubule Dynamics: an Interplay of Biochemistry and Mechanics. *Nat. Rev. Mol. Cel Biol* 19 (7), 451–463. doi:10.1038/s41580-018-0009-y

Chanez, B., Gonçalves, A., Badache, A., and Verdier-Pinard, P. (2015). Eribulin Targets a Ch-TOG-dependent Directed Migration of Cancer Cells. *Oncotarget* 6 (39), 41667–41678. doi:10.18632/oncotarget.6147

Dias-Zamboni, J., Paravani, E. V., Adur, J. F., and Casco, V. (2007). Implementation of an Iterative Deconvolution Algorithm and its Evaluation on Three-Dimensional Images of Fluorescent Microscopy. *Acta Microscopica* 16 (1-2), 8–15.

Dragestein, K. A., van Cappellen, W. A., van Haren, J., Tsibidis, G. D., Akhmanova, A., Knoch, T. A., et al. (2008). Dynamic Behavior of GFP-CLIP-170 Reveals Fast Protein Turnover on Microtubule Plus Ends. *J. Cel. Biol.* 180 (4), 729–737. doi:10.1083/jcb.200707203

Duellberg, C., Cade, N. I., Holmes, D., and Surrey, T. (2016). The Size of the EB Cap Determines Instantaneous Microtubule Stability. *eLife* 5, e13470. doi:10.7554/eLife.13470

Galjart, N. (2010). Plus-end-tracking Proteins and Their Interactions at Microtubule Ends. *Curr. Biol.* 20, R528–R537. doi:10.1016/j.cub.2010.05.022

Gard, D. L., and Kirschner, M. W. (1987). Microtubule Assembly in Cytoplasmic Extracts of Xenopus Oocytes and Eggs. *J. Cel. Biol.* 105 (5), 2191–2201. doi:10.1083/jcb.105.5.2191

Grimaldi, A. D., Maki, T., Fitton, B. P., Roth, D., Yampolsky, D., Davidson, M. W., et al. (2014). CLASPs Are Required for Proper Microtubule Localization of End-Binding Proteins. *Developmental Cel.* 30 (3), 343–352. doi:10.1016/j.devcel.2014.06.026

Guesdon, A., Bazile, F., Buey, R. M., Mohan, R., Monier, S., García, R. R., et al. (2016). EB1 Interacts with Outwardly Curved and Straight Regions of the Microtubule Lattice. *Nat. Cel Biol* 18 (10), 1102–1108. doi:10.1038/ncb3412

Hiller, G., and Weber, K. (1978). Radioimmunoassay for Tubulin: a Quantitative Comparison of the Tubulin Content of Different Established Tissue Culture Cells and Tissues. *Cell* 14 (4), 795–804. doi:10.1016/0092-8674(78)90335-5

Matov, A., Applegate, K., Kumar, P., Thoma, C., Krek, W., Danuser, G., et al. (2010). Analysis of Microtubule Dynamic Instability Using a Plus-End Growth Marker. *Nat. Methods* 7 (9), 761–768. doi:10.1038/nmeth.1493

Maurer, S. P., Cade, N. I., Bohner, G., Gustafsson, N., Boutant, E., and Surrey, T. (2014). EB1 Accelerates Two Conformational Transitions Important for Microtubule Maturation and Dynamics. *Curr. Biol.* 24 (4), 372–384. doi:10.1016/j.cub.2013.12.042

Mitchison, T., and Kirschner, M. (1984). Dynamic Instability of Microtubule Growth. *Nature* 312 (5991), 237–242. doi:10.1038/312237a0

Mustyatsa, V. V., Kostarev, A. V., Tvorogova, A. V., Ataullakhanov, F. I., Gudimchuk, N. B., and Vorobjev, I. A. (2019). Fine Structure and Dynamics of EB3 Binding Zones on Microtubules in Fibroblast Cells. *MBoC* 30 (17), 2105–2114. doi:10.1091/mbc.E18-11-0723

Nakamura, S., Grigoriev, I., Nogi, T., Hamaji, T., Cassimeris, L., and Mimori-Kiyosue, Y. (2012). Dissecting the Nanoscale Distributions and Functions of Microtubule-End-Binding Proteins EB1 and Ch-TOG in Interphase HeLa Cells. *PLoS One* 7 (12), e51442. doi:10.1371/journal.pone.0051442

Odde, D. J. (1997). Estimation of the Diffusion-Limited Rate of Microtubule Assembly. *Biophysical J.* 73 (1), 88–96. doi:10.1016/S0006-3495(97)78050-0

Rickman, J., Duellberg, C., Cade, N. I., Griffin, L. D., and Surrey, T. (2017). Steady-state EB Cap Size Fluctuations Are Determined by Stochastic Microtubule Growth and Maturation. *Proc. Natl. Acad. Sci. USA* 114 (13), 3427–3432. doi:10.1073/pnas.1620274114

Schuyler, S. C., and Pellman, D. (2001). Microtubule "Plus-End-Tracking Proteins". *Cell* 105 (4), 421–424. doi:10.1016/s0092-8674(01)00364-6

Seetapun, D., Castle, B. T., McIntyre, A. J., Tran, P. T., and Odde, D. J. (2012). Estimating the Microtubule GTP Cap Size *In Vivo*. *Curr. Biol.* 22 (18), 1681–1687. doi:10.1016/j.cub.2012.06.068

Serikbaeva, A., Tvorogova, A., Kauanova, S., and Vorobjev, I. A. (2018). Analysis of Microtubule Dynamics Heterogeneity in Cell Culture. *Methods Mol. Biol. (Clifton, N.J.)* 1745, 181–204. doi:10.1007/978-1-4939-7680-5_11

Skube, S. B., Chaverri, J. M., and Goodson, H. V. (2010). Effect of GFP Tags on the Localization of EB1 and EB1 Fragments *In Vivo*. *Cytoskeleton* 67 (1), 1–12. doi:10.1002/cm.20409

Vasquez, R. J., Howell, B., Yvon, A. M., Wadsworth, P., and Cassimeris, L. (1997). Nanomolar Concentrations of Nocodazole Alter Microtubule Dynamic Instability *In Vivo* and *In Vitro*. *MBoC* 8 (6), 973–985. doi:10.1091/mbc.8.6.973

Vorobjev, I. A., Rodionov, V. I., Maly, I. V., and Borisy, G. G. (1999). Contribution of Plus and Minus End Pathways to Microtubule Turnover. *J. Cel Sci.* 112 (14), 2277–2289. doi:10.1242/jcs.112.14.2277

Watanabe, T., Noritake, J., Kakeno, M., Matsui, T., Harada, T., Wang, S., et al. (2009). Phosphorylation of CLASP2 by GSK-3β Regulates its Interaction with IQGAP1, EB1 and Microtubules. *J. Cel. Sci.* 122 (Pt 16), 2969–2979. doi:10.1242/jcs.046649

Zanic, M., Widlund, P. O., Hyman, A. A., and Howard, J. (2013). Synergy between XMAP215 and EB1 Increases Microtubule Growth Rates to Physiological Levels. *Nat. Cel Biol* 15 (6), 688–693. doi:10.1038/ncb2744

Zhang, B., Zerubia, J., and Olivo-Marin, J.-C. (2007). Gaussian Approximations of Fluorescence Microscope point-spread Function Models. *Appl. Opt.* 46 (10), 1819–1829. doi:10.1364/ao.46.001819

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.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 21 January 2022
doi: 10.3389/fmolb.2021.800819

Prediction of Residue-specific Contributions to Binding and Thermal Stability Using Yeast Surface Display

*Shahbaz Ahmed[1], Munmun Bhasin[1], Kavyashree Manjunath[2] and Raghavan Varadarajan[1]**

[1]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India,* [2]*Institute for Stem Cell Science and Regenerative Medicine, Bangalore, India*

Accurate prediction of residue burial as well as quantitative prediction of residue-specific contributions to protein stability and activity is challenging, especially in the absence of experimental structural information. This is important for prediction and understanding of disease causing mutations, and for protein stabilization and design. Using yeast surface display of a saturation mutagenesis library of the bacterial toxin CcdB, we probe the relationship between ligand binding and expression level of displayed protein, with *in vivo* solubility in *E. coli* and *in vitro* thermal stability. We find that both the stability and solubility correlate well with the total amount of active protein on the yeast cell surface but not with total amount of expressed protein. We coupled FACS and deep sequencing to reconstruct the binding and expression mean fluorescent intensity of each mutant. The reconstructed mean fluorescence intensity (MFI_{seq}) was used to differentiate between buried site, exposed non active-site and exposed active-site positions with high accuracy. The MFI_{seq} was also used as a criterion to identify destabilized as well as stabilized mutants in the library, and to predict the melting temperatures of destabilized mutants. These predictions were experimentally validated and were more accurate than those of various computational predictors. The approach was extended to successfully identify buried and active-site residues in the receptor binding domain of the spike protein of SARS-CoV-2, suggesting it has general applicability.

Keywords: mutational scanning, residue burial, saturation mutagenesis, free energy, protein stability

OPEN ACCESS

Edited by:
Paulo Ricardo Batista,
Oswaldo Cruz Foundation (Fiocruz), Brazil

Reviewed by:
Mattia Miotto,
Sapienza University of Rome, Italy
Alessandro Angelini,
Ca' Foscari University of Venice, Italy

***Correspondence:**
Raghavan Varadarajan
varadar@iisc.ac.in

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *24 October 2021*
Accepted: *14 December 2021*
Published: *21 January 2022*

Citation:
Ahmed S, Bhasin M, Manjunath K and Varadarajan R (2022) Prediction of Residue-specific Contributions to Binding and Thermal Stability Using Yeast Surface Display. Front. Mol. Biosci. 8:800819. doi: 10.3389/fmolb.2021.800819

INTRODUCTION

Mutagenesis is often used to generate variants of proteins with improved biophysical properties such as solubility and activity and to understand protein function. The advancement of high-throughput mutagenesis techniques has enabled the generation of a large number of variants of a protein in a short span of time, in a massively parallelizable manner (Zheng et al., 2004; Jain and Varadarajan, 2014; Wrenbeck et al., 2016). If an appropriate functional assay to score protein activity *in vivo* exist, it is possible to infer the relative activity of each variant in the library, through library screening coupled to next generation sequencing (Fowler et al., 2010; Adkar et al., 2012; Matreyek et al., 2018).

Abbreviations: YSD, yeast surface display; SSM, site saturation mutagenesis; FACS, fluorescence-activated cell sorting; DFE, distribution of fitness effects; RBD, receptor binding domain; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; ACE-2, angiotensin-converting enzyme 2.

However, there is a dearth of efficient, high-throughput methods to measure the solubility and stability of multiple protein variants in parallel, and to discriminate between buried and active-site residues solely using mutational data (Bhasin and Varadarajan, 2021).

Yeast surface display (YSD) is commonly used as a tool to identify protein variants with improved biophysical properties (Schweickhardt et al., 2003; Jones et al., 2006). YSD is preferable to bacterial expression for disulfide containing or glycosylated proteins. Agglutinin based Aga2p is the most widely used system to display proteins on the yeast cell surface (Shusta et al., 2008). Aga2p is a small protein (7.5 kDa), covalently linked via disulphide linkages to the yeast cell surface protein Aga1p (Boder and Wittrup, 1997). Previous studies have shown that the amount of protein displayed on the yeast cell surface is directly correlated to the amount of protein secreted by the cells, as well as the thermal stability of the protein (Shusta et al., 1999). However, in other studies where the secretion efficiency (Hagihara and Kim, 2002) or yeast cell surface expression of proteins was measured, no such correlation was observed (Park et al., 2006; Piatesi et al., 2006). Proteolysis of yeast surface displayed proteins has also been used to differentiate properly folded, stable variants from unstructured variants or molten globules, as a proxy for stabilization (Chevalier et al., 2017; Rocklin et al., 2017; Basanta et al., 2020). However, this has primarily been applied to relatively small proteins (Chevalier et al., 2017; Rocklin et al., 2017; Dou et al., 2018; Basanta et al., 2020).

A previous study which showed correlation between stability and expression levels was carried out on a limited number of mutants, that were studied individually. In addition, the WT protein itself had a very low T_m (Shusta et al., 1999). It has also been suggested that if the stability of a protein crosses a certain threshold, its expression does not increase linearly with increase in stability and it is therefore difficult to distinguish stable mutants from less stable ones, using only expression as the criterion (Traxlmayr and Shusta, 2017). With a very high level of yeast surface expression for unstable variants, the yeast quality control system may not be able to differentiate between properly folded, unfolded or molten globule like proteins. However, once displayed on the yeast cell surface such mutants may unfold or aggregate and hence will not bind to a tertiary structure specific ligand or cognate partner.

To verify the above hypothesis, we used *Escherichia. coli* (*E.coli*) CcdB as a model protein. CcdB is the toxin component of the CcdAB toxin-antitoxin (TA) module which binds both free DNA Gyrase and the DNA Gyrase-DNA complex, these are referred to as inhibition and poisoning respectively. Formation of the poisoned CcdB:DNA Gyrase: DNA ternary complex stalls replication and causes cell death (Bernard and Couturier, 1992). The other component of this TA module codes for an antitoxin CcdA, which neutralizes the toxicity of the CcdB toxin upon binding to CcdB. A mutation of Arginine to Cysteine in the DNA Gyrase subunit A (GyrA) at residue 462 can abolish the binding of Gyrase to CcdB (Bernard and Couturier, 1992). The CSH501 *E. coli* strain carries this mutation in the gene of the *gyrA* subunit which makes it insensitive to CcdB (Bajaj et al., 2008). In a previous study, a single-site saturation mutagenesis library of CcdB was generated and the mutants were scored based on their *in vivo* growth phenotype (MS_{seq} score) (Adkar et al., 2012). In *E. coli*, a good correlation was found between the MS_{seq} score of ~70 mutants with either ΔT_m of purified protein (r = 0.65) or *in vivo* solubility in *E. coli* (r = 0.69) (Tripathi et al., 2016). In contrast to plate based phenotypes, YSD provides greater flexibility and improved quantitation. We therefore wished to explore the correlation between the amount of surface expression or ligand binding seen with YSD, with thermal stability and *E. coli in vivo* solubility using this large set of characterized mutants, which had a range of *in vitro* thermal stability and *in vivo* solubility.

We initially examined 30 different variants of CcdB. Mutants were chosen so as to have varying solubility (when expressed in *E. coli*), *in vitro* thermal stability, accessibility and residue depth. Fewer mutants were chosen for exposed residues, where most mutants are tolerated. Residue V18 is one of the most highly buried residues in CcdB and several mutants which span a range of thermal stability and *in vivo* solubility were chosen at this position. The *in vivo* solubility of these mutants ranged from completely soluble to insoluble. We did not find a good correlation between total expressed protein amount on the yeast cell surface and either *in vivo* solubility in *E. coli,* or *in vitro* determined thermal stability. However, a better correlation was observed between the amount of active protein on the yeast cell surface (i.e., the amount of bound ligand) with *in vivo* solubility/thermal stability. In the yeast cell surface display system (Chao et al., 2006), activity was monitored by measuring the extent of binding of yeast cell surface displayed CcdB to a FLAG tagged fragment of GyrA14 as described previously (Sahoo et al., 2015).

Multiple rounds of sorting enrich mutants which have the highest expression and binding on the yeast cell surface. Sorting in such a way may lead to the identification of mutants with better biophysical properties, however, it does not give any information about the relative activity of all the mutants in a library. We coupled FACS and deep sequencing to reconstruct the MFI (MFI_{seq}) of each mutant in the Site Saturation Mutagenesis (SSM) library of CcdB, using single round FACS sorting methodology. We use this parameter MFI_{seq}, to rank all the mutants based on their activity to generate the mutational landscape or distribution of fitness effects (DFE). We found that the DFE generated using binding was more accurate than the DFE generated using expression. Overall, our MFI_{seq} scoring parameter could readily discriminate between stable and destabilized mutants of CcdB in a highly multiplexed manner.

It is well known that mutations that affect activity occur primarily at either surface exposed residues directly involved in binding or catalysis or at buried residues important for folding and stability. It has been difficult to distinguish between these two classes of residues, solely from mutational data (Bhasin and Varadarajan, 2021). We show here that by examining the effects of charged substitution on surface expression we can discriminate between the two classes of residues. To further validate the approach described above, we analyzed previously published saturation mutagenesis YSD expression and binding data for the receptor binding domain (RBD) of SARS-CoV-2 to its ligand ACE-2 (Starr et al., 2020). We could successfully predict both binding-site and buried residues solely from the mutational data in this system as well.

MATERIALS AND METHODS

Bacterial Strains, Yeast Strains and Plasmids

E.coli CSH501 strain carries a mutation in the gyrA gene which abolishes inhibition and poisoning by CcdB (Bajaj et al., 2008). The EBY100 strain of Saccharomyces cerevisiae has the aga1 gene under the Gal1 promoter for inducible expression and a TRP1 auxotrophic mutation. The strain lacks the aga2 gene, so only Aga2p fused protein expressed from the plasmid, will form a complex with the Aga1p for yeast cell surface display (Boder and Wittrup, 2000). The ccdB gene was cloned in the pBAD24 plasmid for controllable expression in *E. coli*. ccdB mutants were cloned in the pPNLS shuttle vector for yeast cell surface expression (Najar et al., 2017).

Cloning of WT and Mutant ccdB in *E.coli*

ccdB mutants in pBAD24 were generated using three fragment Gibson assembly. Briefly, ccdB was amplified in two fragments using two sets of oligos. For each fragment one of the oligos binds to the vector and the other binds to the gene. The primer of both fragments which bind to the gene were completely overlapping and contained the desired mutation. The fragments were gel extracted and Gibson assembled with NdeI and HindIII digested pBAD24 vector. The Gibson assembled product was electroporated in *E. coli* CSH501 strain and positive transformants were selected on LB agar media containing ampicillin (100 μg/ml). The sequence was confirmed by Sanger sequencing. Sequence confirmed WT or mutant ccdB in pBAD24 vector was used as a template for PCR to amplify the ccdB gene by Vent DNA polymerase. The PCR amplified product was co-transformed with SfiI digested pPNLS vector in the EBY100 strain of S*accharomyces cerevisiae* using LiAc/SS carrier DNA/PEG method for *in vivo* recombination (Gietz and Schiestl, 2007). Positive transformants were selected on SDCAA Tryptophan dropout media plates and the sequence was confirmed by Sanger sequencing.

Protein Purification

WT and mutant CcdB was purified as described previously (Chattopadhyay and Varadarajan, 2019). Briefly, an overnight culture was diluted 100-fold in LB media containing ampicillin (100 μg/ml) and induced with L-arabinose (0.2% w/v) at an OD_{600} of ~0.5. Following induction for 3 h, cells were harvested and lysed by sonication. The soluble fraction was separated using centrifugation and incubated with CcdA peptide (residues 45–72nd) coupled to Affigel-15 at 4°C. The unbound fraction was removed and the column was washed with bicarbonate buffer (50 mM $NaHCO_3$, 500 mM NaCl, pH 8.5). The bound protein was eluted with 200 mM glycine (pH 2.5) and collected in an equal volume of 400 mM HEPES buffer (pH 8) to neutralize the acidity of glycine.

GyrA14 was purified as described previously (Dao-Thi et al., 2004). Briefly, an overnight culture was diluted 100-fold in LB media containing ampicillin (100 μg/ml) and induced with IPTG (1 mM) at an OD_{600} of ~0.5. Following induction for 3 h, cells were harvested and resuspended in TES buffer (0.2 M Tris, pH 7.5, 0.5 mM EDTA, 0.5 M sucrose and 1 mM PMSF). Cells were lysed and the soluble fraction was separated using centrifugation. The soluble fraction was incubated with pre-equilibrated Ni-NTA beads for 2 h at 4°C. The unbound fraction was removed, and the column was washed with 100 column volumes of wash buffer (50 mM imidazole in 0.05 M Tris, pH 8, 0.5 M NaCl). The protein was eluted with 500 mM imidazole in 0.05 M Tris, pH 8, 0.5 M NaCl and dialysed against 1x PBS.

Estimation of Solubility of WT and Mutant CcdB in *E.coli*

E.coli CSH501 strain, transformed with pBAD24 plasmid containing WT or mutant ccdB, was grown in media containing ampicillin for 16 h at 37°C and 180 RPM. A secondary culture was grown by diluting overnight grown culture 100-fold. Upon reaching an OD_{600} of 0.4–0.5, CcdB variants were induced with Arabinose at a final concentration of 0.2% (w/v) for 3 h. The cells were harvested from 1.5 ml culture and lysed in 500 μl 1X PBS, using sonication. Supernatant and pellet fractions were separated by centrifugation at 13,000 RPM at 4°C. The pellet fraction was resuspended in 500 μl 1X PBS and equal volumes of pellet and supernatant fractions were loaded on Tricine-SDS-PAGE to measure the relative amounts of protein in each fraction.

Protein Thermal Stability Measurement Using Thermal Shift Assay

The thermal shift assay was conducted in an iCycle iQ5 Real Time Detection System (Bio-Rad, Hercules, CA). A solution of total volume 20 μl containing 10 μM of the purified CcdB protein and 2.5X Sypro orange dye in suitable buffer (200 mM HEPES, 100 mM glycine), pH 7.5 was added to a well of a 96-well iCycler iQ PCR plate. The plate was heated from 15°C to 90°C with a 0.5°C increment every 30 s. The normalized fluorescence data was plotted against temperature and T_m measured as described (Niesen et al., 2007; Tripathi et al., 2016).

Yeast Surface Expression of WT and Mutant CcdB Proteins in EBY100 Cells and Flow Cytometric Analysis

Saccharomyces cerevisiae EBY100 cells containing WT ccdB or mutant in pPNLS plasmids were grown in 3 ml SDCAA media (glucose 20 g/L, yeast nitrogen base 6.7 g/L, casamino acid 5 g/L, citrate 4.3 g/L, sodium citrate dihydrate 14.3 g/L) for 16 hours. Grown cells were diluted to an OD_{600} of 0.2 in 3 ml SDCAA media and grown till the OD_{600} reached two. Thirty million cells were harvested using centrifugation and resuspended in 3 ml

SGCAA induction media (galactose 20 g/L, yeast nitrogen base 6.7 g/L, casamino acid 5 g/L, citrate 4.3 g/L, sodium citrate dihydrate 14.3 g/L) for 16 hours at 30°C, 250 RPM (Chao et al., 2006). One million cells were used for flow cytometric analysis. The amount of total protein expressed on the yeast cell surface was estimated by incubating the induced cells in 20 µl FACS buffer (1X PBS and 0.5% BSA), containing chicken anti-HA antibodies from Bethyl labs (1:600 dilution) for 30 min at 4°C. This was followed by washing the cells twice with 100 µl FACS buffer at 4°C. Washed cells were incubated with 20 µL FACS buffer containing goat anti-chicken antibodies conjugated to Alexa Fluor 488 (1:300 dilution), for 20 min at 4°C. Fluorescence of yeast cells was measured by flow-cytometric analysis. The total amount of active protein on the yeast cell surface was estimated by incubating the induced cells in 20 µl FACS buffer containing 100 nM GyrA14 for 45 min at 4°C. Cells were washed and incubated with 20 µl mouse anti-FLAG antibodies (1:300). This was followed by washing the cells twice with FACS buffer, followed by incubating with 20 µl rabbit anti-mouse antibodies conjugated to Alexa Fluor 633 (1:1,600 dilution). The flow-cytometric analysis was carried out on BD Accuri or BD Aria III instruments.

Yeast Surface Expression and Sorting of CcdB Single-Site Saturation Mutagenesis Library

Previously, an SSM library of ccdB was generated in the pBAD24 vector (Adkar et al., 2012; Tripathi et al., 2016). The library was PCR amplified using primers having homology to the pPNLS vector. The PCR amplified library was gel extracted and cloned in pPNLS vector using yeast *in vivo* recombination.

A similar protocol was used for sample preparation of the library for FACS as described above for the single mutants with slight modifications. Briefly, ten million cells were taken for FACS sample preparation and the reagents were used in 10X higher volumes compared to the earlier flowcytometric analysis. Two different concentrations of GyrA14 (100 nM, 5 nM) were used for sorting CcdB mutants based on the binding in the 1D histogram. The cells were sorted in 11 and 10 different populations (bins) in case of binding with GyrA14 at concentrations of 100 and 5 nM respectively. Additionally, 11 different populations (bins) were sorted from the expression histogram. The experiment was repeated in a biological replicate. The sorting of CcdB libraries was performed using a BD Aria III cell sorter.

Sample Preparation for Deep Sequencing

Sorted populations were grown on SDCAA agar plates for 48 h. Colonies were scraped and plasmids were extracted from the cells. The ccdB gene was PCR amplified using primers which bind upstream and downstream of the ccdB sequence and had multiplex identifier (MID) sequence to segregate the reads from different sorted bins. The DNA was amplified for 15 cycles using PCR and the amplified product was gel extracted and purified. Equal amounts of DNA from each sorted population were pooled, and the library was generated using the TruSeq™ DNA PCR-Free kit from Illumina. The sequencing was done on an Illumina HiSeq 2,500 250 PE platform at Macrogen, South Korea after incorporating 20% ϕX174 DNA in the library.

Analysis of Deep Sequencing Data

Deep sequencing data for the ccdB mutants obtained from the Hiseq 2,500 platform was processed using a pipeline developed by adopting certain aspects from an already existing in-house protocol (https://github.com/skshrutikhare/cys_library_analysis). The latter method involved the alignment with wild type sequence followed by merging of the paired-end reads, while in the modified protocol, the reads are first merged and then aligned with the wild-type sequence. The present methodology consists of the following steps: assembling the paired end reads, quality filtering, binning, alignment and mutant identification. All these steps were incorporated in a pipeline and made executable from a single command using a parameter file unique to a given data-set. In the first step, paired end reads were assembled using the PEAR v0.9.6 (Paired-End Read Merger) tool (Zhang et al., 2014). The "quality filtering" step involved deletion of terminal "NNN" residues in the reads, and removal of reads, not containing the relevant MID and/or primers, along with the reads having mismatched MID's. Finally, only those reads having bases with Phred score ≥20 are retained. A binning step involved further filtering, which eliminated all those reads having incorrectly placed primers, truncated MIDs/primers (due to quality filtering) and shorter/longer sequences than the length of the wild type sequences. The remaining reads were binned according to the respective MIDs. In the alignment step, reads were aligned with the wild type ccdB sequence using the Water v6.4.0.0 program (Smith and Waterman, 1981) and reformatted. The default values of all parameters, except the gap opening penalty, which was changed to 20, was used. In the final step of "substitution", reads were classified based on insertions, deletions and substitutions (single, double etc mutants).

Mean Fluorescence Intensity Reconstruction From Deep Sequencing Data

Reads of each mutant were normalized across different bins individually (**Equation 1**), and the fraction of each mutant (*Xi*) distributed amongst the different bins was calculated (**Equation 2**). The reconstructed MFI for an individual mutant was calculated by the summation of the product, obtained upon multiplying the fraction (*Xi*) of the mutant in a particular bin 1) with the MFI of the corresponding bin obtained from the FACS experiment (*Fi*), across the various bins populated by the respective mutant (**Equation 3**).

$$\text{Normalized read of mutant in bin } i\ (Ni) = \frac{\textbf{No. of reads of mutant } \boldsymbol{i} \textbf{ in bin } \boldsymbol{i}}{\sum \textbf{reads in bin } i} \qquad \text{Equation 1}$$

$$\text{Fraction of mutant in each gate}\,(Xi) = \frac{\boldsymbol{Ni}}{\sum_1^n \boldsymbol{Ni}} \qquad \text{Equation 2}$$

$$\text{Reconstructed MFI} = \sum_{1}^{n} Fi*Xi \qquad \text{Equation 3}$$

The MFI_{seq} of the biological replicates were different so the MFI_{seq} of one of the replicates was adjusted using "m" and "c" obtained from the correlation between the replicates and then averaged.

$$\text{Average MFI}_{\textbf{seq}} = \frac{\textbf{MFIseq (replicate 1)} + (\textbf{m} * \textbf{MFIseq (replicate 2)} + C)}{2}$$

Maximum Likelihood Mean Fluorescence Intensity) Calculation

Reads of each mutant were normalized within and across the bins. The fraction of each mutant (*Xi*), distributed amongst the different bins, was calculated as explained in the above section. The fraction (*Xi*) was multiplied with a scaling factor to convert the data into integers as this is required by the package below. The mlMFI was calculated using a maximum likelihood method using the fitdistrplus R package as explained earlier (Starr et al., 2020). The "fitdistcens" function in the fitdistplus R package helps in the estimation of fluorescence values for such observations using a maximum likelihood approach, where the values are transformed into a data frame of two columns left and right, describing each observed value as an interval and assuming a normal distribution of values. The left column contains the left bound of the interval and the right column contains the right bound of the interval for interval-censored observations, based on the fluorescence boundaries of each bin. The maximum likelihood approach was used to estimate the MFI of binding and expression for each mutant, based on its distribution of reads across the sorted bins, and the fluorescence boundaries of each sorted bin.

Mean Fluorescence Intensity Calculations After Bins Merging

The bins were merged following which mlMFI amd MFI_{seq} were calculated for GyrA14 binding (100 nM) for replicate 1. The fraction of each mutant in each bin was calculated as explained in the earlier sections. To merge bins for a given mutant, fractions present in each of the bins to be merged were added arithmetically. For mlMFA calculation, the minimum and maximum fluorescent boundary of the merged bin was set at the lowest and highest value of the fluorescent boundary for that set of bins. The mlMFI of CcdB mutants was calculated as explained above. In the case of MFI_{seq}, the mean fluorescent intensity of merged bins was determined by making a new bin spanning the set of merged bins. The MFI_{seq} of CcdB mutants was then calculated as explained above.

Depth, Accessibility and RankScore Calculations

Depth was calculated using the server DEPTH (Chakravarty and Varadarajan, 1999; Tan et al., 2011). Accessibility was calculated using the program NACCESS (Hubbard SJ, 1993). In both cases, the input co-ordinates were homodimeric CcdB (PDB ID 3VUB). RankScore and MS_{seq} are measures of mutational sensitivity in *E. coli*. Values were obtained from Adkar et al. (Adkar et al., 2012). Buried residues were those with <10% accessibility in 3VUB. Active-site residues were those with ΔASA>0. ΔASA difference between the solvent accessible surface area of CcdB residues in the free (3VUB) and GyrA14-bound forms (1X75) respectively (Aghera et al., 2020).

Deep Mutational Scanning of SARS COV-2 Receptor Binding Domain

The deep mutational scanning data was taken from a recent report (Starr et al., 2020) in which two independent libraries of RBD were generated and sorted in four different bins based on expression or binding to ACE-2. In the MFI of binding and expression for individual mutants was reconstructed in that study using a maximum likelihood method using fitdistrplus R package. The expression MFI [Sortseq (expr)] data was shared by the authors in a repository (https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS). We reconstructed the binding MFI [Sortseq (bind)] at an ACE-2 concentration of 100 pM (TiteSeq_09). For Sortseq (bind) estimation we used the script provided by the authors (https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS/blob/master/results/summary/compute_expression_meanF.md). The authors used data from both single and multiple mutants, together with a model to account for epistatic effects to infer the MFI values for individual mutants. We modified the script to change the input data required to calculate Sortseq (bind). For both Sortseq (bind) and Sortseq (expr), we analyzed only single mutant data to avoid any artifacts that might arise from the epistatic model and took the average of delta Sortseq MFI {log (Sortseq (WT))—log [Sortseq (mutant)]} of mutants which had multiple barcodes. The Sortseq MFI values of mutants were averaged between the two libraries and the antilog was calculated for delta Sortseq MFI to analyse the ratio of Sortseq (bind) or Sortseq (expr) of mutants with respect to WT.

Statistical Analysis

The correlations and *p* values for its significance were calculated using the GraphPad Prism software 9.0.0 (* indicates $p < 0.05$, ** indicates $p < 0.01$, **** indicates $p < 0.0001$). The weighted correlations were calculated using the weights function of R. For the computation of weighted correlation, a weight of 1/(σ/μ) was used on the mean values of replicates.

RESULTS

Yeast Surface Display of CcdB Mutants

Yeast surface display (YSD) has become an increasingly popular tool for protein engineering and library screening applications (Pepper et al., 2008). Aga2p mating adhesion receptor of *Saccharomyces cerevisiae* is used as a fusion protein for yeast surface display. For surface expression, we used a vector in which CcdB is fused at the C-terminus of Aga2 (Sahoo et al., 2015). We generated (**Supplementary Figure S1**) and individually

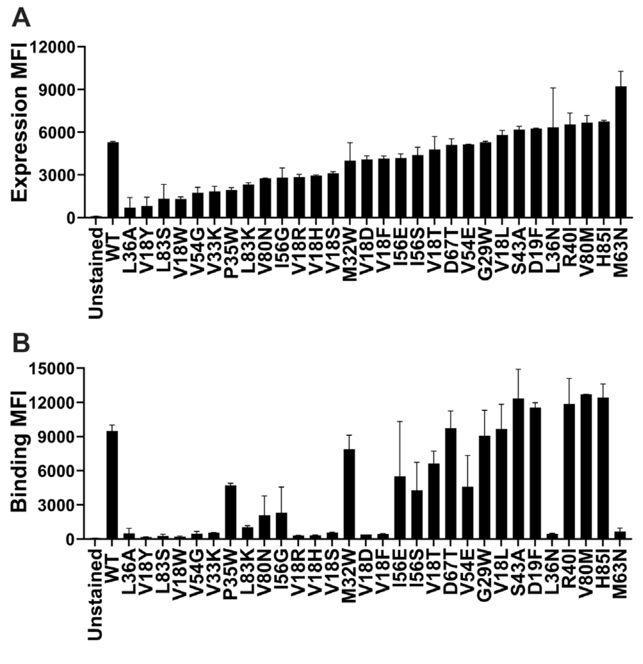


FIGURE 1 | Comparison of the level of expression and binding of CcdB mutants on the yeast cell surface. **(A)** The expression and **(B)** binding to GyrA14 of individual mutants. Errors are calculated from two biological replicates. Most mutants expressed at high levels, however, the amount of active protein varied widely. A few mutants which showed a high level of expression did not show any binding to GyrA14. In both panels, mutants are arranged in order of increasing expression level.

characterized 30 CcdB variants on the yeast cell surface. Most CcdB mutants had similar levels of expression to the WT protein (**Figure 1A**). However, the mutants showed different amounts of active protein as assayed by binding to the FLAG tagged GyrA14 compared to the WT protein (**Figure 1B**). Previously, we have characterized the *in vitro* thermal stability and *in vivo* solubility of several CcdB mutants (Tripathi et al., 2016). The amounts of total and active protein were estimated using antibodies against the HA-tag at the N-terminal of the yeast surface displayed CcdB and the C-terminal FLAG tag of GyrA14 respectively. The correlation coefficient (r) between amount of total protein on the yeast cell surface with *in vivo* solubility or T_m of the corresponding purified protein were 0.31 and 0.70 respectively (**Figures 2A,B**). It is unclear why mutants which have very low solubility in *E. coli* are highly expressed on the yeast cell surface. It was previously hypothesized that the protein folding quality control system in yeast is not as effective as in mammalian systems, therefore partially folded/molten globule/aggregated protein may exist on the surface of yeast (Park et al., 2006). A correlation of r = 0.80 was found between the amount of active protein on the yeast cell surface with its *in vivo* solubility determined in *E. coli* (**Figure 2C**). We also found a better correlation (r = 0.90) between amount of active CcdB protein on the yeast cell surface and its *in vitro* thermal stability (**Figure 2D**), compared to that between total CcdB protein on the yeast cell surface and thermal stability.

Deep Sequencing Analysis of CcdB Library and Mean Fluorescence Intensity Calculation for CcdB Mutants

To extend these results, an SSM library of ccdB was expressed on the yeast cell surface. Different populations based on extent of binding to gyrase or cell surface expression were sorted. A total of 32 different populations were sorted at two different concentrations of GyrA14 (100 nM, 5 nM) as a function of either surface expression level or the extent of binding to GyrA14 (**Supplementary Figure S2**). The lower concentration of GyrA14 was chosen to be around the K_D of CcdB-GyrA binding (**Supplementary Figure S3**), the higher concentration was one where WT CcdB approaches saturation in binding with GyrA14 on the yeast cell surface. We hypothesized that at lower concentrations of GyrA14, the binding on the yeast cell surface will be a function of both stability as well as binding affinity. However, at saturating concentration of GyrA14, the binding on the yeast cell surface will largely be a function of amount of correctly folded protein that in turn might be a function of protein stability, rather than the K_d of the mutant(s). MFI was calculated for each mutant as explained in the Methods section. The MFI was calculated at different stringencies (where the stringency refers to the sum of reads for a given mutant over each gate of the histogram), namely 25, 50, 100, 150, and 200 reads. All mutants with a total read number less than the stringency value were removed from the analysis. As the stringency increased, the pairwise correlation between the biological replicates increased (**Supplementary Figure S4**, **Supplementary Table S1**). The data was analysed with a stringency of 50 reads, since at higher stringencies, correlation did not improve significantly, but the number of mutants reduced. Reconstructed Binding and Expression MFI from deep sequencing data are hereafter referred to as MFI_{seq} (bind) and MFI_{seq} (expr) respectively.

Mean Fluorescence Intensity Reconstruction and its Correlation With Stability, Solubility and Residue Burial

A few published studies have described estimation of MFI values using deep sequencing of sorted populations and are therefore similar to our experimental strategy. However, the procedure for MFI reconstruction in these reports was relatively complicated compared to that used here (Sharon et al., 2012; Noderer et al., 2014; Peterman and Levine, 2016; Cambray et al., 2018). In those studies, the fractions of reads were calculated in each bin for all the mutants and MFI (mlMFI) of mutants were calculated by fitting the data to a maximum likelihood distribution of the histogram. We found that if mutants are present in only one bin (highly destabilized or nonsense mutants) then this method is unable to perform the MFI calculation (Starr et al., 2020). For the remaining mutants we found a good correlation between MFI_{seq} and mlMFI for binding at 5 and 100 nM GyrA14, and for expression (**Supplementary Figure S5**). For mutants with over 50 reads, we could calculate the MFI of 11,153 mutants using the

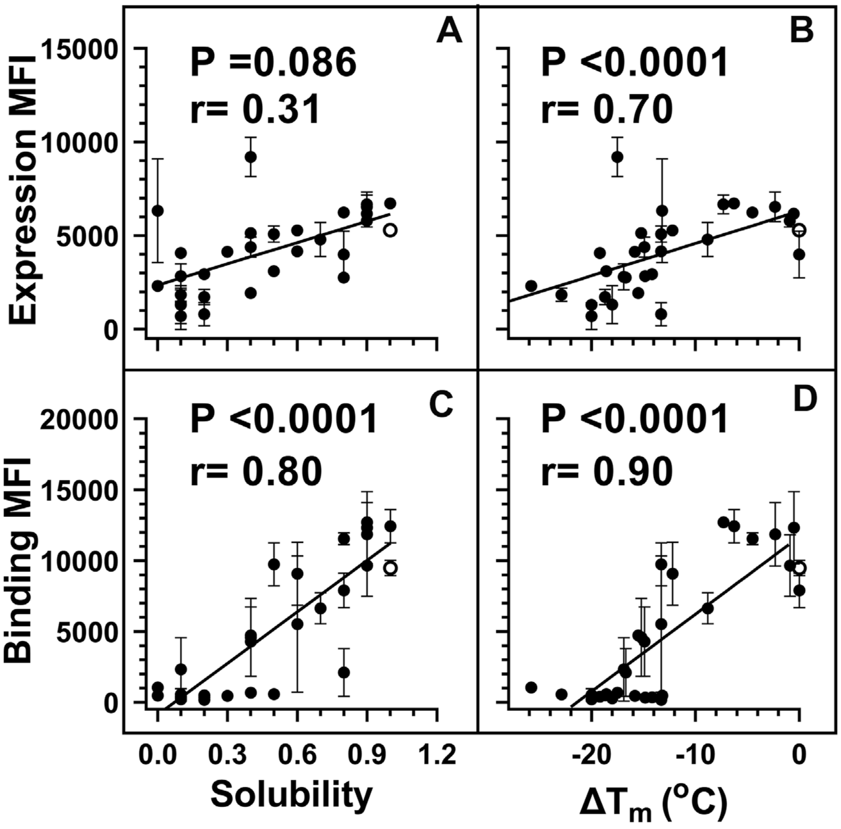


FIGURE 2 | Weighted correlations of E. coli *in vivo* solubility and *in vitro* thermal stability with the amount of total and active protein respectively, on the yeast cell surface. For individual mutants, MFI's of expression and binding were estimated by probing the HA tag on surface expressed protein and the FLAG tag on cell surface bound GyrA14 respectively. For weighted correlation calculation, a weight of 1/(σ/μ) was used. Here σ and μ represent the standard deviation and mean values for each point respectively. Weighted correlation of the total amount of protein (Expression MFI) displayed on the yeast cell surface with **(A)** *in vivo* solubility and **(B)** ΔT_m [T_m (mutant)-T_m (WT)] of CcdB mutants. Weighted correlation of the amount of active protein (Binding MFI) on the yeast cell surface with **(C)** *E. coli in vivo* solubility and **(D)** ΔT_m of CcdB mutants. A better correlation was observed between biophysical parameters with binding MFI rather than expression MFI. In the figure, the ΔT_m of WT was increased by 1°C to remove overlap with another point. Data for *E. coli in vivo* solubility and thermal stability was taken from Tripathi et al. (Tripathi et al., 2016). WT data is shown in open circles. *p* values indicate the significance for non-zero slope values in all the correlations.

maximum likelihood method and 11,436 mutants using our method. We also found that progressively reducing the number of bins from eleven to six, does not significantly affect the estimated MFI values, however a further reduction to four bins results in a noticeable change in the estimated values using either method (**Supplementary Figure S6**). A good correlation was also found between the MFI of individually analysed mutants and their corresponding MFI_{seq} values, validating our approach of MFI reconstruction (**Supplementary Figure S7A, 7B**). Individually analysed mutants showed a good correlation between the amount of active protein on the cell surface and *in vitro* measured thermal stability of the purified protein. Similarly, we also found a good correlation between MFI_{seq} (bind) of mutants inferred from deep sequencing, and thermal stability as well as *in vivo* solubility for the selected mutants (**Supplementary Figure S7C, 7D**).

For the exposed residues (>10% accessibility) (**Supplementary Figure S8**), mutations did not affect the degree of surface expression and binding to GyrA14 (**Figures 3A,B**). Expression was also unaffected by mutations in the active-site residues (identified from PDB ID:1X75) (**Figure 3C**, **Supplementary Figure S8**). However, many buried site mutants showed very low expression, possibly because of aggregation and degradation inside cells or during export (**Figure 3C**). In the case of binding for buried and active-site residues, a very high mutational sensitivity was found (**Figure 3D**) similar to the previous report of CcdB mutants in *E. coli* (Tripathi et al., 2016). We also found a very high mutational sensitivity of binding for a few non-interacting residues in the loop connecting beta strands S2 and S3 at both 5 and 100 nM GyrA14 concentration (**Supplementary Figure S9**). The residues I24, I25 and D26 in this loop are directly involved in interacting with Gyrase and mutation at non-interacting residues (22, 23 and 27) in the loop might restrict or alter the conformation of the loop, thus reducing the affinity of CcdB mutants to GyrA14. However, there was no effect on the expression of the mutants in this loop, indicating that the mutant proteins are not destabilized (**Supplementary Figure S9**). We did not find a high correlation between MFI_{seq} (bind) and either accessibility or depth, because many mutations at both buried and active-site residues have high mutational

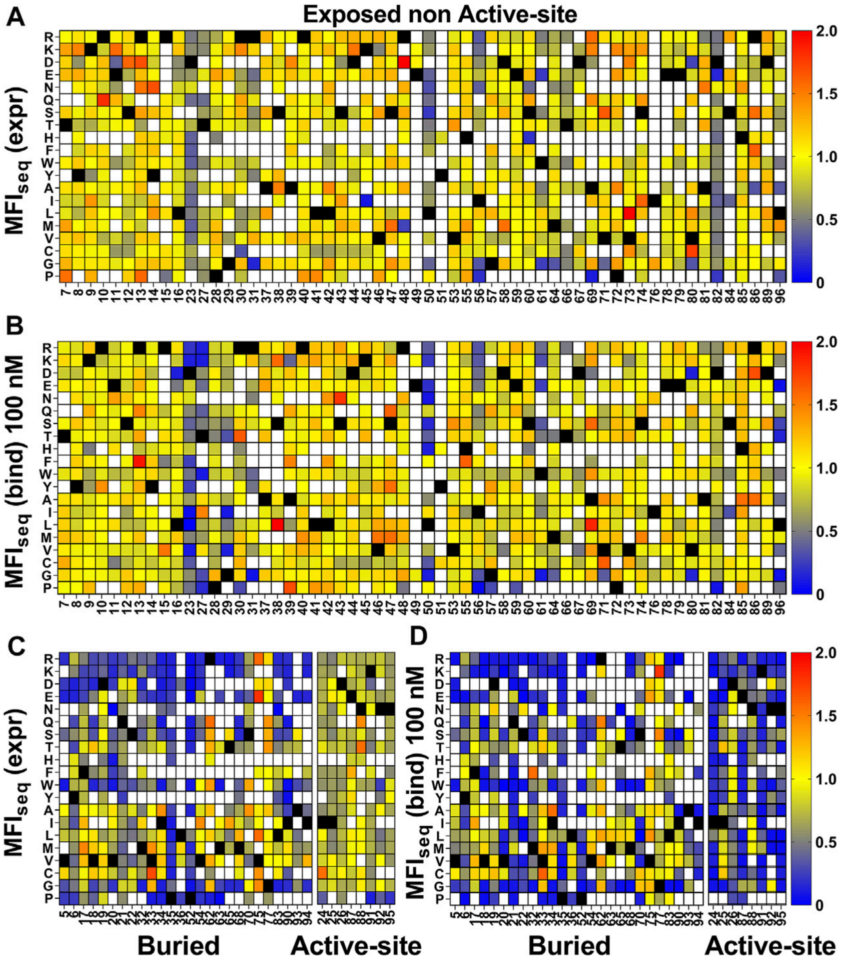


FIGURE 3 | Heatmap of normalized MFI_{seq} values for CcdB mutants. MFI_{seq} value of mutant was divided by the MFI_{seq} value of WT to normalize it. **(A)** MFI_{seq} (expr) and **(B)** MFI_{seq} (bind) at 100 nM GyrA14 for exposed non active-site residues. **(C)** MFI_{seq} (expr) and **(D)** MFI_{seq} (bind) for buried and active-site residues. Exposed, buried (PDB ID:3VUB) and active-site (PDB ID:1X75) residues are segregated based on the crystal structure. Residues which had accessibility greater than 10% were considered exposed, all remaining residues were considered buried, and active-site mutants in contact with GyrA14 were identified as explained the Methods section. Blue to red colour represents increasing normalized MFI_{seq} values, black colour shows the WT residue at the corresponding position. White colour indicates that the mutant is not available. The buried site residues have very high mutational sensitivity both in case of expression and binding. The active-site residues show mutational sensitivity only with respect to Gyrase binding. Information about the mutational sensitivity of expression and binding can be used to differentiate exposed, buried and active-site residues.

sensitivity (**Supplementary Table S2**). The previously described parameter RankScore, is a measure of mutant activity in *E. coli* (Adkar et al., 2012) with high RankScore denoting lower activity. We found a poor correlation between the MFI_{seq} (bind) values of CcdB mutants at both exposed non active-site as well as active-site residues, and RankScore. In *E. coli,* most of the exposed non active-site residues do not show any mutational sensitivity, i.e., they have the same RankScore values as WT. However, in the present case many such CcdB mutants show lower binding to GyrA14 compared to WT. The loss of binding could be attributed to the decrease in the affinity between CcdB and Gyrase, or destabilization due to mutation. We defined a new parameter MrMFI (mean residue MFI) which is the mean of the MFI values of all the mutants at a certain position. MrMFI (expr) and MrMFI (bind) at 100 nM GyrA14, show a good correlation with RankScore (**Supplementary Table S2**). MrMFI (expr) also showed good correlation with Depth which is a structural measure of residue burial (Chakravarty and Varadarajan, 1999). However, in the case of binding at 5 nM, a weaker correlation of MrMFI (bind) with the aforementioned parameters was observed (**Supplementary Table S2**). In previous studies, identification of the active-site residues solely from the deep sequencing data was not very efficient (Adkar et al., 2012; Bhasin and Varadarajan, 2021), this is presumably because *in vivo* activity is often governed by threshold effects, and because mutations at buried residues also affect activity. The current methodology removes such drawbacks. We could distinguish between buried and active-site residues by comparing the MFI_{seq} (bind) and MFI_{seq} (expr). Most buried site residues showed low values of both MFI_{seq} (bind) and MFI_{seq} (expr) compared to WT. However, the active-site residues showed low MFI_{seq} (bind) but similar MFI_{seq} (expr) compared to WT. We found that the average MFI_{seq} values of charged residues are a good predictor to discriminate between buried and active-site residues. For calculating $MrMFI_{charged}$ of charged WT residues, we only consider mutants with opposite charge. For some mutants at buried positions, we found a very low $MrMFI_{charged}$ (expr) but the mutants were absent in $MrMFI_{charged}$ (bind). We found that such mutants had very high reads, suggesting that the values of $MrMFI_{charged}$ (expr) are correct. We anticipated that such mutants lack binding and are therefore present only in the bin which had a background level of binding signal, the presence of mutant in only that gate led to the removal of such mutants due to the stringency set for the analysis. Hence, such mutants were assigned a $MrMFI_{charged}$ (bind) similar to other buried positions. $MrMFI_{charged}$ had a bimodal distribution (**Supplementary Figure S10**), so k-means clustering was performed to identify the mean (μ) and standard deviation (σ) of each distribution. The distributions were named D1 (higher $MrMFI_{charged}$) and D2 (lower $MrMFI_{charged}$). Buried site residues were assigned to be those which have $MrMFI_{charged}$ (bind) and MFI_{seq} (expr) less than the set threshold (μ+0.5*σ) for distribution D2. Active-site residues were assigned as those which had $MrMFI_{charged}$ (bind) less than (μ+σ) of the D2 distribution and MFI_{seq} (expr) higher than (μ−2*σ) of distribution D1 (**Figure 4**). The accuracy, specificity and sensitivity of prediction of exposed non active-site, buried and exposed active-site residues are mentioned in **Supplementary Table S3**. We also compared our prediction results derived from saturation mutagenesis phenotypes with those of an *in silico* predictor, PROF (Rost and Sander, 1994). For a residue to be classified as buried by PROF, the relative solvent accessibility cut-off used is < 12. We observed a slightly lower specificity and accuracy for CcdB, and lower sensitivity in the case of RBD when predictions were made using PROF (**Supplementary Table S4**), relative to our predictions. We also examined the performance of PROF with other proteins and found that the specificity of the predictions was higher than 0.8 in all the cases except for CcdB. However, the sensitivity of the predictions was lower than 0.8 in all the cases except for CcdB, Gal4 and Ubiquitin. The accuracy for the PROF prediction was 0.77 and 0.78 for CcdB and RBD respectively, comparable but slightly lower than the corresponding values of 0.92 and 0.8 for CcdB and RBD respectively, from the saturation mutagenesis predictions in this work.

Selection and Characterization of Putative Stabilized Mutants From Deep Sequencing Data

In the previous section, we discussed the correlation between protein biophysical properties such as thermal stability and *in vivo* solubility with either the amount of active protein or the ratio of active protein to total protein on the yeast cell surface for a few (30) mutants. However, most of these mutants were destabilized with respect to the WT protein. To confirm whether this correlation also holds for mutants that have stability similar or greater than WT, we selected a few CcdB mutants based on either the MFI_{seq} (bind) or MFI_{seq} (ratio) [MFI_{seq} (bind)/MFI_{seq} (expr)] for *in vitro* characterization of thermal stability. We examined the average and standard deviation of expression for all mutants and selected only those mutants based on MFI_{seq} (ratio) which cross a minimum cut-off (μ+0.5*σ) for MFI_{seq} (expr) to remove the bias created by mutants which have very low expression. No threshold for expression was set for selection of mutants based on their MFI_{seq} (bind). No selection of the mutants was performed based solely on the MFI_{seq} (expr).

Six mutants were characterized using the criteria MFI_{seq} (bind) at 5 nM GyrA14, none of them showed a higher T_m than WT (**Figure 5A**); whereas two of the mutants selected on the basis of MFI_{seq} (ratio) showed a significantly higher T_m than WT (**Figure 5B**). A subset of seven mutants was selected based on MFI_{seq} (bind) at 100 nM GyrA14, none of the mutants showed higher stability than WT CcdB (**Figure 5C**). Ten mutants were selected based on MFI_{seq} (ratio) and characterized, four showed higher stability, two mutants were similar to WT and the remaining four were less stable than WT CcdB (**Figure 5D**). We therefore hypothesize that if the stability of a mutant crosses a threshold then its expression will not increase further. To confirm this hypothesis, we measured the amount of active protein on the yeast cell surface for seven individual mutants which had T_m's ranging from 60°C to 70°C, and found that the expression and binding for these mutants are similar to each other and to WT (**Supplementary Figure S11**).

Prediction of Thermal Stabilities of Putative Destabilized Mutants

For destabilized mutants we observed a good correlation between MFI_{seq} (bind) and T_m of individual mutants (**Supplementary Figure S7D**). Using this correlation, we next predicted the T_m of each mutant for an additional set of (n = 28) previously described CcdB mutants (Tripathi et al., 2016) based on their MFI_{seq} (bind). We found a good correlation (r = 0.82) between predicted and *in vitro* measured T_m for this set of CcdB mutants as well (**Supplementary Figure S12A**). This now allows us to identify putative destabilized mutants and accurately predict the extent of destabilization for all such mutants in the CcdB YSD library. We also predicted the thermal stability of CcdB mutants using the *in silico* predictor, HoTMuSiCv1.0 (Pucci et al., 2020), however, we did not find a good correlation between measured and predicted T_m (**Supplementary Figure S12B**). It has been shown that *in vitro* protein thermal stability and free energy of unfolding are correlated (Chen et al., 2000; Prajapati et al., 2007; Tripathi et al., 2016). We therefore predicted the free energy of unfolding for CcdB mutants using SDM (Pandurangan et al., 2017), mCSM (Pires et al., 2014b), PoPMuSiC (Dehouck et al., 2011), DynaMut (Rodrigues et al., 2018), DUET (Pires et al., 2014a), MAESTROweb (Laimer et al., 2016), DeepDDG (Cao et al., 2019), CUPSAT (Parthiban et al., 2006), PremPS (Chen et al., 2020) and INPS-MD (Savojardo et al., 2016). We found moderate correlations, with DeepDDG performing the best (r = 0.59), but still poorer compared to our prediction from YSD data (r = 0.82). For a more detailed comparison we analysed the predictions of stability by DeepDDG, since this showed the highest correlation

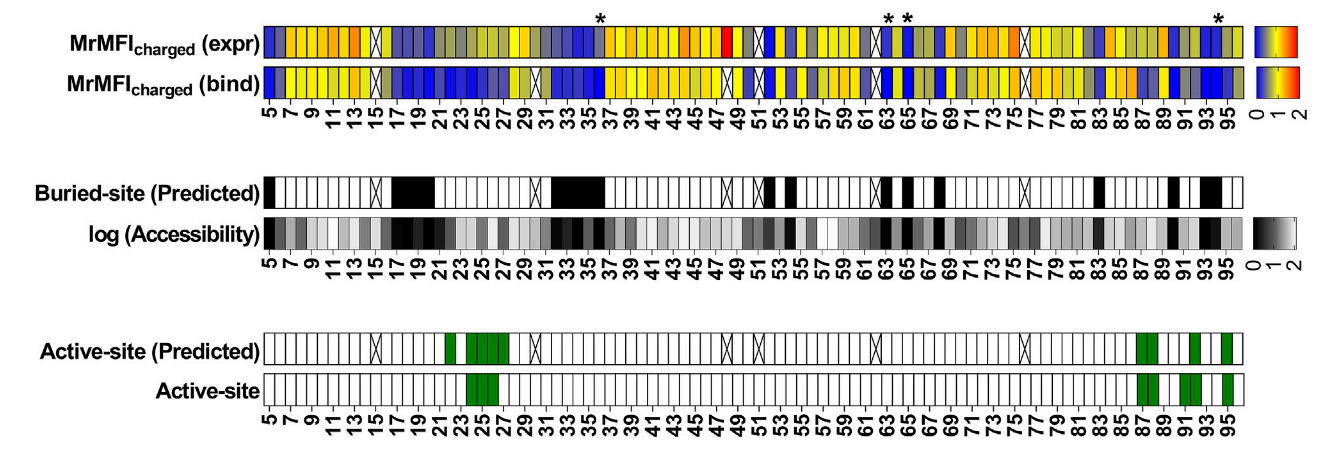


FIGURE 4 | Identification of buried and active-site residues from $MrMFI_{charged}$ (bind) and $MrMFI_{charged}$ (expr). Side chain accessibilities in dimeric CcdB (PDB: 3VUB), darker to lighter shade indicate increasing accessibility, accessibility is reported as log accessibility. The mutants were clustered into two bins based on the distribution of $MrMFI_{charged}$ and k-means and standard deviations were calculated for both distributions. The distributions were named D1 (higher $MrMFI_{charged}$) and D2 (lower $MrMFI_{charged}$). Residues which had $MrMFI_{charged}$ (binding) and $MrMFI_{charged}$ (expr) lower than (μ+0.5*σ) of distribution D2 were characterized as buried. The false negatives were Y6, D19, Q21, S22, S70, V75 and G77, the polar side chains of these residues are pointing towards the surface. Active-site residues were identified as those in contact with GyyrA14 (PDB ID 1X75). Residues which had $MrMFI_{charged}$ (binding) less than (μ+σ) of D2 distribution and $MrMFI_{charged}$ (expr) higher than (μ-2*σ) of distribution D1 were predicted as active-site. We obtained a few putative false positives. However, these residues are likely involved in functional aspects of activity that cannot be inferred from the CcdB:GyrA14 crystal structure. The same residues were seen to be important for CcdB activity *in vivo* in *E. coli* (Tripathi et al., 2016). Some positions could not be categorized due to lack of reads, such positions are indicated with an 'X'. Positions indicated with '*' are the ones where $MrMFI_{charged}$ (expr) was observed and the mutants had high read counts but the mutants were absent in $MrMFI_{charged}$ (bind), such positions were assigned $MrMFI_{charged}$ (bind) values similar to other buried positions.

with measured stability of individual mutants at non active-site residues. We excluded residues 21, 22, 23 and 27 as these positions behaved like active-site residues. We found that trends for ΔΔG predicted by DeepDDG for exposed non active-site residues are similar to those obtained from MFI_{seq} (bind) (**Figures 6A,B**). However, we observed some mutant specific differences at residues 8, 16, 50, 53 and 96. Mutations at residues 50 and 96 have highly deleterious effects which reduced GyrA14 binding to yeast surface displayed protein, these are only partially predicted by DeepDDG. In the case of charged and polar mutations at residue 8, 16 and 53 we did not observe a reduction in binding, but the software predicted them to be destabilizing. In the case of buried positions, we found mutation specific effects at 35, 52 and 94 where DeepDDG predicted changes were significantly smaller than the experimentally observed ones. We also found that most of the phenylalanine, tryptophan and arginine mutations were highly destabilizing and the mutants did not bind to GyrA14, however the software gave a lower stability penalty for these substitutions (**Figures 6C,D**). Our MFI based measurements suggested greater destabilization for several mutants relative to DeepDDG prediction. While the overall trends were similar, as discussed above, there are several differences between MFI based and DeepDDG based stability predictions.

Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain

To examine the generality of our approach, we also analyzed recently reported deep mutational scanning data of the SARS-CoV-2 receptor binding domain (Starr et al., 2020). In this study two separate libraries were generated and individually sorted based on expression and binding to ACE-2. The binding [Sortseq (bind)] or expression [Sortseq (expr]) MFIs relative to WT for barcoded mutants were calculated from the deposited NGS data as explained in the Methods section. Additionally, we analyzed binding at only one concentration of ACE-2 (100 pM, TiteSeq_09) at which the binding started to saturate. Buried residues were those with <10% side chain accessibility in chain C of PDB ID 7KMH (Jones B. E. et al., 2020). ACE-2 binding (active-site) residues were assigned as those contacting ACE-2 (Malladi et al., 2021). To identify the active-site and buried residues from Sortseq data, we calculated the $MrMFI_{charged}$ for each position. Similar to CcdB, we observed a bimodal distribution for both $MrMFI_{charged}$ (bind) and $MrMFI_{charged}$ (expr) (**Supplementary Figure S13**) and k-means and standard deviation were calculated for both the distribution D1 (higher $MrMFI_{charged}$) and D2 (lower $MrMFI_{charged}$). As described above for CcdB, buried residues were identified as those which had $MrMFI_{charged}$ (bind) and $MrMFI_{charged}$ (expr) less than the set threshold (μ+0.5*σ) for distribution D2. The active-site positions were identified as those which had $MrMFI_{charged}$ (bind) lower than the set threshold (μ+σ) for population D2 and $MrMFI_{charged}$ (expr) values higher then (μ-2*σ) for population D1. We accurately identified most of the buried residues, however there were some false positive and false negative predictions relative to the crystal structure information (**Figure 7**). We found 21 positions to be false negative buried positions. We categorized these false negatives into two categories, namely, glycine and the side chains which are pointing towards the surface. The accessibility calculated by

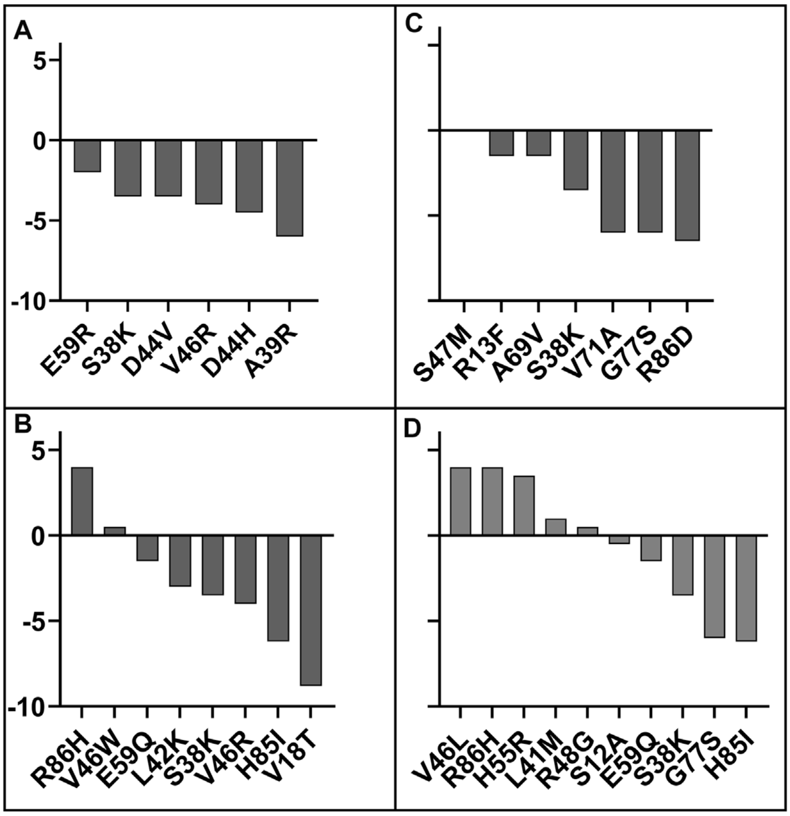


FIGURE 5 | ΔT_m of putative stabilized CcdB mutants. Mutants were identified from **(A)** MFI$_{seq}$ (bind) at 5 nM GyrA14, **(B)** MFI$_{seq}$ (ratio) at 5 nM GyrA14, **(C)** MFI$_{seq}$ (bind) at 100 nM GyrA14, **(D)** MFI$_{seq}$ (ratio) at 100 nM GyrA14. The mutants were randomly selected from a subset of forty mutants which showed the highest MFI$_{seq}$ (bind) or the highest MFI$_{seq}$ (ratio) and had MFI$_{seq}$ (expr) > 6,672.

DEPTH server for glycine was zero and we therefore expected glycine to fall into the false negative buried category. Thirteen positions out of twenty-one false negative were glycine. Another six positions, 336, 348, 361, 443 and 480 had their side chains pointing towards the protein surface. We also found similar false negative buried residues in CcdB where the side chain hydrophilic group was pointing towards the protein surface. Position 363 and 365 in RBD had accessibility <10% and were pointing towards the core of the protein in the PDB (7KMH) used to calculate accessibility. However, we found that these positions have high accessibility (>30%) in another structure (PDB ID 7D2Z). All the available RBD structures are in complex with other molecules, this might be responsible for variation in the accessibility of residues in different RBD structures. We found 17 false positive buried residue predictions, seven of them were aromatic, seven are charged or polar, two are prolines and one is an aliphatic residue. These positions have both reduced expression and binding for charged residue substitutions (**Supplementary Figure S14A, 14D**) similar to the buried residues (**Supplementary Figure S14B, 14E**). The specificity, sensitivity and accuracy of prediction is mentioned in **Supplementary Table S3**. Active site residues were identified with very high accuracy (**Supplementary Table S3**), though there were a few false negative and false positive predictions. Additionally, we found several positions which had Sortseq (expr) like WT, however, they had very low Sortseq (bind) (**Supplementary Figure S14A, 14D**). We hypothesize that these positions are also assisting in the maintenance of proper RBM conformation and enabling its binding to ACE-2. Residues 447, 448, 473 and 476 which gave false positive results, 447 and 476 are part of the receptor binding motif (RBM) and contain glycine in a conformation which is available only for glycine. Hence mutation to a non-Gly residue will likely disrupt the conformation of the RBM thus decreasing binding to ACE-2. Mutations at positions 446, 453, 493 and 498 gave false negative results. Of these false negative positions, 446 is again glycine. We found that the Arg mutants at N493 and N498 positions have very little effect on expression and binding (**Supplementary Figure S14C, 14F**). We hypothesized that these positions may not have the most optimal WT residue, or they may show no mutational penalty for binding to ACE-2. A recent report showed that the affinity of Q498R to ACE-2 is higher than WT RBD (Xue et al., 2020) and was enriched as double mutant Q498R/N501Y when selection was performed for RBD mutants having high affinity towards ACE-2 (Zahradník et al., 2021). It has also been reported that when chimeric virus evolved in the presence of neutralizing antibodies C121 and C141,

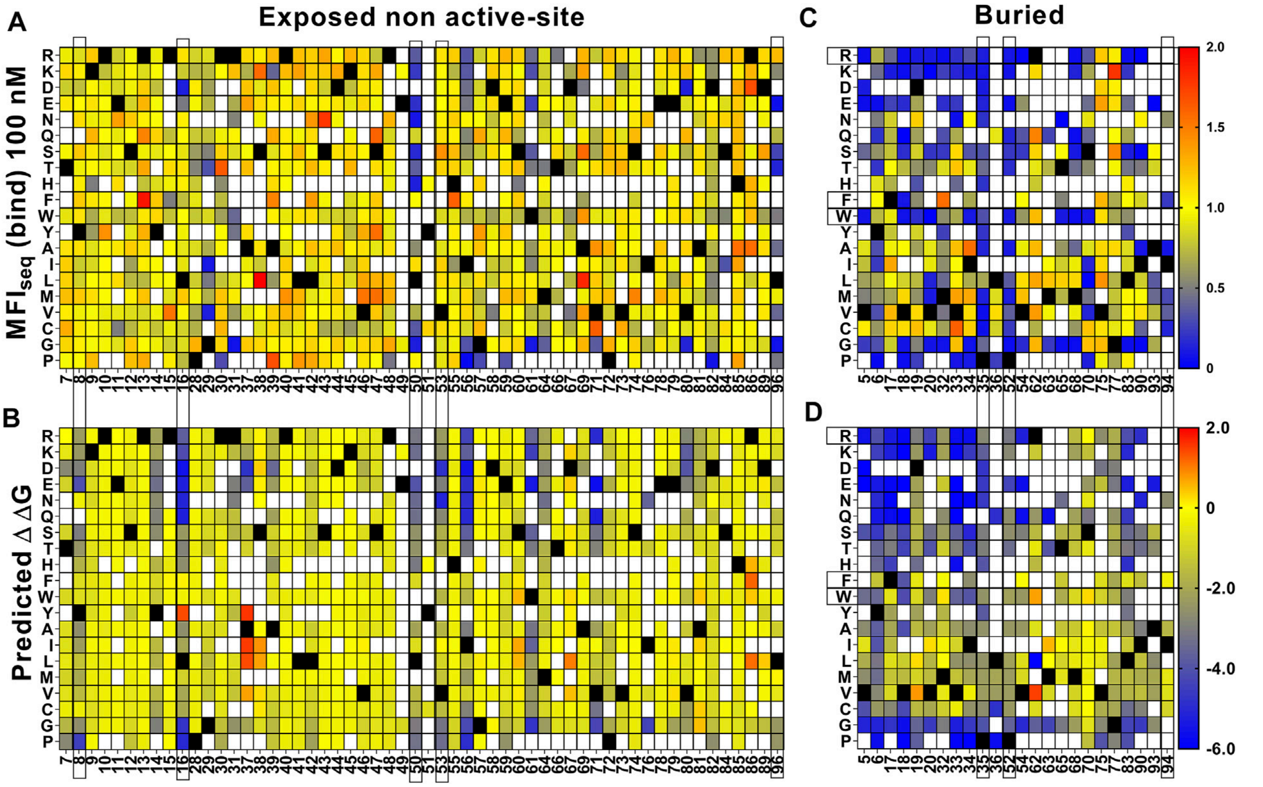


FIGURE 6 | Comparison of stabilities estimated by DeepDDG and yeast surface display. Heat maps for **(A,C)** MFI$_{seq}$ (bind) normalized to WT and **(B,D)** ΔΔG predicted by DeepDDG. Residue positions or specific amino acid mutations showing significantly different predicted stabilities by the two methods are highlighted by a box. Blue to red colour corresponds to increasing stability.

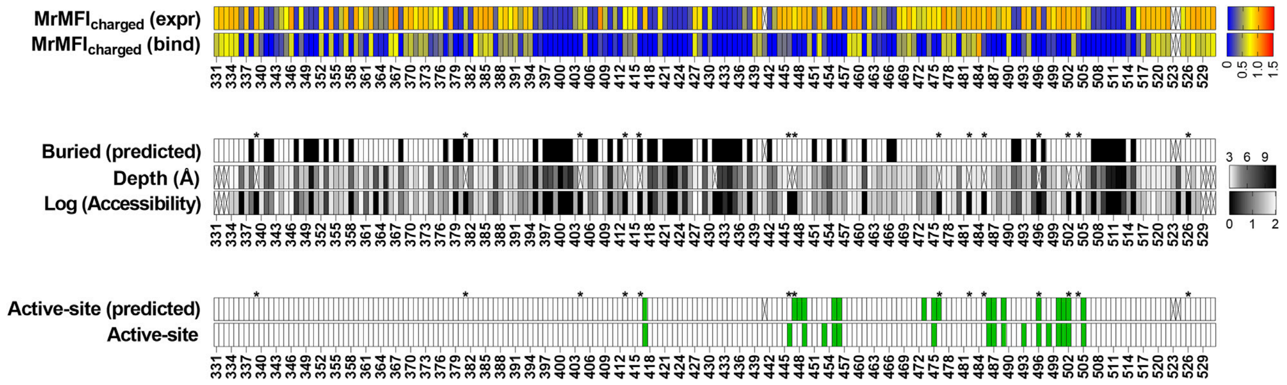


FIGURE 7 | Prediction of buried and active-site positions in SARS-CoV-2 RBD from Sortseq data. Buried residues were identified from chain C of PDB ID 7KMH, residues which had <10% side chain accessibility were categorized as buried. The accessibility and depth was calculated using DEPTH server (Tan et al., 2011). Active-site residues were identified from PDB ID 6M0J as explained earlier (Malladi et al., 2021). Criteria used to predict buried and active-site positions from MFI data were identical to those used for CcdB. Positions which did not have MrMFI data or could not be assigned to either buried or active-site categories are highlighted with "X". Accessibility calculated by DEPTH server for glycine is zero and these are marked with a "*".

this enriched for the Q493R mutation. The mutant virus grows to high PFU titers similar to WT, and infectivity is also inhibited by a chimeric ACE-2 analog, similar to WT (Weisblum et al., 2020). The specificity, sensitivity and accuracy of prediction is mentioned in **Supplementary Table S3**.

DISCUSSION

With the advancement of mutagenesis and directed evolution methodologies, proteins with modified traits and function can be developed in a relatively short duration of time (Chen and

Arnold, 1991; Winter et al., 1994; Bornscheuer et al., 2019). *E. coli* remains an expression host of choice for many proteins and high level, soluble *E. coli* expression is a desirable attribute. When eukaryotic or unstable prokaryotic proteins are overexpressed in bacteria, they often tend to form insoluble aggregates called inclusion bodies (IB). Formation of IBs often results in low yields of purified soluble protein. Designing improved variants of a protein by increasing half-life, stability and activity is an ongoing requirement of most pharmaceutical and biotechnology industries. However, a reliable, high-throughput, efficient and rapid method is required for solubility and stability analysis of engineered proteins. Previously, several high-throughput methods to select for soluble expression have been developed based on fusion to a reporter protein. These rely on the reporter activity, which is perturbed if an aggregation prone protein is fused (Maxwell et al., 1999; Waldo et al., 1999; Wigley et al., 2001; Fisher, 2006). These methods can be used to isolate protein variants with enhanced solubility but cannot reveal if the fused protein is properly folded. In some cases, such unstable proteins may also form soluble aggregates (Tripathi et al., 2016). Since many of these reporter screens employ cytoplasmic expression and use bacterial hosts, disulphide rich or glycosylated proteins, or those binding to complex ligands cannot be studied. Yeast surface display coupled to FACS, has been widely used to evolve such targets. Typically, populations are sorted for multiple rounds to enrich for stable binders to a target of interest (Kieke et al., 1999; Esteban and Zhao, 2004; Kim et al., 2006; Traxlmayr and Obinger, 2012). While this approach readily selects for high affinity binders, selecting for stable proteins is more difficult. In some cases, this methodology has also been used to isolate stable variants of proteins (Pepper et al., 2008) and a good correlation was observed between surface expression and improved biophysical parameters. However, other studies in different systems did not find such a correlation (Park et al., 2006; Piatesi et al., 2006).

In the present work we utilize YSD to measure the amount of total protein as well as total active protein displayed on the yeast cell surface. A good correlation was found between the amount of active CcdB mutant on the yeast surface and corresponding *in vivo* solubility in *E. coli* (r = 0.85) or T_m (r = 0.80). A recent report also suggests that the amount of active protein on the yeast cell surface can be used as a criterion to isolate stable mutants (Traxlmayr and Shusta, 2017). In the present study, no correlation was found between the amount of total protein on the yeast cell surface and the biophysical properties of mutants. A few mutants which have very low solubility in *E. coli* showed very high expression, but there was a negligible amount of active protein on the yeast surface. It has been previously suggested that the quality control system in yeast is not able to discriminate these mutants from properly folded ones or alternatively that the folded conformation is maintained by chaperones in the ER (Park et al., 2006). Once these mutants are exported to the cell surface they may start to unfold. This could be one reason why some groups including ours did not find a good correlation of surface expression with the stability or solubility of these proteins. In previous studies (Shusta et al., 1999), a very limited number of proteins were used for surface expression studies, it is possible that in this small number, mutants which had high surface expression or secretion but lower stability than WT were not observed.

Yeast surface display coupled to FACS typically requires multiple rounds of sorting to enrich variants with desired activity and phenotype. Here, we have performed a single round of sorting and developed a rapid, uncomplicated procedure of estimating MFI's of individual mutants of CcdB combining FACS and deep sequencing. This MFI_{seq} was shown to correlate well with the corresponding experimentally measured MFIs for several individual mutants. The MFI_{seq} was used to generate the mutational landscape of expression and binding of a mutant library. We showed that such data can be used to accurately discriminate between buried, exposed non active-site and exposed active-site residues both for CcdB and an unrelated protein, RBD of the spike protein of SARS-CoV-2. Highly destabilizing charged mutations in the core of the protein decreased both expression and binding, while the active-site residues showed reduction in binding alone for charged mutations. Relative to an earlier study which assayed *in vivo* activity in *E. coli* (Adkar et al., 2012), the present methodology is better able to identify and distinguish between the two categories of mutationally sensitive residues, namely buried and exposed, active-site residues. Identification of active-site residues of interacting partners through charged mutation scanning provides a better alternative to alanine and cysteine scanning mutagenesis. In general, mutations that affect total activity *in vivo* can do so by affecting specific activity without changing the amount of folded protein, decrease the amount of folded protein without affecting specific activity or a combination of the above. The present analysis distinguishes between the above possibilities, and is therefore able to distinguish buried from exposed, active-site positions. This is useful for applications that attempt to use saturation mutagenesis data for protein model discrimination and structure prediction (Khare et al., 2019; Jones E. M. et al., 2020) as well as interpreting clinical data on disease causing mutations (Findlay et al., 2018; Livesey and Marsh, 2020).

MFI_{seq} (bind) was also used to predict the T_m of CcdB mutants. We found a good correlation between predicted and measured ΔT_m for a subset of CcdB mutants. We also compared the accuracy of *in silico* approaches used to predict the stability of mutants and found that these predictors had lower accuracy relative to our approach. We used experimental stability measurements for a small number of destabilized mutations, combined with MFI_{seq} measurement to predict stabilities of all destabilized mutants in the saturation mutagenesis library. We could readily identify destabilized mutants of CcdB, however, the recovery of mutants more stable than WT was lower, but still significant, considering the rarity of such mutations. This is likely due to the possibility that if the stability of the protein crosses a threshold, additional increments in stability do not result in enhanced expression or binding.

A limitation of the present approach is that it requires an epitope tagged or fluorescently labelled conformation specific binding partner. Another limitation could be differential relative stability of proteins upon yeast cell surface display compared to expression in the native host and/or intracellular expression. For glycosylated proteins, the stability of mutants may also be altered because of hyper glycosylation of protein on the yeast cell surface compared to proteins expressed in mammalian systems or prokaryotic systems where glycosylation is absent. The presence of glycosylation may also affect the binding to a cognate partner which in turn may give rise to false results. This does not appear to be the case for the SARS-CoV-2 RBD which contains two glycans at residues 331 and 343, but may be an issue for proteins with multiple glycosylation sites. We are examining these possibilities in ongoing studies. Despite these caveats, the present study suggests that the proposed methodology can accurately distinguish buried from active-site residues, quantitatively estimate thermal stabilities of destabilized mutants in large libraries, and also be used with moderate accuracy to identify stabilized mutants.

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

RV and SA designed the experiments. SA performed all the experiments, RV and SA analyzed all the data. KM wrote the software and carried out the processing of the deep sequencing data. MB calculated the MFI of CcdB mutants using maximum likelihood method. RV and SA wrote most of the manuscript.

FUNDING

This work was funded by grants to RV from the Department of Science and Technology, grant number-EMR/2017/004054, DT.December 15, 2018), Government of India, Department of Biotechnology, grant no. BT/COE/34/SP15219/2015 DT. November 20, 2015, Ministry of Science and Technology, Government of India and Bill and Melinda Gates Foundation (United States) (INV-005948). We also acknowledge funding for infrastructural support from the following programs of the Government of India: DST FIST, UGC Centre for Advanced study, Ministry of Human Resource Development (MHRD), and the DBT IISc Partnership Program. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

ACKNOWLEDGMENTS

SA acknowledges Council of Scientific and Industrial Research for his fellowship (SPM-07/079(0218)/2015-EMR-I). KM is thankful to Department of Science and Technology (DST) Science and Engineering Research Board for financial support, sanction order no: PDF/2017/002641. MB acknowledges Council of Scientific and Industrial Research for her fellowship (SRF-09/079(2766)/2017-EMR-I). RV is a J. C. Bose Fellow of DST. Aparna Asok is duly acknowledged for FACS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.800819/full#supplementary-material

REFERENCES

Adkar, B. V., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P., et al. (2012). Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing. *Structure* 20, 371–381. doi:10.1016/j.str.2011.11.021

Aghera, N. K., Prabha, J., Tandon, H., Chattopadhyay, G., Vishwanath, S., Srinivasan, N., et al. (2020). Mechanism of CcdA-Mediated Rejuvenation of DNA Gyrase. *Structure* 28, 562–572.e4. doi:10.1016/j.str.2020.03.006

Bajaj, K., Dewan, P. C., Chakrabarti, P., Goswami, D., Barua, B., Baliga, C., et al. (2008). Structural Correlates of the Temperature Sensitive Phenotype Derived from Saturation Mutagenesis Studies of CcdB. *Biochemistry* 47, 12964–12973. doi:10.1021/bi8014345

Basanta, B., Bick, M. J., Bera, A. K., Norn, C., Chow, C. M., Carter, L. P., et al. (2020). An Enumerative Algorithm for De Novo Design of Proteins with Diverse Pocket Structures. *Proc. Natl. Acad. Sci. USA* 117, 22135–22145. doi:10.1073/pnas.2005412117

Bernard, P., and Couturier, M. (1992). Cell Killing by the F Plasmid CcdB Protein Involves Poisoning of DNA-Topoisomerase II Complexes. *J. Mol. Biol.* 226, 735–745. doi:10.1016/0022-2836(92)90629-X

Bhasin, M., and Varadarajan, R. (2021). Prediction of Function Determining and Buried Residues through Analysis of Saturation Mutagenesis Datasets. *Front. Mol. Biosci.* 8, 635425. doi:10.3389/fmolb.2021.635425

Boder, E. T., and Wittrup, K. D. (1997). Yeast Surface Display for Screening Combinatorial Polypeptide Libraries. *Nat. Biotechnol.* 15, 553–557. doi:10.1038/nbt0697-553

Boder, E. T., and Wittrup, K. D. (2000). [25] Yeast Surface Display for Directed Evolution of Protein Expression, Affinity, and Stability. *Meth. Enzym.* 328, 430–444. doi:10.1016/s0076-6879(00)28410-3

Bornscheuer, U. T., Hauer, B., Jaeger, K. E., and Schwaneberg, U. (2019). Directed Evolution Empowered Redesign of Natural Proteins for the Sustainable Production of Chemicals and Pharmaceuticals. *Angew. Chem. Int. Ed.* 58, 36–40. doi:10.1002/anie.201812717

Cambray, G., Guimaraes, J. C., and Arkin, A. P. (2018). Evaluation of 244,000 Synthetic Sequences Reveals Design Principles to Optimize Translation in *Escherichia coli*. *Nat. Biotechnol.* 36, 1005–1015. doi:10.1038/nbt.4238

Cao, H., Wang, J., He, L., Qi, Y., and Zhang, J. Z. (2019). DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. *J. Chem. Inf. Model.* 59, 1508–1514. doi:10.1021/acs.jcim.8b00697

Chakravarty, S., and Varadarajan, R. (1999). Residue Depth: a Novel Parameter for the Analysis of Protein Structure and Stability. *Structure* 7, 723–732. doi:10.1016/s0969-2126(99)80097-5

Chao, G., Lau, W. L., Hackel, B. J., Sazinsky, S. L., Lippow, S. M., and Wittrup, K. D. (2006). Isolating and Engineering Human Antibodies Using Yeast Surface Display. *Nat. Protoc.* 1, 755–768. doi:10.1038/nprot.2006.94

Chattopadhyay, G., and Varadarajan, R. (2019). Facile Measurement of Protein Stability and Folding Kinetics Using a Nano Differential Scanning Fluorimeter. *Protein Sci.* 28, 1127–1134. doi:10.1002/pro.3622

Chen, K., and Arnold, F. H. (1991). Enzyme Engineering for Nonaqueous Solvents: Random Mutagenesis to Enhance Activity of Subtilisin E in Polar Organic Media. *Nat. Biotechnol.* 9, 1073–1077. doi:10.1038/nbt1191-1073

Chen, J., Lu, Z., Sakon, J., and Stites, W. E. (2000). Increasing the Thermostability of Staphylococcal Nuclease: Implications for the Origin of Protein Thermostability. *J. Mol. Biol.* 303, 125–130. doi:10.1006/jmbi.2000.4140

Chen, Y., Lu, H., Zhang, N., Zhu, Z., Wang, S., and Li, M. (2020). PremPS: Predicting the Impact of Missense Mutations on Protein Stability. *PLOS Comput. Biol.* 16, e1008543. doi:10.1371/journal.pcbi.1008543

Chevalier, A., Silva, D.-A., Rocklin, G. J., Hicks, D. R., Vergara, R., Murapa, P., et al. (2017). Massively Parallel De Novo Protein Design for Targeted Therapeutics. *Nature* 550, 74–79. doi:10.1038/nature23912

Dao-Thi, M.-H., Van Melderen, L., De Genst, E., Buts, L., Ranquin, A., Wyns, L., et al. (2004). Crystallization of CcdB in Complex with a GyrA Fragment. *Acta Crystallogr. D Biol. Cryst.* 60, 1132–1134. doi:10.1107/S0907444904007814

Dehouck, Y., Kwasigroch, J. M., Gilis, D., and Rooman, M. (2011). PoPMuSiC 2.1: A Web Server for the Estimation of Protein Stability Changes upon Mutation and Sequence Optimality. *BMC Bioinformatics* 12, 151. doi:10.1186/1471-2105-12-151

Dou, J., Vorobieva, A. A., Sheffler, W., Doyle, L. A., Park, H., Bick, M. J., et al. (2018). De Novo design of a Fluorescence-Activating β-barrel. *Nature* 561, 485–491. doi:10.1038/s41586-018-0509-0

Esteban, O., and Zhao, H. (2004). Directed Evolution of Soluble Single-Chain Human Class II MHC Molecules. *J. Mol. Biol.* 340, 81–95. doi:10.1016/j.jmb.2004.04.054

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., et al. (2018). Accurate Classification of BRCA1 Variants with Saturation Genome Editing. *Nature* 562, 217–222. doi:10.1038/s41586-018-0461-z

Fisher, A. C. (2006). Genetic Selection for Protein Solubility Enabled by the Folding Quality Control Feature of the Twin-Arginine Translocation Pathway. *Protein Sci.* 15, 449–458. doi:10.1110/ps.051902606

Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., et al. (2010). High-resolution Mapping of Protein Sequence-Function Relationships. *Nat. Methods* 7, 741–746. doi:10.1038/nmeth.1492

Gietz, R. D., and Schiestl, R. H. (2007). High-efficiency Yeast Transformation Using the LiAc/SS Carrier DNA/PEG Method. *Nat. Protoc.* 2, 31–34. doi:10.1038/nprot.2007.13

Hagihara, Y., and Kim, P. S. (2002). Toward Development of a Screen to Identify Randomly Encoded, Foldable Sequences. *Proc. Natl. Acad. Sci.* 99, 6619–6624. doi:10.1073/pnas.102172099

Hubbard SJ, T. J. (1993). *NACCESS*. London: Dep. Biochem. Mol. Biol. Univ. Coll. London.

Jain, P. C., and Varadarajan, R. (2014). A Rapid, Efficient, and Economical Inverse Polymerase Chain Reaction-Based Method for Generating a Site Saturation Mutant Library. *Anal. Biochem.* 449, 90–98. doi:10.1016/j.ab.2013.12.002

Jones, L. L., Brophy, S. E., Bankovich, A. J., Colf, L. A., Hanick, N. A., Garcia, K. C., et al. (2006). Engineering and Characterization of a Stabilized α1/α2 Module of the Class I Major Histocompatibility Complex Product Ld. *J. Biol. Chem.* 281, 25734–25744. doi:10.1074/jbc.M604343200

Jones, B. E., Brown-Augsburger, P. L., Corbett, K. S., Westendorf, K., Davies, J., Cujec, T. P., et al. (2020a). LY-CoV555, a Rapidly Isolated Potent Neutralizing Antibody, Provides protection in a Non-human Primate Model of SARS-CoV-2 Infection. *Biorxiv Prepr. Serv. Biol.* doi:10.1101/2020.09.30.318972

Jones, E. M., Lubock, N. B., Venkatakrishnan, A., Wang, J., Tseng, A. M., Paggi, J. M., et al. (2020b). Structural and Functional Characterization of G Protein-Coupled Receptors with Deep Mutational Scanning. *Elife* 9, e61312. doi:10.7554/eLife.54895

Khare, S., Bhasin, M., Sahoo, A., and Varadarajan, R. (2019). Protein Model Discrimination Attempts Using Mutational Sensitivity, Predicted Secondary Structure, and Model Quality Information. *Proteins* 87, 326–336. doi:10.1002/prot.25654

Kieke, M. C., Shusta, E. V., Boder, E. T., Teyton, L., Wittrup, K. D., and Kranz, D. M. (1999). Selection of Functional T Cell Receptor Mutants from a Yeast Surface-Display Library. *Proc. Natl. Acad. Sci.* 96, 5651–5656. Availableat: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=21915&tool=pmcentrez&rendertype=abstract. doi:10.1073/pnas.96.10.5651

Kim, Y.-S., Bhandari, R., Cochran, J. R., Kuriyan, J., and Wittrup, K. D. (2006). Directed Evolution of the Epidermal Growth Factor Receptor Extracellular Domain for Expression in Yeast. *Proteins* 62, 1026–1035. doi:10.1002/prot.20618

Laimer, J., Hiebl-Flach, J., Lengauer, D., and Lackner, P. (2016). MAESTROweb: a Web Server for Structure-Based Protein Stability Prediction. *Bioinformatics* 32, 1414–1416. doi:10.1093/bioinformatics/btv769

Livesey, B. J., and Marsh, J. A. (2020). Using Deep Mutational Scanning to Benchmark Variant Effect Predictors and Identify Disease Mutations. *Mol. Syst. Biol.* 16, e9380. doi:10.15252/msb.20199380

Malladi, S. K., Singh, R., Pandey, S., Gayathri, S., Kanjo, K., Ahmed, S., et al. (2021). Design of a Highly Thermotolerant, Immunogenic SARS-CoV-2 Spike Fragment. *J. Biol. Chem.* 296, 100025. doi:10.1074/jbc.RA120.016284

Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., et al. (2018). Multiplex Assessment of Protein Variant Abundance by Massively Parallel Sequencing. *Nat. Genet.* 50, 874–882. doi:10.1038/s41588-018-0122-z

Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D., and Davidson, A. R. (1999). A Simple *In Vivo* Assay for Increased Protein Solubility. *Protein Sci.* 8, 1908–1911. doi:10.1110/ps.8.9.1908

Najar, T. A., Khare, S., Pandey, R., Gupta, S. K., and Varadarajan, R. (2017). Mapping Protein Binding Sites and Conformational Epitopes Using Cysteine Labeling and Yeast Surface Display. *Structure* 25, 395–406. doi:10.1016/j.str.2016.12.016

Niesen, F. H., Berglund, H., and Vedadi, M. (2007). The Use of Differential Scanning Fluorimetry to Detect Ligand Interactions that Promote Protein Stability. *Nat. Protoc.* 2, 2212–2221. doi:10.1038/nprot.2007.321

Noderer, W. L., Flockhart, R. J., Bhaduri, A., Diaz de Arce, A. J., Zhang, J., Khavari, P. A., et al. (2014). Quantitative Analysis of Mammalian Translation Initiation Sites by FACS -seq. *Mol. Syst. Biol.* 10, 748. doi:10.15252/msb.20145136

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., and Blundell, T. L. (2017). SDM: A Server for Predicting Effects of Mutations on Protein Stability. *Nucleic Acids Res.* 45, W229–W235. doi:10.1093/nar/gkx439

Park, S., Xu, Y., Stowell, X. F., Gai, F., Saven, J. G., and Boder, E. T. (2006). Limitations of Yeast Surface Display in Engineering Proteins of High Thermostability. *Protein Eng. Des. Sel.* 19, 211–217. doi:10.1093/protein/gzl003

Parthiban, V., Gromiha, M. M., and Schomburg, D. (2006). CUPSAT: Prediction of Protein Stability upon point Mutations. *Nucleic Acids Res.* 34, W239–W242. doi:10.1093/nar/gkl190

Pepper, L., Cho, Y., Boder, E. T., and Shusta, E.V. (2008). A Decade of Yeast Surface Display Technology: where Are We Now? *Cchts* 11, 127–134. doi:10.2174/138620708783744516

Peterman, N., and Levine, E. (2016). Sort-seq under the Hood: Implications of Design Choices on Large-Scale Characterization of Sequence-Function Relations. *BMC Genomics* 17, 206. doi:10.1186/s12864-016-2533-5

Piatesi, A., Howland, S. W., Rakestraw, J. A., Renner, C., Robson, N., Cebon, J., et al. (2006). Directed Evolution for Improved Secretion of Cancer-Testis Antigen NY-ESO-1 from Yeast. *Protein Expr. Purif.* 48, 232–242. doi:10.1016/j.pep.2006.01.026

Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014a). DUET: A Server for Predicting Effects of Mutations on Protein Stability Using an Integrated Computational Approach. *Nucleic Acids Res.* 42, W314–W319. doi:10.1093/nar/gku411

Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014b). MCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures. *Bioinformatics* 30, 335–342. doi:10.1093/bioinformatics/btt691

Prajapati, R. S., Das, M., Sreeramulu, S., Sirajuddin, M., Srinivasan, S., Krishnamurthy, V., et al. (2007). Thermodynamic Effects of Proline Introduction on Protein Stability. *Proteins* 66, 480–491. doi:10.1002/prot.21215

Pucci, F., Kwasigroch, J. M., and Rooman, M. (2020). Protein Thermal Stability Engineering Using HoTMuSiC. *Methods Mol. Biol.* 2112, 59–73. doi:10.1007/978-1-0716-0270-6_5

Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., et al. (2017). Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing. *Science* 357, 168–175. doi:10.1126/science.aan0693

Rodrigues, C. H., Pires, D. E., and Ascher, D. B. (2018). DynaMut: Predicting the Impact of Mutations on Protein Conformation, Flexibility and Stability. *Nucleic Acids Res.* 46, W350–W355. doi:10.1093/nar/gky300

Rost, B., and Sander, C. (1994). Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. *Proteins* 19, 55–72. doi:10.1002/prot.340190108

Sahoo, A., Khare, S., Devanarayanan, S., Jain, P. C., and Varadarajan, R. (2015). Residue Proximity Information and Protein Model Discrimination Using Saturation-Suppressor Mutagenesis. *Elife* 4, e09532. doi:10.7554/eLife.09532

Savojardo, C., Fariselli, P., Martelli, P. L., and Casadio, R. (2016). INPS-MD: a Web Server to Predict Stability of Protein Variants from Sequence and Structure. *Bioinformatics* 32, 2542–2544. doi:10.1093/bioinformatics/btw192

Schweickhardt, R. L., Jiang, X., Garone, L. M., and Brondyk, W. H. (2003). Structure-expression Relationship of Tumor Necrosis Factor Receptor Mutants that Increase Expression. *J. Biol. Chem.* 278, 28961–28967. doi:10.1074/jbc.M212019200

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., et al. (2012). Inferring Gene Regulatory Logic from High-Throughput Measurements of Thousands of Systematically Designed Promoters. *Nat. Biotechnol.* 30, 521–530. doi:10.1038/nbt.2205

Shusta, E. V., Kieke, M. C., Parke, E., Kranz, D. M., and Wittrup, K. D. (1999). Yeast Polypeptide Fusion Surface Display Levels Predict thermal Stability and Soluble Secretion Efficiency 1 1Edited by J. A. Wells. *J. Mol. Biol.* 292, 949–956. doi:10.1006/jmbi.1999.3130

Shusta, E., Pepper, L., Cho, Y., and Boder, E. (2008). A Decade of Yeast Surface Display Technology: Where Are We Now? *Cchts* 11, 127–134. doi:10.2174/138620708783744516

Smith, T. F., and Waterman, M. S. (1981). Identification of Common Molecular Subsequences. *J. Mol. Biol.* 147, 195–197. doi:10.1016/0022-2836(81)90087-5

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., et al. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* 182, 1295–1310.e20. doi:10.1016/j.cell.2020.08.012

Tan, K. P., Varadarajan, R., and Madhusudhan, M. S. (2011). DEPTH: a Web Server to Compute Depth and Predict Small-Molecule Binding Cavities in Proteins. *Nucleic Acids Res.* 39, W242–W248. doi:10.1093/nar/gkr356

Traxlmayr, M. W., and Obinger, C. (2012). Directed Evolution of Proteins for Increased Stability and Expression Using Yeast Display. *Arch. Biochem. Biophys.* 526, 174–180. doi:10.1016/j.abb.2012.04.022

Traxlmayr, M. W., and Shusta, E. V. (2017). "Directed Evolution of Protein thermal Stability Using Yeast Surface Display," in *Methods in Molecular Biology* (New York, NY: Humana Press), 45–65. doi:10.1007/978-1-4939-6857-2_4

Tripathi, A., Gupta, K., Khare, S., Jain, P. C., Patel, S., Kumar, P., et al. (2016). Molecular Determinants of Mutant Phenotypes, Inferred from Saturation Mutagenesis Data. *Mol. Biol. Evol.* 33, 2960–2975. doi:10.1093/molbev/msw182

Waldo, G. S., Standish, B. M., Berendzen, J., and Terwilliger, T. C. (1999). Rapid Protein-Folding Assay Using green Fluorescent Protein. *Nat. Biotechnol.* 17, 691–695. doi:10.1038/10904

Weisblum, Y., Schmidt, F., Zhang, F., DaSilva, J., Poston, D., Lorenzi, J. C., et al. (2020). Escape from Neutralizing Antibodies by SARS-CoV-2 Spike Protein Variants. *Elife* 9, e61312. doi:10.7554/eLife.61312

Wigley, W. C., Stidham, R. D., Smith, N. M., Hunt, J. F., and Thomas, P. J. (2001). Protein Solubility and Folding Monitored *In Vivo* by Structural Complementation of a Genetic Marker Protein. *Nat. Biotechnol.* 19, 131–136. doi:10.1038/84389

Winter, G., Griffiths, A. D., Hawkins, R. E., and Hoogenboom, H. R. (1994). Making Antibodies by Phage Display Technology. *Annu. Rev. Immunol.* 12, 433–455. doi:10.1146/annurev.iy.12.040194.002245

Wrenbeck, E. E., Klesmith, J. R., Stapleton, J. A., Adeniran, A., Tyo, K. E. J., and Whitehead, T. A. (2016). Plasmid-based One-Pot Saturation Mutagenesis. *Nat. Methods* 13, 928–930. doi:10.1038/nmeth.4029

Xue, T., Wu, W., Guo, N., Wu, C., Huang, J., Lai, L., et al. (2020). Single point Mutations Can Potentially Enhance Infectivity of SARS-CoV-2 Revealed by In Silico Affinity Maturation and SPR Assay. *bioRxiv*. doi:10.1101/2020.12.24.424245

Zahradník, J., Marciano, S., Shemesh, M., Zoler, E., Chiaravalli, J., Meyer, B., et al. (2021). SARS-CoV-2 RBD *In Vitro* Evolution Follows Contagious Mutation Spread, yet Generates an Able Infection Inhibitor. *bioRxiv*. doi:10.1101/2021.01.06.425392

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: A Fast and Accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620. doi:10.1093/bioinformatics/btt593

Zheng, L., Baumann, U., and Reymond, J.-L. (2004). An Efficient One-step Site-Directed and Site-Saturation Mutagenesis Protocol. *Nucleic Acids Res.* 32, e115. doi:10.1093/NAR/GNH110

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.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 27 January 2022
doi: 10.3389/fmolb.2021.787368

Molecular Architecture of the Antiophidic Protein DM64 and its Binding Specificity to Myotoxin II From *Bothrops asper* Venom

Barbara S. Soares[1], Surza Lucia G. Rocha[1], Viviane A. Bastos[1], Diogo B. Lima[2], Paulo C. Carvalho[3], Fabio C. Gozzo[4], Borries Demeler[5,6,7], Tayler L. Williams[5], Janelle Arnold[8], Amy Henrickson[6], Thomas J. D. Jørgensen[9], Tatiana A. C. B. Souza[3], Jonas Perales[1], Richard H. Valente[1], Bruno Lomonte[10], Francisco Gomes-Neto[1]* and Ana Gisele C. Neves-Ferreira[1]*

[1]Laboratory of Toxinology, Oswaldo Cruz Institute, Rio de Janeiro, Brazil, [2]Department of Chemical Biology, Leibniz Forschungsinstitut für Molekulare Pharmakologie (FMP), Berlin, Germany, [3]Laboratory for Structural and Computational Proteomics, Carlos Chagas Institute, Curitiba, Brazil, [4]Dalton Mass Spectrometry Laboratory, University of Campinas, Campinas, Brazil, [5]Department of Biochemistry and Structural Biology, University of Texas Health Science Center at San Antonio, San Antonio, TX, United States, [6]Department of Chemistry and Biochemistry, University of Lethbridge, Lethbridge, AB, Canada, [7]Department of Chemistry and Biochemistry, University of Montana, Missoula, MT, United States, [8]Department of Environmental Science, Princeton University, Princeton, NJ, United States, [9]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark, [10]Clodomiro Picado Institute, University of Costa Rica, San José, Costa Rica

DM64 is a toxin-neutralizing serum glycoprotein isolated from *Didelphis aurita*, an ophiophagous marsupial naturally resistant to snake envenomation. This 64 kDa antitoxin targets myotoxic phospholipases A_2, which account for most local tissue damage of viperid snakebites. We investigated the noncovalent complex formed between native DM64 and myotoxin II, a myotoxic phospholipase-like protein from *Bothrops asper* venom. Analytical ultracentrifugation (AUC) and size exclusion chromatography indicated that DM64 is monomeric in solution and binds equimolar amounts of the toxin. Attempts to crystallize native DM64 for X-ray diffraction were unsuccessful. Obtaining recombinant protein to pursue structural studies was also challenging. Classical molecular modeling techniques were impaired by the lack of templates with more than 25% sequence identity with DM64. An integrative structural biology approach was then applied to generate a three-dimensional model of the inhibitor bound to myotoxin II. I-TASSER individually modeled the five immunoglobulin-like domains of DM64. Distance constraints generated by cross-linking mass spectrometry of the complex guided the docking of DM64 domains to the crystal structure of myotoxin II, using Rosetta. AUC, small-angle X-ray scattering (SAXS), molecular modeling, and molecular dynamics simulations indicated that the DM64-myotoxin II complex is structured, shows flexibility, and has an anisotropic shape. Inter-protein cross-links and limited hydrolysis analyses shed light on the inhibitor's regions involved with toxin interaction, revealing the critical participation of the first, third, and fifth domains of DM64. Our data showed that the fifth domain of DM64 binds to myotoxin II amino-terminal and beta-wing regions. The third domain of the inhibitor acts in a complementary way to the fifth domain. Their binding to these toxin regions

OPEN ACCESS

Edited by:
Paulo Ricardo Batista,
Oswaldo Cruz Foundation (Fiocruz),
Brazil

Reviewed by:
Santosh Panjikar,
Australian Synchrotron, Australia
André Zelanis,
Federal University of São Paulo, Brazil

***Correspondence:**
Francisco Gomes-Neto
gomes.netof@gmail.com
Ana Gisele C. Neves-Ferreira
anagextra@gmail.com

Specialty section:
This article was submitted to
Structural Biology,
a section of the journal
Frontiers in Molecular Biosciences

Received: *30 September 2021*
Accepted: *07 December 2021*
Published: *27 January 2022*

Citation:
Soares BS, Rocha SLG, Bastos VA, Lima DB, Carvalho PC, Gozzo FC, Demeler B, Williams TL, Arnold J, Henrickson A, Jørgensen TJ, Souza TACB, Perales J, Valente RH, Lomonte B, Gomes-Neto F and Neves-Ferreira AGC (2022) Molecular Architecture of the Antiophidic Protein DM64 and its Binding Specificity to Myotoxin II From Bothrops asper *Venom.*
Front. Mol. Biosci. 8:787368.
doi: 10.3389/fmolb.2021.787368

presumably precludes dimerization, thus interfering with toxicity, which is related to the quaternary structure of the toxin. The first domain of DM64 interacts with the functional site of the toxin putatively associated with membrane anchorage. We propose that both mechanisms concur to inhibit myotoxin II toxicity by DM64 binding. The present topological characterization of this toxin-antitoxin complex constitutes an essential step toward the rational design of novel peptide-based antivenom therapies targeting snake venom myotoxins.

Keywords: cross-linking (XL), mass spectrometry, immunoglobulin fold, structural biology, toxin neutralisation, protein inhibitor, snake envenomation, antiophidic activity

1 INTRODUCTION

DM64 is made up of five immunoglobulin-like (Ig-like) domains and, to our knowledge, is the only mammalian inhibitor of snake venom myotoxic phospholipases A_2 (PLA_2) (Rocha et al., 2002; Rocha et al., 2017). This antitoxin is homologous to DM43, a three Ig-like domain-protein inhibitor grouped into the MEROPS I43 family of Ig-related proteins that binds snake venom metalloendopeptidases (SVMP) with nanomolar affinity (Neves-Ferreira et al., 2002; Rocha et al., 2009; Brand et al., 2012). Both toxin-scavenging proteins are homologous to human α1B-glycoprotein (Ishioka et al., 1986), a cysteine-rich secretory protein 3 (CRISP3) binding protein (Udby et al., 2004) that has been associated with several pathologic conditions (Clerc et al., 2016).

DM64 and DM43 constitute biochemical defenses employed by the South American marsupial *Didelphis aurita* to avoid the toxic activity of viperid snake venoms [reviewed in (Neves-Ferreira et al., 2009; Bastos et al., 2016; Neves-Ferreira et al., 2017)]. Since opossums are prey and predators of snakes, toxin resistance is a doubly advantageous evolutionary trait (Arbuckle et al., 2017). Despite the complex composition of viperid venoms (Gutiérrez et al., 2017), serum-based inhibitors seem restricted to PLA_2 and SVMP toxin families, which suffices against the harmful effects of bothropic envenomation. Therefore, unveiling the molecular basis of such resistance might open important biotechnological perspectives.

The rational design of peptide-based antivenom therapeutics inspired by natural inhibitors of venom toxins is a long-sought goal that has been hampered by the lack of structural information on these proteins. The large immunoglobulin superfamily (CATH 2.60.40.10) is the most diverse regarding the number of structural domains (Bork et al., 1994; Sillitoe et al., 2019). The Ig fold is structurally very stable and shows remarkable plasticity in terms of binding specificity. Proteins that use this fold perform many functions related to immunological recognition, and they are also associated with several developmental and homeostatic phenomena (Bork et al., 1994; Natarajan et al., 2015). Among 7,053 unique PDB structures currently classified within this superfamily (CATH + release 4.3), none shares more than 25% sequence identity with DM64, precluding *in silico* structure prediction by homology modeling. The recently published AlphaFold 2 prediction algorithm (Jumper et al., 2021) may generate more accurate models of the Ig-like domains of DM64, but their correct spatial positioning remains challenging. Previous attempts to crystallize DM64 have failed, most likely due to the negative impact of its glycan moieties, preventing structural analysis by X-ray crystallography from being performed.

In such a challenging scenario, one can benefit from integrative structural strategies, where different lower-resolution techniques may be combined to generate complementary structural information on proteins or protein complexes. Mass spectrometry (MS) and associated methods, such as cross-linking (XL) and hydrogen-deuterium exchange (HDX), are among the most widely used integrative approaches. Such experimental results can be integrated into computational modeling/docking pipelines, substantially improving structural model accuracy (Liu and Heck, 2015; Faini et al., 2016; Koukos and Bonvin, 2019).

In this study, distance restraints from XL-MS provided unprecedented insights into the overall topology of the complex made by DM64 and myotoxin II from *Bothrops asper* venom. This basic toxin of ~14 kDa is a catalytically-inactive PLA_2 homolog (Lys49 variant, Group IIA) that induces significant local myotoxicity (Lomonte and Gutiérrez, 1989; Francis et al., 1991). Complementary information derived from limited proteolysis, analytical ultracentrifugation (AUC), small-angle X-ray scattering (SAXS), and molecular dynamics simulations improved our confidence in the proposed assembly model.

2 MATERIALS AND METHODS

2.1 DM64 and Myotoxin II

Mature DM64 was isolated from *Didelphis aurita* serum as described (Rocha et al., 2002). For simplicity, the protein sequence without the signal peptide (residues 25–504 of Q8MIS3) was renumbered 1–480 in the present study. Dr. Paulo Sergio D'Andrea undertook wild animals' capture and handling protocols (Laboratory of Biology and Parasitology of Wild Mammals Reservoirs, Oswaldo Cruz Institute), with the approval of the Federal Environment Agency (Permanent License 13373-1) and the Oswaldo Cruz Institute Ethics Commission on Animal Use (CEUA/IOC L-036/2018). All research activities carried out with the Brazilian genetic heritage were registered in the National System of Genetic Resource Management and

Associated Traditional Knowledge (SisGen AF0A111). Myotoxin II (P24605) was isolated from *Bothrops asper* venom as published (Mora-Obando et al., 2014a). Protein quantitation was based on 280 nm molar extinction coefficients: DM64 57,005 $M^{-1}cm^{-1}$ and myotoxin II 21,275 $M^{-1}cm^{-1}$ (Pace et al., 1995).

2.2 Molecular Mass Determination by Mass Spectrometry

DM64 was desalted, concentrated, and introduced into the mass spectrometer by a system composed of a dual pump Agilent 1200 HPLC system, a Rheodyne manual injection valve, and a VALCO 10-port valve. Fifty picomoles of the inhibitor were initially applied to a Waters MassPREP micro desalting column (2.1 × 5.0 mm; 20 μm; 1,000 Å) previously equilibrated with 0.23% (v/v) formic acid in water. Sample application/desalting/concentration proceeded at a flow rate of 300 μL/min for 5 min using the aforementioned mobile phase. The column was then set in line with the second HPLC pump so that the sample could be eluted from the column directly into the mass spectrometer. In this case, the flow rate was set to 75 μL/min, and the eluents were 0.23% (v/v) formic acid in 95% water (mobile phase A) and 0.23% (v/v) formic acid in 95% acetonitrile (mobile phase B). Dimethyl sulfoxide (DMSO) at a level of 5% was added to both solutions to increase analyte charging (Iavarone and Williams, 2003). The column was equilibrated with 5% B for 1.5 min, followed by a linear gradient from 5 to 50% B lasting 13.5 min, and 50–95% B for 2.0 min. Sample elution was monitored online on a Waters Synapt G1 HDMS instrument set as follows: source voltages—capillary (3.5 kV), sampling cone (40.0 V), and extraction cone (4.0 V); temperatures—source (100 °C) and desolvation (250°C); MCP detector voltage (1,700 V). Data were submitted for analysis using the MassLynx software package (Waters). Background subtraction settings for "polynomial order", "below curve (%)", and "tolerance" were set to 25, 5, and 0.01, respectively. A representative "zoomed" 1,280–1,420 m/*z* range was then chosen for processing with MaxEnt 1, a maximum entropy deconvolution software (Ferrige et al., 1992). Regarding MaxEnt 1 settings, the resolution was set for 0.50 Da/channel for a "uniform Gaussian damage model" with minimum intensity ratios of 33% (left and right) and allowing the algorithm to iterate until convergence. DM64 had a final molecular mass range iteration across 64,000 to 67,000 with a width at half height of 1.00 Da.

2.3 Analytical Ultracentrifugation

Oligomerization properties of DM64 and complex formation of DM64 with myotoxin II were studied by sedimentation velocity. DM64 and myotoxin II control experiments were performed at three different loading concentrations to examine the potential for mass action-induced oligomerization. For DM64, the concentration range spanned roughly one order of magnitude: 0.81 μM (measured at 230 nm), 4.2, and 9.0 μM (measured at 280 nm). For myotoxin II, we examined 25.7 μM (measured at 280 nm), 82.6, and 119 μM (measured at 295 nm). Mixtures of both proteins were measured at three molar ratios (1:1, 2:1, and 4:1 myotoxin:DM64, all measured at 230 nm). Sedimentation experiments were performed on a Beckman Proteomelab XLI AUC at the Center for Analytical Ultracentrifugation of Macromolecular Assemblies (CAUMA) at the University of Texas Health Science Center at San Antonio (UTHSCSA), and on a Beckman Optima AUC instrument at the Canadian Center for Hydrodynamics at the University of Lethbridge. All samples were measured at 20°C and 55 krpm by UV intensity detection using an An60Ti rotor and standard 2-channel titanium centerpieces (Nanolytics, Potsdam, Germany) or standard 2-channel epon-filled centerpieces (Beckman-Coulter, Indianapolis). All data were analyzed with UltraScan 4.0 release 6,153 (Demeler and Gorbet, 2016). All samples were measured in a 10 mM sodium phosphate buffer containing 50 mM NaCl, pH 7.4. Hydrodynamic buffer density (1.0012 g/cm^3) and viscosity (1.0065 cp) were estimated with UltraScan (Demeler and Gorbet, 2016).

Sedimentation velocity (SV) data were analyzed as reported earlier (Demeler, 2010). Optimization was performed by 2-dimensional spectrum analysis [2DSA, (Brookes et al., 2010)] with simultaneous removal of time- and radially invariant noise contributions and fitting of boundary conditions. 2DSA solutions were subjected to parsimonious regularization by genetic algorithm analysis (Brookes and Demeler, 2007) and, where applicable, by the enhanced van Holde–Weischet analysis (Demeler and van Holde, 2004). A further refinement using Monte Carlo analysis (Demeler and Brookes, 2008) was also applied to determine confidence limits for the determined parameters. The calculations were carried out on high-performance computing platforms at the Texas Advanced Computing Center and the San Diego Supercomputing Center (Brookes and Demeler, 2008).

For hydrodynamic radius (R_h) estimates, SV experiments were fitted to finite element solutions of the Lamm equation to obtain a solute's sedimentation and diffusion coefficients from the experimental data (Cao and Demeler, 2008). The diffusion coefficient, D, is given by:

$$D = \frac{RT}{Nf}$$

Where R is the universal gas constant, T is the absolute temperature, N is Avogadro's number, and f is the translational frictional coefficient of the solute. Combining the diffusion coefficient with the Stokes-Einstein relationship (Einstein, 1905) allows us to calculate the hydrodynamic radius, R_h:

$$R_h = \frac{RT}{6\pi\eta ND}$$

Where η is the viscosity of the solvent.

2.4 Size Exclusion Chromatography (SEC)

The stoichiometry of the interaction between DM64 and myotoxin II was analyzed by SEC, as described (Bastos et al., 2020), with modifications. A fixed amount of DM64 was titrated with increasing concentrations of myotoxin II (1:0.5; 1:1; 1:2; 1:4 mol/mol) in 0.1 M Tris-HCl pH 7.5, containing 0.5 M NaCl. The mixtures were incubated at 25°C for 15 min and then injected on a Superdex 200 Increase 5/150 GL column (GE Healthcare), previously equilibrated with the same buffer. The chromatogram peaks were integrated, and the areas corresponding to free myotoxin or to free DM64 co-eluted with DM64-myotoxin complex were plotted. The experiment was run in duplicate.

2.5 Small-Angle X-Ray Scattering

SAXS data were collected at the SAXS2 beamline (Brazilian Synchrotron Light Laboratory, Campinas, Brazil). The radiation wavelength was set to 1.48 Å. The scattering vector ranged from 0.1 to 1.5 nm^{-1}. DM64 was mixed with myotoxin II (1:1 mol/mol) in 20 mM Tris-HCl pH 7.5, containing 20 mM $CaCl_2$ and 150 mM NaCl. Frames with an exposure time of 300 or 600 s were recorded. To guarantee an accurate solvent correction, buffer baselines were collected under identical conditions before and after sample data collection. Background scattering was subtracted from the protein scattering pattern, which was then normalized and corrected. Data were processed and analyzed with the ATSAS package (Franke et al., 2017). Experimental data fitting and evaluation of the pair-distance distribution function P(r) were performed using the program GNOM (Svergun, 1992a). The radius of gyration (Rg) was estimated by the indirect Fourier transform method, and the maximum dimension Dmax was calculated from the actual space P(r) function as the distance r, where the P(r) value reaches zero. The theoretical scattering curve and radius of gyration of DM64-myotoxin complex were calculated using CRYSOL (Svergun et al., 1995). The comparison of Kratky plots for free DM64 and the complex DM64-myotoxin II was performed normalizing the scattering data for I (0) = 1 and multiplying q by Rg in order to remove the protein size information and keep the shape information (Kikhney and Svergun, 2015).

2.6 Cross-Linking Reaction

DM64 (20 μg) and myotoxin II were dissolved in 20 mM HEPES pH 7.5 and incubated at 25°C for 15 min (1:1 and 1:2 mol/mol). The noncovalent complex formed between both proteins was stabilized with BS^3 (bis(sulfosuccinimidyl)suberate, Thermo Fisher Scientific) for 90 min at 25°C. The reaction was stopped by quenching in 20 mM ammonium bicarbonate. Protein (5–10 μM final concentration) to cross-linker ratios ranging from 1:500 to 1:2,800 mol/mol (the latter corresponding to 1:20 w/w) were tested. DM64 and myotoxin II controls were individually submitted to the same experimental procedure, using the higher protein to cross-linker ratio. Aliquots (10%) of each sample were analyzed by native polyacrylamide gel electrophoresis (PAGE) and sodium dodecyl sulfate (SDS-PAGE) under reducing conditions (12%T, silver staining) (Laemmli, 1970; Heukeshoven and Dernick, 1985). Molecular mass markers were from GE Healthcare. The remaining aliquots (90%) were injected on a Superdex 200 HR 10/30 column (GE Healthcare) to remove excess BS^3 and protein aggregates. Finally, the samples were digested in solution with Lys-C endopeptidase (1:100 E:S, w/w) and trypsin (1:50 E:S, w/w) as previously described (Bastos et al., 2020). The experiment was run in duplicate.

2.7 Interaction Between Cross-linked DM64 and Native Myotoxin II

The functional integrity of DM64 cross-linked with BS^3 was inferred based on its ability to interact with native myotoxin II, following incubation at 25°C for 15 min (1:1 mol/mol). Electrophoretic mobility shift assay on nondenaturing polyacrylamide gels was used to monitor complex formation. Protein bands were excised from gels, destained in 50% acetonitrile/25 mM ammonium bicarbonate pH 8.0, and further submitted to in-gel digestion (Shevchenko et al., 1996), modified as follows: 1) protein reduction was performed in 65 mM DTT in 100 mM NH_4HCO_3 pH 8.0 for 30 min at 56°C; 2) proteins were alkylated in 200 mM iodoacetamide in 100 mM NH_4HCO_3 pH 8.0 for 30 min at ambient temperature in the dark; 3) gel bands were swollen in a digestion buffer containing 40 mM NH_4HCO_3 pH 8.0 and 20 ng/μL trypsin (Promega V511); 4) for peptide extraction, gel bands were initially subjected to an ultrasonic bath treatment for 10 min, followed by vortexing for 20 s before transferring the supernatant volume to a new tube. An additional extraction step was performed following the addition of 30 μL of 5% formic acid in 50% acetonitrile to the gel bands, which were sequentially submitted to vortexing (for 20 s), incubation at ambient temperature (15 min), ultrasonic treatment (2 min) and vortexing (20 s). This extraction cycle was repeated one more time. Recovered supernatants were combined and dried on a vacuum centrifuge. Dried samples were redissolved in 1% formic acid before peptide identification by MS/MS on a QExactive Plus orbitrap (Thermo Scientific) (Fioramonte et al., 2018).

PatternLab for Proteomics V (PLV) (Carvalho et al., 2016) was used for peptide-spectrum matching and filtering, using a database consisting of the sequences of DM64 and myotoxin II included in the PLV standard contaminant library (total of 125 target sequences). Database searches considered semi-tryptic specificity, maximum of two missed cleavages, and the following modifications: fixed carbamidomethylation of cysteine (+57.02146 Da), variable oxidation of methionine (+15.9949 Da), and variable hydrolyzed BS^3 (dead-end) in lysine, serine, and N-terminal (+156.0786 Da). The results were filtered for a maximum error of 10 ppm (precursor and fragment-ions), accepting an FDR of 1%.

2.8 Mass Spectrometry Analysis of Cross-Linked Peptides

Cross-linked peptides were analyzed by nLC-nESI-MS/MS on a Dionex UltiMate™ 3,000 RSLCnano system coupled online to a QExactive Plus orbitrap mass spectrometer (Thermo Scientific) (Fioramonte et al., 2018), modified as follows: 1) a PicoFrit column (75 μm inner diameter, 15 μm tip) (New Objective) packed in-house (40 cm length) with ReproSil-PurC18-AQ (1.9 μm resin, 200 Å pore size) (Dr. Maisch GmbH) was used in all chromatographic runs; 2) Higher-energy Collisional

Dissociation (HCD) was performed with stepped Normalized Collision Energy (sNCE) at 30 and 40. At least two technical replicates of each biological replicate were analyzed. The software *Spectrum Identification Machine for Cross-linked Peptides* (SIM-XL) (Lima et al., 2015; Fioramonte et al., 2018) was used. It searched uninterpreted MS/MS raw files against a local database, which comprised the FASTA sequences of DM64 and mature myotoxin II, with no decoy sequences included. The following parameters were used: full enzymatic specificity (Lys-C + trypsin), ≤3 missed cleavages allowed, mass tolerance filters of 5 ppm for both MS1 and MS2, fixed modification of cysteine residues (carbamidomethyl, +57.02146 Da), and hydrolyzed BS^3 dead-end (+156.0786 Da) as variable modification. A mass shift of +138.068100 Da was assumed to be associated with peptides cross-linked with BS^3. The software considered only MS/MS spectra that contained at least one of the following XL reporter ions: *m/z* 222.149, *m/z* 239.1759, *m/z* 240.159, or *m/z* 305.2229. The assumed reaction site specificities for BS^3 were KK, KS, SS, KN-TERM, SN-TERM. Identified spectra were submitted to automatic filtering (SIM-XL score ≥3; annotated fragment-ions/peptide chain ≥3), followed by careful manual verification (Iacobucci and Sinz, 2017). All curated cross-links met the following criteria: 1) High signal-to-noise ratio spectrum, where only minor peaks were unassigned; 2) Good sequence coverage for both cross-linked chains; 3) Unequivocal positioning of the cross-linking site; 4) Homogeneous mass error distribution for fragment-ions.

2.9 Limited Protein Hydrolysis and Affinity Purification of Interacting Peptides

Hitrap® NHS-activated affinity column (1 ml) containing immobilized myotoxin II was prepared as described (Rocha et al., 2002). It was used as bait to fish potential interacting peptides generated by enzymatic hydrolysis of DM64. Briefly, DM64 (200 μg) was dissolved in 0.4 M ammonium bicarbonate and 8 M urea, reduced with DTT and alkylated with iodoacetamide (Camacho et al., 2016). Using a ZebaTM Spin column (Thermo, 7 KMWCO), the sample was desalted, the buffer was exchanged to Tris-HCl 25 mM pH 8.5 containing 1 mM EDTA, and protein digestion with Lys-C endopeptidase (Promega) proceeded for 20 h at 37°C (1:100 E:S, w/w). The hydrolysate was chromatographed through the affinity column coupled with the toxin. Bound and unbound peptide fractions were desalted on self-packed Poros® 20 R2 (Applied Biosystems) tip columns, dried on a vacuum centrifuge, redissolved in 1% formic, and further analyzed by MS/MS (Camacho et al., 2016). Each fraction was analyzed in technical duplicate.

Alternatively, affinity fractions were directly injected onto a Vydac C18 column (2.1 mm × 15 cm). The collected peaks were dried, redissolved in 1% formic acid, and individually analyzed by MS/MS. Reversed-phase chromatography was performed at 0.2 ml/min, employing 0.1% TFA (solvent A) and 0.1% TFA in acetonitrile (mobile phase B) and the following elution gradient: 5% B for 5 min, 45%B at 55 min, 75% B at 60 min. A similar approach was used for fishing myotoxin II peptides generated by Glu-C endopeptidase (Promega), with the following modifications: after denaturation, reduction and alkylation, the toxin was desalted on a C18 spin column (Harvard Apparatus) equilibrated with 25 mM ammonium carbonate pH 7.8; the toxin was enzymatically digested with Glu-C for 20 h at 25°C at a 1:20 E:S ratio (w/w); the affinity column containing immobilized DM64 was prepared as described (Rocha et al., 2017).

Peaks X algorithm (Bioinformatics Solutions Inc.) was used to search MS/MS spectra against a database consisting of DM64 and myotoxin II sequences, embedded in a set of 20,533 human proteins retrieved from SwissProt, in addition to common contaminants (https://www.thegpm.org/crap). The following parameters were used for database searching: error tolerance of 10 ppm and 0.02 Da respectively for precursor- and fragment-ions; up to two missed cleavages allowed and fully-tryptic cleavage specified; carbamidomethylation of cysteines (+57.02146 Da) as fixed modification and no variable modification included. Identification results were filtered to accept ≤1% FDR at the spectral, peptide, and protein levels.

2.10 Molecular Modeling

The amino acid sequence of DM64 (Q8MIS3) was retrieved from the Uniprot database. The structures of the mature inhibitor or its isolated five Ig-like domains were initially obtained through homology modeling calculations with the automated server "Iterative Threading ASSEmbly Refinement" (I-TASSER) (Roy et al., 2010; Yang et al., 2015). Then, models were visually inspected, and images were rendered using Visual Molecular Dynamics version 1.9.3 (Humphrey et al., 1996).

Cross-linked residues identified by mass spectrometry were converted into distance restraints, assigning a Euclidean distance between Cβ atoms of each pair of reactive residues. Euclidean and topological distances were calculated applying TopoLink v1.18 (Ferrari et al., 2019). The software output classified all cross-linking results, correlating Euclidean and topological distances and defining distance limits for each cross-link. For the cross-linking agent BS^3, the maximum extended Euclidean distances (D^{Euclid}) was defined as 1) Lys-Lys, 21.8 Å; 2) Lys-Ser, 18.0 Å; Ser-Ser, 14.1 Å.

2.11 Molecular Docking

The interaction between DM64 and myotoxin II was studied by molecular docking, applying "Rosetta Docking Protocol" (Gray et al., 2003; Wang et al., 2005) guided by distance constraints derived from cross-linking experiments. First, the crystal structure of the myotoxin II monomer was retrieved from the PDB databank under the identification code 1CLP (chain A). Next, myotoxin II structure and DM64 Ig-like domains' models were relaxed with standard protocols by "Rosetta's short Relax Protocol," constraining backbone and side-chain heavy atoms based on the input structure and turning off-ramp down constrains. One hundred structures were generated for each protein, selecting the lowest total energy structure as input to the molecular docking (Nivon et al., 2013; Conway et al., 2014).

The complete view of the interaction between DM64 and myotoxin II was obtained in five steps: 1) docking D5-myoII, 2) docking D4-D5myoII, 3) docking D3-D4D5myoII, 4) docking

D2-D3D4D5myoII, and 5) docking D1-D2D3D4D5myoII. Each docking step consisted of three phases: in the first one, a single PDB file was created manually, adding a pair of proteins (or a protein and a partial complex) displaced 20 Å away from each other; in the second phase, 150,000 runs of global docking were performed, aiming to generate an initial model for the interaction based on cross-linking distance restraints. In this phase, docking partners' positions were spun and randomized before each run, avoiding a bias created by choosing initial coordinates for the docking. Docking models were analyzed by the TopoLink software and further classified by the number of violated constraints. The group with the lowest number of violated constraints was clustered according to each constraint (validated or not) sorted by Rosetta's total energy. The lowest total energy score model was then chosen as a starting coordinate for the local docking step.

In the third phase, 150,000 runs of local docking were performed, sampling the conformational space around the previously selected conformation, searching for the best agreement with experimental data. In this phase, the first docking partner (myotoxin II alone or in the complex) was fixed, and the incoming DM64 domain was perturbed by 3 Å translation and 8° rotation before each run. TopoLink was then used to analyze models and classify them according to the number of violated constraints. The group with the lowest number of violated constraints was clustered according to each constraint (validated or not) and sorted by the Rosetta energy terms: TotalScore, InterfaceScore (I_sc), and ConstraintScore (AtomPair). The lowest ConstraintScore was chosen as the final docking model and used as starting coordinate for docking the next DM64 domain.

2.12 Molecular Dynamics Simulation

The docked model of the DM64-myotoxin II complex was used as starting coordinates for the molecular dynamics (MD) simulations. Simulations were performed by Gromacs version 2020.3 (Berendsen et al., 1995) with Amber99SB force field (Jorgensen et al., 1996) and TIP3P water model (Jorgensen et al., 1983). The Propka module determined the ionization states of side chains for DM64 and myotoxin II in the PDB2PQR server at pH 7.5 (Dolinsky et al., 2007). The starting model was centered in a cubic water box at 2 nm of the box limits in any dimension. The charges were neutralized with counterions.

Energy minimization was applied using steepest descent and conjugated gradient algorithms with a final maximum force of 1,000 kJ/mol/nm (0.01 nm step size, cutoff of 1.2 nm for neighbor list, Coulomb interactions, and van der Waals interactions). Before the production runs, sequential equilibration processes in NVT (1.0 ns) and NPT (1.0 ns) were performed to adjust the systems into the desired temperatures and volumes, under position restraints with the force constants of 1,000 kJ/(mol·nm^2) applied to the protein complex. LINCS algorithm and virtual sites were used with a time step of 2 fs (cutoff of 1.4 nm neighbor list, Coulomb interactions, and van der Waals interactions). The temperature was stabilized by a V-rescale thermostat (Bussi et al., 2007), and pressure was coupled to a Parrinello-Rahman barostat (Parrinello and Rahman, 1981).

The position restraints were released gradually in 26 NVT equilibrium steps (200 ps each) with the force constants 1,000, 800, 600, 400, 300, 250, 200, 175, 150, 125, 100, 75, 50, 25, 23, 20, 18, 16, 14, 12, 10, 8, 6, 4, 2 and 0 kJ/(mol·nm2). After the preparation phase, 600 ns of MD simulation was accumulated and fitted, considering myotoxin II as the system's center.

Gromacs native package was used to analyze: 1) root mean square deviation (RMSD); 2) root mean square fluctuation (RMSF) of atomic positions in the trajectory; RMSD cluster analysis using Gromos clustering algorithm with 2 nm cutoff; 4) radius of gyration (Rg), and 5) hydrogen bonds with a maximum distance of 3.5 Å between donor and acceptor and a maximum angle of 30° between donor, donor-hydrogen and acceptor. The percentual time a hydrogen bond was present in the simulation was extracted from the interprotein hydrogen bond existence matrix, using a bash script that extracts the total number of occurrences of a hydrogen bond in the total number of frames of the MD simulation.

Principal component analysis (PCA) evaluated the conformational space sampled in the MD simulation of the DM64-myotoxin II complex. First, Gromacs module "covar" was applied to calculate and diagonalize the (mass-weighted) covariance matrix. Second, Gromacs module "anaeig" was used to analyze the eigenvectors of the covariance matrix. The main conformational motions were represented by the projections of the first and second principal components (PC1 and PC2, respectively). Trajectories were visualized, and images were rendered using Visual Molecular Dynamics software version 1.9.3

3 RESULTS

3.1 DM64-Myotoxin Interaction: Molecular Mass and Stoichiometry Analyses

The molecular mass of DM64 determined by ESI-Q-TOF was 65.37 kDa (**Figures 1A,B**). Sedimentation velocity experiments were performed to analyze the oligomerization state of DM64 and myotoxin II. AUC measurements resulted in molar masses in excellent agreement with the known molar masses of monomeric toxin and inhibitor (**Table 1**). Sedimentation profiles from increasing concentrations of those two proteins (**Figures 2A,B**) did not produce any change in the sedimentation behavior or molar mass, suggesting the absence of mass action in the examined concentration range.

For DM64, glycosylation (~18.5%) was expected to affect the partial specific volume. Using the ESI-Q-TOF result for DM64 (65.37 g/mol), the ExPASy computed average molar mass from protein sequence alone (53.3 g/mol, https://web.expasy.org/compute_pi), and the partial specific volume predicted with UltraScan based on protein sequence (0.7342 ccm/g for DM64 and 0.7266 ccm/g for myotoxin II), together with an average partial volume of 0.622 ccm/g for glycosylation reported by (Lewis and Junghans, 2000), we arrived at a modified partial specific volume of 0.7134 ml/g for the inhibitor, and 0.7161 ml/g

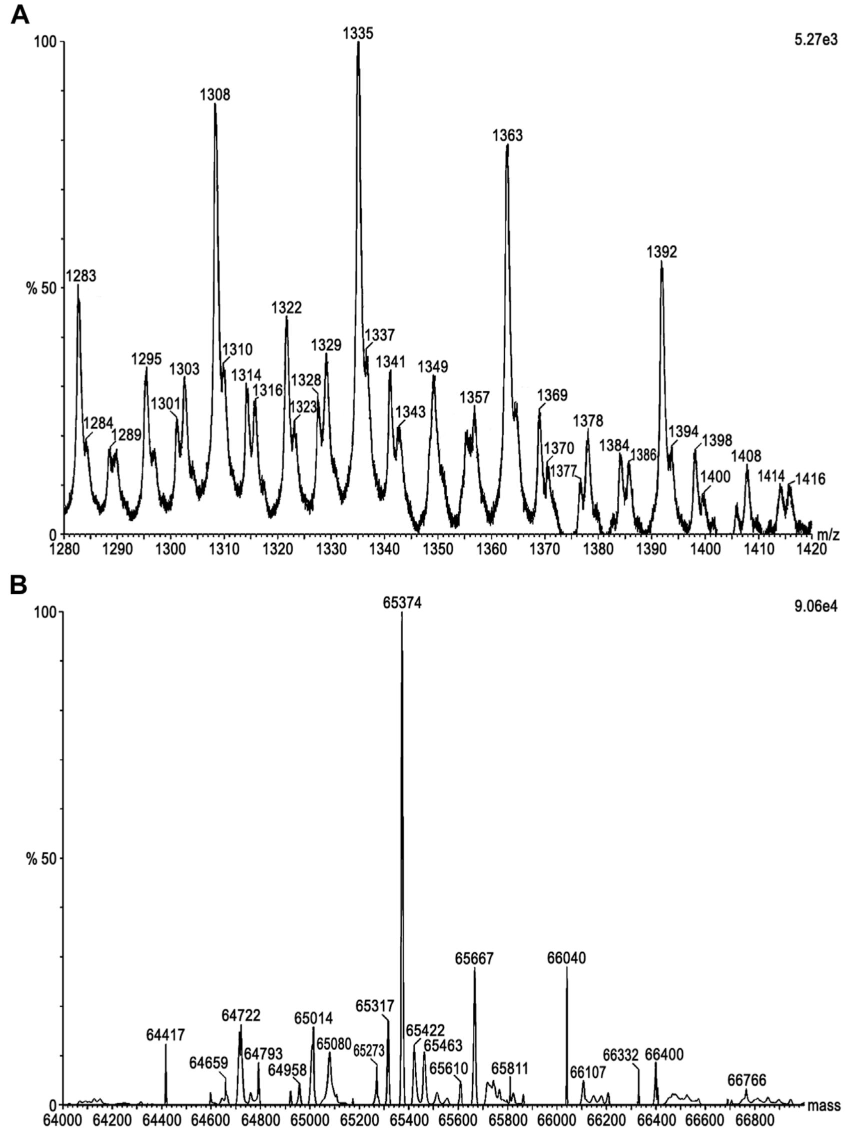


FIGURE 1 | Molecular mass determination of DM64 by ESI-MS. The setup was composed of an Agilent 1200 HPLC system coupled to a Waters Synapt G1 mass spectrometer. DM64 (50 pmol) was desalted on a Waters MassPREP microdesalting column (2.1 × 5.0 mm; 20 µm; 1,000 Å) followed by direct elution into the ion source for MS analysis. Acquired data were submitted to analysis with the MassLynx software, including **(A)** background subtraction, choice of *m/z* range for further processing, and **(B)** deconvolution using MaxEnt 1 algorithm to generate the final mass values.

TABLE 1 | Molecular mass determination by sedimentation velocity experiments.

	Sedimentation coefficient (s, x10^{-13})	Diffusion coefficient (cm^2/sec, x10^{-7})	Hydrodynamic radius (nm)	Molar mass (Dalton, x10^3)	PSV (ml/g)	f/f$_0$
Myotoxin II (60 µM)	1.84	12.10	1.77	13.5	0.7266	1.13
DM64 (1 µM)	4.03	5.41	3.96	63.0	0.7134	1.52
DM64:myotoxin II 1:1	4.84	5.08	4.22	81.5	0.7161	1.48

for the 1:1 DM64-myotoxin II complex. Molar mass transformations based on the adjusted partial specific volumes are shown in **Table 1**.

When mixed at stoichiometric amounts, myotoxin II and DM64 formed a homogeneous species consistent with a 1:1 complex, and sedimented at 4.84 s, faster than the monomeric

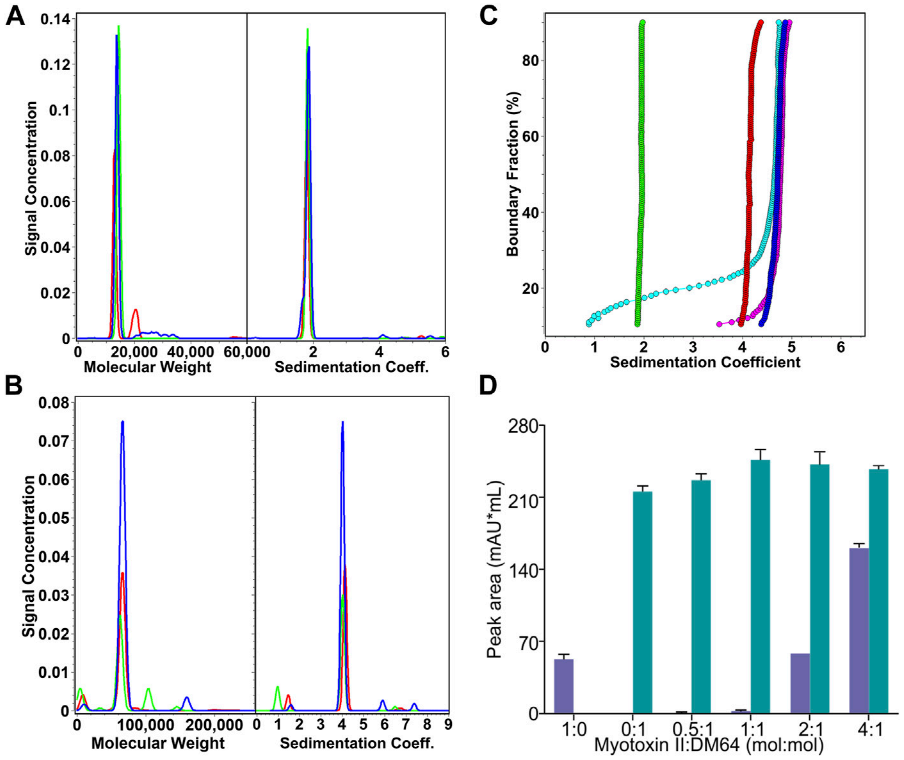


FIGURE 2 | Analysis of the stoichiometry of the interaction between DM64 and myotoxin II. **(A)** Sedimentation coefficient distributions and molar mass estimates of myotoxin II controls (25.7 µM, green; 82.6 µM, red; 119 µM, blue). **(B)** Sedimentation coefficient distributions and molar mass estimates of DM64 controls (0.81 µM, red; 4.2 µM, green; 9.0 µM, blue) **(C)** van Holde-Weischet integral G(s) distributions of myotoxin II:DM64 mixtures (1:1, blue; 2:1, magenta; 4:1, cyan) and myotoxin II (green) and DM64 (red) monomeric controls. **(D)** Size exclusion chromatography analysis of myotoxin II and DM64 mixed at different molar ratios. The peak areas in the chromatograms were integrated and plotted. Because the retention times for DM64 alone and the complex DM64-myotoxin II are very similar, they were plotted as the sum of co-eluted peaks (green bars). Free myotoxin II eluted at a different elution time (blue bars).

DM64 (4.03 s) and myotoxin II (1.84 s) controls. The 1:1 DM64: myotoxin II mixture was measured at a protein absorbance of 0.29 OD at 230 nm, corresponding to a molar concentration of 800 nM for each protein. This concentration is near the detection limit of the analytical ultracentrifuge as buffer interference increases at lower wavelengths. Our analysis clearly shows that all protein molecules are in a complex at this concentration (**Figure 2C**, dark blue line). A slower species representing either free DM64 or free myotoxin II could not be detected. When the relative molar ratio of the toxin was increased to 2:1 or 4:1, the same complex was observed with an unchanged sedimentation coefficient. Still, additional slower species, representing free myotoxin II, were observed (**Figure 2C**, cyan and magenta lines). These results indicate that the K_D is at most 800 nM, but most likely lower, reflecting a very strong interaction. DM64 binds with an equimolar amount of myotoxin II, and excess myotoxin II accumulates as free monomer. The molar mass transformation based on the adjusted partial specific volume for the 1:1 myotoxin II-DM64 complex is shown in **Table 1**. It is in excellent agreement with the sum of monomeric DM64 and monomeric myotoxin II.

The equimolar stoichiometry was corroborated by analytical size exclusion chromatography (**Figure 2D**). The peak area corresponding to co-eluting free DM64 and the toxin-antitoxin complex increased until a 1:1 myotoxin:DM64 molar ratio was reached. At a 2:1 myotoxin:DM64 molar ratio, no further increase in the area of the co-eluting peaks was detected; instead, a peak corresponding to free myotoxin was observed, and its area was equivalent to *ca.* 1-fold molar excess of myotoxin. When a 4:1 myotoxin-DM64 molar ratio was assayed, the peak area of the free toxin was proportional to *ca.* 3-fold molar excess of myotoxin II.

3.2 Mapping DM64-Myotoxin II Complex by Cross-Linking Mass Spectrometry

The complex made of DM64 and myotoxin II was stabilized using BS^3. This cross-linker reacts with nucleophilic groups, such as primary amines and, to a lesser extent, hydroxyl groups (Sinz, 2006). Control samples of myotoxin II and DM64 were individually submitted to the same protocol. Using BS^3, the number of cross-linkable residues (Lys + Ser) in DM64 represented 12% of the total number of residues. In contrast, in the very basic myotoxin II, the number of potential target residues was higher, covering 19.8% of the protein sequence.

To ensure high cross-linking yields with acceptable structural distortion, DM64 was initially reacted with increasing BS^3 concentrations (1,000- to 2,800-fold molar excess over

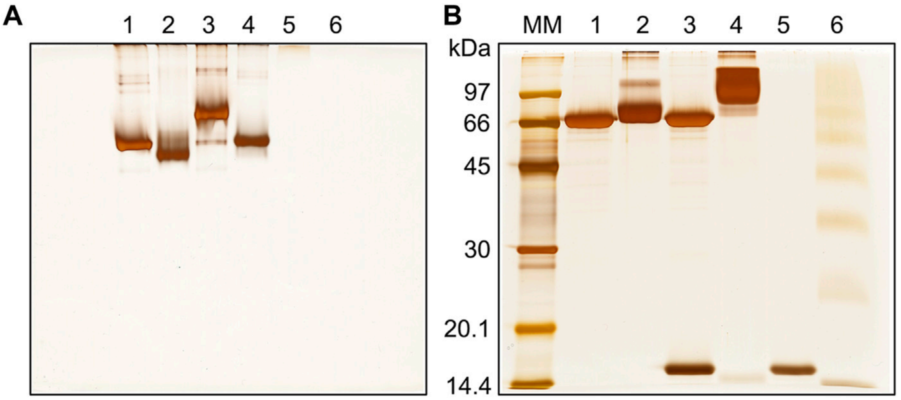


FIGURE 3 | Gel electrophoresis analysis of the cross-linking reaction between DM64 and myotoxin II (1:1 mol/mol). The toxin-antitoxin noncovalent complex was stabilized with BS^3 (90 min reaction time at 25°C, protein to cross-linker ratio of 1:2,800 mol/mol) and analyzed by **(A)** native PAGE and **(B)** SDS-PAGE under reducing conditions. All samples were run on 12% T gels stained with silver nitrate. MM: molecular mass markers; lane 1: DM64; lane 2: DM64 + BS^3; lane 3: DM64 + myotoxin II (mtx II); lane 4: (DM64 + mtx II) + BS^3; lane 5: mtx II; lane 6: mtx II + BS^3.

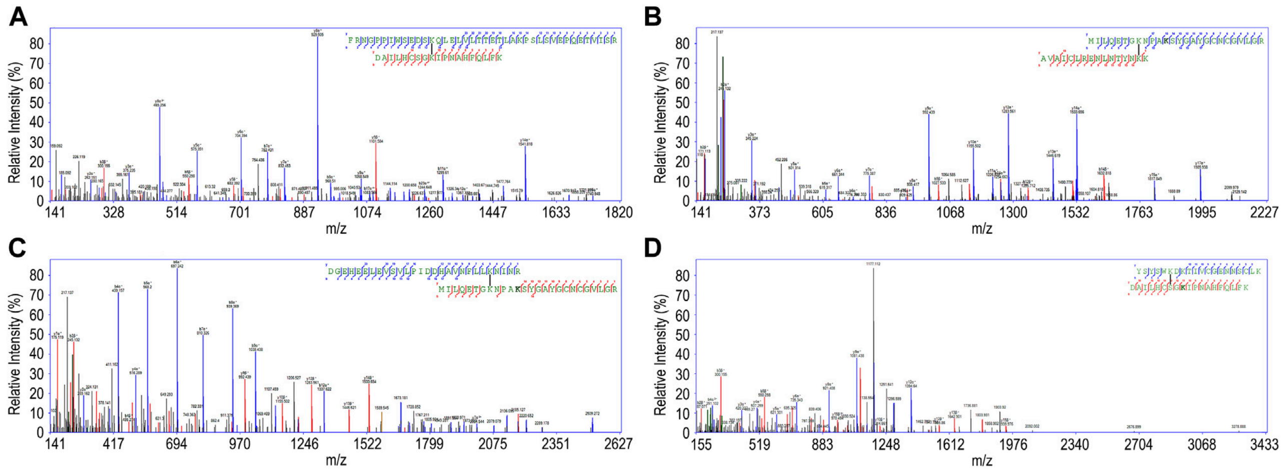


FIGURE 4 | Representative good-quality fragmentation spectra of peptides cross-linked with BS^3, interpreted by the SIM-XL software, and manually verified. **(A)** Intra-protein link from DM64 (scan # 27,856, SIM-XL primary score 6.74); **(B)** Intra-protein link from myotoxin II (scan # 20,073, score 4.62); Inter-protein link from DM64-myotoxin II complex: **(C)** Scan # 28,261, score 5.86; **(D)** Scan # 21,754, score 4.32. For each spectrum, the sequences of the cross-linked peptides are shown, along with all observed b- and y-ions (blue and red lines, respectively for α- and β-peptides) and the position of the linker (black line).

protein). Inference on protein fold integrity was based on the ability of cross-linked DM64 to interact with unmodified myotoxin II. Such interaction was monitored by nondenaturing gel electrophoresis (**Supplementary Figure S1**). Gel bands suspected to represent the toxin-antitoxin complex were analyzed by MS/MS. Under all experimental conditions, sequences corresponding to myotoxin II and DM64 were obtained from the same gel band. Thus, the comigration of toxin and antitoxin was interpreted as evidence of protein interaction. Such evidence suggests that eventual structural disturbances due to the chemical cross-linking of DM64 were somewhat limited.

To further optimize the experimental conditions, the toxin-antitoxin complex was stabilized with different concentrations of BS^3. No significant artifactual aggregation/oligomerization was observed by native PAGE as the cross-linker concentration increased (**Supplementary Figure S2A**). On the other hand, the SDS-PAGE profile showed that the band broadening effect, typical of the cross-linking reaction (Leitner et al., 2014), was more pronounced when higher concentrations of BS^3 were used (**Supplementary Figure S2B**). Therefore, all subsequent cross-linking experiments were performed with the highest molar excess of this cross-linker.

BS^3-stabilized toxin-antitoxin complex migrated as a single sharp band on the native gel, with slower mobility than cross-linked DM64 (**Figure 3A**). Due to its basic nature (pI > 9) (Lomonte and Rangel, 2012), control myotoxin II did not enter the stacking gel (**Figure 3A**, lanes 5–6). The complex stabilized

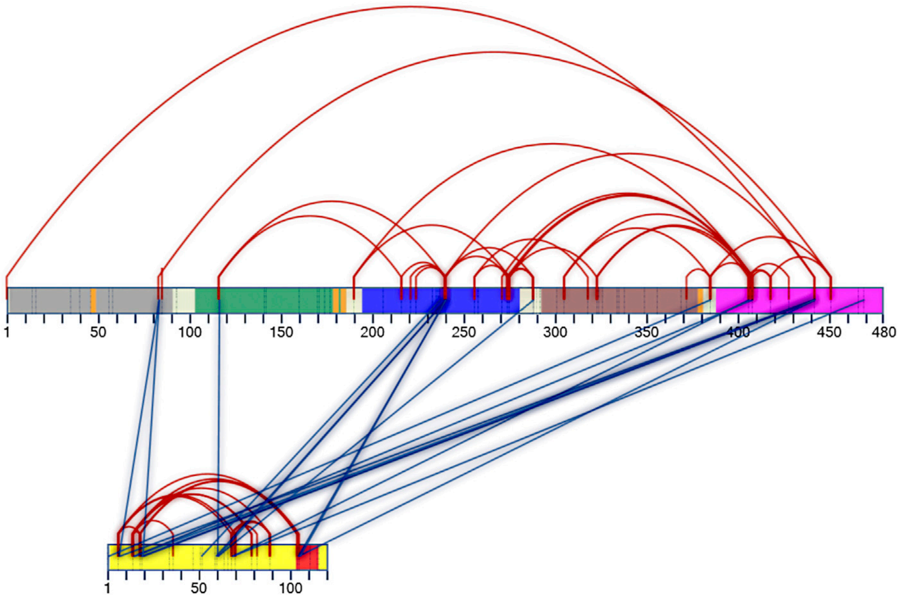


FIGURE 5 | Two-dimensional interaction map of DM64 in complex with myotoxin II. Protein sequences are represented by numbered rectangles in which vertical dashed lines indicate all potential target residues for cross-linking reaction. DM64 protein is colored according to Ig-like domains: D1: gray; D2: green; D3: blue; D4: brown; D5: pink. The four consensus sequences for N-glycosylation are shown in orange, and inter-domain linker regions are shown in white. The protein sequence of myotoxin II is represented by a yellow rectangle, with the C-terminal cationic/hydrophobic myotoxic site colored in red. Intra-protein cross-links are represented by red lines, whereas inter-protein cross-links are shown in blue. Following identification by the SIM-XL software and manual verification, only good-quality spectra were used to build the 2D map.

with BS^3 was visualized as a broad higher molecular mass band on reducing SDS-PAGE, and virtually no bands corresponding to free myotoxin or DM64 could be observed (**Figure 3B**, lane 4). The high number of lysines in myotoxin II led to several higher-order oligomers when the toxin was cross-linked with BS^3 in the absence of DM64 (**Figure 3B**, lane 6), thus precluding the subsequent analyses of this sample.

Several spectra corresponding to cross-linked peptides from the complex made of DM64-and myotoxin II were identified by the SIM-XL software (representative examples in **Figure 4**). They were manually inspected to allow the selection of high-confidence cross-linking-spectrum matches (CSM). Sequence coverages of the DM64-myotoxin II complex, based on the primary structures of intra- (type 1) and inter- (type 2) cross-linked peptides (Schilling et al., 2003), were: 84.3% for myotoxin II and 69.8% for DM64 (not shown). In the case of DM64, non-covered regions corresponded to the surrounding sequences of the four N-glycosylation sites. Similar results were obtained when DM64 was cross-linked without the toxin.

Ideally, cross-links should be validated on experimentally determined high-resolution structures before targeting unknown 3D structures (Merkley et al., 2014). Therefore, as a proof-of-principle test, all manually curated intra-myotoxin II BS^3 links (observed when the toxin was in complex with DM64) were mapped on the crystal structure of the toxin (1CLP, chain A) (**Supplementary Figure S3**). No aberrant cross-links could be detected, although 4 out of 15 unique links (73% validation) exceeded the maximum allowed topological distance between the β-carbon of cross-linked residues, calculated from the X-ray structure (**Supplementary Tables S1A,C**). They involve residues located in regions of the toxin with higher average B-factor values and may reflect the structural flexibility of the protein in solution.

Regarding the connections within DM64, 32 unique cross-links were identified when the protein was complexed with myotoxin II (red lines in **Figure 5** and **Supplementary Table S1D**). A similar pattern was observed when DM64 was cross-linked in its free form (not shown), and 39 unique cross-links were confidently identified (**Supplementary Table S1E**). In both cases, several unique cross-links connected residues far apart in the primary structure of the inhibitor, clearly indicating the close spatial proximity of the corresponding cross-linking sites.

Eighteen unique inter-protein cross-links were confidently identified between DM64 and myotoxin II (blue lines in **Figure 5** and **Supplementary Table S1F**). Half of the links connected the fifth domain of DM64 (residues Lys386, Ser407, Lys409, Lys443, and Ser470) with myotoxin II, five of which were centered at Lys443, a cross-linking hotspot in the inhibitor sequence. Most cross-linked residues in the fifth domain of DM64 (6 out of 9 cross-links) connected the inhibitor to the N-terminal region of the toxin. The third domain of DM64 showed the second-highest number of cross-links with myotoxin II. In total, five cross-links were confidently identified, almost always involving the inhibitor's residue Lys241 and the middle (Lys60 and Lys61) or the C-terminal (Lys105 and Lys106) regions of the toxin.

3.3 Identification of Potential Interaction Regions Between DM64 and Myotoxin II

The myotoxin II hydrolysate bound to the DM64 affinity column was subjected to off-line C18 reversed-phase fractionation, and the collected fractions were identified by MS/MS. Only one toxin peptide, corresponding to the C-terminal end (99-NLNTYNKKYRYYLKPLCKKADAC-121), was confidently identified. The same result was obtained when the bound fraction was directly analyzed by MS/MS, without pre-fractionation. Similarly, using both methodological strategies, one DM64 peptide was identified in the hydrolysate bound to myotoxin II. Its sequence was located in the fifth domain of the inhibitor (420-DGEHEELEVSVLPIDDHAVNFLLK-443). Two other DM64 peptides were confidently identified only when MS/MS directly analyzed the bound hydrolysate: one peptide located in the third domain of the inhibitor (258-YSCRYRFRNGPPIWSEDSK-276) and a second one corresponding to its C-terminal end (453-YRCRYTTREDPILESEMSDPAELQVTGQ-480). All peptides bound to the affinity columns were deemed too long to be used as constraints in the docking strategies. Nevertheless, they served as additional evidence to further validate the toxin-antitoxin model, as discussed later.

3.4 Molecular Modeling

3.4.1 DM64: Complete Primary Structure

In a first attempt, DM64 was modeled based on its whole amino acid sequence, applying the algorithms I-TASSER (Yang et al., 2015) or Rosetta (Yang et al., 2020). The models produced by each strategy were convergent, showing an extended spatial distribution of the five Ig-like domains (not shown). All models were then submitted to topological validation using experimentally derived structural data on the isolated inhibitor generated by cross-linking mass spectrometry. The model with the highest number of validated cross-links was chosen to represent this strategy.

Intra-domain cross-links (13 of 39 observed cross-links) were used to validate individual Ig folds (**Supplementary Table S1B**). In contrast, inter-domain cross-links (26 out of 39) were used to analyze the model's global structure. In the first case, an overall validation rate of 61.5% was obtained, as follows: 75.0% validation within the third domain of DM64 (3 validated links/4 observed links); 0% in D4 (0 validation/1 observation); 62.5% in D5 (5 validations/8 observations). Regarding inter-domain cross-links, only one could be validated out of 26 identified links (3.8%).

3.4.2 DM64: Individual Ig-like Domains

The modeling strategy was further optimized to incorporate myotoxin II structure as a reference point to guide the spatial positioning of DM64 domains. Initially, individual DM64 Ig-like domains were modeled applying the same algorithms used for the whole protein. Each domain extension was defined in this first step based on the Ig-like fold, and the transition regions for consecutive domains were identified. **Supplementary Figure S4** shows the primary structure of DM64, with residues delimiting the N- and C-terminal regions of each domain colored in yellow. Conserved connecting regions that allow flexibility and plasticity between the domains could be identified for each domain pair and are highlighted in cyan. The sequence of each Ig-like domain was then adjusted to include four to five residues of the connecting region, including a security range in each domain's terminal region. **Table 2** shows the final extension of all modeled domains, together with their main structural features. Next, individually modeled domains were used to build the model structure of DM64 bound to myotoxin II, following a series of molecular docking steps guided by distance restraints generated by cross-linking mass spectrometry. Only cross-links observed in the toxin-complexed inhibitor were considered.

3.5 Molecular Docking Between Individual Ig-like Domains of DM64 and Myotoxin II

For the sake of clarity during the molecular docking process, each individual Ig-like domain of DM64 was indexed by a letter, as follows: D1 (chain A), D2 (chain B), D3 (chain C), D4 (chain D), D5 (chain E). Myotoxin II (myoII) was referred to as chain F. To facilitate interpreting the results, all main structural features of the myotoxin II crystal structure were highlighted in the ribbon representation shown in **Supplementary Figure S5**. Detailed results of all docking steps performed between DM64 and myotoxin II described below are shown in **Supplementary Table S1A**.

3.5.1 First Docking Step: D5-Myotoxin II

Nine cross-links were identified between the fifth domain of DM64 and myotoxin II (**Supplementary Figure S6**), the second-highest number of cross-links observed. Only two were not validated in the model structure generated by molecular docking: Lys386E-Ser1F (N-terminal residue of myotoxin II) and Lys409E-Lys15F (residue located in the small helix of myotoxin II). On the other hand, the remaining seven observed cross-links were topologically validated using the same reference structure, yielding an overall validation rate of 77.8%. The docking results indicate that D5 is spatially close to the N-terminal helix of myotoxin II (cross-link observed between Lys443E-Lys7F) and the small helix region (Lys443E-Lys15F, Lys443E-Lys19F, and Lys443E-Ser20F). D5 is likely also close to the beta-wing region (Ser407E-Lys69F and Ser470E-Lys69F) (Arni and Ward, 1995). Interestingly, the topologically validated cross-link Lys443E-Lys105F approximates D5 and myotoxin Lys105, the first residue of the myotoxic peptide located in the C-terminal region of the toxin.

3.5.2 Second Docking Step: D4-D5Myotoxin II

Of six cross-links involving the fourth domain of DM64, four could be topologically validated in the available model structure (66.7% validation rate) (**Supplementary Figure S7**). A single cross-link was identified between D4 and myotoxin II: Lys289D-Lys61F was spaced by a valid topological distance connecting D4 and the frontal loop of the toxin. This region leads to the beta-wing, which is implicated in the interaction between myotoxin II monomers. The remaining five cross-links connect the fourth and the fifth domains of DM64, three of which were topologically

TABLE 2 | Extension and main characteristics of the immunoglobulin-like domains of DM64.

Domain	Start	End	Cys-Cys	Glycosylation site	Residues
D1	001-LAMET-005	091-VTGKE-095	027–073	019-PWT**N**VTL-025	95
D2	099-APLLR-103	183-VVIPD-187	120–162	155-**N**NTG**N**YS-161	89
D3	191-KPDFH-195	281-VLTTE-285	212–260	—	95
D4	289-KPSLS-293	377-EIRVE-381	311–358	351-YDTG**N**FS-357	93
D5	386-KPTLH-390	476-QVTGQ-480	406–455	—	95

*Domain, Sequential name of DM64 domains; Start, starting sequence and residue number for each domain; End, ending sequence and residue number for each domain; Cys-Cys, disulfide bond identified by residue numbers; Glycosylation Site, N-glycosylation sites in consensus sequence (**N**X-S/T) and residue numbering; Residues, number of residues in the domain.*

validated (Lys306D-Lys386E, Lys324D-Ser407E, and Lys324D-Lys409E). It is worth mentioning that, in the final model of the complexed toxin-antitoxin structure, the N-glycosylation site in the fourth domain of DM64 (Asn355) (**Table 2**) was exposed at the surface as expected.

The docking models did not include the transition sequence between the fourth and fifth domains of DM64 (residues 382-GLLP-385) (**Supplementary Figure S4**, underlined residues). Nevertheless, when considering an entirely extended conformation, the expected maximum distance between Cα atoms of residues Glu381 and Lys386 in DM64 is 15.87 Å. This distance corresponded to 14.33 Å in the final docked model, thus compatible with the proposed covalent topology of the molecule.

3.5.3 Third Docking Step: D3-D4D5Myotoxin II

The third domain of DM64 showed the highest density of unique cross-links. Nine out of 15 distance restraints were topologically validated, corresponding to an overall validity rate of 60% (**Supplementary Figure S8**). Five inter-protein cross-links connected D3 and two distinct regions of myotoxin II. Three cross-links (Lys238C-Lys52F, Lys241C-Lys60F, and Lys241C-Lys61F) comprise the toxin's frontal loop and could be topologically validated. On the other hand, both cross-links (Lys241C-Lys105F and Lys241C-Lys106F) connecting the inhibitor and the first and second residues of the myotoxic C-terminal peptide did not fall below the distance limit imposed by the BS^3 cross-linker.

Ten out of the 15 cross-links observed in D3 were classified as intra-protein. Five of them connected D3 to D4, including three validated links (Lys241C-Lys289D, Ser275C-Lys289D, and Lys276C-Lys289D) and two non-validated ones (Lys257C-Lys319D and Lys276C-Lys324D). The remaining five intra-protein cross-links were observed between D3 and D5. Three (Lys191C-Lys409E, Ser275C-Lys409E, and Lys276C-Lys409E) were within an acceptable topological distance, while two cross-links (Lys241C-Lys443E and Ser272C-Lys409E) did not comply with the model structure of the inhibitor. For all violated distances within DM64, measured Euclidean distances between the beta-carbons of cross-linked residues were under 34 Å.

The modeled structures of the inhibitor did not include the sequence between D3 and D4 (residues 286-TLA-288) (**Supplementary Figure S4**, underlined residues). The Cα-Cα distance between Glu285 and Lys289 in the fully extended conformation is 14.5 Å. This measurement in the docking model corresponded to 19.96 Å, thus lightly exceeding the maximum expected distance.

3.5.4 Fourth Docking Step: D2-D3D4D5Myotoxin II

D2 encloses two glycosylation sites (**Table 2**) and showed the lowest number of experimental restrictions. Only three unique cross-links were identified, all of which centered at Lys117 and in good agreement with the maximum bound distances defined by the docked structure of the toxin-antitoxin complex (**Supplementary Figure S9**). A single inter-protein restriction (Lys117B-Lys61F) connected D2 to the frontal loop of myotoxin II. Two other cross-links within DM64 linked D2 to D3 (Lys117B-Ser217C and Lys117B-Lys241C). As expected, considering the presence of two N-linked glycan antennas, both glycosylation sites in the second domain of DM64 (Asn155 and Asn159) were exposed on the complex surface.

The transition sequence between D2 and D3 (residues 188-LLP-190) was not included in the docking models. In the fully extended conformation, the maximum distance between the alpha-carbons of Asp187 and Lys191 was 12.41 Å, whereas, in the final docking model, the two residues were 15.18 Å apart. Despite exceeding the distance limit, both N- and C- terminal regions of D2 and D3 are not structured, allowing certain flexibility regarding the maximum distance limit.

3.5.5 Fifth Docking Step: D1-D2D3D4D5Myotoxin II

Four unique cross-links were identified in the N-terminal domain of DM64, three of which could be validated in the final docking model (75% validation) (**Supplementary Figure S10**). Two valid inter-protein cross-links (Ser84A-Lys7F and Ser84A-Ser20F) connected D1 and the N-terminal region of myotoxin II (residues Lys7 and Ser20, located at the N-terminal helix and the small helix, respectively). Two cross-links were identified within DM64, connecting D1 and D5: the topologically validated Leu1A-Lys443E and the non-validated Ser84A-Lys452E. Finally, the glycosylation site Asn22 (**Table 2**) was exposed at the complex surface in the final docking model, allowing the attachment of a glycan antenna.

The transition region between D1 and D2 (96-PLP-98) was not present in the docking models. Although a maximum distance of 10.67 Å was expected between the alpha-carbons of Glu95 and Ala99, the final docking model showed a larger maximum distance of 23.03 Å.

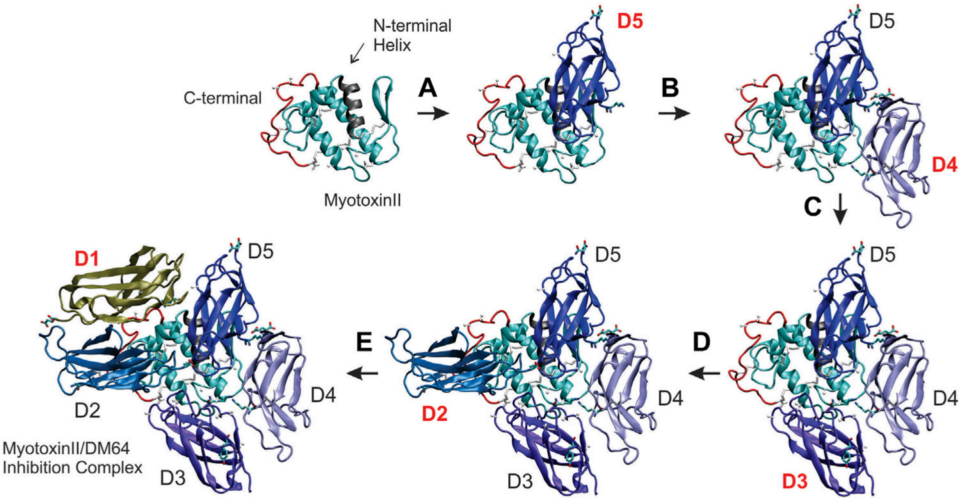


FIGURE 6 | Building the structural model of DM64 docked with the crystal structure of myotoxin II. DM64 Ig-like domains were individually modeled and sequentially docked to myotoxin II (PDB ID 1CLP). Each docking step consisted of a global phase and a local phase when the interaction surface was refined. The first image in the upper left panel of the figure shows the ribbon representation of myotoxin II, with its highlighted N-terminal helix (residues 1 to 14; colored in gray) and the C-terminal region (residues 100 to 121; colored in red). **(A)** In the first docking step, the fifth Ig-like domain of DM64 (D5, colored in dark blue) was docked and interacted mainly with the N-terminal region of myotoxin II. **(B)** D4 (light blue) was docked at the frontal loop of myotoxin II, followed by **(C)** D3 (violet). **(D)** D2 (blue) was docked over the C-terminal region of myotoxin II, together with **(E)** D1 (tan).

3.6 DM64-Myotoxin II Model Construction

Individual docking steps were sequentially shown in **Figure 6** to better visualize DM64-myotoxin II interaction in the same spatial orientation. First, myotoxin II (1CLP, chain A) was represented by a ribbon cartoon, where the N-terminal helix (gray) and the C-terminal region (red) were highlighted to serve as reference points across the five docking steps. All Ig-like domains of DM64 were docked at the opposite side of the helices plane of myotoxin II, following the clockwise direction starting from domain D5 (**Figure 6**, step A).

In the final docking model of DM64-myotoxin II (**Figure 7A**), 26 out of 37 observed cross-linking restraints were within valid topological distances, corresponding to 70% overall validation. Roughly half of the cross-links connected DM64 and myotoxin II (14 validated links/18 observed links, 78% validation), whereas the remaining links were within DM64 (12 validations/19 observations, 63% validation). The main interactions between myotoxin II and D1, D3, and D5 from DM64 are illustrated in **Figures 7B–D**. As discussed below, these appear to be the most dominant regions for the interaction with myotoxin II.

3.7 SAXS Analysis

SAXS data were collected for native DM64 in complex with myotoxin II to complement our structural analysis further. The scattering curve (**Figure 8A**) showed no evidence of aggregation of the sample. The linear Guinier plot (**Figure 8A**, inset) indicated that the system was essentially monodisperse. The distance distribution P(r) profile (**Figure 8B**) showed a non-gaussian shape, with a radius of gyration (Rg) of 3.71 nm and a maximum dimension (Dmax) of 12.02 nm. For comparison, the Rg value calculated for free DM64 was 3.52 nm. The normalized Kratky plot (Glatter and Kratky, 1982) for the complex was compatible with a folded protein presenting regions of flexibility (**Figure 8C**, blue dots). Interestingly, free DM64 showed similar compactness, with increased flexibility (**Figure 8C**, red dots).

3.8 Molecular Dynamics Simulation

Free molecular dynamics was used to assess the stability of the interaction between DM64 and myotoxin II. The PCA1 and PCA2 projections over 600 ns of MD simulation time of DM64-myotoxin II complex are shown in **Supplementary Figure S11**. The first 100 ns (grey dots) were considered an equilibrium period, whereas the black and blue dots represent the remaining 500 ns of MD simulation. From the equilibrium period to 300 ns, there was a transition in the conformational space (indicated by blue dots), followed by a stabilization in the PCA region (black dots farthest to the figure’s right).

To better understand the transition sampled in the PCA analysis, the topological validation of the identified cross-links was analyzed throughout the simulation time (**Figure 9A**). MD simulation started with the coordinates of the docking model of DM64-myotoxin II complex, initially containing 26 topologically validated cross-links. A cyclic fluctuation of validated cross-links was observed during the simulation, reflecting the breathing of the interaction in the time scale, originated by movements of the domains while interacting with myotoxin II and with the vicinal DM64 domains. The MD simulation showed a decrease in the total number of validated cross-links in the first nanoseconds; for 88.5% of the simulation time, a validation rate ranging from 15 to 19 validations was observed. The most populated state corresponded to models showing 17 topological cross-linking validations (**Figure 9A**, red line, and **Figure 9B**). A maximum of 23 validated cross-links was observed in a single model, and 22

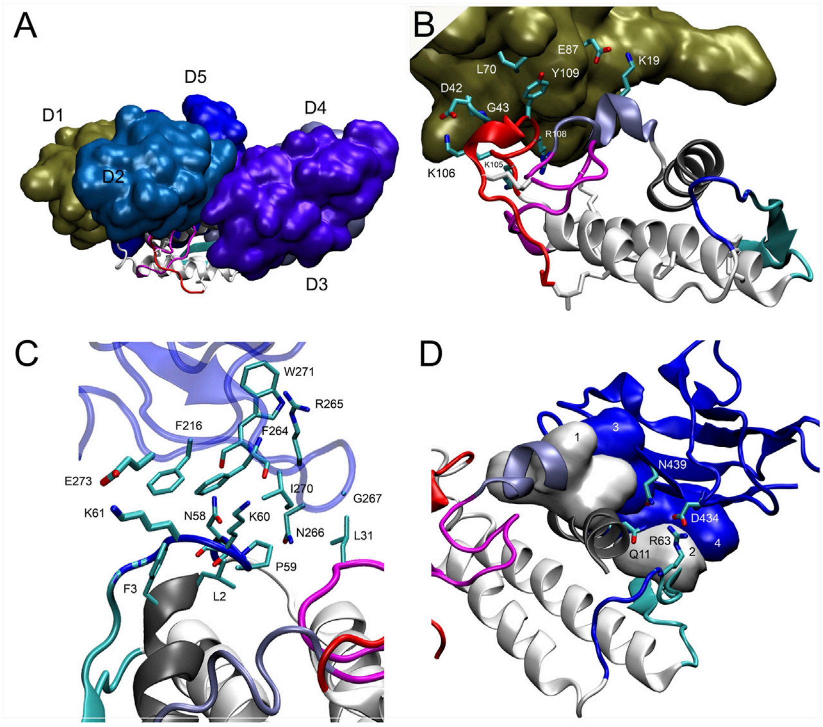


FIGURE 7 | Interaction of myotoxin II and DM64 domains. **(A)** Lateral view of the complex model following the same color scheme of **Figure 6**. DM64 interacts throughout the curved interface of myotoxin II. Despite the complementarity of the surfaces, myotoxin II C-terminal peptide (colored in red) remains free. **(B)** Domain D1 (chain A) interacts with the MDoS residues K19, K105, and R108 of myotoxin II (chain F). The interaction surface encompasses hydrophobic contacts between Tyr109F/Leu70A and two long-lived hydrogen bonds between Lys106F/Asp42A and Lys19F/Glu87A. **(C)** Domain D3 (chain C) interacts with the frontal loop, preceding the beta-wing. The surface encompasses hydrophobic contacts between myotoxin II residues Phe3, Leu2, Leu31, Pro59, and D3 residues Phe216, Phe264, Arg265, Asn266, Gly267, and Ile270. **(D)** Domain D5 (chain E) interacts with the N-terminal helix (gray) and beta-wing (cyan) of myotoxin II (Chain F). Hydrophobic contacts at the myotoxin II interaction surface were represented as numbered white surfaces 1 (Leu10, Gly14 Lys15, and Asn16) and 2 (Trp68). On domain D5, the interaction region comprises blue surfaces 3 (Leu441 and Ile403) and 4 (His390, Val392, and His393). Additionally, several hydrogen bonds showing long-lived nature were detected on the interface, such as Arg63F-Asp434E and Gln11F-Asn439E.

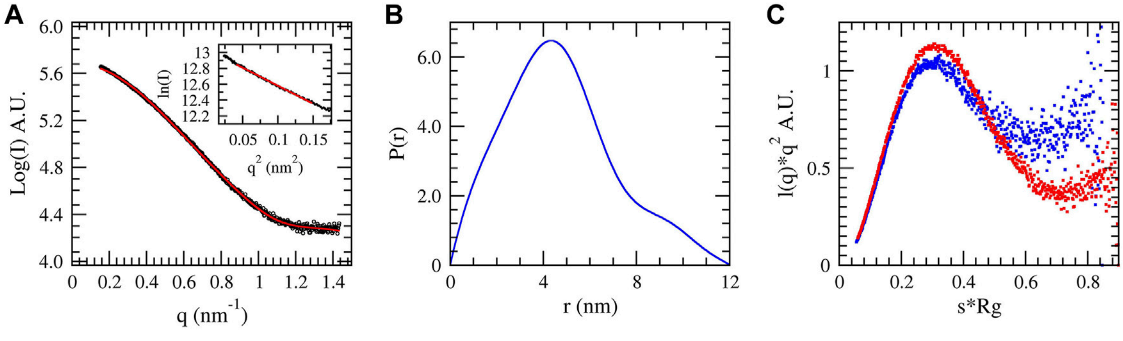


FIGURE 8 | Small-angle X-ray scattering of the complex DM64-myotoxin II in solution. **(A)** Theoretical (red) and experimental (blue) scattering curves of the complex. The Guinier region is shown in the inset. **(B)** Pair distribution function P(r) of the complex. **(C)** Normalized Kratky plot obtained for DM64-myotoxin II complex (blue dots) and free DM64 (red dots).

validations were observed in only 52 models, both cases representing minor states sampled in the MD simulation.

To further investigate the reduction in the number of topologically validated cross-links during MD simulation, validations within each DM64 domain were individually analyzed. D1, D3, and D5 showed a relatively stable number of validations as a function of simulation time (**Figures 10A,E,I**). In the case of D1, of the three cross-links identified, two remained

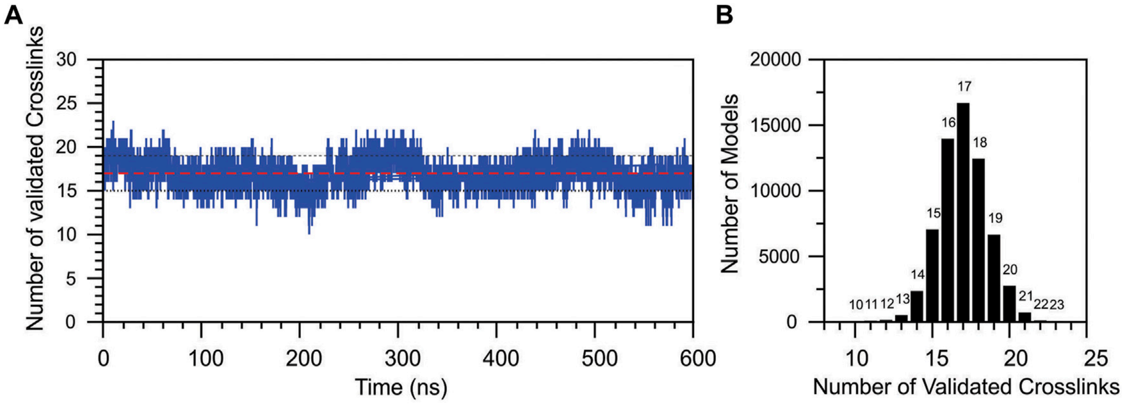


FIGURE 9 | Topological validation of all cross-linking results during the molecular dynamics simulation of DM46-myotoxin II complex. **(A)** The number of validated cross-links was plotted as a function of the MD simulation time. The traced red line represents the average validation, and the black dotted lines represent the average fluctuation. **(B)** The total number of models produced in the MD simulation as a function of the number of validated cross-links.

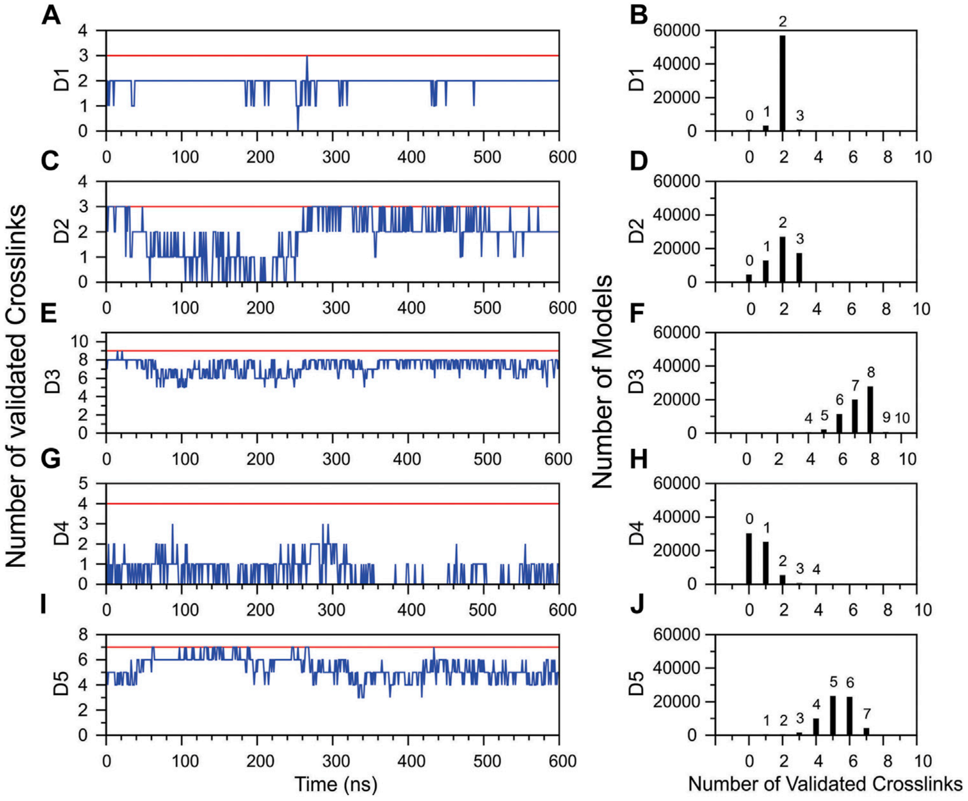


FIGURE 10 | Topological validation of the cross-linking results within each Ig-like domain of DM64 during the molecular dynamics simulation of DM64-myotoxin II complex. On the left **(A,C,E,G,I)**, the blue line represents the average number of validations calculated over 100 frames as a function of simulation time. The red line indicates the maximum number of validated crosslinks identified in the DM64-myotoxin II docking model. On the right **(B,D,F,H,J)**, distribution of the total number of models produced in the MD simulation as a function of the number of validated cross-links.

within topologically validated distances over 90.0% of the simulation time (**Figure 10A**). Almost all the ensemble models showed two validations, while the validation of all three cross-links was sampled for a minor group of 657 models (**Figure 10B**).

In D3, seven and eight cross-links remained valid for 34.4 and 39.9% of the simulation time, respectively. Full validation of the cross-links was observed in very few models: 0.30% of the models showed nine validations, and 0.01% included ten links encompassing valid topological distances (**Figure 10F**). In D5, five validations remained constant for 38.5% of the simulation time, whereas six validations were observed for 35.0% of this interval. Full validation of the cross-links (all seven links) was observed during 6.9% of the simulation time (**Figures 10I,J**).

The glycosylated domains D2 and D4 were mainly responsible for decreasing the total validation rate (**Figures 10D,H**). D2 showed a validation drop in the first 250 ns of simulation, followed by a complete recovery immediately afterward (**Figure 10C**). Interestingly, the PCA analysis of MD simulations showed a transition in this time (**Supplementary Figure S11**, blue dots). On the other hand, D4 suffered a remarkable reduction in the number of validated cross-links at the beginning of the simulation, with no subsequent recovery to initial values. The number of validated links fluctuated between 0 and 1 for about half of the total simulation time (**Figure 10G**).

Hydrogen bond occurrences between individual DM64 domains and myotoxin II were monitored as a function of the simulation time (**Supplementary Table S2** and **Figure 7**). Only the most prevalent hydrogen bonds (i.e., those with the most extended lifetimes) were used to infer stable interactions between protein surfaces. They were mainly observed in D5, where several significantly long-lived hydrogen bonds were detected, strikingly between Asn439 of D5-DM64 and Gln11 of myotoxin II. The lifetime of this hydrogen bond corresponded to 83% of the total simulation time. In D1, two long-lived hydrogen bonds were observed. They encompass myotoxin residues located at the toxic, membrane-destabilizing region (Asp42A-Lys106F) and the small helix (Glu87A-Lys19F). Finally, only short-lived hydrogen bonds were observed for D2, D3, and D4, with lifetimes corresponding to less than 10% of the total simulation time.

4 DISCUSSION

Bothrops genus accounts for most snakebites in Latin America. Severe tissue damage is a major concern in bothropic envenomation, particularly in remote areas of the Amazon region. Even in cases not causing the victim's death, these snakebites can frequently lead to the loss or disablement of limbs, with substantial physical, psychological, and socioeconomic impacts (Gutiérrez et al., 2006; Chippaux, 2017; Fan and Monteiro, 2018). Local toxicity includes edema, blistering, hemorrhage, dermonecrosis, and myonecrosis. Such pathological alterations are induced mainly by metalloendopeptidases and phospholipases A_2 (Gutiérrez et al., 2017), which may exert synergistic toxic effects (Bustillo et al., 2012; Mora-Obando et al., 2014b). These venom proteins display fast-acting toxicity and wide antigenic variability, and their neutralization by conventional antivenom immunotherapy remains challenging (Gutiérrez et al., 1998).

The global burden of snake envenomation has been estimated at approximately six million disability-adjusted life years (DALYs), superior to most neglected tropical diseases (Habib et al., 2015). The development of new therapeutics is one of the urgent needs identified by the World Health Organization to reduce the suffering associated with snakebites (Williams D J et al., 2019). In line with this demand, the growing field of antivenom research has been devoted to studying the problem from various perspectives, including 1) the development of auxiliary diagnostic tests to identify the type/degree of envenomation; 2) the use of selected immunogens for improved antivenom efficacy; 3) the expansion of the treatment toolbox, including alternatives to be used in the field immediately after the bite (Laustsen et al., 2016; Williams H F et al., 2019; Knudsen et al., 2021).

From a therapeutic point of view, we argue that naturally occurring toxin inhibitors are valuable molecules. DM64, the only known myotoxin inhibitor of mammalian origin, binds to a wide variety of basic PLA_2 (Rocha et al., 2017). Unlike the reptile PLI classes described in the literature (Ohkura et al., 1997; Campos et al., 2016), DM64 exerts its antitoxic activity without interfering with the catalytic activity of Asp49-PLA_2 (Rocha et al., 2002). The present study sought to structurally explore the interaction between DM64 and a Lys49-PLA_2 target, aiming to advance the knowledge about the mechanism of action of this naturally-occurring toxin inhibitor.

The primary structure of DM64 comprises 480 amino acid residues, adding up to 53.3 kDa. Its glycan moiety, corresponding to approximately 13 kDa, is responsible for the molecular heterogeneity observed in this molecule. The glycoprotein comprises five Ig-like domains (80–90 amino acid residues, including a central disulfide bond) connected by relatively conserved sequences of 12–13 amino acid residues (**Supplementary Figure S4**). The molecular masses here determined for DM64 by ESI-Q-TOF and AUC were in good agreement with previously determined values by MALDI-TOF MS (63.6 kDa) and SDS-PAGE under reducing conditions (66.5 kDa) (Rocha et al., 2002). It is worth mentioning that AUC results indicated a high anisotropy for DM64 (**Table 1**). Accordingly, protein glycosylation is consistent with higher frictional ratios (f/f_0), resulting from increased hydration near hydrophilic sugar moieties. The presence of glycans could explain the anomalous behavior of DM64 previously observed by size exclusion chromatography (86–110 kDa) (Rocha et al., 2002), which had since then been primarily attributed to the dimerization of DM64 under native solution conditions. However, AUC can measure sedimentation and diffusion transport, providing a more reliable molar mass estimate that does not rely on reference standards or spherical shape assumptions. Therefore, it now seems clear that monomeric DM64 elutes earlier than expected in SEC due to glycosylation (León et al., 2012), leading to overestimating molar mass (Rocha et al., 2002). As already reported for DM64, an elevated anisotropy was also observed by AUC for the complex DM64-myotoxin II (**Table 1**).

The toxin used as a DM64 target in this study was myotoxin II from *Bothrops asper* venom, a well-known Lys49-PLA_2 whose schematic structure is shown in **Supplementary Figure S5** (Lomonte and Gutiérrez, 1989). Calcium-binding is prevented in myotoxin II due to a few amino acid modifications, mainly the critical substitution of Asp by Lys at position 49. Seven disulfide bonds stabilize myotoxin II structure, and substantial fluctuation in secondary or tertiary structure, even at the N-terminal and

C-terminal regions, is not expected. The myotoxic activity of myotoxin II is mainly associated with its C-terminal sequence (105-KKYRYYLKPLCKK-117), as shown by studies with synthetic peptides. A combination of lysine and tyrosine residues forms a highly exposed cationic/hydrophobic region in the protein's three-dimensional structure. The amphipathic character of this region is critical to the biological activity, and a triple tyrosine to tryptophan substitution in the C-terminal synthetic peptide of myotoxin II drastically enhanced its myotoxicity [reviewed in (Lomonte et al., 2003; Lomonte and Rangel, 2012)].

In myotoxin II's apo form (PDB ID 1CLP), the N-terminal helix and the loop preceding the first β-strand of the beta-wing are implicated in the dimerization interface. This conventional dimer is stabilized by intermolecular hydrogen bonds involving Glu12, Trp68, and Lys71 [numbered Glu12, Trp77, and Lys80 in (Arni and Ward, 1995)]. On the other hand, the dimer adopted a novel conformation when myotoxin II was bound to the polyanionic compound suramin (Murakami et al., 2005). In this alternative dimer (PDB ID 1Y4L), the dimerization interface is formed by the hydrophobic surfaces surrounding the putative active site entry. Both oligomeric conformations are possible biological dimers, although bioinformatics analyses indicate that the alternative conformation is more stable in solution (Dos Santos et al., 2009; Fernandes et al., 2013).

A comprehensive toxic mechanism involving the alternative conformation of Lys49 myotoxins has been proposed based on structural data (Fernandes et al., 2013; Fernandes et al., 2014). Hydrophobic ligands (e.g., fatty acids binding in the hydrophobic channel of the toxin) would trigger a conformational shift that "activates" the dimer. The stabilization of this active conformation depends on interactions between residues in the putative Ca^{2+} binding loop and the C-terminal region. In this final conformation, the "membrane docking site" (MDoS, mainly formed by the cationic residues in the C-terminal region) and a cluster of critical nearby hydrophobic residues (MDiS, standing for "membrane disruption site") in both monomers are perfectly positioned. The optimum interaction of the toxin MDoS with target anionic sites is followed by bilayer disruption through its MDiS. Consequently, there is a loss of control of ion influx across the membrane, irreversibly compromising the integrity of the muscle fiber (Gutiérrez and Ownby, 2003; Lomonte and Rangel, 2012). In myotoxin II from *Bothrops asper* venom, Lys19, Lys105, and Arg108 would form the putative MDoS. In contrast, MDiS would be formed by Leu111 and Leu114 [respectively residues 19, 115, 118, 121 and 124 in (Fernandes et al., 2014)], although a definitive identification to pinpoint the functionally critical residues of this toxin awaits experimental confirmation.

Because it is extremely sensitive (order of femtomoles) and relatively more tolerant to sample heterogeneity, mass spectrometry is gaining popularity in structural biology (Borch et al., 2005; Sinz, 2014). XL-MS involves stabilizing proteins/protein complexes with cross-linking agents, which covalently connect the side chains of specific residues. For the reaction to occur, these residues must be close enough in space for a sufficiently long time (Leitner et al., 2010). Cross-linking agents can vary according to the characteristic of the reactive groups (homo-/hetero-bifunctional), the reaction specificity (basic/acidic side chains or nonspecific/photoreactive), and the spacer chain length (0–20 Å). Additional options are also available in the cross-linker toolbox, such as cleavable, trifunctional, or isotopically labeled reagents (Sinz, 2003; Sinz, 2006).

After the cross-reaction, the proteins/protein complexes are enzymatically digested, and the cross-linked residues can be identified by high-resolution mass spectrometry. In this way, information about the maximum allowable distance between residue pairs is generated that increases structure prediction accuracy when used with computational modeling/docking tools (Sinz, 2014). This novel structural approach can provide valuable information regarding the spatial orientation of proteins and their connectivity. XL-MS results are often analyzed with data generated by complementary low-resolution structural methodologies (e.g., HDX-MS and SAXS) (Sali et al., 2015). This hybrid approach, known as integrative structural biology, has subsidized the formulation of consistent hypotheses connecting structure to biological function (Yu and Huang, 2018). Recently, XL-MS, HDX-MS, molecular modeling, protein-protein docking, and molecular dynamics simulations were successfully employed in the structural characterization of the first toxin-antitoxin complex. The study focused on the interaction between the endogenous circulating inhibitor BJ46a and the snake venom metalloendopeptidase jararhagin, both isolated from the South American snake *Bothrops jararaca* (Bastos et al., 2020).

A similar strategy was used here to advance our understanding of the molecular basis underlying the neutralizing activity of DM64 against a Lys49-PLA_2 myotoxin. The use of a traditional lysine-reactive reagent allowed for the generation of valuable through-space distance information. Initially, free DM64 was modeled based on its whole amino acid sequence, and the homology model resulted in a structure enclosing the fundamental aspects of this protein class. Each Ig-like domain was structured in two β-sheets, composed of three and four β-strands, connected centrally by a disulfide bond. The loops connecting the β-strands in each sheet are the regions implicated in the interaction of the Ig-like domains with their molecular partners (Bork et al., 1994). The glycosylation sites were modeled exposed on the protein surface, allowing the addition of the glycan antennas. When cross-linking restraints were used to validate the final candidate model, distance violation rates indicated that the homology modeling succeeded well at the domain level (61.5% of intra-domain validation) (**Supplementary Table S1B**).

On the other hand, only 3.8% of the inter-domain cross-links were within the cross-linker maximum bound. In several cases, inter-domain cross-links spanning topological distances greater than 100 Å were measured (**Supplementary Table S1B**), clearly indicating that the global fold of the model did not correspond to the conformations experimentally sampled in the cross-linking experiment. The absence of suitable templates (i.e., encompassing consecutive Ig-like domains with high sequence identity) in the PDB data bank can explain the weak modeling performance. Even the recently published AlphaFold2 algorithm (Jumper et al.,

2021) could not significantly improve the quality of the models compared to I-TASSER (not shown).

An alternative approach was then attempted to advance the structural understanding of the inhibitor and its interaction with myotoxin II. Distance constraints determined by cross-linking mass spectrometry were used to guide molecular modeling/docking strategies on the heterocomplex. The model was built based on observed cross-links connecting: 1) different domains of DM64; 2) DM64 domains and myotoxin II.

Structural determination methods are always influenced by time and conformational averaging effects (Markwick et al., 2008; Sun et al., 2019; Kulik et al., 2021). Such limitation also applies to structural methods using chemical cross-linking and high-resolution mass spectrometry. The identification of cross-linked peptides results from a chemical reaction with a diverse population of proteins in the solution, resulting in a dataset containing sub ensembles of cross-links that will never be validated simultaneously in a unique structural model. In docking analysis, we are chasing states of full validation of the experimental data. However, depending on the biological model, this scenario can never be reached. Therefore, free MD simulations were performed to sample the Ig-like domains' dynamic behavior in the complex's docked model. Given the low target–template sequence identity, the level of confidence of our model for the inhibitor is limited (if one considers the finer details of fold and surface features). Despite this, molecular dynamics simulation contributed to refining the results generated by the docking strategy, increasing our confidence in the final proposed model. Information on hydrogen bonding profiling was used with caution, being restricted to the most striking observations.

As the DM64-myotoxin II model was built after a series of docking steps, the choice of the initial docking partners was critical. D1 and D2 are glycosylated and showed low sequence coverages by mass spectrometry and the lowest number of identified cross-links. Most inter-domain cross-links were concentrated in D3 and D5. Interestingly, these domains were the source of the myotoxin-interacting peptides identified in the limited hydrolysis/affinity experiment. Although D3 showed the highest number of identified cross-links, D5 was most abundant in links connecting the inhibitor to the toxin, which was used as a reference point for the docking procedure. Hence, our first docking step focused on the interaction between D5 and myotoxin II.

D5 was docked to myotoxin II in the region of the N-terminal helix, the frontal loop, and the beta-wing (**Supplementary Figure S6**). Such interaction prediction encompassed the highest number of validated cross-links (77.8%) among all docked domains, indicating high interaction stability of this domain. An ensemble of anticorrelated cross-links was observed during the procedure: cross-links with N-terminal residues 1 and 15 in myotoxin II were never validated simultaneously with the remaining cross-links (**Supplementary Table S1A**). This result may indicate a second possibility for D5 interaction, leading to an alternative pathway for building the heterocomplex model structure. The alternative conformation represented a minor subset among 150,000 docking runs, and all attempts to proceed with docking the D4 domain resulted in low validation of inter-domain (D4-D5) cross-links.

The interaction between D5 and myotoxin II was analyzed in the MD simulation of DM64 and myotoxin II complex. Of seven validated cross-links in the initial docking model, between 5 and 6 links remained valid for 73.4% of the simulation time, indicating a highly stable interaction (**Figures 10I,J**). The partial loss of validated cross-links was expected and was attributed to structural relaxation during the model preparation phase (minimization, equilibration, and stabilization). D5-myotoxin II interaction surface also showed a higher number of long-lived hydrogen bonds, and they were verified in the N-terminal helix, the frontal loop, and the beta-wing (**Supplementary Table S2E**).

Residues involved in the most long-lived hydrogen bonds in DM64 (>40% of the total simulation time) were His405, Asn439, and Asp434. The latter is located in a loop (Val430-His436) connecting two β-strands, a region of Ig-like domains predicted to interact with molecular partners (Bork et al., 1994). The limited hydrolysis/affinity experiment identified two peptides from D5 binding to immobilized myotoxin II. The first one (residues 420–443) faces myotoxin II in the docked model of D5-myotoxin II (**Supplementary Figure S6B**, white surface). D5 seems critical for the interaction of the inhibitor with the toxin and the consequent neutralization of its toxic activity. It contributes to blocking the dimerization interface of the myotoxin (assuming a conventional dimeric conformation), thus preventing the oligomeric conformation that seems to be relevant for toxicity. Once bound to D5, myotoxin II would lose the ability to interact with the muscle membrane.

The second DM64 peptide fished by the myotoxin affinity column (residues 453–480) constitutes the C-terminal of the inhibitor. The docked model of DM64-myotoxin II did not show any contact region between this peptide and the toxin. Accordingly, no interaction could be observed in the MD, even when the simulation time was extended to 1 μs (not shown). The binding of the acidic 28 residues long peptide to the column may indicate that the docking model needs further refinement, although it can explain most of the experimental data obtained thus far. Another possibility involves alternative conformational states of the DM64-myotoxin II complex in solution, compatible with the observed interaction result. The remarkable sensitivity of mass spectrometry allows a most compelling exploration of the conformational space under analysis. This pattern is commonly observed in XL-MS experiments, where ambiguous data preclude all results' simultaneous validation. Finally, we cannot rule out the possibility of nonspecific interaction between this C-terminal peptide and the column.

D4 was the first glycosylated (Asn355) domain added to the complex D5-myotoxin II (**Supplementary Figure S7**). The final docking model showed four validated cross-links (**Supplementary Table S1A**). However, the validation rate dropped early in the MD simulation and remained between zero and one valid link 50.0% of the time (**Figures 10G,H**). This behavior agrees with the hydrogen bond analysis, where long-lived hydrogen bonds could not be detected

(**Supplementary Table S2D**). This scenario can be extrapolated to a dynamic interaction over time, including quick contact and detachment cycles from the myotoxin II surface.

Residues in domain D3 were engaged in 15 cross-links with D4, D5, and myotoxin II, nine of which could be validated in the final docking model (**Supplementary Table S1A**). Validated inter-protein cross-links indicate that D3 interacts with the toxin's frontal loop (**Supplementary Figure S8**). MD simulation started with eight validated cross-links, and this number fluctuated between 6 and 8 for 95.0% of the simulation time, indicating a stable interaction (**Figures 10E,F**). Although only short-lived hydrogen bonds were observed in D3 (**Supplementary Table S2C**), they mainly involved DM64 residues Arg265 and Asn266 located in the D3 peptide 258–276, which was bound to myotoxin II in the limited hydrolysis/affinity experiment. A closer analysis of the interaction surface between D3-DM64 and myotoxin II also shows an important contribution of hydrophobic interactions.

We hypothesize that the interaction between the DM64 domains assumes acute angles, creating a structure in zigzag. In this conformation, while D5 interacts on the myotoxin II surface, D4 is pushed out in the direction of the solvent. Sequentially, another acute angle bend turns possible the return of D3 to the myotoxin II surface, probably allowing a higher interaction area with this domain. A similar structure was also described for murine paired immunoglobulin receptor B (PirB), leukocyte immunoglobulin-like receptor (LILR), and killer-cell immunoglobulin-like receptor (KIR) (Vlieg et al., 2019).

Alternative conformations upon the interaction of the third domain of DM64 may explain the more considerable discrepancy observed in the length of the transition region connecting the D3 and D4 domains (19.96 Å measured distance x 14.50 Å maximum expected distance). This result could also be derived from local conformational changes of DM64 upon binding, aiming to maximize the surface buried in the interaction, which would not be considered in molecular modeling. This possibility is supported by the consistent non-validation of cross-links between D3 and D4/D5 and the two non-validated crosslinks correlating D3 to myotoxin II C-terminal residues Lys105F and Lys106F (**Supplementary Table S1A**).

D2 has two glycosylation sites (Asn155 and Asn159) and established only one valid cross-link with the Lys61 residue of myotoxin II. Two additional valid links were observed with D3 (**Supplementary Figure S9** and **Supplementary Table S1A**). A drop in the number of validated cross-links was observed after 60 ns and remained for 250 ns (**Figures 10C,D**). First, there was a shift from two to one validation and then from one to zero. After this period, the validation increased again and stabilized between 2 and 3 during the 300 ns of MD simulation. Domain D2 did not show hydrogen bonds with myotoxin II living in the "on" state longer than 0.5% of the MD simulation time (**Supplementary Table S2B**).

Finally, D1 was added to the docking model of D2D3D4D5-myotoxinII (**Supplementary Figure S10**). The experimental data acquired for this domain was limited, probably due to the high glycosylation content enclosed in the N-terminal domains D2 and D1. The final docking model shows three validated cross-links in D1, two connecting the inhibitor to Lys7 and Ser20 of myotoxin II (**Supplementary Table S1A**). The distance between domains D1 and D2 was about twice the expected maximum (23.03 × 10.67 Å), indicating that the model topology in this region is not sufficiently accurate. As MD simulation started, only two validated cross-links remained for 90.0% of the simulation time (**Figures 10A,B**).

Interestingly, the C-terminal peptide of the toxin (residues 99–121) was identified in the hydrolysate fraction bound to the DM64 affinity column. The docked model showed a significant interaction surface between this toxin region and the first two domains of DM64. In D1, long-lived hydrogen bonds were observed between residue Asp42 of the inhibitor and residue Lys106 of the toxin (**Supplementary Table S2A**). Despite the scarcity of experimental data on the first domain of DM64, the docking model, the MD simulation, and the hydrolysis experiment data support the hypothesis that D1 shows a relatively stable interaction with myotoxin-II. This result indicates that D1 likely blocks the MDoS site, impairing the anchoring of the toxin to the membrane.

Homology modeling of free DM64 always pointed to more elongated, U- or L-shaped structures. This shape is consistent with the non-gaussian distribution of pairwise distances derived from SAXS data and the high anisotropy observed by AUC. MD simulations showed that the connecting regions (10–12 residues) between the five Ig-like domains of DM64 are likely flexible. We hypothesize that the domains bend over myotoxin II, increasing the interaction surface area and the stability of the interaction. Accordingly, the normalized Kratky plot indicated a structured protein with pronounced flexibility when the inhibitor was free. After the binding, the complex showed nearly the same level of compactness, with a significant decrease in flexibility. The SAXS-derived Rg values for the complex (3.71 nm) and free DM64 (3.53 nm) were similar and correlated well with the hydrodynamic radii calculated by AUC (4.22 nm for the complex and 3.96 nm for DM64).

Crysol was used to quantify the goodness of fit between the experimental scattering data and those calculated from the MD simulation conformations. The correlation was weak when the final docking model was used as reference (Chi2 37.98, Rg = 2.84 nm). After 600 ns of MD simulation, the last coordinate showed a more relaxed protein structure, reducing the Chi2 value to 19.64 (Rg = 3.04 nm). Throughout the simulation time, the five most populated clusters showed Chi2 values ranging from 28.6 to 11.03, and minor represented conformations even reached 9.32. The best correlation was obtained by extending the simulation time to 1 μs (Chi2 = 4.40, Rg = 3.20 nm) (data not shown). This last structural model showed 16 validated cross-links, close to the mean validation rate observed in the MD analysis (17 validations, **Figure 9**). These results indicate that the MD cluster analysis can further improve the quality of the DM64-myotoxin II model. Although not ideal yet, the model is reasonably consistent with the experimental data and has supported valuable structural and functional predictions.

From a mechanistic point of view, DM64 inhibits myotoxin II from *Bothrops asper* venom in two presumed ways: 1) The fifth and the third domains interact with the toxin's (conventional) dimerization surface. The binding precludes its quaternary structure assembly, which is relevant for myotoxicity; 2) The first domain directly interacts with the MDoS region, preventing the toxin's anchoring to the muscle cell. Crystallographic studies with several small inhibitory ligands of PLA_2-like toxins (e.g., varespladib, suramin) indicated the hydrophobic channel and the MDoS and MDis regions of the toxin as preferential functional targets (Salvador et al., 2019). Based on our topological data, it is difficult to speculate on the obstruction of the toxin hydrophobic channel by DM64. Previously, we have shown that DM64 inhibits the myotoxic activity of myotoxin I from *B. asper* yet cannot interfere with its catalytic activity (Rocha et al., 2002). This basic Asp49-PLA_2 toxin and myotoxin II from *B. asper* show 61% sequence identity and conserved three-dimensional structures (α chains TM-score = 0.909) (Salvador et al., 2017). Thus, it seems reasonable to assume that the hydrophobic channel of myotoxin II likely does not participate in the mechanism of inhibition by DM64.

DM64 antitoxic protein scaffold has been naturally shaped by extensive trial and error experiments (Voss and Jansa, 2012). Therefore, unveiling the structural determinants of myotoxicity inhibition by DM64 can contribute insights into the rational design of therapeutic alternatives to treat snakebites. Notably, developing new peptide-based drugs as the first line of defense against venom-induced tissue damage is an attractive possibility (Erak et al., 2018). Compared to small molecules (<500 Da), peptide-based drugs are easier to produce and show higher potency and selectivity. Several alternatives to overcome possible limitations related to half-life, stability, solubility, and bioavailability are continuously being developed (Craik et al., 2013; Fosgerau and Hoffmann, 2015; Lau and Dunn, 2018). At least 60 peptide drugs are currently approved by the U.S. Food and Drug Administration (FDA) agency, and more than 600 are in (pre)clinical testing phases (Erak et al., 2018). The field of next-generation antivenom should not fail to explore innovative possibilities in such a blooming area.

5 CONCLUSION

Snake venoms are complex biological mixtures that induce a diversity of pathological effects. Yet, it seems possible to counteract their toxicity by inhibiting only a critical subset of toxins. For bothropic venoms, myotoxic PLA_2s are priority targets. Using integrative structural biology, we have derived topological information on the complex made of a PLA_2-like myotoxin and DM64. The latter is an endogenous protein that cross-neutralizes several homologous myotoxins. This study provides crucial insights towards understanding critical features of the inhibitor's structure-function relationship.

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: http://www.proteomexchange.org/, PXD028522.

AUTHOR CONTRIBUTIONS

AN-F, FG-N, JP, and RV contributed to the conception and design of the study. AH, BS, JA, RV, SR, TS, TW, and TJ performed the experiments. BL, DL, PC, and FG contributed new reagents or analytic tools. AN-F, AH, BD, BL, BS, FG-N, JA, RV, TJ, TS, TW, and VB analyzed data. AN-F and FG-N wrote the first draft of the paper. BD, BL, RV, and TS wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was supported by Brazilian grants from Fiocruz (INOVA GC VPPCB-007-FIO-18-2-9), FAPERJ (APQ1 E-26/010.001929/2019), and CNPq (Universal 426290/2018-6). The study was also funded by the NIH grant GM120600 and NSF grant NSF-ACI-1339649 (to BD). Supercomputer calculations were performed on Comet at the San Diego Supercomputing Center (support through NSF/XSEDE grant TG-MCB070039N to BD) and Lonestar-5 at the Texas Advanced Computing Center (supported through UT grant TG457201 to BD). VB is supported by an early-career researcher small grant from the Royal Society of Tropical Medicine and Hygiene (RSTMH, England, United Kingdom) in partnership with the Wellcome Trust.

ACKNOWLEDGMENTS

SAXS experiments were performed at the Brazilian Synchrotron Light Laboratory (LNLS), an open national facility operated by the Brazilian Centre for Research in Energy and Materials (CNPEM) for the Brazilian Ministry for Science, Technology, Innovations, and Communications (MCTIC). The SAX02 beamline staff is acknowledged for their assistance during the experiments. We acknowledge the support of the San Antonio Cancer Institute grant P30 CA054174 for support of the Center for Analytical Ultracentrifugation of Macromolecular Assemblies at the University of Texas Health Science Center at San Antonio. We are grateful to Heloisa M. N. Diniz (Image Production and Treatment Service, IOC, Fiocruz) for helping with image processing. Finally, the authors thank the Fiocruz Network of Technology Platforms for the core facilities for proteomics and mass spectrometry (RPT02A, Rio de Janeiro). RV is a fellow from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant number 304523/2019-4.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.787368/full#supplementary-material

Supplementary Table S1 | Cross-links identified by the software SIM-XL after applying the BS^3 strategy on DM64 and myotoxin II complex. The tables list all high-confidence cross-links and the analysis of Euclidean and topological distances (Cβ-Cβ atoms) calculated with the software TopoLink.

Supplementary Table S2 | Free MD hydrogen bond lifetimes in the interaction surfaces between each Ig-like domain of DM64 and myotoxin II.

Supplementary Figure S1 | Analysis of complex formation between DM64 cross-linked with BS^3 and native myotoxin II. The binding property of cross-linked DM64 toward native myotoxin II was monitored on a non-denaturing 12% T gel stained with Coomassie blue R250. Lane 1: DM64 (native control); lane 2: DM64 cross-linked with BS^3 (1:2,800 mol/mol); lane 3: DM64 + BS^3 (1:2,000 mol/mol); lane 4: DM64 + BS^3 (1:1,000 mol/mol); lane 5: DM64 (native control) + mtx-II; lane 6: [DM64 + BS^3 (1:2,800 mol/mol)] + mtx-II; lane 7: [DM64 + BS^3 (1:2,000 mol/mol)] + mtx-II; lane 8: [DM64 + BS^3 (1:1,000 mol/mol)] + mtx-II; lane 9: mtx-II.

Supplementary Figure S2 | Stabilization of the complex made by the antitoxin DM64 and myotoxin II using different concentrations of BS^3. **(A)**: Electrophoresis under non-denaturing conditions; **(B)**: Electrophoresis in the presence of SDS and DTT. All samples were run on 12% T gels stained with silver nitrate. MM: molecular mass markers; lane 1: DM64 (native control); lane 2: DM64 cross-linked with BS^3 (1: 2,800 mol/mol); lane 3: mtx-II (native control); lane 4: mtx-II + BS^3 (1:2,800 mol/mol); lane 5: (DM64 + mtx-II) + BS^3 (1:2,800 mol/mol); lane 6: (DM64 + mtx-II) + BS^3 (1: 2,000 mol/mol); lane 7: (DM64 + mtx-II) + BS^3 (1:1,000 mol/mol); lane 8: (DM64 + mtx-II) + BS^3 (1:500 mol/mol).

Supplementary Figure S3 | Myotoxin II cross-linked with BS^3. Ribbon diagram of myotoxin II (1CLP, chain A) showing the topologically validated cross-linking connections colored in green, the N-terminal α-helix (residues 1 to 15, colored in gray), the analogous region for calcium coordination site in catalytically active enzymes (residues 28 to 33, colored in magenta) and the myotoxic peptide (residues 100 to 121, colored in red). The active site residues His47, Tyr51, and Asp89, were represented in sticks. The cross-linking reaction was performed on myotoxin II in complex with DM64.

Supplementary Figure S4 | Primary structure of the inhibitor DM64. A sequence of five amino acid residues was highlighted at each Ig-like domain's start and end (yellow boxes). Four conserved regions formed by 12-13 amino acid residues (cyan boxes) connected sequential Ig-like domains. Except for the underlined sequences, the connection regions were included in the modeling of the respective domain, as detailed in **Table 2**.

Supplementary Figure S5 | Ribbon representation of the crystal structure of myotoxin II (1CLP, chain A) with the main structural elements highlighted in color. Myotoxin II structure is built over a helix plane formed by two long antiparallel α-helices (white). The N-terminal is structured in an α-helix (residues 1-14, gray) bent to create a small helix (residues 16 to 24, blue). The sequence analogous to the calcium coordination loop in catalytically active phospholipases A_2 was colored magenta (residues 25 to 39). Sequentially, residues from 40 to 57 form the first long α-helix (white). The region from residues 58 to 73 corresponds to the frontal loop (blue) that connects to a short antiparallel β-sheet (residues 63-86, cyan) named beta-wing, followed by the second long α-helix (89-108, white). Finally, the myotoxic C-terminal peptide extends from residue 109 to 133 (red). The numbers in parenthesis indicate the following disulfide bonds (yellow): (1) Cys50-Cys88, (2) Cys43-Cys95, (3) Cys57-Cys81, (4) Cys28-Cys44, (5) Cys49-Cys121, (6) Cys26-Cys115, and (7) Cys75-Cys86.

Supplementary Figure S6 | Ribbon representation of the fifth Ig-like domain of DM64 docked with the crystal structure of myotoxin II. **(A)** D5 (blue) interacting with myotoxin II (cyan), with the C-terminal myotoxic peptide (residues 100 to 120) colored in red and the N-terminal helix (residues 1 to 14) highlighted in grey. All five validated cross-links involving Lys443 in D5 and myotoxin II are indicated with dotted green lines (topological distances). The relative position of both proteins in the complex was fixed by cross-link Ser407E-Lys69F, which restricted the movements of the interaction surface. **(B)** The same as in panel A, with the peptide 420-443 from DM64 interacting with myotoxin II, as shown by the limited hydrolysis/affinity experiment, represented in white surface.

Supplementary Figure S7 | Ribbon representation of the fourth Ig-like domain of DM64 docked with D5Myotoxin II. D4 (light blue) interacts with D5 (dark blue) and myotoxin II (cyan), with the C-terminal myotoxic peptide (residues 100 to 120) colored in red and the N-terminal helix (residues 1 to 14) highlighted in grey. Myotoxin residue Lys61 lies in front of the N-terminal helix and is involved in a cross-link with Lys289D, stabilizing D4 and myotoxin II interaction. All validated cross-links following this second docking step are indicated with dotted green lines (topological distances).

Supplementary Figure S8 | Ribbon representation of the third Ig-like domain of DM64 docked with D4D5Myotoxin II. DM64 domain D3 (violet) interacting with D4 (light blue), D5 (dark blue), and myotoxin II (cyan). Residues Lys52, Lys60, and Lys 61 are in the loop connecting the two helices of the myotoxin II structure.

Supplementary Figure S9 | Docking the second Ig-like domain of DM64 with D3D4D5Myotoxin II. Residue Lys117 in D2 (navy blue cartoon) was cross-linked to residues Lys217 and Lys241in D3 (violet cartoon), and myotoxin II (cyan cartoon), through residue Lys61, located in the loop connecting the two long helices. Domains D5 (dark blue) and D4 (light blue) were represented as surfaces for clarity.

Supplementary Figure S10 | Docking the first Ig-like domain of DM64 with D2D4D5Myotoxin II. D1 (tan cartoon) is interacting with D2 (navy blue surface), D3 (violet surface), D4 (light blue surface), D5 (dark blue surface), and myotoxin II (cyan cartoon). Residue Ser84 connects Myo-II residues Lys7 and Ser20 located at the N-terminal helix. Leu1 residue at D1 N-terminal connects residue Ly443 at the C-terminal portion of D5. In addition, the glycosylation site Asn22 is exposed on the D1 surface.

Supplementary Figure S11 | Principal Component Analysis (PCA) of molecular dynamics simulations. The equilibrium period (0–100 ns) was represented by gray dots, whereas black dots represent the production phase. A shift was observed in the time window up to 300 ns (blue dots).

REFERENCES

Arbuckle, K., Rodríguez de la Vega, R. C., and Casewell, N. R. (2017). Coevolution Takes the Sting Out of it: Evolutionary Biology and Mechanisms of Toxin Resistance in Animals. *Toxicon* 140, 118–131. doi:10.1016/j.toxicon.2017.10.026

Arni, R. K., Ward, R. J., Gutierrez, J. M., and Tulinsky, A. (1995). Structure of a Calcium-independent Phospholipase-like Myotoxic Protein fromBothrops Aspervenom. *Acta Cryst. D* 51, 311–317. doi:10.1107/s0907444994011455

Bastos, V. A., Gomes-Neto, F., Perales, J., Neves-Ferreira, A. G., and Valente, R. H. (2016). Natural Inhibitors of Snake Venom Metalloendopeptidases: History and Current Challenges. *Toxins (Basel)* 8, 250. doi:10.3390/toxins8090250

Bastos, V. A., Gomes-Neto, F., Rocha, S. L. G., Teixeira-Ferreira, A., Perales, J., Neves-Ferreira, A. G. C., et al. (2020). The Interaction between the Natural Metalloendopeptidase Inhibitor BJ46a and its Target Toxin Jararhagin Analyzed by Structural Mass Spectrometry and Molecular Modeling. *J. Proteomics* 221, 103761. doi:10.1016/j.jprot.2020.103761

Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995). GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.* 91, 43–56. doi:10.1016/0010-4655(95)00042-e

Borch, J., Jørgensen, T. J., and Roepstorff, P. (2005). Mass Spectrometric Analysis of Protein Interactions. *Curr. Opin. Chem. Biol.* 9, 509–516. doi:10.1016/j.cbpa.2005.08.013

Bork, P., Holm, L., and Sander, C. (1994). The Immunoglobulin Fold Structural Classification, Sequence Patterns and Common Core. *J. Mol. Biol.* 242, 309–320. doi:10.1016/s0022-2836(84)71582-8

Brand, G. D., Salbo, R., Jørgensen, T. J. D., Bloch, C., Jr., Erba, E. B., Robinson, C. V., et al. (2012). The Interaction of the Antitoxin DM43 with a Snake Venom Metalloproteinase Analyzed by Mass Spectrometry and Surface Plasmon Resonance. *J. Mass. Spectrom.* 47, 567–573. doi:10.1002/jms.2990

Brookes, E., Cao, W., and Demeler, B. (2010). A Two-Dimensional Spectrum Analysis for Sedimentation Velocity Experiments of Mixtures with Heterogeneity in Molecular Weight and Shape. *Eur. Biophys. J.* 39, 405–414. doi:10.1007/s00249-009-0413-5

Brookes, E., and Demeler, B. (2008). Parallel Computational Techniques for the Analysis of Sedimentation Velocity Experiments in UltraScan. *Colloid Polym. Sci.* 286, 138–148. doi:10.1007/s00396-007-1714-9

Brookes, E., and Demeler, B. (2007). Parsimonious Regularization Using Genetic Algorithms Applied to the Analysis of Analytical Ultracentrifugation Experiments. in GECCO '07 Proceedings of the 9th annual conference on

Genetic and evolutionary computation. July 2007. London, England (New York: ACM), 361–368.

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* 126, 014101. doi:10.1063/1.2408420

Bustillo, S., Gay, C. C., García Denegri, M. E., Ponce-Soto, L. A., Joffé, E. B. D. K., Acosta, O., et al. (2012). Synergism between Baltergin Metalloproteinase and Ba SPII RP4 PLA2 from Bothrops Alternatus Venom on Skeletal Muscle (C2C12) Cells. *Toxicon* 59, 338–343. doi:10.1016/j.toxicon.2011.11.007

Camacho, E., Sanz, L., Escalante, T., Pérez, A., Villalta, F., Lomonte, B., et al. (2016). Novel Catalytically-Inactive PII Metalloproteinases from a Viperid Snake Venom with Substitutions in the Canonical Zinc-Binding Motif. *Toxins (Basel)* 8, 292. doi:10.3390/toxins8100292

Campos, P. C., De Melo, L. A., Dias, G. L. F., and Fortes-Dias, C. L. (2016). Endogenous Phospholipase A2 Inhibitors in Snakes: a Brief Overview. *J. Venom Anim. Toxins Incl Trop. Dis.* 22, 37. doi:10.1186/s40409-016-0092-5

Cao, W., and Demeler, B. (2008). Modeling Analytical Ultracentrifugation Experiments with an Adaptive Space-Time Finite Element Solution for Multicomponent Reacting Systems. *Biophysical J.* 95, 54–65. doi:10.1529/biophysj.107.123950

Carvalho, P. C., Lima, D. B., Leprevost, F. V., Santos, M. D. M., Fischer, J. S. G., Aquino, P. F., et al. (2016). Integrated Analysis of Shotgun Proteomic Data with PatternLab for Proteomics 4.0. *Nat. Protoc.* 11, 102–117. doi:10.1038/nprot.2015.133

Chippaux, J.-P. (2017). Incidence and Mortality Due to Snakebite in the Americas. *Plos Negl. Trop. Dis.* 11, e0005662. doi:10.1371/journal.pntd.0005662

Clerc, F., Reiding, K. R., Jansen, B. C., Kammeijer, G. S. M., Bondt, A., and Wuhrer, M. (2016). Human Plasma Protein N-Glycosylation. *Glycoconj J.* 33, 309–343. doi:10.1007/s10719-015-9626-2

Conway, P., Tyka, M. D., Dimaio, F., Konerding, D. E., and Baker, D. (2014). Relaxation of Backbone Bond Geometry Improves Protein Energy Landscape Modeling. *Protein Sci.* 23, 47–55. doi:10.1002/pro.2389

Craik, D. J., Fairlie, D. P., Liras, S., and Price, D. (2013). The Future of Peptide-Based Drugs. *Chem. Biol. Drug Des.* 81, 136–147. doi:10.1111/cbdd.12055

Demeler, B. (2010). Methods for the Design and Analysis of Sedimentation Velocity and Sedimentation Equilibrium Experiments with Proteins. *Curr. Protoc. Protein Sci.* Chapter 7, Unit, 13. doi:10.1002/0471140864.ps0713s60

Demeler, B., and Brookes, E. (2008). Monte Carlo Analysis of Sedimentation Experiments. *Colloid Polym. Sci.* 286, 129–137. doi:10.1007/s00396-007-1699-4

Demeler, B., and Gorbet, G. E. (2016). "Analytical Ultracentrifugation Data Analysis with UltraScan-III," in *Analytical Ultracentrifugation*. Editors S. Uchiyama, F. Arisaka, W. Stafford, and T. Laue (Tokyo: Springer), 119–143. doi:10.1007/978-4-431-55985-6_8

Demeler, B., and van Holde, K. E. (2004). Sedimentation Velocity Analysis of Highly Heterogeneous Systems. *Anal. Biochem.* 335, 279–288. doi:10.1016/j.ab.2004.08.039

Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., et al. (2007). PDB2PQR: Expanding and Upgrading Automated Preparation of Biomolecular Structures for Molecular Simulations. *Nucleic Acids Res.* 35, W522–W525. doi:10.1093/nar/gkm276

Dos Santos, J., Fernandes, C., Magro, A., and Fontes, M. (2009). The Intriguing Phospholipases A2 Homologues: Relevant Structural Features on Myotoxicity and Catalytic Inactivity. *Ppl* 16, 887–893. doi:10.2174/092986609788923310

Einstein, A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik* 322, 549–560. doi:10.1002/andp.19053220806

Erak, M., Bellmann-Sickert, K., Els-Heindl, S., and Beck-Sickinger, A. G. (2018). Peptide Chemistry Toolbox - Transforming Natural Peptides into Peptide Therapeutics. *Bioorg. Med. Chem.* 26, 2759–2765. doi:10.1016/j.bmc.2018.01.012

Faini, M., Stengel, F., and Aebersold, R. (2016). The Evolving Contribution of Mass Spectrometry to Integrative Structural Biology. *J. Am. Soc. Mass. Spectrom.* 27, 966–974. doi:10.1007/s13361-016-1382-4

Fan, H. W., and Monteiro, W. M. (2018). History and Perspectives on How to Ensure Antivenom Accessibility in the Most Remote Areas in Brazil. *Toxicon* 151, 15–23. doi:10.1016/j.toxicon.2018.06.070

Fernandes, C. A. H., Borges, R. J., Lomonte, B., and Fontes, M. R. M. (2014). A Structure-Based Proposal for a Comprehensive Myotoxic Mechanism of Phospholipase A2-like Proteins from Viperid Snake Venoms. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1844, 2265–2276. doi:10.1016/j.bbapap.2014.09.015

Fernandes, C. A. H., Comparetti, E. J., Borges, R. J., Huancahuire-Vega, S., Ponce-Soto, L. A., Marangoni, S., et al. (2013). Structural Bases for a Complete Myotoxic Mechanism: crystal Structures of Two Non-catalytic Phospholipases A2-like from Bothrops Brazili Venom. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1834, 2772–2781. doi:10.1016/j.bbapap.2013.10.009

Ferrari, A. J. R., Clasen, M. A., Kurt, L., Carvalho, P. C., Gozzo, F. C., and Martínez, L. (2019). TopoLink: Evaluation of Structural Models Using Chemical Crosslinking Distance Constraints. *Bioinformatics* 35, 3169–3170. doi:10.1093/bioinformatics/btz014

Ferrige, A. G., Seddon, M. J., Green, B. N., Jarvis, S. A., Skilling, J., and Staunton, J. (1992). Disentangling Electrospray Spectra with Maximum Entropy. *Rapid Commun. Mass. Spectrom.* 6, 707–711. doi:10.1002/rcm.1290061115

Fioramonte, M., De Jesus, H. C. R., Ferrari, A. J. R., Lima, D. B., Drekener, R. L., Correia, C. R. D., et al. (2018). XPlex: An Effective, Multiplex Cross-Linking Chemistry for Acidic Residues. *Anal. Chem.* 90, 6043–6050. doi:10.1021/acs.analchem.7b05135

Fosgerau, K., and Hoffmann, T. (2015). Peptide Therapeutics: Current Status and Future Directions. *Drug Discov. Today* 20, 122–128. doi:10.1016/j.drudis.2014.10.003

Francis, B., Gutierrez, J. M., Lomonte, B., and Kaiser, I. I. (1991). Myotoxin II from *Bothrops asper* (Terciopelo) Venom Is a Lysine-49 Phospholipase A2. *Arch. Biochem. Biophys.* 284, 352–359. doi:10.1016/0003-9861(91)90307-5

Franke, D., Petoukhov, M. V., Konarev, P. V., Panjkovich, A., Tuukkanen, A., Mertens, H. D. T., et al. (2017). ATSAS 2.8: a Comprehensive Data Analysis Suite for Small-Angle Scattering from Macromolecular Solutions. *J. Appl. Cryst.* 50, 1212–1225. doi:10.1107/s1600576717007786

Glatter, O., and Kratky, O. (1982). *Small Angle X-ray Scattering*. London: New York Academic Press.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., et al. (2003). Protein-protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. *J. Mol. Biol.* 331, 281–299. doi:10.1016/s0022-2836(03)00670-3

Gutiérrez, J. M., Calvete, J. J., Habib, A. G., Harrison, R. A., Williams, D. J., and Warrell, D. A. (2017). Snakebite Envenoming. *Nat. Rev. Dis. Primers* 3, 17063. doi:10.1038/nrdp.2017.63

Gutiérrez, J. M., León, G., Rojas, G., Lomonte, B., Rucavado, A., and Chaves, F. (1998). Neutralization of Local Tissue Damage Induced by *Bothrops asper* (Terciopelo) Snake Venom. *Toxicon* 36, 1529–1538. doi:10.1016/s0041-0101(98)00145-7

Gutiérrez, J. M., and Ownby, C. L. (2003). Skeletal Muscle Degeneration Induced by Venom Phospholipases A_2: Insights into the Mechanisms of Local and Systemic Myotoxicity. *Toxicon* 42, 915–931. doi:10.1016/j.toxicon.2003.11.005

Gutiérrez, J. M., Theakston, R. D., and Warrell, D. A. (2006). Confronting the Neglected Problem of Snake Bite Envenoming: the Need for a Global Partnership. *Plos Med.* 3, 727–731. doi:10.1371/journal.pmed.0030150

Habib, A. G., Kuznik, A., Hamza, M., Abdullahi, M. I., Chedi, B. A., Chippaux, J.-P., et al. (2015). Snakebite Is under Appreciated: Appraisal of Burden from West Africa. *Plos Negl. Trop. Dis.* 9, e0004088. doi:10.1371/journal.pntd.0004088

Heukeshoven, J., and Dernick, R. (1985). Simplified Method for Silver Staining of Proteins in Polyacrylamide Gels and the Mechanism of Silver Staining. *Electrophoresis* 6, 130–112. doi:10.1002/elps.1150060302

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

Iacobucci, C., and Sinz, A. (2017). To Be or Not to Be? Five Guidelines to Avoid Misassignments in Cross-Linking/Mass Spectrometry. *Anal. Chem.* 89, 7832–7835. doi:10.1021/acs.analchem.7b02316

Iavarone, A. T., and Williams, E. R. (2003). Mechanism of Charging and Supercharging Molecules in Electrospray Ionization. *J. Am. Chem. Soc.* 125, 2319–2327. doi:10.1021/ja021202t

Ishioka, N., Takahashi, N., and Putnam, F. W. (1986). Amino Acid Sequence of Human Plasma Alpha 1B-Glycoprotein: Homology to the Immunoglobulin Supergene Family. *Proc. Natl. Acad. Sci.* 83, 2363–2367. doi:10.1073/pnas.83.8.2363

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

Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* 118, 11225–11236. doi:10.1021/ja9621760

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

Kikhney, A. G., and Svergun, D. I. (2015). A Practical Guide to Small Angle X-ray Scattering (SAXS) of Flexible and Intrinsically Disordered Proteins. *FEBS Lett.* 589, 2570–2577. doi:10.1016/j.febslet.2015.08.027

Knudsen, C., Jürgensen, J. A., Føns, S., Haack, A. M., Friis, R. U. W., Dam, S. H., et al. (2021). Snakebite Envenoming Diagnosis and Diagnostics. *Front. Immunol.* 12, 661457. doi:10.3389/fimmu.2021.661457

Koukos, P. I., and Bonvin, A. M. J. J. (2019). Integrative Modelling of Biomolecular Complexes. *J. Mol. Biol.* 432, 2861–2881. doi:10.1016/j.jmb.2019.11.009

Kulik, M., Mori, T., and Sugita, Y. (2021). Multi-Scale Flexible Fitting of Proteins to Cryo-EM Density Maps at Medium Resolution. *Front. Mol. Biosci.* 8, 631854. doi:10.3389/fmolb.2021.631854

Laemmli, U. K. (1970). Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature* 227, 680–685. doi:10.1038/227680a0

Lau, J. L., and Dunn, M. K. (2018). Therapeutic Peptides: Historical Perspectives, Current Development Trends, and Future Directions. *Bioorg. Med. Chem.* 26, 2700–2707. doi:10.1016/j.bmc.2017.06.052

Laustsen, A. H., Engmark, M., Milbo, C., Johannesen, J., Lomonte, B., Gutiérrez, J. M., et al. (2016). From Fangs to Pharmacology: The Future of Snakebite Envenoming Therapy. *Curr. Pharm. Des.* 22, 5270–5293. doi:10.2174/1381612822666160623073438

Leitner, A., Walzthoeni, T., and Aebersold, R. (2014). Lysine-specific Chemical Cross-Linking of Protein Complexes and Identification of Cross-Linking Sites Using LC-MS/MS and the xQuest/xProphet Software Pipeline. *Nat. Protoc.* 9, 120–137. doi:10.1038/nprot.2013.168

Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F., Rinner, O., Beck, M., et al. (2010). Probing Native Protein Structures by Chemical Cross-Linking, Mass Spectrometry, and Bioinformatics. *Mol. Cell Proteomics* 9, 1634–1649. doi:10.1074/mcp.r000001-mcp201

León, I. R., Neves-Ferreira, A. G. C., Rocha, S. L. G., Trugilho, M. R. O., Perales, J., and Valente, R. H. (2012). Using Mass Spectrometry to Explore the Neglected Glycan Moieties of the Antiophidic Proteins DM43 and DM64. *Proteomics* 12, 2753–2765.

Lewis, M. S., and Junghans, R. P. (2000). Ultracentrifugal Analysis of Molecular Mass of Glycoproteins of Unknown or Ill-Defined Carbohydrate Composition. *Methods Enzymol.* 321, 136–149. doi:10.1016/s0076-6879(00)21191-9

Lima, D. B., De Lima, T. B., Balbuena, T. S., Neves-Ferreira, A. G. C., Barbosa, V. C., Gozzo, F. C., et al. (2015). SIM-XL: A Powerful and User-Friendly Tool for Peptide Cross-Linking Analysis. *J. Proteomics* 129, 51–55. doi:10.1016/j.jprot.2015.01.013

Liu, F., and Heck, A. J. (2015). Interrogating the Architecture of Protein Assemblies and Protein Interaction Networks by Cross-Linking Mass Spectrometry. *Curr. Opin. Struct. Biol.* 35, 100–108. doi:10.1016/j.sbi.2015.10.006

Lomonte, B., Angulo, Y., and Calderón, L. (2003). An Overview of Lysine-49 Phospholipase A2 Myotoxins from Crotalid Snake Venoms and Their Structural Determinants of Myotoxic Action. *Toxicon* 42, 885–901. doi:10.1016/j.toxicon.2003.11.008

Lomonte, B., and Gutiérrez, J. (1989). A New Muscle Damaging Toxin, Myotoxin II, from the Venom of the Snake *Bothrops asper* (Terciopelo). *Toxicon* 27, 725–733. doi:10.1016/0041-0101(89)90039-1

Lomonte, B., and Rangel, J. (2012). Snake Venom Lys49 Myotoxins: From Phospholipases A2 to Non-enzymatic Membrane Disruptors. *Toxicon* 60, 520–530. doi:10.1016/j.toxicon.2012.02.007

Markwick, P. R. L., Malliavin, T., and Nilges, M. (2008). Structural Biology by NMR: Structure, Dynamics, and Interactions. *Plos Comput. Biol.* 4, e1000168. doi:10.1371/journal.pcbi.1000168

Merkley, E. D., Rysavy, S., Kahraman, A., Hafen, R. P., Daggett, V., and Adkins, J. N. (2014). Distance Restraints from Crosslinking Mass Spectrometry: Mining a Molecular Dynamics Simulation Database to Evaluate Lysine-Lysine Distances. *Protein Sci.* 23, 747–759. doi:10.1002/pro.2458

Mora-Obando, D., Díaz, C., Angulo, Y., Gutiérrez, J. M., and Lomonte, B. (2014a). Role of Enzymatic Activity in Muscle Damage and Cytotoxicity Induced byBothrops asperAsp49 Phospholipase A2myotoxins: Are There Additional Effector Mechanisms Involved? *PeerJ* 2, e569. doi:10.7717/peerj.569

Mora-Obando, D., Fernández, J., Montecucco, C., Gutiérrez, J. M., and Lomonte, B. (2014b). Synergism between Basic Asp49 and Lys49 Phospholipase A2 Myotoxins of Viperid Snake Venom *In Vitro* and *In Vivo*. *PLoS One* 9, e109846. doi:10.1371/journal.pone.0109846

Murakami, M. T., Arruda, E. Z., Melo, P. A., Martinez, A. B., Calil-Eliás, S., Tomaz, M. A., et al. (2005). Inhibition of Myotoxic Activity of *Bothrops asper* Myotoxin II by the Anti-trypanosomal Drug Suramin. *J. Mol. Biol.* 350, 416–426. doi:10.1016/j.jmb.2005.04.072

Natarajan, K., Mage, M. G., and Margulies, D. H. (2015). "Immunoglobulin Superfamily," in *Encyclopedia of Life Sciences* (Chichester: John Wiley & Sons). doi:10.1002/9780470015902.a0000926.pub2

Neves-Ferreira, A. G. C., Perales, J., Fox, J. W., Shannon, J. D., Makino, D. L., Garratt, R. C., et al. (2002). Structural and Functional Analyses of DM43, a Snake Venom Metalloproteinase Inhibitor from Didelphis marsupialisSerum. *J. Biol. Chem.* 277, 13129–13137. doi:10.1074/jbc.m200589200

Neves-Ferreira, A. G. C., Valente, R. H., Domont, G. B., and Perales, J. (2017). "Natural Inhibitors of Snake Venom Metallopeptidases," in *Handbooks of Toxinology*. Editor P. Gopalakrishnakone (New York, United States: Springer). doi:10.1007/978-94-007-6452-1_19

Neves-Ferreira, A. G. C., Valente, R. H., Perales, J., and Domont, G. B. (2009). "Natural Inhibitors: Innate Immunity to Snake Venoms," in *Reptile Venoms and Toxins*. Editor S. P. Mackessy (New York: Taylor & Francis/CRC Press), 259–284.

Nivón, L. G., Moretti, R., and Baker, D. (2013). A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS One* 8, e59004. doi:10.1371/journal.pone.0059004

Ohkura, N., Okuhara, H., Inoue, S., Ikeda, K., and Hayashi, K. (1997). Purification and Characterization of Three Distinct Types of Phospholipase A2 Inhibitors from the Blood Plasma of the Chinese Mamushi, Agkistrodon Blomhoffii Siniticus. *Biochem. J.* 325, 527–531. doi:10.1042/bj3250527

Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995). How to Measure and Predict the Molar Absorption Coefficient of a Protein. *Protein Sci.* 4, 2411–2423. doi:10.1002/pro.5560041120

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

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., et al. (2019). The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. *Nucleic Acids Res.* 47, D442–D450. doi:10.1093/nar/gky1106

Rocha, S. L. G., Lomonte, B., Neves-Ferreira, A. G. C., Trugilho, M. R. O., Junqueira-De-Azevedo, I. d. L. M., Ho, P. L., et al. (2002). Functional Analysis of DM64, an Antimyotoxic Protein with Immunoglobulin-like Structure fromDidelphis Marsupialisserum. *Eur. J. Biochem.* 269, 6052–6062. doi:10.1046/j.1432-1033.2002.03308.x

Rocha, S. L. G., Neves-Ferreira, A. G. C., Trugilho, M. R. O., Angulo, Y., Lomonte, B., Valente, R. H., et al. (2017). Screening for Target Toxins of the Antiophidic Protein DM64 through a Gel-Based Interactomics Approach. *J. Proteomics* 151, 204–213. doi:10.1016/j.jprot.2016.05.020

Rocha, S. L. G., Neves-Ferreira, A. G. C., Trugilho, M. R. O., Chapeaurouge, A., León, I. R., Valente, R. H., et al. (2009). Crotalid Snake Venom Subproteomes Unraveled by the Antiophidic Protein DM43. *J. Proteome Res.* 8, 2351–2360. doi:10.1021/pr800977s

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a Unified Platform for Automated Protein Structure and Function Prediction. *Nat. Protoc.* 5, 725–738. doi:10.1038/nprot.2010.5

Sali, A., Berman, H. M., Schwede, T., Trewhella, J., Kleywegt, G., Burley, S. K., et al. (2015). Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* 23, 1156–1167. doi:10.1016/j.str.2015.05.013

Salvador, G. H. M., Gomes, A. A. S., Bryan-Quirós, W., Fernández, J., Lewin, M. R., Gutiérrez, J. M., et al. (2019). Structural Basis for Phospholipase A2-like Toxin Inhibition by the Synthetic Compound Varespladib (LY315920). *Sci. Rep.* 9, 17203. doi:10.1038/s41598-019-53755-5

Salvador, G. H. M., Dos Santos, J. I., Lomonte, B., and Fontes, M. R. M. (2017). Crystal Structure of a Phospholipase A2 from *Bothrops asper* Venom: Insights into a New Putative "myotoxic Cluster". *Biochimie* 133, 95–102. doi:10.1016/j.biochi.2016.12.015

Schilling, B., Row, R. H., Gibson, B. W., Guo, X., and Young, M. M. (2003). MS2Assign, Automated Assignment and Nomenclature of Tandem Mass Spectra of Chemically Crosslinked Peptides. *J. Am. Soc. Mass. Spectrom.* 14, 834–850. doi:10.1016/s1044-0305(03)00327-1

Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996). Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels. *Anal. Chem.* 68, 850–858. doi:10.1021/ac950914h

Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., et al. (2019). CATH: Expanding the Horizons of Structure-Based Functional Annotations for Genome Sequences. *Nucleic Acids Res.* 47, D280–D284. doi:10.1093/nar/gky1097

Sinz, A. (2003). Chemical Cross-Linking and Mass Spectrometry for Mapping Three-Dimensional Structures of Proteins and Protein Complexes. *J. Mass. Spectrom.* 38, 1225–1237. doi:10.1002/jms.559

Sinz, A. (2006). Chemical Cross-Linking and Mass Spectrometry to Map Three-Dimensional Protein Structures and Protein-Protein Interactions. *Mass. Spectrom. Rev.* 25, 663–682. doi:10.1002/mas.20082

Sinz, A. (2014). The Advancement of Chemical Cross-Linking and Mass Spectrometry for Structural Proteomics: from Single Proteins to Protein Interaction Networks. *Expert Rev. Proteomics* 11, 733–743. doi:10.1586/14789450.2014.960852

Sun, Z., Liu, Q., Qu, G., Feng, Y., and Reetz, M. T. (2019). Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem. Rev.* 119, 1626–1665. doi:10.1021/acs.chemrev.8b00290

Svergun, D., Barberato, C., and Koch, M. H. J. (1995). CRYSOL- a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Cryst.* 28, 768–773. doi:10.1107/s0021889895007047

Svergun, D. I. (1992a). Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. *J. Appl. Cryst.* 25, 495–503. doi:10.1107/s0021889892001663

Udby, L., Sørensen, O. E., Pass, J., Johnsen, A. H., Behrendt, N., Borregaard, N., et al. (2004). Cysteine-Rich Secretory Protein 3 Is a Ligand of α1B-Glycoprotein in Human Plasma. *Biochemistry* 43, 12877–12886. doi:10.1021/bi048823e

Vlieg, H. C., Huizinga, E. G., and Janssen, B. J. C. (2019). Structure and Flexibility of the Extracellular Region of the PirB Receptor. *J. Biol. Chem.* 294, 4634–4643. doi:10.1074/jbc.ra118.004396

Voss, R. S., and Jansa, S. A. (2012). Snake-venom Resistance as a Mammalian Trophic Adaptation: Lessons from Didelphid Marsupials. *Biol. Rev. Camb Philos. Soc.* 87, 822–837. doi:10.1111/j.1469-185x.2012.00222.x

Wang, C., Schueler-Furman, O., and Baker, D. (2005). Improved Side-Chain Modeling for Protein-Protein Docking. *Protein Sci.* 14, 1328–1339. doi:10.1110/ps.041222905

Williams, D. J., Faiz, M. A., Abela-Ridder, B., Ainsworth, S., Bulfone, T. C., Nickerson, A. D., et al. (2019). Strategy for a Globally Coordinated Response to a Priority Neglected Tropical Disease: Snakebite Envenoming. *Plos Negl. Trop. Dis.* 13, e0007059. doi:10.1371/journal.pntd.0007059

Williams, H. F., Layfield, H. J., Vallance, T., Patel, K., Bicknell, A. B., Trim, S. A., et al. (2019). The Urgent Need to Develop Novel Strategies for the Diagnosis and Treatment of Snakebites. *Toxins (Basel)* 11, 363. doi:10.3390/toxins11060363

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved Protein Structure Prediction Using Predicted Interresidue Orientations. *Proc. Natl. Acad. Sci. USA* 117, 1496–1503. doi:10.1073/pnas.1914677117

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* 12, 7–8. doi:10.1038/nmeth.3213

Yu, C., and Huang, L. (2018). Cross-linking Mass Spectrometry: an Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* 90, 144–165. doi:10.1021/acs.analchem.7b04431

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 declared a shared affiliation with several of the authors B.S., S.R., V.B., J.P., R.V., F.G-N., A.N.F. at time of review.

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.

frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 04 February 2022
doi: 10.3389/fmolb.2022.832847

Sampling of Protein Conformational Space Using Hybrid Simulations: A Critical Assessment of Recent Methods

***Burak T. Kaynak*[1‡], *James M. Krieger*[1†‡], *Balint Dudas*[1,2,3‡], *Zakaria L. Dahmani*[1], *Mauricio G. S. Costa*[4], *Erika Balog*[3], *Ana Ligia Scott*[5], *Pemra Doruker*[1*], *David Perahia*[2*] and *Ivet Bahar*[1*]**

[1]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States, [2]Laboratoire de Biologie et Pharmacologie Appliquée, Ecole Normale Supérieure Paris-Saclay, Gif-sur-Yvette, France, [3]Department of Biophysics and Radiation Biology, Semmelweis University, Budapest, Hungary, [4]Programa de Computação Científica, Vice-Presiden̂cia de Educação, Informação e Comunicação, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, [5]Laboratory of Bioinformatics and Computational Biology, Center of Mathematics, Computation and Cognition, Federal University of ABC-UFABC, Santo André, Brazil

OPEN ACCESS

Edited by:
Massimiliano Bonomi,
Institut Pasteur, France

Reviewed by:
James Seckler,
Case Western Reserve University,
United States
Igor N. Berezovsky,
*Bioinformatics Institute (A*STAR),*
Singapore

***Correspondence:**
Ivet Bahar
bahar@pitt.edu
David Perahia
david.perahia@ens-paris-saclay.fr
Pemra Doruker
doruker@pitt.edu

†Present address:
James M. Krieger,
Biocomputing Unit, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Madrid, Spain

‡These authors share first authorship

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *10 December 2021*
Accepted: *12 January 2022*
Published: *04 February 2022*

Citation:
Kaynak BT, Krieger JM, Dudas B, Dahmani ZL, Costa MGS, Balog E, Scott AL, Doruker P, Perahia D and Bahar I (2022) Sampling of Protein Conformational Space Using Hybrid Simulations: A Critical Assessment of Recent Methods.
Front. Mol. Biosci. 9:832847.
doi: 10.3389/fmolb.2022.832847

Recent years have seen several hybrid simulation methods for exploring the conformational space of proteins and their complexes or assemblies. These methods often combine fast analytical approaches with computationally expensive full atomic molecular dynamics (MD) simulations with the goal of rapidly sampling large and cooperative conformational changes at full atomic resolution. We present here a systematic comparison of the utility and limits of four such hybrid methods that have been introduced in recent years: MD with excited normal modes (MDeNM), collective modes-driven MD (CoMD), and elastic network model (ENM)-based generation, clustering, and relaxation of conformations (ClustENM) as well as its updated version integrated with MD simulations (ClustENMD). We analyzed the predicted conformational spaces using each of these four hybrid methods, applied to four well-studied proteins, triosephosphate isomerase (TIM), 3-phosphoglycerate kinase (PGK), HIV-1 protease (PR) and HIV-1 reverse transcriptase (RT), which provide extensive ensembles of experimental structures for benchmarking and comparing the methods. We show that a rigorous multi-faceted comparison and multiple metrics are necessary to properly assess the differences between conformational ensembles and provide an optimal protocol for achieving good agreement with experimental data. While all four hybrid methods perform well in general, being especially useful as computationally efficient methods that retain atomic resolution, the systematic analysis of the same systems by these four hybrid methods highlights the strengths and limitations of the methods and provides guidance for parameters and protocols to be adopted in future studies.

Keywords: conformational landscape/space, normal mode analysis, molecular simulations, elastic network models, HIV-1 protease, triosephosphate isomerase, 3-phosphoglycerate kinase, HIV-1 reverse transcriptase

INTRODUCTION

Under physiological conditions, proteins sample a distribution of conformations while retaining their native fold. Indeed, the dynamic equilibrium of accessible conformations often underlies the regulation of protein function and allosteric mechanisms or their adaptability to bind various ligands or drugs (Haliloglu and Bahar, 2015; Zhang et al., 2020; Wingert et al., 2021). Several studies in the last decade have confirmed the importance of structural dynamics in facilitating, if not driving, the interactions and function of biomolecular systems in the cell (Bahar et al., 2010; Orellana, 2019; Thirumalai et al., 2019; Resende-Lara et al., 2020). In particular, the role of structural dynamics in supporting catalytic activity is a topic of interest (Yon et al., 1998; Bahar et al., 2010; Jiang et al., 2011), with the understanding that enzymes are mechanochemical entities (Yang and Bahar, 2005) and conformational mechanics often complement chemical events by enabling domain or loop movements required for activation.

The determination of 3D coordinates of proteins and their complexes/assemblies has accelerated in recent years thanks to advances in experimental methodologies. Specifically, the developments in cryo-electron microscopy (cryo-EM) and X-ray free-electron laser (FEL) crystallography have revealed multiple snapshots of flexible and complex molecular systems (Branden and Neutze, 2021). In parallel with rapidly growing structural data, theoretical and computational methods that exploit those data toward gaining insights into mechanisms of function have gained importance. While traditional methods, exemplified by molecular dynamics (MD) simulations work as primary tools for studying dynamic events at full atomic details, they still fall short of providing an adequate description of cooperative events at time scales beyond microseconds for multi-domain/multi-subunit systems. On the other hand, analytical methods and coarse-grained (CG) models, exemplified by Normal Mode Analysis (NMA) with elastic network models (ENMs), permit us to solve for the spectrum of modes uniquely accessible to supramolecular systems, providing mathematically exact and physically plausible information on cooperative events, albeit neglecting anharmonicity and atomic details.

Several approaches have been developed aiming to increase the specificity of CG approaches while retaining the high resolution of full atomic simulations. Many ENM-based approaches have focused on the optimization of the basic parameters, spring constants and cutoff distances/functions for inter-residue interactions (Hinsen, 1998; Yang et al., 2009; Kaynak et al., 2018; Kaynak and Doruker, 2019; Koehl et al., 2021), but such studies fall short of providing atomic level information. Instead, another research line, the development of the so-called hybrid methods that combine MD and NMA (using either ENMs or full atomic models) proved to be useful in recent years. These methods have demonstrated two key advantages: 1) an accurate description of cooperative changes in structure, usually described by low frequency normal modes (NMs), and 2) providing atomic details and incorporating local non-linear effects from MD simulations that ‘recalibrate’ these conformational changes (Krieger et al., 2020). Such methods are also beneficial for flexible fitting to cryo-EM maps (Costa et al., 2020) where methods employing either MD or NMA are typically used (Miyashita and Tama, 2018).

In this article, we provide a comparative analysis of such hybrid methods developed for efficient sampling of the conformational space and the possible transitions between functional states. We focus on four methods: ClustENM (Kurkcuoglu et al., 2016), its recent extension ClustENMD (Kaynak et al., 2021), MD with excited NMs (MDeNM) (Costa et al., 2015), and collective MD (CoMD) (Gur et al., 2013). ClustENM produces successive generations of conformers by deforming along low frequency modes, clustering the conformers, and performing energy minimization at full atomic scale. ClustENM conformers have been effectively used in ensemble docking studies for protein-ligand, protein-protein and protein-DNA/RNA pairs (Kurkcuoglu and Doruker, 2016; Can et al., 2017; Kurkcuoglu and Bonvin, 2020), including supramolecules like the ribosome. The recent extension, ClustENMD, uses short MD simulations for the refinement of the generated conformers. The MDeNM method is a multi-replica protocol designed to enhance conformational exploration in a subspace defined by a set of low-frequency NMs, also including the couplings with localized motions occurring within the Cartesian space. In this method, additional atomic velocities are introduced along different linear combinations of NMs. Even though NMs are usually computed in vacuum, they are used as privileged directions in MD simulations with an explicit representation of the surrounding medium. MDeNM has demonstrated its power in conformational sampling in several studies revealing important protein functional movements (Dudas et al., 2020; Dudas et al., 2021a; Fagnen et al., 2020; Fagnen et al., 2021) and has also been successfully used in ensemble docking studies (Dudas et al., 2021b). CoMD provides a combination of ENM-NMA and targeted MD, coupled with energy minimization to adaptively generate a series of conformers.

The metrics for evaluating the performance of these methods deserve attention. For example, while the ability to reproduce crystallographic B-factors has been adopted as a metric in many studies, the comparison of ENM-NMA predictions with the covariance derived from MD simulations were reported to enable a more accurate assessment (Fuglebakk et al., 2013). Here we use the data from both MD and experiments to evaluate the principal components (PCs) of structural changes observed in experiments and predicted by the hybrid methods. The idea, independently introduced in two original studies (Yang et al., 2008; Bakan and Bahar, 2009), is to consider the ensemble of structures resolved for a given protein (e.g. multiple X-ray structures resolved in the presence of different drugs for HIV-1 protease), and examine whether this ‘experimental space’ of conformations matches that predicted computationally. This is a rigorous comparison, unbiased by the selection of conformers used as reference.

We perform our comparative analysis for four well-studied enzymes: triosephosphate isomerase (TIM), 3-phosphoglycerate kinase (PGK), HIV-1 protease (PR), and HIV-1 reverse transcriptase (RT) (**Figure 1**). **Table 1** lists the properties of these enzymes, including the reference structure, the size and oligomeric state of the protein, and the number of experimentally resolved structures used in our comparative analysis along with the corresponding threshold for pairwise sequence identity. Overall, the study serves two major purposes: it provides a

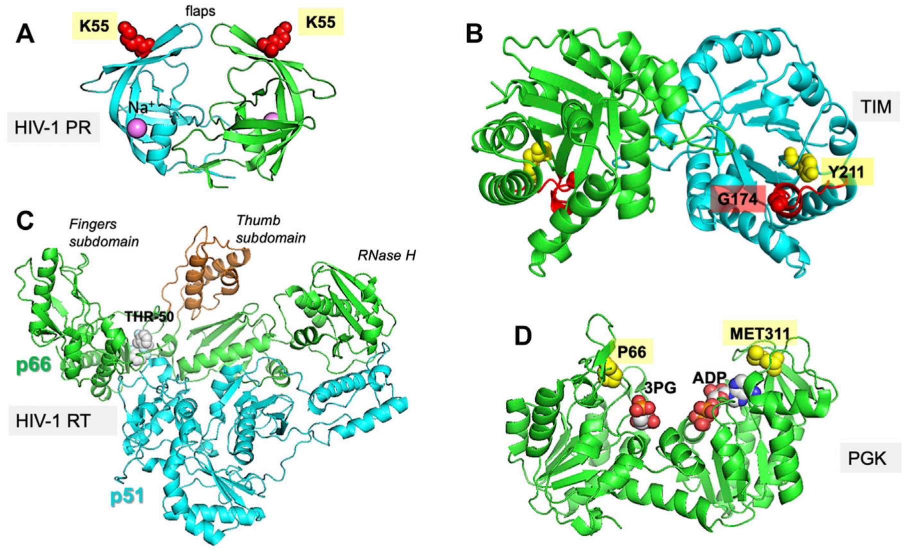


FIGURE 1 | Proteins investigated in the present study. The figure displays the experimentally resolved X-ray structures also used as initial structures for simulations. **(A)** HIV-1 protease (PR) (PDB id: 1tw7) is a wide-open, apo structure. The residue K55 on each subunit of the homodimer isused to probe the opening/closure of the flaps. **(B)** TIM (PDB id:1tcd) is a homodimeric enzyme, for which the catalytic loop is shown in *red* on both subunits of the apo state. The distance between the catalytic loop tip residue G174 and Y211 defines loop opening/closure motion in each subunit. **(C)** HIV-1 RT (PDB id: 2b6a) is a heterodimer composed of p51 and p66 subunits. The current structure is in complex with THR-50. The distance between the fingers and thumb subdomains, both located on the p66 subunit, indicate a transition between closed and open conformations of the region between these two subdomains. **(D)** PGK (PDB id: 2xe7) in the presence of the two substrates, 1,3-bisphosphoglycerate (bPG) and ADP. The distance between P66 and M311, two residues located at the tips of the N- and C-domains near the ligands, probes the opening/closing movement of the enzyme required for its catalytic activity.

TABLE 1 | Proteins used as case studies, and corresponding structural data from experiments.

Protein name (acronym)	PDB structure used as References/initial State	Total # of residues[a] and functional oligomerization state	Number of experimentally resolved structures (and their sequence identity threshold[b])
HIV-1 protease (PR)	1tw7 Martin et al. (2005); Apo	198 (Homodimer; 99 residues/monomer)	768 (90%)
3-phosphoglycerate kinase (PGK)	2xe7 Zerrad et al. (2011); Complex[c]	413 (Monomer; 416 residues)	35 (90%)
triosephosphate isomerase (TIM)	1tcd Maldonado et al. (1998); Apo	497 (Homodimer; 251 residues per monomer)	57 (50%)
HIV-1 reverse transcriptase (RT)	2b6a Morningstar et al. (2007); Complex with a NNRTI[d]	978 (Heterodimer; 560 residues in p66 subunit, 440 in p51)	365 (90%)

[a] *The number of residues resolved in the reference structure. The actual number is written in parenthesis).*
[b] *The percentage in parentheses is the sequence identity threshold after optimal multiple sequence alignment.*
[c] *The ternary complex with 3 PG and ADP, in the open state of PGK, was used as reference structure.*
[d] *NNRTI: non-nucleoside RT inhibitor.*

rigorous comparison of the performance of the hybrid methods revealing their limitations and advantages, and it helps determine the optimal parameters used in these methods thus permitting us to build fully automated algorithms that can be readily adopted for future applications.

METHODS

We present below a brief description of the hybrid methods examined here along with the methods used for comparative analysis. **Table 2** provides a summary of the parameters and protocols used in each hybrid technique, along with the number of conformers generated for each studied system. Overall, *six ensembles* of structures or conformations are studied: those resolved experimentally, those sampled by MD, and those predicted by four hybrid methods, ClustENM, ClustENMD, MDeNM, and coMD. All methods are applicable to proteins, DNA, and RNA molecules and their complexes.

ClustENM and ClustENMD

ClustENM (Kurkcuoglu et al., 2016) is a fully automated conformational sampling method composed of multiple generations/cycles consisting each of the following steps: 1)

TABLE 2 | Parameters, protocols, and outputs of investigated hybrid techniques.

Hybrid method	# Of modes	RMSD[a] or T (K)[b]	# Of runs	Specification of the methods	# Of conformers generated by the hybrid methods			
					PR	PGK[c]	TIM	RT
CoMD	3	1 Å	9/9/9/10 for PR/PGK/TIM/RT	● 50 cycles of targeted MD guided by ANM ● Monte Carlo-metropolis criteria for selecting new conformers[e] ● Explicit (TIP3) water and ion	459	450	459	510
ClustENM[d]	3	1 Å	3	● 5 generations of ANM sampling ● Energy minimization (EM) ● Implicit solvent	898	900	903	903
ClustENMD[d]	3	1 Å	3	● 5 generations of ANM sampling ● MD (for heating up the system) ● Implicit solvent	902	903	903	903
MDeNM	3	2/2/3/3 K for PR/PGK/TIM/RT	1	● Excited MD using atomic NMA ● 20/8/5/10 re-excitations for PR/PGK/TIM/RT ● Explicit (TIP3) water and ion	1,560	1,056	415	600

[a] *Deformation size or RMSD (Å) is used in coMD, for setting target structures; whereas it is used to generate the conformers in each generation of ClustENM(D) followed by relaxation.*

[b] *MDeNM uses an excitation temperature for adding extra velocities to atoms along modes direction. Temperature values separated by slashes correspond to different systems ordered as PR, PGK, TIM, and RT, respectively.*

[c] *PGK, ligands were not included in ClustENM, or ClustENMD, but were included for other methods.*

[d] *For PR and PGK, a few conformers with high energies have been automatically excluded from the ensemble. The maximum number of conformers from each run is 301.*

conformer generation by deforming along global NMs calculated using the anisotropic network model (ANM) (Atilgan et al., 2001), 2) clustering of generated conformers, and 3) relaxation of cluster representatives. The cluster representatives will be the parent conformers that are passed onto the next generation of sampling. The ANM modes are updated for each parent conformer in each generation. A series of deformations along the ± directions of a few global modes (3–5) are carried out by targeting a specific root-mean-square deviation (RMSD) for deformation (*Step 1*). Representative conformers selected for computing the next generation of conformers are relaxed by energy minimization (EM) in implicit solvent (*Step 3*).

In the extended version of ClustENM, called ClustENMD (Kaynak et al., 2021), MD simulations using OpenMM (Eastman et al., 2017) are performed for conformational relaxation in *Step 3*. Relaxation can be performed either in implicit solvent (Onufriev et al., 2004) with the Amber99SB force field (Lindorff-Larsen et al., 2010) or explicit solvent. The former is used in this study. In ClustENMD, ANM sampling (*Step 1*) enables deformations along random combinations of global modes with a specified average RMSD from each parent conformer. Both ClustENM and ClustENMD have been implemented in the application programming interface (API) *ProDy* (Bakan et al., 2011; Zhang et al., 2020).

The present analysis allows us to compare two types of relaxation (*Step 3*) using 1) only EM (ClustENM) and 2) EM followed by heating up the system to a desired temperature (here 303.15 K) by MD simulations of about 3 ps at neutral pH (ClustENMD). The same parameters are used for all proteins, namely average RMSD of 1 Å for deformation and five generations of sampling composed each of random combinations of the first three global modes (**Table 2**).

Collective Molecular Dynamics

Collective Molecular Dynamics (coMD) simulations were run using scripts generated by an updated version of our previous coMD plugin (Gur et al., 2013; Gur et al., 2015) for VMD (Humphrey et al., 1996) available at https://github.com/prody/coMD. Like the old version, the current coMD plugin uses VMD to prepare the simulation system and interpret the output Tcl script. CoMD uses *ProDy* (Bakan et al., 2011; Zhang et al., 2020) to calculate NMs based on the ANM (Atilgan et al., 2001; Eyal et al., 2015), and NAMD (Phillips et al., 2020) for energy minimization and targeted MD (TMD) (Schlitter et al., 1994; Swift and McCammon, 2008). As previously described (Gur et al., 2013; Gur et al., 2015), ANM modes are selected in a Monte Carlo scheme (ANM-MC) with their probabilities and amplitudes based on their eigenvalues, such that the lowest frequency, or the energetically most favorable, global modes dominate the sampling. coMD can be used to either sample the transition path between two endpoints (with the help of a MC/Metropolis algorithm) or explore the conformational space in the vicinity of a starting conformer (by setting the Metropolis acceptance probability equal to 1). In the current implementation, we adopted the second procedure, given that our goal was to explore the conformational space in the absence of any bias.

Method parameters include the number of modes, maximum deviation per mode, and the total RMSD with respect to the starting conformer at each ANM-MC cycle. A combination of three modes, 0.1 Å deviation, and 1 to 1.5 Å RMSD per cycle was found to give the best compromise between sampling a reasonably large conformational space and avoiding unrealistic deformations such as unfolding. We used a TMD duration of 4 ps to be comparable to other methods. The CHARMM36m (C36m) force field (Huang et al., 2017) was used for all systems. In this study, we used CHARMM-GUI (Jo et al., 2008) systems set up in the other methods (rather than the CoMD VMD plugin) and the Tcl running script was adapted accordingly. This included the liganded complex for PGK as in MDeNM. All other parameters were kept at their default values including 20,000 kcal/mol/Å^2 for the TMD spring constant.

MD with Excited Normal Modes

The all-atom NMs required for MDeNM simulations (Costa et al., 2015) were calculated with CHARMM (Brooks et al., 2009) in conjunction with the additive C36m force field (Huang et al. 2017), considering the conformation obtained after an initial short equilibration MD run. The potential energy of the examined structure was minimized till an energy RMS gradient of 10^{-5} kcal/mol/Å was reached. Then, NMs were calculated using the VIBRAN module of CHARMM. The equilibrated solvated structures were considered as starting points for MDeNM simulations. Each MDeNM replica consisted of the conformational exploration along a single linear combination of the selected modes. An RMSD-based filtering was performed before the simulations. Briefly, random linear combinations of the three lowest frequency NMs were followed by 1 Å geometric displacements yielding a deformed structure that must present an RMSD greater than a given threshold (referred to as RMSD threshold) from others previously accepted. If the RMSD of a given generated structure is lower than the threshold, the combination is rejected, and another is generated. This procedure maximizes the variability observed in the excitation directions, therefore covering the defined NM space. These directions are then used to excite the protein kinetically. The excitations are applied periodically each after a given relaxation time during MD simulation, by increasing the atomic velocities along the given excitation direction. As the excitation energy rapidly dissipates, multiple excitations are needed (referred to as the number of excitations). Each excitation increases the temperature of the system by a given amount (called the excitation temperature). Each excitation step is followed by a relaxation time. In line with other studies, we found that excitation energies of 2–3 K coupled with relaxation times ranging from 4 to 8 ps (Costa et al., 2015; Floquet et al., 2015; Dudas et al., 2020, 2021a, 2021b) are broadly applicable. The number of cycles varies by system (**Table 2**). The intermediate conformers at the end of each excitation-relaxation cycle are collected to define the MDeNM ensembles. The other parameters are the same as those described below for MD simulations.

MD simulations

We carried out MD simulations in NAMD for comparison of the conformers sampled in MD trajectories with those generated by the hybrid techniques. For each protein, we carried out three independent runs of 200 ns each, explicit solvent at 303.15 K. The systems were prepared using CHARMM-GUI (Jo et al., 2008) and default parameters were used for MD simulations. A set of 6,000 snapshots have been collected at 100 ps intervals for each protein (2,000/run × 3 runs), forming the MD ensembles.

Comparison Between Computationally Predicted and Experimentally Resolved Ensembles

To compare the simulation outputs with the available experimental data, ensembles of experimentally resolved structures were compiled and subjected to PCA for each of the four proteins studied here (**Table 1**) using methods developed within the *ProDy* API (Bakan et al., 2011; Bakan et al., 2014; Zhang et al., 2019; Zhang et al., 2020; Zhang et al., 2021). The structures were projected onto the reduced space spanned by the first two PCs defined by the *e*xperimental dataset (*e*PC1 and *e*PC2) for each studied protein, using *ProDy* and visualized by Matplotlib (Hunter, 2007). The corresponding conformers generated by computational methods were also projected onto that subspace, thus allowing for comparison of their distribution in the backdrop of experimentally observed conformational space. Continuous population density plots of conformers were generated for each ensemble using kernel density estimate (KDE) plots from the Seaborn Python package (Waskom, 2021). We used *ProDy* for calculating the RMSDs and distance measures representing the departure from the different functional states of the proteins. The variances of computationally predicted conformers along *e*PC1–*e*PC10 were determined by projecting them onto these *e*PCs and evaluating their standard deviation from the mean.

We also determined the simulated PCs (sPCs) for each ensemble of conformations generated by MD, ClustENM, ClustENMD, MDeNM, and coMD, to determine the dominant modes of conformational changes predicted by the simulations or hybrid methods. Finally, to quantify the extent of similarity between the major structural variations observed in experiments and those predicted by simulations, we evaluated the correlation cosines between experimentally sampled top four PCs (*e*PC1-4) and those sampled in simulations (*s*PC1-4). Likewise, similar correlation cosines were evaluated for pairs of outputs from different computational methods. The results are presented in heat maps (6 × 6 super-matrices), the super-elements of which (4 × 4 blocks) describe the pairwise correlation cosines, or the so-called *overlaps* between pairs of PCs from different methods.

Furthermore, we used the root-weighted square inner product (RWSIP) (Carnevale et al., 2007) as another metric to assess the overall consistency between the spectrum of structural changes observed in simulations and those from experiments, defined as

$$RWSIP = \left[\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i^u \lambda_j^v \left(\boldsymbol{u}_i \cdot \boldsymbol{v}_j \right)^2}{\sum_{i=1}^{N} \lambda_i^u \lambda_i^v} \right]^{1/2} \tag{1}$$

Here λ_i^u and λ_j^v are the eigenvalues of the covariance matrices corresponding to their respective PCs, $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ (*s*PCs and *e*PCs). *N* is the number of the PCs (*N* = 4 in our analysis). RWSIP takes into account the relative contribution (eigenvalue) of each PC, thereby giving larger weights to the more collective modes.

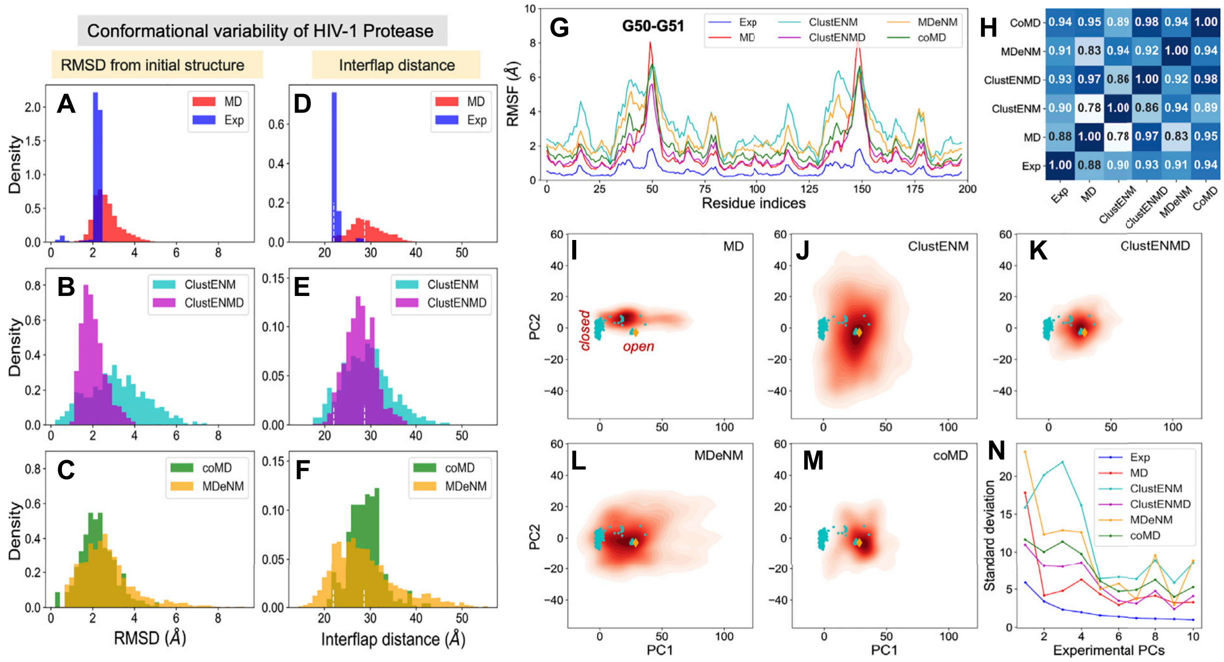


FIGURE 2 | Comparison of computational and experimental ensembles of HIV-1 PR. Histograms: **(A–C)** RMSDs with respect to the starting open structure (PDB id: 1tw7). **(D–F)** Distances between α-carbons of K55 on different subunits that monitors the opening/closure of flaps. *Dashed lines* indicate the distances of the initial open structure and a closed crystal structure (PDB id: 1bve). The distribution for each ensemble, namely experimental (*blue*), MD (*red*), ClustENM (*cyan*), ClustENMD (*magenta*), coMD (*green*), and MDeNM (*orange*), is shown on the shared x-axis. **(G)** RMSFs as a function of residue index for each ensemble, **(H)** Pearson correlation coefficients between all pairs of RMSF profiles. **(I–M)** Population distributions of ensembles of conformers generated by simulations, projected onto the subspace spanned by experimental *e*PC1 and *e*PC2, shown for **(I)** MD simulations, **(J)** ClustENM, **(K)** ClustENMD, **(L)** MDeNM, and **(M)** coMD. *Cyan circles* represent the experimental structures, and the *orange diamond* is the initial structure (PDB id: 1tw7). **(N)** Standard deviations of the conformers projected along the experimental PCs.

RESULTS AND DISCUSSION

HIV-1 Protease (PR)

HIV-1 protease (PR) is a homodimeric enzyme consisting of two symmetrically positioned monomers of 99 amino acids each, with the substrate-binding site at the interface of the monomers. The access to this site is mediated by two opposing β hairpins known as flaps (**Figure 1A**). Several studies have pointed to the significance of the coupled movements of the two PR monomers in relation to catalytic activity. Three regions recognized to be functionally important in each monomer are: 1) the N- and C-terminal residues 1–4 and 95–99, essential to dimer assembly; 2) the central region (residues 10–32 and 63–85) of each monomer containing the catalytic site, and also involved in dimerization; and 3) the highly flexible glycine–rich flaps exposed to solvent (residues 33–62) (Scott and Schiffer, 2000; Henzler-Wildman et al., 2007; Palese, 2017). The opening/closing of the flaps and the twist motion of the two monomers with respect to each other serve as collective motions that support the enzymatic function, coupled to the catalytic dyad dynamics (Batista et al., 2011; Badaya and Sasidhar, 2020).

Conformational Variability From Experiments and Computations

PR is one of the most thoroughly studied enzymes as a target for HIV-1 drug development, with over 750 structures resolved to date in different forms, in the presence of different ligands/drugs. **Figure 2A** provides information on the conformational variability of the 768 PDB structures used here as the experimental dataset. The *blue* histogram in panel **A** displays the RMSDs (based on C$^{\alpha}$ atoms) of these structures from the wide-open form [PDB id: 1tw7 (Martin et al., 2005)] used as reference, showing that the crystallographic structures are rather narrowly distributed (within 2.3 Å RMSD). The ensemble of conformations sampled during MD simulations (*red histogram*) exhibits a broader distribution (up to 4.2 Å), comparable to those generated by ClustENMD and CoMD but narrower than those generated by ClustENM and MDeNM (**Figures 2B,C**).

As mentioned above, flap opening is required for the substrate to access the active site, and its closure for proteolytic cleavage to occur. The degree of opening of the flaps can be evaluated through the distance between the C$^{\alpha}$ atom of residue K55 in the two monomers (**Figure 1A**). **Figures 2D–F** shows distributions of this distance obtained for each ensemble. The *white vertical dashed lines* indicate the distances corresponding to the open and closed states of the flap, represented by the reference structure (open) and 1bve (Yamazaki et al., 1996) (closed). Most of the experimentally resolved structures are closed conformers in the presence of a bound ligand. The PR stability is increased in the bound form such that it is better protected against self-cleavage and its crystallization is facilitated. Again, ClustENM and

MDeNM sample significantly more open conformations compared to MD, while ClustENMD and coMD yield inter-flap distances comparable to those sampled in MD.

The conformers generated by the hybrid methods encompass both the open and closed states of the enzyme. The distributions of conformers are continuous and unimodal in MD, ClustENM and ClustENMD, while coMD and MDeNM yield distinct peaks separating the closed form (**Figures 2E,F**). Notably, the open state is more populated than the closed in all simulations, in contrast to experimental structures. This could be attributed to the fact that simulations were carried out using a wide-open unliganded, drug-resistant mutant (PDB id: 1tw7) as the initial conformation. Note that ligand binding usually favors closed conformers; whereas the unliganded PR preferentially adopts open conformers predisposed to ligand-binding.

Residue Fluctuation Profiles

While the RMSDs and inter-flap distances point to broadly distributed ensembles of conformers predicted by the hybrid methods (especially ClustENM and MDeNM), it is of interest to assess whether the residue fluctuation profiles exhibited by those ensembles differ from those observed in experiments and in MD simulations. **Figure 2G** displays the root-mean-square-fluctuations (RMSFs) of C$^{\alpha}$ atoms from their average positions in each ensemble. As expected, all computations yield higher RMSFs than those observed in the X-ray crystallographic ensemble (reflecting the restricted residue mobilities in the crystals), and ClustENM and MDeNM ensembles exhibit the highest RMSFs. However, the RMSF shapes (profiles as a function of residue index) deduced from experiments and simulations are very similar, as quantified by the pairwise Pearson correlation coefficients (**Figure 2H**). All four hybrid methods exhibit correlation coefficients higher than or equal to 0.9 with experimental data (and among themselves), showing that a robust pattern of residue fluctuations, albeit the increased amplitudes, is captured by all ensembles. We note that MD simulations yield sharp peaks around G50-G51. This region corresponds to the tips of the flaps, indicating that MD may overestimate these local motions, relative to others that move concertedly. However, the correlation between MD and experiments is still strong (0.88), and those with ClustENMD and coMD are remarkably high (≥0.95).

Conformational Landscape

The above analyses compared the conformational diversity and residue flexibilities of the ensembles. Next, we proceed to a closer inspection of the conformational space explored by each method. To this aim, we first determined the subspace spanned by the principal components *e*PC1 and *e*PC2 obtained from the PCA of known structures. The known structures projected onto this subspace are displayed in **Figures 2I–M** by the *cyan dots*, each dot representing a PDB structure. The cluster of dots on the left refers to closed structures, and the reference (open) structure is displayed by the *orange diamond*. The origin of the plot represents the "average" structure, which lies in the region occupied by the closed structures due to the predominance of closed structures in the experimental ensemble. Next, we evaluated the distribution of conformers for each computationally generated ensemble, projected onto the same subspace. These distributions are displayed by contour plots (*orange-to-red shades*) in **Figures 2I–M**. The shading/levels get darker as the population density increases. The color-coded contour plots exhibit features consistent with the RMSDs in panels **A**-**C**.

The subspace spanned by the experimentally derived *e*PC1 and *e*PC2 provides only a partial view of the spread of conformers generated by computations, as some of the conformers may be broadly dispersed along other *e*PCs. As a measure of the variance of computationally generated conformers along additional *e*PCs, we evaluated the standard deviation of the distribution of each computed ensemble of conformers projected along the first 10 *e*PCs. The results are presented in **Figure 2N**. Highest variations are observed along *e*PC1 followed by either *e*PC3 (ClustENM, MDeNM and coMD) or *e*PC4 (ClustENMD and MD) for all computed ensembles, while the variations along higher modes drop sharply in all cases. These results suggest that the first four *e*PCs are sufficient to describe to a good approximation the diversity of experimentally resolved structures as well as a divergence in the computed conformers (e.g., by ClustENM) with respect to experiments.

Comparison of Global Modes/Principal Directions of Motion

Given the important contribution of the top four PCs, we carried out a detailed comparison of the overlap between *e*PC1-4 from experiments, and *s*PC1-4 from each simulation (MD and four hybrid methods). The heatmap in **Supplementary Figure S1** provides information on the overlap between these six sets of PCs, organized in a super-matrix of 6 × 6 blocks. Each block (4 × 4 matrix) describes the correlation cosine between the top four PCs corresponding to a pair of ensembles. This way, one can trace back the similarities in the observed conformational heterogeneities to similarities between top-ranking PCs. The *bottom row* shows that *e*PC1 strongly correlates with *s*PC1 from MD and MDeNM (with respective correlation cosines of 0.84 and 0.76), and with *s*PC2 from coMD (0.73). As to the *e*PC2, we note its high correlations with ClustENM *s*PC2 (0.71) and MDeNM *s*PC3 (0.73). Likewise, the second *block-row* from bottom shows the moderate correlations between MD and hybrid methods, and the top four block-rows show strong correlations between the hybrid methods. Thus, even though the order of the PCs may differ, the six ensembles of structures/conformations exhibit equivalent pairs of PCs which predominantly define the observed distributions of conformers. We have furthermore evaluated the RWSIP values [**Eq. (1)**] as an additional metric for comparison. All hybrid methods as well as MD simulations yield satisfactory correlation (varying as 0.62 ≤ RWSIP ≤0.83) with the experimental ensemble (**Table 3**) in line with a previous study showing close correspondence between NMs and ePCs for this system (Yang et al., 2008).

The structural variations described by the first four *e*PCs are schematically described by the color-coded ribbon diagrams in the upper panel of **Figure 3**. The lower panel displays the first four *s*PCs generated by MDeNM. The *s*PCs are reordered to highlight (*by boxes*) the *s*PCs equivalent to the *e*PCs. Notably, *e*PC3 also shows a high correlation (0.75) with MDeNM *s*PC2 and exhibits a bending of the whole structure where the flaps

TABLE 3 | Comparative assessment of the performance of hybrid methods.

Protein	ClustENM	ClustENMD	MDeNM	coMD	MD	Experimental
Closest distance of approach (Å) (Figure 2, Figure 4, Figure 6, and Figure 8; panels D–F)[a]						
PR (K55-K55)_	**17.7**	**17.6**	**14.3**	**20.8**	**21.0**	21.5
TIM (G174-Y211)	**13.8**	**12.0**	14.9	**13.4**	**13.4**	12.9
PGK (P66-M311)	**17.4**	**20.4**	**21.5**	**18.1**	24.9	23.6
RT (thumb-fingers)	28.8	29.9	**26.8**	33.4	39.2	26.9
Minimum RMSD (Å) from the closed structure[b]						
Initial structure[c]	Attained in simulations, starting from the open state					Initial RMSD (exp)[d]
PR (1tw7)	2.2	**1.7**	**1.8**	2.3	**<2**	2.7 (1bve)
PGK (2xe7)	**1.4**	**1.7**	2.9	2.5	3.2	3.6 (2wzb)
RT (2b6a)	3.5	3.8	4.3	3.8	4.7	5.2 (3kli)
RWSIP[e] **with respect to experimental PCs**						
PR	0.62	0.75	0.77	0.72	**0.83**	1.00
TIM	0.54	**0.58**	0.55	0.54	0.52	1.00
PGK	0.87	**0.89**	0.78	0.78	0.74	1.00
RT	**0.72**	0.70	0.56	0.56	0.43	1.00
Average	0.69	**0.73**	0.66	0.64	0.63	—

[a]*Those entries with Δd < 1 Å [where Δd = d (comp)—d (exp)] are shown in boldface.*

[b]*Minimum RMSD, of each ensemble with respect to the closed structure reflects the extent of approach from open-to-closed state (those values below 2.0 Å are highlighted in bold). TIM is not included as experimental structures do not show a global transition between open and closed states but just loop motions (RMSD within ~ 1 Å RMSD).*

[c]*PDB id for open structure is in brackets.*

[d]*PDB id for closed structure is in brackets.*

[e]*The highest (best) RWSIP value observed for each protein is highlighted in bold.*

move together. These functional movements along MDeNM *s*PC1-3 can be viewed in **Supplementary Movie S1**. Earlier studies have shown that the first two collective modes of PR describe internal movements allowing for substrate binding and catalysis (Scott and Schiffer, 2000; Batista et al., 2011; Palese, 2017). We see here that the first two modes, *e*PC1/*s*PC1 and *e*PC2/*s*PC3 capture these flap opening/closure and twisting events, as well as the coupled twisting and bending of the monomers; and *s*PC3 accounts for inter-subunit counter-rotation.

Triosephosphate Isomerase

TIM is a homodimeric enzyme, each subunit adopting a TIM barrel fold (**Figure 1B**). It plays a key role in the glycolytic pathway catalyzing the interconversion between two triose phosphate sugars, dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate. The active sites are located at the C-terminal end of each β-barrel. A crucial feature of TIM functional dynamics is the catalytic loop opening/closure on each subunit. Catalysis takes place when the loop is closed protecting the active site from solvent exposure. Loop closure is not ligand-gated, i.e., it takes place in the apo state as well (Williams and McDermott, 1995; Cansu and Doruker, 2008).

Conformational Variability From Experiments and Computations

In contrast to the other examined proteins, homologous TIM structures with ≥90% sequence identity to the Trypanosoma cruzi

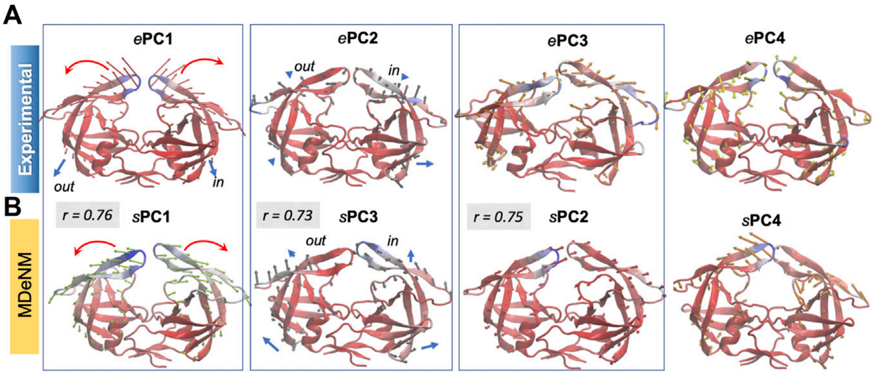


FIGURE 3 | First four principal modes for HIV-1 PR ensembles. The **(A,B)** are based on the PCAs of the experimental and MDeNM ensembles, respectively. Those PCs that exhibit high (>0.70) correlations are enclosed in boxes. See the corresponding movies in **Supplementary Movie S1**.

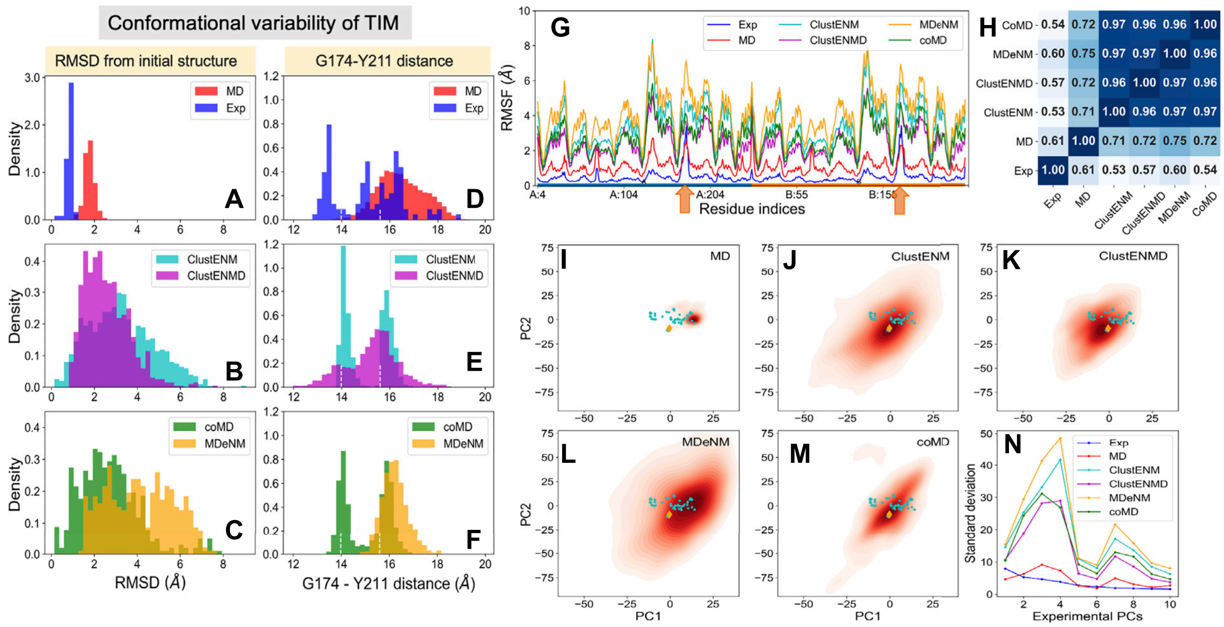


FIGURE 4 | Comparison of the ensembles of conformers experimentally observed and computationally generated for TIM. The panels are in the same format as **Figure 2**. **(A–C)** RMSD histograms with respect to the starting apo structure (PDB id: 1tcd). **(D–F)** Distances between α-carbons of G174 and Y211 that monitors the opening-closing of the catalytic loop. *Dashed lines* indicate the loop distances in the different subunits of starting dimeric structure. **(G)** RMSFs with respect to the initial structure for each ensemble. *Orange arrows* indicate the catalytic loop position in each subunit. **(H)** Pearson correlation coefficients between pairs of RMSF profiles. **(I–M)** Population distributions of computed ensembles (labeled in each panel) projected onto subspace spanned by *e*PC1 and *e*PC2. **(N)** Standard deviations of the conformers projected along ePCs.

structure used here (PDB id: 1tcd) yielded a small set that closely retained the same structure. To increase structural diversity, we have relaxed the threshold sequence identity to 50%, which led to an ensemble of 57 resolved structures for TIM homologs. The *blue* histogram in **Figure 4A** displays their distributions (RMSDs) with respect to the starting conformer [PDB id: 1tcd (Maldonado et al., 1998); **Table 1**]. MD simulations also exhibited a narrowly distributed RMSD histogram (panel A; *red histogram*) while the hybrid methods (panels B–C; labeled) yielded substantially higher RMSDs, pointing to the ability of these methods to sample a broader conformational space, as already seen for HIV-1 PR. Yet, given the high stability of the TIM fold, the relatively lower RMSDs predicted by ClustENMD and coMD could be more realistic.

The catalytic loop motion of TIM can be monitored by the distance change between the C^{α} atoms of G174 (tip residue of loop 6) and Y211 (a relatively immobile residue on the barrel used as reference) (**Figure 1B**). Our previous MD simulations (Kurkcuoglu and Doruker, 2013; Kurkcuoglu et al., 2015) indicated multiple opening/closing events between this pair of residues. **Figures 4D–F** shows the distributions of this distance for both subunits. Experiments show a multimodal distribution varying over a broad range (13–19 Å), with the lower values corresponding to the closed loop. In the reference crystal structure (PDB id: 1tcd), the distances are 14.0 and 15.6 Å for A and B monomers, respectively, shown by *white vertical dashed lines* in each panel. ClustENM and coMD exhibit bimodal distributions with peaks localized around these values. Relaxation by heating up in ClustENMD enhances loop flexibility leading to a broader distribution (panel **E**; *magenta*). MD and MDeNM, on the other hand favor the open state only, missing the closed state of the loop observed in experiments and other hybrid methods.

Residue Fluctuation Profiles

The RMSFs and their Pearson correlation coefficients are presented in the respective panels G and H of **Figure 4**. Consistent with RMSDs, the conformers generated by hybrid methods, and especially MDeNM and ClustENM display higher RMSFs compared to those observed in MD simulations and experiments. The two *orange arrows* along the abscissa in **Figure 4G** indicate the location of the catalytic loops. Thes showed the highest conformational diversity in experiments (*blue curve*). All hybrid methods display similar profiles, but their correlations with experimental ensembles, which vary in the range 0.53–0.60, are much lower than that (0.90–0.94) observed for HIV-1 PR. Their correlations with MD vary from 0.71 to 0.75. The hybrid methods consistently show very high correlations among themselves (>0.95), suggesting that they robustly sample similar motions, beyond those observed in X-ray crystals as will be further elaborated below.

Conformational Landscape

Figures 4I–M display the loci of the 57 experimentally resolved structures (*cyan dots*) in the reduced space spanned by *e*PC1 and *e*PC2, and the distribution of computationally predicted

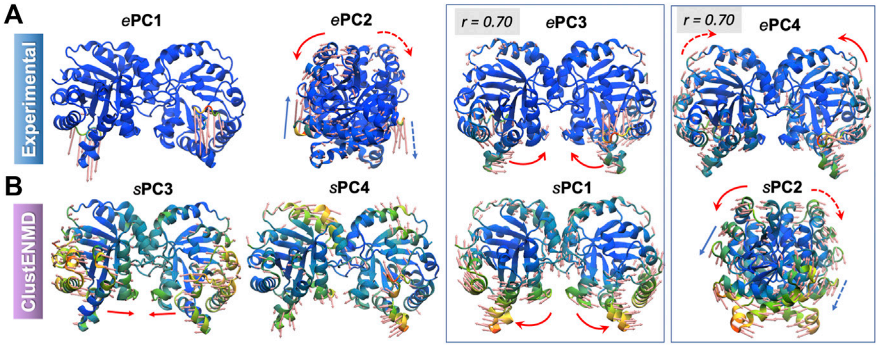


FIGURE 5 | First four principal modes for TIM ensembles. The **(A,B)** refer to experimental and ClustENMD ensembles, respectively. Those PCs that exhibit high (>0.70) correlations are enclosed in boxes. See the corresponding movies in **Supplementary Movie S2**.

ensembles are displayed by the color-coded KDE plots. The origin of each plot corresponds to the mean experimental structure, which lies in proximity to the reference structure (PDB id: 1tcd, *orange diamond*). The experimental structures (*cyan dots*) and MD simulations cover a limited portion of this subspace compared to the hybrid methods with a clear shift of the MD distribution relative to the experimental one in line with the RMSDs and loop distances. **Figure 4N** describes the distributions of the computationally generated conformers along the first 10 *e*PCs. Higher variations in conformations are observed along the *s*PCs 3 and 4 for all hybrid methods. Therefore, the first two experimental PCs, *e*PC1 and *e*PC2, are not sufficient to account for the diversity of the conformations sampled by hybrid methods.

Comparison of Global Modes/Principal Directions of Motion

Supplementary Figure S2 shows the overlaps between the top four PCs for each pair of ensembles. The *s*PCs of the hybrid methods are in close agreement with each other, and also in accord with the first three *s*PCs from MD. As discussed above, the first two *e*PCs do not show significant correlation with the *s*PCs (*bottom two rows*), whereas higher overlaps are evident between *e*PCs 3–4 and *s*PCs 1–2. Not surprisingly, RWSIP values are generally relatively low (0.53–0.58) for this enzyme. Closer examination shows that, *e*PC1 primarily reflects the catalytic loop opening/closure (**Figure 5**), also evident from the large distance change in the loop shown in **Figure 4D**. As such, it is a *local* motion, and it is not among the *s*PC1-4 that the hybrid methods yield (top-ranking *s*PCs usually describe cooperative motions that engage the *entire* protein). *e*PC2, on the other hand, refers to loop motions coupled to relatively more collective motions and shows slightly higher, but still weak, correlations with *s*PCs. In contrast, *e*PC3 and *e*PC4 do exhibit notable correlations with *s*PC1 or *s*PC2 from all simulations. These motions correspond to the counter-rotation and bending of the subunits with respect to each other, coupled to the catalytic loop dynamics, as illustrated in **Figure 5** and **Supplementary Movie S2**. These motions have been identified as the global modes that define the enzyme's putative functional motions (Kurkcuoglu et al., 2006; Cansu and Doruker, 2008; Kurkcuoglu et al., 2015). Notably, *s*PC3 is another highly cooperative motion where the two monomers concertedly bend around an axis perpendicular to that of *s*PC1; and *s*PC4 exhibits a counter-twisting and breathing of the two barrels (**Figure 5**). Overall, hybrid methods point to a broad range of cooperative rearrangements, which cannot be readily discerned upon PCA of structures resolved for TIM homologues which yields either local loop motions (ePC1-2) or highly constrained (small amplitude) global motions (*e*PC3-4).

3-Phosphoglycerate Kinase

PGK is another key glycolytic enzyme, catalyzing the phosphotransfer between 1,3-bisphosphoglycerate (bPG) and ADP. It is a monomeric protein composed of two domains of approximately equal size. The bPG binding site is located on the N-domain, while ADP binds to the C-domain (**Figure 1D**). During its function, the enzyme undergoes a large hinge-bending conformational change bringing the bound substrates into proximity such that the reaction can happen (Palmai et al., 2009; Palmai et al., 2014). The open crystal structure in complex with 3-phosphoglyceric acid (3 PG) and ADP [PDB id: 2xe7 (Zerrad et al., 2011)] is used here as the reference structure for initiating the computations.

Conformational Variability From Experiments and Computations

We considered 35 experimentally resolved structures for PGK, with sequence identity above 90%. **Figures 6A–C** shows the RMSDs of the different conformational ensembles, including the ensemble of experimentally resolved structures and conformers from MD simulations (panel **A**), and those generated by hybrid methods (panels **B–C**) with respect to the initial open structure. The conformers were superposed onto the mean experimental structure using the C^{α} coordinates in both domains. There are two separate groups of experimentally resolved structures (**Figure 6A**), with the lower RMSD group corresponding to the open structures, and that centered around 3.7 Å corresponding to closed structures. All hybrid methods

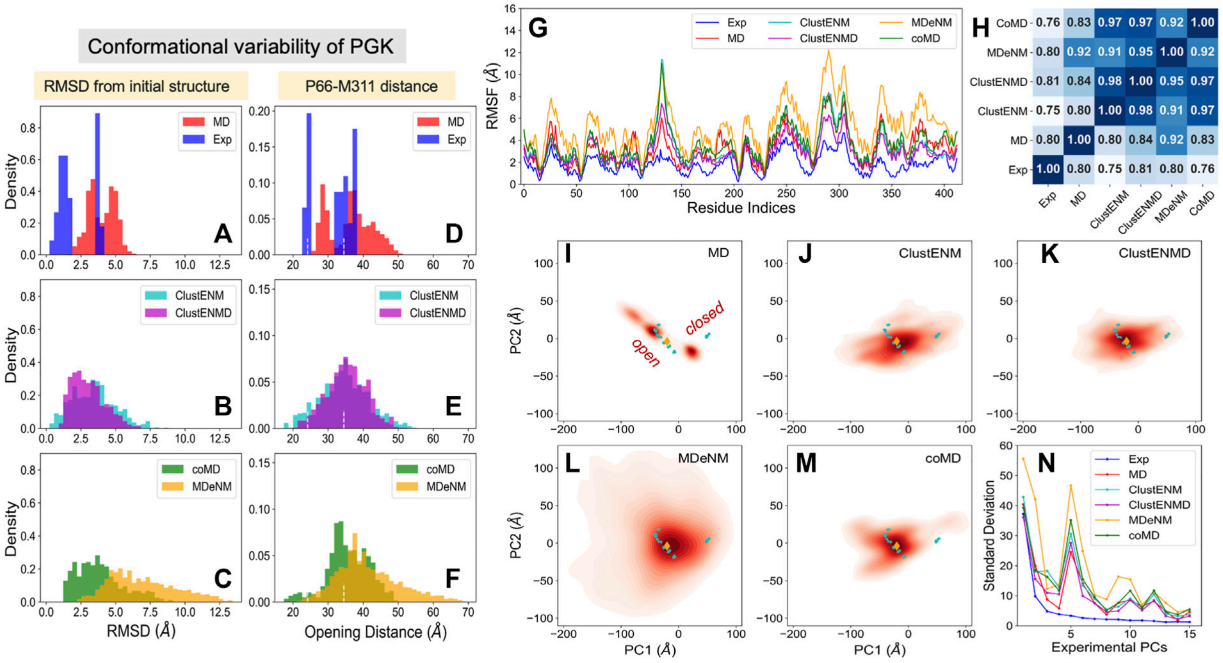


FIGURE 6 | Comparison of computational and experimental ensembles of PGK. The panels are in the same format as **Figure 2 (A–C**) RMSDs with respect to the starting open structure (PDB id: 2xe7). **(D–F)** Distances between the α-carbons of P66 and M311 that monitors the opening-closing motion of PGK. Dashed lines indicate the distances of the initial open, and a catalytically fully closed crystal structure (PDB id: 2wzb). **(G)** RMSFs for each ensemble**, (H)** Pearson correlation coefficients between all pairs of RMSF profiles. **(I–M)** Population distributions of ensembles of conformers generated by simulations, projected onto the subspace spanned by *e*PC1 and *e*PC2, shown for **(I)** MD simulations, **(J)** ClustENM, **(K)** ClustENMD, **(L)** MDeNM, and **(M)** coMD. **(N**) Standard deviations of the conformers projected along the experimental PCs.

yielded wide unimodal distributions for RMSDs in contrast to the bimodal distributions exhibited by experimental structures and MD conformers. Larger RMSD regions correspond to further opening/relaxation of the protein, beyond that observed in the crystals and/or accessed in MD simulations.

The opening-closing motion of PGK can be monitored through the distance between the α-carbons of P66 and M311, two residues located at the tips of the N- and C-domains in the vicinity of the ligands. **Figures 6D–F** provides the distributions of the interdomain distance probed by these two residues. For reference, the distance in the initial open structure (34.5 Å) and that assumed in a catalytically active fully closed crystal structure (PDB id: 2wzb (Cliff et al., 2010), 24.2 Å) are indicated by the *white dashed lines*. **Figure 6D** clearly distinguishes between the closed and open experimental structures. The MD conformers exhibit further opening as well as moderate closing of PGK but do not cover the region of fully closed experimental structures. On the other hand, all hybrid methods successfully detect the fully closed region albeit to different extents. In contrast to the other simulation methods, MD, coMD, and MDeNM included ADP and 3 PG in the binding pocket. This hindered the sampling of conformations beyond the fully closed experimental structures (under 23 Å), while ClustENM and ClustENMD suggested that the interdomain distance could become lower than 20 Å. CoMD, which included the ligands in the TMD and EM stages but not in the C^{α}-based ANM for the NMA, still sampled these conformations to a small extent.

Residue Fluctuation Profiles

The RMSF in α-carbons with respect to their mean positions are presented in **Figure 6G**. Like HIV-1 PR and TIM, computations yielded higher fluctuations compared to experiments. In contrast to **Figure 4G** for TIM, the MD-generated ensemble for PGK exhibited RMSFs falling within the same range as for the hybrid methods. This could be due to the bimodal distribution of the opening distances (**Figure 6D**) displayed by the three independent MD runs. In contrast, the hybrid methods showed broad unimodal distributions (**Figures 6E,F**). Yet, the predicted RMSF profiles (**Figure 6G**) remained comparable to that obtained by MD. We note that MDeNM (*yellow curve*) yielded the largest fluctuations, consistent with the sampling of widely open conformers (see the corresponding long tails in panels **C** and **F**); however, as shown in **Figure 6G**, the overall profile indicated by experiments and MD simulations were robustly reproduced by all hybrid methods. The pairwise Pearson correlation coefficients presented in **Figure 6H** show that all hybrid methods exhibited a fairly strong correlation among themselves (varying from 0.91 to 0.98), similar to the results observed in HIV-1 PR and TIM. This time, we also observe a relatively strong correlation between the hybrid methods and experiments (0.75–0.81) and MD runs (0.80–0.92). MDeNM

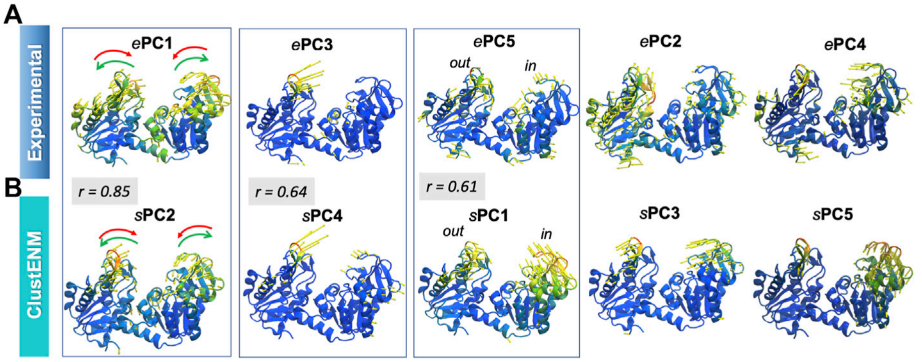


FIGURE 7 | First five principal modes for PGK ensembles. The **(A,B)** are based on PCA of experimental and ClustENM ensembles, respectively. Those pairs of PCs that exhibit relatively high (>0.60) correlations are enclosed in boxes. See the corresponding **Supplementary Movie S3**.

exhibits a remarkably high correlation (0.92) with MD revealing that the relative flexibilities of the residues are accurately accounted for, even though the absolute sizes of the motions (uniformly and significantly) differ.

Conformational Landscape

Figures 6I–M display the projections of the computationally generated conformers onto the reduced space spanned by *e*PC1 and *e*PC2. Two distinct sets of experimentally resolved structures (*cyan dots*) are discerned: one corresponding to open structures (including the reference structure denoted by the *orange diamond*), the other to closed structures. MD simulations starting from the open structure could not sample the region of closed conformations, and only partially covered the space sampled by the open experimental structures. All hybrid methods, on the other hand, covered the space occupied by both groups of structures. Considerable opening of PGK is visible in ClustENM-generated conformers and an even greater opening is predicted by MDeNM.

Comparison of Global Modes/Principal Directions of Motion

Figure 6N describes the standard deviation (or square root of variance) of the conformers along the first 15 *e*PCs. All computationally generated ensembles show a wide dispersion along *e*PC1 and *e*PC5, with *e*PC5 showing a clear peak. **Supplementary Figure S3** provides the overlap matrices (first five modes) among each pair of ensembles. The *bottom row* clearly shows that the *s*PC2 predicted by all four hybrid methods strongly correlate with *e*PC1, with correlation cosines ranging from 0.76 (coMD) to 0.89 (MDeNM), hence the broad dispersion of the computationally predicted conformers along *e*PC1 (driven by their *s*PC2). The high variance of predicted conformers along *e*PC5, on the other hand, apparently originates from the overlap with the first mode, *s*PC1, predicted by all hybrid methods. The overlaps are moderate in this case, varying from 0.52 for MDeNM to 0.61 for ClustENM. Notably, the *s*PCs predicted by MD show the weakest correlations with experiments among all computational methods, but still exhibit modest overlaps with *e*PC1 and *e*PC5. Likewise, MD shows the lowest RWSIP in **Table 3** while ClustENM and ClustENMD show the highest.

Figure 7 illustrates, using color-coded ribbon diagrams and arrows, the first five ePCs (*upper panel*) and *s*PCs from ClustENM (*lower panel*). Both experiments and ClustENM describe the opening-closing (hinge-bending) motion as well as some breathing motions of the two domains. *e*PC1, counterpart of *s*PC2 as discussed above, represents the hinge-bending motion mediated by the interdomain helix. *e*PC3 corresponds to large conformational changes at the loop (residues 130–140) located at the tip of the N-domain, also expressed in large RMSF values in **Figure 6G**. Notably ClustENM *s*PC4 approximates the same movement (with a correlation cosine 0.64), also shown in **Figure 7**. ClustENM *s*PC1 and its approximate counterpart *e*PC5 induce out-of-plane motions, inward and outward, in the two respective domains, even though the size of motions along *e*PC5 is smaller. **Supplementary Movie S3** illustrates the ClustENM *s*PC1, *s*PC2 and *s*PC4, as the three functional mechanisms of motions accessible to PGK.

HIV-1 Reverse Transcriptase

Like HIV-1 protease, RT has been, and continues to be, an important target for HIV-1 drug discovery (Esposito et al., 2012; Gu et al., 2020). RT is a heterodimer composed of subunits p66 and p51 (**Figure 1C**). The p66 subunit contains the DNA polymerase and RNase H domains, thus performing dual enzymatic activity, while the p51 subunit serves as a scaffold. The DNA polymerase domain itself consists of four subdomains: fingers, thumb, palm, and connection. The former two are distinguished by their high mobility required to bind the nucleotide oligomer; the palm serves as a hinge center, and the connection forms the base connecting to the p51 subunit. Nucleoside/nucleotide RT inhibitors (NRTIs) were the first class of antiretroviral drugs approved for therapeutic use, followed by non-nucleoside/nucleotide RT inhibitors (NNRTIs). Most NNRTIs (Namasivayam et al., 2019) bind a pocket at the palm interface with the thumb or fingers, impairing the hinge movements of these two subdomains essential to polymerase activity. Moreover, inhibitors have been designed that control the global movements of the RNAse H (Ilina et al., 2012), and even

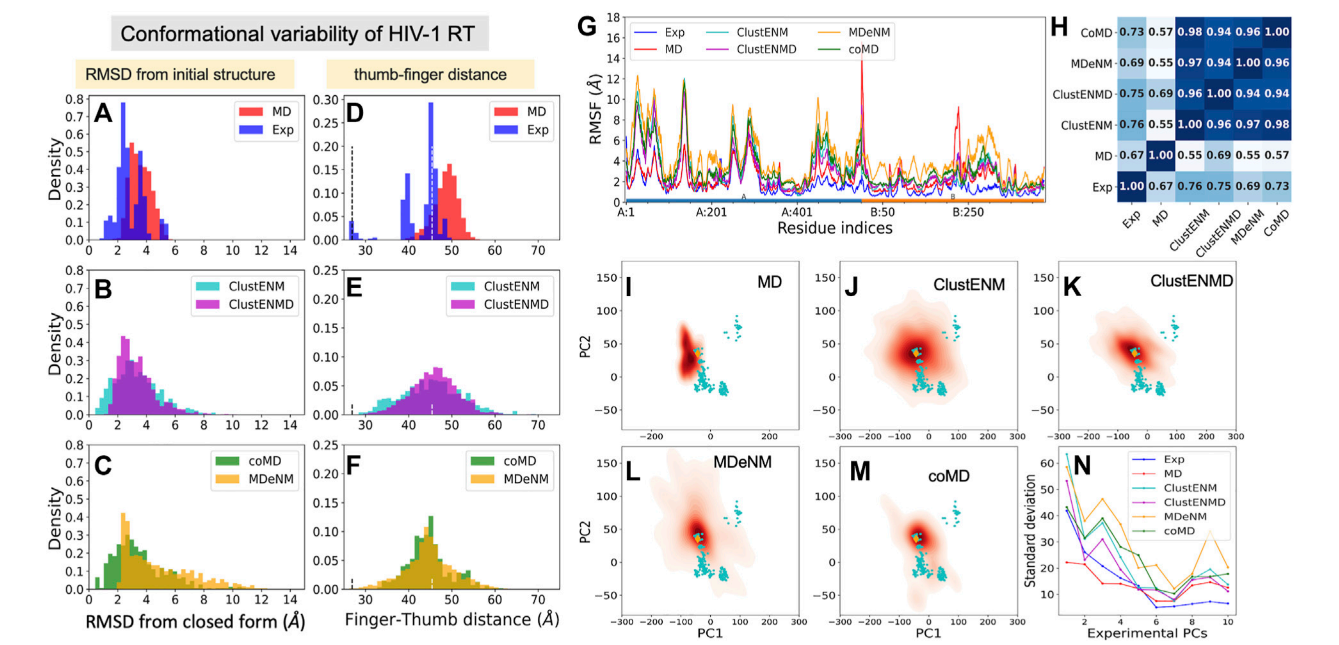


FIGURE 8 | Comparison of computational and experimental ensembles of HIV-1 RT. **(A–C)** RMSDs with respect to the starting closed ligand-free structure (PDB id: 3kli). **(D–F)** Distances between the centers of mass of the fingers (residues 1–85 to 112–155) and thumb (residues 242–312) subdomains on the p66 subunit (chain A) that describe open-to-closed fluctuations of the two subdomains. *Dashed lines* indicate the distances in the open and closed reference structures (PDB id: 2b6a and 3kli, respectively). **(G)** RMSFs for each ensemble. The *bars* underneath indicate the p66 subunit (chain A; *blue*) and the p51 subunit (chain B; *orange*). **(H)** Pearson correlation coefficients between pairs of RMSF profiles. **(I–M)** Population distributions for ensembles generated by **(I)** MD, **(J)** ClustENM, **(K)** ClustENMD, **(L)** MDeNM, and **(M)** coMD. **(N)** Standard deviations of the conformers projected along the experimental PCs.

those having dual actions on both enzymatic activity exist (Corona et al., 2016).

Conformational Variability From Experiments and Computations

We used as reference a complex with a NNRTI (THR-50) [PDB id: 2b6a (Morningstar et al., 2007)], **Figure 1C**, which represents an *open* form of the fingers-thumbs subdomain of RT. **Figures 8A–C** displays the RMSDs with respect to the closed reference structure [PDB id: 3kli (Tu et al., 2010)] observed in experiments and computations. The resolved structures (365 included here) show a conformational variability (1 ≤ RMSD ≤ 6 Å) wider than those of the other three proteins studied, and the hybrid methods show even broader distributions. MD simulations show the narrowest distributions (2 ≤ RMSD ≤ 6 Å), unable to sample the closed forms. MDeNM shows the highest RMSDs (up to 12 Å) but cannot sample the closed forms with RMSD < 2 Å while ClustENM and coMD satisfactorily sample both closed and open forms (**Figures 8A–C**). The ability of ENM-based hybrid methods to sample the broad range of subdomain and domain rearrangements of RT originates from the ability of ENMs to describe the RT global dynamics (Bahar et al., 1999; Sluis-Cremer et al., 2004).

Toward understanding the origin of these large RMSDs, we examined the distance between the mass centers of the thumb and fingers subdomains (**Figures 8D–F**), which is a determinant of conformational variability. The *vertical dashed lines* indicate the distances corresponding to the closed and open states (d = 26.9 and 45.4 Å, respectively). As noted above, MD conformers only sample open conformers (within 40 ≤ d ≤55 Å), whereas the hybrid methods exhibit broader distributions encompassing a wide distribution of thumb-finder distances.

Residue Fluctuation Profiles

Residue fluctuation profiles are presented in **Figure 8G**, along with their Pearson correlation coefficients in panel H. Despite their large RMSFs, RMSF profiles of all hybrid methods as well as MD simulations and experiments show close similarities. The correlations of the hybrid methods with experiments vary from 0.69 (MDeNM) to 0.76 (ClustENM), while those with MD are lower (0.55–0.69). This, and the lower correlation between MD and experiments (0.67), indicates that the large movements undergone by the experimental structures adhere to the intrinsic dynamics of RT constrained by its overall fold topology as predicted by hybrid methods, while MD simulations of 200 ns fall short of an adequate sampling of conformational space for this large (1,000 residues) protein.

Conformational Landscape

The results are shown in **Figures 8I–N**, in the same format as before. Hybrid methods show sampling power superior to that of MD: ClustENM (panel **J**), coMD (panel **M**) and MDeNM (panel **L**) cover a space large enough to include a significant share of experimental structures (*cyan dots*) in contrast to MD. Notably, the space sampled by ClustENM encompasses almost all experimental structures projected onto the subspace spanned by *e*PC1 and 2.

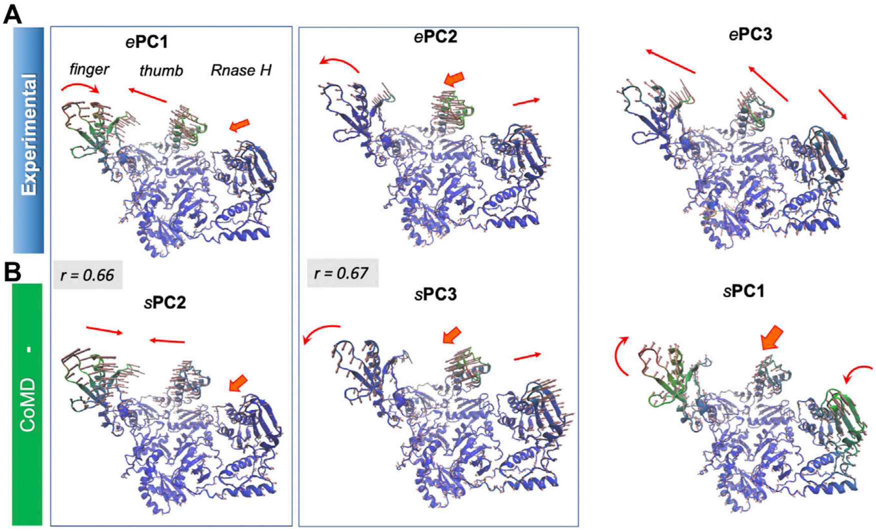


FIGURE 9 | First three principal modes for HIV-1 RT ensembles. The **(A,B)** are based on PCA of experimental and coMD ensembles, respectively. Pairs that exhibit relatively high (>0.65) correlations are enclosed in boxes. See the corresponding **Supplementary Movie S4**.

Comparison of Global Modes/Principal Directions of Motion

Figure 8N describes the standard deviation (or square root of variance) of the conformers along the first 10 *e*PCs. The computationally generated ensembles excluding MD show a wide dispersion along *e*PC1 to *e*PC4, supporting the wide diversity of the generated structures. ClustENM and MDeNM standout as the two hybrid methods that yield the largest dispersion of conformers along *e*PC1-2. ClustENM also had the highest RWSIP value (0.72; **Table 3**), while MD yields the lowest (0.43).

Supplementary Figure S4 provides the overlap matrix between the top four PCs for all pairs of conformational ensembles. The *bottom row* shows that *e*PC1 correlates with the *s*PC2 of all four hybrid methods, with correlation cosines varying from 0.59 (MDeNM) to 0.66 (coMD and ClustENM), in addition to *s*PC1 of ClustENMD (0.58) and ClustENM (0.56) and *s*PC3 from MDeNM (0.57). **Figure 9** and **Supplementary Movie S4** show that this PC describes the opening-closing of the thumb and finger subdomains with respect to each other, accompanied by concerted reorientation of RNase H domain. Likewise, *e*PC2 shows a good correlation with *s*PC3 (e.g. 0.67 for coMD). In this case, the finger and thumb subdomains of the DNA polymerase domain undergo anticorrelated movements with respect to RNAse H (**Figure 9**). Notably, the computationally predicted first mode of motion (*s*PC1), which is in remarkably strong agreement between all four hybrid methods (correlation cosines >0.90), is not accounted for by *e*PC1-4. This essential motion (relative movements of the fingers and RNase H accompanied by out-of-plane movements of the thumb), also supported by MD, is also illustrated in **Supplementary Movie S4**.

CONCLUSION

Recent years have seen an increase in the number and complexity of hybrid methods developed for investigating the conformational space accessible to proteins (Krieger et al., 2020), and especially to complexes or multimeric proteins. This has been accompanying the advances in experimental technologies that allow for the elucidation of multiple conformers, and the increasing need to map the accessible conformational space toward elucidating the mechanisms of function. While such hybrid methods appear to provide tools for exploring large-scale conformational motions at atomic resolution, it is of interest to assess their limitations as well as their advantages in a comparative analysis. Here we focused on four such methods and used as benchmarks sets of experimentally resolved structures and MD-sampled conformers for four well-studied proteins. Our analysis simultaneously revealed that these two latter sets suffer from limitations themselves, as discussed below. Overall, six ensembles of conformers were compared in each case, including those observed experimentally, simulated by MD, and predicted by hybrid methods. The analysis used four criteria/metrics, and the performance of the methods vis-à-vis each metric is discussed below.

The overall breadth of conformational space predicted by hybrid methods is significantly larger than that observed in X-ray structures or sampled by MD simulations. In all cases, the RMSDs of the conformers generated by the hybrid methods exhibited a much broader distribution than those experimentally observed, as illustrated in panels **A-C** of **Figure 2**, **Figure 4**, **Figure 6**, and **Figure 8**. This is most striking in the TIM analysis. Despite the inclusion of TIM sequence homologs with >50% sequence identity, the maximum RMSD between these 57 structures remained 1.1 Å (**Figure 4A**). Thus, the crystal structures resolved for this dimeric enzyme exhibit minimal global conformational variability, which is presumably partly due to constraints in the crystal environment and partly to the particular highly stable α/β-barrel fold. Local functional changes evidenced by large fluctuations in inter-residue distances (G174-Y211) are observed in this case with minimal domain/monomer movements. The RMSD of conformers predicted by

ClustENMD and coMD remained generally lower than 4 Å with respect to the initial structure, which is plausible for a dimeric enzyme of ~500 residues. MDeNM and ClustENM, on the other hand, led to up to 8 Å RMSDs, and it remains to be seen if such large conformational changes are accessible to TIM family members. In contrast, the X-ray structures for RT showed a much broader variation (up to 6 Å RMSD), and all hybrid methods satisfactorily reproduced the breadth of conformational variation (**Figures 8A–C**).

Hybrid methods predict conformers that comply with functional changes in conformations. For each studied system, we selected specific distances that probe functional movements (e.g. flap opening/closure in PR, thumb-finger distance in RT, catalytic loop motion in TIM, interdomain distance in PGK) and investigated whether hybrid methods could produce conformers consistent with experimentally detected changes (**Figure 2**, **Figure 4**, **Figure 6** and **Figure 8**; panels D–F). The challenge in most simulations is to capture closed conformations, or the closest distances, which are not entropically favored. **Table 3** lists these closest distances of approach for residue pairs selected to reflect functional movements (*first column*), observed in computations and experiments. Except for RT, the hybrid methods yielded conformers that satisfied the closest distances of approach, even exceeding in some cases the closest distances observed in experiments. In the case of RT, the closest thumb-fingers distance observed in experiments is captured by MDeNM, and approached by ClustENM and ClustENMD but not by coMD and MD. We note that MD also failed to sample the closest interdomain distances in PGK; and MDeNM could sample only the extended forms of TIM catalytic loop.

As another metric we examined whether the closed form could be attained when initiating the simulations from the open form. RMSDs between the reference open and closed forms of PR, PGK and RT were 2.7, 3.6 and 5.2A, respectively. Hybrid methods demonstrated a substantially higher ability to attain the closed state compared to MD simulations (**Table 3**).

Residue fluctuations (RMSFs) exhibit robust profiles, despite significant (uniformly distributed) changes in the overall sizes of motions. A striking observation repeatedly observed in all four proteins and quantified by Pearson correlations (of >0.86) was the robustness of RMSF profiles predicted by all four hybrid methods, despite their differences in the absolute RMSFs (**Figure 2**, **Figure 4**, **Figure 6** and **Figure 8**, panels G-H). Even more interesting was their strong correlation with the RMSF profiles extracted from aligned experimental structures, despite the significant suppression of fluctuations in crystal structures. For HIV-1 PR, the correlations between experimental RMSFs and those predicted by hybrid methods fell in the range 0.92 ± 0.02; for PGK and RT, they vary as 0.78 ± 0.03 and 0.73 ± 0.03. In contrast, TIM exhibited significantly lower correlations (0.56 ± 0.04). As pointed out above, the ensembles of structures resolved for TIM are very narrowly distributed, and so are the residue RMSFs. The RMSFs extracted from these highly similar crystal structures may not reflect the full conformational spectrum.

The correlations between RMSF profiles predicted by hybrid methods and MD simulations, varied in the ranges 0.88 ± 0.09, 0.74 ± 0.02, 0.86 ± 0.06 for PR, TIM and PGK, respectively, while that of RT was much lower (0.59 ± 0.09). Given that hybrid results gave a significantly higher correlation with experiments, this low correlation indicates the sampling inaccuracy of MD (of 100 s of nanoseconds) for this protein of 1,000 residues. Finally, the comparison of the Pearson correlations between experimental and simulated RMSFs showed that ClustENMD and MDeNM simulations achieved slightly higher performances (0.77 and 0.75 respectively, averaged over the four case studies), followed by coMD, MD and ClustENM which showed comparable performance (~0.74).

Closer look at principal changes in conformation points to the conservation of dominant modes of motion, supported by both experiments and computations. Dissection of the spectrum of collective motions upon PCA of the generated conformers in each ensemble revealed close similarities, as shown in the overlap matrices presented in **Supplementary Figures S1–S4** for top-ranking 4 PCs (or 5 for PGK). Given that PCs usually define cooperative mechanisms relevant to function, it is important to assess to what extent the *s*PCs derived from hybrid methods agree with experimental PCs (*e*PCs). Using RWSIP (Carnevale et al., 2007) as a metric (**Table 3**), we found that ClustENMD performed the best among the examined five computational methods, followed by ClustENM and MDeNM. These two analyses also allowed us to identify commonalities and divergences between methods to reveal the most salient features for each system, revealing the benefit of using multiple methods together.

We also observed significant variations of generated conformers along experimental PCs other than *e*PC1 and *e*PC2 (panel **N** in **Figure 2**, **Figure 4**, **Figure 6** and **Figure 8**). For example, TIM shows higher variations along the third and fourth *e*PCs compared to those along the first two *e*PCs in all simulations including MD. Likewise, an essential mechanism of motion of RT, robustly predicted as *s*PC1 by all computational methods and known to be essential to function (Jernigan et al., 2000; Sluis-Cremer et al., 2004) eludes the top-ranking *e*PCs. These observations show some global modes/deformations are suppressed in the X-ray structures presumably due to tight packing or symmetry requirements imposed by crystallization.

Other considerations: Optimization of parameter sets and computing efficiency. The parameters used in ClustENM, ClustENMD, and coMD were the same across all the systems, where those of MDeNM were adjusted for different proteins (**Table 2**). These parameters for the former three methods seem to perform satisfactorily for the studied proteins, except for RT in terms of reaching the closed structure (e.g., PDB id: 3kli) starting from an open form. If the extent of conformational flexibility (e.g. RMSD between endpoints) were to be known in advance, the parameters could be adjusted for more precise sampling of conformers. Obviously, generation of additional cycles enables the sampling of a broader conformational space, which may be more appropriate for larger proteins. A systematic study of the size of experimentally observed conformational space as a function of the size and packing density of protein, or different structural classes may provide guidance for selecting parameters based on the system properties. Given that the hybrid methods tested here are relatively recent, there is plenty of room for subsequent studies aiming to define optimal parameters for different systems.

To provide initial insights into the influence of these sampling parameters, we analyzed the progression of RWSIP values as a function of the number of generations/cycles/excitations depending upon the hybrid method for PGK as an illustrative case. **Supplementary Figure S5** shows the progression of RWSIP values as a function of the number of generations for ClustENM and ClustENMD, where the conformers of the current generation are added to the ensemble of previous ones in each successive generation. The RWSIP values of the three independent runs, the average values, as well as the values of the combined ensemble comprising all three runs, are observed to start converging after the second generation and saturate in both cases. This indicates that the intrinsic dynamics encoded by the experimental structures is achieved in early generations. However, the conformers obtained in the later generations allow for approaching the closed conformer of PGK.

Supplementary Figure S6 shows the equivalent progression of RWSIP values for MDeNM. The number of excitations does not influence the RWSIP values in this case, as the same directions of motion are excited each time for any given replica, but the number of excitations is again important for the degree of conformational space sampled including the approach towards closed conformers. The number of replicas is a key parameter that determines the directional coverage and the RWSIP value rises with the number of replicas. Some transition points are observed at about 2, 5 and 10 replicas along with a slow convergence after 30 to 40 replicas.

Supplementary Figure S7 shows the equivalent progression of RWSIP values for coMD. In this case, the direction changes every cycle, and the number of cycles thus makes a much bigger difference. Interestingly, we observe two convergence regimes. Firstly, about 15 cycles is required to converge upon an optimal RWSIP value. However, after about 25–30 cycles, this value decreases as additional directions are explored and the RWSP converges on a new, lower value. The RWSIP is therefore a useful criterion for assessing how many cycles are beneficial for coMD, just like the number of replicas for MDeNM. Looking at how the RWSIP changes as a function of the number of runs, it is clear that there can be substantial variation between runs with some having much higher RWSIP values than others. It is therefore necessary to include a large enough number of runs (e.g., 5 or more) to obtain a sufficient coverage of motion directions.

Finally, an important advantage of hybrid methods is their computational efficiency, and this is without compromising their accuracy as the current comparison with experimental and MD data demonstrates. In particular, the efficiency of ClustENM and ClustENMD are reflected by run times on the order of minutes (**Supplementary Table S1**), while coMD is of the order of hours. The computational efficiency of ClustENM and ClustENMD stems from the usage of implicit solvent during the EM or MD steps, in addition to the adoption of ENMs for predicting the NMs. Given that ENM-based methods provide accuracy levels comparable to those based on full atomic models (MDeNM and MD) with significant savings in computing time, further development of MDeNM methodology using ENM-based NMA, as opposed to full atomic NMA, is currently in progress.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

BK ran ClustENM and ClustENMD for all proteins and MD for TIM and HIV-PR, created HIV-RT and TIM experimental ensembles, wrote analysis code used throughout, did analysis for TIM and PR, and contributed to the text. JK ran coMD simulations for all systems and MD for HIV-RT, helped with MDeNM, generated initial experimental ensembles, and wrote coMD text. BD ran MDeNM for PGK, HIV-PR and TIM and MD for PGK and HIV-PR, generated the PGK experimental ensemble, performed PGK analysis, and wrote MDeNM and PGK text (with EB). ZD ran HIV-RT MDeNM simulations (with MC), performed HIV RT analysis and wrote the associated text. MC helped ZD with HIV-RT MDeNM and text. EB helped BD with PGK text. AS ran MDeNM and MD simulations for some systems, performed HIV-PR analysis, and wrote the associated text (with DP). PD ran MD simulations for TIM and HIV-PR (with AS), contributed to the analysis procedure (with BK), and wrote many parts of the text. DP helped with PGK MDeNM and HIV-PR analysis and text, and provided guidance on MDeNM. IB provided the overall vision and guidance throughout and played a large part in the writing. All authors reviewed complete versions of the manuscript.

ACKNOWLEDGMENTS

All authors thank Gengkon Lum and Afshin Eskandari Nasrabad for their invaluable help with the computer cluster in the Department of Computational and Systems Biology at the University of Pittsburgh where most of the calculations were carried out. Support from National Institutes of Health grant R01GM139297 is gratefully acknowledged by IB and PD. Support from the Molecular Sciences Software Institute *via* a Seed Software Fellowship and from the European Union (EU) and Horizon 2020 through the Marie Sklodowska-Curie Fellowship grant EnLaCES (MSCA IF 2020; Proposal: 101024130) are gratefully acknowledged by JK, Support from Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq grant 310081/2020-3 is gratefully acknowledged by AS. Support from the Brazilian-French collaborative project CAPES/COFECUB is gratefully acknowledged by MGSC and DP. BD highly appreciates the support provided by the Campus Mundi traineeship program (CM-SMP-KA107/299032/2018 and CM-SMP-KA107/1371/2018) financed by the European Union (EU).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.832847/full#supplementary-material

REFERENCES

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophysical J.* 80, 505–515. doi:10.1016/s0006-3495(01)76033-x

Badaya, A., and Sasidhar, Y. U. (2020). Inhibition of the Activity of HIV-1 Protease through Antibody Binding and Mutations Probed by Molecular Dynamics Simulations. *Sci. Rep.* 10, 5501. doi:10.1038/s41598-020-62423-y

Bahar, I., Erman, B., Jernigan, R. L., Atilgan, A. R., and Covell, D. G. (1999). Collective Motions in HIV-1 Reverse Transcriptase: Examination of Flexibility and Enzyme Function. *J. Mol. Biol.* 285, 1023–1037. doi:10.1006/jmbi.1998.2371

Bahar, I., Lezon, T. R., Yang, L.-W., and Eyal, E. (2010). Global Dynamics of Proteins: Bridging between Structure and Function. *Annu. Rev. Biophys.* 39, 23–42. doi:10.1146/annurev.biophys.093008.131258

Bakan, A., and Bahar, I. (2009). The Intrinsic Dynamics of Enzymes Plays a Dominant Role in Determining the Structural Changes Induced upon Inhibitor Binding. *Proc. Natl. Acad. Sci.* 106, 14349–14354. doi:10.1073/pnas.0904214106

Bakan, A., Dutta, A., Mao, W., Liu, Y., Chennubhotla, C., Lezon, T. R., et al. (2014). Evol and ProDy for Bridging Protein Sequence Evolution and Structural Dynamics. *Bioinformatics* 30, 2681–2683. doi:10.1093/bioinformatics/btu336

Bakan, A., Meireles, L. M., and Bahar, I. (2011). ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* 27, 1575–1577. doi:10.1093/bioinformatics/btr168

Batista, P. R., Pandey, G., Pascutti, P. G., Bisch, P. M., Perahia, D., and Robert, C. H. (2011). Free Energy Profiles along Consensus Normal Modes Provide Insight into HIV-1 Protease Flap Opening. *J. Chem. Theor. Comput.* 7, 2348–2352. doi:10.1021/ct200237u

Brändén, G., and Neutze, R. (2021). Advances and Challenges in Time-Resolved Macromolecular Crystallography. *Science* 373. doi:10.1126/science.aba0954

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., et al. (2009). CHARMM: the Biomolecular Simulation Program. *J. Comput. Chem.* 30, 1545–1614. doi:10.1002/jcc.21287

Can, M. T., Kurkcuoglu, Z., Ezeroglu, G., Uyar, A., Kurkcuoglu, O., and Doruker, P. (2017). Conformational Dynamics of Bacterial Trigger Factor in Apo and Ribosome-Bound States. *PLoS One* 12, e0176262. doi:10.1371/journal.pone.0176262

Cansu, S., and Doruker, P. (2008). Dimerization Affects Collective Dynamics of Triosephosphate Isomerase. *Biochemistry* 47, 1358–1368. doi:10.1021/bi701916b

Carnevale, V., Pontiggia, F., and Micheletti, C. (2007). Structural and Dynamical Alignment of Enzymes with Partial Structural Similarity. *J. Phys. Condens. Matter* 19, 285206. doi:10.1088/0953-8984/19/28/285206

Cliff, M. J., Bowler, M. W., Varga, A., Marston, J. P., Szabo, J., Hounslow, A. M., et al. (2010). Transition State Analogue Structures of Human Phosphoglycerate Kinase Establish the Importance of Charge Balance in Catalysis. *J. Am. Chem. Soc.* 132 (18), 6507–6516. doi:10.1021/ja100974t

Corona, A., Meleddu, R., Esposito, F., Distinto, S., Bianco, G., Masaoka, T., et al. (2016). Ribonuclease H/DNA Polymerase HIV-1 Reverse Transcriptase Dual Inhibitor: Mechanistic Studies on the Allosteric Mode of Action of Isatin-Based Compound RMNC6. *PLoS One* 11, e0147225. doi:10.1371/journal.pone.0147225

Costa, M. G. S., Batista, P. R., Bisch, P. M., and Perahia, D. (2015). Exploring Free Energy Landscapes of Large Conformational Changes: Molecular Dynamics with Excited normal Modes. *J. Chem. Theor. Comput.* 11, 2755–2767. doi:10.1021/acs.jctc.5b00003

Costa, M. G. S., Fagnen, C., Vénien-Bryan, C., and Perahia, D. (2020). A New Strategy for Atomic Flexible Fitting in Cryo-EM Maps by Molecular Dynamics with Excited Normal Modes (MDeNM-EMfit). *J. Chem. Inf. Model.* 60, 2419–2423. doi:10.1021/acs.jcim.9b01148

Dudas, B., Merzel, F., Jang, H., Nussinov, R., Perahia, D., and Balog, E. (2020). Nucleotide-Specific Autoinhibition of Full-Length K-Ras4B Identified by Extensive Conformational Sampling. *Front. Mol. Biosci.* 7, 145. doi:10.3389/fmolb.2020.00145

Dudas, B., Perahia, D., and Balog, E. (2021a). Revealing the Activation Mechanism of Autoinhibited RalF by Integrated Simulation and Experimental Approaches. *Sci. Rep.* 11, 10059. doi:10.1038/s41598-021-89169-5

Dudas, B., Toth, D., Perahia, D., Nicot, A. B., Balog, E., and Miteva, M. A. (2021b). Insights into the Substrate Binding Mechanism of SULT1A1 through Molecular Dynamics with Excited normal Modes Simulations. *Sci. Rep.* 11, 13129. doi:10.1038/s41598-021-92480-w

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, e1005659. doi:10.1371/journal.pcbi.1005659

Esposito, F., Corona, A., and Tramontano, E. (2012). HIV-1 Reverse Transcriptase Still Remains a New Drug Target: Structure, Function, Classical Inhibitors, and New Inhibitors with Innovative Mechanisms of Actions. *Mol. Biol. Int.* 2012, 586401. doi:10.1155/2012/586401

Eyal, E., Lum, G., and Bahar, I. (2015). The Anisotropic Network Model Web Server at 2015 (ANM 2.0). *Bioinformatics* 31, 1487–1489. doi:10.1093/bioinformatics/btu847

Fagnen, C., Bannwarth, L., Oubella, I., Forest, E., De Zorzi, R., de Araujo, A., et al. (2020). New Structural Insights into Kir Channel Gating from Molecular Simulations, HDX-MS and Functional Studies. *Sci. Rep.* 10, 8392. doi:10.1038/s41598-020-65246-z

Fagnen, C., Bannwarth, L., Zuniga, D., Oubella, I., Zorzi, R. D., Forest, E., et al. (2021). Unexpected Gating Behaviour of an Engineered Potassium Channel Kir. *Front. Mol Biosci* 8, 538. doi:10.3389/fmolb.2021.691901

Floquet, N., Costa, M. G. S., Batista, P. R., Renault, P., Bisch, P. M., Raussin, F., et al. (2015). Conformational Equilibrium of CDK/Cyclin Complexes by Molecular Dynamics with Excited Normal Modes. *Biophysical J.* 109, 1179–1189. doi:10.1016/j.bpj.2015.07.003

Fuglebakk, E., Reuter, N., and Hinsen, K. (2013). Evaluation of Protein Elastic Network Models Based on an Analysis of Collective Motions. *J. Chem. Theor. Comput.* 9, 5618–5628. doi:10.1021/ct400399x

Gu, S.-X., Zhu, Y.-Y., Wang, C., Wang, H.-F., Liu, G.-Y., Cao, S., et al. (2020). Recent Discoveries in HIV-1 Reverse Transcriptase Inhibitors. *Curr. Opin. Pharmacol.* 54, 166–172. doi:10.1016/j.coph.2020.09.017

Gur, M., Madura, J. D., and Bahar, I. (2013). Global Transitions of Proteins Explored by a Multiscale Hybrid Methodology: Application to Adenylate Kinase. *Biophysical J.* 105, 1643–1652. doi:10.1016/j.bpj.2013.07.058

Gur, M., Zomot, E., Cheng, M. H., and Bahar, I. (2015). Energy Landscape of LeuT from Molecular Simulations. *J. Chem. Phys.* 143, 243134. doi:10.1063/1.4936133

Haliloglu, T., and Bahar, I. (2015). Adaptability of Protein Structures to Enable Functional Interactions and Evolutionary Implications. *Curr. Opin. Struct. Biol.* 35, 17–23. doi:10.1016/j.sbi.2015.07.007

Henzler-Wildman, K. A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., et al. (2007). Intrinsic Motions along an Enzymatic Reaction Trajectory. *Nature* 450, 838–844. doi:10.1038/nature06410

Hinsen, K. (1998). Analysis of Domain Motions by Approximate normal Mode Calculations. *Proteins* 33, 417–429. doi:10.1002/(sici)1097-0134(19981115)33:3<417:aid-prot10>3.0.co;2-8

Huang, J., and MacKerell, A. D., Jr. (2013). CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* 34, 2135–2145. doi:10.1002/jcc.23354

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, 71–73. doi:10.1038/nmeth.4067

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual Molecular Dynamics. *J. Mol. Graph* 14 (33-38), 33–38. doi:10.1016/0263-7855(96)00018-5

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9, 90–95. doi:10.1109/mcse.2007.55

Ilina, T., LaBarge, K., Sarafianos, S. G., Ishima, R., and Parniak, M. A. (2012). Inhibitors of HIV-1 Reverse Transcriptase-Associated Ribonuclease H Activity. *Biology* 1, 521–541. doi:10.3390/biology1030521

Jernigan, R. L., Bahar, I., Covell, D. G., Atilgan, A. R., Erman, B., and Flatow, D. T. (2000). Relating the Structure ofHIV-1 Reverse Transcriptaseto its Processing Step. *J. Biomol. Struct. Dyn.* 17, 49–55. doi:10.1080/07391102.2000.10506603

Jiang, J., Shrivastava, I. H., Watts, S. D., Bahar, I., and Amara, S. G. (2011). Large Collective Motions Regulate the Functional Properties of Glutamate Transporter Trimers. *Proc. Natl. Acad. Sci.* 108, 15141–15146. doi:10.1073/pnas.1112216108

Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008). CHARMM-GUI: a Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* 29, 1859–1865. doi:10.1002/jcc.20945

Kaynak, B. T., and Doruker, P. (2019). Protein-Ligand Complexes as Constrained Dynamical Systems. *J. Chem. Inf. Model.* 59, 2352–2358. doi:10.1021/acs.jcim.8b00946

Kaynak, B. T., Findik, D., and Doruker, P. (2018). RESPEC Incorporates Residue Specificity and the Ligand Effect into the Elastic Network Model. *J. Phys. Chem. B* 122, 5347–5355. doi:10.1021/acs.jpcb.7b10325

Kaynak, B. T., Zhang, S., Bahar, I., and Doruker, P. (2021). ClustENMD: Efficient Sampling of Biomolecular Conformational Space at Atomic Resolution. *Bioinformatics* 37, 3956–3958. doi:10.1093/bioinformatics/btab496

Koehl, P., Orland, H., and Delarue, M. (2021). Parameterizing Elastic Network Models to Capture the Dynamics of Proteins. *J. Comput. Chem.* 42, 1643–1661. doi:10.1002/jcc.26701

Krieger, J. M., Doruker, P., Scott, A. L., Perahia, D., and Bahar, I. (2020). Towards Gaining Sight of Multiscale Events: Utilizing Network Models and normal Modes in Hybrid Methods. *Curr. Opin. Struct. Biol.* 64, 34–41. doi:10.1016/j.sbi.2020.05.013

Kurkcuoglu, O., Jernigan, R. L., and Doruker, P. (2006). Loop Motions of Triosephosphate Isomerase Observed with Elastic Networks. *Biochemistry* 45, 1173–1182. doi:10.1021/bi0518085

Kurkcuoglu, Z., Bahar, I., and Doruker, P. (2016). ClustENM: ENM-Based Sampling of Essential Conformational Space at Full Atomic Resolution. *J. Chem. Theor. Comput.* 12, 4549–4562. doi:10.1021/acs.jctc.6b00319

Kurkcuoglu, Z., and Bonvin, A. M. J. J. (2020). Pre- and post-docking Sampling of Conformational Changes Using ClustENM and HADDOCK for Protein-protein and protein-DNA Systems. *Proteins* 88, 292–306. doi:10.1002/prot.25802

Kurkcuoglu, Z., and Doruker, P. (2016). Ligand Docking to Intermediate and Close-To-Bound Conformers Generated by an Elastic Network Model Based Algorithm for Highly Flexible Proteins. *PLoS One* 11, e0158063. doi:10.1371/journal.pone.0158063

Kurkcuoglu, Z., and Doruker, P. (2013). Substrate Effect on Catalytic Loop and Global Dynamics of Triosephosphate Isomerase. *Entropy* 15, 1085–1099. doi:10.3390/e15031085

Kurkcuoglu, Z., Findik, D., Akten, E. D., and Doruker, P. (2015). How an Inhibitor Bound to Subunit Interface Alters Triosephosphate Isomerase Dynamics. *Biophysical J.* 109, 1169–1178. doi:10.1016/j.bpj.2015.06.031

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., et al. (2010). Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* 78, 1950–1958. doi:10.1002/prot.22711

Maldonado, E., Soriano-García, M., Moreno, A., Cabrera, N., Garza-Ramos, G., Tuena de Gómez-Puyou, M., et al. (1998). Differences in the Intersubunit Contacts in Triosephosphate Isomerase from Two Closely Related Pathogenic Trypanosomes. *J. Mol. Biol.* 283, 193–203. doi:10.1006/jmbi.1998.2094

Martin, P., Vickrey, J. F., Proteasa, G., Jimenez, Y. L., Wawrzak, Z., Winters, M. A., et al. (2005). "Wide-Open" 1.3 Å Structure of a Multidrug-Resistant HIV-1 Protease as a Drug Target. *Structure* 13, 1887–1895. doi:10.1016/j.str.2005.11.005

Miyashita, O., and Tama, F. (2018). Hybrid Methods for Macromolecular Modeling by Molecular Mechanics Simulations with Experimental Data. *Adv. Exp. Med. Biol.* 1105, 199–217. doi:10.1007/978-981-13-2200-6_13

Morningstar, M. L., Roth, T., Farnsworth, D. W., Kroeger Smith, M., Watson, K., Buckheit, R. W., et al. (2007). Synthesis, Biological Activity, and crystal Structure of Potent Nonnucleoside Inhibitors of HIV-1 Reverse Transcriptase that Retain Activity against Mutant Forms of the Enzyme. *J. Med. Chem.* 50, 4003–4015. doi:10.1021/jm060103d

Namasivayam, V., Vanangamudi, M., Kramer, V. G., Kurup, S., Zhan, P., Liu, X., et al. (2019). The Journey of HIV-1 Non-nucleoside Reverse Transcriptase Inhibitors (NNRTIs) from Lab to Clinic. *J. Med. Chem.* 62, 4851–4883. doi:10.1021/acs.jmedchem.8b00843

Onufriev, A., Bashford, D., and Case, D. A. (2004). Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins* 55, 383–394. doi:10.1002/prot.20033

Orellana, L. (2019). Large-Scale Conformational Changes and Protein Function: Breaking the In Silico Barrier. *Front. Mol. Biosci.* 6, 117. doi:10.3389/fmolb.2019.00117

Palese, L. L. (2017). Conformations of the HIV-1 Protease: A crystal Structure Data Set Analysis. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1865, 1416–1422. doi:10.1016/j.bbapap.2017.08.009

Palmai, Z., Chaloin, L., Lionne, C., Fidy, J., Perahia, D., and Balog, E. (2009). Substrate Binding Modifies the Hinge Bending Characteristics of Human 3-phosphoglycerate Kinase: a Molecular Dynamics Study. *Proteins* 77, 319–329. doi:10.1002/prot.22437

Palmai, Z., Seifert, C., Gräter, F., and Balog, E. (2014). An Allosteric Signaling Pathway of Human 3-phosphoglycerate Kinase from Force Distribution Analysis. *Plos Comput. Biol.* 10, e1003444. doi:10.1371/journal.pcbi.1003444

Phillips, J. C., Hardy, D. J., Maia, J. D. C., 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

Resende-Lara, P. T., Perahia, D., Scott, A. L., and Braz, A. S. K. (2020). Unveiling Functional Motions Based on point Mutations in Biased Signaling Systems: A normal Mode Study on Nerve Growth Factor Bound to TrkA. *PLoS One* 15, e0231542. doi:10.1371/journal.pone.0231542

Schlitter, J., Engels, M., and Krüger, P. (1994). Targeted Molecular Dynamics: a New Approach for Searching Pathways of Conformational Transitions. *J. Mol. Graphics* 12, 84–89. doi:10.1016/0263-7855(94)80072-3

Scott, W. R. P., and Schiffer, C. A. (2000). Curling of Flap Tips in HIV-1 Protease as a Mechanism for Substrate Entry and Tolerance of Drug Resistance. *Structure* 8, 1259–1265. doi:10.1016/s0969-2126(00)00537-2

Sluis-Cremer, N., Temiz, N., and Bahar, I. (2004). Conformational Changes in HIV-1 Reverse Transcriptase Induced by Nonnucleoside Reverse Transcriptase Inhibitor Binding. *Curr. HIV Res.* 2, 323–332. doi:10.2174/1570162043351093

Swift, R. V., and McCammon, J. A. (2008). Catalytically Requisite Conformational Dynamics in the mRNA-Capping Enzyme Probed by Targeted Molecular Dynamics. *Biochemistry* 47, 4102–4111. doi:10.1021/bi8000209

Thirumalai, D., Hyeon, C., Zhuravlev, P. I., and Lorimer, G. H. (2019). Symmetry, Rigidity, and Allosteric Signaling: From Monomeric Proteins to Molecular Machines. *Chem. Rev.* 119, 6788–6821. doi:10.1021/acs.chemrev.8b00760

Tu, X., Das, K., Han, Q., Bauman, J. D., Clark, A. D., Hou, X., et al. (2010). Structural Basis of HIV-1 Resistance to AZT by Excision. *Nat. Struct. Mol. Biol.* 17, 1202–1209. doi:10.1038/nsmb.1908

Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Joss* 6, 3021. doi:10.21105/joss.03021

Williams, J. C., and McDermott, A. E. (1995). Dynamics of the Flexible Loop of Triose-Phosphate Isomerase: The Loop Motion Is Not Ligand Gated. *Biochemistry* 34, 8309–8319. doi:10.1021/bi00026a012

Wingert, B., Krieger, J., Li, H., and Bahar, I. (2021). Adaptability and Specificity: How Do Proteins Balance Opposing Needs to Achieve Function? *Curr. Opin. Struct. Biol.* 67, 25–32. doi:10.1016/j.sbi.2020.08.009

Yamazaki, T., Hinck, A. P., Wang, Y.-X., Nicholson, L. K., Torchia, D. A., Wingfield, P., et al. (1996). Three-dimensional Solution Structure of the HIV-1 Protease Complexed with DMP323, a Novel Cyclic Urea-type Inhibitor, Determined by Nuclear Magnetic Resonance Spectroscopy. *Protein Sci.* 5, 495–506. doi:10.1002/pro.5560050311

Yang, L.-W., and Bahar, I. (2005). Coupling between Catalytic Site and Collective Dynamics: a Requirement for Mechanochemical Activity of Enzymes. *Structure* 13, 893–904. doi:10.1016/j.str.2005.03.015

Yang, L., Song, G., Carriquiry, A., and Jernigan, R. L. (2008). Close Correspondence between the Motions from Principal Component Analysis of Multiple HIV-1 Protease Structures and Elastic Network Modes. *Structure* 16, 321–330. doi:10.1016/j.str.2007.12.011

Yang, L., Song, G., and Jernigan, R. L. (2009). Protein Elastic Network Models and the Ranges of Cooperativity. *Proc. Natl. Acad. Sci.* 106, 12347–12352. doi:10.1073/pnas.0902159106

Yon, J. M., Perahia, D., and Ghélis, C. (1998). Conformational Dynamics and Enzyme Activity. *Biochimie* 80, 33–42. doi:10.1016/s0300-9084(98)80054-0

Zerrad, L., Merli, A., Schröder, G. F., Varga, A., Gráczer, É., Pernot, P., et al. (2011). A spring-loaded Release Mechanism Regulates Domain Movement and Catalysis in Phosphoglycerate Kinase. *J. Biol. Chem.* 286, 14040–14048. doi:10.1074/jbc.m110.206813

Zhang, S., Krieger, J. M., Zhang, Y., Kaya, C., Kaynak, B., Mikulska-Ruminska, K., et al. (2021). ProDy 2.0: Increased Scale and Scope after 10 Years of Protein

Dynamics Modelling with Python. *Bioinformatics* 37, 3657. doi:10.1093/bioinformatics/btab187/6211036

Zhang, S., Li, H., Krieger, J. M., and Bahar, I. (2019). Shared Signature Dynamics Tempered by Local Fluctuations Enables Fold Adaptability and Specificity. *Mol. Biol. Evol.* 36, 2053–2068. doi:10.1093/molbev/msz102

Zhang, Y., Doruker, P., Kaynak, B., Zhang, S., Krieger, J., Li, H., et al. (2020). Intrinsic Dynamics Is Evolutionarily Optimized to Enable Allosteric Behavior. *Curr. Opin. Struct. Biol.* 62, 14–21. doi:10.1016/j.sbi.2019.11.002

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.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 11 April 2022
doi: 10.3389/fmolb.2022.826136

Building Biological Relevance Into Integrative Modelling of Macromolecular Assemblies

Anne-Elisabeth Molza[1,2†], Yvonne Westermaier[3], Magali Moutte[4], Pierre Ducrot[3], Claudia Danilowicz[5], Veronica Godoy-Carter[6], Mara Prentiss[5], Charles H. Robert[1,2], Marc Baaden[1,2] and Chantal Prévost[1,2]*

[1]CNRS, Université Paris-Cité, UPR 9080, Laboratoire de Biochimie Théorique, Paris, France, [2]Institut de Biologie Physico-Chimique-Fondation Edmond de Rothschild, PSL Research University, Paris, France, [3]Biophysics and Modelling Department/In Vitro Pharmacology Unit–IDRS (Servier Research Institute), Croissy-sur-Seine, France, [4]Servier Monde, Suresnes, France, [5]Department of Physics, Harvard University, Cambridge, MA, United States, [6]Department of Biology, Northeastern University, Boston, MA, United States

OPEN ACCESS

Edited by:
Paulo Ricardo Batista,
Oswaldo Cruz Foundation (Fiocruz),
Brazil

Reviewed by:
Tahsin F. Kellici,
Evotec, United Kingdom
Lorenzo Stella,
University of Rome Tor Vergata, Italy

***Correspondence:**
Chantal Prévost
chantal.prevost@ibpc.fr

†Present address:
Anne-Elisabeth Molza,
DSIMB, UMR-S1134 INSERM,
Faculté des Sciences et Technologies,
Université de La Réunion, Saint Denis,
La Réunion, France

Specialty section:
This article was submitted to
Structural Biology,
a section of the journal
Frontiers in Molecular Biosciences

Received: 30 November 2021
Accepted: 21 February 2022
Published: 11 April 2022

Citation:
Molza A-E Westermaier Y, Moutte M,
Ducrot P, Danilowicz C,
Godoy-Carter V, Prentiss M,
Robert CH, Baaden M and Prévost C
(2022) Building Biological Relevance
Into Integrative Modelling of
Macromolecular Assemblies.
Front. Mol. Biosci. 9:826136.
doi: 10.3389/fmolb.2022.826136

Recent advances in structural biophysics and integrative modelling methods now allow us to decipher the structures of large macromolecular assemblies. Understanding the dynamics and mechanisms involved in their biological function requires rigorous integration of all available data. We have developed a complete modelling pipeline that includes analyses to extract biologically significant information by consistently combining automated and interactive human-guided steps. We illustrate this idea with two examples. First, we describe the ryanodine receptor, an ion channel that controls ion flux across the cell membrane through transitions between open and closed states. The conformational changes associated with the transitions are small compared to the considerable system size of the receptor; it is challenging to consistently track these states with the available cryo-EM structures. The second example involves homologous recombination, in which long filaments of a recombinase protein and DNA catalyse the exchange of homologous DNA strands to reliably repair DNA double-strand breaks. The nucleoprotein filament reaction intermediates in this process are short-lived and heterogeneous, making their structures particularly elusive. The pipeline we describe, which incorporates experimental and theoretical knowledge combined with state-of-the-art interactive and immersive modelling tools, can help overcome these challenges. In both examples, we point to new insights into biological processes that arise from such interdisciplinary approaches.

Keywords: integrative modelling, biological function, large macromolecular assemblies, molecular dynamics simulations, normal modes, ryanodine receptor, homologous recombination

1 INTRODUCTION

Many biological processes depend on the formation of large transient or permanent macromolecular assemblies. These assemblies may harbor signaling functions (e.g., membrane receptor proteins), act as motors for cell motility, membrane crossing, DNA maintenance, or ATP synthesis, or form scaffolds for the cell wall or for communication networks. In recent years, extraordinary advances have been made in both technology and information processing, leading to an avalanche of experimental 3D structures of large biomolecular complexes (Nogales, 2016; Murata and Wolf, 2018; Ziegler et al., 2021), many with resolutions better than 4 Å (Ghanim et al., 2021). The availability of these structures has enabled unprecedented advances in understanding many important biological processes (Verkhivker et al., 2021). It seems that access to high-resolution structures of all large multimolecular edifices involved in complex biological processes is only a matter of time. Does this mean that we will be able to understand their mechanisms? This question highlights a fundamental challenge for understanding functional biological assemblies: how to consistently integrate experimental structure determination into mechanistic and functional models. From this point of view, integrative modelling can be seen as a tool that goes beyond the structural level to explicitly include questions of biological function and mechanism.

Mechanistic understanding of the function of proteins and other biological macromolecules is usually based on the availability of three-dimensional structural information. The earliest example of this is probably the replication of DNA based on an atomic model of the double helix (Watson and Crick, 1953). Shortly thereafter, the first crystal structure of a protein, myoglobin, appeared (Kendrew et al., 1958), sparking a movement to determine more and more macromolecular structures at ever higher resolution by X-ray diffraction, neutron diffraction, NMR, and cryo-electron microscopy. Currently, there are nearly 200,000 experimental 3D structures of individual proteins and small complexes in the Protein Data Bank. Analysis of this rich source of high-resolution structural data increasingly contributed to the creation of models of proteins for which no experimental structure existed (e.g., Webb and Sali, 2021). This culminates today in the successful application of artificial intelligence to the automatic generation of highly accurate 3D models (Jumper et al., 2021), reflecting the tremendous progress that has been made in the last decade in predicting protein structures from sequences. Taken together, this detailed structural information has helped to elucidate the mechanisms of myriad catalytic reactions and biochemical processes carried out by macromolecules in the cell.

Over the same period, and largely thanks to advances in cryoEM (Nogales, 2016; Murata and Wolf, 2018) and its coupling with integrative modelling (Ziegler et al., 2021), this inexorable push towards high-resolution 3D structural data has attained the scale of large-scale assemblies in the cell. It would thus appear that understanding the mechanisms of these processes is indeed at hand. Yet deciphering the function of large assemblies in many cases defies the “structure illuminates function” logic. This may occur because 3D data is missing for particular regions of otherwise well-resolved structures, thus preventing one from fully exploiting them, or from the absence of essential cofactors or partner proteins needed for their function. Similarly, structural disorder in the complex may be intrinsic to its function, so that a structural average or snapshot provides little insight. The availability of the structures of the endpoints of a given process (reactants, products) and of a few intermediates may not be sufficient *per se* to directly address the underlying biological mechanisms. This could be because the stable intermediates are structurally too far apart to simply infer the transition between them, or because new structural information reveals new gaps in our knowledge. Structural sets could reflect different experimental conditions or different levels of resolution and thus not readily provide a meaningful overview of the biology in question. In such cases, filling the structural gaps requires the additional effort of integrating complementary experimental and numerical approaches.

The term *integrative modelling* is frequently used to describe obtaining 3D structures of macromolecular assemblies based upon medium-to low-resolution data such as cryo-EM maps or SAXS profiles. Additional information obtained from mutagenesis, NMR, or mass-spectrometry and cross-linking (Faini et al., 2016) has frequently been used to reduce structural ambiguities. Docking has been helpful in predicting 3D structures of large macromolecular complexes starting from those of their known constituents (Inbar et al., 2005; Schneidman-Duhovny and Wolfson, 2020), although many such cases remain a challenge due to complex combinatorial problems. Platforms for integrating such diverse information into structural models propose user-friendly software pipelines that take advantage of the known high-resolution 3D-structures of complex components or their homologs sharing a sufficient degree of sequence similarity (Russel et al., 2012; van Zundert et al., 2016; Mirabello and Wallner, 2017). The outcome of such modelling is often considered to represent metastable conformational substates that exhibit robust stability when subjected to hundreds of nanoseconds of molecular dynamics simulations. However, these criteria are not always sufficient to approach functional questions.

Here, we focus on two representative examples of assemblies where structural information alone is not sufficient to obtain mechanistic insights. The first is the ryanodine receptor (RyR). To date, RyRs are the largest known ion channels, with a molecular mass of over 2.3 MDa. Each subunit of the homotetrameric channel contains more than 5,000 residues. The number of RyR cryo-EM datasets is steadily increasing (Baldwin et al., 2018; Spurgeon et al., 2021). However, although many maps are available that span different conformational substates obtained under different conditions, there are still no complete models of this receptor. The maps lack homogeneity and the necessary coherence that would permit gaining insight into the structure-dynamics-function relationships in this critical channel protein and help uncover missing biological significance. The second case concerns the nucleoprotein filament complex that is active in homologous recombination (HR)– a fundamental biological process that aims at faithfully repairing broken DNA strands (Prentiss et al., 2015; Bell and Kowalczykowski, 2016). This complex comprises a long filament of recombination proteins assembled on DNA strands. The whole assembly undergoes very rapid dynamic evolution that is directly linked to its function, hence the difficulty in obtaining reliable structural information on intermediate states.

We illustrate how these challenges can be addressed using a modelling pipeline (**Table 1**) which begins with relatively simple structural modelling and gradually incorporates steps that require more sophisticated methodologies. The process is explicitly oriented towards arriving at dynamic information which may be used simply to verify the metastability of a state, for example, or to explore changes of the system in response to a perturbation

TABLE 1 | Integrative modelling tasks applied to two representative examples.

	RyR1	RecA*
generation of missing atom coordinates	yes	no
side chain reconstruction	yes	no
backbone modelling	yes	yes
ab initio structure prediction	yes	no
interactive loop modelling	yes	yes
interactive engineering	no	yes
cofactor/molecular docking	no	yes
MD refinement	yes	yes
MD (meta-)stability verification	yes	no
MD/NM preliminary (or model) exploration	yes	yes

or relaxation process. A given step in **Table 1** may or may not be applied depending on the biological context. In the first system presented, the ryanodine receptor RyR1, we focus on the modelling of "building blocks" consisting of a single macromolecule but in different conformational states, for which coherent modelling, validation and comparisons can be performed. The second example, a RecA-DNA-polymerase complex that executes the final steps of the HR process, concerns the assembly and coordination of building blocks that are already available into functional biological machines that can be used to predict mechanistic aspects of the process. Although this is conceptually more complex than the construction of the building blocks themselves, the two examples share similar logic at different levels, and employ dynamic exploration as part of the modelling and preliminary exploration steps.

2 METHODS

The elements described here allowed performing the integrative modelling tasks in our modelling pipeline, which are summarized in **Table 1**. We also detail the protocols we developed, introduce nomenclature, and explain key decisions. The overall methodology is general, but the actual details are specific to each system.

2.1 Full-Length RyR1 Models

2.1.1 Multiple Alignments

Three mammalian isoforms of RyR exist in cells: RyR1 predominates in skeletal muscle; RyR2 is the most abundantly expressed isoform in cardiac muscle; and RyR3 is expressed differentially in brain (Takeshima et al., 1989; Nakai et al., 1990; Fill and Copello, 2002; Lanner et al., 2010), endocrine cells and other tissues. At least two of these isoforms, RyR1 and RyR2, share the same structural organization with an overall mushroom-like shape as observed by electron microscopy (Van Petegem, 2015; Zalk et al., 2015; Georges et al., 2016; Peng et al., 2016; Dhindwal et al., 2017). These isoforms have a high percentage (ca. 70%) of sequence identity.

We selected reviewed RyR sequences from the UniProtKB database (Boutet et al., 2016). The entries used for this study were: P11716, P21817, E9PZQ0, P30957, Q92736, E9Q401, Q9TS33, Q15413, and A2AGL3. ClustalO (Sievers et al., 2011) was used for alignments, WebLogo (Crooks et al., 2004) and Skyalign (Wheeler et al., 2014) were used to assess the presence of conserved residues.

2.1.2 Secondary Structure and Transmembrane Domain Predictions

RyR1 homologues have a high percentage of sequence identity (on the order of 70%), and their secondary structures are well conserved. We were particularly interested in the interspersed disordered region between residues 4254 and 4539, which separates the cytoplasmic from the transmembrane domains. To obtain information on this region, which we call "Big Loop" (BL), we first examined its sequence conservation and residue composition. Since we focus on the RyR1 isoform, we used five homologous protein sequences and performed a multiple sequence alignment using the Clustal Omega software. Secondary structures were predicted using several web servers, such as PSIPRED and PSSPRED (Jones, 1999; Yan et al., 2013). Transmembrane regions were predicted using the TOPCONS and CCTOP servers (Dobson et al., 2015; Tsirigos et al., 2015). The ion binding sites were predicted via the IonCom program (Hu et al., 2016). For the BL, we used PEP-Fold3 (Lamiable et al., 2016), particularly for residues 4320 to 4345. Finally, we assessed the presence of an auxiliary transmembrane helix (or helices) and other secondary structure elements by visual inspection of the cryo-EM density maps.

2.1.3 Starting Structures Used for Modelling

The partial structural models deposited in the PDB (Georges et al., 2016) were used as initial data for a series of modelling steps (**Table 2**). Since then, additional structures for the rabbit RyR1 isoform were released, which we may use in subsequent work to compare to the models that we produced. We briefly examined this new structural data, which only became available after our work was completed, and concluded that there was no imperative to repeat the procedure for the purpose of illustrating the integrative modelling pipeline.

2.1.4 Initial Structure Preparation

Due to the inherent conformational flexibility of some regions of RyR1, the cryo-EM structures contain several gaps and are subject to many uncertainties regarding the coordinates of specific residues.

We first examined the coordinates of the "unknown residue" (*UNK*) annotations in the PDB files. We assumed that the Cα coordinates in the PDB files were acceptable starting positions for these residues. These amino acids were assigned initial positions with ROSETTA based on their sequence and then checked visually one by one against the density map. Further, missing side chains and residues of known sequence were modelled with the *de novo* modelling method using the program ROSETTA remodel (Huang et al., 2011).

At the end of the model building process, we checked newly rebuilt residues visually, especially the helical domains such as the junctional solenoid (J-solenoid), using selected buried residues as

TABLE 2 | RyR1 putative functional states, conformations, ligands, PDB and EMDB identifiers used for this work. # AAs stands for number of amino acids in the corresponding sequence. CFF stands for caffeine. All data are for rabbit RyR1.

State	Conformation	Ligands	PDB Id	EMDB Id	# AAs
apo	closed	-	5TB0	8391	18096
primed	closed-like	Ca2+	5T15	8342	19136
Activated/intermediate	intermediate	ATP/CFF	5TAP	8381	18096
activated	open	ATP/CFF/Ca2+	5TA3	8377	18096
locked	open	Ryanodine	5TAW	8387	18096

a guide to assess whether the side chains of the residues fit well into the density maps.

Before refining the entire model, several loops or side chains were optimized by a combination of side-chain reorientation and loop placement using in-house routines for flexible and interactive molecular dynamics implemented in the BioSpring tool (Molza et al., 2014) followed by energy-minimization steps using the YASARA Structure software (version 17.12.24) (Krieger et al., 2002). This procedure was particularly useful for refining interfacial loops between adjacent subunits (residues 4290–4299, 3086–3120, 3067–3075, and 4346–4426).

2.1.5 Model Refinement

For refining the coordinates of the modelled residues and verifying the side-chain placements, Molecular Dynamics Flexible Fitting (MDFF) simulations were performed using the NAMD software (Phillips et al., 2005; Trabuco et al., 2008). Our initial models were refined using the real-space refinement procedure from the PHENIX package (Afonine et al., 2013; Urzhumtsev et al., 2016) in order to avoid the propagation of geometry errors from the input structure (PDB and *de novo* modelling output files) during the MDFF procedure.

The secondary structures of residues folded into α-helices or β-sheets were preserved during simulations by imposing harmonic restraints with force constants of 20 kcal mol^{-1} rad^{-2} for dihedral angles and hydrogen bonds, involving backbone atoms of the same subset of residues using the ssrestraints command.

The NAMD configuration files for the MDFF simulations were generated using the mdff setup command. The MDFF simulation was performed *in vacuo* with NAMD 2.12 using the CHARMM36 force field (Lee et al., 2014). 10,000 minimization steps were performed, followed by slow heating to 300 K and 4 M production steps (4 ns) of molecular dynamics, with a time step of 1 fs/step. A scaling factor of 1 kcal mol^{-1} was used to adjust the strength of the influence of the electron density map on the tetramer model during the fitting process. The remaining parameters were defined based on default values in the mdff_template.namd file. The MD calculations were performed on the ADA supercomputer at the French IDRIS Supercomputing Center.

For reducing remaining outliers, additional refinement steps were performed using PHENIX and YASARA minimization steps under restraints.

2.1.6 Side-Chain Refinement in Ligand Binding Sites

As starting positions, we used the coordinates of the calcium and zinc ions, ATP and caffeine molecules, as previously determined by electron microscopy (Georges et al., 2016). For ryanodine, we used a Ryd model provided by Ngo and the coordinates from their MD study of ligand placement (Ngo et al., 2017). The models were created using YASARA. For each model, the most likely protonation state was calculated at pH 7.0. The program YASARA was used for atom typing and hydrogen atom assignment, followed by virtual titration. Sidechains of residues forming ligand binding sites and ligands were then refined using a YASARA macro within the VINALS method (VINA with Local Search) for local ligand docking plus minimization steps. The simulation cell was enlarged by 10 Åin this step.

2.1.7 Coarse-Grained Molecular Dynamics Simulations

Coarse-grained Molecular Dynamics (CGMD) simulations were performed using the Martini force field (http://www.cgmartini.nl/). The setup uses version v2.2 of the force field (de Jong et al., 2013) in combination with the ELNeDin elastic network (Periole et al., 2009) and the GROMACS 2016.3 simulation engine (Abraham et al., 2015). Proteins were embedded in a membrane composed of DOPC/DOPE at a ratio of 5:3, and the systems were neutralized and then solvated in 150 mM NaCl. More than 100K pseudo atoms were simulated for two systems containing the channel core (with or without the BL region). More than 438K pseudo atoms were simulated for the entire protein. The calculations were performed at the French GENCI "IDRIS" Supercomputing Center and on local GPU clusters.

2.1.8 Normal Mode Analyses

Due to the large size of the RyR1 system, we carried out Normal Modes Analysis (NMA) using the Non-Linear Rigid Block (NOLB by Hoffmann and Grudinin (2017)) software, which uses a coarse-grained block approach (Durand et al., 1994). In addition to efficiently analyzing large assemblies, NOLB decomposes the block motions into instantaneous rotational and translational components, whose nonlinear extrapolation allows the computation of periodic trajectories for visualization that preserve structural integrity better than standard linear approaches. The 100 lowest frequency vibrational modes were calculated for each of the superimposed RyR1 conformations using an all-atom elastic network model with an interaction cutoff distance of 10 Å.

Ligand atoms, when present, were included in the analyses. Vibrational modes were sorted in order of frequency $\sqrt{\lambda_i}$, with λ_i being the eigenvalues of the mass-weighted stiffness matrix (Hoffmann and Grudinin, 2017).

Two NM sets a and b were compared by examining the subspace overlap, or sum of squared projection of each mode vector from set a onto the space spanned by the modes of set b (see, e.g., Batista et al., 2010)]. Only C_α atom movements were used for these comparisons. The comparison subspace was obtained by orthonormalizing the C_α mode vectors from set b using standard techniques. The value of each vector overlap lies between 0 (no subspace overlap) and 1 (perfect overlap).

NOLB was also used to compute transition paths between pairs of RyR1 conformations by minimizing the RMSD between two input structures (an initial and a target structure) obtained by the linear combination of the calculated modes; these paths were achieved using all 100 computed modes. Scripts were written in Python (general numerical analysis and Chimera scripts) and Tcl (in the VMD environment) to analyze the NM results.

2.2 Atomic Model of a RecA*-DinB Complex

2.2.1 Starting Point for the Present Study

The starting point for this system was the results of a preliminary coarse-grained study of the densely packed association complex between the RecA nucleoprotein filament (RecA*) and DNA-polymerase IV (DinB), which had been obtained by interactive simulations using the BioSpring software also used for the RyR1 modelling above (**Section 2.1.4**), and which represents each component of a complex system by a variable network of harmonic springs, allowing interactive dynamic testing of possible component orientations and deformations (Molza et al., 2014; Tashjian et al., 2019). In this case the components were the nucleoprotein filament (RecA and DNA) resulting from our earlier strand exchange simulations (Yang et al., 2015) and the crystal structure of DinB bound to its cognate DNA (PDB code 4IRC) (**Figure 1B**).

2.2.2 Initial Model Refinement

The result of the coarse-grained BioSpring flexible assembly process required refinement, as much of the DinB secondary structure was distorted due to the spring network adaptation to the crowded environment of the RecA* filament. We thus superimposed individual DinB secondary structures taken from the crystal structure of DNA-bound DinB onto the corresponding regions in the preliminary model. One helix showed a partial steric clash with the terminal RecA monomer; the clash was released by laterally displacing the helix as a rigid body by 2 Å with respect to the corresponding BioSpring-positioned segment. Where possible, the loops linking secondary structure regions in the crystal structure were included in the model when this resulted in no steric clashes with RecA proteins; otherwise the BioSpring loop structure was used. All disruptions of the peptide chain structure after this process were removed using energy minimization with NAMD, which restored the standard covalent bond geometry.

2.2.3 Introduction of a Guanine Nucleotide Tag

We introduced a guanine nucleotide (resid 11 in the template strand, noted G_{11}^{templ}) as a tag in the template strand in order to be able to identify the site of nucleotide (CTP) addition. The final simulated complex was formed by 17 RecA monomers, the DinB protein and three DNA strands forming a D-loop inside the RecA filament: the ssDNA or primer, with sequence (5′–3′) $(dT)_{53}$, the template strand $(dA)_{10}dG(dA)_{63}$ partly dissociated from the homologous dsDNA, and the displaced strand $(dT)_{74}$ from the dsDNA.

2.2.4 All-Atom Molecular Dynamics Simulation

MD simulations were performed with NAMD 2.10 (Phillips et al., 2005) and the CHARMM 27 force field with CMAP corrections (Mackerell et al., 2004). A 2 fs integration timestep was used with the SHAKE algorithm, long-range electrostatics were accounted for using the particle-mesh Ewald method, and a Nosé-Hoover-Langevin piston was used for pressure maintenance. The complex was solvated in a TIP3P water box with 0.15 mMol NaCl and progressively heated to 300K. During heating, C_α and P atoms were harmonically restrained to their initial position with a force constant of 0.5 kcal mol^{-1} $Å^{-2}$ that was gradually released during 20 ns equilibration followed by a 270 ns production phase. Calculations were performed at the French GENCI CINES Supercomputing Center.

3 RESULTS

In this section we describe how we created models using our integrative modelling pipeline by detailing the modelling tasks summarized in **Table 1** employed for each of our two example systems. We then present an initial set of interpretations of the models' properties to show how they can provide biologically useful new information. The first "use case," RyR, illustrates how consistent integrative models of a single macromolecular building block can be created. The second use case, RecA, documents how to create larger assemblies from such building blocks and the pitfalls to avoid to create a functionally relevant model. The overall approach, however, is general. The details provided for each system are intended to support the usefulness of the models and to provide avenues for further exploration of these biological systems. These may be reported in future work.

3.1 Modelling the RyR1 Ryanodine Receptor Channel and Its Dynamics

Ryanodine Receptors (RyRs) regulate cytosolic calcium concentration, which is critical in numerous signaling pathways. Mutations in these receptors in muscle cells can lead to severe skeletal muscle and heart disease. RyRs belong to the six-transmembrane helix ion channel superfamily (Lanner et al., 2010) and are localized in the endoplasmic and sarcoplasmic reticulum.

Here, RyR1 provides a case study to demonstrate general aspects of the pipeline that lead from deposited cryo-EM

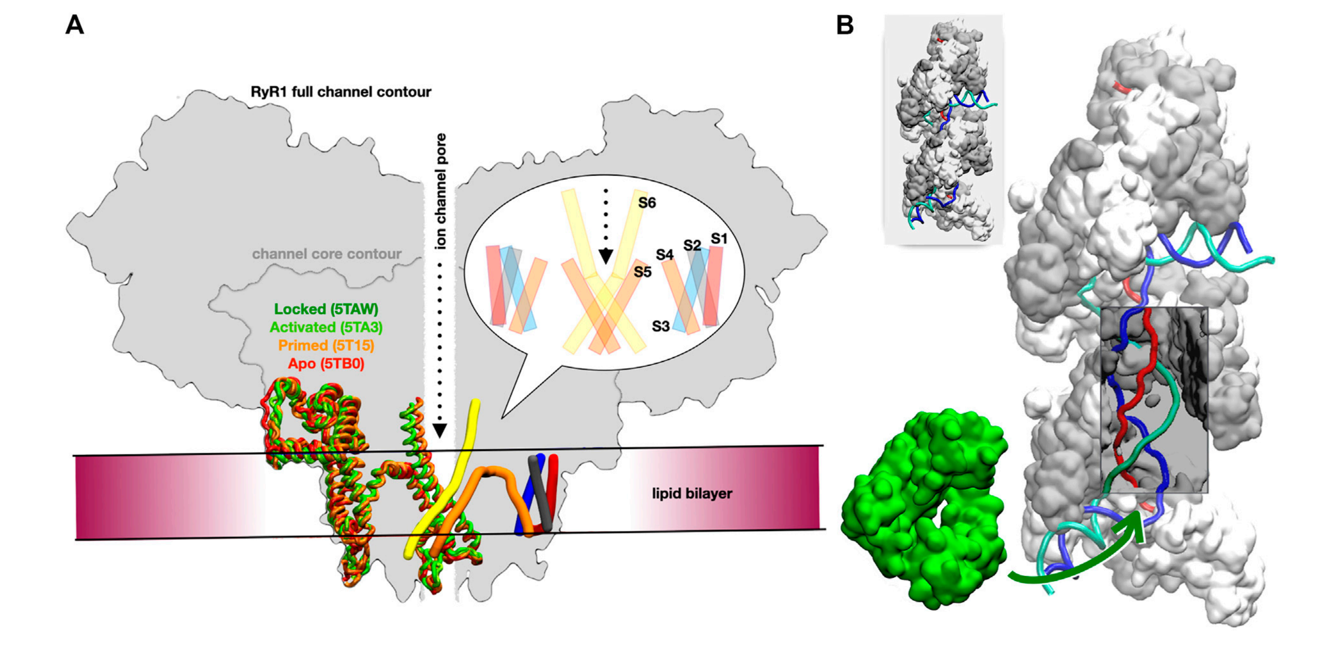


FIGURE 1 | Overview of the two studied systems. **(A)** Schematic cross-sectional view of the full RyR channel model. The approximate location of the lipid bilayer is drawn, as is the central ion channel pore. On the left side of the schematic, the helical sections for a transmembrane domain monomer for each of the four states locked (dark green), activated (light green), primed (orange), and apo (red) are overlaid to highlight the conformational changes in this region. On the right side, the pattern of the six transmembrane helices is shown in cylindrical form: S1 in red, S2 in grey, S3 in blue, S4 and S5 in orange, and S6 in yellow. A bubble inset shows a schematic cross-section of the transmembrane domain, as typically found in the literature, for reference. **(B)** Two turns of the RecA* filament with three bound DNA strands; RecA proteins are represented alternatively in white and grey (surface); the DNA strands are schematically represented in red, blue and cyan; the section that is incorporated in the filament form a D-loop, where the blue strand has exchanged its pairing partner from the cyan to the red strand; a window has been opened in the filament to visualize the otherwise hidden D-loop (a complete filament is represented in the top left insert). The DinB DNA polymerase is represented in green surface; the green arrow points to the region of the filament where DinB binds to the DNA strands and RecA monomers in order to start elongating the red DNA strand.

datasets to complete models that can be used for molecular modelling studies on mechanistic and structural aspects. We were particularly interested in 1) addressing issues related to the channel's large size and 2) benefitting from the availability of maps for multiple functional/conformational sub-states, which provide a good test for the usability of the final models and the level of detail and accuracy that can be expected from them.

We focused on the rabbit RyR1 isoform due to the availability of numerous structural and experimental studies on its gating and activation (Georges et al., 2016). Our goal was to test whether reliable atomistic models of the five resolved channel states (**Table 2**) could be constructed and mechanistically interpreted based on the experimental data. These five states are defined according to Georges et al. (2016): for the apo and primed states, the channel is in a closed conformation, whereas in the activated and ryanodine-locked states, it is in an open conformation. **Figure 1A** provides a schematic view of the different states of the receptor highlighting the transmembrane and adjacent regions (the intermediate state is not shown in order to reduce visual distraction).

3.1.1 Multiple Sequence Alignment

As previous comparative studies have shown (Yan et al., 2015; Efremov et al., 2015), RyR homologs have a significant percentage of sequence similarity, especially in structured regions such as the solenoid domains, with the highest sequence identity in the transmembrane domain (TMD). For example, human and rabbit RyR1 sequences are found to have 96.6% identity. Clustal Omega analyses revealed a high degree of similarity for the RyR isoforms RyR1 to RyR3 with percentages of more than 73%, while the identities are slightly lower at about 66%. In the divergent regions observed in the multiple alignments, more than 100 residues in the RyR3 sequence between the SPRY2 and SPRY3 domains are absent. Interestingly, even within these divergent regions, some conserved amino acids were highlighted. For example, a glutamate-rich loop (residues 1875–1921), in which some charged residues are conserved (particularly in the sequences of the RyR1 isoform) (**Supplementary Figure S1A**) or the Big Loop (BL, residues 4254–4539) connecting the cytoplasmic and transmembrane domains (**Supplementary Figure S2**). DISOPRED (Ward et al., 2004) predicted the BL region to be unstructured. Moreover, it was not resolved in any cryo-EM electron density map available when constructing our models. These regions have a particular amino acid composition that we analyzed further. A sequence logo plot clearly shows pronounced conservation within the glutamate-rich region of the loop at residues 1875 to 1921. For the disordered BL loop (residues 4254–4539), the sequence logo shows a variety of features, for example, the presence of a putative PE_5 Ca^{2+}binding site motif, hydrophobic and charged residues with a polyarginine repeat (that are either conserved or share considerable similarity), and many glycine residues known for their flexibility (**Supplementary Figure S1B**). The multiple sequence alignment with Clustal Omega revealed a high percentage of residue conservation for the BL region in homologs: for example, a 94% sequence identity for RyR1_BL human and RyR1_BL rabbit.

3.1.2 Clues for an Additional Transmembrane Helix in the Big Loop Region

The multiple sequence alignment raises the question of the structural interpretation of the BL residue range (4254–4539). We wondered about the role of this disordered region connecting cytoplasmic and transmembrane parts. Since no experimental structural data was available, we performed an *in silico* study of this region. To this end, we analyzed the intrinsic disorder and predicted possible secondary structure elements as well as the transmembrane region, which indicated the possibility of an additional transmembrane helix in the BL region (**Supplementary Figure S3**). When we began this study, ambiguous information about the presence of a transmembrane helix was available from cryo-EM. Below, we review the various observations from our investigations that support such a finding.

3.1.2.1 Secondary Structures and Propensity to Embed into the Membrane

PPSIPRED predicted five putative alpha-helices within the big disordered region (see **Supplementary Figures S3A,B**). The TOPCONS and CCTOP server predictions for the transmembrane region predicted multiple transmembrane parts. Interestingly, most tools predicted the same auxiliary transmembrane helix in the region of residues 4320 to 4336 as shown in **Supplementary Figure S3C**. To evaluate these results, we attempted to model some parts of this loop using a *de novo* approach with the PEP-FOLD3 package (Lamiable et al., 2016).

3.1.2.2 Structural Predictions Through De Novo *Folding*

The Pep-Fold3 predictions indicate that the conserved regions (residues 4310–4318 and other parts of the BL region previously shown in the sequence logos in **Supplementary Figure S1B**) are most likely in helical conformation (**Supplementary Figure S3D**).

3.1.2.3 Unravelling the Role of the BL Region

The results and predictions for the transmembrane region are consistent with the literature, which suggests that some residues in this region tend to fold into an alpha-helix and possibly embed into the membrane (Van Petegem, 2015; Zalk et al., 2015; Du et al., 2002; Van Petegem, 2017; Santulli et al., 2018; Efremov et al., 2015) (**Supplementary Figure S3C**). An additional arginine-rich helix could be localized at the surface of the bilayer (Hristova and Wimley, 2011; Ulmschneider et al., 2017). The BL region could be interacting with other partner molecules or phospholipids or even be implicated in the oligomeric interactions of the RyR. The propensity of some residues to fold into transient TM helices could also impact function, but this aspect remains somewhat unclear. These hypotheses should be verified experimentally. Comparing the rabbit and human RyR sequences, we can speculate about the role of the BL region. Because of the high identity and similarity percentages, we could use the initial models for this region as structural templates for constructing full-length RyR2 and even RyR3 models.

3.1.3 Model-Building Procedure

Despite many available maps for RyR, complete models of the protein are not yet available. This lack of completed models hampers many efforts to obtain insight into the structure-dynamics-function relationship. In addition, RyR serves here as a use case to show how information from the pool of deposited cryo-EM datasets can be used to complete these models. Given that the number of relatively high-resolution datasets is continuously increasing, we hope that the model building procedures developed here will provide a general reference protocol that can be further adapted and optimized for exploring mechanistic and structural aspects of other systems under investigation.

Figure 2 provides a global schematic overview of the entire modelling procedure. In the following sections, the individual steps for the reconstruction of the receptors are described in detail.

3.1.3.1 Full-Length Models of the Receptor (RyR1) at the Atomic Level

To model missing parts of the protein, several software packages were tested. For the specific use case of the RyR protein, given the challenges presented by its size, the best results were obtained by *de novo* modelling using Rosetta (https://www.rosettacommons.org/) to complete the models. The process is shown schematically in **Supplementary Figure S4**.

To optimize the geometry of these initial models for further computations after *de novo* modelling, we applied some model refinement routines using minimization steps within the Yasara software (Yasara Structure 17.12.24; http://www.yasara.org/) to remove possible clashes, knots, and some bond errors. We then refined the models for all atoms using the MDFF method (https://www.ks.uiuc.edu/Research/mdff/). This step is intended to optimize the newly created loops and side-chain coordinates based on the volumetric data from the experimental electron density maps.

We set up an MDFF routine that allows us to extract a partial volume corresponding to a monomer from a map to fit the model (**Figure 3**). After applying the MDFF procedure to a monomer, the different partial volumes fit well into the full map. However, we found that long loops (e.g., the region of residues 4254–4539) did not converge into the map despite many MDFF runs (**Figure 3** and **Supplementary Figure S5A**). We attribute this lack of convergence to either the large size of these loop regions or to missing densities. Using the BioSpring tool (Molza et al., 2014) at a later stage (see below) improved convergence.

A variant of the MDFF method was required due to the system's size and the computational resources needed. We experimented with a more computationally intensive protocol

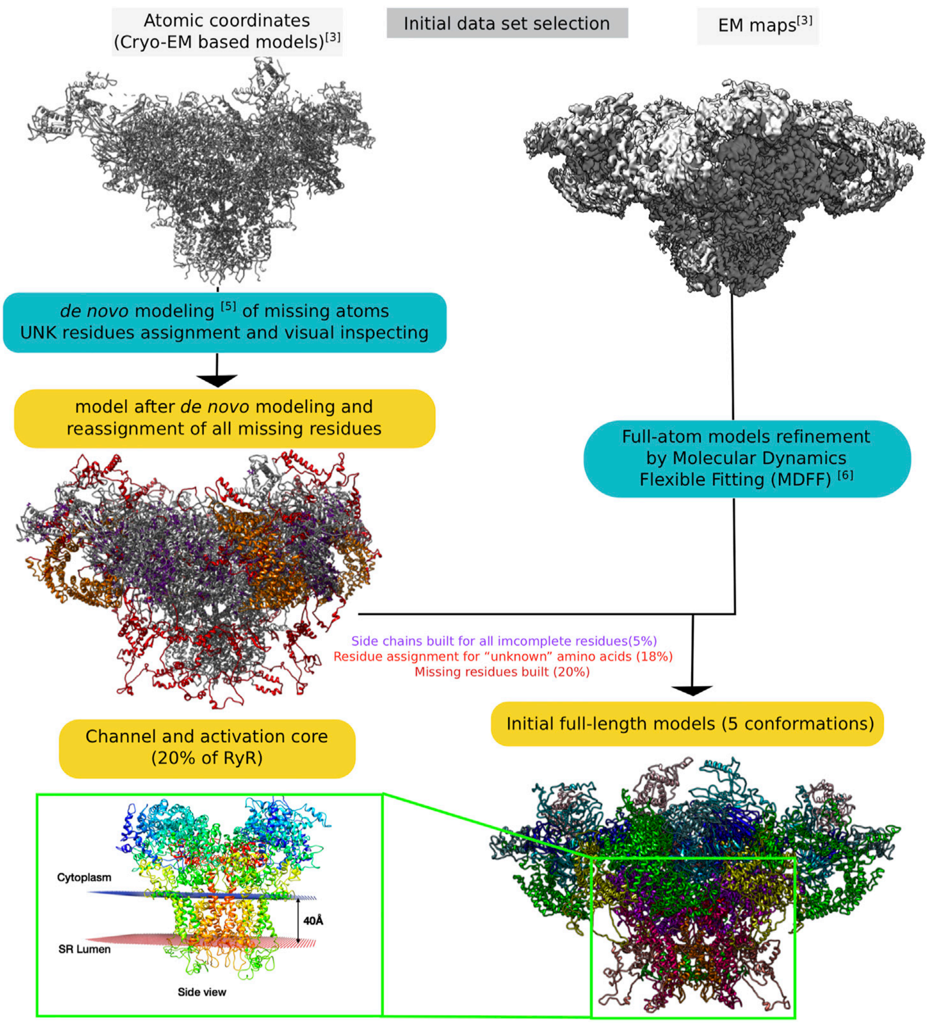

Atomic coordinates
(Cryo-EM based models)[3]
Initial data set selection
EM maps[3]
de novo modeling [5] of missing atoms
UNK residues assignment and visual inspecting
model after de novo modeling and
reassignment of all missing residues
Full-atom models refinement
by Molecular Dynamics
Flexible Fitting (MDFF) [6]
Side chains built for all imcomplete residues(5%)
Residue assignment for "unknown" amino acids (18%)
Missing residues built (20%)
Initial full-length models (5 conformations)
Channel and activation core
(20% of RyR)
Cytoplasm
40Å
SR Lumen
Side view

FIGURE 2 | Schematic representation of the entire process of reconstructing the RyR1 model from raw electron microscopy maps to full-length whole-atom models.

to fit the entire tetramer model into the map. Fitting each monomer into a subvolume is a good choice for a big system. However, if sufficient technical resources are available, fitting the whole system is preferable, as it also accounts for the interactions between the subunits.

The protocol is shown in **Figure 3**. It can be summarized for the monomer case as follows: Preparing the subvolume, fitting the monomer model to the subvolume and generating the tetramer model, checking that there is no overlap between subunits, and making any necessary corrections. For modelling the full tetramer (without the option of imposed symmetry), one first has to prepare the tetramer model and files for MDFF, then fit the model to the experimental map (in our experiments, this took about eight times longer than in the monomer case), check that there are no "aberrant" fits, and adjust the coordinates if necessary.

This step was followed by a visual inspection of the residues, especially residues modelled from "UNK" residue types. Some limitations are due to low-density regions or difficulties in fitting long loops. When necessary, some loops (e.g., in residue regions 4290–4299, 3064–3133) were shifted using the interactive and flexible method BioSpring previously developed in the laboratory (Molza et al., 2014), as shown in **Supplementary Figure S5B**. Namely, these regions were aligned to the corresponding densities of the neighbouring subunits.

In our final models used for subsequent analysis, the additional transmembrane helix was included for the apo, activated and locked states based on the evidence we found in the cryo-EM densities, whereas there was no density to model it into the primed and activated/intermediate states, despite the fact that the helix was present in PDB 5TAP.

3.1.3.2 Assessment of Stereochemical Model Quality

Molprobity (Williams et al., 2018) was used to assess the structural quality of the models compared to the starting

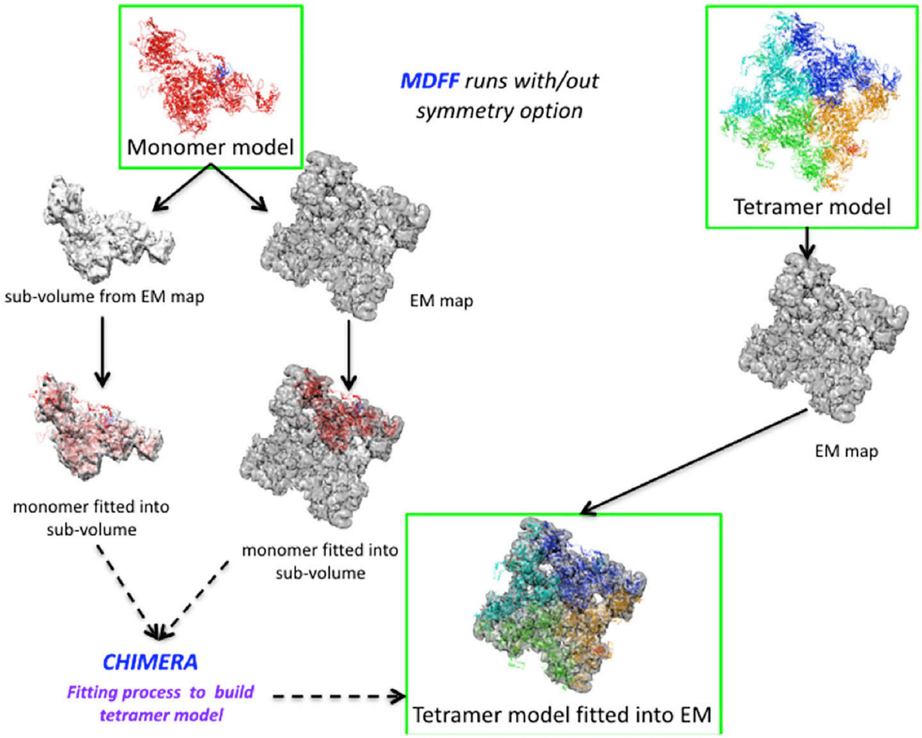

Monomer model
MDFF runs with/out symmetry option
Tetramer model
sub-volume from EM map
EM map
EM map
monomer fitted into sub-volume
monomer fitted into sub-volume
CHIMERA
Fitting process to build tetramer model
Tetramer model fitted into EM

FIGURE 3 | Runs performed to produce the MDFF calculations. Two different configurations are described in the Figure, either by fitting the monomer model to a partial volume or by fitting the tetramer model to the entire density map. The MDFF runs outlined with green rectangles represent a good compromise between the fit achieved and time to solution (left path) or a more direct, but computationally intensive way (right path) to obtain a tetramer model fitted to experimental maps.

TABLE 3 | Molprobity validation results (% residues) for the models derived from the five structural RyR1 templates used.

State	Structure	Favored regions	Allowed	Outliers	Rotamer outliers	C_β outliers
apo	5tb0	88.80	11.08	0.12	0.50	0
	noloops model	88.52	11.20	0.28	0.20	0
	full model	86.92	12.75	0.33	0.19	0
primed	5t15	88.24	11.49	0.27	0.45	0
	noloops model	91.25	8.51	0.24	0.06	0
	full model	89.42	10.33	0.25	0.05	0
activated/ intermediate	5tap	89.92	9.90	0.18	0.54	4
	noloops model	89.21	10.46	0.33	0.26	0
	full model	87.18	12.47	0.35	0.22	0
activated	5ta3	89.09	10.76	0.15	0.54	0
	noloops model	89.40	10.36	0.24	0.17	0
	full model	87.32	12.41	0.27	0.17	0
locked	5taw	89.53	10.30	0.17	0.39	0
	noloops model	90.85	8.95	0.20	0.21	0
	full model	88.96	10.77	0.27	0.23	0

PDB structures, distinguishing the full models from those (“no-loops”) without modelled loops (red regions of **Supplementary Figure S4**). The results are summarized in **Table 3** and show systematic improvement for some metrics, whereas others fluctuate or degrade. The C_β outliers in the 5tap conformation were corrected by our procedure, and rotamer outliers were systematically reduced for all models and fall within the recommended limit of 0.3. While the backbone outliers would normally be less than 0.2%, we observed several cases with higher values. The percentage in favoured Ramachandran regions should be above 98%, but does not go beyond 91.25% in the deposited structures in the PDB or in our models. In summary, despite the addition of a large number of missing structural data, the model quality did not significantly degrade, and even systematically improved for some metrics.

3.1.3.3 Assessment of the Overall Assembly Through Coarse-Grained Simulations

The modelling steps described above resulted in all-atom structures of the RyR channel that were significantly enhanced with respect to the starting structures (**Supplementary Figure S6A**).

An important question is whether the molecular assemblies generated constitute a viable tetrameric channel in a membrane environment. To address this question, we performed Coarse-Grained Molecular Dynamics (CGMD) simulations. Two types of models of different sizes were used for this study: a **channel core model** centered on the RyR channel and a **full-length model** of the receptor. The channel core includes the core solenoid, the transmembrane (including the pore), and the C-terminal domains; the interaction of all these domains forms the ligand-binding sites of RyR1. This system comprises residues 3667 to 5037 (biological numbering) of each monomer. The complete model contains all residues in the RyR protein (residues 1 to 5037 for each monomer). To avoid bias in the coarse-grained simulations, we experimented with the presence or absence of the unstructured BL region that connects the cytoplasmic with the transmembrane domain. No connections or restraints were added between the monomers of the channel in either model.

Microsecond-scale CGMD simulations of the full RyR channel in DOPC/DOPE membrane and water simulations (500,000 particles, **Supplementary Figure S6B**) showed good stability over this period. A plot of the center-of-mass distance for all pairs of subunits is shown in **Supplementary Figure S6C**. Extended CGMD simulations (1–10 μs) were performed for the RyR1 channel core (3667–5037 of each monomer), the channel core excluding the BL region (deleted residues: 4254–4539) for the five RyR channel states of **Table 2**, and the full-length apo-state model. The protein architecture remained stable during the simulations, as shown in **Supplementary Figure S6**, confirming the interactions between the subunits of the tetramer and thus indirectly validating the model reconstruction and monomer interfaces. Specifically, the six helices (**Figure 1**) maintained their integrity during MD, with the highest fluctuation amplitude in the loop regions, especially in the S1-S2 loop. The distance between each monomer of the channel remained stable, with the radius of gyration varying between 10.6 and 10.8 nm (the variation may be due to the interaction with lipids). The bilayer thickness fluctuated little (38.7 ± 0.3 Å). In the future, we plan to transfer the systems from CG to the atomistic level and perform molecular dynamics for all atoms to achieve an accurate structural characterization of the RyR1 dynamics.

3.1.4 Exploration of the Dynamics of the RyR Channel Models

3.1.4.1 Studying RyR1 Deformations Through Normal Mode Analyses

Normal modes analysis (NMA) allows one to examine the dynamics of a given structure given the 3D coordinates of its component atoms, which in this case are obtained from our models. NMA is based on a harmonic approximation of the molecule's potential energy in the region of a minimum in the potential energy surface, which is expressed as a function of changes in the $3N$ coordinates of the structure's N atoms. It provides a detailed description of the vibrational dynamics associated with small perturbations of the minimum-energy structure. NMA notably provides a set of normal mode vectors, each of which consists of $3N$ components representing the amplitudes of atom movements for a single vibrational mode, plus the set of associated frequencies or deformation energies. In practice, the low-frequency (or low-energy) modes are of the most interest, as they have been extensively correlated with directions of conformational changes and dynamics in biological macromolecules (e.g., Tama and Sanejouand, 2001; Rueda et al., 2007).

Here we used NMA to better understand the dynamics of our models of the ryanodine receptor in its different conformational states and how these movements differ from one state to another. For such large molecules, the analyses were facilitated through the use of the NOLB software (Hoffmann and Grudinin, 2017). We performed NMA on the full-length RyR-1 models and the restricted channel regions after removing the modelled loop regions (structure "no-loops" in **Table 3**). These calculations were carried out on the five conformational states of RyR-1 (**Table 2**): the closed or closed-like states apo and primed, the intermediate state, and the open states activated and locked.

A variety of definitions for RyR's motions have been used in the literature. Here we defined specific geometric characteristics of protein subdomain movements that characterized RyR vibrational modes. Three types of movements in particular were used to categorize the motions observed when visualizing the NM vibrational modes, and are designated here as twisting, breathing, and blooming motions. **Supplementary Figure S7** shows schematic diagrams of these motions in RyR.

3.1.4.2 Contributions of NM Movements to RyR Conformational Transitions

For the following analyses, we used the models restricted to the channel region, which encompasses residues 3667 to 5037 of each chain. The structural similarity of the different channel states was assessed by the C_α RMSD after superposition for all model structure pairs (**Table 4**). Of the channel regions of the five models, the primed and intermediate structures are the closest (1.4 ÅC_α distance), followed by the models of the locked and activated states (1.6 Å), respectively. All other model structure pairs are between 2.4 and 2.9 Å apart.

NMA suggests how the vibrational dynamics of the macromolecule may contribute to a given conformational transition. One way to see this is by calculating the linear combination of low-frequency vibrational mode vectors (atom movements) of the starting structure A that minimizes the least-squares distance to the final structure B (Hoffmann and Grudinin, 2017; Moal and Bates, 2010). **Figure 4A** shows an example of using this approach to studying the activated → intermediate transition The linear combination of normal-mode vectors for the 100 lowest-frequency modes of the activated conformation reduces the RMS distance to the intermediate

TABLE 4 | C_α rmsd in Angstroms for pairs of RyR model channel core models (calculated without the flexible BL region).

	Apo	Primed	Activated/intermediate	Activated	Locked
Apo	-	2.8	2.5	2.7	2.9
Primed		-	1.4	2.5	2.6
Activated/intermediate			-	2.4	2.5
Activated				-	1.6

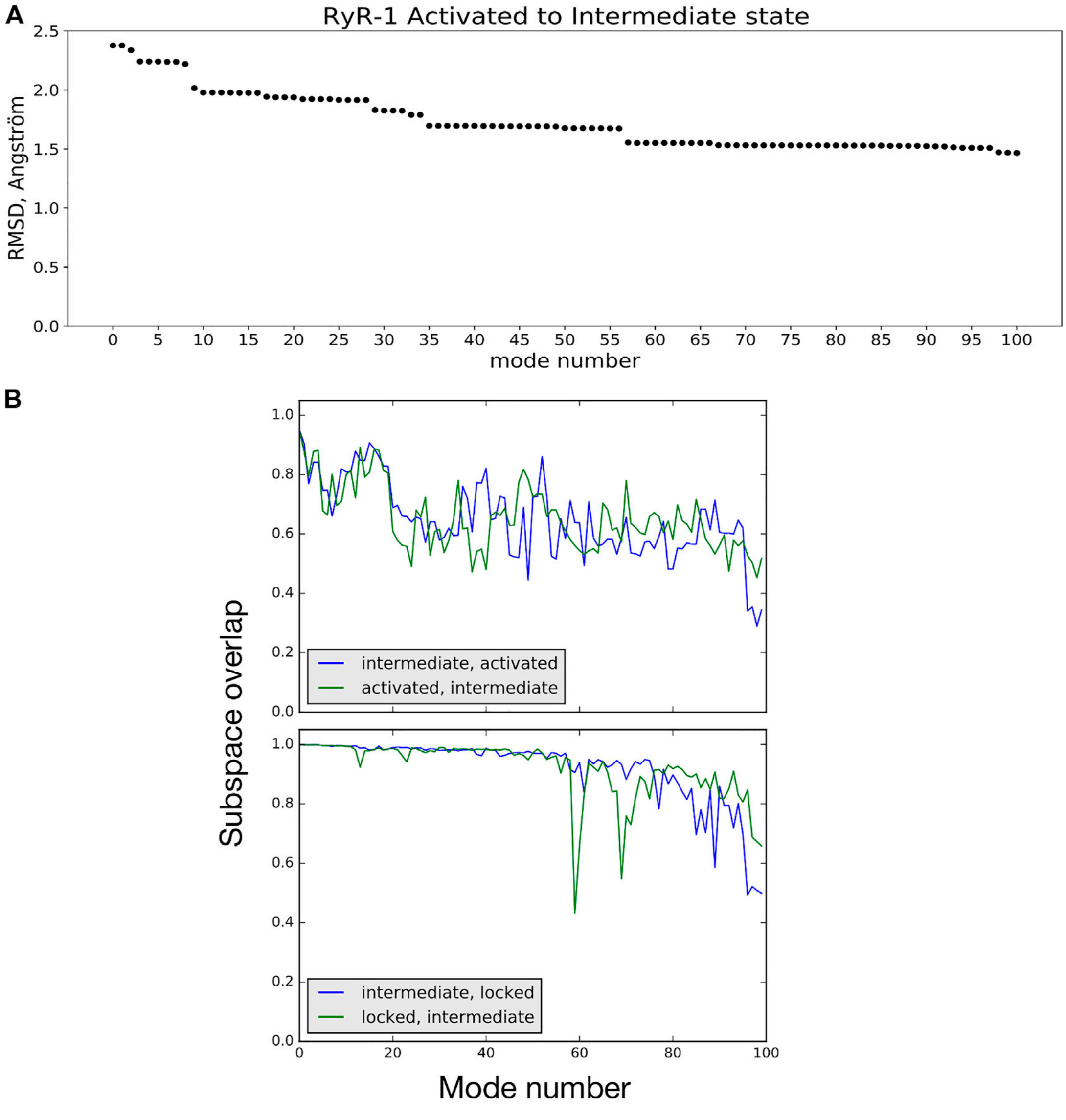


FIGURE 4 | (A) RMSD between the activated and intermediate conformational states is reduced by 38% by displacing atoms along a linear combination of NM vectors. Mode 0 refers to the initial RMSD between the two model structures. **(B)** Subspace-overlap analysis for the 100 lowest-frequency normal mode vectors of selected pairs of conformational states of the RyR1 receptor. *Top:* In blue, overlap of each intermediate state mode on the subspace of the 100 activated-state modes. In green, overlap of each activated state mode on the subspace of the 100 intermediate-state modes. *Bottom:* Same coloring as in *Top*, but for the intermediate and locked state modes.

conformation from 2.38 to 1.46 Å (38%), while the first 50 reduce the RMSD by 29%. As is often the case for conformational changes in biological macromolecules (Moal and Bates, 2010), a relatively small number of low-frequency NM movements participate in this transition; mode 9 alone contributes 10% of the total RMSD decrease. A short animation of the motion

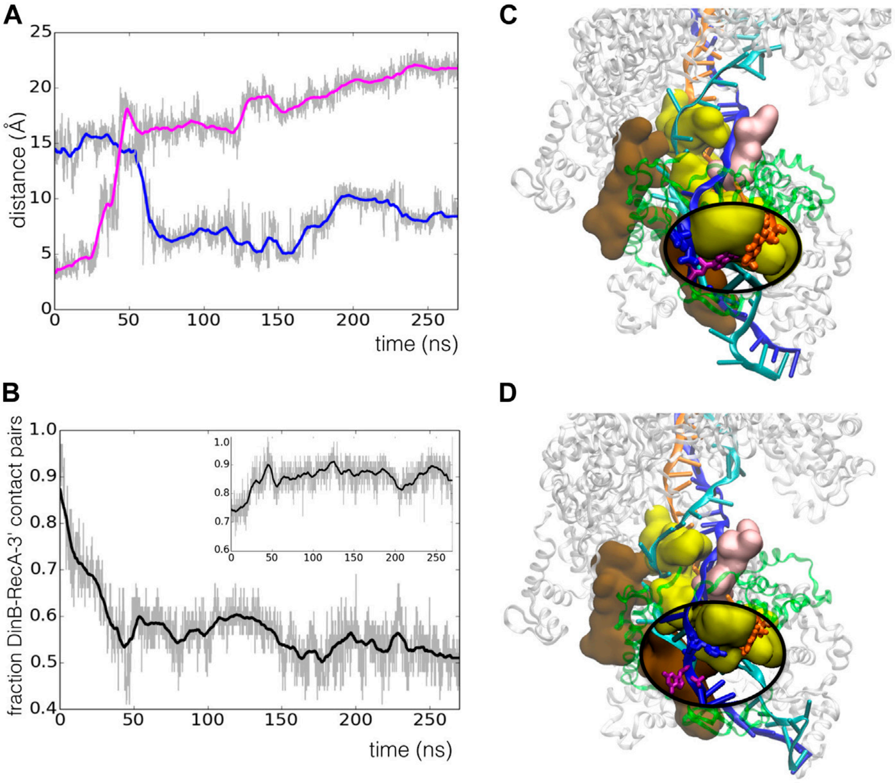


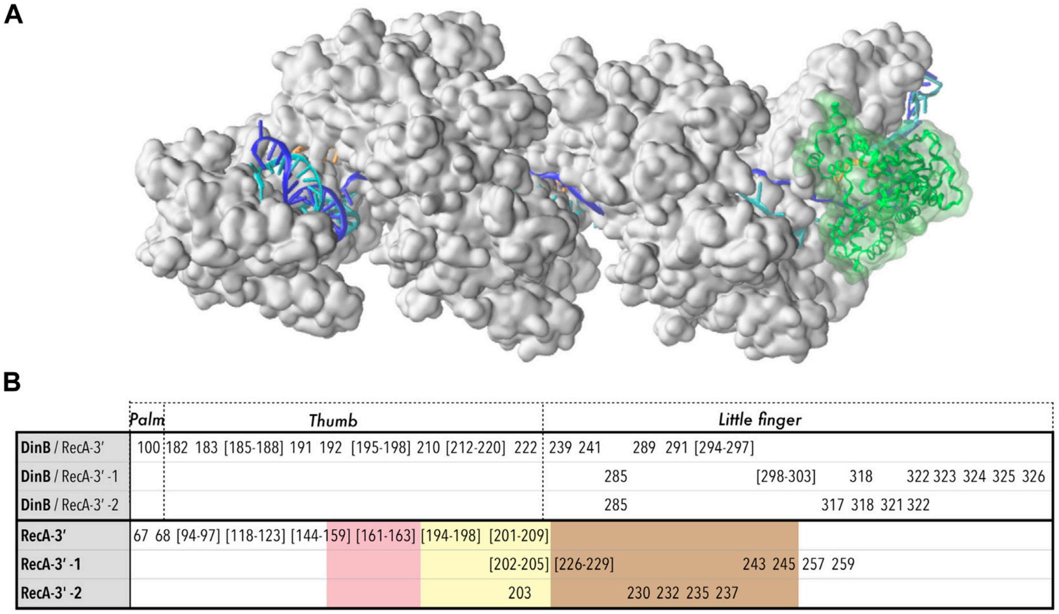
FIGURE 5 | Spontaneous register shift to the correct base pairing in the RecA*-DinB complex during 270 ns MD simulation. **(A)** Time evolution of the distance between atom O4 of T_{53} in the primer strand and atom N1 of either A_{12} (blue line) or G_{11} (magenta line) in the template strand; G_{11}^{templ}, initially close to T_{53}^{prim}, rapidly separates from that thymine after from 30 ns, while A_{12}^{templ} initiates a quick approximation to the thymine starting from 50ns simulation time; T_{53}^{prim}:A_{12}^{templ} is the last base pair between the primer and the template strands in 3′, while G_{11} is the first base of the template strand that will be used in the primer elongation process. **(B)** Time evolution of the fraction of residue-residue contacts between DinB and the last RecA monomer in the initial structure, that is conserved during the MD run; the inset shows the evolution with respect to the structure at 45 ns (vertical broken line) **(C,D)** Snapshots of the system taken at 20 ns **(C)** and 145 ns **(D)** simulation time. RecA proteins and DinB are, respectively, represented with white and green ribbons; motifs from the last three RecA proteins that are involved in the interface with DinB are shown in surface representation and colored using the codes given in **Figure 6B**; the DNA strands are represented in cartoon with color codes as described in **Figure 6A**. In both **(C)** and **(D)**, an oval insert with magnifying glass effect enables visualizing the bases referred to in **(A)**, represented in licorice with T_{53}^{prim} in orange, A_{12}^{templ} in blue and G_{11}^{templ} in magenta; in order to permit visualizing the bases which are deeply buried in DinB catalytic cleft, DinB is not represented in the inserts.

described by the low-frequency NM contributions to this transition is available (Robert and Molza, 2021). These NM contributions describe a longitudinal compression of the channel: bending of helix S6c and movement of the C-terminal domain act to reduce the opening on the cytoplasmic side, while the transmembrane domain twists about the channel axis and descends relative to the rest of the receptor, reducing the volume of the central cavity enclosed by helix S6 (breathing motion).

The NMA thus showed that the intrinsic, low-energy vibrational movements of the RyR channel in the active conformation already encode atom movements that tend to lead to a conformational transition to the intermediate state. This could be then used to obtain further insight into associated features such as domain movement, ligand-binding cavity opening/closing, etc. The NM motions thus reveal an energetic and dynamic rationale for previous depictions of conformational differences between cryo-EM structures (Georges et al., 2016; Willegems and Efremov, 2018), depictions that alone do not provide information for judging the energetic or dynamic favorability of inferred movements.

3.1.4.3 Comparing Vibrational Motions of the RyR1 Channel as a Function of State

NMA can also be employed to assess the similarity of two conformational states of a protein in terms of their vibrational dynamics. To demonstrate this, we extracted the movements of common C_α atoms from the all-atom modes of different channel models calculated using NOLB. These will be referred to as C_α modes in the following. The plots shown in **Figure 4B** compare the intermediate and activated state models (top) and the intermediate and locked state models (bottom). In these plots, each of the 100 C_α mode vectors from one model *i*, arranged in order of increasing frequency, is projected onto the space spanned by the first 100 vibrational modes of the second model *j*, and vice versa. The intermediate and active state models can be seen to share about 70% of their low-frequency mode subspaces. On the other hand, the intermediate and locked states share a higher



	Palm	Thumb	Little finger
DinB / RecA-3′	100	182 183 [185-188] 191 192 [195-198] 210 [212-220] 222	239 241 289 291 [294-297]
DinB / RecA-3′ -1			285 [298-303] 318 322 323 324 325 326
DinB / RecA-3′ -2			285 317 318 321 322
RecA-3′	67 68 [94-97] [118-123] [144-159] [161-163] [194-198] [201-209]		
RecA-3′ -1		[202-205]	[226-229] 243 245 257 259
RecA-3' -2		203	230 232 235 237

FIGURE 6 | Model structure of the RecA*-DinB complex, after 270 ns molecular dynamics simulation. **(A)** The protein part of the RecA* filament (close to three filament turns) is formed by 17 RecA proteins shown in surface representation, in white. The filament binds three DNA strands (cartoon representation) organized as a D-loop: the ssDNA bound to the filament in site I, or primer (orange) binds the complementary or template strand (blue) in the region where that strand is incorporated in the filament, forming a stretched/unwound heteroduplex; beyond the initial dsDNA entry point in the filament in 5′ and its exit point in 3′, the complementary strand remains paired with its initial partner or leading strand (cyan) in a relaxed B-form. The junctions between the stretched and relaxed regions form kinks. The DinB protein is represented in cartoon and transparent surface, in green. **(B)** Residues involved in the interface between DinB and the last three RecA monomers starting from the 3′ extremity (RecA-3′, RecA-3′ -1, RecA-3′ -2). For DinB, correspondence with the "Palm", "Thumb" and "Little finger" domains is reported; for RecA, the residues belonging to loops L1 (pink), L2 (yellow) or the LexA binding loop (marron), are indicated using color code shading. The same color code is used in **Figure 5**.

percentage of their low-frequency vibrational movements, particularly for the first 50 modes. This example demonstrates that comparing the vibrational dynamics of coherent models of the different states of the RyR receptor clearly provides a richer comparison than simple rmsd: the intermediate state is about the same rms distance from the activated state (2.4 Å) as it is from the locked state (2.5 Å), but the differences in vibrational dynamics of the models are more nuanced. Further investigation into comparisons of the RyR1 vibrational dynamics as a function of state are ongoing.

3.2 Integrative Modelling of DNA Polymerase Binding to RecA Nucleofilaments

We now provide an example in which refinement and initial validation of a structural model and dynamic exploration of the model go hand in hand. Here, the challenge is to validate the construction of an intermediate species in the final stage of homologous recombination (HR): the coupling of RecA-induced DNA strand exchange to DNA synthesis. As will be seen, the dynamic exploration is performed using classical molecular dynamics simulations rather than the harmonic dynamics (normal modes) approach employed above for the ryanodine receptor. Molecular dynamics appears to be well adapted to probing the modelled intermediate in this highly dynamic process because of the strong mechanical coupling among the components.

It is useful at this point to briefly present the HR process, largely conserved from prokaryotes to humans, along with the key players of this process relevant to the present modelling. Homologous recombination is a highly complex, multicomponent process that intervenes in the cell to assure the repair of double-strand DNA breaks, which are typically lethal (Bell and Kowalczykowski, 2016). The two broken ends are first processed to form a single-stranded DNA tail (ssDNA) on either side of the break. To repair the break, a helical nucleoprotein filament is formed by the oligomerization of protein monomers on these ssDNA tails. In bacterial HR, a well-studied model system, this protein is RecA. The nucleoprotein filament systematically binds and searches double-stranded DNA (dsDNA) in the genome to find a sequence match with the ssDNA. Within the HR filament, the two DNA species (ssDNA and dsDNA) are positioned near each other, allowing mutual probing of their sequences for homology, followed by base pairing and strand exchange in the case of homology (Prentiss et al., 2015; Bell and Kowalczykowski, 2016). This probing and proofreading process takes advantage of frequent reversals of strand exchange (Danilowicz et al., 2021), but this phase terminates when a DNA polymerase takes over, pushing the strand exchange reaction towards irreversibility (Lu et al., 2019) by using the dsDNA complementary strand as a template to start elongating the ssDNA at its 3′ extremity. The subject of the modelling in this example is the interaction of the RecA-DNA filament with the DNA polymerase DinB.

3.2.1 Building-Block Scaffold

The modelling steps described here consist of the refinement and validation of a coarse-grained scaffold structure defined in a preliminary modelling study (Tashjian et al., 2019) of the RecA*-DinB complex, which itself was built on experimental evidence that the coupling between recombination and synthesis occurs via direct binding of the DNA polymerase DinB to the RecA* filament (RecA nucleoprotein filament with three bound DNA strands) (Henrikus et al., 2020) (see also Godoy et al., 2007). This structure was a coarse-grained model obtained using BioSpring, an interactive modelling facility also used in the RyR1 modelling presented above. The structure also integrated a structural building block RecA* derived from the model of early RecA-ssDNA-dsDNA intermediate created in 2015 (Yang et al., 2015). Indeed, the striking similarity between that RecA* model and the recently published Cryo-EM structure of short post-strand-exchange intermediates (Yang et al., 2020) provided further impetus for the current integrated modelling.

3.2.2 Coarse-Grained to All-Atom

While the coarse-grained representation used in BioSpring presents a direct correspondence to all-atom representation, which allows easy recovery of an all-atom model, that model needed refinement. This was notably the case for the DinB building block, for which the spring network used in BioSpring had allowed distortions in the secondary structures during the tight assembly process.

We reintroduced individual secondary structure elements and loops taken from the DinB crystal structure as described in the *Methods* section; thereafter energy minimization was sufficient to regularize their geometries.

3.2.3 Mechanical Coupling Within RecA*-DinB Complex

We subjected the all-atom starting model to 270 ns of molecular dynamics simulation in explicit solvent. Over the timecourse of the simulation, the filament flexes somewhat, with the overall end-to-end distance reducing by 10%, but no tendency for dissociation of the polymerase was observed (**Figure 6A**); the total number of DinB-RecA* interface contacts decreased by about 15% in the first 70 ns, then remained stable throughout the last 200 ns of the trajectory (**Supplementary Figure S8**). This indicates that the large majority of DinB/RecA contact regions from the preliminary modelled structure (Tashjian et al., 2019) were preserved in the highly crowded environment of the complex.

However, we observed spontaneous relaxation that took place during the simulation trajectory through concerted rearrangements of RecA motifs and DNA strands. The starting model presented a mismatch in the last heteroduplex base pair in the 3′ end due to an artificial shift in the heteroduplex register resulting from the preliminary modelling process, which had shifted the last several bases of the template strand with respect to the primer, such that G_{11}^{templ} interacted with T_{53}^{prim}, the last 3′-base of the primer strand (**Figure 5B**). From the point of view of sequence, G_{11} is the first base of the template strand that should assist the synthesis of the first added nucleotide in the polymerase catalytic site to extend the primer sequence (**Figure 5C**). **Figure 5A** shows how the mismatched T_{53}^{prim}:G_{11}^{templ} heteroduplex base pair spontaneously separated after about 50 ns, rapidly followed by the formation of the correct T_{53}^{prim}:A_{12}^{templ} base pair (shown in **Figure 5D**). This pairing shift comes with a sliding of the DNA template strand backbone (in blue in **Figures 5C,D** along the L2 loop of the last RecA monomer (in yellow in **Figures 5C,D**). This loop is deeply inserted in the DinB catalytic cleft. **Figure 5B** shows that this sliding occurs simultaneously with a shift in the interface between DinB and the terminal RecA monomer, that also takes place at 50 ns. At the same time, the interface loses up to 50% of its initial residue pair contacts, stabilizing to a new network of residue-residue contacts that is 86% conserved on average during the last 225 ns of the simulation (**Figure 5B**, insert). Recent studies, both on RecA filaments and on heterodimers, showed that stable interfaces typically conserve 70–90% of their interaction contact network when simulated at 300 K, due to thermal movements (Boyer et al., 2019; Prévost and Sacquin-Mora, 2021).

3.2.4 RecA*-DinB Interface

Figure 6B lists the residues that participate in the interface between DinB and the last three RecA monomers at the filament 3′ extremity. The list has been established from a MD structure at 120 ns simulation time, therefore after the reorganization took place. Two DinB domains mainly participate in the three DinB/RecA interfaces, namely the thumb and the little finger domains, with the palm domain showing only marginal contribution (for definitions of these domains see **Figure 6**). Decomposing the contributions of DinB domains and RecA motifs testifies to the intricate association of DinB within the filament interior. While the thumb domain only interacts with the last RecA monomer at the 3′ end, the little finger domain contributes to all three DinB-RecA interfaces. Three residues of the little finger domain, namely Arg85, Lys318 and Trp322, simultaneously interact with the two RecA monomers situated 5′ of the terminal RecA. On the RecA side, the interface between DinB and the terminal RecA encompasses most of the L1 and L2 loops, flexible loops that participate in the binding of the DNA strands in the HR filament. Although the L2 loop is deeply inserted in the DinB catalytic cleft, its presence remains compatible with that of the primer and template strands while allowing strand adjustments during the MD trajectory as described above. L2 loops of the two non-terminal RecA monomers also partly contribute to the interface with DinB, which also includes the so-called LexA-binding loop, a hairpin motif that contributes to forming the secondary DNA binding site in the filament.

The model we present here of a putative intermediate of association between the RecA nucleoprotein filament and the DinB DNA polymerase offers a link between the recombination and the DNA synthesis activities in HR. *In vitro* studies have previously demonstrated that DinB associates with isolated RecA proteins, and the DinB interface with RecA has been characterized (Godoy et al., 2007) (See also **Supplementary Figure S9**). Most interestingly, the regions (or patches) of the DinB surface that interact strongly with RecA were mainly observed within the catalytic cleft, in the little finger (patch *Pa1*: 287–298) and in the palm domains (patch *Pa2*: 145–160). Our model of DinB interacting with RecA*/DNA identified patch *Pa1* as a major contributor to the interface between DinB and the last RecA monomer (**Figure 6**). More exactly, *Pa1* interacts with the L2 loop of that monomer, which is deeply buried in the catalytic cleft. In addition, while patch *Pa2* does not directly contribute to the interface in our model, some of its residues such as Phe147 are found within 6 Åof that same L2 loop. Further approach of the *Pa2* patch to the L2 loop would appear to be sterically hindered by the neighboring RecA monomers. These observations suggest that DinB interacts with RecA in a similar way whether the RecA is part of the nucleoprotein filament or free in solution (Godoy et al., 2007).

4 DISCUSSION

The integrative modelling pipeline we describe here, as summarized in **Table 1**, aims at achieving coherence, completeness, and (meta-)stability in the generated models.

Coherence was a constant concern for building the RyR1 models starting from Cryo-EM data generated in different conditions. It was ensured by separately applying the modelling pipeline to each of the input data sets, resulting in five complete all-atom reference models of functional substates ranging from open to closed conformations. In order to take into account the enormous size of assemblies such as the RyR1 receptor, our protocol focuses on providing a general approach with computationally efficient steps such as interactive modelling and coarse-grained refinement. The pipeline also provides the possibility to limit the computational cost (e.g., by fitting to cryo-EM sub-maps) when performance might be critical. It should be noted that our protocol differs from that of Heinz et al. (2018), whose focus was on quantifying ion permeation and identifying pathways through the channel core. That work used molecular dynamics simulations as the main tool to support the refinement process, which is computationally expensive. It focused on modelling the closed stated based on cryo-EM data from (Efremov et al., 2015; Yan et al., 2015; Zalk et al., 2015), while an open-state model restricted to the central and channel domains was generated from the closed-state model using MD guided by the cryo-EM data of Georges et al. (2016).

4.1 Missing Regions and Interactions

The RyR1 receptor also illustrates how this pipeline addresses the challenges of obtaining complete models from incomplete structural data. All available experimental density maps of RyR1 lack density for large regions of the sequence. This may indicate disorder or flexibility in these regions. However it is now well recognized that even intrinsically disordered, flexible protein regions may play important functional roles. In tackling this problem, we observed that fully automated protocols could not handle such situations. For example, the use of MDFF was not always sufficient for generating the missing regions. Human interventions were required to make decisions and optimize the models, using available experimental data and additional sequence analyses. This phase of the modelling also highlighted the important question of the role of the BL region in RyR1, which exhibits distinct patterns and a high degree of conservation but for which no structural data is available. We hypothesize that this region contributes to the interaction with partner molecules, phospholipids, or other RyR1 subunits. These steps in the pipeline provided evidence in RyR1 for an additional transmembrane region, consistent with suggestions in the literature (Georges et al., 2016; Van Petegem, 2017). The additional TM helix was stable in the MD simulation step and appeared to reinforce intra- and inter-subunit packing of the transmembrane region by providing additional bridging interactions between neighboring subunits as well as between subunits and the membrane. Our initial models for this region could be used in further studies as structural templates for building full-length RyR2 and even RyR3 models.

We emphasize the power that interactive manipulation and visualization offers to this part of the integrative modelling pipeline, in addition to traditional modelling and visualization tools, when confronting such complex problems. Interactive approaches have existed for some time, but have not typically been present (or acknowledged) in integrative modelling studies. In cases such as the RyR1 and the RecA homologous recombination systems described here, the overall modelling pipeline involves results from automatic procedures that are explicitly coupled to, interpreted, and filtered through human interaction. For example, both the RyR1 receptor and the RecA-DNA-DinB complex required assembling entangled regions in the modelled complex. The release of steric clashes in many such cases is a tedious task but one that can be handled efficiently using interactive simulations. It might be pointed out that, in modelling, human input is ubiquitous and nearly unavoidable; a common example is the choice of restraints to be applied in modelling, even in such simple cases as energy minimization. The underlying hypothesis is that the ability to manipulate parts of the model as it is being built– to incorporate feedback on the modelling and to make decisions when there are multiple options or ambiguities– will enable the development of better models. Here, we made use of our in-house tool BioSpring at several points in the pipeline. Such approaches have occasionally been documented in the literature (Molza et al., 2014; Croll, 2018), but currently remain marginal. Since integrative modelling often means seeking the best compromise from different and sometimes conflicting sources of information about a given target, the pipeline we have developed explicitly acknowledges that the human operator in the modelling loop may be relied upon to make informed decisions in accordance with the biological data and the physical laws at play in the chosen modelling approach, e.g., in molecular dynamics simulation, constrained minimization, etc. (Lanrezac et al., 2021).

4.2 Roles of Assessment and Exploration of Model Dynamics

Simulating the dynamic nature of structural intermediates in the modelling process also plays an important role at different steps of our pipeline, where it serves multiple but distinct purposes. Molecular dynamics simulation, whether obtained using all-atom or coarse-grained representations, allows checking the (meta)-stability of a generated model, of a folded protein for example, over a chosen timescale, as in the RyR1 receptor modelling. It also permits identifying dynamic evolution of models, which reflects plausible local dynamics and structural changes, as proved valuable in the RecA* system. However, even without costly exploratory MD simulations, normal modes analysis of model structures can provide insight into differences in their intrinsic vibrational dynamics depending on their functional state. For example, comparing the low-frequency vibrational modes for two different conformational states can provide information on large-amplitude, low-frequency mobility differences (e.g., Thomas et al., 1996; Batista et al., 2011). In the case of RyR1, we could examine such differences by exploiting the coherency obtained by modelling the different states of this receptor in order to establish comparisons. As shown, NMA also offers insight into the directions of conformational transitions between two conformational states, which reflects the power of looking to vibrational mode directions for understanding conformational change in biological macromolecules (Tama and Sanejouand, 2001) and predicting protein-ligand and protein-protein interactions (Moal and Bates, 2010).

Dynamic exploration by MD adds further value to integrative modelling through refinement of modelled structures, which can itself provide both better models and valuable insights into function. For example, when multiple building blocks, each refined independently of the others, are combined into a higher-order assembly, there is unavoidably a stress that develops at the interfaces. Such stress may be classified into two categories. The first is what might be termed a “residual” stress that must be resolved in order to obtain a stable interaction. Examples of this were seen in the steric clashes in the RyR receptor, and similarly in the interaction of the DinB crystal structure with RecA-DNA, when assembling subunits in which loops and other motifs in the interface regions had been modelled independently. In our work these stresses were resolved by interactive molecular dynamics approaches that induced little overall change in the geometry of the assembly itself. In contrast, the RecA system revealed a second category of stress that may be linked to the function of the RecA-DNA-DinB assembly (Lu et al., 2019). Evidence of this “mechanical stress” was seen in the resolution of the mismatch of the last basepair of the heteroduplex DNA bound to RecA through unbiased MD simulation, via a register shift in the DNA pairing which resulted from a collective movement that spontaneously reduced the mechanical stress on the stretched DNA strands in the complex. This stress has been shown to play a major role in the recombination process through single molecule experiments (Conover et al., 2011; Danilowicz et al., 2012, 2014). A precedent for the functional role of stress was also seen in our earlier modelling study of the RecA-DNA interaction at the initial stage of the HR process, in which dynamic exploration of an integrative model resulted in spontaneous strand exchange (Yang et al., 2015), essentially mimicking the very fast event that stands at the heart of the HR process. Such functional conversion of mechanical stress could not have been deduced from static models such as the Cryo-EM structures recently published in Yang et al. (2020). The MD results obtained in the present study similarly suggest that tension-induced collective rearrangements in the DinB-RecA* system may play a role during DNA synthesis in the last step of homologous recombination.

Our modelling pipeline is seen thus to place fluid boundaries between validation, refinement and dynamic exploration steps. Indeed, information obtained in the course of an integrative modelling workflow can be useful to orient further exploration of functional mechanisms related to the complex under study. This was the case for both systems presented here despite their very different characteristics. In the case of the RyR1 channel, validation of the ensemble of conformational substates through the NM analysis provides information about the dynamic relationship between these conformational states and evidence of transitions between them. This information is a step in understanding the nature of the transition pathways between the functional states. In the case of the HR complex, the observations that we made during refinement appear to provide important indications for further mechanistic exploration, while validating the model in terms of its capacity to demonstrate the transmission of stress in the DNA to other components of the assembly.

More generally, workflows such as the one presented here provide insights not only into the structure and sequence properties of studied complexes, but also into their dynamics and function, as is required for drug design. Models obtained in this way can be used, for example, to infer putative drug-binding pockets. But the dynamics of pocket accessibility affects its druggability, and thus dynamic information clearly adds insight, for example for the development of new molecules aimed at restoring the closure of the mutant RyR1 channel without altering its function.

5 CONCLUSION

The increasing flow of experimental structural data calls for a new class of integrative modelling workflows combining a broad range of tools. We have shown that integrative modelling of individual macromolecular building blocks, even for very large systems such as the RyR ion channel, is now possible. Ultimately, the results obtained from such approaches on the RyR family will pave the way to improve our understanding of allosteric long-range gating of channel opening and ligand binding effects, which are essential for drug development to treat RyR channelopathies. Further integrative modelling applications involve how such building blocks are assembled into higher-level organizations. Our case study of the complex homologous recognition system shows that such approaches, requiring tailored modelling tools, can go very far in the exploration of structure-function questions in a reliable fashion, as the building blocks are mechanically coupled. This procedure permits emitting testable hypotheses for challenging systems that are difficult to address experimentally.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

MP, CD, VG-C and CP contributed to the conception of the RecA*-DinB study, A-EM and CP produced the data and analyses on this complex; A-EM, YW, MM, PD, CR and MB contributed to the conception of the RyR study; A-EM and CR produced the data; A-EM, MB and CR performed the analyses on RyR; PD and MB co-supervised A-EM's work, with contributions by MM and YW. A-EM, MB, CHR and CP wrote sections of the manuscript; YW, PD, MM, CD, MP and VG-C participated in the writing; CR, MB and CP supervised the writing; all authors revised the manuscript and approved the submitted version.

FUNDING

We acknowledge the Servier group for funding part of this project. HPC resources from GENCI were obtained via the grant numbers A0070701714, A0080707438 and A0100707438. This work was supported by the French "Initiative d'Excellence" program (Grants "DYNAMO", ANR-11-LABX-0011, and "CACSICE", ANR-11-EQPX-0008), the Chu Family Foundation to MP, and VG-C research was supported in part by a grant to the Northeastern University Skills for Capacity and Inclusion Program by the Howard Hughes Medical Institute through the Science Education Program. MB thanks "Sesame Ile-de-France" for co-funding the display wall used for data analysis.

ACKNOWLEDGMENTS

We are grateful for the cryo-EM data provided to us by Andrew R. Marks. Moreover, we also thank Sergei Grudinin for useful help with and update on the NOLB software. Computational work was performed using HPC resources from GENCI and LBT-HPC (thanks to the support from Geoffrey Letessier).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.826136/full#supplementary-material

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

Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D., and Urzhumtsev, A. (2013). Bulk-solvent and Overall Scaling Revisited: Faster Calculations, Improved Results. *Acta Crystallogr. D* 69, 625–634. doi:10.1107/S0907444913000462

Baldwin, P. R., Tan, Y. Z., Eng, E. T., Rice, W. J., Noble, A. J., Negro, C. J., et al. (2018). Big Data in cryoEM: Automated Collection, Processing and Accessibility of EM Data. *Curr. Opin. Microbiol.* 43, 1–8. doi:10.1016/j.mib.2017.10.005

Batista, P. R., Pandey, G., Bisch, P. M., Pascutti, P. G., Perahia, D., and Robert, C. H. (2011). Free Energy Profiles along Consensus normal Modes Provide Insight into HIV-1 Protease Flap Opening. *J. Chem. Theor. Comput.* 7, 2348–2352. doi:10.1021/ct200237u

Batista, P. R., Robert, C. H., Maréchal, J.-D., Ben Hamida-Rebaï, M., Pascutti, P., Bisch, P. M., et al. (2010). Consensus Modes, a Robust Description of Protein Collective Motions from Multiple-Minima normal Mode Analysis–Application to the HIV-1 Protease. *Phys. Chem. Chem. Phys.* 12, 2850–2859. doi:10.1039/b919148h

Bell, J. C., and Kowalczykowski, S. C. (2016). RecA: Regulation and Mechanism of a Molecular Search Engine. *Trends Biochem. Sci.* 41, 491–507. doi:10.1016/j.tibs.2016.04.002

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., et al. (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods Mol. Biol.* 1374, 23–54. doi:10.1007/978-1-4939-3167-5_2

Boyer, B., Danilowicz, C., Prentiss, M., and Prévost, C. (2019). Weaving DNA Strands: Structural Insight on ATP Hydrolysis in RecA-Induced Homologous Recombination. *Nucleic Acids Res.* 47, 7798–7808. doi:10.1093/nar/gkz667

Conover, A. J., Danilowicz, C., Gunaratne, R., Coljee, V. W., Kleckner, N., and Prentiss, M. (2011). Changes in the Tension in dsDNA Alter the Conformation of RecA Bound to dsDNA–RecA Filaments. *Nucleic Acids Res.* 39, 8833–8843. doi:10.1093/nar/gkr561

Croll, T. I. (2018). ISOLDE : a Physically Realistic Environment for Model Building into Low-Resolution Electron-Density Maps. *Crystallogr. D* 74, 519–530. doi:10.1107/S2059798318002425

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a Sequence Logo Generator. *Genome Res.* 14, 1188–1190. doi:10.1101/gr.849004

Danilowicz, C., Feinstein, E., Conover, A., Coljee, V. W., Vlassakis, J., Chan, Y.-L., et al. (2012). RecA Homology Search Is Promoted by Mechanical Stress along the Scanned Duplex DNA. *Nucleic Acids Res.* 40, 1717–1727. doi:10.1093/nar/gkr855

Danilowicz, C., Peacock-Villada, A., Vlassakis, J., Facon, A., Feinstein, E., Kleckner, N., et al. (2014). The Differential Extension in dsDNA Bound to Rad51 Filaments May Play Important Roles in Homology Recognition and Strand Exchange. *Nucleic Acids Res.* 42, 526–533. doi:10.1093/nar/gkt867

Danilowicz, C., Vietorisz, E., Godoy-Carter, V., Prévost, C., and Prentiss, M. (2021). Influences of ssDNA-RecA Filament Length on the Fidelity of Homologous Recombination. *J. Mol. Biol.* 433, 167143. doi:10.1016/j.jmb.2021.167143

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. Theor. Comput.* 9, 687–697. doi:10.1021/ct300646g

Dhindwal, S., Lobo, J., Cabra, V., Santiago, D. J., Nayak, A. R., Dryden, K., et al. (2017). A Cryo-EM-Based Model of Phosphorylation- and FKBP12.6-mediated

Allosterism of the Cardiac Ryanodine Receptor. *Sci. Signal.* 10, eaai8842. doi:10.1126/scisignal.aai8842

Dobson, L., Reményi, I., and Tusnády, G. E. (2015). CCTOP: a Consensus Constrained TOPology Prediction Web Server. *Nucleic Acids Res.* 43, W408–W412. doi:10.1093/nar/gkv451

Du, G. G., Sandhu, B., Khanna, V. K., Guo, X. H., and MacLennan, D. H. (2002). Topology of the Ca2+ Release Channel of Skeletal Muscle Sarcoplasmic Reticulum (RyR1). *Proc. Natl. Acad. Sci. U S A.* 99, 16725–16730. doi:10.1073/pnas.012688999

Durand, P., Trinquier, G., and Sanejouand, Y.-H. (1994). A New Approach for Determining Low-Frequency normal Modes in Macromolecules. *Biopolymers* 34, 759–771. doi:10.1002/bip.360340608

Efremov, R. G., Leitner, A., Aebersold, R., and Raunser, S. (2015). Architecture and Conformational Switch Mechanism of the Ryanodine Receptor. *Nature* 517, 39–43. doi:10.1038/nature13916

Faini, M., Stengel, F., and Aebersold, R. (2016). The Evolving Contribution of Mass Spectrometry to Integrative Structural Biology. *J. Am. Soc. Mass. Spectrom.* 27, 966–974. doi:10.1007/s13361-016-1382-4

Fill, M., and Copello, J. A. (2002). Ryanodine Receptor Calcium Release Channels. *Physiol. Rev.* 82, 893–922. doi:10.1152/physrev.00013.2002

Georges, A. D., Clarke, O. B., Zalk, R., Yuan, Q., Condon, K. J., Grassucci, R. A., et al. (2016). Structural Basis for Gating and Activation of RyR1. *Cell* 167, 145–157. e17. doi:10.1016/j.cell.2016.08.075

Ghanim, G. E., Fountain, A. J., van Roon, A.-M. M., Rangan, R., Das, R., Collins, K., et al. (2021). Structure of Human Telomerase Holoenzyme with Bound Telomeric DNA. *Nature* 593, 449–453. doi:10.1038/s41586-021-03415-4

Godoy, V. G., Jarosz, D. F., Simon, S. M., Abyzov, A., Ilyin, V., and Walker, G. C. (2007). UmuD and RecA Directly Modulate the Mutagenic Potential of the Y Family DNA Polymerase DinB. *Mol. Cel* 28, 1058–1070. doi:10.1016/j.molcel.2007.10.025

Heinz, L. P., Kopec, W., de Groot, B. L., and Fink, R. H. A. (2018). In Silico assessment of the Conduction Mechanism of the Ryanodine Receptor 1 Reveals Previously Unknown Exit Pathways. *Sci. Rep.* 8, 6886. doi:10.1038/s41598-018-25061-z

Henrikus, S. S., Henry, C., McGrath, A. E., Jergic, S., McDonald, J. P., Hellmich, Y., et al. (2020). Single-molecule Live-Cell Imaging Reveals RecB-dependent Function of DNA Polymerase IV in Double Strand Break Repair. *Nucleic Acids Res.* 48, 8490–8508. doi:10.1093/nar/gkaa597

Hoffmann, A., and Grudinin, S. (2017). NOLB: Nonlinear Rigid Block normal-mode Analysis Method. *J. Chem. Theor. Comput* 13, 2123–2134. doi:10.1021/acs.jctc.7b00197

Hristova, K., and Wimley, W. C. (2011). A Look at Arginine in Membranes. *J. Membr. Biol* 239, 49–56. doi:10.1007/s00232-010-9323-9

Hu, X., Dong, Q., Yang, J., and Zhang, Y. (2016). Recognizing Metal and Acid Radical Ion-Binding Sites by Integrating Ab Initio Modeling with Template-Based Transferals. *Bioinformatics* 32, 3260–3269. doi:10.1093/bioinformatics/btw396

Huang, P.-S., Ban, Y.-E. A., Richter, F., Andre, I., Vernon, R., Schief, W. R., et al. (2011). RosettaRemodel: a Generalized Framework for Flexible Backbone Protein Design. *PLoS One* 6, e24109. doi:10.1371/journal.pone.0024109

Inbar, Y., Benyamini, H., Nussinov, R., and Wolfson, H. J. (2005). Combinatorial Docking Approach for Structure Prediction of Large Proteins and Multi-Molecular Assemblies. *Phys. Biol.* 2, S156–S165. doi:10.1088/1478-3975/2/4/S10

Jones, D. T. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *J. Mol. Biol.* 292, 195–202. doi:10.1006/jmbi.1999.3091

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

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-ray Analysis. *Nature* 181, 662–666. doi:10.1038/181662a0

Krieger, E., Koraimann, G., and Vriend, G. (2002). Increasing the Precision of Comparative Models with YASARA NOVA–a Self-Parameterizing Force Field. *Proteins* 47, 393–402. doi:10.1002/prot.10104

Lamiable, A., Thévenet, P., Rey, J., Vavrusa, M., Derreumaux, P., and Tufféry, P. (2016). PEP-FOLD3: Faster De Novo Structure Prediction for Linear Peptides in Solution and in Complex. *Nucleic Acids Res.* 44, W449–W454. doi:10.1093/nar/gkw329

Lanner, J. T., Georgiou, D. K., Joshi, A. D., and Hamilton, S. L. (2010). Ryanodine Receptors: Structure, Expression, Molecular Details, and Function in Calcium Release. *Cold Spring Harb Perspect. Biol.* 2, a003996. doi:10.1101/cshperspect.a003996

Lanrezac, A., Férey, N., and Baaden, M. (2021). Wielding the Power of Interactive Simulations. *WIREs Comput. Mol. Sci.*, e1594. doi:10.1002/wcms.1594

Lee, S., Tran, A., Allsopp, M., Lim, J. B., Hénin, J., and Klauda, J. B. (2014). CHARMM36 United Atom Chain Model for Lipids and Surfactants. *J. Phys. Chem. B* 118, 547–556. doi:10.1021/jp410344g

Lu, D., Danilowicz, C., Tashjian, T. F., Prévost, C., Godoy, V. G., and Prentiss, M. (2019). Slow Extension of the Invading DNA Strand in a D-Loop Formed by RecA-Mediated Homologous Recombination May Enhance Recognition of DNA Homology. *J. Biol. Chem.* 294, 8606–8616. doi:10.1074/jbc.RA119.007554

Mackerell, A. D., Jr, Feig, M., and Brooks, C. L., 3rd (2004). Extending the Treatment of Backbone Energetics in Protein Force fields: Limitations of Gas-phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* 25, 1400–1415. doi:10.1002/jcc.20065

Mirabello, C., and Wallner, B. (2017). InterPred: A Pipeline to Identify and Model Protein–Protein Interactions. *Proteins* 85, 1159–1170. doi:10.1002/prot.25280

Moal, I. H., and Bates, P. A. (2010). SwarmDock and the Use of normal Modes in Protein-Protein Docking. *Int. J. Mol. Sci.* 11, 3623–3648. doi:10.3390/ijms11103623

Molza, A.-E., Férey, N., Czjzek, M., Le Rumeur, E., Hubert, J.-F., Tek, A., et al. (2014). Innovative Interactive Flexible Docking Method for Multi-Scale Reconstruction Elucidates Dystrophin Molecular Assembly. *Faraday Discuss.* 169, 45–62. doi:10.1039/c3fd00134b

Murata, K., and Wolf, M. (2018). Cryo-electron Microscopy for Structural Analysis of Dynamic Biological Macromolecules. *BBA-Gen Subjects* 1862, 324–334. doi:10.1016/j.bbagen.2017.07.020

Nakai, J., Imagawa, T., Hakamat, Y., Shigekawa, M., Takeshima, H., and Numa, S. (1990). Primary Structure and Functional Expression from cDNA of the Cardiac Ryanodine Receptor/calcium Release Channel. *FEBS Lett.* 271, 169–177. doi:10.1016/0014-5793(90)80399-4

Ngo, V. A., Perissinotti, L. L., Miranda, W., Chen, S. R. W., and Noskov, S. Y. (2017). Mapping Ryanodine Binding Sites in the Pore Cavity of Ryanodine Receptors. *Biophys. J.* 112, 1645–1653. doi:10.1016/j.bpj.2017.03.014

Nogales, E. (2016). The Development of Cryo-EM into a Mainstream Structural Biology Technique. *Nat. Methods* 13, 24–27. doi:10.1038/nmeth.3694

Peng, W., Shen, H., Wu, J., Guo, W., Pan, X., Wang, R., et al. (2016). Structural Basis for the Gating Mechanism of the Type 2 Ryanodine Receptor RyR2. *Science* 354, aah5324. doi:10.1126/science.aah5324

Periole, X., Cavalli, M., Marrink, S.-J., and Ceruso, M. A. (2009). Combining an Elastic Network with a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J. Chem. Theor. Comput* 5, 2531–2543. doi:10.1021/ct9002114

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* 26, 1781–1802. doi:10.1002/jcc.20289

Prentiss, M., Prévost, C., and Danilowicz, C. (2015). Structure/function Relationships in RecA Protein-Mediated Homology Recognition and Strand Exchange. *Crit. Rev. Biochem. Mol. Biol.* 50, 453–476. doi:10.3109/10409238.2015.1092943

Prévost, C., and Sacquin-Mora, S. (2021). Moving Pictures: Reassessing Docking Experiments with a Dynamic View of Protein Interfaces. *Proteins* 89, 1315–1323. doi:10.1002/prot.26152

Robert, C., and Molza, A.-E. (2021). Low-frequency normal Modes Contribution to the Activated to Intermediate State Transition in the Ryanodine Receptor RyR1. Available at: https://figshare.com/s/c0ec203c4662f1a39c1f.

Rueda, M., Chacón, P., and Orozco, M. (2007). Thorough Validation of Protein normal Mode Analysis: a Comparative Study with Essential Dynamics. *Structure* 15, 565–575. doi:10.1016/j.str.2007.03.013

Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., et al. (2012). Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLOS Biol.* 10, 1–5. doi:10.1371/journal.pbio.1001244

Santulli, G., Lewis, D., des Georges, A., Marks, A. R., and Frank, J. (2018). “Ryanodine Receptor Structure and Function in Health and Disease,” in *Membrane Protein Complexes: Structure and Function*. Editors J. R. Harris and E. J. Boekema (Singapore: Springer Singapore), 87, 329–352. Series Title: Subcellular Biochemistry. doi:10.1007/978-981-10-7757-9_11

Schneidman-Duhovny, D., and Wolfson, H. J. (2020). Modeling of Multimolecular Complexes. *Methods Mol. Biol.* 2112, 163–174. doi:10.1007/978-1-0716-0270-6_12

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi:10.1038/msb.2011.75

Spurgeon, S. R., Ophus, C., Jones, L., Petford-Long, A., Kalinin, S. V., Olszta, M. J., et al. (2021). Towards Data-Driven Next-Generation Transmission Electron Microscopy. *Nat. Mater.* 20, 274–279. doi:10.1038/s41563-020-00833-z

Takeshima, H., Nishimura, S., Matsumoto, T., Ishida, H., Kangawa, K., Minamino, N., et al. (1989). Primary Structure and Expression from Complementary DNA of Skeletal Muscle Ryanodine Receptor. *Nature* 339, 439–445. doi:10.1038/339439a0

Tama, F., and Sanejouand, Y. (2001). Conformational Change of Proteins Arising from normal Mode Calculations. *Protein Eng.* 14, 1–6. doi:10.1093/protein/14.1.1

Tashjian, T. F., Danilowicz, C., Molza, A.-E., Nguyen, B. H., Prévost, C., Prentiss, M., et al. (2019). Residues in the Fingers Domain of the Translesion DNA Polymerase DinB Enable its Unique Participation in Error-Prone Double-Strand Break Repair. *J. Biol. Chem.* 294, 7588–7600. doi:10.1074/jbc.RA118.006233

Thomas, A., Field, M. J., and Perahia, D. (1996). Analysis of the Low-Frequency normal Modes of the R State of Aspartate Transcarbamylase and a Comparison with the T State Modes. *J. Mol. Biol.* 261, 490–506. doi:10.1006/jmbi.1996.0478

Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible Fitting of Atomic Structures into Electron Microscopy Maps Using Molecular Dynamics. *Structure* 16, 673–683. doi:10.1016/j.str.2008.03.005

Tsirigos, K. D., Peters, C., Shu, N., Käll, L., and Elofsson, A. (2015). The TOPCONS Web Server for Consensus Prediction of Membrane Protein Topology and Signal Peptides. *Nucleic Acids Res.* 43, W401–W407. doi:10.1093/nar/gkv485

Ulmschneider, M. B., Ulmschneider, J. P., Freites, J. A., von Heijne, G., Tobias, D. J., and White, S. H. (2017). Transmembrane Helices Containing a Charged Arginine Are Thermodynamically Stable. *Eur. Biophys. J.* 46, 627–637. doi:10.1007/s00249-017-1206-x

Urzhumtsev, A., Afonine, P. V., Van Benschoten, A. H., Fraser, J. S., and Adams, P. D. (2016). From Deep TLS Validation to Ensembles of Atomic Models Built from Elemental Motions. Addenda and Corrigendum. *Acta Crystallogr. D* 72, 1073–1075. doi:10.1107/S2059798316013048

Van Petegem, F. (2017). Ligand Binding to Ryanodine Receptors Revealed through Cryo-Electron Microscopy. *Cell Calcium* 61, 50–52. doi:10.1016/j.ceca.2016.10.004

Van Petegem, F. (2015). Ryanodine Receptors: Allosteric Ion Channel Giants. *J. Mol. Biol.* 427, 31–53. doi:10.1016/j.jmb.2014.08.004

van Zundert, G., Rodrigues, J., Trellet, M., Schmitz, C., Kastritis, P., Karaca, E., et al. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* 428, 720–725. doi:10.1016/j.jmb.2015.09.014

Verkhivker, G. M., Agajanian, S., Oztas, D. Y., and Gupta, G. (2021). Comparative Perturbation-Based Modeling of the SARS-CoV-2 Spike Protein Binding with Host Receptor and Neutralizing Antibodies: Structurally Adaptable Allosteric Communication Hotspots Define Spike Sites Targeted by Global Circulating Mutations. *Biochemistry* 60 (19), 1459–1484. doi:10.1021/acs.biochem.1c00139

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The DISOPRED Server for the Prediction of Protein Disorder. *Bioinformatics* 20, 2138–2139. doi:10.1093/bioinformatics/bth195

Watson, J. D., and Crick, F. H. (1953). Genetical Implications of the Structure of Deoxyribonucleic Acid. *Nature* 171, 964–967. doi:10.1038/171964b0

Webb, B., and Sali, A. (2021). Protein Structure Modeling with Modeller. *Methods Mol. Biol.* 2199, 239–255. doi:10.1007/978-1-0716-0892-0_14

Wheeler, T. J., Clements, J., and Finn, R. D. (2014). Skylign: a Tool for Creating Informative, Interactive Logos Representing Sequence Alignments and Profile Hidden Markov Models. *BMC Bioinformatics* 15, 7. doi:10.1186/1471-2105-15-7

Willegems, K., and Efremov, R. G. (2018). Influence of Lipid Mimetics on Gating of Ryanodine Receptor. *Structure* 26, 1303–1313. e4. doi:10.1016/j.str.2018.06.010

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., et al. (2018). MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Sci.* 27, 293–315. doi:10.1002/pro.3330

Yan, R., Xu, D., Yang, J., Walker, S., and Zhang, Y. (2013). A Comparative Assessment and Analysis of 20 Representative Sequence Alignment Methods for Protein Structure Prediction. *Sci. Rep.* 3, 2619. doi:10.1038/srep02619

Yan, Z., Bai, X.-c., Yan, C., Wu, J., Li, Z., Xie, T., et al. (2015). Structure of the Rabbit Ryanodine Receptor RyR1 at Near-Atomic Resolution. *Nature* 517, 50–55. doi:10.1038/nature14063

Yang, D., Boyer, B., Prévost, C., Danilowicz, C., and Prentiss, M. (2015). Integrating Multi-Scale Data on Homologous Recombination into a New Recognition Mechanism Based on Simulations of the RecA-ssDNA/dsDNA Structure. *Nucleic Acids Res.* 43, 10251–10263. doi:10.1093/nar/gkv883

Yang, H., Zhou, C., Dhar, A., and Pavletich, N. P. (2020). Mechanism of Strand Exchange from RecA-DNA Synaptic and D-Loop Structures. *Nature* 586, 801–806. doi:10.1038/s41586-020-2820-9

Zalk, R., Clarke, O. B., des Georges, A., Grassucci, R. A., Reiken, S., Mancia, F., et al. (2015). Structure of a Mammalian Ryanodine Receptor. *Nature* 517, 44–49. doi:10.1038/nature13950

Ziegler, S. J., Mallinson, S. J., St. John, P. C., and Bomble, Y. J. (2021). Advances in Integrative Structural Biology: Towards Understanding Protein Complexes in Their Cellular Context. *Comput. Struct. Biotech. J.* 19, 214–225. doi:10.1016/j.csbj.2020.11.052

Conflict of Interest: MM was employed by Servier Monde.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership