EXPERIMENTS AND SIMULATIONS: A PAS DE DEUX TO UNRAVEL BIOLOGICAL FUNCTION

EDITED BY: Massimiliano Bonomi, Edina Rosta, Maya Topf and Gregory Bowman

PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-089-5
DOI 10.3389/978-2-88974-089-5

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

EXPERIMENTS AND SIMULATIONS: A PAS DE DEUX TO UNRAVEL BIOLOGICAL FUNCTION

Topic Editors:
Massimiliano Bonomi, Institut Pasteur, France
Edina Rosta, King's College London, United Kingdom
Maya Topf, Leibniz Institute of Experimental Biology and UKE, Germany
Gregory Bowman, Washington University School of Medicine in St. Louis, United States

Citation: Bonomi, M., Rosta, E., Topf, M., Bowman, G., eds. (2022). Experiments and Simulations: A Pas de Deux to Unravel Biological Function. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-089-5

Table of Contents

frontiers in Molecular Biosciences

EDITORIAL
published: 29 November 2021
doi: 10.3389/fmolb.2021.799406

Editorial: Experiments and Simulations: A *Pas de Deux* to Unravel Biological Function

Maya Topf[1], *Edina Rosta*[2,3], *Gregory R. Bowman*[4] *and Massimiliano Bonomi*[5,6]*

[1]Center for Structural Systems Biology (CSSB), Leibniz-Institut für Experimentelle Virologie (HPI) and Universitätsklinikum Hamburg-Eppendorf (UKE), Hamburg, Germany, [2]Department of Chemistry, King's College London, London, United Kingdom, [3]Department of Physics and Astronomy, University College London, London, United Kingdom, [4]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St Louis, MO, United States, [5]Structural Bioinformatics Unit, Department of Structural Biology and Chemistry, CNRS UMR 3528, Institut Pasteur, Paris, France, [6]USR3756 Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Paris, France

Keywords: modeling, molecular dynamic (MD), integrative approaches, functional dynamics, experimental-computational method, molecular simulation

Editorial on the Research Topic

Experiments and Simulations: A Pas de Deux to Unravel Biological Function

Understanding the molecular mechanisms used by biological systems to perform their functions is often essential to rationally target associated diseases. In many cases, the determination of the three-dimensional structure of these systems provides precious insights. However, it is more often the interplay between structural and dynamical properties that determines the behavior of complex systems (Henzler-Wildman and Kern, 2007; Orozco, 2014). While both experimental and computational methods are invaluable tools to study protein structure and dynamics, limitations in each individual technique can hamper their predictive capabilities (Schneidman-Duhovny et al., 2014). On one hand, determining structural models solely from experimental data is challenging as these data often come from time and ensemble-averaged measurements over conformationally heterogeneous states, provide sparse and sometimes ambiguous information, and are always subject to random and systematic errors. On the other hand, structural models determined by computational approaches such as protein structure prediction methods and/or molecular dynamics (MD) are limited by the inaccuracies of the force fields used as well as by the challenge of exhaustively sampling the conformational landscape of complex systems (Bonomi et al., 2017). Combining experiments and simulations is therefore a successful strategy to overcome the limitations of the individual approaches and to accurately characterize the behavior of biological systems (Bottaro and Lindorff-Larsen, 2018; Rout and Sali, 2019).

The goal of this Research Topic is to present some representative examples of synergistic use of experimental and computational techniques aimed at accurately characterizing the structure, dynamics, and ultimately function of biological systems. This Research Topic of 14 articles explores different areas of experimental-computational integration: the use of computational approaches to assist the interpretation of existing experimental data or to predict the outcome of new measurements, the experimental validation of computational predictions, and the incorporation of experimental data to drive and/or refine molecular simulations. A wide spatial spectrum of systems will be covered, encompassing ordered and disordered peptides and proteins, small-molecules interacting with proteins, protein complexes, nucleic acids, and entire cells. The integration of molecular simulations with different types of experimental data will be illustrated, including cryo-electron microscopy (cryo-EM) and tomography, super-resolution microscopy, Nuclear Magnetic Resonance (NMR) spectroscopy, biochemical, and Small Angle X-ray Scattering (SAXS) data.

OPEN ACCESS

Edited and reviewed by:
Francesco Luigi Gervasio, University College London, United Kingdom

***Correspondence:**
Massimiliano Bonomi
massimiliano.bonomi@pasteur.fr

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

Received: *21 October 2021*
Accepted: *01 November 2021*
Published: *29 November 2021*

Citation:
Topf M, Rosta E, Bowman GR and Bonomi M (2021) Editorial: Experiments and Simulations: A Pas de Deux to Unravel Biological Function. Front. Mol. Biosci. 8:799406. doi: 10.3389/fmolb.2021.799406

One of the areas of research in which computationally approaches, and particularly MD simulations, have traditionally been used to complement experimental measurements is the prediction and/or rationalization of the effect of mutations on the structure and dynamics of biological systems. Crnjar et al. use MD simulations to shed light into the effect of a single-point mutation in the outer lipid-facing helix (M4) of the 5-HT3A pentameric ligand-gated ion channels. The mutation of a tyrosine (Y441) in this area has been experimentally shown to inhibit the function of the receptor. The MD simulations reported in this paper reveal a network of interactions that connects Y441 to the ion channel hydrophobic gate on helix M2, thus rationalizing the effect of the mutation that has experimentally been observed. Ochoa et al. build a set of scoring matrices using structural observables extracted from MD simulations to predict the effect of single-point mutations on peptide binders to the Major Histocompatibility Complex (MHC) class II receptors. The method developed integrates sequence, structural and dynamical information and can be used to guide the design of novel peptide binders to the MHC class II receptors.

Another area in which molecular simulations can complement experiments is the characterization of conformations that are often difficult to observe directly, such as short-lived conformational states and disordered motifs. Nierzwicki and Palermo illustrate the case of the CRISPR-Cas9 genome editing machinery, of which the catalytically active structure has been predicted by MD simulations and subsequently validated by high-resolution cryo-EM data. This paper also provides an overview of computational approaches that can be used to refine both single-structure models and conformational ensembles given a cryo-EM map. Lambrughi et al. use enhanced-sampling MD simulations to study the Ubiquitin Interacting Motif (UIM), a conserved, highly-dynamic segment used by several multi-domain proteins to interact with ubiquitin. While existing X-ray data could not capture the structural heterogeneity of UIM, the MD ensembles revealed an equilibrium between ordered and disordered states, in agreement with NMR chemical shifts data.

Computational techniques have become over the years an invaluable tool in the drug discovery field (Brogi et al., 2020). In this Research Topic, De Felice et al. present an *in silico* approach named "Computational Profiling for GPCRs" that repurposes a GPCR-binding ligand for a different GPCR. The method is tested on 3 different GPCR receptors and validated using docking calculations and pharmacological data. Sgrignani and Cavalli use computational docking and molecular simulations to investigate the mode of binding of bromhexine to the transmembrane serine protease TMPRSS2, an enzyme involved in the activation of several coronaviruses, including SARS-CoV-2. Their analysis reveals the existence of an allosteric pocket involved in the binding of bromhexine to TMPRSS2 and in its inhibition. Finally,Prerovská et al. perform MD simulations to predict the structure of the complex formed by β-D-galactosyl Yariv reagent and oligo β-D-(1→3)-galactan and ultimately to shed light into the structural basis of arabinogalactan protein precipitation by Yariv.

Over the last decade, a novel class of methods that incorporate experimental information into molecular simulations has flourished. These so-called integrative or hybrid modeling approaches use experimental data to either guide or refine *a posteriori* (Rangan et al., 2018) molecular simulations in order to determine individual structural models or conformational ensembles consistent with the available information. These techniques are often based on 1) Bayesian frameworks to properly balance the information provided by different types of experiments with prior, physico-chemical knowledge of the system; and 2) the Maximum Entropy Principle to resolve the ambiguity of determining conformational distributions based on the knowledge of ensemble-averaged experimental observations (Ravera et al., 2016; Bonomi et al., 2017; Bottaro and Lindorff-Larsen, 2018; Cesari et al., 2018). In this Research Topics, six examples of this type of hybrid computational-experimental approaches are illustrated.

The papers by Ahmed et al., Spill et al., and Paissoni et al. present different ways to characterize conformational ensembles of dynamic systems by integrating MD simulations with SAXS data. Ahmed et al. use the refinement Bayesian/Maximum Entropy (BME) technique to determine structural ensembles of α-synuclein starting from MD simulations performed with different force fields. Spill et al. propose a Bayesian weighting approach for SAXS data coupled with a selection of an ensemble of minimal size to characterize the conformational heterogeneity of a tandem of domains from the protein PTPN4. Paissoni et al. perform an exhaustive analysis of the statistical precision of the metainference technique coupled with metadynamics using as test system the chignolin peptide. Voelz et al. present a review of the features, advantages over other integrative approaches, and shortcomings of their Bayesian Inference of Conformational Populations (BICePs) method. Gaalswyk et al. illustrate how their Modeling Employing Limited Data (MELD) approach can be used to determine protein structural ensembles using sparse NMR data combined with physical modeling. Finally, Vakili and Habeck introduce a Bayesian technique to address the problem of reconstructing, in tomography, a 3D structure from 2D views along unknown random directions.

The majority of applications of combined computational-experimental techniques presented in this Research Topic involve biological systems ranging in size, from small molecules to protein complexes. However, integrative approaches can in principle be applied to larger spatial scales, provided that appropriate experimental data are available. One illustrative example presented here is the work of Bianchi et al. In this paper, the authors combine single molecule super-resolution microscopy in cells with computational modeling to study how a small RNA (SgrS) regulates glucose-phosphate stress, or "sugar shock" in *E. coli*. Their stochastic simulations guided by the in cell experimental data enable the description of the cellular heterogeneity observed in the *E. coli* sugar shock response network.

To summarize, this Research Topic demonstrates how a synergistic use of experiments and simulations can be a powerful strategy to study structure, dynamics, and function of biological systems across a variety of spatial and temporal

scales. These successes have been made possible by the continuous improvements in both experimental and computational techniques, the development of open-source software and web-servers for integrative studies (Rout and Sali, 2019), the dissemination of protocols in open-source notebooks and public repositories (The PLUMED Consortium, 2019), and the availability of raw data and structural models on public databases, such as the PDB-Dev[1](Burley et al., 2017). There is a bright future ahead for integrative studies, especially these days, when a new class of methods based on artificial intelligence, such as DeepMind's AlphaFold (Jumper et al., 2021) and RoseTTA fold (Baek et al., 2021), has joined the game and already taken it by storm, reaching accuracy in structure prediction comparable to experimental techniques.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science.* 373, 871–876. doi:10.1126/science.abj8754

Bonomi, M., Heller, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of Protein Structural Ensemble Determination. *Curr. Opin. Struct. Biol.* 42, 106–116. doi:10.1016/j.sbi.2016.12.004

Bottaro, S., and Lindorff-Larsen, K. (2018). Biophysical Experiments and Biomolecular Simulations: A Perfect Match? *Science.* 361, 355–360. doi:10.1126/science.aat4010

Brogi, S., Ramalho, T. C., Kuca, K., Medina-Franco, J. L., and Valko, M. (2020). Editorial: In Silico Methods for Drug Design and Discovery. *Front. Chem.* 8, 612. doi:10.3389/fchem.2020.00612

Burley, S. K., Kurisu, G., Markley, J. L., Nakamura, H., Velankar, S., Berman, H. M., et al. (2017). PDB-dev: a Prototype System for Depositing Integrative/Hybrid Structural Models. *Structure.* 25, 1317–1318. doi:10.1016/j.str.2017.08.001

Cesari, A., Reisser, S., and Bussi, G. (2018). Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments. *Computation.* 6, 15. doi:10.3390/computation6010015

Henzler-Wildman, K., and Kern, D. (2007). Dynamic Personalities of Proteins. *Nature.* 450, 964–972. doi:10.1038/nature06522

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

Orozco, M. (2014). A Theoretical View of Protein Dynamics. *Chem. Soc. Rev.* 43, 5051–5066. doi:10.1039/c3cs60474h

Rangan, R., Bonomi, M., Heller, G. T., Cesari, A., Bussi, G., and Vendruscolo, M. (2018). Determination of Structural Ensembles of Proteins: Restraining vs Reweighting. *J. Chem. Theor. Comput.* 14, 6632–6641. doi:10.1021/acs.jctc.8b00738

Ravera, E., Sgheri, L., Parigi, G., and Luchinat, C. (2016). A Critical Assessment of Methods to Recover Information From Averaged Data. *Phys. Chem. Chem. Phys.* 18, 5686–5701. doi:10.1039/c5cp04077a

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

Schneidman-Duhovny, D., Pellarin, R., and Sali, A. (2014). Uncertainty in Integrative Structural Modeling. *Curr. Opin. Struct. Biol.* 28, 96–104. doi:10.1016/j.sbi.2014.08.001

The PLUMED Consortium (2019). Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods.* 16, 670–673. doi:10.1038/s41592-019-0506-8

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, though no other collaboration, with one of the authors ER at the time of the 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.

[1]https://pdb-dev.wwpdb.org

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 23 December 2020
doi: 10.3389/fmolb.2020.593826

Stochastic Analysis Demonstrates the Dual Role of Hfq in Chaperoning *E. coli* Sugar Shock Response

David M. Bianchi[1,2], *Troy A. Brier*[1,2], *Anustup Poddar*[2,3,4], *Muhammad S. Azam*[5†], *Carin K. Vanderpool*[5], *Taekjip Ha*[2,3,4] *and Zaida Luthey-Schulten*[1,2*]

[1] Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [2] Center for the Physics of Living Cells, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [3] Department of Biophysics, Johns Hopkins University, Baltimore, MD, United States, [4] HHMI Investigator Program, Howard Hughes Medical Institute, Chevy Chase, MD, United States, [5] Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL, United States

OPEN ACCESS

Edited by:
Edina Rosta,
King's College London,
United Kingdom

Reviewed by:
Carlo Camilloni,
University of Milan, Italy
Ilpo Vattulainen,
University of Helsinki, Finland

***Correspondence:**
Zaida Luthey-Schulten
zan@illinois.edu

†Present address:
Muhammad S. Azam,
Department of Biochemistry and Molecular Biophysics, University of Chicago, Chicago, IL, United States

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

Received: 11 August 2020
Accepted: 23 November 2020
Published: 23 December 2020

Citation:
Bianchi DM, Brier TA, Poddar A, Azam MS, Vanderpool CK, Ha T and Luthey-Schulten Z (2020) Stochastic Analysis Demonstrates the Dual Role of Hfq in Chaperoning E. coli Sugar Shock Response. Front. Mol. Biosci. 7:593826. doi: 10.3389/fmolb.2020.593826

Small RNAs (sRNAs) play a crucial role in the regulation of bacterial gene expression by silencing the translation of target mRNAs. SgrS is an sRNA that relieves glucose-phosphate stress, or "sugar shock" in *E. coli*. The power of single cell measurements is their ability to obtain population level statistics that illustrate cell-to-cell variation. Here, we utilize single molecule super-resolution microscopy in single *E. coli* cells coupled with stochastic modeling to analyze glucose-phosphate stress regulation by SgrS. We present a kinetic model that captures the combined effects of transcriptional regulation, gene replication and chaperone mediated RNA silencing in the SgrS regulatory network. This more complete kinetic description, simulated stochastically, recapitulates experimentally observed cellular heterogeneity and characterizes the binding of SgrS to the chaperone protein Hfq as a slow process that not only stabilizes SgrS but also may be critical in restructuring the sRNA to facilitate association with its target *ptsG* mRNA.

Keywords: stochastic biology, cell simulations, small RNA, single-molecule techniques, super-resolution microscopy, gene regulatory networks, cellular stress response

1. INTRODUCTION

The ability of living cells to modulate their gene expression in response to changing environmental conditions is critical to their growth and continued development. Many bacteria use the phosphoenolpyruvate phosphotransferase (PTS) system to transport and phosphorylate incoming sugars to prepare them for subsequent glycolytic metabolism. The uptake of phosphosugars must be balanced with their breakdown in order to prevent metabolic stress. In *E. coli*, a stress response induced by unbalanced glucose-phosphate transport and metabolism or "sugar shock," is referred to as glucose-phosphate stress response. A primary activity of this stress response is RNA silencing of *ptsG*, a gene coding for the glucose transport protein of the same name (also known as EIICBGlc in *E. coli*), by the small RNA (sRNA) SgrS. Small RNAs are usually non-coding RNA molecules that act by base pairing with target messengers to regulate translation or mRNA stability and have been observed across all domains of life (Babski et al., 2014). *sgrS* is upregulated by a transcriptional activator (SgrR) when the cell is under a state of glucose-phosphate stress. SgrS regulates *ptsG* post-transcriptionally by a mechanism where SgrS binds to *ptsG* messenger RNA (mRNA) and prevents

its translation to protein by blocking access of the ribosome to the mRNA (Vanderpool and Gottesman, 2004; Maki et al., 2010). This also enhances the co-degradation of *ptsG* mRNA and SgrS via enzymes responsible for the removal of bulk RNA such as ribonuclease E (RNase E) (Kawamoto et al., 2006; Maki et al., 2010). This co-degradation reduces the number of PtsG sugar transporter proteins that are produced and thus reduces the impact of glucose-phosphate stress, since fewer transport proteins are available to bring sugar into the cell.

SgrS and *ptsG* mRNA associate via complementary base pairing that occludes the ribosome binding site on the mRNA. Recently, this mechanism has been analyzed in conjunction with binding of the Sm-like chaperone protein Hfq to SgrS, which has been proposed to stabilize the sRNA, and facilitate the interaction between the sRNA and its mRNA target (Ishikawa et al., 2012). Hfq also promotes SgrS–dependent regulation of other targets involved in sugar shock such as *manXYZ*, and *yigL* in *E. coli*. In this study, we focus only on the primary regulatory target *ptsG* mRNA and do not consider the other targets of the SgrS regulon, which are described in Bobrovskyy et al. (2019).

Previous experimental and theoretical work (Jones et al., 2014; Peterson et al., 2015) has demonstrated the necessity of accounting for gene replication over the course of the cell cycle in order to capture the population variation observed in messenger RNA abundance. The additional noise emanating from transcription at multiple gene loci manifests itself in the broad mRNA copy number distributions observed in a population of cells. The aforementioned work also demonstrated that including the effect of gene regulation by transcription factors can be critical in order to appropriately describe stochastic dynamics. The effect of transcriptional regulation is apparent in the SgrS–*ptsG* mRNA system, where the expression of SgrS is maintained by the regulator SgrR, which activates *sgrS* and autorepresses its own expression during glucose-phosphate stress conditions (Vanderpool and Gottesman, 2004, 2007).

Recently, Fei et al. (2015) presented a deterministic kinetic model of the SgrS mediated regulation of *ptsG* mRNA in *E. coli*. Using single-molecule fluorescence experiments (smFISH and STORM), SgrS and *ptsG* mRNA copy numbers in cells were measured, which produced distributions of RNA at various time points after the induction of sugar stress across a population of fast-growing *E. coli*. However, it is important to note that both the *ptsG* mRNA and the SgrS regulating it are present in low copy number (a few to tens of particles) and therefore exhibit intrinsically noisy behavior in both their gene expression and regulatory behaviors. For this reason it is most appropriate to treat the regulatory processes via stochastic simulation in order to quantify the variation that is observed across a population of cells, which has been demonstrated previously (Elowitz et al., 2002; Raser, 2005; Earnest et al., 2018).

Here, we have developed a stochastic model, to our knowledge the first of its kind for an RNA silencing network, that captures the mRNA and sRNA distributions experimentally observed in a population of hundreds of *E. coli* cells. The stochastic model additionally incorporates the following features that extend the platform given by Fei et al. (2015): (1) accounting for gene replication, (2) transcriptional gene regulation of *sgrS* by its activator SgrR and (3) explicit representation of the SgrS stabilization via the Hfq chaperone protein. This model robustly describes experimentally observed RNA distributions, closely matching regulatory dynamics from immediately after induction until a steady state is reached 20 min later. We also utilize this model to analyze the effects of the size of the pool of Hfq chaperone protein available to SgrS, to decouple the rate of Hfq stabilization of SgrS and its subsequent activity in enhancing association to the target, *ptsG* mRNA, and to study the effect of an *sgrS* point mutation in the SgrS-Hfq binding region on regulatory dynamics.

2. MATERIALS AND METHODS

2.1. Model and Computational Methods

The previous kinetic model for SgrS regulation of *ptsG* mRNA (Fei et al., 2015) utilized simple mass-action kinetics to describe the target search process and modeled gene expression as a constitutive process, with RNA species originating from a single gene copy. Despite its simplicity, this model captures average regulatory network behavior and also gives insight into many of the parameters required for the more descriptive stochastic model that is the focus of this work. For example, since an overall binding rate for SgrS to *ptsG* mRNA was established in Fei et al. (2015) we are now able to complexify the model by the addition of the chaperone protein Hfq, which allowed us to predict (by fitting to the experimental data) the size of the pool of Hfq available to stabilize SgrS and the rate at which it binds the sRNA (separate from its association to *ptsG* mRNA).

The kinetic model was implemented and solved stochastically as a well-mixed Chemical Master Equation (CME) in the Lattice Microbes (LM) simulation software suite (Peterson et al., 2013; Roberts et al., 2013; Hallock et al., 2014; Hallock and Luthey-Schulten, 2016). The corresponding rate constants (**Table 1**) were adapted from the kinetic model described in **Figure 1**. One important feature added to the model is the explicit presence of the chaperone protein Hfq, which has been shown to both stabilize SgrS (substantially increasing its half-life) and to facilitate the association of SgrS to *ptsG* mRNA (Vanderpool and Gottesman, 2004; Hopkins et al., 2011; Wagner, 2013; Santiago-Frangos and Woodson, 2018). In order to capture the cell-to-cell heterogeneity due to the small number of particles (e.g., gene copies) involved in transcription, it is critical to account for transcriptional regulation of the genes involved in the glucose-phosphate stress response. For this reason, we include the transcriptional activation of *sgrS* by the transcription factor SgrR, which has been shown to upregulate *sgrS* expression in the presence of αMG (the unmetabolizable inducer used in place of glucose for our experiments) (Vanderpool and Gottesman, 2004, 2007). Regulation of *ptsG* by the transcriptional repressor Mlc was not included in the model since repression is relieved in the presence of glucoside sugars. With αMG present, Mlc is sequestered at the membrane by binding the EIIB subunit of the PtsG transporter protein complex (Lee, 2000; Seitz et al., 2003; Nam et al., 2008), relieving repression and resulting in high levels of *ptsG* transcriptional activity (Balasubramanian and Vanderpool, 2013). Since the decay time of PtsG proteins is

TABLE 1 | The list of parameters used for the kinetic model.

Parameter	Value	Unit	Source
$k_{t,p}$	0.12 ± 0.01	s^{-1}	Experimentally measured
β_p	(3.7 ± 0.5) × 10^{-3}	s^{-1}	Experimentally measured
$k_{on,Ds}$	(3.0 ± 0.1) × 10^{-2}	s^{-1}	Fit
$k_{off,Ds}$	(9.5 ± 0.1) × 10^{-3}	s^{-1}	Fit
$k_{t,s}$	0.33 ± 0.01	s^{-1}	Fit
k_{ds}	0.022 ± 0.002	s^{-1}	Δ*hfq* decay rate of SgrS
k_{bind}	0.063[a] ± 0.014	s^{-1}	Fit
k_{unbind}	0.0018 ± 0.0004	s^{-1}	SgrS decay rate
k_{on}	(3.1 ± 0.2) × 10^{-4}	$molec^{-1}s^{-1}$	Fei et al., 2015
k_{off}	0.22 ± 0.02	s^{-1}	Fei et al., 2015
k_{cat}	0.3 ± 0.1	s^{-1}	Fei et al., 2015
% high, low gene state *sgrS*	25 ± 12, 75 ± 12	%	Fit
% high, low gene state *ptsG*	46 ± 20, 54 ± 20	%	Fit
Hfq pool size (available to SgrS Regulon)	250 ± 167	*molec*	Fit

The % in each gene state refers to percentage of the cellular population with the gene being in a low or high gene copy state as described in section 2.1.1. (a) k_{bind} is given as a Pseudo first order rate accounting for the average expected pool size of Hfq participating in SgrS stabilization and enhancement (250). When converted to the corresponding bulk second order rate with 250 Hfq present k_{bind} agrees well with the range of Hfq binding rates measured for other sRNA reviewed in Santiago-Frangos and Woodson (2018) and discussed further in section 3. Confirmation of k_{on} and k_{off} as the same values given in Fei et al. (2015) is discussed in section 2.2. Calculation and analysis of parameter uncertainty values by Markov Chain Monte Carlo analysis is discussed in ***Supplementary Section 6****.*

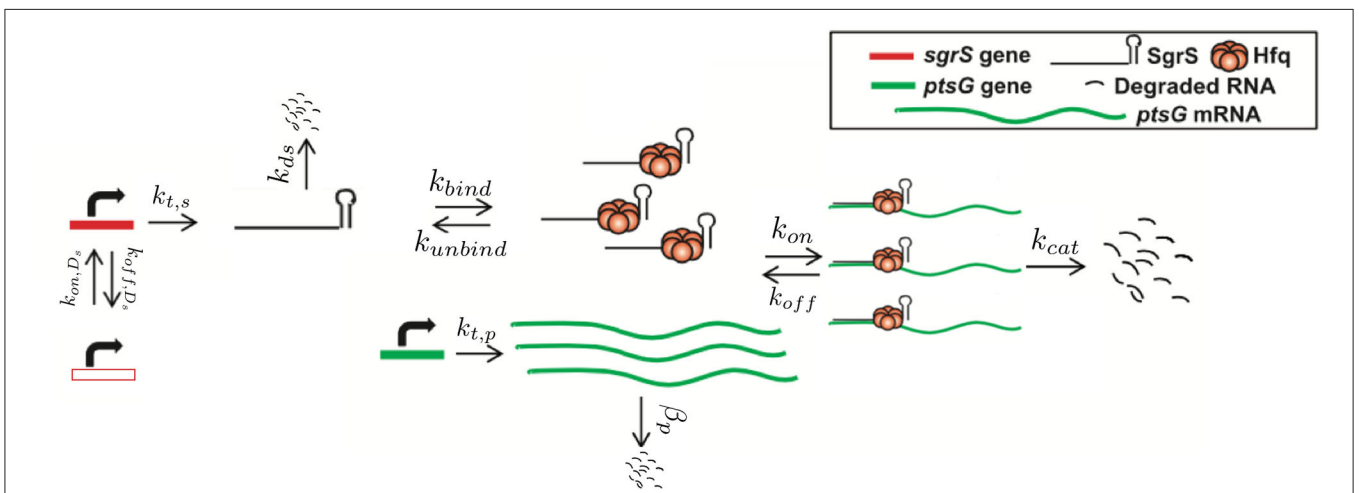


FIGURE 1 | Schematic of the kinetic model as described in the text. The RNA species are transcribed from a sampled genome state with *sgrS* capable of switching between an "ON" and "OFF" state. Explicitly represented Hfq can bind and unbind with SgrS, and then the Hfq–SgrS complex binds (and potentially unbinds) with *ptsG* mRNA. All RNA degradation events are carried out by the enzyme RNase E. See **Figure 4** for the kinetic equations described above.

expected to be approximately on the order of 8 h (Maier et al., 2011), much longer than the timescale of mRNA decay, Mlc repressors are likely still sequestered by the transporters at the membrane 20 min post-induction and have little effect on the SgrS regulatory process. Rates for the association of the Hfq-SgrS complex to *ptsG* mRNA (k_{on}) and the dissociation of the Hfq-SgrS-*ptsG* mRNA complex (k_{off}) were obtained from Fei et al. (2015), which did not include Hfq explicitly but provides the corresponding association and dissociation reaction rates. The value for the co-degradation rate of SgrS and *ptsG* mRNA from the Hfq-SgrS-*ptsG* mRNA complex by RNase E (k_{cat}) is also obtained from Fei et al. (2015) (see section 2.2 for confirmation of k_{on}, k_{off}, and k_{cat} values).

2.1.1. Calculation of Gene Copy Number

Finally, and critically, in order to appropriately capture regulatory effects on gene expression of SgrS and *ptsG* mRNA, it is important to account for gene duplication, as we have previously shown (Peterson et al., 2015). As illustrated by Jones et al. (2014) since the time to replicate the *E. coli* genome (approximately 40 min, Cooper and Helmstetter, 1968) is longer than the fast-growing *E. coli* cell division time of 20 min (or the 35

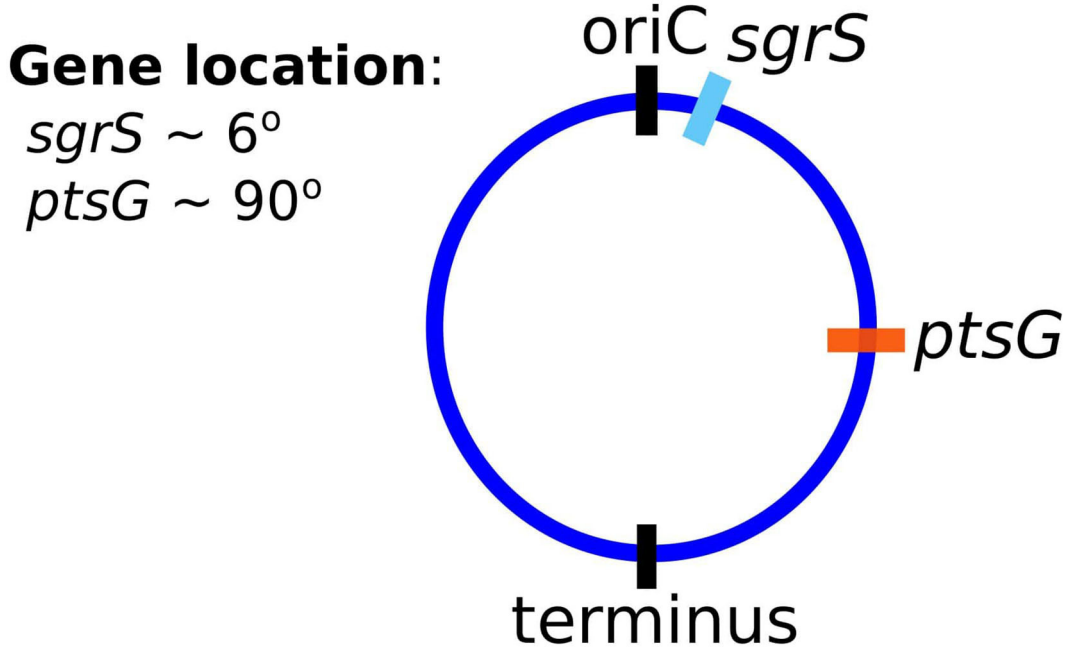


FIGURE 2 | The gene location for SgrS and *ptsG* mRNA relative to the origin of replication (*oriC*) are shown on the circular genome of the *E. coli* cells used for this study. As it is closer to the origin of replication *sgrS* (cyan) is likely to be present in higher gene copy number than *ptsG* (orange), which is farther away from the *oriC*.

min observed in our experiments), the cell has nested replication forks that are already replicating the genomes of daughter and granddaughter cells prior to cell division. In particular, genes close to the origin of replication are likely to have multiple copies present over much of the cell cycle. This phenomenon has been shown previously for genes near the origin in *E. coli* by both isotopic labeling of nucleotides and imaging of fluorescent chromosome markers (Cooper and Helmstetter, 1968; Youngren et al., 2014). Due to the position of *sgrS* (only 6° away along the circular chromosome) very near to the origin of replication, it is likely that multiple gene copies are accessible for transcription over the course of the cell cycle. About half-way between the origin and terminus of replication (at approximately 90°) *ptsG* is also likely to have multiple gene copies present at some point over the course of the cell cycle, although at lower copy number than *sgrS*. **Figure 2** depicts the two genes and their location along the circular *E. coli* genome.

The experimentally measured cells were unsynchronized and should have multiple replication forks present over the course of the 20 min post-induction, our measurement window. To account for gene duplication effects in a population of unsynchronized cells, we sample the percentage of the cellular population in either a low or high gene state, which corresponds to the expected distribution of the number of genes present over the course of the cell cycle after induction. In this way, we effectively flip a coin to decide whether a simulation replicate corresponding to an individual experimentally imaged *E. coli* cell has 2 copies (low gene state) or 4 copies (high gene state) of *sgrS* and similarly 1 or 2 copies of *ptsG*. This allows us to account for the effect of gene duplication in generating mRNA noise over the heterogeneous population of hundreds of *E. coli* cells that were observed experimentally. We assume that all gene copies are transcribed independently from one another and at the same rate, a notion that Wang et al. (2019) has recently examined in *E. coli* under various growth conditions. Under similar growth conditions to ours [MOPS glucose-based medium with a doubling time of 35 min (see section 2.2)], the data from Wang et al. (2019) suggest that transcription does appear to be independent and uncorrelated between copies of the same gene.

Figure 3 illustrates the reasoning for the specific choices of high and low state gene copy numbers for *ptsG* and *sgrS* in an *E. coli* cell growing faster than the expected time necessary for replication (approximately 40 min, compared to an experimentally observed generation time of approximately 35 min) (Cooper and Helmstetter, 1968; Youngren et al., 2014).

Stochastic simulations were performed by sampling the CME for the model given in **Figure 1** with the widely used Gillespie Direct Method of the Stochastic Simulation Algorithm (SSA), which is implemented in the publicly available Lattice Microbes (LM) software suite (version 2.3 was used) and its python interface pyLM (Peterson et al., 2013; Roberts et al., 2013; Hallock et al., 2014; Hallock and Luthey-Schulten, 2016). We ran 2,000 replicate simulations for 25 min after αMG induction of glucose-phosphate stress in order to match the corresponding 20 min smFISH-STORM experiments. Initial conditions for basal SgrS (1–3 copies) and *ptsG* mRNA (30–40 copies) copy number were sampled from the experimentally measured distributions and rounded to the nearest integer particle number (a necessity for stochastic representation). Simulations

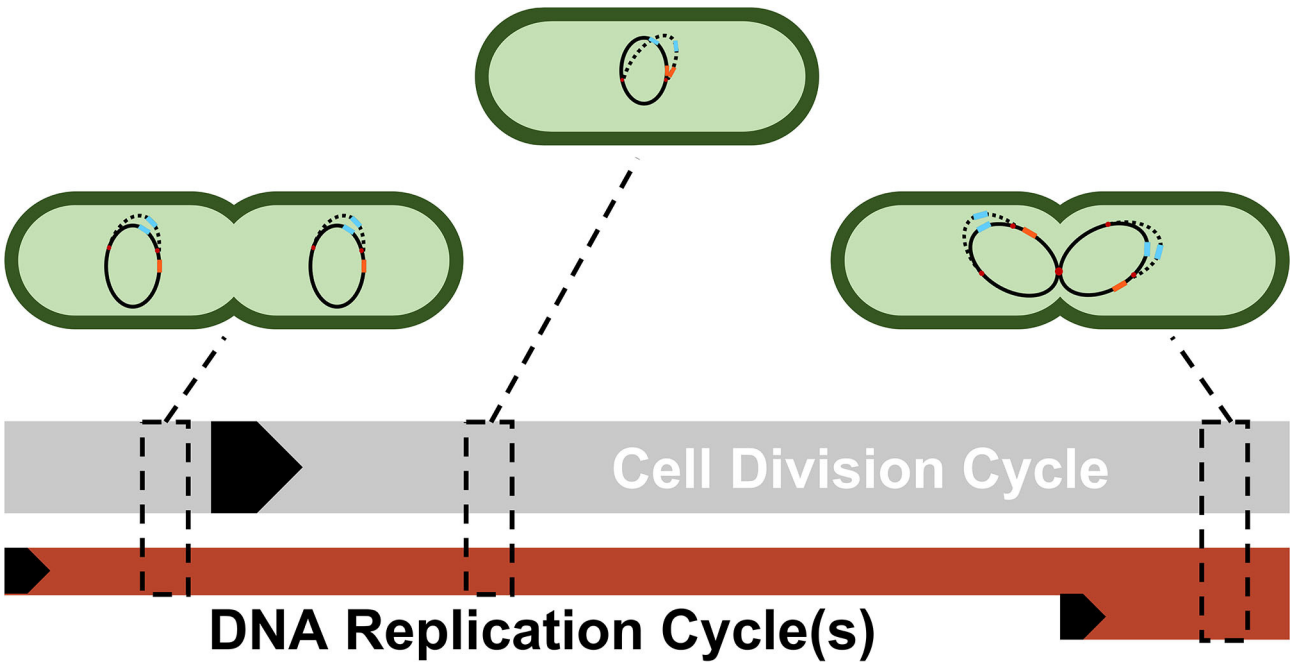


FIGURE 3 | A simplified depiction of possible cellular states throughout a single DNA replication cycle. Each cell shows a snapshot of the gene state of a cell given its progression through the DNA replication and cell division cycle. Due to the difference in lengths of the cell division cycle (∼35 min) and DNA replication cycle (∼40 min), DNA replication and cell division are not completely in sync. Multiple replication forks (red dots) can form on the genome in order to ensure DNA is duplicated properly in these fast-growing cells. As a result, genes closer to the origin such as *sgrS* (blue) are duplicated in the same timeframe that replication is initiated (resulting in 2 or 4 gene copies), while genes closer to the terminus such as *ptsG* (orange) are replicated later in the C period, the period when a majority of DNA is duplicated (resulting in 1 or 2 gene copies). The black arrows denote the start of a cycle.

were computed on a local cluster containing AMD Opteron Interlagos cores.

2.1.2. SgrS Regulatory Network Kinetic Model

The kinetic model describing the reactions characterizing the *E. coli* glucose-phosphate response network by the small RNA SgrS is given in **Figure 4**. Simulation files are available in Jupyter Notebook format to be simulated via the Lattice Microbes (LM) Software Package at http://faculty.scs.illinois.edu/schulten/research/sgrs-2020/.

2.2. Experimental Methods and Materials

Wild type *E. coli* cells (DJ480) were grown overnight at 37 °C, 250 rpm in LB Broth. The SgrS U224G mutant was grown in LB Broth with 50 μg/ml spectinomycin (Spec) (Sigma-Aldrich). The next day, overnight cultures were diluted 100-fold into MOPS EZ rich defined medium with 0.2% glucose and the cells were grown until OD_{600} reached 0.15–0.25. α-methyl D-glucopyranoside (αMG) (Sigma Aldrich) was then added to provoke glucose-phosphate stress and induce SgrS expression response. Specific volumes of liquid were removed from the culture at 0, 2, 4, 6, 8, 10, 15, and 20 min after induction and mixed with formaldehyde (Fisher Scientific) to a final concentration of 4% for cell fixation prior to single molecule experiments. See **Supplementary Table 1** for a description of the cellular strains utilized for these experiments.

Following fixation, the cells were incubated and washed, before being permeabilized with 70% ethanol, to allow for fluorescence *in situ* hybridization (FISH). Stellaris Probe Designer was used to design the smFISH oligonucleotide probes that were ordered from Biosearch Technologies (https://www.biosearchtech.com/). See **Supplementary Table 2** for a table of the probes used in this work. Each sRNA was labeled with 9 Alexa Fluor 647 probes while each *ptsG* mRNA was labeled with 28 CF 568 probes. The labeled RNA molecules were then imaged via the super-resolution technique STORM (Stochastic Optical Reconstruction Microscopy). A density-based clustering analysis algorithm (DBSCAN) (Daszykowski et al., 2001) was utilized to calculate RNA copy numbers. The algorithm used was the same as previously published (Fei et al., 2015), but the Nps and Eps values were updated for the SgrS and *ptsG* mRNA images, since CF 568 was used instead of Alexa Fluor 568 and a 405 nm laser was used to reactivate the dyes. The SgrS (9 probes labeled with AlexaFluor 647) images were clustered using $Nps = 3$ and $Eps = 15$ and the *ptsG* mRNA (28 probes labeled with CF 568) images were clustered using $Nps = 10$ and $Eps = 25$ and these numbers were empirically chosen. A MATLAB code was used for cluster analysis.

The raw data was acquired using the Python-based acquisition software and it was analyzed using a data analysis algorithm which was based on work previously published by Babcock

$$\mathbf{D_{on,p_{\{1,2\}}} \xrightarrow{k_{t,p}} \textit{ptsG}\ mRNA} \qquad \textbf{rxn 1.1}$$

$$\mathbf{\textit{ptsG}\ mRNA \xrightarrow{\beta_p} \emptyset} \qquad \textbf{rxn 1.2}$$

$$\mathbf{D_{off,s} \underset{k_{off,D_s}}{\overset{k_{on,D_s}}{\rightleftharpoons}} D_{on,s}} \qquad \textbf{rxn 2.0}$$

$$\mathbf{D_{on,s_{\{2,4\}}} \xrightarrow{k_{t,s}} SgrS} \qquad \textbf{rxn 2.1}$$

$$\mathbf{SgrS \xrightarrow{k_{ds}} \emptyset} \qquad \textbf{rxn 2.2}$$

$$\mathbf{Hfq + SgrS \underset{k_{unbind}}{\overset{k_{bind}}{\rightleftharpoons}} Hfq{:}SgrS} \qquad \textbf{rxn 3}$$

$$\mathbf{Hfq{:}SgrS + \textit{ptsG}\ mRNA \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} Hfq{:}SgrS{:}\textit{ptsG}\ mRNA} \qquad \textbf{rxn 4}$$

$$\mathbf{Hfq{:}SgrS{:}\textit{ptsG}\ mRNA \xrightarrow{k_{cat}} Hfq} \qquad \textbf{rxn 5}$$

FIGURE 4 | Kinetic Equations of the SgrS regulatory network. $D_{on,p_{1,2}}$ refers to the gene (or DNA) for *ptsG* in 1 (low state) or 2 (high state) copies and $D_{on,s_{2,4}}$ corresponds to the gene for *sgrS* in 2 (low state) or 4 (high state) copies. $D_{on,s}$ corresponds to *sgrS* when it is in the "ON" state due to activated or solute bound transcriptional activator SgrR being bound (Vanderpool and Gottesman, 2007). k_{ds} corresponds to the experimentally measured degradation rate of SgrS when cellular Hfq is not present and k_{unbind} corresponds to the experimentally measured degradation of SgrS when Hfq was present.

et al. (2013). The peak identification and fitting were performed using the method described previously (Fei et al., 2015). The z-stabilization was done by the CRISP system and the horizontal drift was calculated using Fast Fourier Transformation (FFT) on the reconstructed images of subsets of the super-resolution image, comparing the center of the transformed images and corrected using linear interpolation.

The *ptsG* mRNA degradation rates were calculated via a rifampicin-chase experiment. The wild type (DJ480) *E. coli* cells and Δ*hfq* mutant strain SA1816 [DJ480, *laclg*, *tetR*, *spec*, Δ*hfq::kan*] cells were grown in LB Broth with the respective antibiotics at 37 °C, 250 rpm overnight. They were used to calculate the RNA degradation rates. The Δ*hfq::kan* allele was moved to create strain SA1816 constructed by P1 transduction (Miller, 1972). When the OD_{600} reached 0.15–0.25, rifampicin (Sigma-Aldrich) was added to cultures to a final concentration of 500 μg/ml. The cells were labeled by smFISH probes and analyzed by the same process described above, taking the time of rifampicin addition or αMG removal as the 0 time point. Aliquots were taken after 0, 2, 4, 6, 8, 10, 15, and 20 min (0, 2, 4, 6, and 8 min for ΔHfq strains). For the purpose of background subtraction, ΔSgrS and Δ*ptsG* mRNA strains were grown, labeled with probes and imaged in the same manner to be used for the measurement of the background signal due to the non-specific binding of Alexa Fluor 647 and CF 568. The natural logs of the RNA copy numbers were plotted against time and the slope of the linear fitting was used to calculate the RNA lifetime and then the degradation rates. SgrS degradation rates were obtained from Fei et al. (2015), where they were measured by stopping the transcription of *sgrS* by removing αMG from the media and then were imaged and analyzed to calculate the degradation rates in the same manner as was described for *ptsG* mRNA. The values for k_{cat}, k_{on}, and k_{off} for WT cells were confirmed to be within the errors reported for the values given in (Fei et al., 2015) by fitting to the experimentally measured RNA counts with the simplified model given in that work. The transcription rate of *ptsG* was determined using $k_{t,p} = \beta_p \times [p]_0$, [as described in Fei et al. (2015)], where $[p]_0$ was the average initial level of *ptsG* mRNA before stress induction. The transcription rate obtained was unchanged between the wild-type and the U224G mutant cells.

3. RESULTS

Figure 5 demonstrates the ability of our newly developed kinetic model to capture the average cellular copy number of SgrS and *ptsG* mRNA over the course of the 20 min period post-induction.

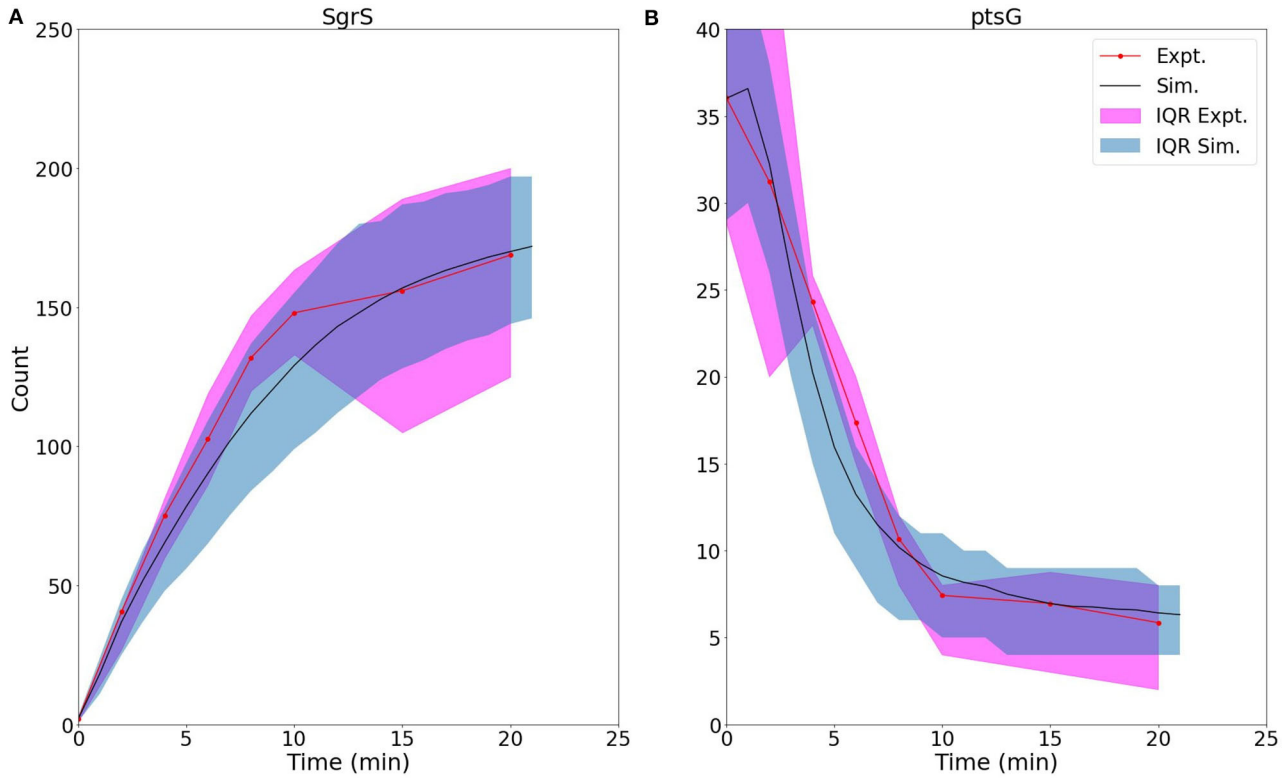


FIGURE 5 | Average time trace and interquartile range (IQR) of **(A)** labeled SgrS and **(B)** *ptsG* mRNA from both 85–169 cells from smFISH experiments (red, circled line) and 2,000 replicates from kinetic model simulations (blue, solid line). The kinetic model shows strong agreement, especially at long times (10–20 min) after induction and captures overall response behavior. An available pool of 250 Hfq and the kinetic parameters given in **Table 1** were utilized. Results considering both lower and higher available Hfq pools are discussed in **Supplementary Figure 1**.

The overlap of the interquartile range (IQR) of both the experimental and simulated cellular populations demonstrates the agreement over a variety of cells [at different gene states (i.e., high/low copy number), and RNA expression levels].

The ability of our improved kinetic model to capture population-level statistics of single cell copy number distributions of SgrS and *ptsG* mRNA is demonstrated in **Figure 6**. Kernel Density Estimates (KDE), which are used to estimate the probability densities of distributions of approximately 100–200 experimentally measured cells and 2,000 simulated cells are displayed, along with dashed vertical lines giving the average RNA copy numbers observed. KDEs were utilized to provide a reasonable comparison to the experimental values despite the fact that there were a relatively low number of cells measured at each time point (approximately 100–200) compared to the number of replicates required for appropriate stochastic simulation (2,000) (Histograms of experimental RNA counts measured before KDE imposition are given in **Supplementary Figure 7**). The distributions obtained from both experiment and the kinetic model show strong agreement (especially in the case of *ptsG* mRNA), which can be seen quantitatively by the starred line showing the Kullback–Leibler Divergence (KL Divergence) in **Figure 7**. The KL Divergence (Equation 2), which was minimized to fit to experimental RNA distributions over all time points, is a statistical measure used to characterize the difference between a probability distribution (the KDE of simulated cells) and a reference distribution (the KDE of experimentally measured cells).

The parameters obtained from the fitting process give some insight into the role of stabilization by Hfq in the SgrS-*ptsG* mRNA target search process and the role of transcriptional regulation by SgrR in the regulatory network. The pseudo first order rate of Hfq binding to SgrS (k_{bind}) is $0.063 \pm 0.014\ s^{-1}$, while the degradation rate of SgrS (k_{ds}), obtained from Δ*hfq* strain experiments (described in section 2.2), is $0.022 \pm 0.002\ s^{-1}$. The available Hfq pool size of 250 ± 167 predicted by fitting to the kinetic model seems reasonable in that average proteomics values have been found to be on the order 1,500 (Taniguchi et al., 2010; Santiago-Frangos and Woodson, 2018) and unique sRNAs have been shown to be bound to 10 to 1,000 copies of Hfq in *E. coli* (Melamed et al., 2020) (Further discussion of range of Hfq copy number is given in **Supplementary Section 1**). Additionally, the aforementioned SgrS-Hfq binding rate k_{bind} corresponds well to experimentally measured *in vitro* values for sRNA-Hfq binding for sRNA of its approximate size (Fender et al., 2010; Hopkins et al., 2011; Santiago-Frangos and Woodson, 2018).

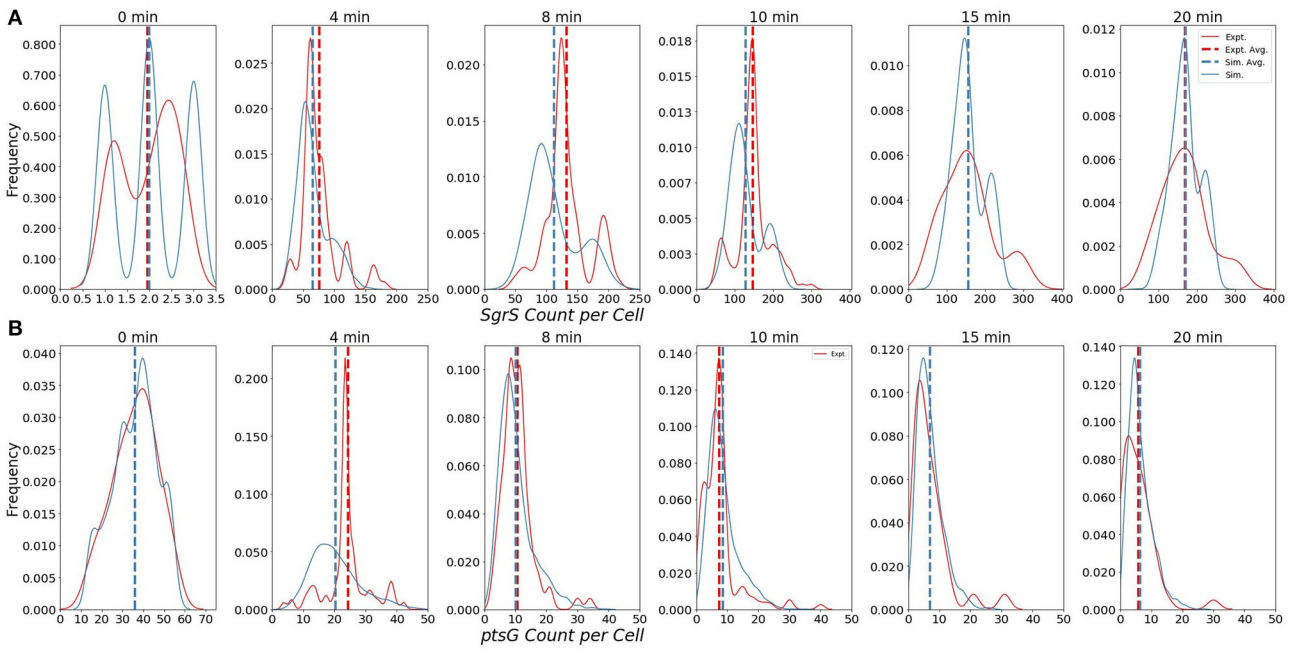


FIGURE 6 | Distributions of **(A)** Wild-Type SgrS (top) and **(B)** *ptsG* mRNA (bottom) at various time points from 0 to 20 min post-induction. Data from smFISH-STORM experiments (red, 100–200 cells per time point) and stochastic simulations (blue, 2,000 cells per time point) are shown as kernel density estimates. The effect of number of cell replicates is studied further in **Supplementary Figure 3**. Average copy number at each time point is are displayed with dashed vertical lines.

If the pseudo first order rate for k_{bind} reported in **Table 1** is converted to a bulk second order rate by incorporating the Hfq concentration at the predicted available pool size of 250, we obtain a binding rate of $1.5 \times 10^5\ M^{-1}\ s^{-1}$. This value (on the order of $1–3\ 10^5\ M^{-1}\ s^{-1}$ within the uncertainty reported in **Table 1**) agrees better with the reported value of approximately (Santiago-Frangos and Woodson, 2018) $10^6\ M^{-1}\ s^{-1}$ for long RNAs binding to Hfq (Lease and Woodson, 2004; Fender et al., 2010) than $10^8\ M^{-1}\ s^{-1}$ reported for short, unstructured RNAs binding to Hfq (Hopkins et al., 2011). Since SgrS is a relatively long sRNA (sRNA have typically been found to be between 37 and 300 nt Wang et al., 2015a with a length of 227 nucleotides, the slow sRNA-Hfq binding rate obtained by fitting seems appropriate. This type of slow sRNA association process has been suggested to be characterized by RNA restructuring (by which Hfq remodels sRNA regions in order to make its secondary structure more accessible for target mRNA base pairing) (Antal et al., 2004; Soper and Woodson, 2008; Soper et al., 2011; Bordeau and Felden, 2014), which has been proposed to occur for SgrS (Maki et al., 2010). k_{bind} is also much greater than the Hfq-SgrS unbinding rate (k_{unbind}) of $0.0018 \pm 0.0004\ s^{-1}$ which was obtained from fitting to the degradation rate of SgrS in a cell where Hfq was expressed (distinct from the Δhfq rate) by assuming that Hfq-SgrS unbinding is the rate-limiting step in the degradation of free SgrS represented in **Figure 4** (**Rxn 2.2**). These results seem reasonable in that SgrS should associate with Hfq at a rate comparable to its degradation as well as that SgrS-Hfq binding should happen at a significantly higher rate than their dissociation for sRNA chaperone stabilization by Hfq to be effective.

The kinetic values for transcriptional regulation by the activator SgrR also seem reasonable with a $k_{on,Ds}$ of $3.0 \times 10^{-2}\ s^{-1}$ and a $k_{off,Ds}$ of $9.5 \times 10^{-3}\ s^{-1}$. The gene switching parameters correspond to *sgrS* activation via SgrR binding occurring approximately 30 s after initiation of induction, with all *sgrS* genes assumed to start in the "OFF" state (the effect of starting genes in the "OFF" vs. the "ON" state is analyzed in **Supplementary Figure 2**). This seems reasonable since SgrS sRNA moves from a basal level of a few copies to greater than 40 copies on average in 2 min time (**Figure 5**). The fact that $k_{on,Ds}$ is 3 times greater than $k_{off,Ds}$ means that activation happens more frequently than deactivation from unbinding of SgrR. This relative behavior is somewhat expected as sugar shock has been induced and SgrR is believed to be transformed to its active conformation as a transcription factor for *sgrS* by binding to a small molecule at its C-terminus (Vanderpool and Gottesman, 2004, 2007). While the available evidence suggests that the activity of SgrR due to solute binding rather than *sgrR* expression affects activation of *sgrS*, it has been demonstrated that SgrR is negatively autoregulated (Vanderpool and Gottesman, 2007) which may lead to a ceiling on the level of *sgrS* activation that can occur even after glucose-phosphate stress is fully induced. Thus, we incorporate constant rates of $k_{on,Ds}$ and $k_{off,Ds}$ for *sgrS* activation in our model, instead of a time variant rate constant for either parameter.

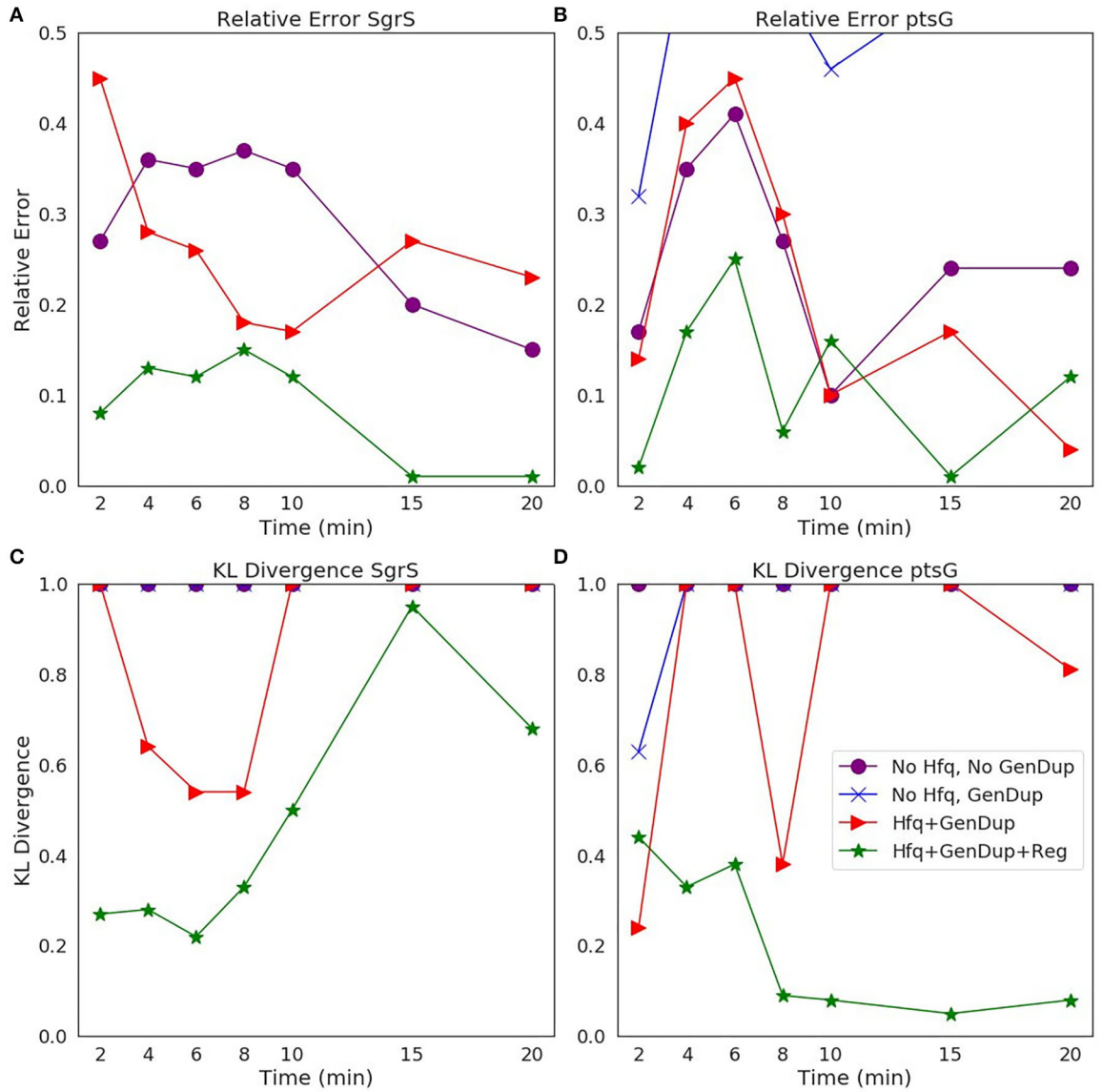


FIGURE 7 | Statistical analysis of the agreement of **(A,C)** *SgrS* sRNA and **(B,D)** *ptsG* mRNA copy number between experiment and theory on both **(A,B)** an average (Relative Error) and **(C,D)** distribution (Kullback–Leibler: KL Divergence) level. KL Divergence values for the model with no Hfq stabilization nor Gene Duplication are not shown as the values obtained are at 1.0, corresponding to significant disagreement in that model variant and experiment. GeneDup refers to a model with Gene Duplication for both SgrS and *ptsG* implemented and Reg refers to a model with transcriptional regulation of SgrS by SgrR in place. The green line (with star markers) indicates the full kinetic model used for this study, which provides the best fit to both average and population level data for both SgrS and *ptsG* mRNA.

3.1. Comparison of Goodness of Fit Based on Model Complexity

To illustrate the improvement of the kinetic model to describe cellular populations, we compare simulation results sequentially as each level of complexity (i.e., transcriptional regulation by SgrR, gene replication, and stabilization by the chaperone protein Hfq) is added to the original reduced model presented in Fei et al. (2015). **Figure 7** demonstrates the improvement in descriptiveness at both an average and population level with progression to a more fine-grained kinetic model. The relative error (Equation 1) of the average copy number of SgrS and *ptsG* mRNA gives the capability of the model to reproduce experiments on an average level, while the Kullback-Leibler Divergence (KL Divergence) (Equation 2) shows the agreement between the experimentally observed and simulation distributions of RNA copy numbers at a series of times from 0 to 20 min post induction.

The Relative Error used to illustrate the agreement between the experimentally measured average RNA copy number and the theoretical value is given by:

$$\eta = \left| \frac{Exp_{avg} - Sim_{avg}}{Exp_{avg}} \right| \quad (1)$$

where Exp_{avg} is the experimentally measured average RNA copy number at a given time point and Sim_{avg} is the simulated average RNA copy number at the same time point.

The KL Divergence used to compare agreement between experimental and simulated distributions is given by:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \quad (2)$$

where $P(i)$ is the continuous probability distribution given by the Gaussian KDE of the experimental copy number distribution of RNA (SgrS or *ptsG* mRNA) and $Q(i)$ is the analogous simulated RNA copy number distribution.

It is clear that the decrease in the KL Divergence (**Figures 7C,D**), describing the ability of the kinetic model to accurately describe cell-to-cell variation, is most substantial in the final model presented in this work (star markers). Accounting for transcriptional regulation by SgrR, ongoing gene replication, and the stabilizing effect of Hfq allows for a more faithful description of the noise observed in a cellular population in the process of sugar shock response.

3.2. Characterizing the Effects of SgrS Point Mutation on Association to Hfq and ptsG *mRNA*

The stochastic model we have presented can also be utilized to characterize the effects of *sgrS* point mutations on the regulatory network as a whole. The polyU tail region of *sgrS* comprising the final 8 residues of the 5' end (all of which are uridine in the sRNA) has previously been shown to be an important site for Hfq recruitment (Otaka et al., 2011). When the polyU tail is truncated or similarly disrupted, there is a noticeable decrease in SgrS regulatory efficiency. With this in mind, we used the previously defined kinetic model (See **Figure 4**) to characterize the effect of a point mutation resulting in a U to G change in SgrS at position 224 (in the polyU tail region, hereafter referred to as U224G) of the sRNA on regulatory kinetics. This point mutation is well downstream of the seed region (nucleotides 168–187) where SgrS-*ptsG* mRNA base pairing occurs (Maki et al., 2010; Bobrovskyy and Vanderpool, 2014) so it should not directly interfere with sRNA-mRNA interactions. It is also important to consider the possible structural effects arising from polyU tail mutation. Through *in silico* folding with the RNA structure prediction tool mFold (Zuker, 2003), we confirmed that the stability of the U224G with a ΔG of $-17.60\ kcal/mol$ is unchanged from the predicted wild-type value of $-17.60\ kcal/mol$, and also indicated that sRNA structure is conserved **Supplementary Figure 5**) and the measured wild-type ΔHfq degradation rate (see section 2.2) is appropriate for use in fitting the U224G mutant data (as a rate for **Figure 1**, **rxn 2.2**).

We then fit to the experimentally measured SgrS and *ptsG* mRNA distributions using the previously determined kinetic model (Given in **Figure 1** and **Table 1**). A robust fit describing both average behavior as well as population level variation (**Figure 8**, **Supplementary Figure 4**) was achieved primarily by modulating the rates of SgrS to Hfq binding and unbinding (k_{bind} and k_{unbind}) and the *ptsG* mRNA annealing rates k_{on} and k_{off} (which were also free parameters in this treatment) to a much lesser extent, which further demonstrates the role of the polyU tail in Hfq chaperone recruitment. The changes in the kinetic parameters of the model used to fit mutant U224G relative to the wild-type cells (WT) illustrate that the effects of this mutation on SgrS-Hfq association are much larger, relative to the subsequent annealing of SgrS to its target *ptsG* mRNA (**Table 2**) (Further discussion of mutant U224 structure is given in **Supplementary Section 4**).

The 48% decrease in the SgrS-Hfq binding rate k_{bind} and 66% increase in the unbinding rate of the sRNA and chaperone complex k_{unbind} highlight the effects of polyU tail disruption, and support previous conclusions that this is an important site for Hfq stabilization of SgrS (Otaka et al., 2011), and the regulatory efficiency of the network as a whole. The smaller relative changes in the Hfq-SgrS-*ptsG* mRNA annealing rates k_{on} and k_{off} by 32% and 22% respectively may be due to altered interactions with Hfq that impair Hfq–dependent annealing of SgrS and *ptsG* mRNA (**Supplementary Section 4**). In light of the previously discussed slow SgrS–Hfq association process, it is reasonable that RNA restructuring of Hfq may be disrupted by mutation U224G, thus leading to slower and weaker annealing to *ptsG* mRNA. One possible explanation for the disturbance of regulation in mutant U224G is the disruption of orderly transcription termination (the polyU tail is at the 3' end of *sgrS*). Such readthrough transcription has previously been ascribed to decrease the efficiency of SgrS binding to Hfq (Morita et al., 2015, 2017). Even considering values near the ceiling of the uncertainties reported in **Table 2** it seems clear that both k_{bind} and k_{on} decrease and that both k_{unbind} and k_{off} increase due to the disruption of the polyU tail at U224G, highlighting the importance of Hfq in both stabilizing SgrS and in promoting the association of SgrS to *ptsG* mRNA.

4. DISCUSSION

The construction of a stochastic kinetic model including gene replication, transcriptional regulation, and the role of the Hfq chaperone protein demonstrates the utility of combining single cell experiments with stochastic modeling. The SgrS Regulatory Network is a noisy system characterized by small numbers of sRNA and mRNA, as well as gene copy numbers that vary from cell-to-cell. This leads to the population level heterogeneity that can then be used to parameterize a kinetic model for analysis of the role of specific molecular actors, such as the chaperone Hfq, and the effects of point mutation on sRNA silencing of mRNA.

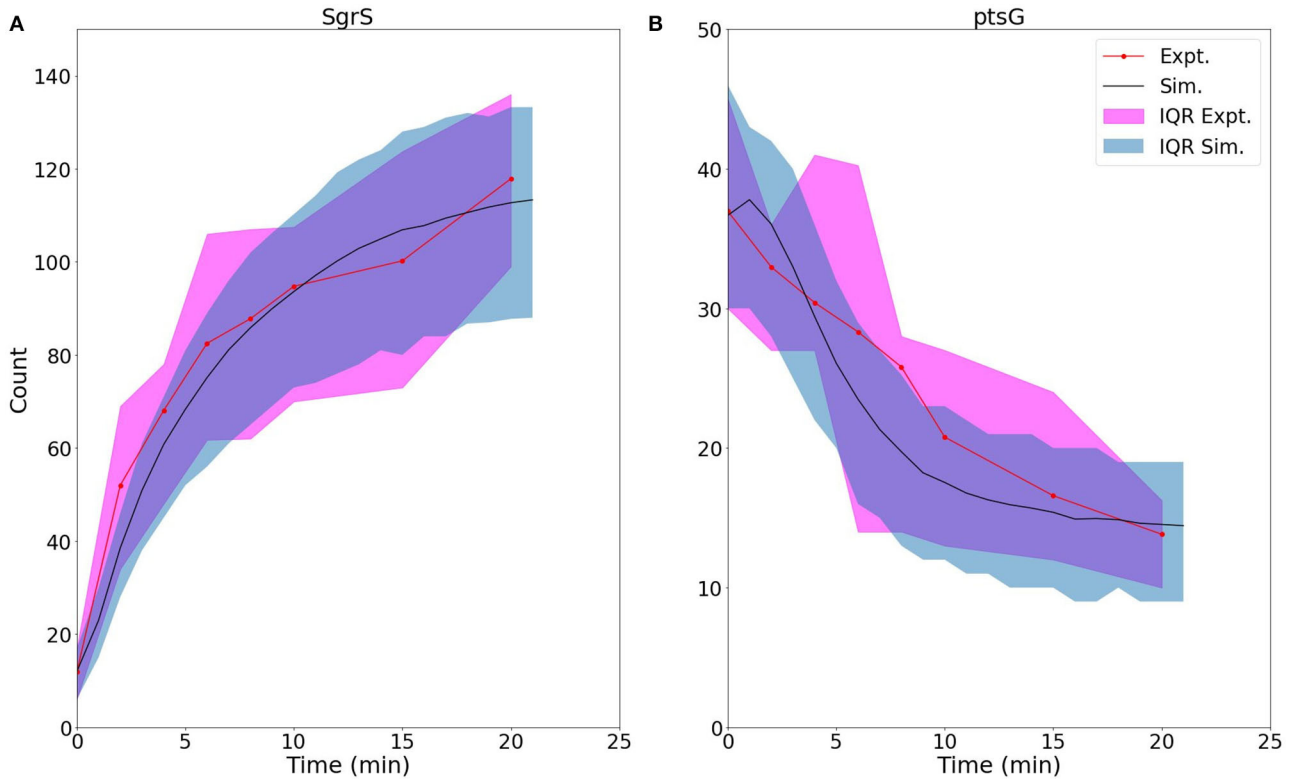


FIGURE 8 | For U224G mutant cells, average time trace and interquartile range (IQR) of **(A)** labeled SgrS and **(B)** *ptsG* mRNA from both 83–110 cells from smFISH experiments (red, circled line) and 2,000 replicates from kinetic model simulations (blue, solid line). The kinetic model shows strong agreement, especially at long times (10–20 min) after induction and captures overall response behavior. An available pool of 250 Hfq and the kinetic parameters given in **Table 1** were utilized, other than changes to SgrS-Hfq binding and unbinding rates and *ptsG* mRNA annealing and dissociation rates given in **Table 2**.

The average number of Hfq hexamers present in an *E. coli* cell has been reported to be on the order of 1,400–10,000 (2–15 μM) (Taniguchi et al., 2010; Mancuso et al., 2012; Wiśniewski and Rakus, 2014; Wang et al., 2015b; Santiago-Frangos and Woodson, 2018). It is worth noting that an extensive microfluidic-based, single-cell proteomics study that analyzed over 4,000 individual *E. coli* cells grown in similar media conditions as our study (Taniguchi et al., 2010) found a mean Hfq level of 1500. Additional immunoprecipitation and sequencing studies (by RIL-Seq) have shown the number of various individual mRNAs and sRNAs being bound to Hfq to range from 10 to 1,000 in *E. coli* (Melamed et al., 2020). Thus, our prediction (from fitting) that a pool of approximately 250 $\pm$ 167 Hfq (approximately 0.5 μM) are available to bind with SgrS sRNA at any time in the simulation of sugar shock regulation seems reasonable.

In addition, our approach allowed us to characterize the rate of Hfq-SgrS association compared to values reported for Hfq stabilization of other regulatory sRNAs. If the pseudo first order Hfq binding rate k_{bind} reported in **Table 1** is converted to a bulk second order rate we obtain a binding rate of $1.5 \times 10^5\ M^{-1}\ s^{-1}$ which agrees reasonably well with the reported value (Santiago-Frangos and Woodson, 2018) of approximately $10^6\ M^{-1}\ s^{-1}$ for long RNAs binding to Hfq (Lease and Woodson, 2004; Fender et al., 2010) (compared to the value of to $10^8\ M^{-1}\ s^{-1}$ for short, unstructured RNAs binding to Hfq Hopkins et al., 2011). SgrS is a relatively long sRNA with a length of 227 nucleotides (sRNAs have been observed with 37-300 nt Wang et al., 2015a), therefore the slow sRNA-Hfq binding process that we describe does seem likely. We suggest that this could be due to RNA restructuring of SgrS (Antal et al., 2004; Soper and Woodson, 2008; Maki et al., 2010; Soper et al., 2011; Bordeau and Felden, 2014) by Hfq in order to promote binding with *ptsG* mRNA. It is thought that cellular sRNA and mRNA are present in large excess over Hfq (Wagner, 2013), so nearly all cellular Hfq hexamers are thought to be bound to RNA. Since cellular mRNA in *E. coli* are found to be on the order of approximately 2,000–8,000 copies (Bartholomäus et al., 2016) (much greater than the highest measured SgrS sRNA value of 200) the available Hfq pool size that we present is representative of the relative competitiveness (and time-dependent cycling) of SgrS for Hfq relative to the other particles that interact with the chaperone.

The study of mutant U224G shows the importance of Hfq stabilization in the SgrS regulatory network as a whole and seems to corroborate previous findings (Otaka et al., 2011) that highlight the importance of the polyU tail for Hfq association with SgrS. The substantial decrease of the Hfq-SgrS binding rate and increase in the related unbinding rate relative to the *ptsG*

TABLE 2 | The list of kinetic parameters for SgrS-Hfq association (k_{bind} and k_{unbind}) and annealing with *ptsG* mRNA (k_{on} and k_{off}) for wild-type (WT) cells as well as SgrS mutant U224G (Reactions in **Figure 4**).

Parameter	Mutant	Value	% Difference from WT
k_{bind}	U224G	$0.033 \pm 0.010\ s^{-1}$	−48%
	WT	$0.063 \pm 0.014\ s^{-1}$	
k_{unbind}	U224G	$0.003 \pm 0.002\ s^{-1}$	+66%
	WT	$0.0018 \pm 0.0004\ s^{-1}$	
k_{on}	U224G	$(2.1 \pm 1.0) \times 10^{-4}\ molec^{-1}s^{-1}$	−32%
	WT	$(3.1 \pm 0.2) \times 10^{-4}\ molec^{-1}s^{-1}$	
k_{off}	U224G	$0.27 \pm 0.11\ s^{-1}$	+22%
	WT	$0.22 \pm 0.02\ s^{-1}$	

The substantial differences between WT and U224G for the values of k_{bind} and k_{unbind} demonstrate the disruption of Hfq binding that accompanies the mutation in the polyU tail, which has been observed previously (Otaka et al., 2011). The smaller relative changes in the ptsG mRNA annealing rates may be due to disruption of RNA restructuring (Antal et al., 2004; Soper and Woodson, 2008; Soper et al., 2011; Bordeau and Felden, 2014) of SgrS by Hfq that hampers association to the mRNA target. Calculation and analysis of parameter uncertainty values by Markov Chain Monte Carlo analysis is discussed in ***Supplementary Section 6****.*

mRNA annealing rates further down the network obtained from fitting confirms this point (**Table 2**). The changes in the SgrS-*ptsG* mRNA annealing rates k_{on} and k_{off} seem to support conclusions from the wild-type cells that Hfq-SgrS binding may result in some restructuring of the sRNA that makes this a slow process. This may explain the lower efficiency in *ptsG* mRNA association observed in mutant U224G, since Hfq cannot bind SgrS as effectively due to mutation at the polyU tail. Therefore, the predicted restructuring of SgrS by Hfq necessary to facilitate *ptsG* mRNA association is also hampered, resulting in slower and less stable mRNA binding (a decrease in k_{on} and an increase in k_{off}).

While this work is useful in describing the role of Hfq in the SgrS regulatory network and in capturing the stochastic nature of regulation over a population of replicating cells, it does not consider certain cellular processes that may affect network dynamics. First, the various other SgrS mRNA targets that may be present in a living *E. coli* cell under certain growth conditions may affect the SgrS pool available to regulate *ptsG* mRNA. In addition, other factors such as sRNA recycling (i.e., SgrS not being co-degraded with its target mRNA) (Soper et al., 2011; Santiago-Frangos and Woodson, 2018), which have been proposed for some sRNA and are now under study for SgrS, were not included, but can be incorporated into the model. Also, the process of spatial target search (via RNA and protein diffusion) of SgrS-Hfq for *ptsG* mRNA and RNase E (which may be localized in ribonucleoprotein bodies Al-Husini et al., 2018 or near the membrane Moffitt et al., 2016) for the entire protein-RNA complex as it seeks to degrade the RNA is not explicitly considered in our model (as it a well-stirred model). The potential of binding of the SgrS to *ptsG* mRNA as soon as the sRNA binding site on the mRNA is transcribed [i.e., co-transcriptional regulation which has been posited previously by Chen et al. (2019)], may be of interest to add to the model, since the model assumes only post-transcriptional binding of *ptsG* mRNA to the SgrS-Hfq complex. A further experiment that would be useful in the study of these processes would be an RIL-Seq experiment (Melamed et al., 2020) that quantifies the interactions of Hfq with other RNA (such as *yigL* or *manX*) relative to its interactions with SgrS, to better understand the pool of Hfq available for the SgrS stress response network.

In conclusion, by incorporating gene replication, stabilization by the chaperone protein Hfq, and transcriptional gene regulation of *sgrS* we have developed a kinetic model capable of describing the cellular heterogeneity observed in the *E. coli* sugar shock response network. Stochastic simulation of the kinetic model allows us to take full advantage of the single-molecule fluorescence data that illustrates cell-to-cell variability in a collection of hundreds of cells. While the post-transcriptional regulation and silencing of *ptsG* mRNA by the sRNA is the critical feature, accounting for gene replication, transcriptional regulation, and stabilization gives a more robust picture of the regulatory network as a whole. In addition, complexifying the model highlights the importance of stabilization by Hfq and chaperone proteins in general in RNA silencing networks and allowed for a prediction of the rate of association of SgrS and Hfq (as a slow process, characterized by restructuring), the effective available Hfq pool size for the SgrS regulon under sugar stress conditions, as well as an analysis of an SgrS point mutation in one of the presumed Hfq binding modules (the polyU tail). The model presented in this work establishes a framework for models analyzing other sRNA mediated gene regulatory networks, and can be extended to spatially-resolved models describing SgrS target search kinetics.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s. The Jupyter notebook containing the model used in this study as well as plotting and analysis of sample stochastic simulations can be found at: http://faculty.scs.illinois.edu/schulten/research/sgrs-2020/. Experimental smFISH-STORM data is also available at the preceding web address and is shown in **Supplementary Data Sheet 1**.

AUTHOR CONTRIBUTIONS

DB, TB, and ZL-S: writing–original draft and writing–reviewing and editing. AP: performed single-molecule experiments. MA: cultivated cell strains and performed sRNA lifetime experiments. DB: design and simulation of stochastic model. TB: writing of Jupyter notebook. AP, MA, CV, and TH: reviewing. DB, TB, AP, ZL-S, CV, and TH: conceived research plan. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from National Institutes of Health (NIGMS Grant R01 GM112659 and R35 GM122569)

and through the National Science Foundation Physics Frontiers Center: The Center for the Physics of Living Cells (CPLC) (NSF PHY 1430124).

ACKNOWLEDGMENTS

The authors acknowledge Dr. Marie Ma for participation in helpful discussions regarding SgrS stability. Substantial article content was published at the pre-print server Biorxiv at https://www.biorxiv.org/content/10.1101/2020.06.30.178566v2.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2020.593826/full#supplementary-material

REFERENCES

Al-Husini, N., Tomares, D. T., Bitar, O., Childers, W. S., and Schrader, J. M. (2018). α-proteobacterial RNA degradosomes assemble liquid-liquid phase-separated RNP bodies. *Mol. Cell* 71, 1027–1039.e14. doi: 10.1016/j.molcel.2018.08.003

Antal, M., Bordeau, V., Douchin, V., and Felden, B. (2004). A small bacterial RNA regulates a putative ABC transporter. *J. Biol. Chem*. 280, 7901–7908. doi: 10.1074/jbc.M413071200

Babcock, H. P., Moffitt, J. R., Cao, Y., and Zhuang, X. (2013). Fast compressed sensing analysis for super-resolution imaging using l1-homotopy. *Opt. Exp*. 21:28583. doi: 10.1364/OE.21.028583

Babski, J., Maier, L.-K., Heyer, R., Jaschinski, K., Prasse, D., Jäger, D., et al. (2014). Small regulatory RNAs in archaea. *RNA Biol*. 11, 484–493. doi: 10.4161/rna.28452

Balasubramanian, D., and Vanderpool, C. K. (2013). Deciphering the interplay between two independent functions of the small RNA regulator SgrS in salmonella. *J. Bacteriol*. 195, 4620–4630. doi: 10.1128/JB.00586-13

Bartholomäus, A., Fedyunin, I., Feist, P., Sin, C., Zhang, G., Valleriani, A., et al. (2016). Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci*. 374:20150069. doi: 10.1098/rsta.2015.0069

Bobrovskyy, M., Azam, M. S., Frandsen, J. K., Zhang, J., Poddar, A., Ma, X., et al. (2019). Determinants of target prioritization and regulatory hierarchy for the bacterial small RNA SgrS. *Mol. Microbiol*. 4:61. doi: 10.1111/mmi.14355

Bobrovskyy, M., and Vanderpool, C. K. (2014). The small RNA SgrS: roles in metabolism and pathogenesis of enteric bacteria. *Front. Cell. Infect. Microbiol*. 4:61. doi: 10.3389/fcimb.2014.00061

Bordeau, V., and Felden, B. (2014). Curli synthesis and biofilm formation in enteric bacteria are controlled by a dynamic small RNA module made up of a pseudoknot assisted by an RNA chaperone. *Nucleic Acids Res*. 42, 4682–4696. doi: 10.1093/nar/gku098

Chen, J., Morita, T., and Gottesman, S. (2019). Regulation of transcription termination of small RNAs and by small RNAs: molecular mechanisms and biological functions. *Front. Cell. Infect. Microbiol*. 9:201. doi: 10.3389/fcimb.2019.00201

Cooper, S., and Helmstetter, C. E. (1968). Chromosome replication and the division cycle of *Escherichia coli*. *J. Mol. Biol*. 31, 519–540. doi: 10.1016/0022-2836(68)90425-7

Daszykowski, M., Walczak, B., and Massart, D. (2001). Looking for natural patterns in data. *Chenom. Intell. Lab. Syst*. 56, 83–92. doi: 10.1016/S0169-7439(01)00111-3

Earnest, T. M., Cole, J. A., and Luthey-Schulten, Z. (2018). Simulating biological processes: stochastic physics from whole cells to colonies. *Rep. Prog. Phys*. 81:052601. doi: 10.1088/1361-6633/aaae2c

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science* 297, 1183–1186. doi: 10.1126/science.1070919

Fei, J., Singh, D., Zhang, Q., Park, S., Balasubramanian, D., Golding, I., et al. (2015). Determination of *in vivo* target search kinetics of regulatory noncoding RNA. *Science* 347, 1371–1374. doi: 10.1126/science.1258849

Fender, A., Elf, J., Hampel, K., Zimmermann, B., and Wagner, E. G. H. (2010). RNAs actively cycle on the sm-like protein hfq. *Genes Dev*. 24, 2621–2626. doi: 10.1101/gad.591310

Hallock, M. J., and Luthey-Schulten, Z. (2016). "Improving reaction kernel performance in lattice microbes: particle-wise propensities and run-time generated code," in *Parallel and Distributed Processing Symposium Workshop (IPDPSW), 2016 IEEE International* (Chicago, IL), 428–343. doi: 10.1109/IPDPSW.2016.118

Hallock, M. J., Stone, J. E., Roberts, E., Fry, C., and Luthey-Schulten, Z. (2014). Simulations of reaction diffusion processes over biologically-relevant size and time scales using multi-GPU workstations. *Parallel Comput*. 40, 86–99. doi: 10.1016/j.parco.2014.03.009

Hopkins, J. F., Panja, S., and Woodson, S. A. (2011). Rapid binding and release of hfq from ternary complexes during RNA annealing. *Nucleic Acids Res*. 39, 5193–5202. doi: 10.1093/nar/gkr062

Ishikawa, H., Otaka, H., Maki, K., Morita, T., and Aiba, H. (2012). The functional hfq-binding molecule of bacterial sRNAs consists of a double or single hairpin preceded by a u-rich sequence and followed by a 3' poly(u) tail. *RNA* 18, 1062–1074. doi: 10.1261/rna.031575.111

Jones, D. L., Brewster, R. C., and Phillips, R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science* 346, 1533–1536. doi: 10.1126/science.1255301

Kawamoto, H., Koide, Y., Morita, T., and Aiba, H. (2006). Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by hfq. *Mol. Microbiol*. 61, 1013–1022. doi: 10.1111/j.1365-2958.2006.05288.x

Lease, R. A., and Woodson, S. A. (2004). Cycling of the SM-like protein hfq on the DSRA small regulatory RNA. *J. Mol. Biol*. 344, 1211–1223. doi: 10.1016/j.jmb.2004.10.006

Lee, S.-J. (2000). Signal transduction between a membrane-bound transporter, PtsG, and a soluble transcription factor, Mlc, of *Escherichia coli*. *EMBO J*. 19, 5353–5361. doi: 10.1093/emboj/19.20.5353

Maier, T., Schmidt, A., Güell, M., Kühner, S., Gavin, A.-C., Aebersold, R., et al. (2011). Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol. Syst. Biol*. 7:511. doi: 10.1038/msb.2011.38

Maki, K., Morita, T., Otaka, H., and Aiba, H. (2010). A minimal base-pairing region of a bacterial small RNA SgrS required for translational repression of ptsG mRNA. *Mol. Microbiol*. 76, 782–792. doi: 10.1111/j.1365-2958.2010.07141.x

Mancuso, F., Bunkenborg, J., Wierer, M., and Molina, H. (2012). Data extraction from proteomics raw data: an evaluation of nine tandem MS tools using a large orbitrap data set. *J. Proteomics* 75, 5293–5303. doi: 10.1016/j.jprot.2012.06.012

Melamed, S., Adams, P. P., Zhang, A., Zhang, H., and Storz, G. (2020). RNA-RNA interactomes of ProQ and hfq reveal overlapping and competing roles. *Mol. Cell* 77, 411–425.e7. doi: 10.1016/j.molcel.2019.10.022

Miller, J. (1972). *Experiments in Molecular Genetics*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory.

Moffitt, J. R., Pandey, S., Boettiger, A. N., Wang, S., and Zhuang, X. (2016). Spatial organization shapes the turnover of a bacterial transcriptome. *eLife* 5:e13065. doi: 10.7554/eLife.13065.029

Morita, T., Nishino, R., and Aiba, H. (2017). Role of the terminator hairpin in the biogenesis of functional hfq-binding sRNAs. *RNA* 23, 1419–1431. doi: 10.1261/rna.060756.117

Morita, T., Ueda, M., Kubo, K., and Aiba, H. (2015). Insights into transcription termination of hfq-binding sRNAs of *Escherichia coli* and characterization of readthrough products. *RNA* 21, 1490–1501. doi: 10.1261/rna.051870.115

Nam, T.-W., Jung, H. I., An, Y. J., Park, Y.-H., Lee, S. H., Seok, Y.-J., et al. (2008). Analyses of mlc-IIBGlc interaction and a plausible molecular mechanism of

mlc inactivation by membrane sequestration. *Proc. Natl. Acad. Sci. U.S.A*. 105, 3751–3756. doi: 10.1073/pnas.0709295105

Otaka, H., Ishikawa, H., Morita, T., and Aiba, H. (2011). PolyU tail of rho-independent terminator of bacterial small RNAs is essential for hfq action. *Proc. Natl. Acad. Sci. U.S.A*. 108, 13059–13064. doi: 10.1073/pnas.1107050108

Peterson, J. R., Cole, J. A., Fei, J., Ha, T., and Luthey-Schulten, Z. A. (2015). Effects of DNA replication on mRNA noise. *Proc. Natl. Acad. Sci. U.S.A*. 112, 15886–15891. doi: 10.1073/pnas.1516246112

Peterson, J. R., Hallock, M. J., Cole, J. A., and Luthey-Schulten, Z. A. (2013). "A problem solving environment for stochastic biological simulations," in *PyHPC 2013, Supercomputing 2013* (Denver, Co).

Raser, J. M. (2005). Noise in gene expression: origins, consequences, and control. *Science* 309, 2010–2013. doi: 10.1126/science.1105891

Roberts, E., Stone, J. E., and Luthey-Schulten, Z. (2013). Lattice Microbes: high-performace stochastic simulation method for the reaction-diffusion master equation. *J. Comp. Chem*. 3, 245–255. doi: 10.1002/jcc.23130

Santiago-Frangos, A., and Woodson, S. A. (2018). Hfq chaperone brings speed dating to bacterial sRNA. *Wiley Interdiscipl. Rev. RNA* 9:e1475. doi: 10.1002/wrna.1475

Seitz, S., Lee, S.-J., Pennetier, C., Boos, W., and Plumbridge, J. (2003). Analysis of the interaction between the global regulator Mlc and EIIBGlcof the glucose-specific phosphotransferase system in *Escherichia coli*. *J. Biol. Chem*. 278, 10744–10751. doi: 10.1074/jbc.M212066200

Soper, T. J., Doxzen, K., and Woodson, S. A. (2011). Major role for mRNA binding and restructuring in sRNA recruitment by hfq. *RNA* 17, 1544–1550. doi: 10.1261/rna.2767211

Soper, T. J., and Woodson, S. A. (2008). The rpoS mRNA leader recruits hfq to facilitate annealing with DsrA sRNA. *RNA* 14, 1907–1917. doi: 10.1261/rna.1110608

Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., et al. (2010). Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329, 533–538. doi: 10.1126/science.1188308

Vanderpool, C. K., and Gottesman, S. (2004). Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. *Mol. Microbiol*. 54, 1076–1089. doi: 10.1111/j.1365-2958.2004.04348.x

Vanderpool, C. K., and Gottesman, S. (2007). The novel transcription factor SgrR coordinates the response to glucose-phosphate stress. *J. Bacteriol*. 189, 2238–2248. doi: 10.1128/JB.01689-06

Wagner, E. G. H. (2013). Cycling of RNAs on hfq. *RNA Biol*. 10, 619–626. doi: 10.4161/rna.24044

Wang, J., Liu, T., Zhao, B., Lu, Q., Wang, Z., Cao, Y., et al. (2015a). sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria. *Nucleic Acids Res*. 44, D248–D253. doi: 10.1093/nar/gkv1127

Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015b). Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15, 3163–3168. doi: 10.1002/pmic.201400441

Wang, M., Zhang, J., Xu, H., and Golding, I. (2019). Measuring transcription at a single gene copy reveals hidden drivers of bacterial individuality. *Nat. Microbiol*. 4, 2118–2127. doi: 10.1038/s41564-019-0553-z

Wiśniewski, J. R., and Rakus, D. (2014). Quantitative analysis of the *Escherichia coli* proteome. *Data Brief* 1, 7–11. doi: 10.1016/j.dib.2014.08.004

Youngren, B., Nielsen, H. J., Jun, S., and Austin, S. (2014). The multifork *Escherichia coli* chromosome is a self-duplicating and self-segregating thermodynamic ring polymer. *Genes Dev*. 28, 71–84. doi: 10.1101/gad.231050.113

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. 31, 3406–3415. doi: 10.1093/nar/gkg595

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: 30 March 2021
doi: 10.3389/fmolb.2021.636562

Impact of Structural Observables From Simulations to Predict the Effect of Single-Point Mutations in MHC Class II Peptide Binders

Rodrigo Ochoa[1,2], *Roman A. Laskowski*[2], *Janet M. Thornton*[2] *and Pilar Cossio*[1,3]*

[1]Biophysics of Tropical Diseases, Max Planck Tandem Group, University of Antioquia UdeA, Medellin, Colombia, [2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom, [3]Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Frankfurt am Main, Germany

The prediction of peptide binders to Major Histocompatibility Complex (MHC) class II receptors is of great interest to study autoimmune diseases and for vaccine development. Most approaches predict the affinities using sequence-based models trained on experimental data and multiple alignments from known peptide substrates. However, detecting activity differences caused by single-point mutations is a challenging task. In this work, we used interactions calculated from simulations to build scoring matrices for quickly estimating binding differences by single-point mutations. We modelled a set of 837 peptides bound to an MHC class II allele, and optimized the sampling of the conformations using the Rosetta backrub method by comparing the results to molecular dynamics simulations. From the dynamic trajectories of each complex, we averaged and compared structural observables for each amino acid at each position of the 9˚mer peptide core region. With this information, we generated the scoring-matrices to predict the sign of the binding differences. We then compared the performance of the best scoring-matrix to different computational methodologies that range in computational costs. Overall, the prediction of the activity differences caused by single mutated peptides was lower than 60% for all the methods. However, the developed scoring-matrix in combination with existing methods reports an increase in the performance, up to 86% with a scoring method that uses molecular dynamics.

Keywords: MHC class II, single-point mutation, structural bioinformatics, simulations, binding

OPEN ACCESS

Edited by:
Gregory Bowman, Washington University School of Medicine in St. Louis, United States

Reviewed by:
Carlo Camilloni, University of Milan, Italy
Sophie Sacquin-Mora, UPR9080 Laboratoire de Biochimie Théorique, France

***Correspondence:**
Pilar Cossio
pilar.cossio@biophys.mpg.de

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

Received: *01 December 2020*
Accepted: *15 February 2021*
Published: *30 March 2021*

Citation:
Ochoa R, Laskowski RA, Thornton JM and Cossio P (2021) Impact of Structural Observables From Simulations to Predict the Effect of Single-Point Mutations in MHC Class II Peptide Binders. Front. Mol. Biosci. 8:636562. doi: 10.3389/fmolb.2021.636562

INTRODUCTION

The Major Histocompatibility Complex (MHC) class II is a key receptor responsible for recognizing fragments of proteins belonging to external pathogens, as well as recognizing human proteins that can boost the emergence of autoimmune events and immunological processes (Wieczorek et al., 2017). The structures of multiple MHC class II alleles have been elucidated. They are composed of α and β chains split into four sub-units, two of them forming a groove where the peptides bind (Bjorkman, 2015) (see **Supplementary Figure S1**). The peptides contain a core region, which is a fragment of nine amino acids responsible to stabilize the peptide-MHC class II interaction. The peptide-core binds in four key pockets of the receptor that are formed between the α and β chains (Unanue et al., 2016). The available structures of MHC class II bound to peptides provide

information about the binding poses, which are commonly in a polyproline II-like extended conformation (Bermúdez et al., 2014). Understanding the preference of amino acids for certain positions is relevant to comprehending how epitopes can trigger adaptive immune responses (Unanue et al., 2016). Moreover, this structural information allows us to study the physicochemical interactions within key pockets in the binding groove, which is crucial to stabilizing the complexes (Yeturu et al., 2010).

These structural insights are usually not included in the prediction tools of peptides binding to the MHC class II. The lack of structural and dynamical representations of the complexes, as well as the demand on computational resources, are some of the limitations (Zhang et al., 2010). Instead, researchers have focused on generating profiles and motifs of sequences using information from bioactivity datasets (Wang et al., 2008). The main purpose of these tools is to rank peptide-binders by their predicted affinities, and associate the values to a potential immunological response. Among the available approaches, the most common ones are machine learning models trained with a diverse set of peptides bound to different MHC class II alleles (Andreatta et al., 2015; Peters et al., 2020). Some other researchers have focused on creating position-specific scoring matrices that can be implemented to select peptide candidates through simple bioinformatics pipelines, and to predict which core region of the peptide is responsible for the interaction with the main pockets of the receptor (Rapin et al., 2008). The available tools cover a diverse set of MHC class II alleles, providing clues for researchers working on the design of vaccines and immunotherapies (Nandy and Basak, 2016).

One particular challenge about the binding predictions is to evaluate affinity differences for single-point mutations on the peptide. Efforts have been focused to understand the impact of such mutations in the context of protein function, participation in molecular pathways and changes in their physico-chemical properties (Bogan and Thorn, 1998; Tokuriki et al., 2007; Hopf et al., 2017). From a structural perspective, coordinates can be used as input to predict the side chain conformations of the mutated amino acids, and assess their impact from a stability or binding perspective (Li et al., 2014). In the case of MHC class II, sequence-based strategies can be implemented to predict these activity differences, but structural and dynamical insights about the mechanisms behind these modifications are also relevant (Kuhlman and Bradley, 2019; Aranha et al., 2020). Many of these methods rely on energy evaluations to check differences in terms of solvent exposure, generation of hydrogen bonds, electrostatics contributions, backbone and side chain flexibility, and weak interactions such as van der Waals (Sammond et al., 2007; Slutzki et al., 2015; Barlow et al., 2018). Understanding the main drivers of these affinity differences is relevant for the design and discovery of novel peptide binders.

Methods using structural and dynamical information can be implemented to assess the role of the peptide/receptor conformations in the binding affinity and stability (Antunes et al., 2018; Lanzarotti et al., 2018). For MHC class II, the crystal structures show that peptides tend to bind in similar conformations for the available alleles (Wieczorek et al., 2016). Therefore, these structures can be used as templates to model other peptides bound to the receptor, and enable the study of how modifications can affect the binding from a physicochemical perspective (Ochoa et al., 2019). These models can be subjected to conformational sampling to analyze the fluctuations of the complexes in equilibrium (Ferrante et al., 2015) and score the most favourable conformations (Cossio et al., 2012; Sarti et al., 2013).

Among the sampling approaches, molecular dynamics (MD) has proved to be a useful way of studying the conformational space of peptides bound to MHC class II structures (Omasits et al., 2008; Ochoa et al., 2019). However, the scalability is limited by the required computational resources if large sets of peptides are analyzed. One option is to implement Monte Carlo algorithms to obtain representative structures of the complexes in equilibrium (King and Bradley, 2010). This is the case of the backrub method from Rosetta, where the backbone flexibility is modelled based on observations from high-resolution crystal structures (Smith and Kortemme, 2008). The movements are mainly backbone rotations around the axes of C_α atoms that are accepted using a Metropolis criterion based on the minimization of a bond-angle penalty imposed by the chosen force fields (Smith and Kortemme, 2010). The trajectories provide information on the system's intrinsic flexibility, solvent accessibility and the main interactions (e.g., hydrogen bonds, non-bonded contacts) formed by the amino acids.

In this work, we evaluated how two kinds of molecular interactions can aid in the prediction of the affinity-differences for single modifications in the core region of a set of MHC class II peptide binders. For this purpose, we created a set of scoring matrices, as is typically done using sequence analysis, but here derived from structural observables from simulations of a large set of peptides/MHC class II complexes. The matrices allow the estimation of binding differences caused by single-point mutations, and complement current state-of-the-art methods to improve the predictions. We modelled a large set of peptides with binding data available for one representative MHC class II allele. Then, we sampled the conformational space using the backrub method optimized to reproduce the finite-temperature ensemble from molecular dynamics simulations. Hydrogen bonds and contact interactions were used to calculate the scoring-matrices SM-HB and SM-C, respectively, from the structural descriptors per core position in the peptide. The magnitude and stability of these observables were associated to binding differences of single-modified peptides.

In addition, five other approaches, having a wide range of computational costs, as well as accuracy, were assessed to predict the binding differences. Specifically, two sequence-based methods were implemented, which involve the use of a motif matrix to predict the most probable amino acids of the peptide core regions, and a machine learning tool used to predict binding affinities for this system. The third and fourth methods are a previously benchmarked structural/dynamical approach using an MD/scoring and backrub/scoring combination to rank peptides bound to the MHC class II (Ochoa et al., 2019). Finally, a

Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) approach is used to calculate average energies per peptide based on the MD trajectories obtained in the previous strategy. In general, the predictions had an accuracy below 60% for all the methods, but combining the best scoring-matrix SM-HB (i.e., the one generated from the hydrogen bonds) with the existing methods improves the performance.

MATERIALS AND METHODS

In the following, we first explain how we build the scoring-matrices based on the structural observables. Then, we describe how to evaluate their impact on activity differences caused by single-point mutations on the peptide binders. This is followed by a description of the additional methods used for comparison and their combination with the developed structural scoring-matrices.

Structural Scoring-Matrices From Simulations

To evaluate the impact of structural interactions, we created a set of scoring matrices based on hydrogen bonds (SM-HB) and contacts (SM-C) generated between the peptide core region and residues of the MHC class II binding site. For that purpose, we first optimized the conformational sampling of MHC class II structures using the Rosetta backrub method (Davis et al., 2006) in comparison to MD simulations. Then, we modelled a large dataset of known peptide binders of the same MHC class II allele, and with the observables we generated the scoring matrices. A detailed explanation is presented below.

Conformational Sampling Optimization

Before checking the role of the structural interactions, we assessed the conformational sampling of the Monte Carlo backrub method in Rosetta in comparison to MD simulations to explore conformations of crystal structures of peptides bound to MHC class II alleles. We selected a set of 10 peptide-MHC class II crystal structures from the Protein Data Bank (PDB) (Berman et al., 2000) of the most widely studied allele, DRB1*01:01 (see Supplementary Text for details about the structure selection). We used this benchmark to compare molecular dynamics (MD) and Monte Carlo backrub simulations.

Molecular dynamics: Each crystal structure was subjected to MD simulations of 20 nanoseconds (ns) with previous minimization and NVT/NPT equilibration phases, using GROMACS v5.1 (Hess et al., 2008). The main MD parameters are described in the Supplementary Text. A temperature of 350 K was chosen to perform the simulations, allowing a fast exploration of the conformational space. Since we are interested only in the peptide-receptor interactions, all the protein atoms located at a distance greater than 12 Å from any peptide atom were restrained.

Backrub Monte Carlo: The same crystal structures were subjected to Metropolis Monte Carlo simulations using the backrub algorithm (Davis et al., 2006) available in RosettaCommons version 2016.32 (www.rosettacommons.org). A total of 50,000 Monte Carlo trials were run per complex using two kT values: 0.35 and 1.2. C_α atoms were chosen as pivots for all the protein and peptide residues. The minimum backrub segment size in atoms was 3, and the maximum segment was 64. The probabilities for sampling side chain and backbone torsions were set at the default values. The simulations were run over a single core for each complex. An optimal backrub parameter setup was selected in order to reproduce the equilibrium ensemble from MD.

Structural observables

Several structural observables were used to characterize the conformations from the different simulations:

Side chain dihedrals: The side chain dihedrals χ_1 and χ_2 were monitored for all the amino acids belonging to the peptide, and the distributions were compared to that obtained from MD. The Kullback-Leibler divergence metric (Hershey and Olsen, 2007) was implemented to compare the distributions.

Main chain hydrogen bonds: We monitored interactions made by the amino acids of the peptide core region with the receptor. Specifically, we calculated the number of potential hydrogen bonds made by the backbone atoms using the HBPLUS program (McDonald and Thornton, 1994).

Contacts: We also calculated the number of non-bonded contacts with a threshold of 4 Å between the atoms of the peptide and those of the receptor using Biopython modules (Cock et al., 2009).

The latter two observables were also used to calculate the structural descriptors for creating the scoring-matrices.

Modelling and Simulations of a Large Dataset of Peptides Bound to MHC Class II

After establishing the best backrub simulation setup, a set of peptides with available binding data for different MHC class II alleles was modelled and simulated to calculate scoring-matrices from the chosen structural descriptors.

First, we selected as a representative structure of the allele DRB1*01:01 and the crystal structure with PDB id 1T5X, which is co-crystallized with a peptide that we used as template to model peptide binders with bioactivity information available. For the peptides, we used a public dataset containing 44,541 measured affinities covering 26 MHC class II alleles (Wang et al., 2010). We selected a total of 837 15 mer peptides for the allele DRB1*01:01 after applying the filter below, with activities from 1 to 10,000 nM. The filter was the prediction of the 9 mer core region of each selected peptide using two methods. The first was based on available motifs derived from a position-specific scoring matrix published for several MHC class II alleles (Rapin et al., 2008). The sequences were analyzed over windows of nine amino acids, where each fragment was scored to obtain a ranked list of fragments with probabilities of being the core region of the peptide interacting with the MHC class II receptor. For the second method, we implemented the NetMHCIIpan-4.0 tool, which has as its main goal the prediction of affinities for peptides bound to MHC class II molecules, and also the prediction of the 9°mer core regions of the peptide sequences (Andreatta et al.,

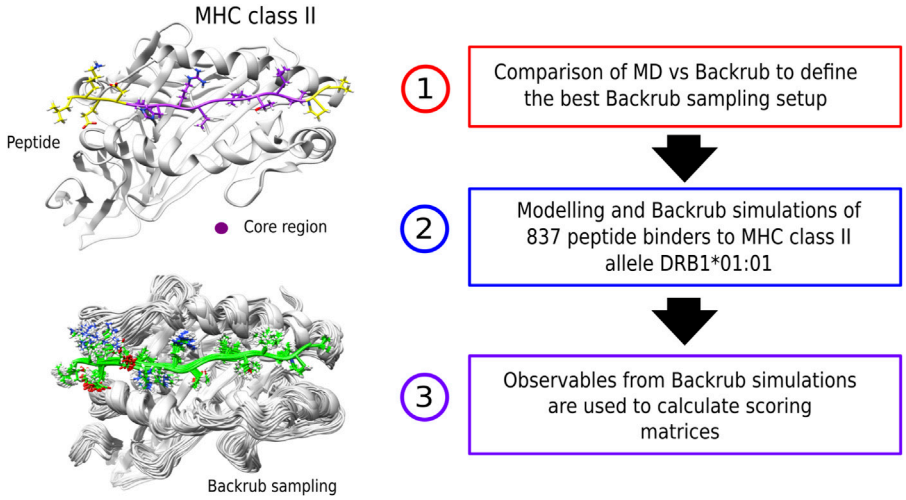


FIGURE 1 | (left) Examples of a peptide bound to an MHC class II receptor and conformations from the Backrub Rosetta simulations. **(right)** Schematic representation of the methodological steps that involve creation of the scoring-matrices. First, an MD vs. Backrub comparison was performed to define the best Backrub setup. Then, the modelling and sampling of a set of known peptide binders was performed to obtain the observables for building the scoring-matrices.

2015). A peptide was selected when both methods predicted identical core regions with the highest scores. As a final step, we aligned the predicted core region with the core from the peptide template. If, after the alignment, we needed to add more than two amino acids for either flanking region (N or C-terminal), the peptide was discarded.

We modelled the selected peptides by iterative single substitutions of the peptide template sequence. The mutations were performed with the package fixbb from Rosetta (Loffler et al., 2017), which was compared in a previous study to other available mutation protocols (Ochoa et al., 2018). The method selects the most probable rotamer from a dictionary of backbone-dependent conformations. After each mutation, the side chain atoms were relaxed with the backbone fixed. The modelling of additional amino acids in the flanking region, when required, was done with the Remodel package from RosettaCommons (Huang et al., 2011), where the new amino acid was subjected to the prediction of the rotamer with relaxation of the side chains.

For each peptide, the backrub simulation from Rosetta was applied with $kT = 1.2$ (Smith and Kortemme, 2008) as found to be optimal (see the Results). Each Monte Carlo simulation had 200,000 trials. We obtained 2,000 frames per simulation, and the previously described interactions (see Methods *Structural observables*) were calculated per amino acid in all the core positions for each frame. We did the same under three other scenarios: (i) using the last 1,000, (ii) using the 1,000 frames with best energy-scores [i.e., backrub scoring function (Alford et al., 2017)], and (iii) using the single frame with the lowest energy. A summary of the modelling and sampling strategies is shown on **Figure 1**.

Definition of the Scoring-Matrices

We calculated averages of the observables per amino acid in each position of the core to define scoring-matrices of the structural descriptors. The averages covered the number of amino acids available in the dataset per position in the core region. For each position in the core region, we calculated a vector with 20 indices (one per each natural amino acid) using the average of the observable from the backrub trajectory. At the *i*th core position for amino acid type *j*, the average observable $\mathcal{O}$ is defined as

$$\mathcal{O}_{ij} = \frac{1}{N_f} \sum_{\alpha} \sum_{f} o_{ij}^{\alpha f}, \quad (1)$$

where o is the observable, f is the frame number, N_f is the total number of frames and α indexes the simulation run (having one simulation for each binding-peptide from the dataset). In the case of using just one frame from the backrub trajectory, only the average across the simulation is calculated. If an amino acid type is not found at a given position for a certain run, the observable is taken as zero. We note that the amino acid distribution is not homogeneous but it is given by the natural-occurring frequencies found in the dataset. These intrinsic frequencies are implicitly taken into account in **Eq. 1**. This allows us to improve the available motifs of peptides binding to MHC class II alleles by adding weights due to the structural observables.

The Scoring-Matrix for a Given Observable is Defined as

$$E_{ij} = -\ln \frac{\mathcal{O}_{ij}}{\sum_j \mathcal{O}_{ij}} \quad (2)$$

which provides a scoring-energy for each amino acid (j) in each core position (i).

To visualize the frequency contribution of each amino acid on the peptide library and the scoring-matrices, logoplots were generated using the WebLogo3 server (Crooks et al., 2004).

Assessment of Single-point Mutation Activity Predictions

The obtained scoring-matrices, SM-HB and SM-C, were compared to other methods based on their capability to predict single-point mutation activity differences. The test consisted of predicting the sign of the experimental $\Delta\Delta G$ for each pair of peptides differing by single-point mutations in the peptide core region. A total of 112 peptides forming 56 pairs were selected and not used to calculate the scoring-matrices from the descriptors. One requirement to select the pairs of peptides is the prediction of identical core regions with high reliability, based on the same criteria used to model the peptides (see *Modelling and simulations of a large dataset of peptides bound to MHC class II*).

Additional Methodologies for Comparison

Five additional methods were used to compare and complement the results with the scoring-matrices. These methods are:

A sequence motif reported for allele DRB1*01:01 was used to compare the probabilities of finding an amino acid in the core region (Rapin et al., 2008). The higher the value in the matrix indicates a higher probability. The difference in probabilities was used to compare the differences in affinity.

The tool NetMHCIIpan was used to predict a numerical affinity per peptide. The sign of the predicted difference is compared to the sign of the experimental values for assessing the performance.

A hybrid MD/scoring approach was also used to predict the sign of the activity difference using structural models of the peptides based on a previously published protocol (Ochoa et al., 2019). In summary, each peptide was subjected to MD simulations of 10 ns using the same MD setup as explained previously. Each frame of the last half of the trajectory was scored using six different scoring functions: Haddock (Dominguez et al., 2003), Vina (Trott and Olson, 2009), a combination of DFIRE and GOAP (DFIRE-GOAP) (Yang and Zhou, 2008; Zhou and Skolnick, 2011), Pisa (Krissinel and Henrick, 2007), FireDock (Andrusier et al., 2007), and the BMF-BLUUES scoring combination (Berrera et al., 2003; Fogolari et al., 2012). If three or more scoring functions predicted the sign of the score differences equal to the sign of the experimental activity differences, it was counted as a match to assess the performance.

A hybrid backrub/scoring approach as explained in the previous strategy, using 50,000 Monte Carlo trials per run with a kT of 1.2. The backrub trajectory was scored using the same scoring functions and consensus criterion to match the sign of the activity difference.

Finally, as the most exhaustive approach, we calculated average energies per peptide complex using the MM-PBSA methodology. For that purpose we used the g_mmpbsa plugin (Kumari et al., 2014) to calculate the solvated and non-solvated terms using as input the MD trajectories of 10 ns calculated in the third strategy.

Combination With the Structural Scoring-Matrices

To improve the performances, we combined the previous approaches with the scoring-matrices results. Specifically, we evaluated if using the scoring-matrices together with other methods can increase the number of predictions after checking by pairs if either of the two methods predicts correctly the sign of the mutation activity difference. This analysis works as a conditional "or" to evaluate how many cases can be covered using more than one method, and subsequently observe how many predictions match the experimental data.

TABLE 1 | Percentage of amino acids per backrub configuration (kT = 0.35 and kT = 1.2) for each side chain dihedral that sampled the conformational space similarly to MD simulations among all the 10 MHC class II crystal structures

Side chain dihedrals	kT = 0.35 (%)	kT = 1.2 (%)
χ_1	19.2	80.8
χ_2	12.9	87.1

RESULTS

To evaluate the impact of interactions in affinity changes caused by single-point mutations in MHC class II peptide binders, a set of scoring-matrices was calculated to assign probabilities for each type of amino acid in each position of the peptide core region. The matrices are created using the main chain hydrogen bonds (SM-HB), and the non-bonded contacts (SM-C) obtained from trajectories of peptides in complex with the MHC class II allele. To optimize the sampling, we first compared the Backrub approach to the results from MD simulations, in order to guarantee enough conformational exploration with computationally efficiency.

Optimization of the Structural Scoring-Matrices

Backrub Simulation Optimization

We optimized the kT parameter used in backrub Rosetta simulations in comparison to finite temperature MD simulations. We used as benchmark a set of 10 peptide-MHC class II structures of allele DRB1*01:01, available in the PDB (see Supplementary Text, **Supplementary Table S1**; **Supplementary Figure S2**). After subjecting the crystal complexes to both sampling methods, the trajectories were analyzed based on several structural observables for different kT. First, we calculated the distributions of the χ_1 and χ_2 for each amino acid of the peptide. Examples comparing the results of MD to backrub sampling using kT = 0.35 and kT = 1.2 are shown in **Figure 2** for two amino acids. **Supplementary Table S1** shows how many amino acids were sampled similarly between the backrub and MD configurations using the side chain-dihedral distributions (see Supplementary Text and **Supplementary Figure S3** for details and additional validations). We find that using the backrub simulations with kT = 1.2 is suitable to efficiently explore the side chain dihedrals in comparison to MD.

We then calculated the average of the number contacts and the number of hydrogen bonds created by the main chain atoms (**Table 2**) for the amino acids located in the core region and for both sampling methodologies (using the optimal kT for backrub Rosetta). The fractional error calculated using standard deviation

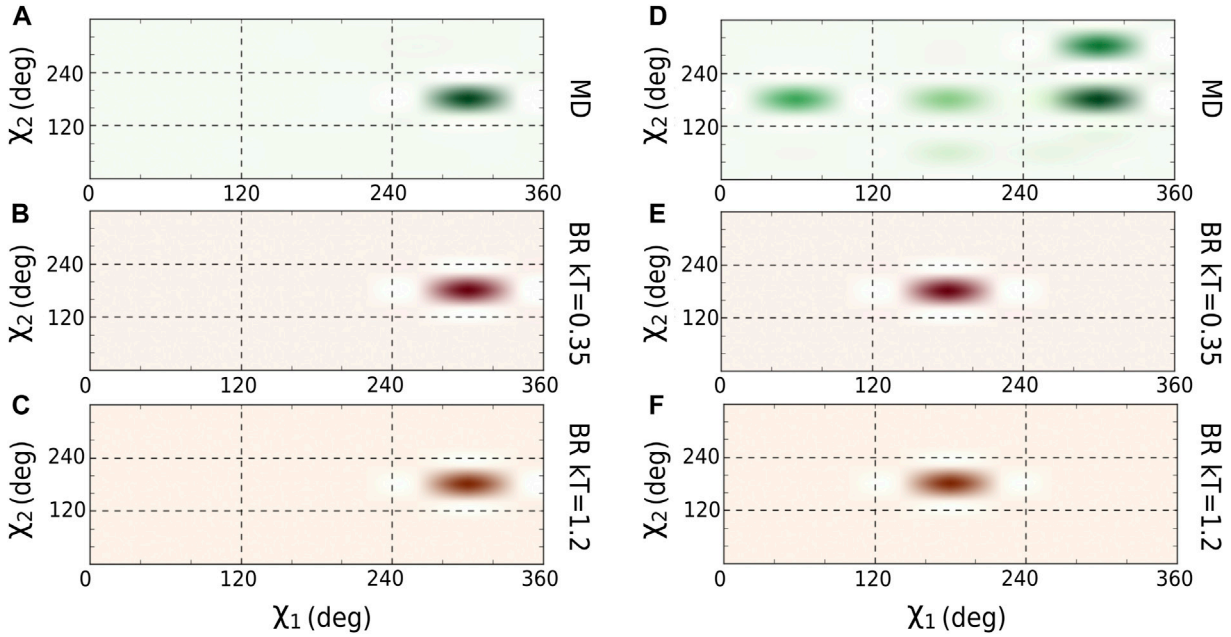


FIGURE 2 | Comparison of χ_1 and χ_2 distributions for amino acid Leu9 from the peptide bound to MHC class II (PKYVKQNTLKLAT PDB id: 1fyt). **(A)** Last 10 ns of MD, **(B)** Backrub using kT = 0.35, and **(C)** Backrub using kT = 1.2. The same analysis was done for amino acid Arg15 of another peptide (AAYSDQATPLLLSPR PDB id 1t5x). **(D)** Last 10 ns of MD, **(E)** Backrub using kT = 0.35, and **(F)** Backrub using kT = 1.2.

TABLE 2 | Average and fractional error of the number of contacts and hydrogen bonds (HB) made by the main chain atoms of the peptide-core amino acids bound to MHC class II, and sampled with MD or backrub (BR) using kT = 1.2. The fractional error was calculated using the standard deviation from the simulations for each peptide core position. The last row shows an average value for all the structures.

PDB	MD contacts	BR contacts	MD HB	BR HB
1fyt	147.9 ± 12.5	148.1 ± 9.8	8.7 ± 1.1	7.4 ± 1.1
1klg	112.6 ± 9.9	104.0 ± 7.8	8.5 ± 1.1	8.3 ± 1.2
1sje	130.3 ± 10.5	104.5 ± 7.3	10.5 ± 0.9	8.6 ± 1.1
1sjh	115.1 ± 10.6	107.8 ± 9.9	8.4 ± 1.1	9.9 ± 1.0
1t5x	135.4 ± 11.8	72.5 ± 10.1	7.8 ± 1.0	3.7 ± 1.0
2fse	126.1 ± 13.0	96.8 ± 8.9	8.9 ± 1.1	6.4 ± 0.8
3pgd	129.9 ± 13.5	133.3 ± 9.4	9.2 ± 1.1	8.7 ± 0.7
4aen	114.6 ± 11.3	99.0 ± 7.6	8.2 ± 1.2	7.4 ± 1.1
4i5b	134.8 ± 10.8	126.5 ± 9.6	8.9 ± 1.2	8.1 ± 1.1
4ov5	161.1 ± 13.8	133.9 ± 11.0	10.2 ± 1.2	8.5 ± 0.7
Average	130.7 ± 11.8	112.6 ± 9.2	8.9 ± 1.1	7.7 ± 1.0

of the simulations is also shown in **Table 2**. We find that the averages for the backrub method are slightly lower than those for MD but within error estimates. Correlations of the values are shown in **Supplementary Figure S4**. The impact of the selected structural descriptors will be discussed in later sections.

Scoring-Matrices From Optimized Backrub Simulations

We selected a total of 837 15-mer peptides from the chosen bioactivity dataset (Nielsen et al., 2010) after filtering, as described in the Methods. The peptides were simulated with backrub using kT = 1.2. The number of hydrogen bonds made by the main chain, and number of non-bonded contacts were calculated from the structure in the trajectories. These observables were averaged and used to calculate the scoring-matrices SM-HB and SM-C per amino acid in the core region according to the equations in the Methods.

These scoring-matrices incorporate the frequency of the structural descriptors obtained from all the sampled peptides, as well as the amino acid distribution of the peptide library. In **Figure 3**, we show the frequency of the amino acid distribution in the set of 837 peptides (**Figure 3A**) and the motif of the peptide core region obtained from the SM-HB as observable (**Figure 3B**). The motif for the SM-C is available in the **Supplementary Figure S5**.

Assessment of the Scoring-Matrices to Predict MHC II- Peptide Activity Differences

We first assessed if the scoring-matrices are able to predict correctly the sign of activity differences by single-point mutations on the peptide core region. The 56 pairs of peptides differing in single amino acids are reported in the **Supplementary Table S2** with the corresponding experimental activities per peptide, and the difference values. This information was obtained from the bioactivity dataset (Wang et al., 2010), which follows experimental gold-standard protocols for binding measurements to MHC receptors, in comparison to other techniques (Kastritis and Bonvin, 2010). We note that this set of peptides was not included during the creation of the scoring-matrices.

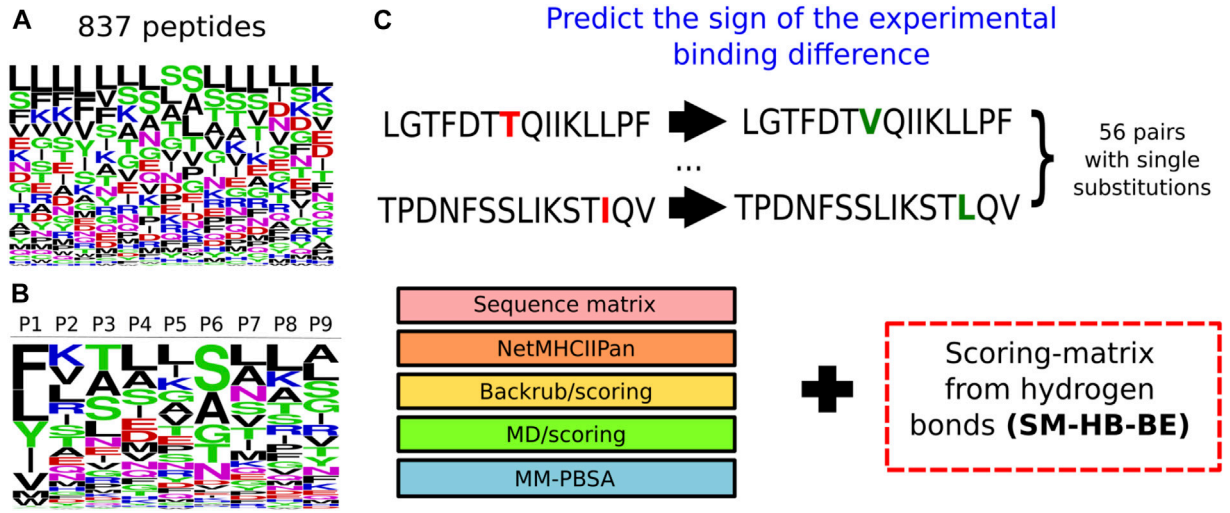


FIGURE 3 | Information used to model peptides bound to the structure of the MHC class II allele DRB1*0101. **(A)** Logo representing the frequency of the amino acids within the 837 15-mer peptides that were modelled bound to the MHC class II structure. The larger the height of the letter the more relevant the amino acid is for improving binding. **(B)** Logo representing the probability of the amino acids at each position of the core region based on the number of hydrogen bonds made by the main chain. The colors represent categories of the amino acids based on physicochemical properties: blue (positive charged), red (negative charged), green (small), fucsia (asparagine) and black (aliphatic). **(C)** Prediction of the sign of the experimental binding differences, for a set of 56 peptides with single substitutions, using the scoring-matrix (SM-HB-BE) in combination with the state-of-the-art methodologies.

TABLE 3 | Prediction of the sign of the experimental activity differences by single-point mutations of the peptide core amino acids using the scoring-matrix calculated based on the hydrogen bonds made by main chain atoms (SM-HB) and the number of non-bonded contacts (SM-C). The comparisons include data for the four strategies to extract information from the 2,000 backrub frames.

Strategy	Matches for SM-HB (%)	Matches for SM-C (%)
All the frames	0.553	0.501
Last half frames	0.518	0.464
Half frames with best energies	0.589	0.501
Best energy frame	0.464	0.518

The prediction results were assessed using the SM-HB and SM-C matrices with a different number and type of frames selected from the backrub trajectories. Specifically, the matrices were obtained using all the frames, the last half of the frames, half frames with best energy-scores and the single best energy frame after optimization (see Methods). A summary of the performances to predict the sign of the activity differences is shown in **Table 3**.

We find that, in general, the observable with the highest number of correct predictions is the SM-HB, in comparison with the SM-C. In particular, for the SM-HB, the best performance was 58.9% using half of the frames with the best predicted energies based on the Rosetta scoring function (henceforth SM-HB-BE with "BE" for best energies). To complement the analysis, we calculated the scoring-matrices six times by dividing the original 837 peptide set into six independent sets. With these matrices, we calculated the mean and standard deviation of the number of matches against the experimental data (see **Supplementary Table S3**). In agreement with the results shown in **Table 3**, we found that the selected SM-HB-BE has the best performance.

Prediction of Activity Differences for Methods That Range in Computational Costs

We compared the best structural scoring matrix (SM-HB-BE), to five previously benchmarked approaches to rank MHC class II peptide binders based on their predicted affinities, or based on the probabilities of finding certain amino acids in the peptide sequence (**Figure 3C**). These methods differ in the theory and, importantly, in their computational cost. In the case of the sequence-based methods, these are able to predict affinities in just a few minutes, but they largely depend on the chemical space of the training data to be successful. The structure/dynamics-based methods range from days to weeks in computational costs. The latter do not rely on training datasets but on physical, chemical and dynamical properties. To assess these diverse

methodologies, we tested them to predict the sign of activity differences by single-point mutations as explained in Methods, and compared their computational cost by running them on an Intel Xeon 24-core server with NVIDIA Titan X GPU acceleration (**Table 4**). In addition, a bootstrapping approach with 50 replicas was ran using randomly, and with repetitions, any pair from the total 56 pairs mutated peptides, in order to obtain a standard deviation of the match for each strategy.

We found that SM-HB-BE has a similar but slightly better match than the main state-of-the-art method (NetMHCIIpan) and the structural MD/scoring and backrub/scoring approaches, but with lower computational times. In the case of MM-PBSA, the results are similar to the backrub/scoring method, but with a computational performance that is 150 times larger than the most efficient sequence-based method (which is inconvenient for large-scale analysis). Based on the results, it is possible to use some of these structural descriptors to pre-select mutations in the core region that could improve the binding affinity requiring low computational costs. We note that the implementation of the scoring-matrices is highly efficient due to its usage as sequence-based descriptors of a particular peptide. The same happens with the sequence-based matrix and the machine learning method. In this sense, using the backrub trajectories to calculate consensus average scores is the most efficient alternative, based on time differences between a few hours to weeks taken by the backrub method and MD simulations (**Table 4**).

We also studied if for certain mutations their activity differences are more difficult to predict. We found that those involving arginine and charged amino acids are more challenging. In addition, amino acids changing drastically in size can misguide the predictions for the majority of the methods. A list of the cases where most of the methods fail is shown in **Supplementary Table S4**. Overall, these results indicate that predicting the activity differences of single-point mutations of peptides bound to MHC class II is challenging, even using extensive calculations such as MM-PBSA.

Combining Structural Scoring-Matrices With Alternative Methodologies Improves the Affinity Difference Prediction

Because there is still room to improve the affinity-difference prediction, we combined the results of each additional method with the SM-HB-BE. The combination consists on checking if either of the two methods predicts a positive mutation, if so then the mutation has a match with the experimental data. This allow us to verify which method complements better with the selected scoring-matrix (**Table 5**). We also included the standard deviations of the matches by following the same bootstrapping approach explained in the previous section.

We found that combining SM-HB-BE with the MD/scoring approach can predict correctly 85.7% of the mutations included in the study, followed by a 78.6% using the backrub/scoring and the MM-PBSA methodologies. As seen, the best results are found after combining the scoring-matrices with structure/dynamic-based strategies, but such combination can be done with the backrub/scoring approach that is more computationally efficient (**Table 4**). In any case, the calculated scoring-matrices can improve information about the frequency of amino acids in core positions using motif representations, and overall the performance is higher than using the sequence-based matrices available in the literature (Rapin et al., 2008).

TABLE 4 | Match values and bootstrapping standard deviations for the prediction of the sign of the experimental activity differences by single-point mutations of the peptide core amino acids for five state-of-the-art methodologies and the SM-HB-BE (i.e., scoring matrix from hydrogen bonds using half of the conformations with best energies). In addition, we include the computational costs, in days, for running the methods with the 56 pairs of mutated peptides. The strategies are the sequence motif matrix, the machine learning tool NetMHCIIpan, the MD/scoring and backrub/scoring approaches, and the MM-PBSA calculations (see Methods).

Complementary strategy	Matched predictions	Computational cost (days)
Sequence matrix	0.393 ± 0.067	0.05
NetMHCIIpan	0.536 ± 0.079	0.1
Backrub/scoring	0.536 ± 0.067	2
MD/scoring	0.571 ± 0.071	15
MM-PBSA	0.518 ± 0.062	15
SM-HB-BE scoring matrix	0.589 ± 0.065	0.05

DISCUSSION

We evaluated the role of structural observables from simulations for predicting activity differences caused by single-point mutation of MHC class II peptide binders. A scoring-matrix derived from counting the number of hydrogen bonds formed by the main chain atoms using the best Rosetta energies (SM-HB-BE), can significantly improve the prediction of these differences if combined with other sequence or simulation-based methodologies.

To deal with the number of modelled peptides, we required running an efficient methodology for sampling the conformations as closely as possible to their equilibrium ensemble. After optimizing the Monte Carlo backrub parameters, we obtained similar conformations to those explored by MD simulations. We note that the MD simulation time was chosen based on previous assessments for exploring well the conformations around the mutated complex, which is around 20 ns for this system (Ochoa et al., 2019; Ochoa et al., 2020). The backrub method tends to perform a similar exploration of the formation of certain contacts and hydrogen bonds, mostly those created by the core region of the peptide. Moreover, the RMSD values between conformations from MD vs backrub are indistinguishable from those of MD vs MD. However, we note that the method is unable to reproduce completely the landscape explored by MD, which can be a limitation. This is why starting from a crystallized bound-conformation is critical for providing more reliable poses of the modelled peptides. Regarding the computational time, the backrub method can sample a similar number of frames as MD in just a few hours, in comparison to days required for MD in high-computing infrastructures (**Table 4**). This facilitates the analysis of a large set of peptides

TABLE 5 | Match values and bootstrapping standard deviations for the prediction of the sign of the experimental activity differences by single point mutations of the peptide core amino acids. The results are for the combination of the additional methodologies with the SM-HB-BE matrix. The strategies are the sequence motif matrix, the machine learning tool NetMHCIIpan, the MD/scoring and backrub/scoring approaches, and the MM-PBSA calculations (see Methods).

Complementary strategy	Matched predictions in combination with the SM-HB-BE matrix
Sequence matrix	0.607 ± 0.062
NetMHCIIpan	0.714 ± 0.064
Backrub/scoring	0.786 ± 0.051
MD/scoring	0.857 ± 0.047
MM-PBSA	0.786 ± 0.048

for this MHC class II allele, and others with structures available in public databases.

The peptide were selected based on criteria that facilitate the initial modelling of the rotamers (Ochoa et al., 2018), and the inclusion, in some cases, of additional flanking amino acids. Moreover, these peptides have available experimental binding data. Therefore, the new descriptors contain intrinsic information about the distribution of amino acids based on binding information, implying that our structural insights are complementing the known sequence-based motifs (Menconi et al., 2008; Andreatta et al., 2012). This is relevant because our protocol does not start from scratch. Instead, its main goal is to exploit the current knowledge of the system, and provide better metrics for the understanding of the MHC class II binding using simulations.

The calculated observables can be compared to the reported MHC class II promiscuity in terms of the intrinsic stability of the interactions between the peptide and the MHC class II binding groove. We found, for example, that the hydrogen bonds created by the main chain atoms are one of the most important structural observables. This claim has also been proposed in other studies, motivated by the stability of the peptide-bound conformation in spite of being completely linear, which is crucial in the molecular editing processes within the antigen presentation pathways (Painter et al., 2008; Yaneva et al., 2010; Ferrante et al., 2015). Therefore, simulating the dynamics of the complex can bring novel insights into the binding nature, and allows us to predict activity differences caused by single-point substitutions on the peptide sequence.

CONCLUSION

Simulations provide structural insights for creating simple scoring-matrices that complement available methods to better predict the effect of single-mutations on the binding of peptides to MHC class II molecules. Integrating sequence, structural and dynamical information is useful to progress in the immunoinformatics field, not only for MHC class II structures, but also for other key components within the immune response pathways.

Moreover, the methodology can contribute to the identification of epitopes for certain alleles using structural and dynamical information. In fact, the method can be expanded to calculate descriptors for other peptide-binding complexes, to design of novel epitopes by single-point substitutions, and to understand the impact of antigen mutations in the immune system, for example, using structural interactions with the *T*-cell receptors, having direct consequence in vaccine design (Purcell et al., 2007). The descriptors can also help to, possibly, discriminate at a reasonable level between good binders and non-binders. However, a better discrimination requires combining multiple methods, or implement more exhaustive approaches to capture the chemical contributions of the peptide residues through more explicit free energy calculations (Wieczorek et al., 2016; Huang et al., 2017).

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/rochoa85/MHC_class_II_matrices, Github.

AUTHOR CONTRIBUTIONS

RO designed and ran the modelling and simulation computational protocols, created the code, and wrote the manuscript. RL tested the scoring matrices and the code, provided feedback and wrote the manuscript. JT provided feedback and wrote the manuscript. PC supported the computational analysis, provided feedback and wrote the manuscript.

FUNDING

This work has been supported by Colciencias, University of Antioquia and Ruta N, Colombia, the Max Planck Society, Germany and the CABANA initiative, United Kingdom.

ACKNOWLEDGMENTS

The authors thank AL for his advice in the methodology and result analysis. RO thanks the CABANA initiative for funding the secondment project during 2019. RO. and PC. were also supported by Colciencias, University of Antioquia, Ruta N,

Colombia and the Max Planck Society, Germany. The computations were performed on the EMBL-EBI cluster and on a local server with an NVIDIA Titan X GPU. PC gratefully acknowledges the support of the NVIDIA Corporation for the donation of this GPU.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.636562/full#supplementary-material.

REFERENCES

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., et al. (2017). The rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theor. Comput.* 13, 3031–3048. doi:10.1021/acs.jctc.7b00125

Andreatta, M., Karosiene, E., Rasmussen, M., Stryhn, A., Buus, S., and Nielsen, M. (2015). Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics* 67, 641–650. doi:10.1007/s00251-015-0873-y

Andreatta, M., Lund, O., and Nielsen, M. (2012). Simultaneous alignment and clustering of peptide data using a gibbs sampling approach. *Bioinformatics* 29, 8–14. doi:10.1093/bioinformatics/bts621

Andrusier, N., Nussinov, R., and Wolfson, H. J. (2007). FireDock: fast interaction refinement in molecular docking. *Proteins* 69, 139–159. doi:10.1002/prot.21495

Antunes, D. A., Devaurs, D., Moll, M., Lizée, G., and Kavraki, L. E. (2018). General prediction of peptide-MHC binding modes using incremental docking: a proof of concept. *Sci. Rep.* 8, 4327. doi:10.1038/s41598-018-22173-4

Aranha, M. P., Jewel, Y. S. M., Beckman, R. A., Weiner, L. M., Mitchell, J. C., Parks, J. M., et al. (2020). Combining three-dimensional modeling with artificial intelligence to increase specificity and precision in peptide–mhc binding predictions. *J.Immunol.* 205, 1962–1977. doi:10.4049/jimmunol.1900918

Barlow, K. A., Ó Conchúir, S., Thompson, S., Suresh, P., Lucas, J. E., Heinonen, M., et al. (2018). Flex ddG: rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation. *J. Phys. Chem. B* 122, 5389–5399. doi:10.1021/acs.jpcb.7b11367

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Bermúdez, A., Calderon, D., Moreno-Vranich, A., Almonacid, H., Patarroyo, M. A., Poloche, A., et al. (2014). Gauche+ side-chain orientation as a key factor in the search for an immunogenic peptide mixture leading to a complete fully protective vaccine. *Vaccine* 32, 2117–2126. doi:10.1016/j.vaccine.2014.02.003

Berrera, M., Molinari, H., and Fogolari, F. (2003). Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinform.* 4, 8. doi:10.1186/1471-2105-4-8

Bjorkman, P. J. (2015). Not second class: the first class II MHC crystal structure. *J. Immunol.* 194, 3–4. doi:10.4049/jimmunol.1402828

Bogan, A. A., and Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* 280, 1–9. doi:10.1006/jmbi.1998.1843

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi:10.1093/bioinformatics/btp163

Cossio, P., Granata, D., Laio, A., Seno, F., and Trovato, A. (2012). A simple and efficient statistical potential for scoring ensembles of protein structures. *Sci. Rep.* 2, 1–8. doi:10.1038/srep00351

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

Davis, I. W., Arendall, W. B., Richardson, D. C., and Richardson, J. S. (2006). The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure* 14, 265–274. doi:10.1016/j.str.2005.10.007

Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. (2003). HADDOCK: a protein protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* 125, 1731–1737. doi:10.1021/ja026939x

Ferrante, A., Templeton, M., Hoffman, M., and Castellini, M. J. (2015). The thermodynamic mechanism of peptide–MHC class II complex formation is a determinant of susceptibility to HLA-DM. *J. Immunol.* 195, 1251–1261. doi:10.4049/jimmunol.1402367

Fogolari, F., Corazza, A., Yarra, V., Jalaru, A., Viglino, P., and Esposito, G. (2012). Bluues: a program for the analysis of the electrostatic properties of proteins based on generalized born radii. *BMC Bioinformatics* 13, S18. doi:10.1186/1471-2105-13-S4-S18

Hershey, J. R., and Olsen, P. A. (2007). Approximating the kullback leibler divergence between gaussian mixture models. *IEEE Int. Conf. Acoust. Speech Signal Process.* 7, 317–320. doi:10.1109/ICASSP.2007.366913

Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). Gromacs 4: algorithms for highly efficient, load balanced, and scalable molecular simulations. *J. Chem. Theor. Comput.* 4, 435–447. doi:10.1021/ct700301q

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., et al. (2017). Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35, 128–135. doi:10.1038/nbt.3769

Huang, M., Huang, W., Wen, F., and Larson, R. G. (2017). Efficient estimation of binding free energies between peptides and an MHC class II molecule using coarse-grained molecular dynamics simulations with a weighted histogram analysis method. *J. Comput. Chem.* 38, 2007–2019. doi:10.1002/jcc.24845

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

Kastritis, P. L., and Bonvin, A. M. J. J. (2010). Are scoring functions in protein protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.* 9, 2216–2225. doi:10.1021/pr9009854

King, C. A., and Bradley, P. (2010). Structure-based prediction of protein-peptide specificity in rosetta. *Proteins* 78, 3437–3449. doi:10.1002/prot.22851

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

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

Kumari, R., Kumar, R., Lynn, A., and Lynn, A. (2014). g_mmpbsa: a gromacs tool for high-throughput mm-pbsa calculations. *J. Chem. Inf. Model.* 54, 1951–1962. doi:10.1021/ci500020m

Lanzarotti, E., Marcatili, P., and Nielsen, M. (2018). Identification of the cognate peptide-MHC target of T cell receptors using molecular modeling and force field scoring. *Mol. Immunol.* 94, 91–97. doi:10.1016/j.molimm.2017.12.019

Li, M., Petukh, M., Alexov, E., and Panchenko, A. R. (2014). Predicting the impact of missense mutations on protein–protein binding affinity. *J. Chem. Theor. Comput.* 10, 1770–1780. doi:10.1021/ct401022c

Löffler, P., Schmitz, S., Hupfeld, E., Sterner, R., Merkl, R., and Hughes, M. (2017). Rosetta:MSF: a modular framework for multi-state computational protein design. *PLoS Comput. Biol.* 13, e1005600. doi:10.1371/journal.pcbi.1005600

McDonald, I. K., and Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* 238, 777–793. doi:10.1006/jmbi.1994.1334

Menconi, F., Monti, M. C., Greenberg, D. A., Oashi, T., Osman, R., Davies, T. F., et al. (2008). Molecular amino acid signatures in the mhc class ii peptide-binding pocket predispose to autoimmune thyroiditis in humans and in mice. *Proc. Natl. Acad. Sci.* 105, 14034–14039. doi:10.1073/pnas.0806584105

Nandy, A., and Basak, S. (2016). A brief review of computer-assisted approaches to rational design of peptide vaccines. *Int. J. Mol. Sci.* 17, 666. doi:10.3390/ijms17050666

Nielsen, M., Lund, O., Buus, S., and Lundegaard, C. (2010). MHC Class II epitope predictive algorithms. *Immunology* 130, 319–328. doi:10.1111/j.1365-2567.2010.03268.x

Ochoa, R., Soler, M., Laio, A., and Cossio, P. (2020). Parce: protocol for amino acid refinement through computational evolution. *Comput. Phys. Commun.* 260, 107716. doi:10.1016/j.cpc.2020.107716

Ochoa, R., Laio, A., and Cossio, P. (2019). Predicting the affinity of peptides to major histocompatibility complex class II by scoring molecular dynamics simulations. *J. Chem. Inf. Model.* 59, 3464–3473. doi:10.1021/acs.jcim.9b00403

Ochoa, R., Soler, M. A., Laio, A., and Cossio, P. (2018). Assessing the capability of in silico mutation protocols for predicting the finite temperature conformation of amino acids. *Phys. Chem. Chem. Phys.* 20, 25901–25909. doi:10.1039/C8CP03826K

Omasits, U., Knapp, B., Neumann, M., Steinhauser, O., Stockinger, H., Kobler, R., et al. (2008). Analysis of key parameters for molecular dynamics of pMHC molecules. *Mol. Simul.* 34, 781–793. doi:10.1080/08927020802256298

Painter, C. A., Cruz, A., López, G. E., Stern, L. J., and Zavala-Ruiz, Z. (2008). Model for the peptide-free conformation of class II MHC proteins. *PLoS One* 3, e2403. doi:10.1371/journal.pone.0002403

Peters, B., Nielsen, M., and Sette, A. (2020). T cell epitope predictions. *Annu. Rev. Immunol.* 38, 123–145. doi:10.1146/annurev-immunol-082119-124838

Purcell, A. W., McCluskey, J., and Rossjohn, J. (2007). More than one reason to rethink the use of peptides in vaccine design. *Nat. Rev. Drug Discov.* 6, 404–414. doi:10.1038/nrd2224

Rapin, N., Hoof, I., Lund, O., and Nielsen, M. (2008). MHC motif viewer. *Immunogenetics* 60, 759–765. doi:10.1007/s00251-008-0330-2

Sammond, D. W., Eletr, Z. M., Purbeck, C., Kimple, R. J., Siderovski, D. P., and Kuhlman, B. (2007). Structure-based protocol for identifying mutations that enhance protein-protein binding affinities. *J. Mol. Biol.* 371, 1392–1404. doi:10.1016/j.jmb.2007.05.096

Sarti, E., Zamuner, S., Cossio, P., Laio, A., Seno, F., and Trovato, A. (2013). Bachscore. a tool for evaluating efficiently and reliably the quality of large sets of protein structures. *Comput. Phys. Commun.* 184, 2860–2865. doi:10.1016/j.cpc.2013.07.019

Slutzki, M., Reshef, D., Barak, Y., Haimovitz, R., and Rotem-Bamberger, S. (2015). Crucial roles of single residues in binding affinity, specificity, and promiscuity in the cellulosomal cohesin-dockerin interface. *J. Biol. Chem.* 290, 13654–13666. doi:10.1074/jbc.M115.651208

Smith, C. A., and Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.* 380, 742–756. doi:10.1016/j.jmb.2008.05.023

Smith, C. A., and Kortemme, T. (2010). Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *J. Mol. Biol.* 402, 460–474. doi:10.1016/j.jmb.2010.07.032

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., and Tawfik, D. S. (2007). The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* 369, 1318–1332. doi:10.1016/j.jmb.2007.03.069

Trott, O., and Olson, A. J. (2009). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461. doi:10.1002/jcc.21334

Unanue, E. R., Turk, V., and Neefjes, J. (2016). Variations in MHC class II antigen processing and presentation in health and disease. *Annu. Rev. Immunol.* 34, 265–297. doi:10.1146/annurev-immunol-041015-055420

Wang, P., Sidney, J., Dow, C., Mothé, B., Sette, A., and Peters, B. (2008). A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput. Biol.* 4, e1000048, doi:10.1371/journal.pcbi.1000048

Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M., et al. (2010). Peptide binding predictions for HLA DR, DP, and DQ molecules. *BMC Bioinform.* 11, 568. doi:10.1186/1471-2105-11-568

Wieczorek, M., Abualrous, E. T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., et al. (2017). Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Front. Immunol.* 8, 1–16. doi:10.3389/fimmu.2017.00292

Wieczorek, M., Sticht, J., Stolzenberg, S., Günther, S., Wehmeyer, C., El Habre, Z., et al. (2016). MHC Class II complexes sample intermediate states along the peptide exchange pathway. *Nat. Commun.* 7, 13224. doi:10.1038/ncomms13224

Yaneva, R., Schneeweiss, C., Zacharias, M., and Springer, S. (2010). Peptide binding to MHC class i and ii proteins: new avenues from new methods. *Mol. Immunol.* 47, 649–657. doi:10.1016/j.molimm.2009.10.008

Yang, Y., and Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins* 72, 793–803. doi:10.1002/prot.21968

Yeturu, K., Utriainen, T., Kemp, G. J., and Chandra, N. An automated framework for understanding structural variations in the binding grooves of MHC class II molecules. *BMC Bioinform.* 11 (2010) S55. doi:10.1186/1471-2105-11-S1-S55

Zhang, H., Wang, P., Papangelopoulos, N., Xu, Y., Sette, A., Bourne, P. E., et al. (2010). Limitations of ab initio predictions of peptide binding to MHC class II molecules. *PLoS One* 5, e9272. doi:10.1371/journal.pone.0009272

Zhou, H., and Skolnick, J. (2011). GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* 101, 2043–2052. doi:10.1016/j.bpj.2011.09.012

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

OPINION
published: 05 April 2021
doi: 10.3389/fmolb.2021.641208

Molecular Dynamics to Predict Cryo-EM: Capturing Transitions and Short-Lived Conformational States of Biomolecules

Łukasz Nierzwicki[1] *and Giulia Palermo*[1,2]*

[1]Department of Bioengineering, University of California, Riverside, CA, United States, [2]Department of Chemistry, University of California, Riverside, CA, United States

Single-particle cryogenic electron microscopy (cryo-EM) has revolutionized the field of the structural biology, providing an access to the atomic resolution structures of large biomolecular complexes in their near-native environment. Today's cryo-EM maps can frequently reach the atomic-level resolution, while often containing a range of resolutions, with conformationally variable regions obtained at 6 Å or worse. Low resolution density maps obtained for protein flexible domains, as well as the ensemble of coexisting conformational states arising from cryo-EM, poses new challenges and opportunities for Molecular Dynamics (MD) simulations. With the ability to describe the biomolecular dynamics at the atomic level, MD can extend the capabilities of cryo-EM, capturing the conformational variability and predicting biologically relevant short-lived conformational states. Here, we report about the state-of-the-art MD procedures that are currently used to refine, reconstruct and interpret cryo-EM maps. We show the capability of MD to predict short-lived conformational states, finding remarkable confirmation by cryo-EM structures subsequently solved. This has been the case of the CRISPR-Cas9 genome editing machinery, whose catalytically active structure has been predicted through both long-time scale MD and enhanced sampling techniques 2 years earlier than cryo-EM. In summary, this contribution remarks the ability of MD to complement cryo-EM, describing conformational landscapes and relating structural transitions to function, ultimately discerning relevant short-lived conformational states and providing mechanistic knowledge of biological function.

Keywords: molecular dynamics, enhanced sampling, cryo-EM, CRISPR-Cas9, structure prediction

OPEN ACCESS

Edited by:
Massimiliano Bonomi, Institut Pasteur, France

Reviewed by:
Slavica Jonic, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), France
Riccardo Pellarin, Institut Pasteur, France

***Correspondence:**
Giulia Palermo
giulia.palermo@ucr.edu

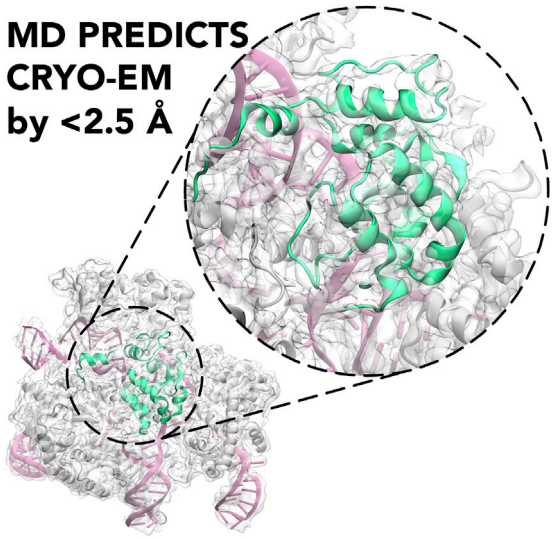
Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *13 December 2020*
Accepted: *15 February 2021*
Published: *05 April 2021*

Citation:
Nierzwicki Ł and Palermo G (2021) Molecular Dynamics to Predict Cryo-EM: Capturing Transitions and Short-Lived Conformational States of Biomolecules. Front. Mol. Biosci. 8:641208. doi: 10.3389/fmolb.2021.641208

STATE-OF-THE-ART CRYO-EM MODELLING THROUGH MOLECULAR DYNAMICS

Single-particle cryogenic electron microscopy (cryo-EM) has revolutionized the field of structural biology, providing an access to the atomic resolution structures of large biomolecular complexes in their near-native environment (Nogales, 2015). The number of macromolecular structures determined by cryo-EM is rapidly increasing, indeed, it is predicted that by 2024 the number of yearly released structures will be higher for cryo-EM than for X-ray crystallography (Callaway, 2020). The cryo-EM technique comprises of three consecutive steps. At first, the sample is frozen over millisecond time scales, what results in both the formation of amorphous ice and in capturing the



GRAPHICAL ABSTRACT | Molecular Dynamics (MD) is shown to predict the cryo-EM structure of the active CRISPR-Cas9 system with an RMSD between the cryo-EM structure and the MD ensemble of <2.5 Å.

biomacromolecule in its near-native conformation through quick undercooling of the sample. The term "near-native" refers to the fact that during cryofixation, limited conformational transitions can result in some non-native conformations within the structural ensemble. Given the timescale of cryofixation (i.e., milliseconds), these transitions should be limited. Next, a number of two-dimensional (2D) electron microscopy (EM) images of the biomacromolecule are collected and, finally, these 2D images are combined into a three-dimensional electrostatic potential map of the biomacromolecule(Guo and Jiang, 2014; Kontziampasis et al., 2019; Cianfrocco and Kellogg, 2020). Today's cryo-EM maps can frequently reach the atomic-level resolution, while often containing a range of resolutions, with conformationally variable regions obtained at 6 Å or worse. The latter can also arise from several other factors, such as radiation damage and image alignment errors. Moreover, considering also that the atomic form factors of cryo-EM maps represent the atomic electrostatic potential, negatively charged moieties might be depleted or not visible, as they scatter electrons more efficiently (Marques et al., 2019). Recent advances in post-processing cryo-EM images also allowed to identify multiple conformational states of the biological complexes (Jin et al., 2019) or even to describe the conformational variability of their single subunits (Bai et al., 2015). These advancements and opportunities introduced by single-particle cryo-EM are paving the way for an explosion of computational methods aimed at processing, refining and interpreting cryo-EM data (Dodd et al., 2020; Fraser et al., 2020; Kim et al., 2020; Palermo et al., 2020).

Molecular dynamics (MD) simulations are known to be powerful in describing in detail the intrinsic dynamics of biomolecules and the energetics that underlie conformational transitions (Karplus and McCammon, 2002). This is why MD simulations are an excellent tool to examine hypotheses posed by the experimental findings of cryo-EM studies. It is also apparent that both techniques can mutually benefit from cooperation, where MD can unveil the atomic details of conformational changes and refine the structure for low resolution regions of cryo-EM maps (Kirmizialtin et al., 2015), while cryo-EM can not only provide the structure of biomolecules (Nogales, 2015), but also describe its near-native conformational ensemble in solution (Jin et al., 2019).

The initial approaches combining MD and cryo-EM methods used MD as a fitting scheme to predict the structure of a biomolecule, using the low-resolution EM map to constrain the protein conformation. For this purposes, two commonly used packages are the MD Flexible Fitting (MDFF) (Trabuco et al., 2008) and the Situs (Kovacs et al., 2018) codes, where the first one guides MD simulation toward the cryo-EM density biasing the MD potential energy form to reduce the gradient of the experimental electronic density, while the second one minimizes the discrepancy between the map derived from the MD model and the original cryo-EM map. Hybrid approaches harnessing docking algorithms have also been developed, such as including a rigid fitting stage followed by a refinement based on MD (Topf et al., 2008), or introducing a coarse-grained force field to allow flexibility during the docking search (de Vries and Zacharias ATTRACT-, 2012). MD-based methods were shown to successfully refine the structure of both isolated proteins (e.g., lactoferrin) and large protein assemblies (up to ribosomes) (Trabuco et al., 2008). Unfortunately, one of the prominent challenges for these methods is structure overfitting to the cryo-EM map, where the derived potential can lead to unphysical conformations of the biomolecule (Trabuco et al., 2009). However, such inconveniences can be overcame by combining a series of restraints derived from the experimental density with enhanced sampling MD techniques, as shown for membrane transporter Escherichia coli efflux-multidrug resistance E (EmrE) (Ovchinnikov et al., 2018). In that study, map-restrained Self-guided Langevin dynamics (Wu et al., 2013) was used with a series of heating and cooling cycles of the EmrE protein during MD run. Such approach allowed to relax both the conformation of the protein backbone and side chains and eventually led to a substantial improvement of the MD structure with respect to cryo-EM map. Enhanced sampling simulations in the structure refinement are also used in more advanced MDFF schemes, namely Cascade MDFF and Resolution Exchange MDFF (Singharoy et al., 2016). The former approach is based on simulated annealing (Brünger, 1988), where the structure is fitted sequentially to maps with higher resolution. In the latter, the Hamiltonian replica-exchange simulations (Sugita et al., 2000) are used, where in each replica the potential affecting the system is derived from the flexible fitting to the projections of the cryo-EM maps that change from low to high resolution. In this way the system is allowed to relax conformationally in low resolution replicas, while the conformations that are both relaxed in the force field and fit well to the cryo-EM maps are preferred to exchange into the high-resolution replicas. Multiple replicas were also used in

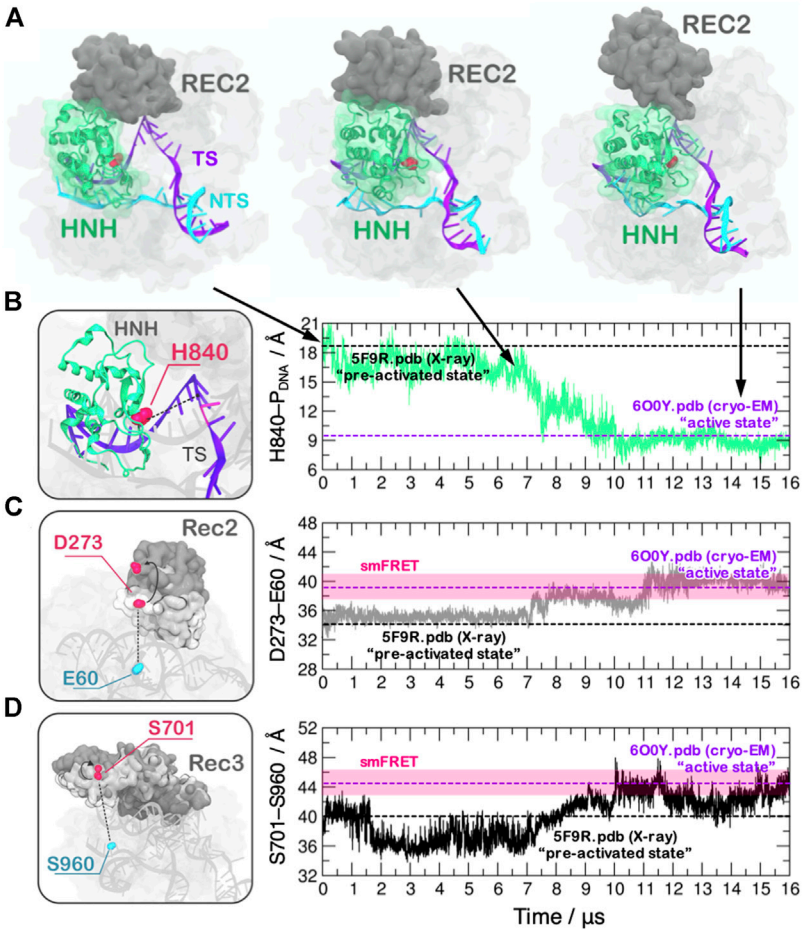


FIGURE 1 | (A) Conformational activation of the HNH domain and structural adaptation of the REC domain during ~16 μs of continuous MD simulations performed on the Anton-2 supercomputer (Palermo et al., 2018). **(B–D)** Time evolution of the distances: **(B)** between H840 and the cleavage site, indicating the docking of HNH at the DNA target strand; **(C)** between E60 and D273 and **(D)** between S960 and S701, indicating the opening of the REC2 and REC3 domains. Horizontal bars are used to indicate the value of the three distances in the X-ray structure of the pre-activated state (PDBid: 5F9R at 3.40 Å resolution (Jiang et al., 2016), starting configuration for MD) and in the structure obtained via cryo-EM (PDBid: 6O0Y at 3.37 Å resolution) (Zhu et al., 2019). Transparent bars indicate the distance range assumed obtained through single molecule Förster Resonance Energy Transfer experiments. Reprinted with permission from Palermo et al. (2018). Copyright 2018 Cambridge University Press. https://doi.org/10.1017/S0033583518000070.

a metainference method, where the restraining force arising from the difference between MD structures and the cryo-EM map is generated in an ensemble-averaged manner(Bonomi et al., 2018; Eshun-Wilson et al., 2019). Such an approach has already been shown to be fruitful in the case of NMR restraints, where the average chemical shifts or coupling constants were not necessarily representative of an heterogenous conformational ensemble present in solution (Camilloni et al., 2012). In the context of cryo-EM, this allows exploring the relevant heterogenous regions of the free energy landscape, while still remaining in agreement with the cryo-EM findings. The most recent approach, implemented in Gromacs 2020 (Igaev et al., 2019), uses a gradient of similarity between a density obtained from MD structure and the experimental density to compute the forces. This approach allows to use a variety of similarity measurements (inner product, relative entropy or cross-correlation the of the densities), enabling to adjust the density-based restraining method. Hence, one can restrain the system without enforcing the trajectory (which could lead to unphysical conformations), which helps reducing the impact of experimental artifacts (Marques et al., 2019) on the conformational dynamics of the simulated biomolecule. The method has been successfully used to unveil the origins of the SARS-CoV-2 spike protein flexibility, allowing to identify the three flexible hinges within the protein (Turoňová et al., 2020). Overall, these examples show how MD simulations guided by cryo-EM data allow for both the structure refinement the interpretation the experimental maps.

Post-processing of MD trajectories to compare the obtained structures with original cryo-EM maps can also be obtained through a variety of visualisation tools, such as e.g., Chimera (Pettersen et al., 2004) that allows for the fitting of experimental and MD derived density maps, also providing a measure for the

fitting quality between densities. The recently released GROmaps tool (Briones et al., 2019) allows to compute the time-averaged MD density map and does expand a set of tools to compare the computed map with the original cryo-EM results. This method in principle can be combined with augmented Markov models (Olsson et al., 2017), where the cryo-EM map could be used as an experimental observable to reweight the simulation ensembles. Such approach increases the credibility of the comparison between cryo-EM maps and MD outcomes without biasing the simulation runs.

CAPTURING TRANSITIONS AND SHORT-LIVED CONFORMATIONAL STATES

MD can also aid cryo-EM experiments by predicting the structure of short-lived conformational states that are both essential for the biomolecular complexes activity and are hard to capture with cryo-EM because of their transient nature. A prominent example is the prediction of the active conformation of the CRISPR-Cas9 (clustered regularly interspaced short palindromic repeat and associated Cas9 proteins) system, which recently emerged as a forefront tool for genome editing (Doudna and Charpentier, 2014). At the molecular level, CRISPR-Cas9 is a large ribonucleoprotein complex, which uses RNA-guided Cas9 endonuclease to recognize and cleave matching sequences of DNA. Biophysical studies have indicated that the catalytic HNH domain is characterized by a "striking plasticity," (Jiang et al., 2016; Palermo et al., 2016), which governs the enzymatic function. This high flexibility, however, initially hampered a definitive characterization of the catalytically competent state through cryo-EM and X-ray crystallography. Early attempts to define the structure of the catalytically active CRISPR-Cas9 employed extensive MD simulations (Palermo et al., 2017; Zuo and Liu, 2017; Palermo et al., 2018). The first effort to determine the structural transitions leading to the active state have been performed using the Gaussian accelerated MD (GaMD) method (Wang et al., 2021) that enables unconstrained enhanced sampling capturing displacements over micro- (μs) to millisecond (ms) timescales, which is of difficult reach through conventional MD. This approach described the activated state (Palermo et al., 2017). Building on this initial study, the Anton-2 supercomputer has been employed to perform unbiased runs of the complex and to determine the continuous dynamics of HNH over multiple μs (Palermo et al., 2018). This characterized the dynamical docking of HNH at the cleavage site, predicting an active conformation that confirmed the initial model obtained through GaMD (**Figure 1**).

This theoretical structure enabled to initiate in-depth studies of the catalysis (Palermo, 2019; Casalino et al., 2020), the allostery (Palermo et al., 2017; East et al., 2020; Nierzwicki et al., 2020) and the system's specificity (Mitchell et al., 2020; Ricci et al., 2019), when no structural information on the active state was available. This helped obtaining information to improve the enzyme catalytic efficiency and to reduce off-target effects, which is a key goal for biomedical applications (Fu et al., 2013). The experimental determination of the catalytically competent state through cryo-EM occurred 2 years after the theoretical model (Zhu et al., 2019), reporting a remarkable agreement with the predicted model (the average RMSD between the cryo-EM structure and the MD ensemble of 2.47 ± 0.14 Å, computed considering the HNH domain and the six nucleotides at the cleavage site). Molecular simulations using Anton-2 further indicated that the recognition regions (REC) of the Cas9 protein would undergo a remarkable opening to allow the process of HNH activation (**Figure 1**), noting also concerted dynamics of the REC-HNH domains (Palermo et al., 2018). These coordinated domain motions were also observed through cryo-EM, revealing their functional role for DNA cleavage (Zhu et al., 2019). Furthermore, a recent single-molecule study probing the conformational dynamics of Cas9 in the post-catalytic state highlighted rapid conformational fluctuations of HNH (Wang et al., 2021), as observed through MD. These results highlight the consistency of the simulations with experimental observations and suggest that state-of-the-art MD can capture short-lived conformational states of biomolecules, which are of difficult reach through structural biophysics techniques.

SUMMARY AND PERSPECTIVES

Here, we highlighted how MD simulations combined with cryo-EM data can provide a deep understanding of key conformational steps that govern the function of biomacromolecules. MD can be used not only to refine cryo-EM structures, especially the low-resolution regions, but also to facilitate interpretation of the experimental findings. Novel MD analysis tools allow also to compute the time-averaged cryo-EM maps from MD trajectories, enabling a reasonable comparison between conformational ensembles determined experimentally and computationally. This overcomes the limitations of comparing single structures, lacking of dynamical information. Finally, MD simulations alone were also shown to be a powerful predicting tool, that allows to characterize the short-lived conformational states of biomolecules hard to capture through cryo-EM.

Ultimately, the rapid development of methods that combine cryo-EM data with MD will further increase the reliability of MD-guided predictions. One can expect that the rigorous comparison between cryo-EM and MD conformational ensembles can be an additional source of the data that can be used to improve the currently available simulation methods. Molecular simulations can also be guided to a conformational ensemble defined as a cryo-EM map rather than a specific structure. This can improve the description of the free energy landscape associated with conformational changes of proteins and nucleic acids, as the cryo-EM map can be used as a reference for the conformational ensemble. Such approach, based on Multi-Map variable method, was very recently released for NAMD (Vant et al., 2020). The initial results for both the steered-MD simulations and free energy methods are encouraging, with the free energy profiles for the conformational transitions comparable to those determined using high-resolution structures as a reference. Overall, non-stop development of cryo-EM–based MD

methods opens novel opportunities for the precise description of biomolecular dynamics.

AUTHOR CONTRIBUTIONS

LN wrote the manuscript. GP conceived research and wrote the manuscript.

FUNDING

This material is based upon work supported by the National Science Foundation under Grant No. CHE-1905374, awarded to G.P. This work was also partially funded by the National Institute of Health through the Grant R01GM141329. This work used the Anton-2 computer that was provided by the Pittsburgh Supercomputing Center through grant PSCA16035P from the NIH. The Anton-2 machine at PSC was generously made available by D.E. Shaw Research. This work also used supercomputing resources with allocation award TG- MCB160059 (to G.P.) through the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Computational resources were also made available by the project M3807 through the National Energy Research Scientific Computing Center (NERSC, to GP), which is a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

REFERENCES

Bai, X. C., Rajendra, E., Yang, G., Shi, Y., and Scheres, S. H. (2015). Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife* 4, e11182. doi:10.7554/eLife.11182

Bonomi, M., Pellarin, R., and Vendruscolo, M. (2018). Simultaneous determination of protein structure and dynamics using cryo-electron microscopy. *Biophys. J.* 114, 1604–1613. doi:10.1016/j.bpj.2018.02.028

Briones, R., Blau, C., Kutzner, C., de Groot, B. L., Aponte-Santamaría, C., and GROmaps, C. (2019). GROmaps: a GROMACS-based toolset to analyze density maps derived from molecular dynamics simulations. *Biophys. J.* 116, 4–11. doi:10.1016/j.bpj.2018.11.3126

Brünger, A. T. (1988). Crystallographic refinement by simulated annealing. Application to a 2.8 A resolution structure of aspartate aminotransferase. *J. Mol. Biol.* 203, 803–816. doi:10.1016/0022-2836(88)90211-2

Callaway, E. (2020). Revolutionary cryo-EM is taking over structural biology. *Nature* 578, 201. doi:10.1038/d41586-020-00341-9

Camilloni, C., Robustelli, P., De Simone, A., Cavalli, A., and Vendruscolo, M. (2012). Characterization of the conformational equilibrium between the two major substates of RNase a using NMR chemical shifts. *J. Am. Chem. Soc.* 134, 3968–3971. doi:10.1021/ja210951z

Casalino, L., Nierzwicki, Ł., Jinek, M., and Palermo, G. (2020). Catalytic mechanism of non-target DNA cleavage in CRISPR-Cas9 revealed by ab initio molecular dynamics. *ACS Catal.* 10, 13596–13605. doi:10.1021/acscatal.0c03566

Cianfrocco, M. A., and Kellogg, E. H. (2020). What could go wrong? A practical guide to single-particle cryo-EM: from biochemistry to atomic models. *J. Chem. Inf. Model.* 60, 2458–2469. doi:10.1021/acs.jcim.9b01178

de Vries, S. J., and ZachariasATTRACT-, M. E. M. (2012). ATTRACT-EM: a new method for the computational assembly of large molecular machines using cryo-EM maps. *PLoS One* 7, e49733. doi:10.1371/journal.pone.0049733

Dodd, T., Yan, C., and Ivanov, I. (2020). Simulation-based methods for model building and refinement in cryoelectron microscopy. *J. Chem. Inf. Model.* 60, 2470–2483. doi:10.1021/acs.jcim.0c00087

Doudna, J. A., and Charpentier, E. (2014). Genome editing. The new Frontier of genome engineering with CRISPR-Cas9. *Science* 346, 1258096. doi:10.1126/science.1258096

East, K. W., Newton, J. C., Morzan, U. N., Narkhede, Y. B., Acharya, A., Skeens, E., et al. (2020). Allosteric motions of the CRISPR-Cas9 HNH nuclease probed by NMR and molecular dynamics. *J. Am. Chem. Soc.* 142, 1348–1358. doi:10.1021/jacs.9b10521

Eshun-Wilson, L., Zhang, R., Portran, D., Nachury, M. V., Toso, D. B., Löhr, T., et al. (2019). Effects of α-tubulin acetylation on microtubule structure and stability. *Proc. Natl. Acad. Sci. USA* 116, 10366–10371. doi:10.1073/pnas.1900441116

Fraser, J. S., Lindorff-Larsen, K., and Bonomi, M. (2020). What will computational modeling approaches have to say in the era of atomistic cryo-EM data?. *J. Chem. Inf. Model.* 60, 2410–2412. doi:10.1021/acs.jcim.0c00123

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., et al. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* 31, 822–826. doi:10.1038/nbt.2623

Guo, F., and Jiang, W. (2014). Single particle cryo-electron microscopy and 3-D reconstruction of viruses. *Methods Mol. Biol.* 1117, 401–443. doi:10.1007/978-1-62703-776-1_19

Igaev, M., Kutzner, C., Bock, L. V., Vaiana, A. C., and Grubmüller, H. (2019). Automated cryo-EM structure refinement using correlation-driven molecular dynamics. *eLife* 8, e43542. doi:10.7554/eLife.43542

Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., et al. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* 351, 867–871. doi:10.1126/science.aad8282

Jin, M., Han, W., Liu, C., Zang, Y., Li, J., Wang, F., et al. (2019). An ensemble of cryo-EM structures of TRiC reveal its conformational landscape and subunit specificity. *Proc. Natl. Acad. Sci. USA* 116, 19513–19522. doi:10.1073/pnas.1903976116

Karplus, M., and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9, 646–652. doi:10.1038/nsb0902-646

Kim, D. N., Gront, D., and Sanbonmatsu, K. Y. (2020). Practical considerations for atomistic structure modeling with cryo-EM maps. *J. Chem. Inf. Model.* 60, 2436–2442. doi:10.1021/acs.jcim.0c00090

Kirmizialtin, S., Loerke, J., Behrmann, E., Spahn, C. M., and Sanbonmatsu, K. Y. (2015). Using molecular simulation to model high-resolution cryo-EM reconstructions. *Meth Enzymol.* 558, 497–514. doi:10.1016/bs.mie.2015.02.011

Kontziampasis, D., Klebl, D. P., Iadanza, M. G., Scarff, C. A., Kopf, F., Sobott, F., et al. (2019). A cryo-EM grid preparation device for time-resolved structural studies. *IUCrJ* 6, 1024–1031. doi:10.1107/S2052252519011345

Kovacs, J. A., Galkin, V. E., and Wriggers, W. (2018). Accurate flexible refinement of atomic models against medium-resolution cryo-EM maps using damped dynamics. *BMC Struct. Biol.* 18, 12. doi:10.1186/s12900-018-0089-0

Marques, M. A., Purdy, M. D., and Yeager, M. (2019). CryoEM maps are full of potential. *Curr. Opin. Struct. Biol.* 58, 214–223. doi:10.1016/j.sbi.2019.04.006

Mitchell, B. P., Hsu, R. V., Medrano, M. A., Zewde, N. T., Narkhede, Y. B., and Palermo, G. (2020). Spontaneous embedding of DNA mismatches within the RNA:DNA hybrid of CRISPR-Cas9. *Front. Mol. Biosci.* 7, 39. doi:10.3389/fmolb.2020.00039

Nierzwicki, Ł., Arantes, P. R., Saha, A., and Palermo, G. (2020). Establishing the allosteric mecanism in CRISPR-Cas9. *Wires Comput. Mol. Sci.* e1503.

Nogales, E. (2015). The development of cryo-EM into a mainstream structural biology technique. *Nat. Methods* 13, 24–27. doi:10.1038/nmeth.3694

Olsson, S., Wu, H., Paul, F., Clementi, C., and Noé, F. (2017). Combining experimental and simulation data of molecular processes via augmented

Markov models. *Proc. Natl. Acad. Sci. USA* 114, 8265–8270. doi:10.1073/pnas.1704803114

Ovchinnikov, V., Stone, T. A., Deber, C. M., and Karplus, M. (2018). Structure of the emre multidrug transporter and its use for inhibitor peptide design. *Proc. Natl. Acad. Sci. USA* 115, 7932–7941. doi:10.1073/pnas.1802177115

Palermo, G., Miao, Y., Walker, R. C., Jinek, M., and McCammon, J. A. (2017a). CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. *Proc. Natl. Acad. Sci. USA* 114, 7260–7265. doi:10.1073/pnas.1707645114

Palermo, G., Miao, Y., Walker, R. C., Jinek, M., and McCammon, J. A. (2016). Striking plasticity of CRISPR-Cas9 and key role of non-target DNA, as revealed by molecular simulations. *ACS Cent. Sci.* 2, 756–763. doi:10.1021/acscentsci.6b00218

Palermo, G., Ricci, C. G., Fernando, A., Basak, R., Jinek, M., Rivalta, I., et al. (2017b). Protospacer adjacent motif-induced allostery activates CRISPR-Cas9. *J. Am. Chem. Soc.* 139, 16028–16031. doi:10.1021/jacs.7b05313

Palermo, G. (2019). Structure and dynamics of the CRISPR-Cas9 catalytic complex. *J. Chem. Inf. Model.* 59, 2394–2406. doi:10.1021/acs.jcim.8b00988

Palermo, G., Sugita, Y., Wriggers, W., and Amaro, R. E. (2020). Faces of contemporary CryoEM information and modeling. *J. Chem. Inf. Model.* 60, 2407–2409. doi:10.1021/acs.jcim.0c00481

Palermo, G., Chen, J. S., Ricci, C. G., Rivalta, I., Jinek, M., Batista, V. S., et al. (2018). Key role of the REC lobe during CRISPR-Cas9 activation by 'sensing', 'regulating', and 'locking' the catalytic HNH domain. *Quart. Rev. Biophys.* 51, e9. doi:10.1017/s0033583518000070

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

Ricci, C. G., Chen, J. S., Miao, Y., Jinek, M., Doudna, J. A., McCammon, J. A., et al. (2019). Deciphering off-target effects in CRISPR-Cas9 through accelerated molecular dynamics. *ACS Cent. Sci.* 5, 651–662. doi:10.1021/acscentsci.9b00020

Singharoy, A., Teo, I., McGreevy, R., Stone, J. E., Zhao, J., and Schulten, K. (2016). Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *eLife* 5, e16105. doi:10.7554/eLife.16105

Sugita, Y., Kitao, A., and Okamoto, Y. (2000). Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* 113, 6042–6051. doi:10.1063/1.1308516

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure* 16, 295–307. doi:10.1016/j.str.2007.11.016

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

Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B., and Schulten, K. (2009). Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* 49, 174–180. doi:10.1016/j.ymeth.2009.04.005

Turoňová, B., Sikora, M., Schürmann, C., Hagen, W. J. H., Welsch, S., Blanc, F. E. C., et al. (2020). *In situ* structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science* 370, 203–208. doi:10.1126/science.abd5223

Vant, J. W., Sarkar, D., Streitwieser, E., Fiorin, G., Skeel, R., Vermaas, J. V., et al. (2020). Data-guided multi-map variables for ensemble refinement of molecular movies. *J. Chem. Phys.* 153, 214102. doi:10.1063/5.0022433

Wang, J., Arantes, P. R., Bhattarai, A., Hsu, R. V., Pawnikar, S., Huang, Y. M., et al. (2021). Gaussian accelerated molecular dynamics (GaMD): principles and applications. *Wires Comput. Mol. Sci.* e1521. doi:10.1002/WCMS.1521

Wang, Y., Mallon, J., Wang, H., Singh, D., Hyun Jo, M., Hua, B., et al. (2021). Real-time observation of Cas9 postcatalytic domain motions. *Proc. Natl. Acad. Sci. USA* 118, e2010650118. doi:10.1073/pnas.2010650118

Wu, X., Subramaniam, S., Case, D. A., Wu, K. W., and Brooks, B. R. (2013). Targeted conformational search with map-restrained self-guided Langevin dynamics: application to flexible fitting into electron microscopic density maps. *J. Struct. Biol.* 183, 429–440. doi:10.1016/j.jsb.2013.07.006

Zhu, X., Clarke, R., Puppala, A. K., Chittori, S., Merk, A., Merrill, B. J., et al. (2019). Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. *Nat. Struct. Mol. Biol.* 26, 679–685. doi:10.1038/s41594-019-0258-2

Zuo, Z., and Liu, J. (2017). Structure and dynamics of Cas9 HNH domain catalytic state. *Sci. Rep.* 7, 17271. doi:10.1038/s41598-017-17578-6

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: 22 April 2021
doi: 10.3389/fmolb.2021.654333

Refinement of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods

Mustapha Carab Ahmed[1], *Line K. Skaanning*[2], *Alexander Jussupow*[3], *Estella A. Newcombe*[1,2], *Birthe B. Kragelund*[1], *Carlo Camilloni*[3,4], *Annette E. Langkilde*[2] *and Kresten Lindorff-Larsen*[1]*

[1] Structural Biology and NMR Laboratory, Department of Biology, Linderstrøm-Lang Centre for Protein Science, University of Copenhagen, Copenhagen, Denmark, [2] Department of Drug Design and Pharmacology, University of Copenhagen, Copenhagen, Denmark, [3] Department of Chemistry, Institute for Advanced Study, Technical University of Munich, Munich, Germany, [4] Dipartimento di Bioscienze, Università degli Studi di Milano, Milan, Italy

OPEN ACCESS

Edited by:
Massimiliano Bonomi,
Institut Pasteur, France

Reviewed by:
Paul Robustelli,
Dartmouth College, United States
Jochen Hub,
Saarland University, Germany

***Correspondence:**
Kresten Lindorff-Larsen
lindorff@bio.ku.dk

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

Received: 15 January 2021
Accepted: 12 March 2021
Published: 22 April 2021

Citation:
Ahmed MC, Skaanning LK, Jussupow A, Newcombe EA, Kragelund BB, Camilloni C, Langkilde AE and Lindorff-Larsen K (2021) Refinement of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods. Front. Mol. Biosci. 8:654333. doi: 10.3389/fmolb.2021.654333

The inherent flexibility of intrinsically disordered proteins (IDPs) makes it difficult to interpret experimental data using structural models. On the other hand, molecular dynamics simulations of IDPs often suffer from force-field inaccuracies, and long simulation times or enhanced sampling methods are needed to obtain converged ensembles. Here, we apply metainference and Bayesian/Maximum Entropy reweighting approaches to integrate prior knowledge of the system with experimental data, while also dealing with various sources of errors and the inherent conformational heterogeneity of IDPs. We have measured new SAXS data on the protein α-synuclein, and integrate this with simulations performed using different force fields. We find that if the force field gives rise to ensembles that are much more compact than what is implied by the SAXS data it is difficult to recover a reasonable ensemble. On the other hand, we show that when the simulated ensemble is reasonable, we can obtain an ensemble that is consistent with the SAXS data, but also with NMR diffusion and paramagnetic relaxation enhancement data.

Keywords: small-angle X-ray scattering, molecular dynamics simulation, NMR, protein, intrinsically disordered protein

INTRODUCTION

Intrinsically Disordered Proteins (IDPs) play important roles in a wide range of biological processes including cell signaling and regulation (Uversky et al., 2005; Das et al., 2015; Snead and Eliezer, 2019), and their malfunction or aggregation is linked to neurodegenerative diseases such as Alzheimer's and Parkinson's disease. A key, defining property of IDPs is that they do not adopt well-defined, permanent secondary and tertiary structures under native conditions, and their conformational properties are thus best described in statistical terms.

Due to the dynamic nature of IDPs and their inherent conformational heterogeneity, IDPs are not easily amenable to high-resolution characterization solely through experimental measurements. To characterize their structural and dynamic properties it is often necessary to integrate various biophysical experiments, and particularly nuclear magnetic resonance (NMR)

spectroscopy (Dyson and Wright, 2001), small angle X-ray scattering (SAXS) (Bernado and Svergun, 2012), circular dichroism (Chemes et al., 2012), and single-molecule Förster resonance energy transfer (sm-FRET) (LeBlanc et al., 2018) have been widely used to characterize the structural properties of IDPs. For instance, pulsed-field-gradient NMR diffusion and SAXS experiments are especially useful to quantify the level of compaction of the IDP. Techniques such as sm-FRET and NMR paramagnetic relaxation enhancement (PRE) provide distance information between different residues or regions of the IDP (Dedmon et al., 2005; Eliezer, 2009). Nevertheless, since most experimental methods only convey ensemble-averaged information and are also affected by random and systematic errors, it is difficult to directly extract information on the underlying heterogeneous ensemble of the IDP. To address this problem, theoretical and computational models can be used to extract detailed structural information from these experiments.

Molecular dynamics (MD) simulations that use physics-based force fields may provide high-resolution temporal and spatial information about the structure and dynamics of IDPs. Extensive sampling of a force field with MD simulations can thus be used to generate conformational ensemble of the IDP. The quality of the results, however, depends heavily on the accuracy of the force field employed. For example, it has been shown that many earlier generations of force fields produce overly compact conformations for many IDPs (Piana et al., 2015). It appears that these force fields fail to accurately describe the solvation of the protein by underestimating protein-water interactions (Sun and Kollman, 1995; Nerenberg et al., 2012; Best et al., 2014; Piana et al., 2015). Recently, however, significant advancements have been made to improve force field accuracy and correct the bias toward overly compact conformations (Best et al., 2014; Piana et al., 2015; Song et al., 2017; Robustelli et al., 2018). Adding to these issues, the large conformational phase space of IDPs, requires extensive sampling of the protein in order to generate converged ensembles. To achieve sufficient sampling, and push the sampling capacity of MD simulations, one often employs enhanced sampling methods such as metadynamics (Barducci et al., 2008) or parallel-tempering replica exchange (Sugita and Okamoto, 1999). Notably, force field and sampling problems are expected to be more severe for longer IDPs.

An approach to address the challenges of force-field accuracy is to combine experimental and theoretical information in order to obtain conformational ensembles of IDPs that agree with experimental measurements. In this way, the simulations are used as a tool to interpret experimental measurements. A number of different approaches have been described and can, roughly, be divided into two different classes in which the experimental data is either (i) used for on-the-fly restraining of a simulation to experimental data, or (ii) post-processing ensembles generated by simulations to match experimental data by reweighting or selection methods. Many different such methods exist, and we refer to recent reviews for additional details (Cesari et al., 2018; Orioli et al., 2020).

Because the conformational ensembles are broad and the experimental data often have low information content and may be noisy, Bayesian inference methods (Box and Tiao, 2011) and the maximum entropy principle (Jaynes, 1957) have emerged as particularly successful frameworks for studying IDPs. In these frameworks, an ensemble generated using a prior model is minimally modified to match the experimentally observed data better. An extension of these frameworks for integrative structural ensemble determination is Metainference Metadynamics (M&M) (Bonomi et al., 2016a), that combines multi-replica all-atom molecular dynamics simulations with ensemble averaged experimental data (Bonomi et al., 2016b). In the M&M approach, the metainference (Bonomi et al., 2016a) part is a Bayesian inference method that allows for the integration of experimental information with prior knowledge of the system from, e.g., physics-based force fields, while also dealing with uncertainty and errors as well as conformationally heterogeneous systems. In addition, metainference can be combined with metadynamics (Laio and Parrinello, 2002; Bonomi et al., 2016b) to accelerate sampling further. A related Maximum Entropy approach has also been applied to determine an ensemble of configurations from SAXS data but using a more refined and potentially accurate method for taking solvent effects into account (Hermann and Hub, 2019). While the above approaches apply the bias on the fly, other Bayesian formalisms takes as input simulations that were generated without taking the experimental data into account, and subsequently updates this using statistical reweighting. Such approaches include our Bayesian/Maximum Entropy (BME) protocol (Bottaro et al., 2020), as well as related methods (Hummer and Köfinger, 2015).

Here, we combined ensemble-averaged experimental SAXS data with MD simulations with the aim to achieve structural ensembles of the system which are in agreement with the experimental data. We did so using both metainference and BME. In particular, we used BME to refine ensembles that had previously been generated using MD simulations (Piana et al., 2015; Robustelli et al., 2018), while metainference was applied to restrain experimental SAXS data during MD simulations with an implicit solvent model (Bottaro et al., 2013). We used the intrinsically disordered protein α-synuclein (αSN) protein as a model, as this protein has been studied extensively by various experimental methods including SAXS and NMR measurements, and because of the availability of long MD trajectories generated from a range of force fields and water models. αSN is a 140-residue long IDP that is primarily expressed in the brain and in its monomeric state is known to be disordered and populate multiple conformational states. αSN aggregation into amyloid fibrils is linked to Parkinson's disease and dementia with Lewy bodies (Spillantini and Goedert, 2000; Ulusoy and Di Monte, 2013).

We assessed the quality of existing ensembles before refinement, and the ability of metainference and BME methods to improve them through incorporation of experimental SAXS data, by comparing with independent measurements of the level of compaction (through the hydrodynamic radius, R_h, as probed by NMR) and previously measured paramagnetic relaxation enhancement data (Dedmon et al., 2005). We find that the inclusion of a SAXS-restraint in the M&M simulation resulted in the generation of a reliable and heterogenous conformational ensemble that also improved the agreement with the NMR

diffusion data. The BME reweighting improved the agreement with the experimental data when we applied the approach to simulations with the TIP4P-D water model. For simulations using the TIP3P water model, which were substantially more compact, it was difficult to find a suitably large ensemble compatible with the experimental SAXS data. Together, our result provide insight into how and when experimental SAXS data can be used to refine ensembles of IDPs, and the role played by the force field as a 'prior' in these Bayesian/Maximum entropy approaches.

METHODS AND MATERIALS

Experimental Data

Human *αSN* for SAXS experiments was expressed, purified, and lyophylized as previously described (van Maarschalkerweerd et al., 2014). Prior to SAXS data collection, the lyophilized powder was dissolved in PBS (20 mM Na_2HPO_4, 150 mM NaCl, pH 7.4) and filtered through a 0.22 μm filter to remove larger aggregates. The final sample concentration before SEC-SAXS was determined by A_{280} to be 4.5 mg/mL using an extinction coefficient of 5960 M^{-1} cm^{-1}. SAXS data was collected as SEC-SAXS data on beamline P12 (Blanchet et al., 2015) operated by EMBL Hamburg at the PETRA III storage ring (DESY, Hamburg, Germany). 50 μL 4.5 mg/mL *αSN* in PBS buffer (20 mM Na_2HPO_4, 150 mM NaCl, pH 7.4) was injected on a Superdex 200inc 5/150 GL column with a flowrate of 0.4 mL/min. The column was pre-equilibrated with the running buffer (PBS with 2% (v/v) glycerol). SAXS data were collected at 20 °C, with continuous exposure of 1 s per frame throughout the SEC elution. Data processing was done using CHROMIXS (Panjkovich and Svergun, 2018), averaging sample data from the frames in the monomeric peak and subtracting the buffer signal taken from the flow-through prior to the sample elution to obtain the final scattering profile (**Supplementary Figure 1**).

We purified *αSN* for NMR experiments as previously described (Skaanning et al., 2020). Translational diffusion constants for *αSN* (50μM with 2% (v/v) glycerol) and 1,4-dioxane (0.2% v/v; as internal reference) were determined by fitting peak intensity decay from diffusion ordered spectroscopy experiments (Wu et al., 1995), using the Stejskal-Tanner equation as described (Prestel et al., 2018). Spectra (a total of 64 scans) were obtained over a gradient strength of 2 to 98%, with a diffusion time (Δ) of 200 ms and gradient length (δ) of 3 ms. Diffusion constants were used to estimate the hydrodynamic radius for *αSN* described (Wilkins et al., 1999; Skaanning et al., 2020) (**Supplementary Figure 2**).

We used previously measured PRE data obtained by measuring intensity ratios with spin-labels added at five different positions (residue: 24, 42, 62, 87, and 103) (Dedmon et al., 2005).

Bayesian/Maximum Entropy Reweighting of Unbiased MD Simulations

We used previously generated ensembles of *αSN* obtained by long-timescale MD simulations with different force fields from the CHARMM and Amber families (here abbreviated by C and A, respectively) and water models (Piana et al., 2015; Robustelli et al., 2018) (**Table 1**). The published simulation using Amber ff99SB-*disp* (Robustelli et al., 2018) was later found to be affected by interactions with its periodic image and has here been replaced by a 73 μs long simulation performed using the same setup but in a 160Å box and available directly from D. E. Shaw Research.

We used our Bayesian/Maximum Entropy (BME) protocol (Ahmed et al., 2020; Bottaro et al., 2020) to reweight the initial force field ensembles (**Table 1**) with the experimental SAXS data, thus obtaining ensembles that are in closer agreement to the experimental data. Briefly described, the BME approach is based on a combined Bayesian/Maximum entropy framework, that enables one to refine a simulation using experimental data while also taking into account the potential noise in the data and in the so-called forward model used to calculate observables for the ensemble. The purpose of the reweighting is to derive a new set of weights for each configuration in a previously generated ensemble so that the reweighted ensemble satisfies the following two criteria: (i) it matches the experimental data better than the original ensemble and (ii) it achieves this improved agreement by a minimal perturbation of the original ensemble. The BME reweighting approach seeks to update the weights, w_j, by minimizing the function:

$$\mathcal{L}(w_1 \dots w_n) = \frac{1}{2}\chi^2(w_1 \dots w_n) - \theta S_{\text{rel}}(w_1 \dots w_n) \quad (1)$$

Here, χ^2 quantifies the agreement between the experimental data and the corresponding observable calculated from the reweighted ensemble. $S_{rel} = -\sum_j^n w_j \log\left(w_j/w_j^0\right)$ measures the deviation between the original ensemble weights, w_j^0, in our case taken as $1/n$, and the reweighted ensemble weights. Finally, the hyperparameter θ tunes the balance between the two terms, and needs to be determined, by evaluating the compromise between the two terms in Equation (1) (Orioli et al., 2020). Reweighting and analysis scripts are available at github.com/KULL-Centre/papers/blob/master/2021/aSYN-ahmed-et-al/.

Metainference Metadynamics

We conducted a SAXS-restrained MD simulation using the metainference metadynamics (M&M) method, where we employed the parallel-bias (PBMetaD) flavor of well-tempered metadynamics (Pfaendtner and Bonomi, 2015) in combination with the multiple-walkers scheme (Raiteri et al., 2006). During the M&M simulation, the SAXS back-calculation step utilizes a hybrid-resolution approach, where the SAXS data is calculated on-the-fly using "Martini beads" that are superimposed on the all-atom structures using PLUMED (Bonomi and Camilloni, 2017; Paissoni et al., 2019, 2020; Jussupow et al., 2020). The approach is particularly efficient as the SAXS back-calculation is calculated using the Debye equation from a coarse-grained model and the excess of electron density in the hydration shell is neglected (Niebling et al., 2014; Paissoni et al., 2020). We note here that the Martini model is only used for calculating the SAXS data, and the simulations are performed using an all-atom, implicit solvent model as detailed below.

TABLE 1 | Ensembles analyzed and refined.

Force field	Water model	Time(μs)	R_g Force field(Å)	R_g Reweighted(Å)	R_h Force field(Å)	R_h Reweighted(Å)
A12	TIP3P	5	15.4 ± 0.1	19 ± 1	20.8 ± 0.1	23.0 ± 0.1
A99SB-ILDN	TIP3P	5	15.3 ± 0.2	16.0 ± 0.3	20.6 ± 0.3	21.3 ± 0.3
C22*	TIP3P	6	17.1 ± 0.4	23 ± 1	22.2 ± 0.3	26.1 ± 0.5
A99SB-ILDN	TIP4P-EW	5	17.9 ± 0.8	24 ± 1	22.8 ± 0.6	26.4 ± 0.6
C22*	TIP4P-D	20	23.3 ± 0.6	29.3 ± 0.9	26.7 ± 0.3	29.6 ± 0.4
A99SB-ILDN	TIP4P-D	11	25.7 ± 0.1	31 ± 1	27.2 ± 0.6	30 ± 1
A12	TIP4P-D	11	29.7 ± 0.5	34.1 ± 0.3	29.7 ± 0.2	32 ± 0.5
A03ws	TIP4P/2005	20	30 ± 2	34.3 ± 0.6	29.1 ± 1.1	32 ± 1
A99SB-*disp*	[1]	73	26 ± 1	31.9 ± 0.6	27.7 ± 0.5	30.8 ± 0.4
CHARMM36[2]	EEF1-SB	3.2[3]	46 ± 4	35.4 ± 0.5	38 ± 3	33.1 ± 0.5
Experiment			35.5 ± 0.5		28.6 ± 0.7	

[1] *A99SB-disp uses a modified version of the TIP4P-D water model.*

[2] *CHARMM36 with EEF1-SB was only used for the metainference metadynamics simulations; here "force field" and "reweighted" refers to two different simulations with and without the experimental bias, respectively.* [3] *Metadynamics simulation time.*

We used GROMACS 2018.1 (Abraham et al., 2015) with PLUMED version 2.4 (Tribello et al., 2014) to perform the *M&M* simulations. We used the CHARMM36 force field (Best et al., 2012) with the EEF1-SB implicit solvent model (Bottaro et al., 2013). We used a previously generated structure of αSN bound to micelles (Ulmer et al., 2005) as starting point for an initial 100-ns long high temperature (500 K) simulation, from which we extracted 64 starting conformations for the multi-replica *M&M* simulation. Charged amino acids were neutralized in line with the parameterization of the EEF1 model (Lazaridis and Karplus, 1999; Bottaro et al., 2013), leaving a neutral molecule, and performed a minimization to a maximum force of 100 kJ/mol/nm. The system was further equilibrated for 20 ns per replica with the metainference bias.

We performed production simulations in the NVT ensemble using Langevin dynamics (Goga et al., 2012) with a friction coefficient of 0.5 ps^{-1} at T = 310 K, and a timestep of 2 fs. The Coulomb interactions were evaluated with a distance dependent dielectric constant of $\epsilon = 15r$ (Lazaridis and Karplus, 1999; Bottaro et al., 2013) and a cut-off at 9 Å. Constraints were applied on the hydrogens with the LINCS algorithm (Hess et al., 1997). For the production simulations the sampling of each replica was enhanced by PBMetaD along with twelve collective variables (CVs) consisting of the radius of gyration and 11 AlphaRMSD CVs to enhance sampling of local backbone conformations (Tribello et al., 2014).

Gaussians were deposited every 200 steps with a height of 0.1 kJ/mol/ps, and the σ values were set to 0.2 *nm* for CVrg and 0.010 for all AlphaRMSD CVs, respectively. We rescaled the height of the Gaussians using the well-tempered scheme with a bias-factor of 20 (Barducci et al., 2008).

Because calculation of the SAXS data is limiting in these simulations, we re-binned the experimental SAXS data to a set of 19 SAXS intensities at different scattering vectors, ranging between 0.01 $Å^{-1}$ and 0.20 $Å^{-1}$. Metainference was applied every 10 steps of the simulation. We used a Gaussian noise model, that applies a single Gaussian per SAXS data-point. The scaling factor between experimental and calculated SAXS intensities was sampled with a flat prior between 0.5 and 2.0 (Löhr et al., 2017). We averaged the estimated metainference weights over a time window of 200 steps; this is done to avoid large fluctuations and prevent numerical instabilities due to too high instantaneous forces (Löhr et al., 2017). The Plumed input file is available in the PLUMED-NEST database (Bonomi et al., 2019) (plumID:21.003; www.plumed-nest.org/eggs/21/003/).

Paramagnetic Relaxation Enhancement

Paramagnetic Relaxation Enhancement (PRE) via nitroxide spin-labels has been used extensively to study long-range interactions within IDPs. The measured PRE depends in particular on the distance between a paramagnetic centre and protein nuclei, in this case backbone amides. Because the PRE originates from a dipolar interaction, the observed PRE depends on r^{-6}, and is thus particularly sensitive to transient, short distances. Because simulations were performed without the spin-labels, and because multiple spin-labels were used to probe the structural ensemble of αSN, we used a post-processing approach to estimate the location of the unpaired electron on the nitroxide label. In particular, we used DEER-PREdict (Tesei et al., 2020), which is based on a Rotamer Library Approach to place spin labels on the protein, to estimate PRE rates. We calculated and compared results from five paramagnetic labeling positions (residue: 24, 42, 62, 87, 103) in αSN (Dedmon et al., 2005). Additional details are available in the **Supplementary Information** and in the DEER-PREdict paper (Tesei et al., 2020).

RESULTS AND DISCUSSION

Using αSN as an example, we compared conformational ensembles generated either directly using molecular dynamics simulations with a molecular mechanics force field, or the same ensemble refined using SAXS data. We also analyzed the results of an approach (M&M) that performs this refinement during the simulation. We thus performed (i) a SAXS-restrained multi-replica simulations using metainference metadynamics and (ii) a reference simulation both using CHARMM36 force

field (Best et al., 2012) used with the EEF1-SB implicit solvent model (Bottaro et al., 2013). Both simulations consisted of 64 replicas, with one simulation using metainference to enforce the agreement with experimental SAXS data, whereas a second, reference simulation did not use experimental restraints and thus sampled the force field only. We also analyzed nine previously published multi-μs MD simulations which had been generated using different combinations of proteins force fields and water models (Piana et al., 2015; Robustelli et al., 2018) from the AMBER (Hornak et al., 2006; Best and Hummer, 2009; Lindorff-Larsen et al., 2010; Robustelli et al., 2018) and CHARMM (Piana et al., 2011) families in combination with either standard TIP3P (Jorgensen, 1981), TIP4P-EW (Horn et al., 2004), TIP4P/2005 (Abascal and Vega, 2005), or the TIP4P-D (Piana et al., 2015) water model. **Table 1** summarizes the simulations and below we refer to the prior (not refined) ensemble as the "force field" ensemble and the posterior (refined) ensemble as the "reweighted" ensemble.

Force Field Accuracy and Sampling

Before the refinement procedure we calculated SAXS intensity curves from each structure in the ensembles using PEPSI-SAXS (Grudinin et al., 2017). We also calculated the R_g from the protein coordinates and used them to estimate the hydrodynamic radius (R_h) for each conformation using a previously described empirical relationship (Nygaard et al., 2017; Ahmed et al., 2020) (**Table 1**). The experimental $R_g = 35.5$ Å was obtained through Guinier analysis of the experimental SAXS curve (see Methods), while the experimental $R_h = 29.0$ Å was obtained through NMR diffusion measurements (**Table 1**).

In line with previous observations (Piana et al., 2015; Robustelli et al., 2018), the ensembles show very different levels of compaction depending on the force field and, in particular, water model used (**Table 1** and **Figure 1**). When paired with the TIP3P water model, both the Amber or CHARMM force fields produce very compact conformations and show poor agreement with the experimental value of R_g. On the other hand, when paired with the recently parameterized TIP4P-D water model the force fields give rise to more expanded structures and match the experimental values of R_g and R_h considerably better. The ensemble generated using CHARMM36 with the EEF1-SB implicit solvent model on the other-hand produce more expanded structures (**Table 1**). Of particular relevance to the reweighting described below it is worth noting how the compact ensembles either do not sample any, or at most very few, structures that are expanded as the *average* R_g observed in experiment (**Figure 1**). This observation already suggests that it will be difficult robustly to derive ensembles that are in agreement with the SAXS data as this in particular is sensitive to the R_g.

Ensemble Refinement Using SAXS Data

In the following section we exemplify the BME refinement against the SAXS data using two representative combinations of force field and water models, specifically A12 paired with either the TIP3P or the TIP4P-D water model (**Figure 2**). We also present the results obtained from "on-the-fly" SAXS-restrained simulation with M&M which we compared to an unrestrained simulation with otherwise identical simulation settings (see Methods). Note that while the R_g values for the simulations were calculated using protein coordinates, the experimental value also includes potential contributions from the solvent. The refinement, analysis and plots for the remaining force fields are shown in the supplementary information (**Supplementary Figures 4–10**).

The BME procedure works by assigning weights to a previously generated ensemble so as to fit the experimental data better. For BME to successfully reweight an ensemble it is thus required that the initial prior ensemble contains the most relevant conformational states of the protein, such that the ensemble that gives rise to the experimental data is a sub-ensemble of the initial prior ensemble. Consequently, if the sampling is incomplete or the unbiased ensemble is very far away from the true ensemble, it may not be possible to reweight the ensemble to reach a satisfactory agreement with the experiments. An indication that this is occurring is that BME will effectively down-weight most of the structures in the prior ensemble and the posterior ensemble will be dominated by a few structures with large weights. This can in turn be quantified by calculating the (effective) fraction of structures, $\phi_{eff} = \exp(S_{rel})$, that contribute to the ensemble (Orioli et al., 2020), so that when $\phi_{eff} \approx 1$ most of the structures are retained, whereas $\phi_{eff} \approx 0$ indicates a few structures with very large weights

In the BME reweighting the confidence in the prior ensemble with respect to the experimental data can be tuned by the hyper-parameter θ (Equation 1). One usually does not know the optimal value for θ beforehand. Here, we choose θ by performing an L-curve analysis (Hansen and O'Leary, 1993; Orioli et al., 2020) in which we plot the χ^2_{red} value (quantifying the difference between experiments and calculated value) as a function of ϕ_{eff}, for different values of θ and choose a value corresponding to the "elbow" region (blue region in **Figures 2A,B**). The L-curve analysis for the A12 force field paired with TIP4P-D water model, lead us to choose $\theta = 1,000$, after which the ensemble retains 88% of the initial structures in the final reweighted ensemble, and show much better agreement with the experimental data, indicative by a low χ^2_{red} (**Figure 2A**). In contrast, the analysis for the TIP3P water model, after reweighting with $\theta = 6,000$, show that only 12% of the initial structures are used in the final reweighted ensemble in order to achieve significant improved agreement with the experimental data (**Figure 2B**). Even at a lower θ value there is still a large discrepancy between experimental and calculated SAXS data ($\chi^2_{red} = 17$ at $\theta = 500$). This is a clear example of a poor prior ensemble, which is caused by insufficient overlap between the force field ensemble and that probed by experiment. In fact, the highest value observed (R_g =23 Å) is significantly lower than the experimental value (black). As a consequence, BME 'throws out' most of the structures from the initial force field ensemble, and the final reweighted ensemble mainly consist of a few highly weighted structures (**Figure 2D**).

The ensemble generated with the TIP4P-D water model (**Figure 2C**) contains structures that span a greater range of R_g values, both above and below the experimental value. After refinement, the reweighted ensemble is shifted to give greater weight to more expanded structures and bringing the average R_g

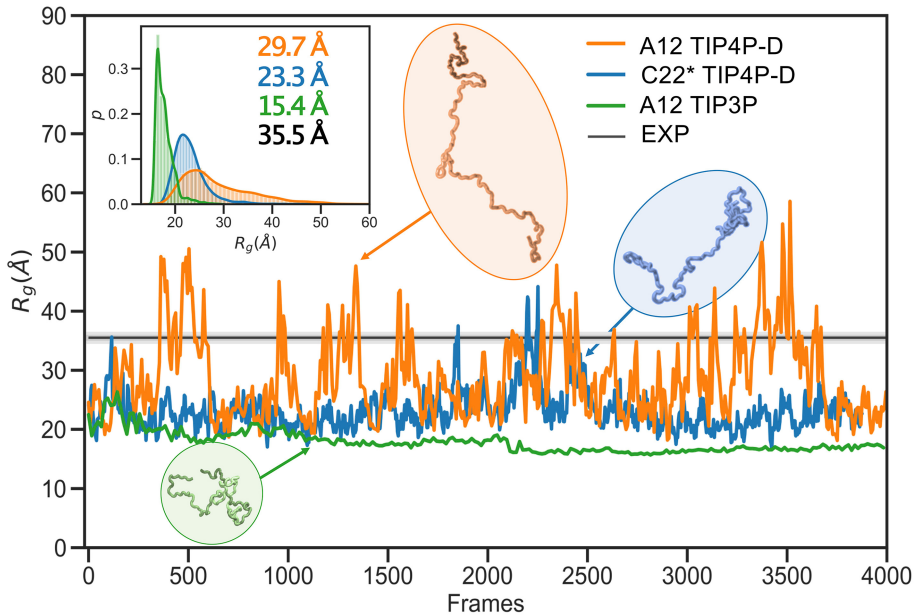


FIGURE 1 | Radius of gyration during simulations with different force fields and water models. As representative examples we show the time-evolution of the radius of gyration for simulations of *αSN* performed with the A12 force field (orange), C22* (blue), and A12 (green) with the TIP4P-D, TIP4P-D, and TIP3P water model, respectively. The experimental value (black) was obtained from a Guinier analysis of the SAXS data. The orange and blue curves have been smoothed to ease visualization. The insert shows probability densities and averages of R_g. Representative structures with different degrees of compaction are also shown. The length of the simulations is 11, 20, and 5 μs, respectively, but are shown here on a normalized timescale to make comparisons easier.

substantially closer to the value estimated from the SAXS data. We note here that we do not fit the R_g value but rather the SAXS data. Because the experimental value of R_g (obtained from a Guinier analyses of the data) contains a contribution from the solvent we do not expect a perfect agreement with the average R_g calculated from the protein coordinates (Henriques et al., 2018). Indeed, this is one of the reasons why we fit the SAXS data directly rather than the R_g.

The effect of reweighting of the two ensembles can also be seen on the distributions of R_h (**Figures 2E,F**). Similar to R_g distributions, the TIP4P-D ensemble is shifted to give greater weight to more expanded structures (**Figure 2E**). As was also evident from the distribution of R_g, the more compact TIP3P ensemble gives rise to a very noisy distribution, because the reweighted ensemble predominantly consists of a few highly weighted structures (**Figure 2F**). To illustrate the consequences of reweighting we also compared the calculated SAXS data from the initial force field and reweighted ensembles to the experimental scattering data (**Figures 2G,H**). As expected, the refined ensembles show better agreement with experiments, in particular for the A12 paired with TIP4P-D. As agreement between experimental and calculated data is the target for BME this observation again just illustrates that the BME method is indeed optimizing agreement.

We repeated these analyses for the remaining combinations of force fields and water models (**Supplementary Figures 4–10**) and summarize the results by assessing how well the ensembles reproduce R_g and R_h before and after refinement (**Figure 3**). We note that the improvement of the R_g observed is due to the use of SAXS data in the refinement, as SAXS intensity curve inherently contains information of the R_g, and that improved agreement with the R_g is thus a sign of the BME approach working rather than a validation of the ensemble.

To evaluate the effectiveness of the SAXS-restrained M&M simulation we monitored the agreement between the back-calculated and the experimental data over the simulation time by monitoring their correlation rather than the χ^2 (Paissoni et al., 2020). Both the SAXS-restrained and the unrestrained reference simulation show a high correlation between back-calculated and experimental data ($>$ 0.98) (**Supplementary Figure 3A**). As expected, the agreement improves substantially when the experimental data is used as a bias in the metainference simulations, confirming the effectiveness of the inclusion of experimental SAXS data (**Supplementary Figure 3A**). Likewise, the average R_g, R_h and the back-calculated SAXS intensity data show improved agreement with the experimental data in the metainference produced ensemble (**Figure 3** and **Supplementary Figure 3**).

In total our analyses show that it is possible to refine MD simulations against SAXS data, though the extent to which agreement can be reached depends on the quality of the input ensemble. For the most compact ensembles we are able to increase the average compaction by fitting to the data, though the average R_g and R_h are still substantially below the experimental values. While the SAXS data (and thus R_g) were used as target values, we also cross-validated with R_h which

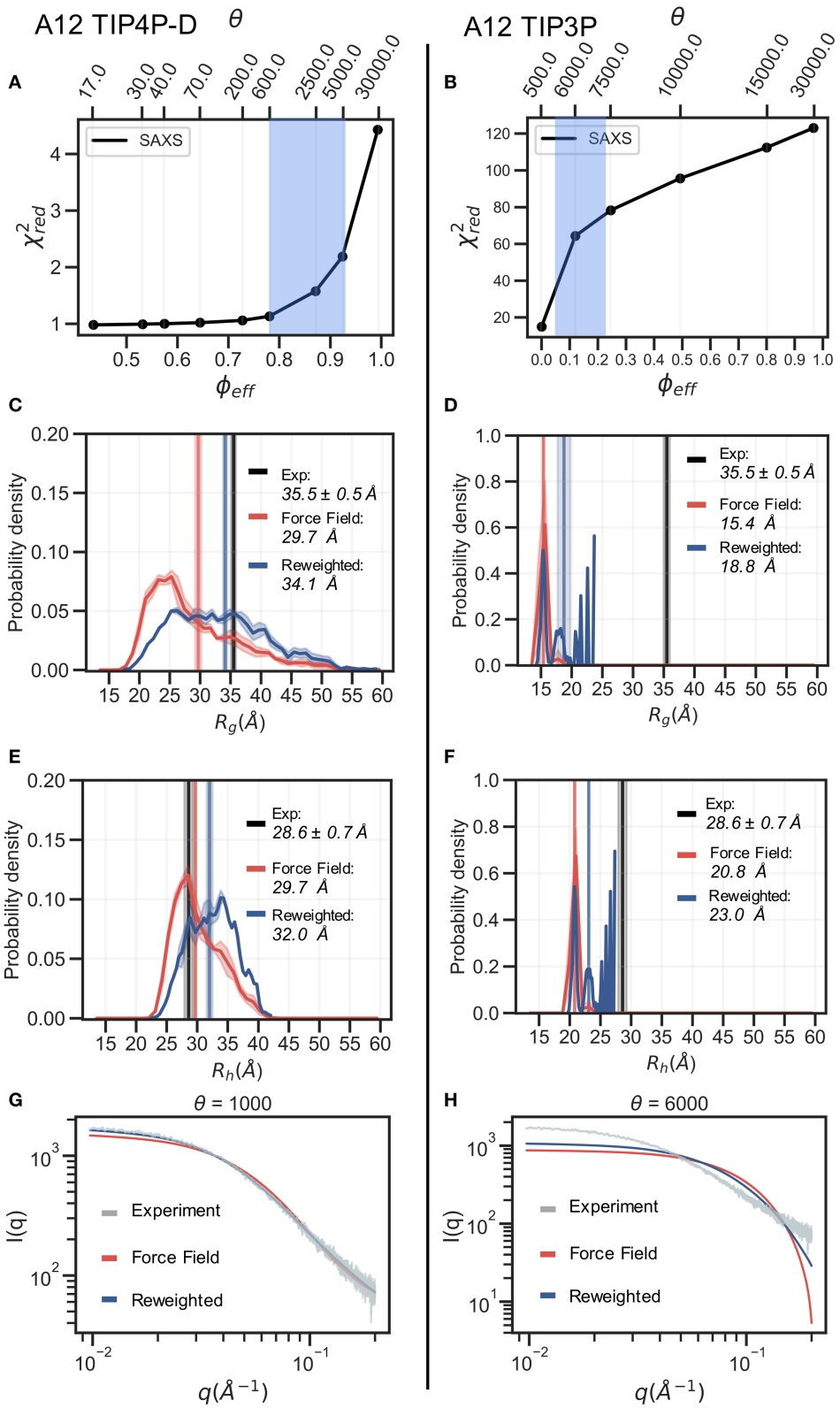


FIGURE 2 | Refinement of two ensembles using BME with SAXS data. SAXS refinement of an ensemble sampled with A12 and either (left) the TIP4P-D water model or (right) the TIP3P water model. (**A,B**) In the L-curve analysis to select the parameter θ we plot χ^2 against ϕ_{eff}. θ balances the prior (force field) and the experimental data, ϕ_{eff} is the effective number of frames used in the final reweighted ensemble. A value of θ is selected from the region marked in blue. We here used θ = 1,000 and

(Continued)

FIGURE 2 | θ = 6,000 for the TIP4P-D ensemble and TIP3P ensemble, respectively. Probability distribution of (**C,D**) R_g and (**E,F**) R_h for the prior (red) and reweighted (blue) ensembles. Solid vertical lines represent the ensemble averaged R_g and R_h. The experimental values are shown in black. The error of the distributions and on the averages (shown as shades) were estimated by block averaging. (**G,H**) Calculated SAXS intensities from the prior ensemble and the reweighted ensembles are compared to the experimental SAXS data.

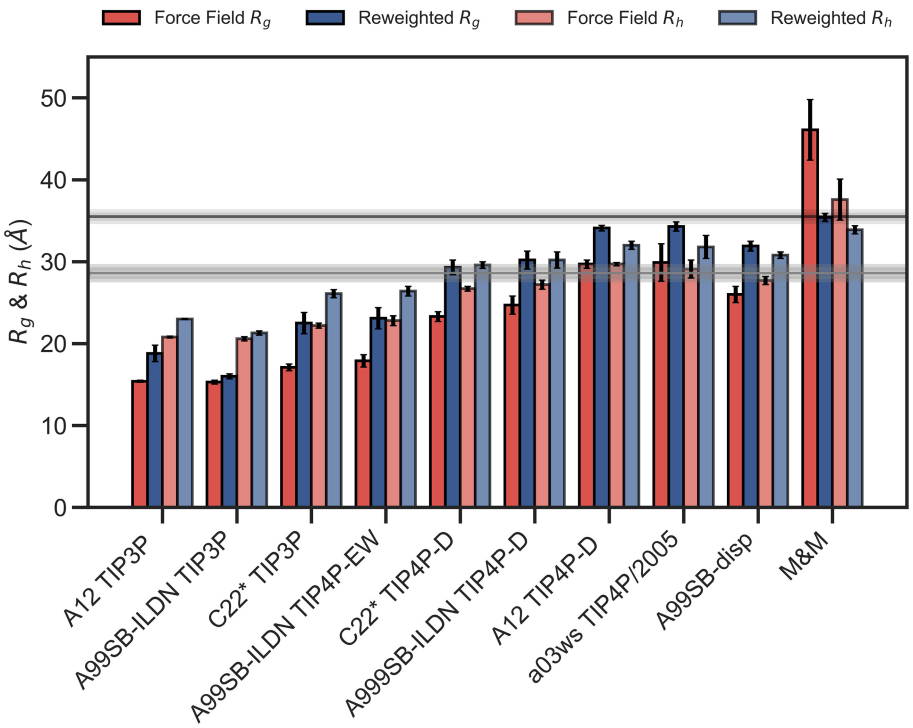


FIGURE 3 | Radius of gyration and hydrodynamic radius calculated from the initial force field ensemble (red) and the experimentally refined ensembles (blue). Experimental values from SAXS (R_g = 35.5Å) and NMR (R_h = 29.0Å) are shown as horizontal lines with the shaded area indicating the error of the experimental values.

was not used in the fitting. Here, the picture is less clear. Overall, for the more compact ensembles, fitting the SAXS data lead to improved prediction of R_h. For other ensembles, such as A12 with TIP4P-D, that show good agreement with R_h before reweighting, the agreement became slightly worse after reweighting. Finally, for the most expanded ensemble obtained with CHARMM36/EEF1-SB, agreement with R_h improved after biasing with the SAXS data. As discussed further below, the approach that we use to estimate R_h from the ensembles is approximate and requires further assessment before these small differences can be interpreted in detail.

Validation With PRE Data

PRE experiments probe the population-weighted average of the distance (as r^{-6}) between a paramagnetic centre and protein nuclei and, given the r^{-6} dependency, is sensitive to the shorter distances even if the populations are small. Here, we compare previously published PREs from spin-labeled αSN (Dedmon et al., 2005) and back-calculated PRE intensity ratios from five labeling sites, for each of the force fields in **Table 1**, before and after refinement (see also Supporting Information). PRE intensity-ratio profiles from a more expanded ensemble generated using A12 with TIP4P-D (**Figure 4A**) and a more compact one generated with A12 with TIP3P (**Figure 4B**) show clear differences in agreement with experiments before refinement with the SAXS data.

BME refinement leads only to small changes in the calculated PRE data for A12/TIP4P-D, whereas the selection of more expanded structures, by applying BME to the ensemble generated with A12/TIP3P, leads to more substantial changes as quantified for example by calculating the RMSD between simulation and experimental data (**Figures 4C,D**). We performed similar calculations and analyses for all ensembles (**Supplementary Figures 11–18**) and summarize the overall RMSD before and after BME (**Figure 4E**). For the force fields paired with TIP3P in particular, we observe many of the long-range contacts diminish after reweighting. These results suggest that the reweighting decreases contributions from structures that are too compact, and that the final reweighted ensemble contains more extended structures. In the TIP4P-D ensembles we still observe that some long-range contacts persist even after reweighting and the better agreement is not alone achieved at the cost of a complete elimination of long-range contacts; nevertheless, the improvements of the PREs are generally small

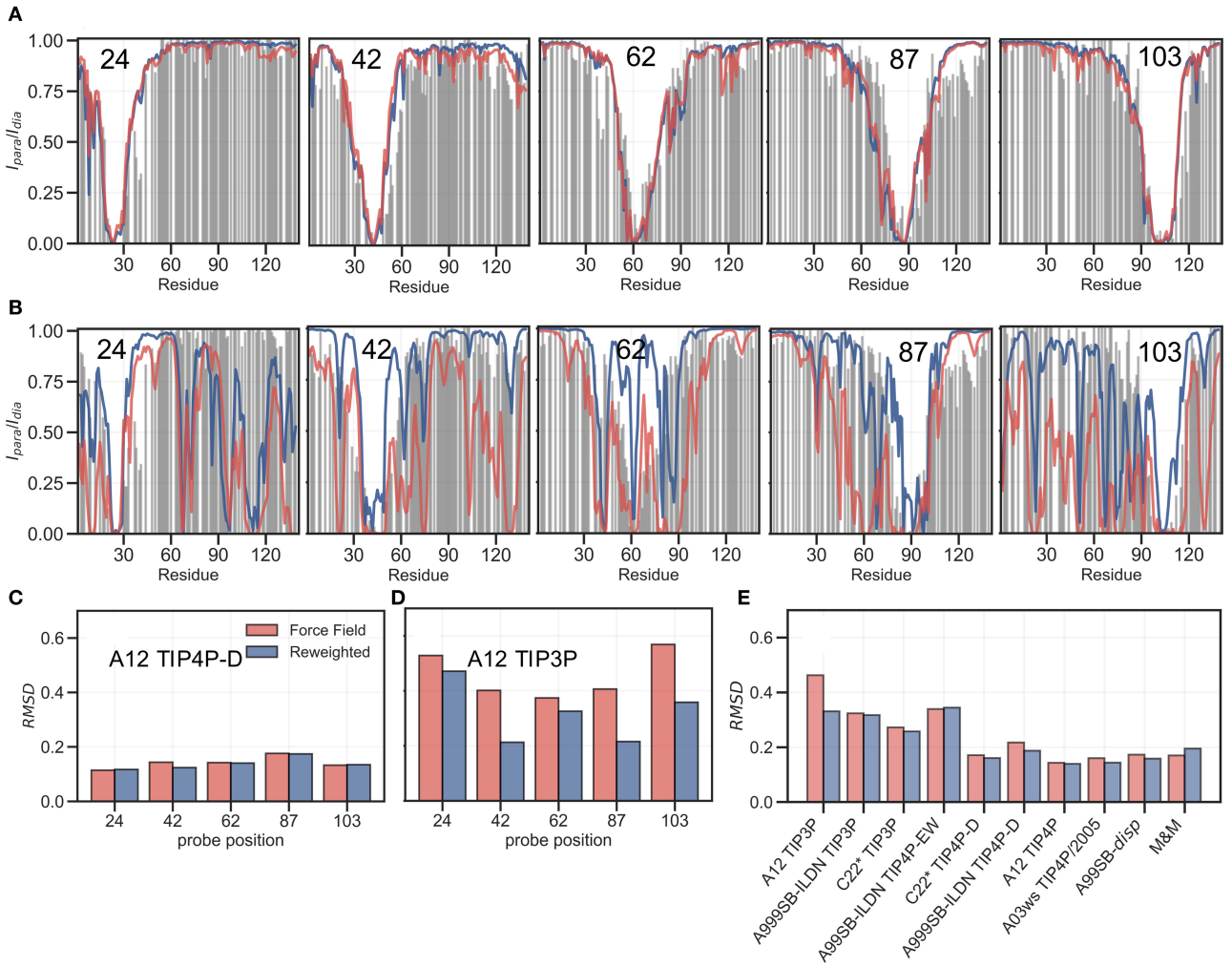


FIGURE 4 | Comparing ensembles to PRE data. We calculated the PRE intensity ratios both from the prior (red) and the reweighted (blue) ensembles and compared to the experimental data (gray). As representative examples we again show results with the A12 protein force field combined with either **(A)** TIP4P-D or **(B)** TIP3P water models, and where the location of the spin label probe is denoted in each plot. Experimental intensity ratios slightly exceeding the value 1 were set to 1 in these plots. **(C,D)** We also calculated the RMSD between the experimental and calculated intensity ratios for each probe and the two force fields both before and after reweighting. **(E)** Finally, we calculated the RMSD between experiment and calculated values over all probe position for and all force fields in **Table 1**.

for these ensembles, and in the case of the metainference ensemble we even observe a small worsening of the agreement.

Comparison of Ensembles

An important question is whether and how much ensembles become more similar to one another after reweighting using experimental data. Clearly, the properties of the final ensembles reflect information both in the prior and in the experimental data. Previously we and others have shown that experimental data make ensembles more similar to one another (Lindorff-Larsen and Ferkinghoff-Borg, 2009; Camilloni et al., 2012; Tiberti et al., 2015; Larsen et al., 2020), though the extent to which this occurs depends on how the ensembles are compared.

The results described above suggest that the description of the level of compaction indeed becomes more similar after reweighting, and this is reflected also in more similar distribution of the radius of gyration (**Supplementary Figure 19**). Nevertheless, it is also clear that differences remain, in particular when the prior gives a very poor description of the data. A more complex situation arises when the ensembles are compared using properties that are only little correlated with those probed by the SAXS experiments, such as for example local (secondary) structure. We therefore used STRIDE (Frishman and Argos, 1995) to calculate the secondary structure in all ensembles, both before and after reweighting with the SAXS data (**Supplementary Figures 19**, **20**). As also previously shown (Robustelli et al., 2018) there is little transient helical structure in these simulations, though with some variation across force fields. Previous analyses suggest that compaction and secondary structure are only weakly coupled in disordered proteins

(Piana et al., 2012; Crehuet et al., 2019; Zerze et al., 2019), and indeed we in general find that reweighting against the SAXS data only has a modest effect on the secondary structure. The M&M simulations, however, do not follow this pattern, but we note here that in contrast to the other simulations, these are two independent simulations. In summary, these analyses demonstrate that inclusion of experimental restraints make ensembles more similar in some properties, but not necessarily in others. Reweighting against a set of experimental data will thus only affect properties that affect, or are otherwise coupled to, the experimental data. As argued previously (Crehuet et al., 2019), this also means that cross-validation is only useful when using types of experiments that probe related molecular properties.

CONCLUSIONS

We have employed "on-the-fly" or "post-facto" integration of MD simulations and SAXS data *αSN* to derive structural ensembles that are in improved agreement with experiments. These approaches take their outset in a Bayesian framework, and thus the results of the posterior distribution may depend on the choice of the prior. Our results clearly show, in line with previous observations (Larsen et al., 2020), that if the prior distribution is a poor model for the experimental data, reweighting becomes noisy. Despite this we find that fitting against SAXS data generally improved or had no effect on the agreement with NMR data (R_h and PREs) that were not target of the optimization. Thus, the inclusion of a SAXS-restraint in the metainference simulation and the BME refinement showed that both methods were able to generate a reliable and heterogenous ensemble that maintained good agreement with independent experimental data. We nevertheless also find that the prior used in such protocols are important, and that more robust analyses are obtained with the best priors.

Our results also reflect an important point when including experimental data to refine ensembles, namely that the ensembles will only be affected along degrees of freedom that are sensitive to the experiments (or vice versa). Thus, as shown by our analyses, while the level of compaction ($p(R_g)$) becomes more similar after inclusion of the SAXS data, this is not the case for the description of the secondary structure. In order to improve the description of both global and local structure one thus needs to include data sensitive to both properties, either individually (such as SAXS and chemical shifts) or combined such as residual dipolar couplings.

Our calculations of R_h and PREs suggest that when the ensembles are "far" away from the experimental data, then improvements driven by the SAXS refinement lead to clear improvements in independent parameters. For ensembles that show better agreement between with the SAXS data to begin with, the picture is less clear. While we on average observe improvements, they are often modest. While some of this is likely because the ensembles are already in reasonably good agreement with the experiment, we also suggest that we are observing the limitations of the forward models for calculating SAXS, R_h and PREs. In particular, we suggest that more research is needed on comparing the accuracy and domains of applicability of existing methods for calculating R_h (Kirkwood and Riseman, 1948; de la Torre et al., 2000; Nygaard et al., 2017; Fleming and Fleming, 2018). Methods for calculating SAXS data (Henriques et al., 2018; Hub, 2018), however, also require choices to be made for how to deal with solvent effects, and calculations of PREs rely on models and parameters to describe effects of dynamics (Tesei et al., 2020). In all cases, further work is needed to make it possible to extract as much as possible information from the data, such as for example the independent information about the moments of the R_g-distribution contained within the SAXS and NMR diffusion measurements (Choy et al., 2002; Ahmed et al., 2020).

Thus, we conclude that in order to obtain improved descriptions of the conformational ensembles of disordered proteins, work is needed in several areas. First, improved force fields and sampling methods give rise to better initial estimates that require less (or no) reweighting. Second, refinement should ideally use data from experiments that are sensitive to as many conformational properties as possible, and at least those that probe the properties of interest. Finally, improved and consistent forward models are required to use this data to provide better models for intrinsically disordered proteins. Importantly, these different aspects work in synergy as accurate prior ensembles are more robust toward reweighting, and that accurate forward models make it possible to extract more information from the experimental data.

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 at: https://github.com/KULL-Centre/papers/blob/master/2021/aSYN-ahmed-et-al/, https://www.plumed-nest.org/eggs/21/003/.

AUTHOR CONTRIBUTIONS

MCA analyzed and performed MD simulations, analyzed the data, wrote the first draft, and made figures. LKS purified αSN, and performed and analyzed SAXS data together with AEL. AJ and CC developed the simulation procedure with MCA, and aided in metainference simulations. EAN purified αSN, and performed and analyzed NMR data together with BBK. KL-L designed the research, supervised MCA, analyzed the data, and revised the article. All authors contributed to the article and approved the submitted version.

FUNDING

We acknowledge support by a grant from the Lundbeck Foundation to the BRAINSTRUC Structural Biology Initiative (R155-2015-2666). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank A. Kikhney and C. Jeffries for assistance during data collection at the P12 SAXS beamline. We thank D. E. Shaw Research for sharing the molecular dynamics trajectories.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.654333/full#supplementary-material

REFERENCES

Abascal, J. L., and Vega, C. (2005). A general purpose model for the condensed phases of water: Tip4p/2005. *J. Chem. Phys*. 123:234505. doi: 10.1063/1.2121687

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1, 19–25. doi: 10.1016/j.softx.2015.06.001

Ahmed, M. C., Crehuet, R., and Lindorff-Larsen, K. (2020). Computing, analyzing, and comparing the radius of gyration and hydrodynamic radius in conformational ensembles of intrinsically disordered proteins. *Methods Mol. Biol*. 2141, 429–445. doi: 10.1007/978-1-0716-0524-0_21

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett*. 100:020603. doi: 10.1103/PhysRevLett.100.020603

Bernado, P., and Svergun, D. I. (2012). Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst*. 8, 151–167. doi: 10.1039/C1MB05275F

Best, R. B., and Hummer, G. (2009). Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B* 113, 9004–9015. doi: 10.1021/jp901540t

Best, R. B., Zheng, W., and Mittal, J. (2014). Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput*. 10, 5113–5124. doi: 10.1021/ct500569b

Best, R. B., Zhu, X., Shim, J., Lopes, P. E., Mittal, J., Feig, M., et al. (2012). Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *J. Chem. Theory Comput*. 8, 3257–3273. doi: 10.1021/ct300400x

Blanchet, C. E., Spilotros, A., Schwemmer, F., Graewert, M. A., Kikhney, A., Jeffries, C. M., et al. (2015). Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *J. Appl. Crystallogr*. 48, 431–443. doi: 10.1107/S160057671500254X

Bonomi, M., Bussi, G., Camilloni, C., and Tribello, G. A. (2019). Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* 16, 670–673. doi: 10.1038/s41592-019-0506-8

Bonomi, M., and Camilloni, C. (2017). Integrative structural and dynamical biology with PLUMED-ISDB. *Bioinformatics* 33, 3999–4000. doi: 10.1093/bioinformatics/btx529

Bonomi, M., Camilloni, C., Cavalli, A., and Vendruscolo, M. (2016a). Metainference: a bayesian inference method for heterogeneous systems. *Sci. Adv*. 2:e1501177. doi: 10.1126/sciadv.1501177

Bonomi, M., Camilloni, C., and Vendruscolo, M. (2016b). Metadynamic metainference: enhanced sampling of the metainference ensemble using metadynamics. *Sci. Rep*. 6:31232. doi: 10.1038/srep31232

Bottaro, S., Bengtsen, T., and Lindorff-Larsen, K. (2020). Integrating molecular simulation and experimental data: a bayesian/maximum entropy reweighting approach. *Methods Mol. Biol*. 2112, 219–240. doi: 10.1007/978-1-0716-0270-6_15

Bottaro, S., Lindorff-Larsen, K., and Best, R. B. (2013). Variational optimization of an all-atom implicit solvent force field to match explicit solvent simulation data. *J. Chem. Theory Comput*. 9, 5641–5652. doi: 10.1021/ct400730n

Box, G. E., and Tiao, G. C. (2011). *Bayesian Inference in Statistical Analysis*, Vol. 40. Hoboken, NJ: John Wiley & Sons.

Camilloni, C., Robustelli, P., Simone, A. D., Cavalli, A., and Vendruscolo, M. (2012). Characterization of the conformational equilibrium between the two major substates of rnase a using NMR chemical shifts. *J. Am. Chem. Soc*. 134, 3968–3971. doi: 10.1021/ja210951z

Cesari, A., Reißer, S., and Bussi, G. (2018). Using the maximum entropy principle to combine simulations and solution experiments. *Computation* 6:15. doi: 10.3390/computation6010015

Chemes, L. B., Alonso, L. G., Noval, M. G., and de Prat-Gay, G. (2012). Circular dichroism techniques for the analysis of intrinsically disordered proteins and domains. *Methods Mol. Biol*. 895, 387–404. doi: 10.1007/978-1-61779-927-3_22

Choy, W.-Y., Mulder, F. A., Crowhurst, K. A., Muhandiram, D., Millett, I. S., Doniach, S., et al. (2002). Distribution of molecular size within an unfolded state ensemble using small-angle X-ray scattering and pulse field gradient NMR techniques. *J. Mol. Biol*. 316, 101–112. doi: 10.1006/jmbi.2001.5328

Crehuet, R., Buigues, P. J., Salvatella, X., and Lindorff-Larsen, K. (2019). Bayesian-maximum-entropy reweighting of IDP ensembles based on NMR chemical shifts. *Entropy* 21:898. doi: 10.3390/e21090898

Das, R. K., Ruff, K. M., and Pappu, R. V. (2015). Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol*. 32, 102–112. doi: 10.1016/j.sbi.2015.03.008

de la Torre, J. G., Huertas, M. L., and Carrasco, B. (2000). Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys. J*. 78, 719–730. doi: 10.1016/S0006-3495(00)76630-6

Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M., and Dobson, C. M. (2005). Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J. Am. Chem. Soc*. 127, 476–477. doi: 10.1021/ja044834j

Dyson, H. J., and Wright, P. E. (2001). Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Methods Enzymol*. 339, 258–270. doi: 10.1016/S0076-6879(01)39317-5

Eliezer, D. (2009). Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol*. 19, 23–30. doi: 10.1016/j.sbi.2008.12.004

Fleming, P. J., and Fleming, K. G. (2018). Hullrad: fast calculations of folded and disordered protein and nucleic acid hydrodynamic properties. *Biophys. J*. 114, 856–869. doi: 10.1016/j.bpj.2018.01.002

Frishman, D., and Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins* 23, 566–579. doi: 10.1002/prot.340230412

Goga, N., Rzepiela, A., De Vries, A., Marrink, S., and Berendsen, H. (2012). Efficient algorithms for Langevin and DPD dynamics. *J. Chem. Theory Comput*. 8, 3637–3649. doi: 10.1021/ct3000876

Grudinin, S., Garkavenko, M., and Kazennov, A. (2017). PEPSI-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallogr. D* 73, 449–464. doi: 10.1107/S2059798317005745

Hansen, P. C., and O'Leary, D. P. (1993). The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput*. 14, 1487–1503. doi: 10.1137/0914086

Henriques, J., Arleth, L., Lindorff-Larsen, K., and Skepö, M. (2018). On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations. *J. Mol. Biol*. 430, 2521–2539. doi: 10.1016/j.jmb.2018.03.002

Hermann, M. R., and Hub, J. S. (2019). SAXS-restrained ensemble simulations of intrinsically disordered proteins with commitment to the principle of maximum entropy. *J. Chem. Theory Comput*. 15, 5103–5115. doi: 10.1021/acs.jctc.9b00338

Hess, B., Bekker, H., Berendsen, H. J., and Fraaije, J. G. (1997). Lincs: a linear constraint solver for molecular simulations. *J. Comput. Chem*. 18, 1463–1472. doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., et al. (2004). Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *J. Chem. Phys*. 120, 9665–9678. doi: 10.1063/1.1683075

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple amber force fields and development

of improved protein backbone parameters. *Proteins* 65, 712–725. doi: 10.1002/prot.21123

Hub, J. S. (2018). Interpreting solution X-ray scattering data using molecular simulations. *Curr. Opin. Struct. Biol.* 49, 18–26. doi: 10.1016/j.sbi.2017.11.002

Hummer, G., and Köfinger, J. (2015). Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* 143, 12B634_1. doi: 10.1063/1.4937786

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.* 106:620. doi: 10.1103/PhysRev.106.620

Jorgensen, W. L. (1981). Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *J. Am. Chem. Soc.* 103, 335–340. doi: 10.1021/ja00392a016

Jussupow, A., Messias, A. C., Stehle, R., Geerlof, A., Solbak, S. M., Paissoni, C., et al. (2020). The dynamics of linear polyubiquitin. *Sci. Adv.* 6:eabc3786. doi: 10.1126/sciadv.abc3786

Kirkwood, J. G., and Riseman, J. (1948). The intrinsic viscosities and diffusion constants of flexible macromolecules in solution. *J. Chem. Phys.* 16, 565–573. doi: 10.1063/1.1746947

Laio, A., and Parrinello, M. (2002). Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* 99, 12562–12566. doi: 10.1073/pnas.202427399

Larsen, A. H., Wang, Y., Bottaro, S., Grudinin, S., Arleth, L., and Lindorff-Larsen, K. (2020). Combining molecular dynamics simulations with small-angle X-ray and neutron scattering data to study multi-domain proteins in solution. *PLoS Comput. Biol.* 16:e1007870. doi: 10.1371/journal.pcbi.1007870

Lazaridis, T., and Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* 35, 133–152. doi: 10.1002/(SICI)1097-0134(19990501)35:2<133::AID-PROT1>3.0.CO;2-N

LeBlanc, S., Kulkarni, P., and Weninger, K. (2018). Single molecule FRET: a powerful tool to study intrinsically disordered proteins. *Biomolecules* 8:140. doi: 10.3390/biom8040140

Lindorff-Larsen, K., and Ferkinghoff-Borg, J. (2009). Similarity measures for protein ensembles. *PLoS ONE* 4:e4203. doi: 10.1371/journal.pone.0004203

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

Löhr, T., Jussupow, A., and Camilloni, C. (2017). Metadynamic metainference: convergence towards force field independent structural ensembles of a disordered peptide. *J. Chem. Phys.* 146:165102. doi: 10.1063/1.4981211

Nerenberg, P. S., Jo, B., So, C., Tripathy, A., and Head-Gordon, T. (2012). Optimizing solute-water van der Waals interactions to reproduce solvation free energies. *J. Phys. Chem. B* 116, 4524–4534. doi: 10.1021/jp2118373

Niebling, S., Björling, A., and Westenhoff, S. (2014). Martini bead form factors for the analysis of time-resolved X-ray scattering of proteins. *J. Appl. Crystallogr.* 47, 1190–1198. doi: 10.1107/S1600576714009959

Nygaard, M., Kragelund, B. B., Papaleo, E., and Lindorff-Larsen, K. (2017). An efficient method for estimating the hydrodynamic radius of disordered protein conformations. *Biophys. J.* 113, 550–557. doi: 10.1016/j.bpj.2017.06.042

Orioli, S., Larsen, A. H., Bottaro, S., and Lindorff-Larsen, K. (2020). How to learn from inconsistencies: integrating molecular simulations with experimental data. *Prog. Mol. Biol. Transl. Sci.* 170, 123–176. doi: 10.1016/bs.pmbts.2019.12.006

Paissoni, C., Jussupow, A., and Camilloni, C. (2019). Martini bead form factors for nucleic acids and their application in the refinement of protein-nucleic acid complexes against SAXS data. *J. Appl. Crystallogr.* 52, 394–402. doi: 10.1107/S1600576719002450

Paissoni, C., Jussupow, A., and Camilloni, C. (2020). Determination of protein structural ensembles by hybrid-resolution SAXS restrained molecular dynamics. *J. Chem. Theory Comput.* 16, 2825–2834. doi: 10.1021/acs.jctc.9b01181

Panjkovich, A., and Svergun, D. I. (2018). Chromixs: automatic and interactive analysis of chromatography-coupled small angle X-ray scattering data. *Bioinformatics* 34, 1944–1946. doi: 10.1093/bioinformatics/btx846

Pfaendtner, J., and Bonomi, M. (2015). Efficient sampling of high-dimensional free-energy landscapes with parallel bias metadynamics. *J. Chem. Theory Comput.* 11, 5062–5067. doi: 10.1021/acs.jctc.5b00846

Piana, S., Donchev, A. G., Robustelli, P., and Shaw, D. E. (2015). Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* 119, 5113–5123. doi: 10.1021/jp508971m

Piana, S., Lindorff-Larsen, K., Dirks, R. M., Salmon, J. K., Dror, R. O., and Shaw, D. E. (2012). Evaluating the effects of cutoffs and treatment of long-range electrostatics in protein folding simulations. *PLoS ONE* 7:e39918. doi: 10.1371/journal.pone.0039918

Piana, S., Lindorff-Larsen, K., and Shaw, D. E. (2011). How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* 100, L47–L49. doi: 10.1016/j.bpj.2011.03.051

Prestel, A., Bugge, K., Staby, L., Hendus-Altenburger, R., and Kragelund, B. B. (2018). Characterization of dynamic IDP complexes by NMR spectroscopy. *Methods Enzymol.* 611, 193–226. doi: 10.1016/bs.mie.2018.08.026

Raiteri, P., Laio, A., Gervasio, F. L., Micheletti, C., and Parrinello, M. (2006). Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B* 110, 3533–3539. doi: 10.1021/jp054359r

Robustelli, P., Piana, S., and Shaw, D. E. (2018). Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* 115, E4758–E4766. doi: 10.1073/pnas.1800690115

Skaanning, L. K., Santoro, A., Skamris, T., Martinsen, J. H., D'Ursi, A. M., Bucciarelli, S., et al. (2020). The non-fibrillating N-terminal of α-synuclein binds and co-fibrillates with heparin. *Biomolecules* 10:1192. doi: 10.3390/biom10081192

Snead, D., and Eliezer, D. (2019). Intrinsically disordered proteins in synaptic vesicle trafficking and release. *J. Biol. Chem.* 294, 3325–3342. doi: 10.1074/jbc.REV118.006493

Song, D., Luo, R., and Chen, H.-F. (2017). The idp-specific force field FF14IDPSFF improves the conformer sampling of intrinsically disordered proteins. *J. Chem. Inform. Model.* 57, 1166–1178. doi: 10.1021/acs.jcim.7b00135

Spillantini, M. G., and Goedert, M. (2000). The α-synucleinopathies: Parkinson's disease, dementia with lewy bodies, and multiple system atrophy. *Ann. N. Y. Acad. Sci.* 920, 16–27. doi: 10.1111/j.1749-6632.2000.tb06900.x

Sugita, Y., and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314, 141–151. doi: 10.1016/S0009-2614(99)01123-9

Sun, Y., and Kollman, P. A. (1995). Hydrophobic solvation of methane and nonbond parameters of the TIP3P water model. *J. Comput. Chem.* 16, 1164–1169. doi: 10.1002/jcc.540160910

Tesei, G., Martins, J. M., Kunze, M. B., Wang, Y., Crehuet, R., and Lindorff-Larsen, K. (2020). Deer-predict: software for efficient calculation of spin-labeling EPR and NMR data from conformational ensembles. *bioRxiv*. doi: 10.1101/2020.08.09.243030

Tiberti, M., Papaleo, E., Bengtsen, T., Boomsma, W., and Lindorff-Larsen, K. (2015). Encore: software for quantitative ensemble comparison. *PLoS Comput. Biol.* 11:e1004415. doi: 10.1371/journal.pcbi.1004415

Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. (2014). PLUMED 2: new feathers for an old bird. *Comput. Phys. Commun.* 185, 604–613. doi: 10.1016/j.cpc.2013.09.018

Ulmer, T. S., Bax, A., Cole, N. B., and Nussbaum, R. L. (2005). Structure and dynamics of Micelle-bound human α-synuclein. *J. Biol. Chem.* 280, 9595–9603. doi: 10.1074/jbc.M411805200

Ulusoy, A., and Di Monte, D. A. (2013). α-Synuclein elevation in human neurodegenerative diseases: experimental, pathogenetic, and therapeutic implications. *Mol. Neurobiol.* 47, 484–494. doi: 10.1007/s12035-012-8329-y

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* 18, 343–384. doi: 10.1002/jmr.747

van Maarschalkerweerd, A., Vetri, V., Langkilde, A. E., Foderà, V., and Vestergaard, B. (2014). Protein/lipid coaggregates are formed during α-synuclein-induced disruption of lipid bilayers. *Biomacromolecules* 15, 3643–3654. doi: 10.1021/bm500937p

Wilkins, D. K., Grimshaw, S. B., Receveur, V., Dobson, C. M., Jones, J. A., and Smith, L. J. (1999). Hydrodynamic radii of native and denatured proteins measured by pulse field gradient nmr techniques. *Biochemistry* 38, 16424–16431. doi: 10.1021/bi991765q

Wu, D., Chen, A., and Johnson, C. S. (1995). An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses. *J. Magn. Reson. A* 115, 260–264. doi: 10.1006/jmra.1995.1176

Zerze, G. H., Zheng, W., Best, R. B., and Mittal, J. (2019). Evolution of all-atom protein force fields to improve local and global properties. *J. Phys. Chem. Lett.* 10, 2227–2234. doi: 10.1021/acs.jpclett.9b00850

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: 30 April 2021
doi: 10.3389/fmolb.2021.666626

Computational Identification of a Putative Allosteric Binding Pocket in TMPRSS2

Jacopo Sgrignani[1]* *and Andrea Cavalli*[1,2]*

[1] Institute for Research in Biomedicine, Università della Svizzera Italiana, Bellinzona, Switzerland, [2] Swiss Institute of Bioinformatics, Lausanne, Switzerland

Camostat, nafamostat, and bromhexine are inhibitors of the transmembrane serine protease TMPRSS2. The inhibition of TMPRSS2 has been shown to prevent the viral infection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and other viruses. However, while camostat and nafamostat inhibit TMPRSS2 by forming a covalent adduct, the mode of action of bromhexine remains unclear. TMPRSS2 is autocatalytically activated from its inactive form, zymogen, through a proteolytic cleavage that promotes the binding of Ile256 to a putative allosteric pocket (A-pocket). Computer simulations, reported here, indicate that Ile256 binding induces a conformational change in the catalytic site, thus providing the atomistic rationale to the activation process of the enzyme. Furthermore, computational docking and molecular dynamics simulations indicate that bromhexine competes with the N-terminal Ile256 for the same binding site, making it a potential allosteric inhibitor. Taken together, these findings provide the atomistic basis for the development of more selective and potent TMPRSS2 inhibitors.

Keywords: TMPRSS2 protein, molecular modeling, allosteric pocket, docking, MD simulation

OPEN ACCESS

Edited by:
Massimiliano Bonomi, Institut Pasteur, France

Reviewed by:
Therese E. Malliavin, Institut Pasteur, France
Matteo Masetti, University of Bologna, Italy

***Correspondence:**
Jacopo Sgrignani
jacopo.sgrignani@irb.usi.ch
Andrea Cavalli
andrea.cavalli@irb.usi.ch

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

Received: *10 February 2021*
Accepted: *01 April 2021*
Published: *30 April 2021*

Citation:
Sgrignani J and Cavalli A (2021) Computational Identification of a Putative Allosteric Binding Pocket in TMPRSS2. Front. Mol. Biosci. 8:666626. doi: 10.3389/fmolb.2021.666626

INTRODUCTION

Since the early days of the pandemic coronavirus disease 2019 (COVID-19) started from the Chinese city of Wuhan, Hubei province, in December 2019, many reports highlighted the crucial role of transmembrane serine protease 2 (TMPRSS2) in the spread and progression of the viral infection (Hoffmann et al., 2020; Sungnak et al., 2020). TMPRSS2 has been identified as one of the proteases responsible for the proteolytic priming of SARS-CoV-2 spike protein which leads to the release of the fusion peptide. In addition to that, TMPRSS2 has been put in relation with the spread of other viruses, such as influenza A viruses, severe acute respiratory syndrome coronavirus 2 (SARS-CoV), and Middle East respiratory syndrome coronavirus (MERS-CoV), and it has been studied as a potential therapeutic target for prostate cancer therapy (Lucas et al., 2014; Shen et al., 2017). Finally, as TMPRSS2 expression is regulated by the androgen receptor, it has been hypothesized that its crucial role in the viral infection might help explain why males have more frequently severe complications and a worse clinical outcome than females and if androgen deprivation therapy (ADT) can have a protective effect against SARS-CoV-2 infection (Montopoli et al., 2020). These observations stimulated intense investigations, and the number of papers with the TMPRSS2 keyword in the title indexed in PubMed during 2020 raised from an average of 80–100/year to 601.

TMPRSS2 is a membrane protein belonging to the type II transmembrane serine protease (TTSP) family. It is functionally classified as a trypsin-like protease (TLP). Like other serine proteases, TMPRSS2 cleaves peptide bonds that are present after positively charged residues (lysine or arginine), and its enzymatic activity depends on the presence of a catalytic triad formed by His296, Asp345, and Ser441. The catalytic selectivity is achieved with the presence of a negatively charged Asp residue at the bottom of a cavity usually indicated as "S1 specificity pocket" (Laporte and Naesens, 2017; Singh et al., 2020).

Structurally, TPMRSS2 is characterized by the presence of a cytoplasmic N-terminal domain, a transmembrane helical domain, and three extracellular domains: low-density lipoprotein (LDL)-receptor class A domain, scavenger receptor cysteine-rich (SRCR) domain, and the peptidase S1 domain, also called serine protease domain (SPD) (**Figure 1A**).

An autocatalytic cleavage between Arg255 and Ile256 activates the 492-residue long TMPRSS2 zymogen. This modification enables the binding of Ile256 into a putative allosteric pocket (A-pocket), which induces a conformational rearrangement of the catalytic site (Bertram et al., 2010). After the cleavage, membrane TLPs, such as TMPRSS2, remain bound to the transmembrane N-terminal domains by a conserved disulfide bond, although a small fraction of the protein can be detected into the extracellular milieu (Szabo et al., 2003; Pászti-Gere et al., 2016; Shen et al., 2017).

Two different species are reported in the literature, one with a mass of ∼55 kDa that corresponds to the full-length protein and one of ∼30 kDa which represents the SPD released in the extracellular space if the disulfide bond is not formed (Afar et al., 2001; Chen et al., 2010).

To date, no atomistic structure of the entire TMPRSS2, or the SPD, is available. However, important information can be derived from the structure of homologous proteins such as matriptase, DESC1, and several kallikreins.

Several inhibitors of TMPRSS2 have been identified in the last years. These include organic compounds such as camostat, nafamostat, and bromhexine (BH) (**Figure 1B**) and peptidomimetics (Meyer et al., 2013; Lucas et al., 2014; Shen et al., 2017; Bestle et al., 2020; Hoffmann et al., 2020; Zang et al., 2020). Of particular note is BH, a component of widely used medicaments against respiratory disorders characterized by viscid or excessive mucus. In fact, following the report of a selective TMPRSS2 inhibition by Lucas et al. (2014), the use of BH for the prevention and therapy of the SARS-CoV-2 infection has been hypothesized (Depfenhart et al., 2020; Habtemariam et al., 2020; Maggio and Corsini, 2020). However, to date, only a limited number of clinical trials have been carried out, and their results remain inconclusive (Ansarin et al., 2020; Li et al., 2020).

In this work, we used computational and experimental methods, such as homology modeling, molecular docking, molecular dynamics (MD), and microscale thermophoresis (MST), to investigate the structure and dynamics of TMPRSS2 and clarify its activation mechanism and the interaction with various inhibitors at an atomistic level of details. We focused, in particular, on the differences in the mode of action of camostat/nafamostat and BH. In fact, while camostat and nafamostat inhibit TMPRSS2 by forming a covalent adduct, the mode of action of BH remains unclear.

Besides the generation of a reliable model of the TMPRSS2 catalytic domain, the results of our investigations confirmed that both camostat and nafamostat are competitive inhibitors efficiently binding the active site. Contrarily, they indicated that the binding of BH to the active site is unlikely, leading us to the identification of a putative allosteric binding pocket.

MATERIALS AND METHODS

Homology Modeling

An atomistic model of the SPD of TMPRSS2 (UniProt code O15393), covering residues from Ile256 to Gly492, was generated by homology modeling. The most suitable templates were identified using the SWISS-MODEL webserver (Waterhouse et al., 2018). This search provided three templates [Protein Data Bank (PDB) codes: 5F8T, 5CE1, and 6O1G], having a sufficient degree of similarity (between 38 and 41%), thus well-suited for an accurate model generation (Xiang, 2006; Cavasotto and Phatak, 2009; Sgrignani et al., 2018). The alignment between the target sequence and the templates was performed using the Prime-STA algorithm, included in the Schrodinger suite for molecular modeling (Schrodinger Suite 2020-1). This algorithm, in addition to the sequence alignment, considers secondary structure matching, providing better alignments also in poorly conserved regions. Next, 10 models were generated for each of the three templates using PRIME, keeping small ligand molecules, such as piperidine-1-carboximidamide (PC1), 2-[6-(1-hydroxycyclohexyl)pyridin-2-yl]-1H-indole-5-carboximidamide (HCP), and N-[(6-amino-2,4-dimethylpyridin-3-yl)methyl]-1-({4-[(1H-pyrazol-1-yl)methyl]phenyl}methyl)-1H-pyrazole-4-carboxamide (7SD) (**Figure 1**), bound to the protein active site in the templates 5F8T, 5CE1, and 6O1G, to preserve their respective conformations.

Finally, the models (subsequently indicated as M-5F8T, M-5CE1, and M-6O1G) with the lowest OPLS3e (Harder et al., 2016) potential energy after minimization were selected for the subsequent calculations.

Molecular Dynamics Simulations

Atomistic models were prepared for MD simulation with the following protocol: (1) the PROPKA program was used to assign the residue protonation state at a reference pH of 7.4 (Olsson et al., 2011) and (2) the structures were solvated in a box of water with a minimal distance from the protein surface of 10 Å. A proper number of counterions were added to the systems to ensure charge neutrality. All the non-solvent molecules were parametrized using the OPLS3 (Harder et al., 2016) force field, while TIP3P model (Jorgensen et al., 1983) was used for water molecules.

Before the MD production runs, the following simulation protocol was used to equilibrate the systems: (1) Brownian dynamics was run for 100 ps in an NVT ensemble (T = 10 K) applying harmonic restraints on solute heavy atoms (force constant 50 kcal/mol/Å^2); (2) NVT (T = 10 K) MD simulation

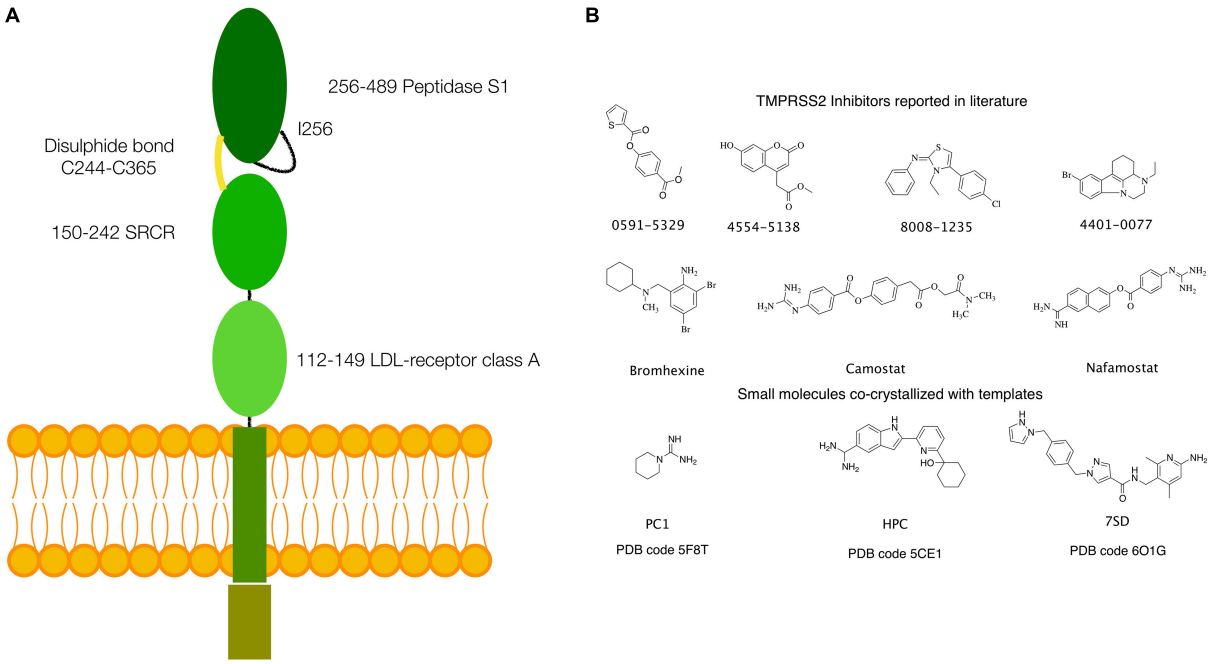


FIGURE 1 | (A) Schematics of the structure of TMPRSS2. **(B)** Small molecules with inhibitory activity on TMPRSS2 reported in the literature and small molecules or fragments co-crystallized in the S1 specificity pocket in the templates used for homology modeling.

of 12 ps in NVT ensemble conserving the same restraints applied in (1); (3) NPT (T = 300 K and P = 1 atm) MD simulation (12 ps) conserving the same restraints applied in (1); and (4) NPT (T = 300 K and P = 1 atm) MD simulation (24 ps) without restraints. The pressure and the temperature were fixed at 300 K and 1 atm by the Martyna–Tobias–Klein barostat (Martyna et al., 1994) and the Nosé–Hoover chain thermostat (Martyna et al., 1992), respectively. All the simulations were run using GPU accelerated DESMOND code. A summary of the simulations run in this work is reported in **Table 1**.

Root mean square deviation (RMSD), root mean square fluctuation (RMSF), and radius of gyration (Rg) analysis were computed using Maestro (Schrodinger Suite 2020-1). Cluster analysis was performed with the program TTClust (Tubiana et al., 2018), focusing on residues belonging to the catalytic site, namely, Cys281, Thr293, Ala294, Ala295, His296, Cys297, Val298, Glu299, Tyr337, Asp338, Ser339, Lys342, Asn343, Ans344, Asp345, Ile346, Ala347, Met424, Cys437, Gln438, Asp440, Ser441, Asp458, Thr459, Ser460, Trp461, and Phe480. Contrarily, the analysis of the loop that regulates the access to the S1 specificity pocket was performed considering all the residues between Gly462 and Val473. The optimal number of clusters was automatically determined using the "elbow" method with k-means (Tibshirani et al., 2001).

Computational Docking of TMPRSS2 Ligands

Computational docking was performed using the software GLIDE (Friesner et al., 2004). The analysis of the structural parameters and the analysis of MD simulations (see section "Results and Discussion") indicated M-5FT8 as having a higher quality and more stable among the generated models.

In analogy to the previous studies (Amaro et al., 2008, 2018; Sgrignani et al., 2009), to account for target flexibility, snapshots from MD simulations of M-5FT8 were selected using the previously described cluster analysis. In particular, four snapshots were selected from the simulations run with positively charged His296 and four from the simulations with His296 protonated on the ε nitrogen (see also section "Results and Discussion").

The grids for docking were centered in the geometric center of all the atoms of the three residues forming the catalytic triads (His296, Asp345, and Ser441). A distinct grid file was generated for all selected snapshots.

Contrarily, for the docking of BH in the putative site predicted by Sitemap, the grid was centered using the corresponding sitepoints. In this context, sitepoints are points in a grid, contiguous, or bridged by short gaps in exposed regions, that define the shape of a putative binding site (Halgren, 2007).

All docking calculations were performed using the standard precision (SP) protocol and GlideScore. Furthermore, docking was performed on all selected snapshots, and, finally, the pose with the best GlideScore, together with the receptor, was saved for the analysis and MD simulations.

The structures of the small molecule ligands were prepared with LIGPREP. In the case of BH, the results indicated a protonation of the ternary amino group; therefore, both enantiomeric molecules (S and R) were considered in docking calculations, but only the complex with best GlideScore was used in MD simulations.

TABLE 1 | Summary of the performed molecular dynamics.

Description of the system	Number of independent simulations	Simulation length (ns)	Ligand
M-5F8T	3 + 3 (His296 protonated on the ε nitrogen)	250	
M-5CE1	3	250	
M-6O1G	3	250	
M-5F8T	1 + 1 (His296 protonated on the ε nitrogen) for both camostat and nafamostat	500	Camostat and nafamostat
M-5F8T	3	2 × 1,000 1 × 500 (the complex decomposed)	BH in Site_1
M-5F8T	1	100	BH in Site_2
Apo-C-M-5F8T	1	1,000	
Apo-M-5F8T	1	1,000	
C-M-5F8T	3	500	(R)-BH in the A-pocket
C-M-5F8T	3	500	(S)-BH in the A-pocket
C-M-5F8T	3	500	(R)-BH in the A-pocket (IFD docking)
C-M-5F8T	3	500	(S)-BH in the A-pocket (IFD docking)

The models from different templates were indicated as M-5F8T, M-5CE1, and M-6O1G. The suffix C indicates the models where the first two residues (Ile256 and Val257) were deleted.

Docking of BH in the A-pocket (see section "Results and Discussion" for a definition) was performed using a representative structure of the open and closed conformations (**Figures 2B–E**) sampled during the MD simulations of C-M-5F8T (see section "Results and Discussion"). In this case, the grid was centered in the COG of the residues Ile381, Ser382, Gly383, Gly385, Ala386, Thr387, Glu388, Asn398, Ala399, Ala400 Asn433, Val434, Asp435, Ser436, Cys437, Asp440, Cys465, and Ala466.

The results of these calculations showed a better GlideScore for the complex in the closed conformation (∼−3.0 kcal/mol vs. ∼ −4.8 kcal/mol for open and closed conformations). However, also in this case, the complex with the best GlideScore dissociated during MD simulations.

Regarding this point, it is important to notice that in M-5F8T, the S-pocket is occupied by a small aminoacidic tail. It is, therefore, reasonable to assume that a side chain rearrangement is needed to accommodate different ligands.

Consequently, the docking was performed again using the induced-fit docking (IFD) protocol of GLIDE, with default input parameters. In particular, only the orientations of the side chains of the residues within a distance of 5 Å from the ligand were optimized. Finally, the complex with the lowest IFD score [a specific score that combines GlideScore, Glide_Ecoul energy, and Prime protein conformation energy (Sherman et al., 2006)] was selected as the best model and used in MD simulations.

Prediction of Putative Allosteric Binding Sites

Several algorithms to detect allosteric pockets in proteins have been developed in the last years (Halgren, 2007, 2009; Yu et al., 2010; Panjkovich and Daura, 2014; Kozakov et al., 2015; Jiménez et al., 2017; Xu et al., 2018; Guarnera and Berezovsky, 2019). Sitemap, proposed by Halgren in 2007 (Halgren, 2007, 2009) and implemented in the Schrodinger suite for molecular modeling, is among the most widely used. Furthermore, it provides a clear assessment of the druggability of the identified pockets.

Consequently, we used this algorithm to investigate the presence of allosteric pockets in both M-5FT8 and C-M-5F8T. All the calculations were performed using default values provided by the Maestro interface (Schrodinger Suite 2020-1). In addition to that, to validate these results, the same structures were analyzed also with other algorithms (PARS, Deepsite, and FTMap), using the respective webservers[1,2,3].

MST Experiments for the TMPRSS2/BH Binding

The binding affinity between TMPRSS2 and BH was measured by MST. Recombinant human TMPRSS2 (106-492aa, 6xHisTag) was acquired from Cusabio (CSB-YP023924HU) and labeled using a His-tag-specific dye (Monolith His-Tag Labeling Kit RED-tris-NTA (MO-L018), NanoTemper® Technologies GmbH, München, Germany), according to manufacturer instructions. A fixed concentration of the labeled TMPRSS2 (5 nM) was mixed with 16 1:1 serial dilution of BH. MST measurements were performed using premium-coated capillary tubes on a NanoTemper instrument.

BH was first dissolved in DMSO at a 5 mM concentration. In all subsequent experiments, both protein and BH were dissolved in Dulbecco's Phosphate-Buffered Saline (PBS; D8537, Sigma Aldrich, Saint Louis, MO, United States).

Two independent experiments were performed to compute the Kd values. Data were analyzed with the NanoTemper analysis software MO.Affinity Analysis (v. 2.3). Kd values were obtained fitting compound concentration-dependent changes in normalized fluorescence (Fnorm).

RESULTS AND DISCUSSION

Homology Modeling of the Serine Protease Domain of TMPRSS2

Considering its relevance for both the drug design and the enzymatic function, we focused our attention on the TMPRSS2 SPD (Ile256 to Gly492).

A search performed with the SWISS MODEL webserver identified three very similar structures (**Figure 3** and **Table 2**) as suitable templates to generate TMPRSS2 models: (1) two structures of the human plasma kallikrein (PK), a serine

[1]http://bioinf.uab.cat/cgi-bin/pars-cgi/pars.pl

[2]https://playmolecule.org/deepsite

[3]http://ftmap.bu.edu/login.php

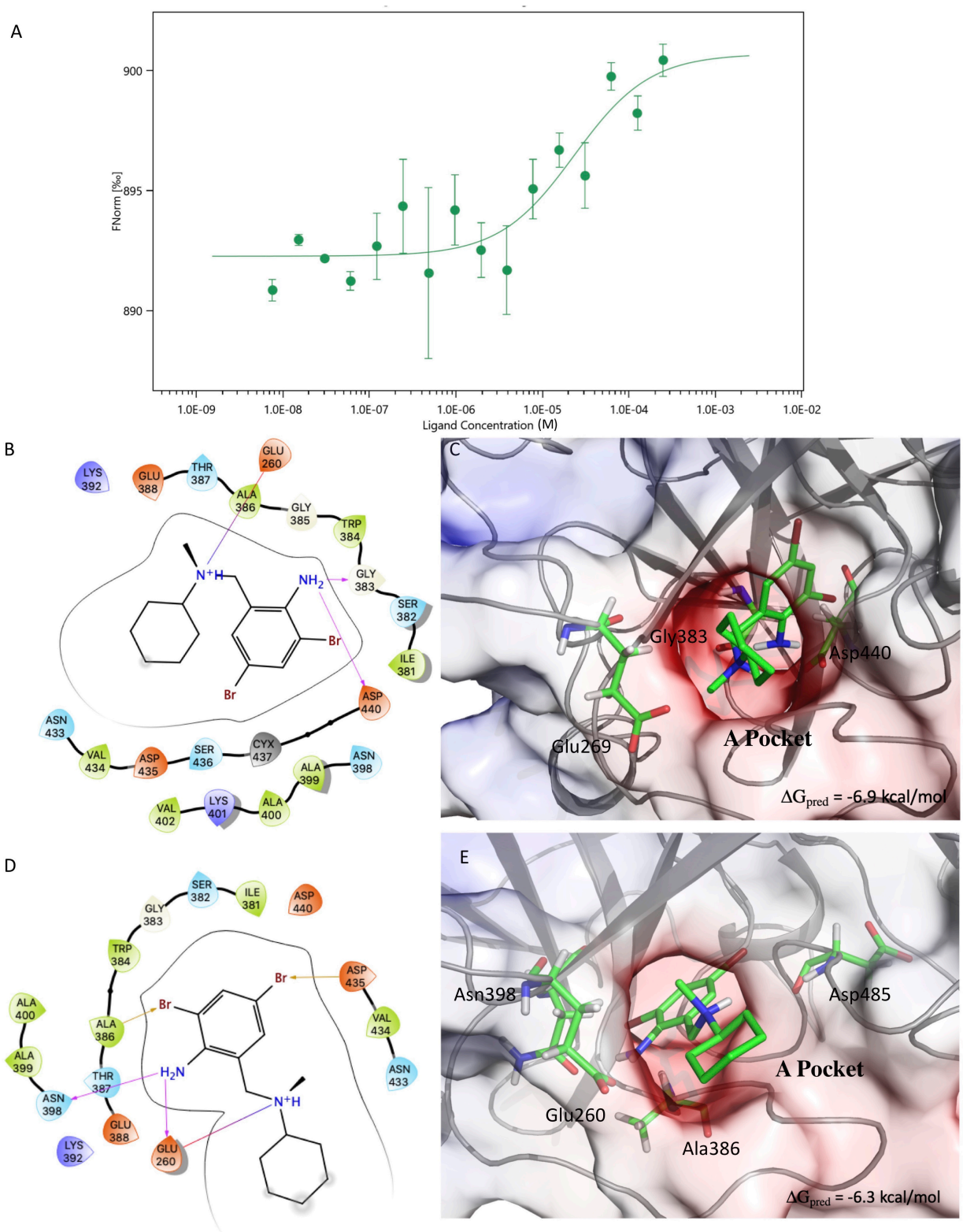


FIGURE 2 | Results of the MST experiments **(A)**. Structures and schemes of the interactions of the S-BH **(B,C)** and R-BH **(D,E)** in complex with C-M-5F8T as resulted from IFD calculations. The protein surface is colored according to the electrostatic potential. The unit of electrostatic potential is kbT/*e* where kb, T, and *e* are the Boltzmann's constant, absolute temperature, and the charge of an electron, respectively. The ΔG_{pred} values reported in the picture are the GlideScore values obtained from docking calculations. IFD, induced-fit docking; MST, microscale thermophoresis.

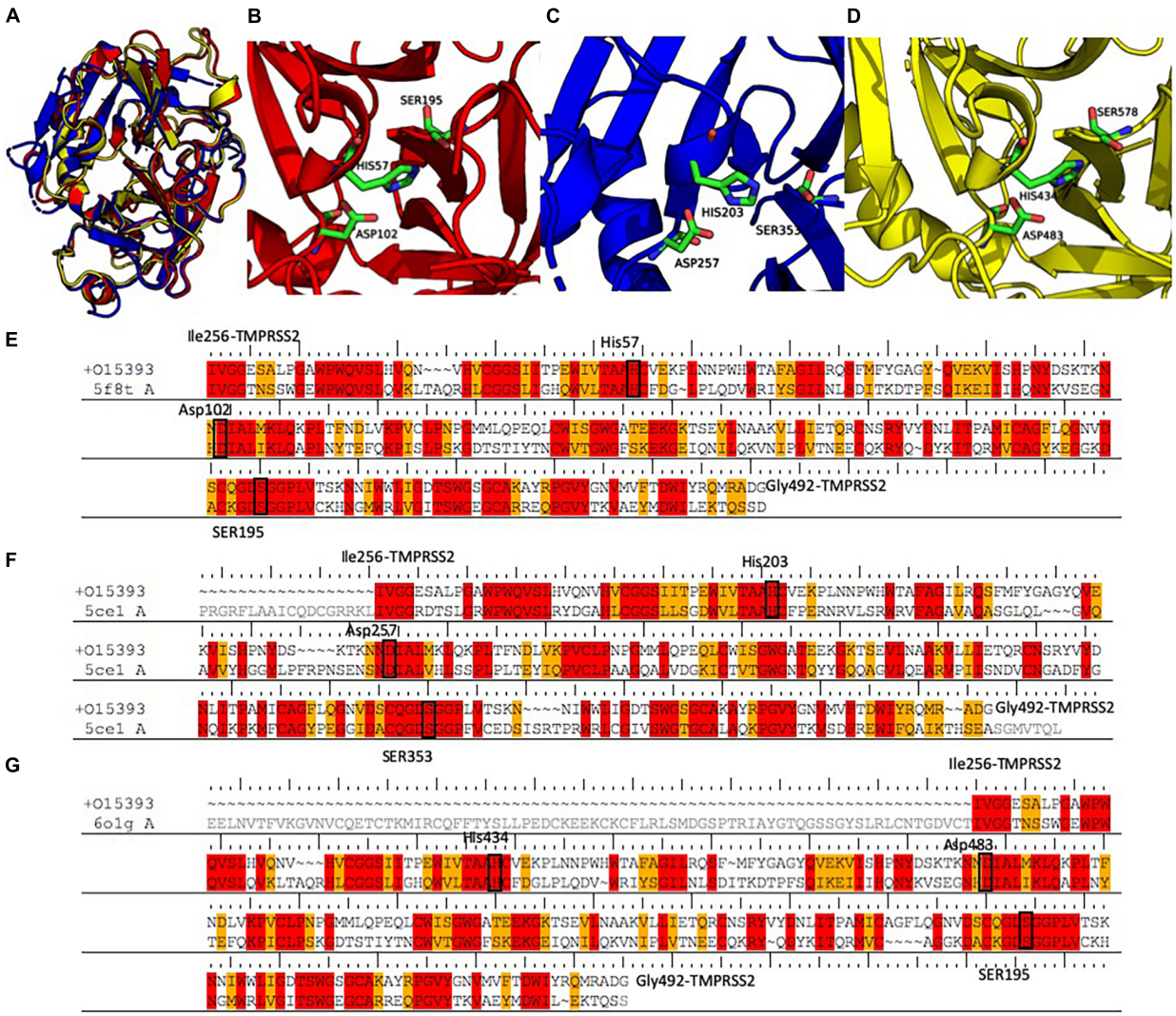


FIGURE 3 | (A) Structural alignment between the three selected templates 5F8T (red), 5CE1 (blue), and 6O1G (yellow). Details of the catalytic sites of the PK structures deposited with the PDB code 5F8T **(B)** and 6O1G **(C)** and of the HP structure deposited with the code 5CE1 **(D)**. **(E–G)** Sequence alignments between the three templates and the SPD of TMPRSS2. Identical residues are colored in red; conserved residues (according to the BLOSUM62 scoring matrix) are colored in orange. Abbreviations: PDB, Protein Data Bank; SPD, serine protease domain.

protease that cleaves high-molecular-weight kininogen (HMWK) to generate bradykinin (BK) (Schmaier, 2013) [PDB codes: 5F8T and 6O1G (Partridge et al., 2019), resolution 1.75 and 2.50 Å] and (2) the structure of hepsin (HP), a membrane-bound serine protease able to catalyze protein cleavage after basic amino-acid residues (PDB code: 5CE1, resolution 2.50 Å). In fact, the pairwise RMSD computed using the Cα atoms and the program ALMOST (Fu et al., 2014) is smaller than 0.5 Å.

The sequences of the three selected templates were aligned to TMPRSS2 using the PRIME-STA procedure (**Figures 3E–G**), and 10 models were generated starting from each template. Finally, the model with the lowest potential energy was selected from the three different groups. As expected, all the three models were very similar, with a pairwise Cα − RMSD below 0.5 Å. Furthermore, visual inspection of the three structures confirmed the similarity between all models, with the exception of the region between Tyr322 and Ser333. In fact, while in the two models derived from PK structures (M-5F8T and M-6O1G), this region

TABLE 2 | Summary of the sequence–sequence alignment between the sequence of the serine protease domain of TMPRSS2 and the three selected templates.

Template PDB code	Score	Identities (%)	Positives (%)	Gaps (%)
5F8T	1,218	41	59	2
5CE1	1,152	39	55	5
6O1G	1,185	41	57	4

The score is the BLAST bit score. Identities is the percentage of residues that are identical between the sequences. Positives is the percentage of residues that are positive matches according to the similarity matrix (BLOSUM62). Gaps is the percentage of gaps in both the query and homolog as returned by BLAST.

is a β-sheet, in the model form HP structure (M-5CE1), it is modeled as a long loop. This is not surprising because in the sequence–sequence alignment between HP and TMPRSS2 used for model generation, this region is characterized by the insertion of three amino acids.

The quality of the models was evaluated with the Protein Structure Quality viewer implemented in Maestro, computing structural parameters widely used in the evaluation of homology models (Sgrignani et al., 2009) and by the PROSA-Web server (Wiederstein and Sippl, 2007; **Table 3**). This analysis did not show any critical points for all generated models. Nevertheless, the number of violations of the allowed regions in the Ramachandran plot, and other violations from the ideal structural parameters were higher for the models generated using 5CE1 and 6O1G.

Molecular Dynamics Simulations of the TMPRSS2 Models

Aimed to (1) understand the overall stability of the generated models, (2) to detect problematic or poorly modeled regions, and (3) to generate an ensemble of protein conformation for docking (Amaro et al., 2008; Sgrignani et al., 2009), we performed three 250 ns long MD simulations for each of the selected models. PROPKA calculations with the model from 5F8T indicated a positively charged catalytic histidine (His296) as the most probable state. However, considering that the same residue was predicted as His-εin the other two models and that this specific protonation state would be required to start the enzymatic reaction (Ishida and Kato, 2003), we simulated this specific residue in both protonation states.

The simulation outputs were analyzed using consolidated observables such as RMSD, Rg, and the per-residue RMSF (**Figure 4**). This analysis highlighted a higher stability of M-5F8T with respect to M-5CE1 and M-6O1G. In particular, the simulations of M-5F8T always converged to a maximum RMSD of <3 Å from the starting model and Rg values similar to the starting one. Contrarily, M-5CE1 and M-6O1G showed continuously increasing RMSD and Rg profiles, suggesting that these models are less stable.

The RMSF profiles of M-5F8T and M-6O1G did not show anything relevant, substantially confirming the stability of M-5F8T. Contrarily, in M-5CE1, the protein region between positions 320 and 350, which contains the Tyr322 and Ser333 loop discussed before, was characterized by high RMSF values.

Small molecules in the catalytic site (PC1, HPC, 7SD, **Figure 1**) of the templates were preserved in the TMPRSS2 models, as the behavior of these molecules during MD simulations provides important hints about their quality and suitability to bind drugs. In the case of M-5F8T, the PC1 molecule remained in the S1 specificity pocket through a salt-bridge with Asp435 (**Figure 5A**). A similar behavior was also observed in the MD simulations of M-5CE1 (**Figure 5B**) for HCP. Contrarily, P4C rapidly dissociated from M-6O1G in all the simulations, probably because of the lack of a positively charged group docking the molecule to the S1 specificity pocket.

The good structural parameters (**Table 3**), the higher stability with respect to the M-5CE1 and M-6O1G (**Figure 4**), and the stable binding observed for PC1 in all the performed simulations suggested M-5F8T as the most reliable TMPRSS2 model. We, therefore, analyzed this model more deeply, focusing on the geometry of the catalytic triad (Asp345, His296, and Ser441). The analysis of distances between the three residues (**Figures 5C–F**) showed that this region of the protein remained stable during all the performed simulations. However, the system with a charged His296 adopted a conformation more similar to the starting model in which the Cγ@Asp345-Cγ@His296 and Cβ@Ser441-Cγ@His296 distances are 5.1 and 4.4 Å, respectively.

Docking of Camostat and Nafamostat to TMPRSS2

As in the MD simulations, docking of camostat and nafamostat in M-5F8T was performed with His296 in two protonation states, that is, positively charged and protonated on ε.

The outcomes of these calculations (**Figure 6**) indicated that camostat adopts a similar binding mode irrespectively to the His296 protonation state. In particular, camostat places its guanidine group in the S1 specificity pocket where it forms a salt bridge with Asp435 orienting the other part of the molecule in the same direction.

Nafamostat is characterized by the presence of a guanidine group and one its isoster. Therefore, while one of these is placed in the S1 specificity pocket, the other forms different interactions depending on the His296 protonation state. In fact, in the model with ε protonated His296, the guanidine moiety forms a salt bridge with Glu299. On the contrary, in the model with the positively charged His296, the isosteric group directly binds Asp345, that is one of the members of the catalytic triad (**Figure 6**).

Interestingly, the binding score is not affected by the His296 protonation state. However, the predicted score is lower for nafamostat than camostat, which is in a qualitative agreement with literature that reports an IC50 value for nafamostat 10 times lower than for camostat (Yamamoto et al., 2016).

TABLE 3 | Results of the structure quality evaluation.

Model name	Ramachandran violation	RMS bond dev.	RMS angle dev.	Backbone	Sidechains	Peptide planar dev.	Sidechains planar dev.	Torsion planar dev.	Z-score
M-5F8T	9	0.020	2.21	5	17	5.80	0.007	1.13	−6.82
M-5CE1	15	0.022	2.27	11	25	6.36	0.008	1.18	−7.12
M-6O1G	19	0.021	3.05	18	23	6.65	0.010	1.45	−6.47

Z-score is a measure of the overall model quality, and it was calculated by the Prosa-webserver (Wiederstein and Sippl, 2007). All the other parameters were calculated by Protein Structure Quality viewer implemented in the Schrodinger suite for molecular modeling.

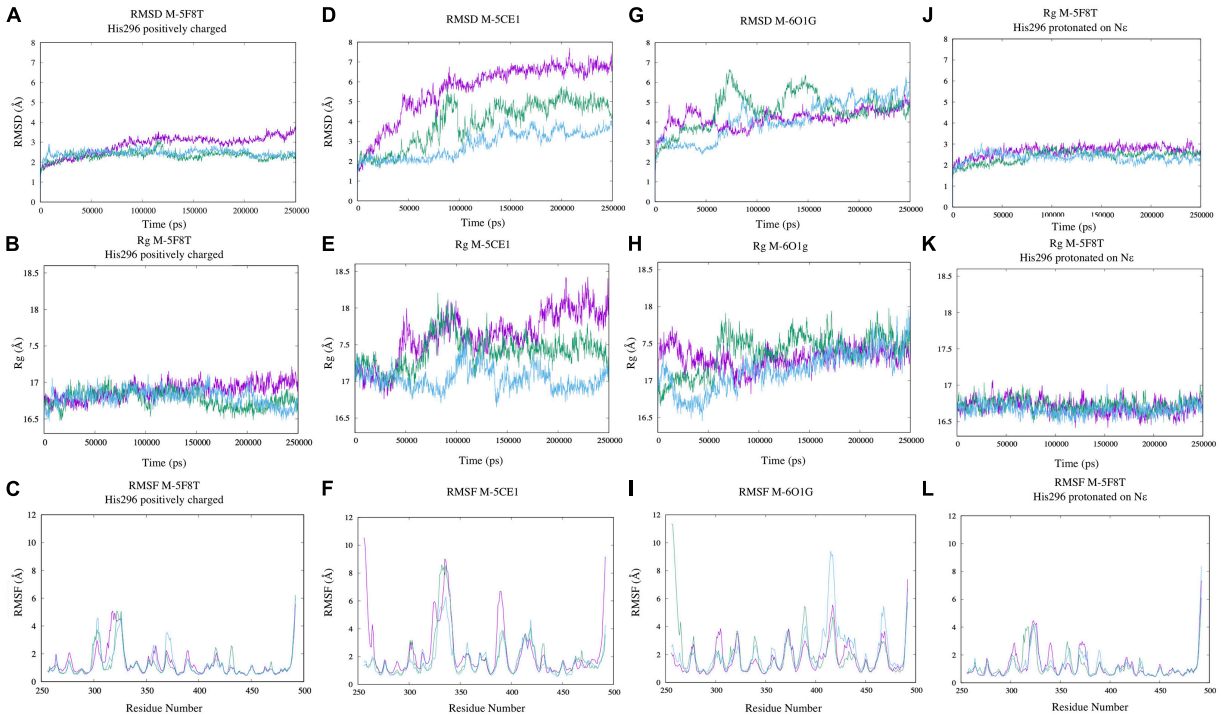


FIGURE 4 | Analysis of the MD simulations of M-5F8T **(A–C)**, M-5F8T-His296 neutral **(D–F)**, M-5CE1 **(G–I)**, and M-6O1G **(J–L)**. The data from the three distinct simulations are depicted in different colors. RMSD, radius of gyration (Rg), and RMSF were calculated considering backbone atoms. RMSD, root mean square deviation; RMSF, root mean square fluctuation.

Bromhexine Binding to TMPRSS2 Investigated by Microscale Thermophoresis

There have been discordant reports on the ability of BH to inhibit TMPRSS2. In fact, while the results of Lucas et al. (2014) appeared robust and convincing, a recent investigation by Hall and coworkers (Shrimp et al., 2020) concluded that BH is completely inactive as a TMPRSS2 inhibitor.

It is, however, important to consider that TMPRSS2 is a membrane protein with a peculiar and poorly understood activation mechanism. The purification of the active form of the enzyme, necessary for the inhibition tests, is thus extremely difficult. Furthermore, we noted that the protein quantity used in TMPRSS2 enzymatic assay is rarely reported (Meyer et al., 2013; Lucas et al., 2014), and that when reported (Shrimp et al., 2020), it extremely high (1 μM) with respect to the 1–2 nM concentrations used for other similar proteases (Hammamy et al., 2013; Ivanova et al., 2017). This suggests that the active species could be only a small fraction of the total protein, making it more difficult to observe a non-covalent weaker inhibition as that of BH. Thus, to better understand the existence and the strength of a BH/TMPRSS2 complex, we performed MST experiments.

MST is a recently developed biophysical technique that enables the investigation of molecular complexes measuring changes, upon binding, of the migration of target proteins in a laser-induced thermal gradient (Jerabek-Willemsen et al., 2011, 2014; Fassi et al., 2019).

The results of the MST experiments confirmed the BH/TMPRSS2 interaction with a Kd of 24 ± 13 μM (**Figure 2A**).

Modeling the Interaction Between BH and TMPRSS2

Motivated by literature data (Lucas et al., 2014) and by the results of our MST experiments, we used computational methods to investigate the BH/TMPRSS2 interaction at an atomistic level.

A closer look at the chemical structure of BH revealed that it cannot form a covalent bond with the protein and, therefore, should have a different inhibition mechanism compared to camostat and nafamostat. We, therefore, performed docking calculations considering the catalytic site as a putative BH binding site. However, when simulated by MD, the ligand–protein complexes dissociated after few nanoseconds suggesting a low reliability of the obtained structures. This observation was validated by performing several simulations starting from slightly different initial poses of BH in the catalytic site obtained using runs with different grids (data not shown): inevitably, the BH-TMPRSS2 complex dissociates in few nanoseconds.

In their 2007 review, Laporte and Naesens (Laporte and Naesens, 2017) suggested that, because of its selectivity for TMPRSS2 over matriptase, trypsin, or thrombin, BH could exert its inhibitory effect binding to an allosteric site.

To better explore this hypothesis, we analyzed our models with Sitemap (Halgren, 2007, 2009), a computational tool already

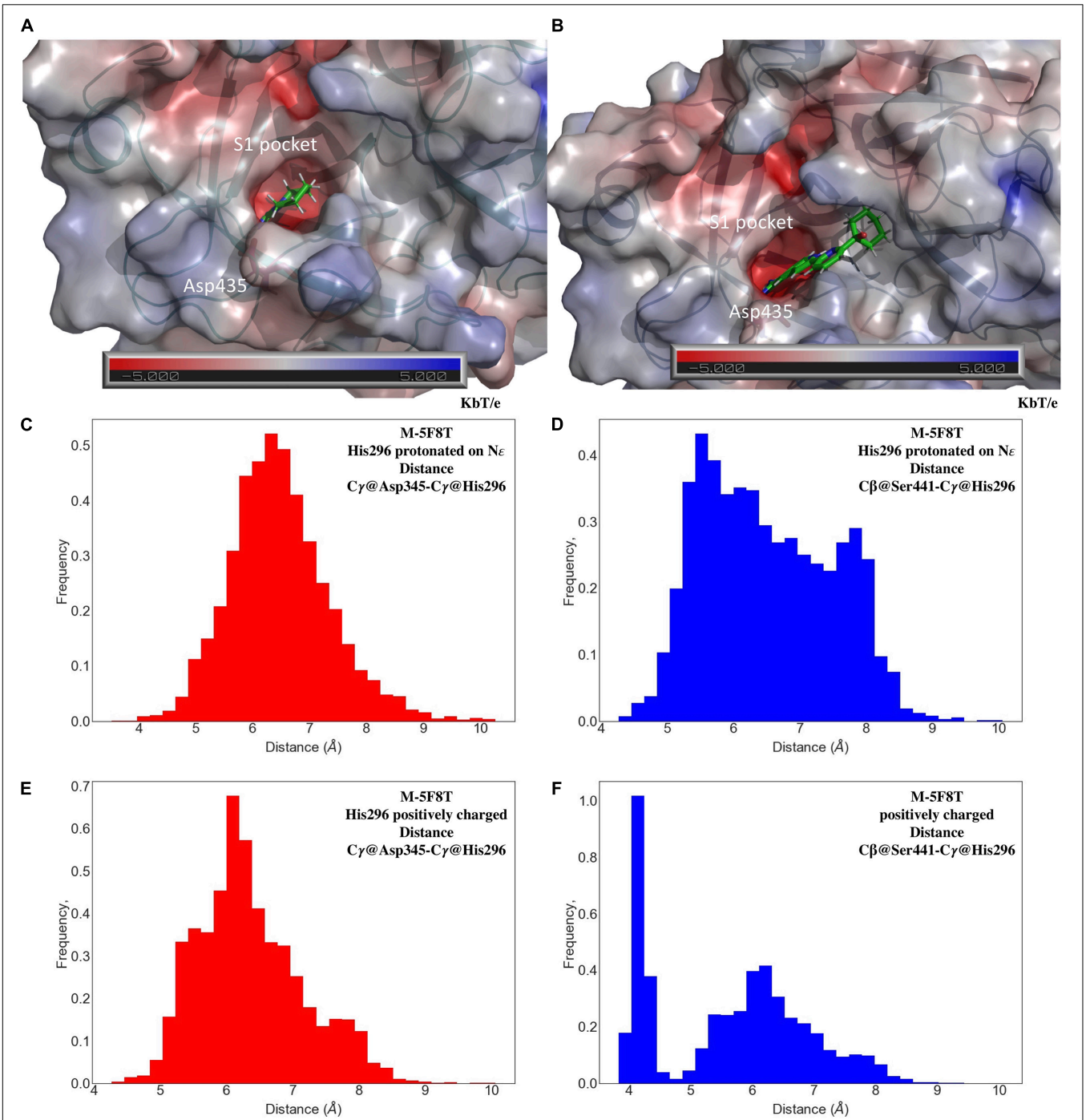


FIGURE 5 | Representative conformations of the MD of M-5F8T **(A)** and M-5CE1 **(B)** with a focus on the interactions of PC1 and HPC with the S1 specificity pocket. The unit of electrostatic potential is kbT/e where kb, T, and *e* are the Boltzmann's constant, absolute temperature, and the charge of an electron, respectively. Histogram analysis of the Cγ@Asp345-Cγ@His296 and Cβ@Ser441-Cγ@His296 distances in the MD simulations of M-5F8T with His296-ε **(C,D)** and His296 positively charged **(E,F)**. MD, molecular dynamics.

applied to the identification of allosteric sites (Sgrignani et al., 2014; Kots et al., 2017; Sanchez-Martin et al., 2020).

This analysis highlighted the existence of two putative drug binding sites (M-5F8T_site_1 and M-5F8T_site_2), for which a SiteScore value of >0.8 was obtained (**Table 4**). To note, while Site_1 describes a zone quite far from the active site, Site_2 includes also a part of the active sites Ser441 and His296.

In recent years, several algorithms for the prediction of putative allosteric sites (see also section "Materials and Methods") have been developed. Therefore, to obtain a more comprehensive

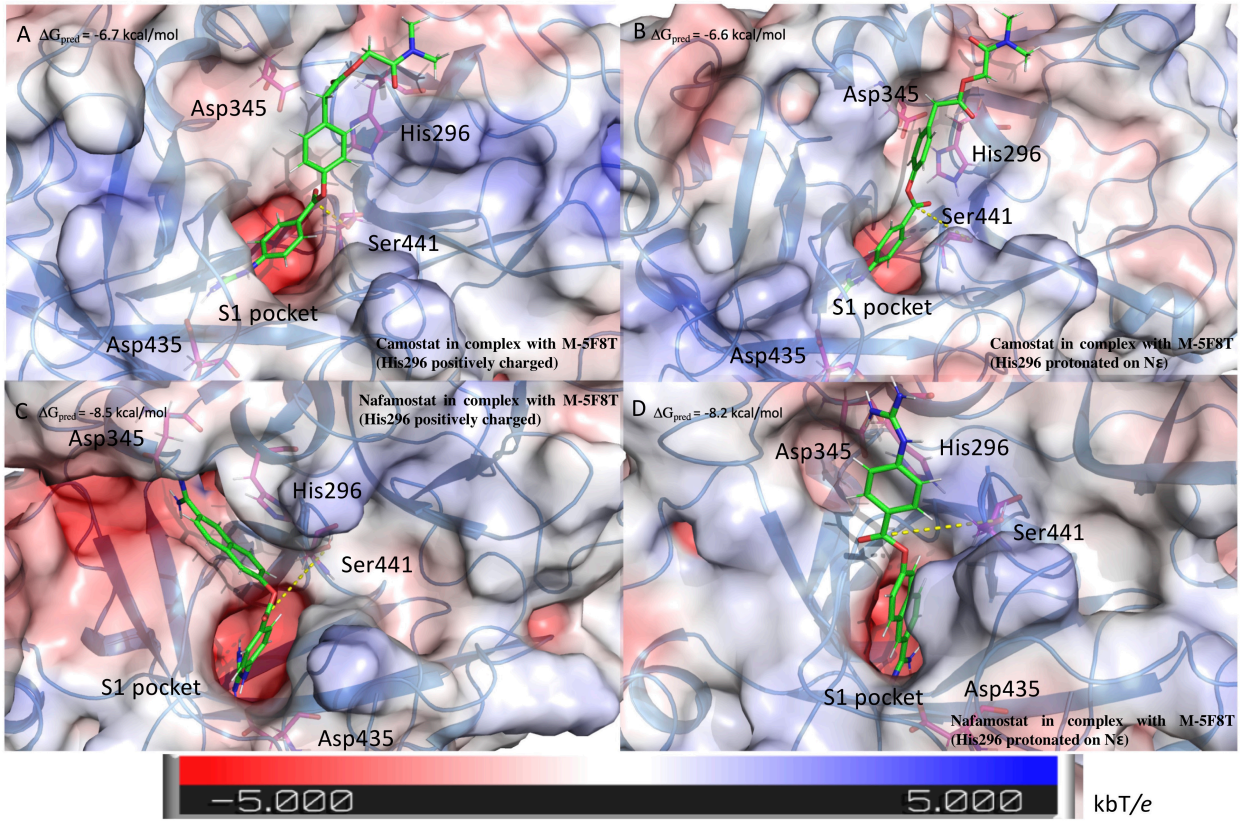


FIGURE 6 | Structure of the complexes between TMPRSS2 and camostat **(A,B)** or nafamostat **(C,D)**. The pictures in the right and left columns refer to the docking calculation ran on M-5F8T considering His296 in its positively charged state or protonated on the ε nitrogen, respectively. The protein surface is colored in the function of the electrostatic potential according to the shown bar. The unit of electrostatic potential is kbT/*e* where kb, T, and *e* are the Boltzmann's constant, absolute temperature, and the charge of an electron, respectively. The ΔG_{pred} values reported in the picture are the GlideScore values obtained from docking calculations.

analysis, we carried out the same calculation using three additional algorithms (PARS, Deepsite, and FTMap). All these calculations confirmed the existence of Site_1, while Site_2 was identified by PARS and FTMap but not Deepsite.

Next, we docked BH in M-5F8T_site_1 and M-5F8T_site_2 and performed MD simulations of the complexes. While the simulations with BH bound to M-5F8T_site_2 resulted in a complex dissociating in the first 100 ns, the complex between BH and TMPRSS2 bound to M-5F8T_site_1 and remained stable for ∼1 μs. In fact, the drug remained close to the protein although it did not find a stable binding mode. Consequently, we performed two additional MD simulations to clarify this point. In the first control simulation, the ligand dissociated in the first 500 ns, while the second control simulation BH remained close to the protein surface without finding a stable binding mode, as in the first run. It should be also noted that in these simulations, while close to the protein surface, BH has a distance of ∼30 Å from the catalytic triad, making it difficult to imagine a direct effect on the catalytic activity from that position.

Summarizing the results of our simulations indicated that M-5F8T_site_1 and M-5F8T_site_2 are unsuitable to bind BH, leaving unsolved the question about the position of the BH allosteric site.

We, therefore, explored the possibility of the existence of a hidden allosteric site.

The Role in Protein Activity of Free Isoleucine at the N-Terminal Side

The essential role for the enzymatic activity of the free isoleucine at the N-terminal side of TLPs has been previously reported (Stubbs et al., 1998; Huber, 2013; Meyer et al., 2013).

From visual inspection of M-5F8T, it can be seen that the N-terminal fragment of the protein occupies a negatively charged cavity (subsequently A-pocket, **Figure 7**) where the positively charged amino group of the N-terminal Ile256 forms a salt bridge with Asp440. Interestingly, Asp440 is contiguous to Ser441, one of the members of the catalytic triads, but oriented in a different direction.

To investigate the importance of this structural feature (i.e., the presence of free isoleucine at N-terminal site) for TMPRSS2, we generated a model of the enzyme deleting the first two residues at N-terminal (Ile256 and Val257) from M-5F8T (this model is subsequently indicated as C-M-5F8T) and performed an MD simulation for 1 μs.

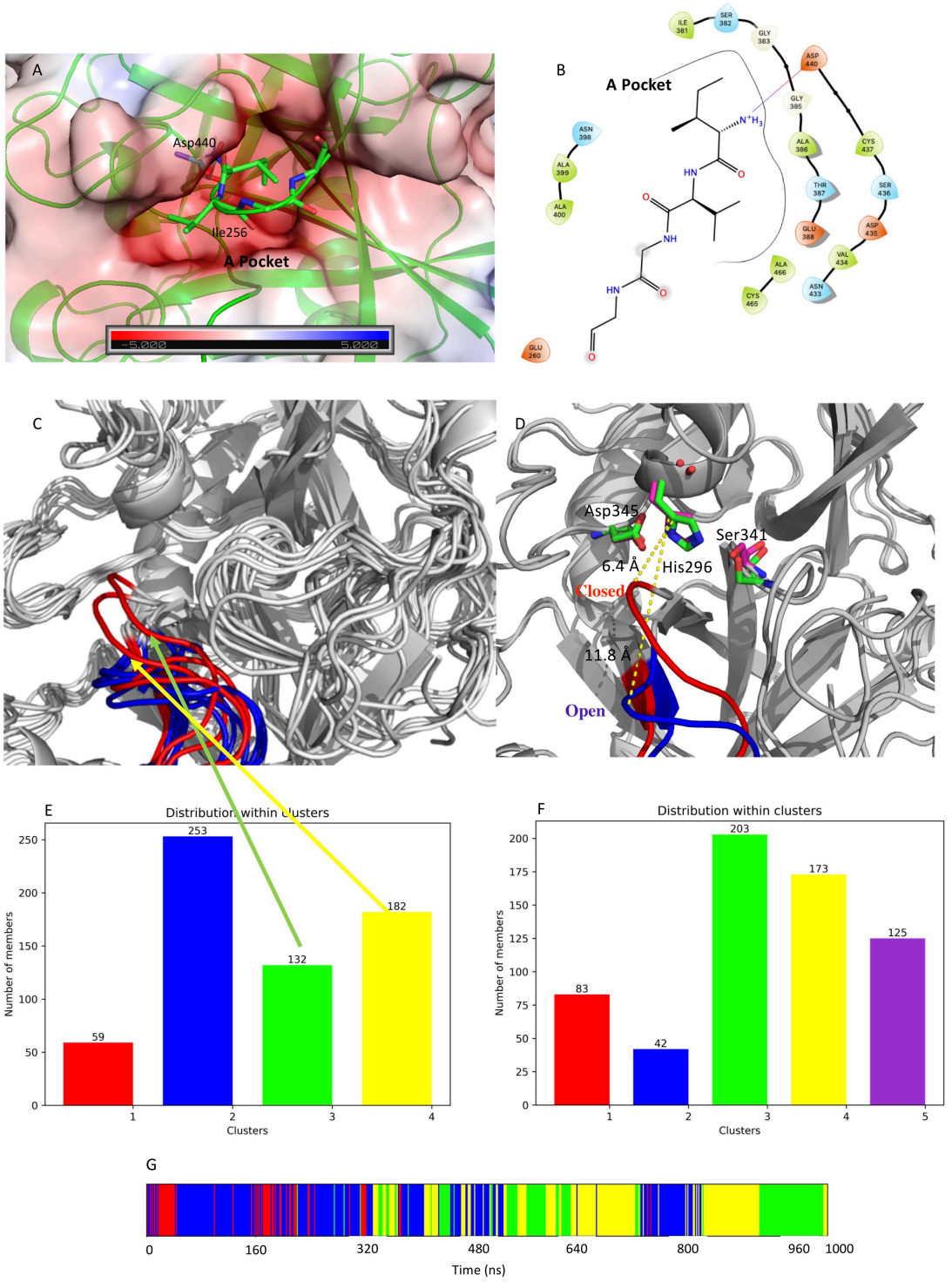


FIGURE 7 | (A) Structure of the negatively charged cavity that hosts the N-terminal tail of the catalytically active TMPRSS2. The protein surface is colored by the electrostatic potential value calculated by the APBS plugin implemented in Pymol-2.3.4. The unit of electrostatic potential is kbT/*e* where kb, T, and *e* are the Boltzmann's constant, absolute temperature, and the charge of an electron, respectively. **(B)** Scheme of the interactions between the N-terminal end and its binding site on the TMPRSS2 structure. **(C)** Conformations of the loop Gly462-Val473 in the representative structures from the identified clusters. The conformations from the simulation of apo-M-5F8T are shown in blue while those from the simulation of C-M-5F8T in red. **(D)** Comparison between the loop conformation assumed in the cluster3 of the C-M-5F8T MD (in red) and that of cluster3 of the M-5F8T MD (in blue). **(E)** Distribution of the C-M-5F8T conformations over the identified clusters. **(F)** Distribution of the M-5F8T conformations over the identified clusters. **(G)** Time evolution of the clusters obtained from the C-M-5F8T MD simulation. MD, molecular dynamics.

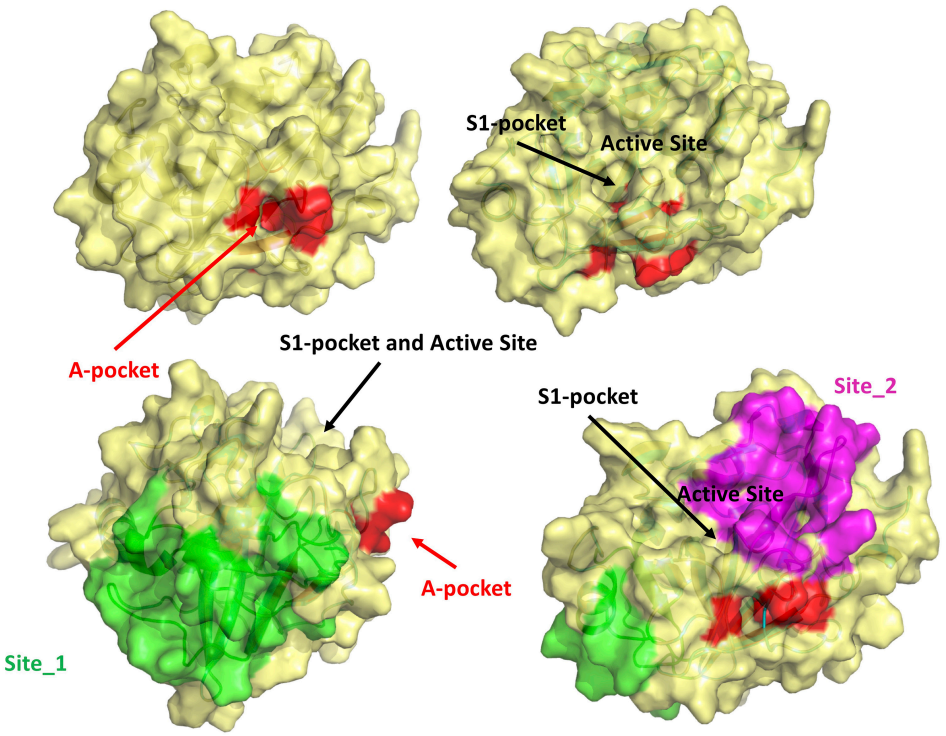


FIGURE 8 | Visual summary of all the possible binding sites investigated in this study. The M-5F8T-C model is represented with different orientations to make clearer the reciprocal positions of the sites.

Next, we compared the outputs of this simulation with an identical simulation of M-5F8T in its apo form.

Given its importance for the substrate recognition in TLPs and its structural proximity to the binding site, we first focused our analysis on the effect of the presence/absence of Ile256-Val257 on the structure of the S1 specificity pocket. Visual inspection of M-5F8T suggested that the Gly462-Val473 loop could regulate the access of the substrates to the S1 specificity pocket. We, therefore, analyzed the effects of the N-terminal truncation on the conformation of this protein region. Interestingly, we observed (**Figure 7C**) that, while in the simulation of the M-5F8T, the loop conserves a conformation similar to that adopted in the starting model; in the second part of the C-M-5F8T simulation, it moves closer to the catalytic triad (**Figure 7D**) occupying a position that should reduce the efficiency of both substrate recognition and catalysis.

Taken together, these observations strongly suggest that the binding of the N-terminal tail into the A-pocket (**Figures 7A,B**, **8**) is important to stabilize the structure of the TMPRSS2 active site and, in particular, of the S1 specificity pocket.

Is the A-Pocket Relevant for Drug Design?

Considering the importance, highlighted by the previously discussed simulations, of the binding of the N-terminal tail into the A-pocket for the stability of the catalytic site, we performed some analysis to explore its druggability.

To this end, we first used Sitemap to analyze the C-M-5F8T model. This analysis showed that the cavity was made accessible by the deletion of the first two residues of M-5F8T and was highly suitable for drug binding, with a value of Sitescore of 0.93 over 1.00, where a druggable cavity should have SiteScore > 0.80.

TABLE 4 | Results of the Sitemap analysis carried out on M-5F8T.

Title	SiteScore	Dscore	Volume (Å^3)	Residues
M-5F8T_site_3	0.924	0.656	68.9	262, 263, 264, 265, 266, 267, 268, 270, 271, 272, 311, 312, 313, 314, 315, 316, 360, 384, 397
M-5F8T_site_2	0.892	0.907	201.3	274, 275, 277, 278, 279, 280, 281, 296, 300, 301, 302, 307, 308, 317, 384, 385, 386, 390, 391, 392, 393, 438, 439, 441
M-5F8T_site_1	0.863	0.872	279.2	369, 370, 371, 372, 373, 374, 376, 377, 403, 404, 405, 406, 407, 409, 413, 421, 422, 424, 425, 428, 429, 430, 469, 471, 476, 478, 479
M-5F8T_site_4	0.738	0.395	50.7	367, 368, 371, 372, 373, 375, 376, 447, 449, 454. 456, 478
M-5F8T_site_5	0.655	0.608	96.7	271, 291, 310, 311, 312, 325, 326, 327, 351, 355

Next, we investigated if this pocket could be a suitable site for the BH binding.

Preliminary calculations (see section "Materials and Methods") showed that the cavity was optimized for the binding of the N-terminal tail and not for the binding of a small molecule. We, therefore, first computed the optimal BH/TMPRSS2 binding pose using the IFD protocol implemented in Glide, followed by three MD independent simulations of 500 ns each.

The results of the docking calculations indicated the both (S)- and (R)-BH bind the A-pocket with a similar predicted affinity (−6.9 and −6.3 kcal/mol, respectively, **Figures 2B–E**).

From the structural point of view, both the molecules place the ring bearing the two bromine atoms in a cavity delimited by Gly363, Ile381, Ser382, Trp384, and Asp440. Moreover, the amino group in position 5 of the same ring establishes h-bond interactions with Asp440 and Gly383 in the case of (S)-BH and with Asn390 and Glu260 for (R)-BH. In both the structures, the positively charged amino group of BH electrostatically interacts with Glu260.

All MD simulations confirmed the stability of the complexes obtained from docking, with BH bound into the A-pocket for the entire simulation time.

Finally, we also performed a cluster analysis to verify the conformation of the Gly462-Val473 loop, which regulates the access to the S1 specificity pocket. This analysis (**Supplementary Figure 1**) clearly showed that the loop conserves a closed conformation in all the representative structures extracted from the simulations of (S)- and (R)-BH inside the A-pocket.

CONCLUSION

TMPRSS2 is an exceptional and intriguing protein (Thunders and Delahunt, 2020), whose precise physiological function remains unknown. Despite this, it has been linked with several human diseases, such as prostate cancer, and has been shown to play a key role in viral infections.

In particular, the SPD of TMPRSS2 is critical for the priming of SARS-Cov-2 spike protein. This prompted us to investigate the interaction between TMPRSS2 various known drugs, using both computational and experimental methods.

While in the case of camostat and nafamostat, our computational studies confirmed that these two molecules bind to the active site of TMPRSS2 and form molecular adducts competent for the formation of covalent complexes; in the case of BH, our studies indicated that a competitive inhibition was unlikely.

On the other side, MST experiments confirmed a BH/TMPRSS2 interaction, leading us to ponder the hypothesis of an allosteric binding. We, therefore, used computer simulations to validate this hypothesis. The MD simulation confirmed that similar to other TLPs, the binding of a free isoleucine residue in the A-pocket is crucial to stabilize the catalytically competent active site conformation. Moreover, our calculations indicated that this cavity (**Figure 8**), fully accessible in the TMPRSS2 zymogen, is suitable to host BH or other more potent drugs that could be identified by virtual screening.

The study presented here provides further understanding of how the catalytic activity of TMPRSS2 can be modulated and new ways to develop more selective and potent antiviral treatments for current and future pandemics.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

JS designed the study, performed and analyzed simulations and experiments, and wrote and revised the manuscript. AC designed the study, analyzed the results of simulations and experiments, and wrote and revised the manuscript. Both authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank Torsten Steinmetzer for useful discussions and Flora Gruner for the generous support. We also thank Fondazione CARIPARO for the financial support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.666626/full#supplementary-material

REFERENCES

Afar, D. E., Vivanco, I., Hubert, R. S., Kuo, J., Chen, E., Saffran, D. C., et al. (2001). Catalytic cleavage of the androgen-regulated TMPRSS2 protease results in its secretion by prostate and prostate cancer epithelia. *Cancer Res.* 61, 1686–1692.

Amaro, R. E., Baron, R., and Mccammon, J. A. (2008). An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput. Aided Mol. Des.* 22, 693–705. doi: 10.1007/s10822-007-9159-2

Amaro, R. E., Baudry, J., Chodera, J., Demir, Ö, Mccammon, J. A., Miao, Y., et al. (2018). Ensemble docking in drug discovery. *Biophys. J.* 114, 2271–2278.

Ansarin, K., Tolouian, R., Ardalan, M., Taghizadieh, A., Varshochi, M., Teimouri, S., et al. (2020). Effect of bromhexine on clinical outcomes and mortality in COVID-19 patients: a randomized clinical trial. *Bioimpacts* 10, 209–215. doi: 10.34172/bi.2020.27

Bertram, S., Glowacka, I., Blazejewska, P., Soilleux, E., Allen, P., Danisch, S., et al. (2010). TMPRSS2 and TMPRSS4 facilitate trypsin-independent spread of influenza virus in Caco-2 cells. *J. Virol.* 84, 10016–10025. doi: 10.1128/jvi.00239-10

Bestle, D., Heindl, M. R., Limburg, H., Van Lam Van, T., Pilgram, O., Moulton, H., et al. (2020). TMPRSS2 and furin are both essential for proteolytic activation of SARS-CoV-2 in human airway cells. *Life Sci. Alliance* 3:e202000786. doi: 10.26508/lsa.202000786

Cavasotto, C. N., and Phatak, S. S. (2009). Homology modeling in drug discovery: current trends and applications. *Drug Discov. Today* 14, 676–683. doi: 10.1016/j.drudis.2009.04.006

Chen, Y.-W., Lee, M.-S., Lucht, A., Chou, F.-P., Huang, W., Havighurst, T. C., et al. (2010). TMPRSS2, a serine protease expressed in the prostate on the apical surface of luminal epithelial cells and released into semen in prostasomes, is misregulated in prostate cancer cells. *Am. J. Pathol.* 176, 2986–2996. doi: 10.2353/ajpath.2010.090665

Depfenhart, M., De Villiers, D., Lemperle, G., Meyer, M., and Di Somma, S. (2020). Potential new treatment strategies for COVID-19: is there a role for bromhexine as add-on therapy? *Intern. Emerg. Med.* 15, 801–812. doi: 10.1007/s11739-020-02383-3

Fassi, E. M. A., Sgrignani, J., D'agostino, G., Cecchinato, V., Garofalo, M., Grazioso, G., et al. (2019). Oxidation state dependent conformational changes of HMGB1 regulate the formation of the CXCL12/HMGB1 Heterocomplex. *Comput. Struct. Biotechnol. J.* 17, 886–894. doi: 10.1016/j.csbj.2019.06.020

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

Fu, B., Sahakyan, A. B., Camilloni, C., Tartaglia, G. G., Paci, E., Caflisch, A., et al. (2014). ALMOST: an all atom molecular simulation toolkit for protein structure determination. *J. Comput. Chem.* 35, 1101–1105. doi: 10.1002/jcc.23588

Guarnera, E., and Berezovsky, I. N. (2019). Toward comprehensive allosteric control over protein activity. *Structure* 27, 866–878.e861.

Habtemariam, S., Nabavi, S. F., Ghavami, S., Cismaru, C. A., Berindan-Neagoe, I., and Nabavi, S. M. (2020). Possible use of the mucolytic drug, bromhexine hydrochloride, as a prophylactic agent against SARS-CoV-2 infection based on its action on the Transmembrane Serine Protease 2. *Pharmacol. Res.* 157:104853. doi: 10.1016/j.phrs.2020.104853

Halgren, T. (2007). New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* 69, 146–148. doi: 10.1111/j.1747-0285.2007.00483.x

Halgren, T. A. (2009). Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inform. Modell.* 49, 377–389. doi: 10.1021/ci800324m

Hammamy, M. Z., Haase, C., Hammami, M., Hilgenfeld, R., and Steinmetzer, T. (2013). Development and characterization of new peptidomimetic inhibitors of the West Nile virus NS2B-NS3 protease. *ChemMedChem* 8, 231–241. doi: 10.1002/cmdc.201200497

Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., et al. (2016). OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* 12, 281–296. doi: 10.1021/acs.jctc.5b00864

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and Is blocked by a clinically proven protease inhibitor. *Cell* 181, 271–280.e278.

Huber, R. (2013). How I chose research on proteases or, more correctly, how it chose me. *Angew. Chem. Int. Ed. Engl.* 52, 68–73. doi: 10.1002/anie.201205629

Ishida, T., and Kato, S. (2003). Theoretical perspectives on the reaction mechanism of serine proteases: the reaction free energy profiles of the acylation process. *J. Am. Chem. Soc.* 125, 12035–12048. doi: 10.1021/ja021369m

Ivanova, T., Hardes, K., Kallis, S., Dahms, S. O., Than, M. E., Künzel, S., et al. (2017). Optimization of substrate-analogue furin inhibitors. *ChemMedChem* 12, 1953–1968. doi: 10.1002/cmdc.201700596

Jerabek-Willemsen, M., André, T., Wanner, R., Roth, H. M., Duhr, S., Baaske, P., et al. (2014). MicroScale thermophoresis: interaction analysis and beyond. *J. Mol. Struct.* 1077, 101–113. doi: 10.1016/j.molstruc.2014.03.009

Jerabek-Willemsen, M., Wienken, C. J., Braun, D., Baaske, P., and Duhr, S. (2011). Molecular interaction studies using microscale thermophoresis. *Assay Drug. Dev. Technol.* 9, 342–353. doi: 10.1089/adt.2011.0380

Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., and De Fabritiis, G. (2017). DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* 33, 3036–3042. doi: 10.1093/bioinformatics/btx350

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, L. M. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935. doi: 10.1063/1.445869

Kots, E. D., Lushchekina, S. V., Varfolomeev, S. D., and Nemukhin, A. V. (2017). Role of protein dimeric interface in allosteric inhibition of N-Acetyl-aspartate hydrolysis by human aspartoacylase. *J. Chem. Inform. Modell.* 57, 1999–2008. doi: 10.1021/acs.jcim.7b00133

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., et al. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* 10, 733–755. doi: 10.1038/nprot.2015.043

Laporte, M., and Naesens, L. (2017). Airway proteases: an emerging drug target for influenza and other respiratory virus infections. *Curr. Opin. Virol.* 24, 16–24. doi: 10.1016/j.coviro.2017.03.018

Li, T., Sun, L., Zhang, W., Zheng, C., Jiang, C., Chen, M., et al. (2020). Bromhexine hydrochloride tablets for the treatment of moderate COVID-19: an open-label randomized controlled pilot study. *Clin. Transl. Sci.* 13, 1096–1102. doi: 10.1111/cts.12881

Lucas, J. M., Heinlein, C., Kim, T., Hernandez, S. A., Malik, M. S., True, L. D., et al. (2014). The androgen-regulated protease TMPRSS2 activates a proteolytic cascade involving components of the tumor microenvironment and promotes prostate cancer metastasis. *Cancer Discov.* 4, 1310–1325. doi: 10.1158/2159-8290.cd-13-1010

Maggio, R., and Corsini, G. U. (2020). Repurposing the mucolytic cough suppressant and TMPRSS2 protease inhibitor bromhexine for the prevention and management of SARS-CoV-2 infection. *Pharmacol. Res.* 157:104837. doi: 10.1016/j.phrs.2020.104837

Martyna, G. J., Klein, M. L., and Tuckerman, M. (1992). Nosé–hoover chains: the canonical ensemble via continuous dynamics. *J. Chem. Phys.* 97, 2635–2643. doi: 10.1063/1.463940

Martyna, G., Tobias, D., and Klein, M. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* 101, 4177–4189. doi: 10.1063/1.467468

Meyer, D., Sielaff, F., Hammami, M., Böttcher-Friebertshäuser, E., Garten, W., and Steinmetzer, T. (2013). Identification of the first synthetic inhibitors of the type II transmembrane serine protease TMPRSS2 suitable for inhibition of influenza virus activation. *Biochem. J.* 452, 331–343. doi: 10.1042/bj20130101

Montopoli, M., Zumerle, S., Vettor, R., Rugge, M., Zorzi, M., Catapano, C. V., et al. (2020). Androgen-deprivation therapies for prostate cancer and risk of infection by SARS-CoV-2: a population-based study (N = 4532). *Ann. Oncol.* 31, 1040–1045. doi: 10.1016/j.annonc.2020.04.479

Olsson, M. H., Sondergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011). PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. *J. Chem. Theory Comput.* 7, 525–537. doi: 10.1021/ct100578z

Panjkovich, A., and Daura, X. (2014). PARS: a web server for the prediction of protein allosteric and regulatory sites. *Bioinformatics* 30, 1314–1315. doi: 10.1093/bioinformatics/btu002

Partridge, J. R., Choy, R. M., Silva-Garcia, A., Yu, C., Li, Z., Sham, H., et al. (2019). Structures of full-length plasma kallikrein bound to highly specific inhibitors describe a new mode of targeted inhibition. *J. Struct. Biol.* 206, 170–182. doi: 10.1016/j.jsb.2019.03.001

Pászti-Gere, E., Czimmermann, E., Ujhelyi, G., Balla, P., Maiwald, A., and Steinmetzer, T. (2016). In vitro characterization of TMPRSS2 inhibition in IPEC-J2 cells. *J. Enzyme Inhib. Med. Chem.* 31, 123–129. doi: 10.1080/14756366.2016.1193732

Sanchez-Martin, C., Moroni, E., Ferraro, M., Laquatra, C., Cannino, G., Masgras, I., et al. (2020). Rational design of allosteric and selective inhibitors of the molecular chaperone TRAP1. *Cell Rep.* 31:107531. doi: 10.1016/j.celrep.2020.107531

Schmaier, A. H. (2013). "Chapter 638 - prekallikrein and plasma kallikrein," in *Handbook of Proteolytic Enzymes*, Third Edition. eds N. D. Rawlings and G. Salvesen (Cambridge, MA: Academic Press), 2885–2892. doi: 10.1016/b978-0-12-382219-2.00638-4

Sgrignani, J., Bon, M., Colombo, G., and Magistrato, A. (2014). Computational approaches elucidate the allosteric mechanism of human aromatase inhibition: a novel possible route to small-molecule regulation of CYP450s activities? *J. Chem. Inf. Mod.* 54, 2856–2868. doi: 10.1021/ci500425y

Sgrignani, J., Bonaccini, C., Grazioso, G., Chioccioli, M., Cavalli, A., and Gratteri, P. (2009). Insights into docking and scoring neuronal alpha4beta2 nicotinic receptor agonists using molecular dynamics simulations and QM/MM calculations. *J. Comput. Chem.* 30, 2443–2454. doi: 10.1002/jcc.21251

Sgrignani, J., Garofalo, M., Matkovic, M., Merulla, J., Catapano, C. V., and Cavalli, A. (2018). Structural biology of STAT3 and its implications for anticancer therapies development. *Int. J. Mol. Sci.* 19:1591. doi: 10.3390/ijms19061591

Shen, L. W., Mao, H. J., Wu, Y. L., Tanaka, Y., and Zhang, W. (2017). TMPRSS2: a potential target for treatment of influenza virus and coronavirus infections. *Biochimie* 142, 1–10. doi: 10.1016/j.biochi.2017.07.016

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., and Farid, R. (2006). Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* 49, 534–553. doi: 10.1021/jm050540c

Shrimp, J. H., Kales, S. C., Sanderson, P. E., Simeonov, A., Shen, M., and Hall, M. D. (2020). An enzymatic TMPRSS2 assay for assessment of clinical candidates and discovery of inhibitors as potential treatment of COVID-19. *ACS Pharmacol. Transl. Sci.* 3, 997–1007. doi: 10.1021/acsptsci.0c00106

Singh, N., Decroly, E., Khatib, A.-M., and Villoutreix, B. O. (2020). Structure-based drug repositioning over the human TMPRSS2 protease domain: search for chemical probes able to repress SARS-CoV-2 Spike protein cleavages. *Eur. J. Pharm. Sci.* 153, 105495–105495. doi: 10.1016/j.ejps.2020.105495

Stubbs, M. T., Renatus, M., and Bode, W. (1998). An active zymogen: unravelling the mystery of tissue-type plasminogen activator. *Biol. Chem.* 379, 95–103.

Sungnak, W., Huang, N., Becavin, C., Berg, M., Queen, R., Litvinukova, M., et al. (2020). SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* 26, 681–687. doi: 10.1038/s41591-020-0868-6

Szabo, R., Wu, Q., Dickson, R. B., Netzel-Arnett, S., Antalis, T. M., and Bugge, T. H. (2003). Type II transmembrane serine proteases. *Thromb. Haemost.* 90, 185–193. doi: 10.1160/th03-02-0071

Thunders, M., and Delahunt, B. (2020). Gene of the month: TMPRSS2 (transmembrane serine protease 2). *J. Clin. Pathol.* 73, 773–776. doi: 10.1136/jclinpath-2020-206987

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *R. Stat. Soc.* 63, 411–423. doi: 10.1111/1467-9868.00293

Tubiana, T., Carvaillo, J.-C., Boulard, Y., and Bressanelli, S. (2018). TTClust: a versatile molecular simulation trajectory clustering program with graphical summaries. *J. Chem. Inform. Modell.* 58, 2178–2182. doi: 10.1021/acs.jcim.8b00512

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303.

Wiederstein, M., and Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucl. Acids Res.* 35, W407–W410.

Xiang, Z. (2006). Advances in homology protein structure modeling. *Curr. Protein Pept. Sci.* 7, 217–227. doi: 10.2174/138920306777452312

Xu, Y., Wang, S., Hu, Q., Gao, S., Ma, X., Zhang, W., et al. (2018). CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Res.* 46, W374–W379.

Yamamoto, M., Matsuyama, S., Li, X., Takeda, M., Kawaguchi, Y., Inoue, J. I., et al. (2016). Identification of nafamostat as a potent inhibitor of middle east respiratory syndrome coronavirus S protein-mediated membrane fusion using the split-protein-based cell-cell fusion assay. *Antimicrob. Agents Chemother.* 60, 6532–6539. doi: 10.1128/aac.01043-16

Yu, J., Zhou, Y., Tanaka, I., and Yao, M. (2010). Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics* 26, 46–52. doi: 10.1093/bioinformatics/btp599

Zang, R., Gomez Castro, M. F., Mccune, B. T., Zeng, Q., Rothlauf, P. W., and Sonnek, N. M. (2020). TMPRSS2 and TMPRSS4 promote SARS-CoV-2 infection of human small intestinal enterocytes. *Sci. Immunol.* 5:eabc3582. doi: 10.1126/sciimmunol.abc3582

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: 30 April 2021
doi: 10.3389/fmolb.2021.644720

A Single Mutation in the Outer Lipid-Facing Helix of a Pentameric Ligand-Gated Ion Channel Affects Channel Function Through a Radially-Propagating Mechanism

Alessandro Crnjar[1†], *Susanne M. Mesoy*[2†], *Sarah C. R. Lummis*[2‡] *and Carla Molteni*[1*‡]

[1] *Physics Department, King's College London, London, United Kingdom,* [2] *Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom*

Pentameric ligand-gated ion channels (pLGICs) mediate fast synaptic transmission and are crucial drug targets. Their gating mechanism is triggered by ligand binding in the extracellular domain that culminates in the opening of a hydrophobic gate in the transmembrane domain. This domain is made of four α-helices (M1 to M4). Recently the outer lipid-facing helix (M4) has been shown to be key to receptor function, however its role in channel opening is still poorly understood. It could act through its neighboring helices (M1/M3), or via the M4 tip interacting with the pivotal Cys-loop in the extracellular domain. Mutation of a single M4 tyrosine (Y441) to alanine renders one pLGIC—the 5-HT_{3A} receptor—unable to function despite robust ligand binding. Using Y441A as a proxy for M4 function, we here predict likely paths of Y441 action using molecular dynamics, and test these predictions with functional assays of mutant receptors in HEK cells and *Xenopus* oocytes using fluorescent membrane potential sensitive dye and two-electrode voltage clamp respectively. We show that Y441 does not act via the M4 tip or Cys-loop, but instead connects radially through M1 to a residue near the ion channel hydrophobic gate on the pore-lining helix M2. This demonstrates the active role of the M4 helix in channel opening.

Keywords: pentameric ligand-gated ion channels, 5-HT_3 receptors, Cys-loop receptors, mutagenesis, molecular dynamics, M4 helix

OPEN ACCESS

Edited by:
Massimiliano Bonomi, Institut Pasteur, France

Reviewed by:
Lucie Delemotte, Royal Institute of Technology, Sweden
Jim Pfaendtner, University of Washington, United States

***Correspondence:**
Carla Molteni
carla.molteni@kcl.ac.uk

[†] *These authors share first authorship*

[‡] *These authors share last authorship*

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

Received: *21 December 2020*
Accepted: *22 February 2021*
Published: *30 April 2021*

Citation:
Crnjar A, Mesoy SM, Lummis SCR and Molteni C (2021) A Single Mutation in the Outer Lipid-Facing Helix of a Pentameric Ligand-Gated Ion Channel Affects Channel Function Through a Radially-Propagating Mechanism. Front. Mol. Biosci. 8:644720. doi: 10.3389/fmolb.2021.644720

1. INTRODUCTION

Pentameric ligand-gated ion channels (pLGICs) are neuroreceptors involved in fast synaptic transmission underlying the physiological processes of muscle action, gut activity, and neurological function. They are present throughout the central and peripheral nervous systems, and mediate the action of biologically active compounds including nicotine, alcohol, and many anesthetics (Nemecz et al., 2016). Their wide range of functions makes them an attractive therapeutic target, if we can understand and modulate their structure and function.

While the transmission of the mechanical signal triggered by agonist binding that culminates in channel opening is not yet fully understood, significant advances can be achieved by means of mutagenesis experiments to pinpoint key residues/mechanisms as well as molecular simulations (Crnjar et al., 2019b), especially because an increasing number of high-resolution structures are

now available in a variety of states (Hilf and Dutzler, 2008; Bocquet et al., 2009; Althoff et al., 2014; Hassaine et al., 2014; Miller and Aricescu, 2014; Sauguet et al., 2014; Du et al., 2015; Huang et al., 2015; Kudryashev et al., 2016; Nys et al., 2016; Basak et al., 2018a,b; Polovinkin et al., 2018; Zhu et al., 2018).

PLGICs are made up of five subunits, with a predominantly β-sheet extracellular domain (ECD), an α-helical transmembrane domain (TMD, containing the ion-permeable pore), and often an intracellular domain (ICD) (**Figure 1**). Neurotransmitters bind at the interface between subunits in the ECD, causing the channel (over 60 Å away) to open, allowing ions into the cell (Lemoine et al., 2012; Nemecz et al., 2016). The TMD is made up of four α-helices, (M1 to M4), with M2 lining the channel pore, M1 and M3 in a second concentric circle, and finally M4 facing the lipid membrane (**Figure 1C**). While this helix is fundamentally an amphipathic barrier to the hydrophobic lipid environment, there is growing evidence that the M4 helix plays a key role in pLGIC function: mutations in the M4 helices of mammalian pLGICs have been shown to reduce or inhibit channel opening (Cory-Wright et al., 2018; Tang et al., 2018; Mesoy et al., 2019), or promote channel function (da Costa Couto et al., 2020) although the exact mechanism is not yet clear.

There are two main proposed mechanisms for M4 affecting channel opening. One mechanism, first outlined by Da Costa et al. (DaCosta and Baenziger, 2009) proposes that the C-terminal end of the M4 helix, which sits at the level of the ECD-TMD interface, is required for the signal transduction through that interface via interactions with the Cys-loop. Loss of the M4 tip or of M4 binding to the rest of the TMD would disrupt this interface. This would not affect ligand binding, but would prevent the channel opening signal from reaching the TMD, resulting in what has been described as an uncoupled receptor. In this case, mutations that alter the pinning of M4 to the rest of the channel could disrupt M4/Cys-loop interactions in what we term a "vertically"-propagating chain of events (i.e., propagating along a direction parallel to the protein axis). This model is structurally appealing, and supported by the fact that allosteric modulation has been shown to propagate from M4 tip residues to the Cys-loop in the cationic $\alpha4\beta2$ nicotinic acetylcholine receptor (nAChR) (Alcaino et al., 2017).

The other mechanism would involve residues of the M4 acting directly on M1/M3, i.e., "radially," and this signal propagating to affect channel opening, e.g., by interactions with M2. A naturally occurring M4 mutation (C418W) in the *Torpedo* nAChR which alters channel function has been shown to be energetically coupled to two M1 residues (S226 and T229) (Domville and Baenziger, 2018). The same work also shows that the C418W mutations does not affect interactions of the M4 C-terminal domain (CTD) with the Cys-loop, supporting the radial mechanism proposal.

In addition to functional characterization showing the involvement of M4 in channel activity, the M4 of the neuromuscular nAChR has been calculated to move as a unit approximately halfway through receptor activation, further supporting its possible role in connecting ligand binding (the first event of activation) to channel opening (Mitra et al., 2004). Together this experimental and theoretical work highlights an intriguing role for the M4 helix in coupling channel opening to

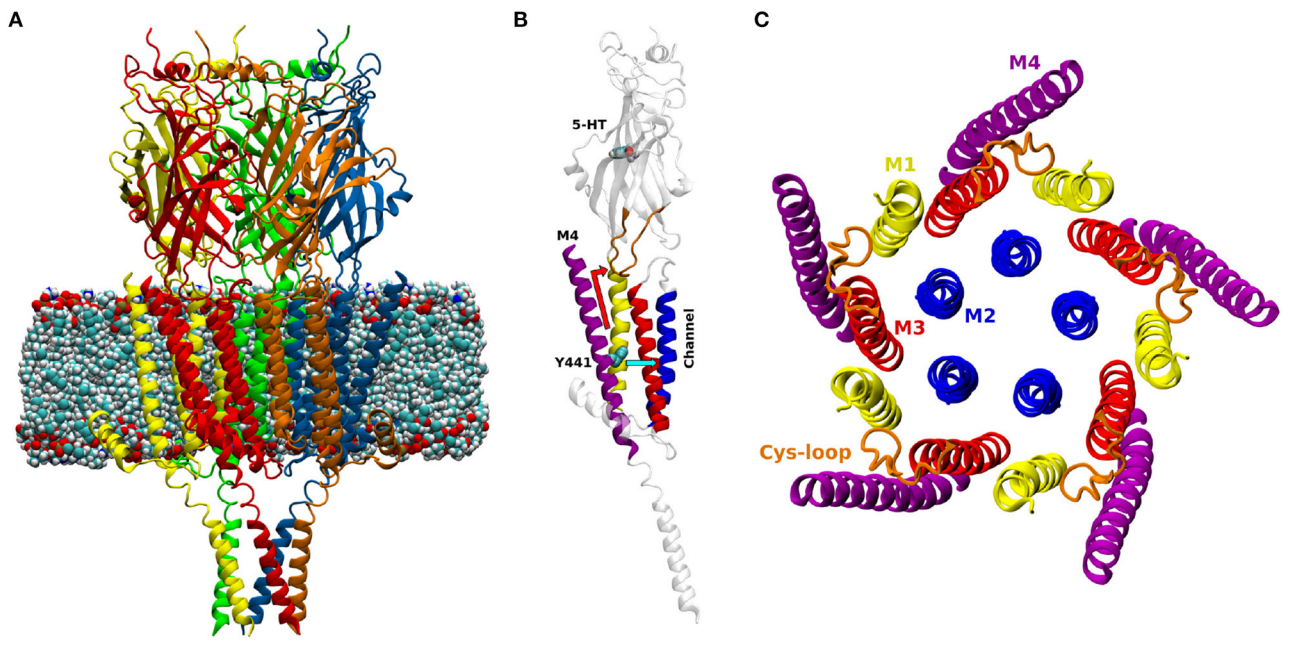


FIGURE 1 | (A) Model of the open structure of the 5-HT_{3A} receptor (PDB ID 6DG8, Basak et al., 2018a) embedded in a lipid membrane. The five subunits are colored in red, yellow, green, blue and orange. **(B)** One subunit of the 5-HT_{3A} receptor; M1: yellow, M2: blue, M3: red, M4: purple, Cys-loop: orange. The atoms of 5-HT and residue Y441 are shown as spheres. The two possible paths of Y441A action are shown in light blue (*radial*) and red (*vertical*). **(C)** View orthogonal to the channel axis of the transmembrane domain of the 5-HT_{3A} receptor.

ligand binding, which could be influenced by mutations and by the lipid membrane (e.g., its composition, thickness, or fluidity).

Among pLGICs, the 5-HT_{3A} receptor (5-HT_{3A}R) constitutes a unique case when it comes to the mechanism of action of the M4 helix. Alanine mutation of a single tyrosine (Y441), located approximately at the center of the M4 (**Figure 1B**) prevents receptor activation but not ligand binding. Of the other 26 M4 residues tested (434-461), 24 can be changed to alanine with little or no detectable effect on channel function (the remaining two abolish receptor expression, so their effect on function is unclear) (Mesoy et al., 2019). We take this to mean that Y441-dependent coupling can be taken as a proxy for the coupling mediated by the entire M4 helix in this channel. Here we use the uncoupling Y441A mutation to probe the role and function of the 5-HT_{3A}R M4 helix by comparing it to the wild-type (WT) in molecular dynamics (MD) simulations, testing proposed mechanisms by mutagenesis and functional assays.

2. MATERIALS AND METHODS

2.1. Molecular Biology

The mouse 5-HT_{3A} cDNA (Q6J1J7) in pcDNA3.1 (for HEK cell transfection) or pGEMHE (for RNA production) was modified by QuikChange site-directed mutagenesis (Agilent Technologies) to create point mutants (verified by sequencing).

2.2. Cell Culture

Human embryonic kidney (HEK) 293 cells were maintained at 37°C at 5% CO_2 in a humidified atmosphere, in Dulbecco's Modified Eagle's Medium/Nutrient Mix F12 (1:1) (Invitrogen, Paisley, UK) with GlutaMAX ^{TM} and 10% fetal bovine serum (GE Healthcare) (DMEM/FBS) and passaged when confluent. Five micrograms of WT or mutant 5-HT_{3A} DNA and 30 μl polyethyleneamine (Polysciences) incubated for 10 min in 1 mL DMEM was added to 60% confluent HEK293 cells for transfection, and cells grown for 2 days before assays.

2.3. FlexStation

As described previously (Price and Lummis, 2005) cells were incubated for 45 min with fluorescent membrane potential-sensitive dye (Membrane Potential Blue kit, Molecular Devices) diluted in Flex buffer (10 mM HEPES, 115 mM NaCl, 1 mM KCl, 1 mM $CaCl_2$, 1 mM $MgCl_2$, and 10 mM glucose, pH 7.4), and subsequently assayed at room temperature for 180 s, with readings every 2 s. 5-HT was added to each well after 20 s. Concentration-response curves were generated by iterative fitting in GraphPad Prism 7 (after normalization to max ΔF) with the equation $y = a + \frac{b-a}{1+10^{(n_H(\log EC_{50}-x))}}$ where y is the fluorescent response, x is log[5-HT] (log of the concentration of ligand), a is the minimum response, b is the maximum response, and n_H is the Hill slope.

2.4. Radioligand Binding

This was performed as previously described. Lummis and Thompson (2013) Briefly receptors in crude HEK293 cell membranes were labeled with the 5-HT_3R antagonist [^{3}H]GR65630 by incubation in 0.5 mL 10 mM HEPES buffer pH 7.5 for 1 h on ice, using 1 μM quipazine to determine non-specific binding. Data were analyzed in GraphPad Prism by iterative curve fitting.

2.5. Two-Electrode Voltage Clamp

Xenopus laevis oocytes from EcoCyte Biosciences (Austin Texas) were injected with 100 pg cRNA (generated with the ThermoFisher mMESSAGE mMACHINE T7 transcription kit) and left in injection media [88 mM NaCl, 2.4 mM $NaHCO_3$, 1 mM KCl, 0.82 mM $MgSO_4$ ·$7H_20$, 5 mM Tris-HCl, 0.33 mM $Ca(NO_3)2{\cdot}4H_2O$, 0.41 mM $CaCl_2{\cdot}2H_2O$, 2.51 mM sodium pyruvate, 0.12 mg/ml theophylline, 0.05 mg/ml gentamicin, pH 7.5] at 16°C for 24 h. Recording was performed at 22°C on a Roboocyte (Multichannel systems, Reulingeen, Germany), using calcium-free ND96 buffer (96 mM NaCl, 2 mM KCl, 1 mM $MgCl_2$, 5 mM HEPES, pH 7.5) and 5-HT solutions applied by a computer-controlled perfusion system. The holding potential was −60 mV, using glass microelectrodes with a resistance of approximately 1 MΩ backfilled with 3M KCl.

2.6. Model and Molecular Dynamics Simulations

A model of the 5-HT_{3A}R was built based on the cryo-EM open structure resolved by Basak et al. at 3.89 Å resolution (pdb entry: 6DG8, Basak et al., 2018a) (**Figure 1A**). The open structure was chosen in part because we assumed that the M4 helices would play their role late in the overall chain of receptor activation (i.e., subsequent to ligand binding), and therefore the effect of mutating Y441A would be better observed in a receptor conformation that is at the end and not the start of the activation process. Additionally, there was experimental evidence that WT and Y441A-containing receptors showed differences with regard to the open state (the WT can attain this state but the mutant receptor does not; Mesoy et al., 2019), suggesting that the open state was more likely to reveal functional differences between the two receptors.

This structure comprises the TMD, the ECD, and part of the ICD: the highly flexible residues 333–396 in the ICD were not resolved experimentally. This unstructured region was not reconstructed, considering not only its considerable length (which would result in a large solvation box), but also its likely lack of influence on the M4 helices. Conversely, the experimentally resolved and structured part of the ICD (the MA helices) was included, as the M4 movements may in principle be affected by the MAs.

The model was protonated at neutral pH, and embedded in a 6:7:7 cholesterol-POPC-POPE lipid membrane (with lipids randomly distributed) using the CHARMM-GUI web-based membrane builder (Jo et al., 2008), resulting in a membrane area of about 124 by 127 Å. The 6:7:7 concentration was chosen to resemble HEK cells membrane composition, resulting in a cholesterol/phospholipid ratio of 0.42, closer to the value of 0.48 in HEK cells (Dawaliby et al., 2016) than to that of 0.6–0.7 in oocytes (Opekarová and Tanner, 2003). This ratio has been used for simulations of membranes with cholesterol (6), POPC (7) and a third lipid (7) for the study of serotonin receptors (Shan et al., 2012; Crnjar and Molteni, 2020; Guros et al., 2020).

Mixed membranes containing POPC and POPE (together with cholesterol), have also been studied in the past (Elmore and Dougherty, 2003; Mahmood et al., 2013; Cao et al., 2015; Patra et al., 2015; Heusser et al., 2018; Oakes and Domene, 2019; Guros et al., 2020). The presence of cholesterol is important as this lipid is present in high concentration in brain cells membranes (Pfrieger, 2003; Chan et al., 2012), and a mixed membrane may prove important for a cooperative modulation of the effects of the Y441A mutation.

The system was then solvated in an orthorhombic supercell, with 52,477 TIP3P water molecules and 0.15 M of Na^+ and Cl^- ions to reproduce physiological conditions, together with 5 Cl^- counterions to counterbalance the positive charge of the five bound 5-HT molecules. The total number of ions was 162 for Na^+, and 142 for Cl^-. PyMOL (Schrödinger, LLC, 2015) was used to turn the five Y441 into alanines, resulting in a mutated receptor (MR) model. In total the WTR model contained 226,082 atoms and the MR model contained 226,027 atoms.

The systems were simulated with the NAMD 2.13 molecular dynamics package (Phillips et al., 2005), the AMBER ff14SB (Maier et al., 2015) and LIPID14 force-field (Dickson et al., 2014). The five 5-HTs in the binding pockets were parameterized as described in the **Supplementary Material**. The simulation time step was 2 fs, and the bonds containing hydrogen were constrained with the SHAKE algorithm. Particle Mesh Ewald was employed for the electrostatic interactions and a cut off of 10 Å was used for the non-bonded interactions.

At the beginning, the WTR and MR models underwent a minimization procedure, a slow heating and a partially restrained equilibration (with the protein α carbons and the 5-HT rings restrained while the lipids were free to diffuse). The equilibration of lipid membranes requires long time windows since their diffusion occurs over times of the order of tens to hundreds of nanoseconds (Kandt et al., 2007; Smith et al., 2018). Thus, the equilibration stage, performed within the isothermal-isobaric (NPT) ensemble, lasted around 150 ns in total, while slowly releasing the chosen restraints. **Supplementary Table 1** reports the full equilibration procedure followed.

After the equilibration, production runs were performed for both models, with 1.0 kcal/mol$Å^2$ restraints on M2, MA, MX α carbons. These restraints were kept during the production, since care must be taken in order to prevent the collapse of open structures of pLGICs, including the possible closure of the channel (Dämgen and Biggin, 2020). Past simulations on this very receptor (Crnjar and Molteni, 2020; Guros et al., 2020) without any restraints applied highlighted how the RMSD of the M4 can go up to 4 Å. This is similar to what was found in our simulations, thus proving that the chosen restraints do not affect the section from residue 441 and above. Moreover, residue 425 is far outside the membrane within the ICD, as shown by **Supplementary Figure 4**.

The production was carried out within the isothermal-isobaric (NPT) ensemble for each model at a temperature of 310 K, which is above gel transition temperature for all lipid species (Silvius, 1982; Kraske and Mountcastle, 2001), and at a pressure of 1 atm. Temperature was controlled by means of a Langevin thermostat with a collision frequency of 1.0 ps^{-1}, and pressure was controlled by means of a Langevin piston barostat with an oscillation period of 200 fs and a damping time constant of 100 fs.

For both WTR and MR, two replicas of 250 ns each were simulated, referred to in the following as R0 and R1. This choice was made as a consequence of the stochastic nature of the Langevin dynamics, which would result in different trajectories in different replicas, particularly affecting the diffusion of lipid molecules, to whom the outermost M4 helices are exposed. Most of the analysis described in the following were performed over the conjunction of the time windows 50-to-250 ns of both R0 and R1 (R01-400). In fact, to improve statistics when performing simulations, long runs or multiple replicas would give qualitative similar results for time- and subunit-averaged quantities (Crnjar and Molteni, 2020). Moreover, the pentameric nature of this pLGIC allows for a simultaneous five-fold sampling of whatever phenomenon occurs within one subunit. For both R0 and R1, the first 50 ns of simulations were excluded from statistics collection in order to mitigate for the use of the same initial geometry and allow for independent equilibration (50 ns being the time window after which the protein RMSD flattens, as shown in **Supplementary Figure 1**). The analysis of quantities which needed to be expressed as functions of time, or that would be dependent on the order of the union of R0 and R1, were performed separately for R0 and for R1.

Trajectories were sampled every 50 ps, and analyzed with the Cpptraj (Roe and Cheatham, 2013) and MDAnalysis (Michaud-Agrawal et al., 2011; Gowers et al., 2016) software. Hydrogen bonds were defined by using a donor-acceptor distance smaller than 3.5 Å and a donor-hydrogen-acceptor angle larger than 120°. These values have been used in several previous works on pLGICs (Melis et al., 2008; McCormack et al., 2010; Comitani et al., 2014, 2015, 2016; Crnjar et al., 2019a,b), and are the defaults of analysis software such as MDAnalysis (Michaud-Agrawal et al., 2011; Gowers et al., 2016). A generic contact between any two given atoms (of two different residues) was considered here when their distance was shorter than a cutoff of 3.5 Å as in previous studies (Deol et al., 2004).

Aromatic interactions were calculated by considering distances and angles involving the vector normal to best-fit-plane to a given single aromatic ring (for residues with multiple aromatic rings, such as tryptophan, we consider the rings separately and then sum the interactions frequencies). π-π interactions consider a ring-ring distance less than 6.0 Å, and normals angle smaller than 45° or greater than 135°. The distance was chosen in order to be 1 Å larger than the optimal one predicted for benzene dimers (Sinnokrot et al., 2002). Anion-π interactions considered a ring-charged atom distance smaller than 5.0 Å, and an angle between ring normal and ring center-to-negative atom distance smaller than 40° or greater than 140° (Lucas et al., 2016).

3. RESULTS

3.1. Vertical Mechanism

We found no difference between the WT and mutant simulation either at or above residue 441 on M4 from examining the local dynamical fluctuations of M4 by evaluating the root mean

square fluctuations (RMSF) of each residue (**Figure 2**). This was calculated over R01-400 with respect to the post-equilibration positions, for the backbone atoms of those residues, and averaged over each of the five subunits. The errors were calculated as maximum semidifferences (half of the difference between maximum and minimum value).

Residue 425 (the last residue of the MA helix) was restrained in both simulations, so we did not consider nearby similarities valid. Around residue 441, the fluctuations were around 1–2 Å in both WTR and MR; on average the dynamics of residue 441 did not seem to be affected by removing the side chain. The fluctuations increased (as did the variation across subunits) toward the top of the helix. The M4 tip is located at the interface with the solvent in the extracellular region, at the same level of the edge of the outer lipid leaflet (**Supplementary Figure 4**), and is therefore allowed to move more freely. These comments hold true for both the WTR and the MR, which show little difference in the dynamics.

The top of the M4 helix therefore appears not to depend on the properties of residue 441. To confirm this, we investigated other factors. Firstly, to better estimate the effect of residue 441 dynamics upwards along the M4, we calculated the time-averaged dynamical correlation of this residue with respect to two representative amino-acids: Y448 and W459 (**Figure 3**). Both are structurally important; mutation of either residue abolishes cell surface expression (Mesoy et al., 2019). Y448 is an integral inwards-facing residue near Y441, and W459 is a good candidate for potential interactions with the Cys-loop, making it an interesting marker of M4 tip behavior. The dynamical correlation C_{ij} between two atoms i and j is defined (Hunenberger et al., 1995) as:

$$C_{ij} = \frac{\langle \vec{r}_i \vec{r}_j \rangle - \langle \vec{r}_i \rangle \langle \vec{r}_j \rangle}{\sqrt{(\langle \vec{r}_i{}^2 \rangle - \langle \vec{r}_i \rangle^2)(\langle \vec{r}_j{}^2 \rangle - \langle \vec{r}_j \rangle^2)}} \quad (1)$$

C_{ij} may take any value between 0 and 1: values close to 1 indicate that the two residues move consistently in the same direction over time, acting like a rigid body, while values close to 0 imply that the dynamics of these residues never display any correlation.

The correlations between residues 441 and 448/459 appear to be subunit- and replica-specific, although the 441-448 correlations are consistently higher than the 441–459 ones (as expected due to residue 448 being closer than 459 to 441). While for the pair 441–448 they reach values up to 0.8 (with an average of 0.69 ± 0.07), for the pair 441–459 they never surpass 0.6 (with an average of 0.24 ± 0.13), implying that 441 is not notably correlated with the top of the M4.

Secondly, we investigated the time- and subunit-averaged interactions (hydrogen bonds, π-π interactions and anion-π interactions) of selected M4 residues, evaluated over R01-400 (**Figures 4**, **5**). Errors were calculated via error propagation from the five standard deviations of data over time for each of the single subunits.

Figure 4 confirms that W459 is the only M4 tip residue that forms hydrogen bonds (although with low frequency) with the Cys-loop (residues 135 to 149). No major differences are noted between WTR and MR in either hydrogen bonds nor aromatic interactions, except for the fact that the removal of the 441 side chain upon mutation prevents it from forming interactions with its neighbors.

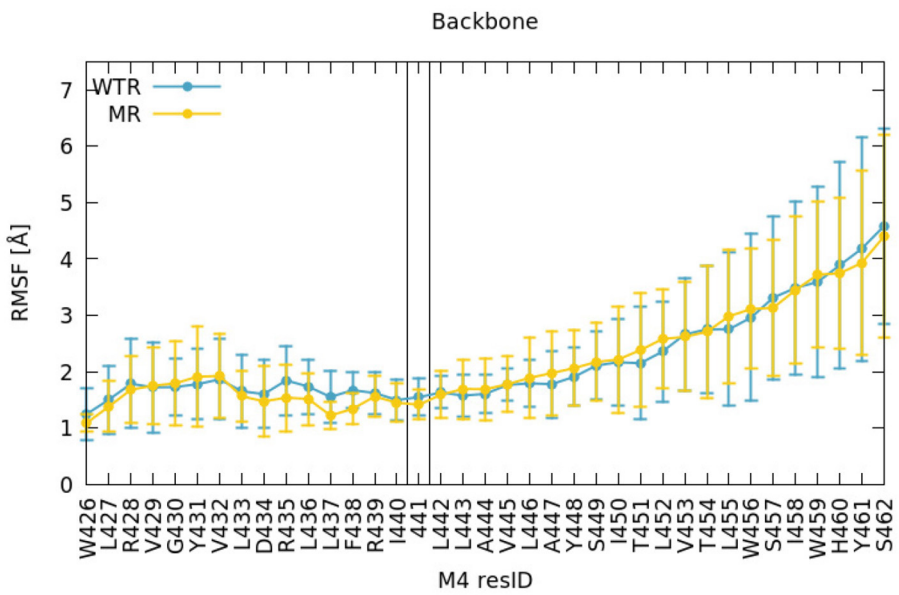


FIGURE 2 | RMSF of backbone atoms of the M4 amino acids, calculated for R01-400 for the WTR (blue) and the MR (yellow).

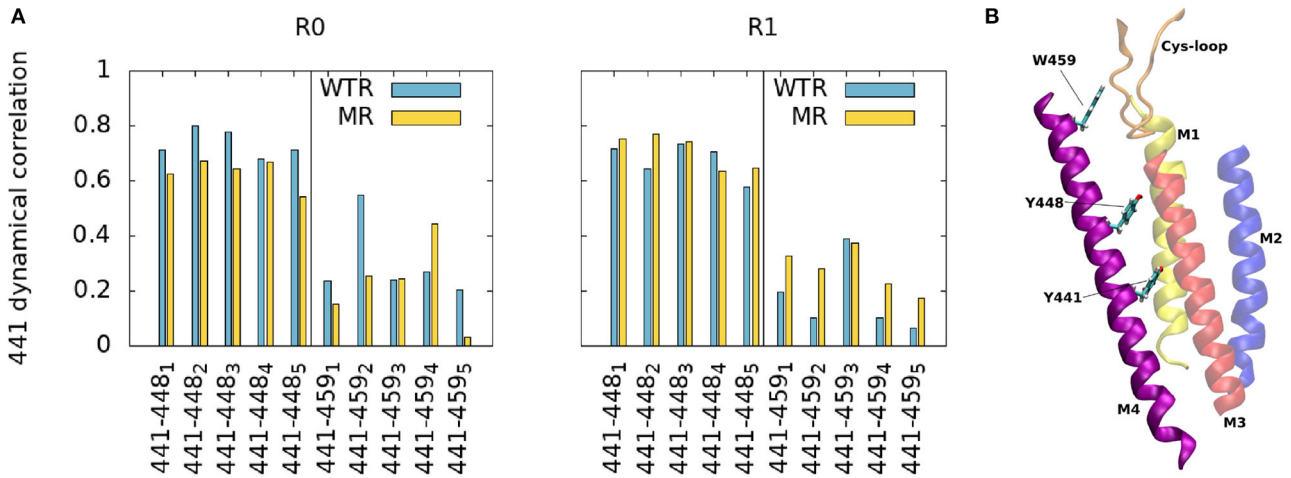


FIGURE 3 | (A) Average dynamic correlation of residue 441 backbone atoms with those of residues 448 and 459, for the WTR (blue) and the MR (yellow) in the two replicas R0 and R1. Subscripts denote the five subunits. **(B)** Position of residues 441, 448, and 459 in the protein.

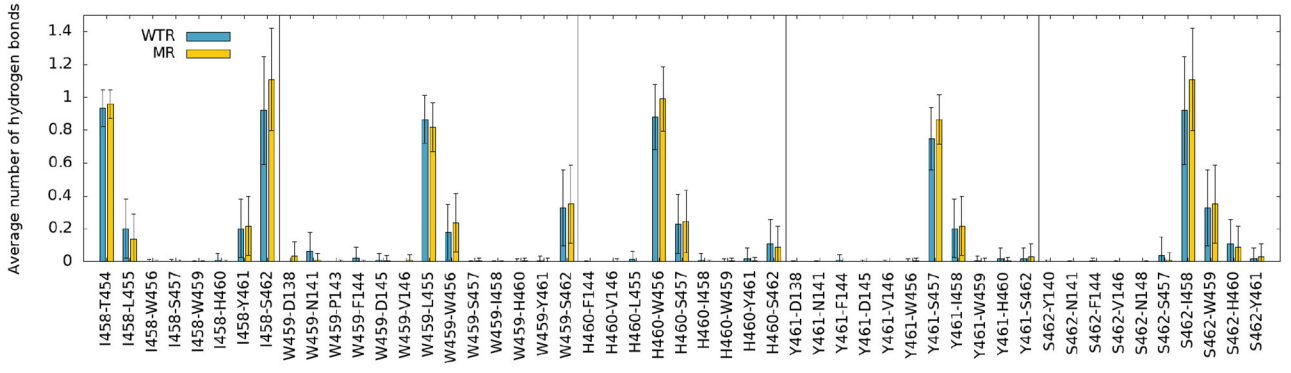


FIGURE 4 | Hydrogen bonds of M4 tip residues (I458, W459, H460, Y461, S462), calculated over R01-400 for the WTR (blue) and the MR (yellow).

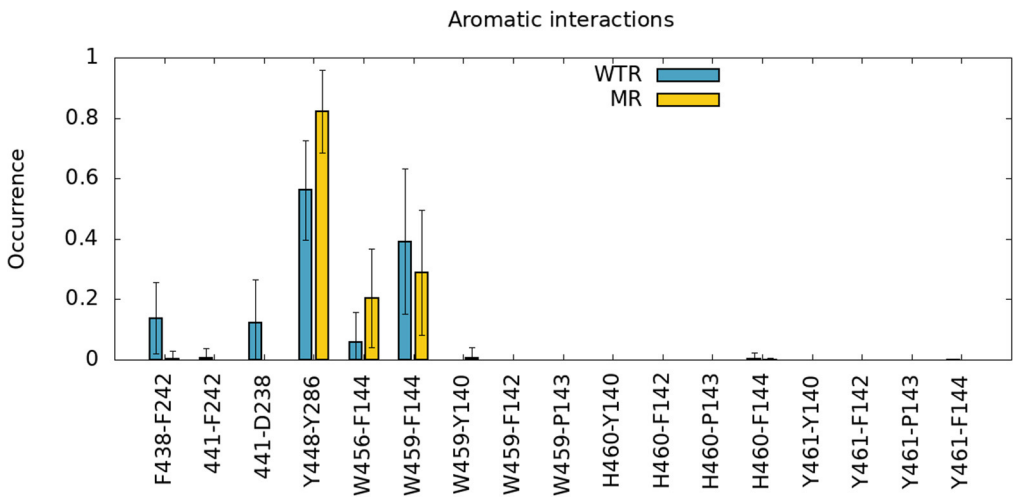


FIGURE 5 | π-π interactions (F438-F424,441-F242,Y448-Y286,W456-F144,W459-F144) and anion-π interactions (441-D238) of different amino-acids of M4, calculated over R01-400 for the WTR (blue) and the MR (yellow).

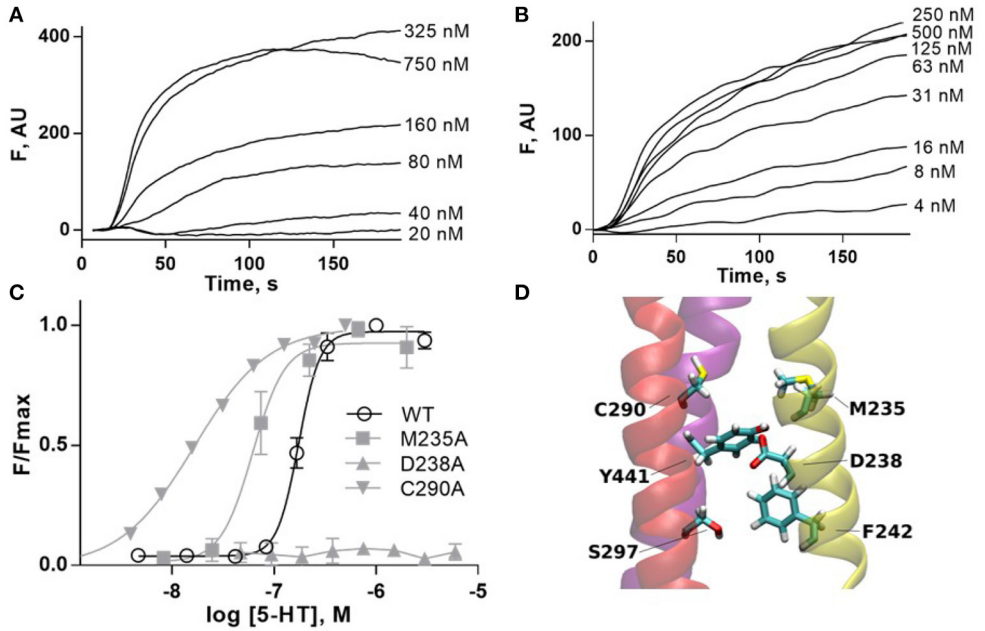


FIGURE 6 | (A,B) Typical fluorescent responses (F in arbitrary units, AU) to addition of 5-HT at 20 s to HEK293 cells expressing 5-HT_{3A} receptors, using a membrane potential sensitive dye. **(A)** WT, **(B)** C290A. **(C)** 5-HT concentration-response curves of mutant receptors in HEK293 cells. For D238A F is compared to WT Fmax. Data is mean ± standard error of the mean (SEM), $n \geq 3$. **(D)** Molecular dynamics snapshot showing Y441 in the M4 helix together with possible interaction partners. M1: yellow, M3: red, M4: purple.

3.2. Radial Mechanism

The radial mechanism was first investigated experimentally. We characterized WT and mutant 5-HT_{3A} receptors expressed in HEK293 cells by measuring responses of a membrane potential-sensitive dye on addition of 5-HT (**Figures 6A–C**). This gave a WT EC_{50} of 0.17 μM (pEC_{50} of 6.76 ± 0.01) and a Hill slope of 3.7 ± 0.3, which is comparable to previous work (Lochner and Lummis, 2010). The WT level of ligand binding at the cell surface was measured with [^{3}H]GR65630, giving a K_d of 0.18 ± 0.03 nM, similar to previous work e.g., (Hovius et al., 2002), and a B_{max} of 1.2 ± 0.8 pmol/mg protein (**Table 1**). Mutants that did not respond in the functional assay are marked as non functional (NF).

Alanine mutations of all residues of interest near Y441 showed that only D238A had a comparable effect to Y441A (**Figure 6** and **Table 1**). Neither Y441A nor D238A responded to application of 5-HT, though they both had high levels of [^{3}H] GR65630 binding sites (Mesoy et al., 2019 and **Table 1**). Their proximity and similar phenotypes on mutation indicate that D238 could be part of the mechanism of Y441 supporting channel function.

To further elucidate the role of the Y441-D238 interaction, we investigated the wider effects of Y441A through D238. In the MD simulations, the time- and subunit-averaged hydrogen bonds of D238 and Y441 with any other residue belonging to the M1, M2, M3, and M4 helices were calculated over R01-400 (**Figure 7**) for both the WTR and the MR. Error bars were calculated via error propagation from the five standard deviations of time-data for each of the single subunits.

TABLE 1 | Mutant receptors in HEK293 cells.

Mutant	EC_{50} (μM)	pEC_{50} (M)	n_H	K_d (nM)	B_{max} (pmol/mg protein)
WT	0.17	6.76 ± 0.01	3.7 ± 0.3	0.18 ± 0.03	1.2 ± 0.8
M235A	0.06	7.21 ± 0.05	2.8 ± 0.9	0.70 ± 0.09	2.4 ± 0.4
D238A		NF		0.50 ± 0.07	1.8 ± 0.4
F242A	0.31	6.51 ± 0.02	3.8 ± 0.8		
C290A	0.02	7.79 ± 0.05	1.2 ± 0.2	0.85 ± 0.10	3.8 ± 0.3
S297A	0.46	6.34 ± 0.03	3.5 ± 0.8		

K_d ± B_{max} measured by saturation binding. Values are mean and SEM, n ≥ 3.

No major differences were observed between the two models, except for the notable lack of hydrogen bonds between residues 441 and 238 in the MR. However, one interesting fact emerges: both residues 238 and 441 are able to make interactions with K255, a lysine that belongs to the M2 helix. This residue is near L260, the hydrophobic gate of the 5-HT_3R (Hassaine et al., 2014; Aryal et al., 2015). The K255 side chain stretches between M1 and M3 and points toward the middle of the four TMD helices, with its terminal nitrogen at a convenient position for the formation of hydrogen bonds with residues in the region (**Figure 8A**).

Residue 441 is within reach of the K255 side chain, so direct hydrogen bonds between these two residues are possible, however they only occurred for a tiny fraction of the simulation time. Conversely, D238 terminal oxygens formed hydrogen bonds for longer.

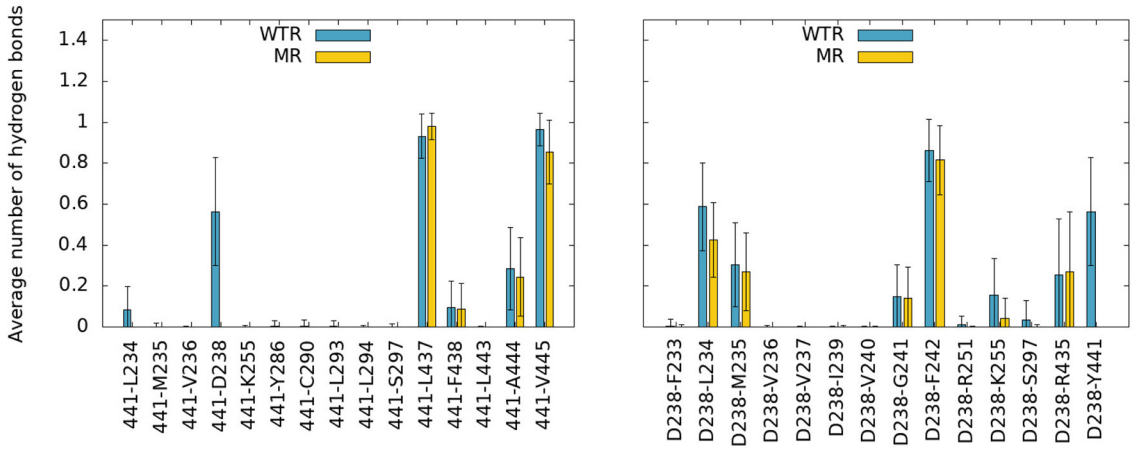


FIGURE 7 | Hydrogen bonds of residue 441 **(left)** and D238 **(right)** with any other protein residue, calculated over R01-400 for the WTR (blue) and the MR (yellow).

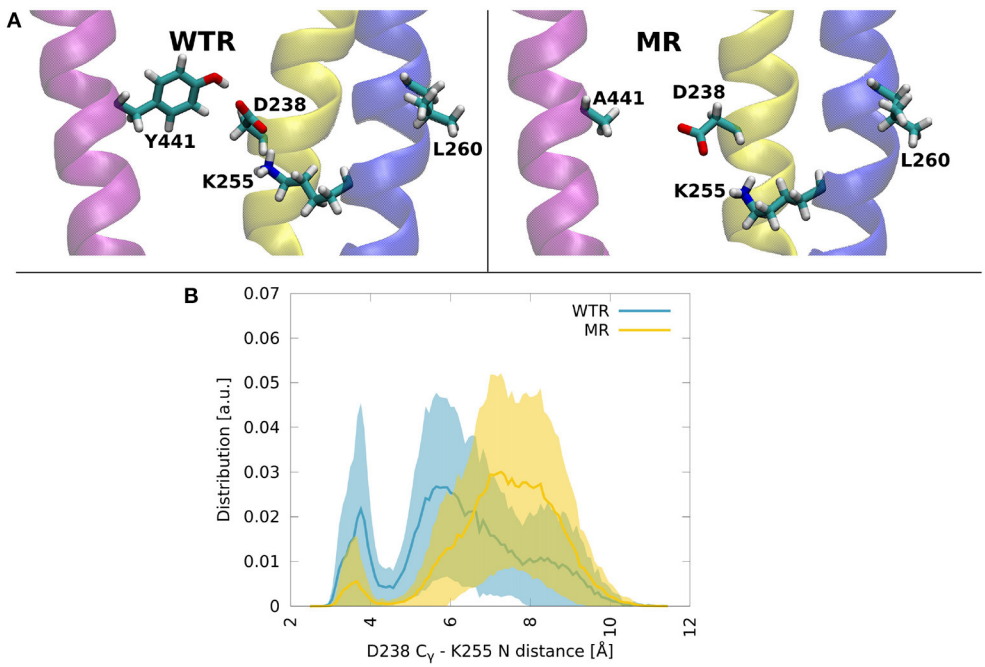


FIGURE 8 | (A) Molecular dynamics snapshots showing how Y441 might help maintain the hydrogen bonding between D238 and K255 in the WTR, while in the MR these two residues are less likely to form interactions. M1: yellow, M2: blue, M4: purple. **(B)** Distributions of the distance between C_γ of D238 and the terminal nitrogen of K255, in the two models. Thick lines: averages; shaded area: error.

The distribution over time and subunits of the distance between the C_γ of D238 and the terminal nitrogen of K255 revealed two peaks in both the WTR and the MR (**Figure 8B**). Errors were calculated via error propagation from the ten standard deviations of data over time for each of the single subunits and for the two replicas R0 and R1. One peak was found at about 7.5 Å for the MR and at about 5.5 Å for the WTR, and shows how the absence of Y441 side chain allows for D238 to be farther away from K255 with respect to the WTR. Another peak was observed around 3.5 Å for both the WTR and the MR, and was much higher for the WTR than for the MR: this indicates that a hydrogen bond may be formed between K255 and D238, which was observed for longer times in the WTR. In the WTR, the distance between C_γ of D238 and the side chain oxygen of Y441 is 4.8 ± 0.4 Å.

To investigate this putative interaction we assayed the effects of mutating K255 in HEK293 cells. K255A is indistinguishable from WT, indicating that K255 is not required for correct channel function. Intriguingly however, K255L is entirely non-responsive to ligand, even though it is expressed, as shown by radioligand binding (**Table 2**). This indicates that K255 may be part of the same interaction chain as Y441 and D238.

3.3. Rescue of Non-functional Receptors

We decided to further probe the most interesting mutants in a different expression system, *Xenopus* oocytes, using two-electrode voltage clamp.

Expressing WT 5-HT_{3A} in *Xenopus* oocytes gave an EC_{50} of 1.7 μM (pEC_{50} = 5.76 ± 0.05) and a Hill slope of 1.8 ± 0.3 (**Figure 9** and **Table 3**), similar to previous work (Lummis et al., 2016).

On expression in *Xenopus* oocytes, both Y441A and K255L (which were non-responsive in HEK cells (Mesoy et al., 2019, **Table 2**)) gave WT-like responses (**Figure 9** and **Table 3**). This demonstrates that Y441, which is required for channel function in HEK293 cells, is not in *Xenopus* oocytes. For the mutants that did function in HEK cells, we observed two different patterns in oocytes. While K255A showed similar shifts relative to WT in both expression systems, C290A was more sensitive to ligand than WT in HEK cells but less sensitive in *Xenopus* oocytes (**Figure 10**).

To assess the impact of lipid composition on channel function, we investigated the effects of lipids on Y441 and nearby residues in the simulated WTR and MR. The two models make use of the same lipid composition (POPC, POPE and cholesterol), but the local lipid environment around each of the five subunits differs due to varied lipid diffusion during the simulation. We found no difference between the two models in any of the measures described below (distributions of lipids around Y441, relative positions of lipids within the membrane with respect to Y441, and hydrogen bonds formed with residues 441 and/or 238).

In order to evaluate the fitness of residue 441 to give rise to a valid lipid binding site, we first evaluated the proximity lifetime distributions of lipids around this residue, for the two replicas (R0 and R1) separately as this quantity strictly depends over the specific replica. The results are reported for phospholipids and cholesterol in **Figure 11**. POPC and POPE are grouped together, since 441 is at the level of phospholipid tails, which are indistinguishable for POPC and POPE. For this calculation, we considered multiple time windows, shifted by 5 ns, and evaluated averages and standard deviations for each 5 ns time period. The statistics available for each residence time decreases for larger times.

These distributions are characterized by very fast decays. While cholesterol only exhibits one binding event for around 15 ns in R1 for the WTR, phospholipids display some binding events up to 25 ns. However, no event is seen for any binding duration beyond this value, meaning that possibly all interactions at the level of residue 441 are relatively weak.

We calculated the z component (where z is the parallel direction to the protein axis) of the distance of center of mass of lipids selections from the center of mass of the five 441 residues, shown in **Supplementary Figure 5**. Residue 441 is at the same height as the center of mass of cholesterol molecules and phospholipid tails of the lower (inner) leaflet. Phospholipid

TABLE 2 | Mutant receptors in HEK293 cells.

Mutant	EC_{50} (μM)	pEC_{50}	n_H	K_d (nM)	B_{max} (pmol/mg protein)
WT	0.17	6.76 ± 0.01	3.6 ± 0.3	0.18 ± 0.03	1.2 ± 0.8
K255A	0.52	6.29± 0.02	2.6± 0.1		
K255L		NF		0.17± 0.02	0.4 ± 0.2
K255Q	0.11	6.95± 0.02	1.9± 0.2		
K255E	0.26	6.59± 0.02	4.4± 1.4		
K255C	0.30	6.52± 0.03	2.3± 0.6		

K_d and B_{max} measured by saturation binding. Values are mean and SEM, n ≥ *3.*

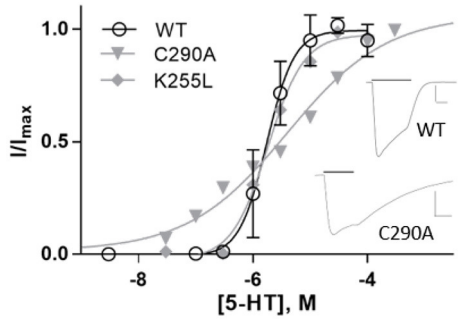


FIGURE 9 | Concentration-response curves of 5-HT_{3A} receptors in *Xenopus* oocytes. Data is mean ± SEM, n≥3. Inset: typical current recordings at 3 μM 5-HT; scale bars are 20 s and 2 μA.

TABLE 3 | Mutant receptors in *Xenopus* oocytes.

Mutant	EC_{50} (μM)	pEC_{50}	n_H
WT	1.7	5.76 ± 0.05	1.8 ± 0.3
Y441A	1.0	5.98 ± 0.10	1.0 ± 0.1
D238A		NF	
K255A	3.3	5.48 ± 0.06	1.9 ± 0.5
K255L	1.8	5.74 ± 0.05	1.4 ± 0.2
K255Q	1.2	5.93 ± 0.04	2.1 ± 0.4
C290A	4.0	5.40 ± 0.15	0.5 ± 0.1

Values are mean and SEM, n ≥ *3.*

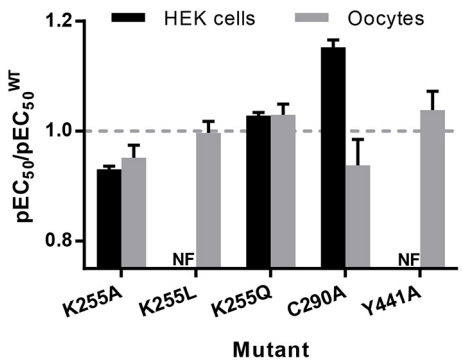


FIGURE 10 | pEC_{50} relative to WT for mutants expressed in HEK293 cells and *Xenopus* oocytes, from **Tables 1**–**3**. Data is mean ± SEM, $n \geq 3$; values less than 1 indicate loss-of-function, and values greater than 1 indicate gain-of-function.

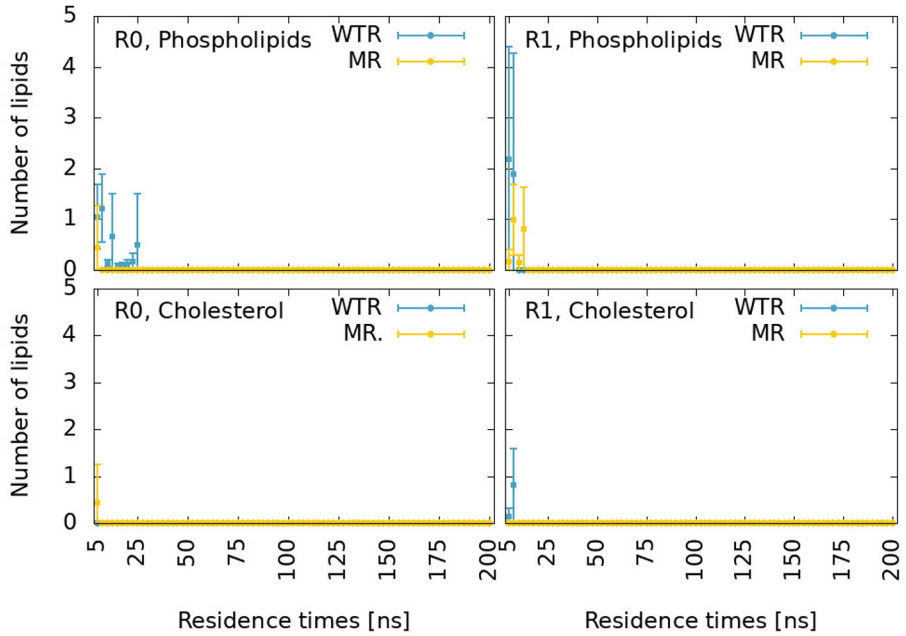


FIGURE 11 | Proximity lifetime distributions of any phospholipid **(top)** and of any cholesterol **(bottom)** in close proximity to residue 441 in the WTR (blue) and in the MR (yellow) for replicas R0 **(left)** and R1 **(right)**. The results are shown for residence times larger than 5 ns.

tails may disrupt interactions between residue 441 and D238, and could intercalate within the subunit (Crnjar and Molteni, 2020). Phospholipid heads are on average quite far from residue 441, but might still form sporadic hydrogen bonds with this residue or with D238. Cholesterol could form π-π interactions with the side chain of residue 441 when present (i.e., in the WTR), sporadic hydrogen bonds with either residue 441 or D238, or conversely could intercalate within the subunit (possibly aided by the lack of the side chain of residue 441 in the MR). A previous *in-silico* study observed cholesterol interacting with Y441 by means of π-π interactions as well as hydrogen bonds (Guros et al., 2020).

The hydrogen bonds formed between residue 441 (or D238) and lipids, calculated over R01-400 for the two models, are displayed in **Supplementary Table 2** and depicted in **Figure 12**.

Only isolated and weak hydrogen bonds were observed between lipids and residues 441 and 238. While their average values are low (or even zero), their effects are still interesting. In the WTR, subunit 5 (**Figure 12A**), a POPE molecule formed hydrogen bonds with both Y441 and D238, pulling D238 away from K255. In the MR, a cholesterol molecule was observed making hydrogen bonds in subunit 1 that pushed D238 outward and away from K255 (**Figure 12B$_1$**), but this did not result in a pulling of D238 at other timepoints (**Figure 12B$_2$**). Similarly, a POPE molecule in subunit 1 engaged in hydrogen bonds with D238 that did not prevent it from also interacting with K255 (**Figure 12C**), but another POPE, in subunit 5, instead pulled D238 outward when interacting with it (**Figure 12D**). Overall, no lipid interactions in this region were observed to promote or inhibit intrasubunit interactions at the level of Y441, nor to prefer any particular conformation or interaction over any other, although further work beyond the scope of this study is required to confirm this, e.g., by using coarse-grained methods, in order to produce a larger statistics of possible lipid interactions with or near Y441.

4. DISCUSSION AND CONCLUSIONS

The aim of this work was to probe the effects of the M4 helix in 5-HT$_{3A}$R function using the non-functional Y441A mutation. Our data suggest that Y441 connects to K255 on the pore-lining M2 helix via the M1 residue D238, and that this interaction is necessary for receptor function in HEK cells but not in *Xenopus* oocytes. We found no indications that Y441 or Y441A-mediated uncoupling affect the M4 tip or the Cys-loop.

4.1. Vertical Mechanism of Connection

In our simulations of both the WTR and the MR, the C-terminal domain appeared to move independently of residue 441 (**Figure 2**), indicating that Y441A does not act through large-scale shifts in M4 movement. We further confirmed the independence of the CTD from the Y441 by dynamical correlation measurements (**Figure 3**), and found no differences in hydrogen bonds from residues other than 441 between the two models (**Figures 4**, **7**). From this we conclude that despite the structural appeal and the early indications pointing to a vertical mechanism of action, M4-mediated coupling does not occur via interactions of the M4 tip with the Cys-loop, as the CTD of M4 is unaffected by the mutation that disconnects channel opening from ligand binding.

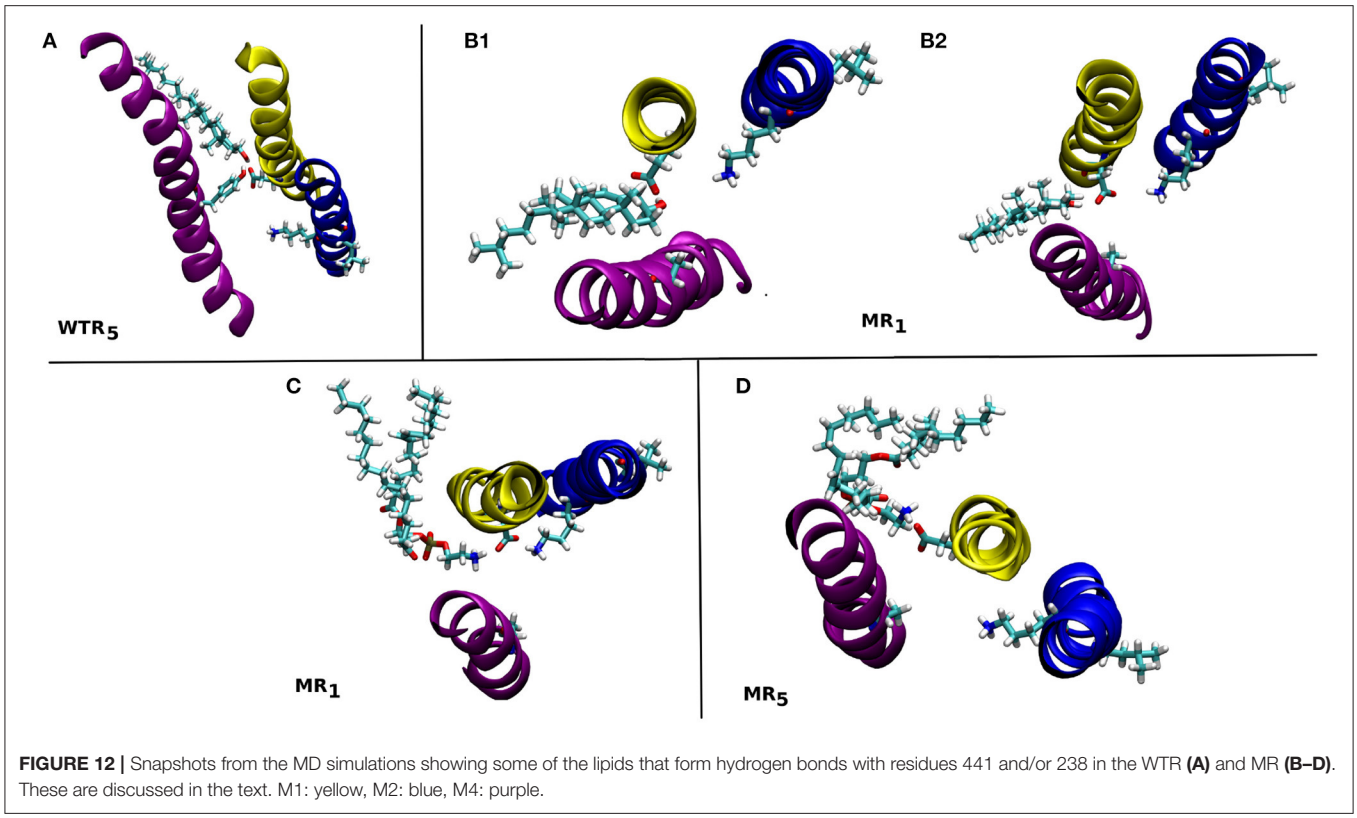


FIGURE 12 | Snapshots from the MD simulations showing some of the lipids that form hydrogen bonds with residues 441 and/or 238 in the WTR **(A)** and MR **(B–D)**. These are discussed in the text. M1: yellow, M2: blue, M4: purple.

4.2. Radial Mechanism of Connection

The possibility of a radial mechanism of connection from M4 to the channel pore is appealing: substitution of a sufficiently different residue for Y441, D238, or K255 individually abolished receptor function but not ligand binding in HEK293 cells (Mesoy et al., 2019, **Tables 1**, **2**). However, these substitutions are not equal. The aromatic group of Y441 must be key to its function, as Y441F has WT-like function where Y441A does not. Likewise some specific property of D238 is required for correct channel function, as evidenced by the lack of function of D238A. K255 is subtly different: the data suggest the residue at this position must be polar or charged if it is large: K255E, C, and Q are all functional, but K255L—bulky and uncharged—results in non-functional receptors. However, a large polar residue is not strictly required here, as evidenced by the WT-like function of K255A.

We propose that a major requirement of residue 255 may be to allow the displacement of M2 on channel opening. The movement of M2 on channel opening in 5-HT_{3A} has been described as a rotation and outward displacement (Basak et al., 2018b). Polovinkin et al. show an outward movement of M2 (Polovinkin et al., 2018), especially the lower half, as well as a rotation to clear the restricting L9' (L260) residues from the center of the pore. We propose that residue 255 requires a polar character here to allow this outward movement if the residue is large (hence K255L blocking channel opening), perhaps involving the observed hydrogen bond to D238 (**Figure 7**). However mutation of K255 to a small residue like alanine may also allow outward movement of this helix, explaining the mutation pattern here.

Conversely the removal by alanine mutation of either Y441 or D238 abolishes channel function. This, too, may be related to the movement of M2 on channel opening (and hence explain why these mutations prevent it). Figure 2C in Basak et al. (2018b) and the "morph" videos in Polovinkin et al. (2018) show particularly well the movement of M1 and M4 on channel opening—outward and, for M4, upward. It seems likely that Y441 and D238 may be required for this movement of their respective helices, and that this helical movement is required for the outward channel-opening movement of M2 discussed above.

While each residue may act individually, it is also striking that these three residues in close proximity give the same functional phenotype on mutation. A Y441-D238 interaction in particular is likely to be required for channel function, though a hydrogen bond is not (as Y441F is WT-like). K255 is not specifically required, though a large non-polar residue at this location is disruptive. While K255A does function, the requirement for any larger residue at position 255 to have some polarity—be it a positive charge, a negative charge, or only a polar group—does point intriguingly toward the hydrogen bond noted with D238 (**Figure 7**). Determining precisely which of these residues interact and how will be key to understanding the wider mechanism of channel opening in 5-HT_3R receptors.

4.3. Rescue of Non-functional Receptors

The stark difference between the uncoupled state of Y441A and K255L in HEK293 cells and their WT-like behavior in *Xenopus* oocytes is intriguing. Putting this in the context of the literature, a wider pattern emerges where mutations in cationic pLGICs M4

helices that affect EC_{50} are slightly beneficial in *Xenopus* oocytes, but detrimental in HEK293 cells:

Many (11 out of 24) alanine mutations in the α7 nAChR M4 helix improve function in oocytes (da Costa Couto et al., 2020). In contrast, several (8 out of 27) alanine mutations in the $\alpha4\beta2$ nAChR M4 helix abolish function (but not ligand binding) when expressed in HEK293 cells (Mesoy and Lummis, 2021). Looking only at the C-terminal end of the M4 helix, it seems that while it can be deleted without ablating function in both ELIC (Hénault et al., 2015) and the *Torpedo* nAChR (Tobimatsu et al., 1987), deletion or alanine mutation of individual C-terminal residues abolish function in the 5-HT_{3A}R and in the $\alpha4\beta2$ nAChR (Pons et al., 2004; Butler et al., 2009; Mesoy et al., 2019; Mesoy and Lummis, 2021). However we note that the ELIC and nAChR studies were performed in *Xenopus* oocytes, and the other four in HEK293 cells, indicating that the requirement for the C-terminal domain may be more a function of the expression system than of the specific channel. An exception to this pattern is that alanine mutations in the M4 of the α subunit of the muscle nAChR expressed in oocytes show both gains and losses of function (Thompson et al., 2020). We note that these mutations are only present in 2 out of 5 subunits per channel, and what would happen in a muscle AChR with all 5 positions mutated is as yet unknown.

This variation in channel function of M4 mutants between expression systems has not been observed in anionic or bacterial channels. In anionic channels, alanine mutations (especially of aromatic residues) are generally detrimental to channel function, regardless of expression system (Haeger et al., 2010; Cory-Wright et al., 2018; Tang et al., 2018). Mutations in M4 have opposite effects in two bacterial pLGICs assayed in the same system (*Xenopus* oocytes): Many (15 out of 25) alanine mutations in the GLIC M4 are detrimental to channel function, while a majority (26 out of 31) of alanine mutations in the ELIC M4 improve channel function (Hénault et al., 2015).

We suggest that there exists a functional mechanism in cationic pLGICs requiring the M4 helix (including the C-terminus) which is necessary in HEK cells but not in *Xenopus* oocytes. If so, firstly conclusions about M4 function from studies in *Xenopus* oocytes cannot be extended to other expression systems, specifically not HEK cells, and vice versa. Secondly, this would indicate that some factor is either present in oocytes that can rescue these mutants, or present in HEKs that inhibits them. Due to the locations of these mutations in the lipid bilayer, we suggest that this is unlikely to be an intracellular factor or a post-translational modification. The proximity of these mutations to the lipid bilayer, along with the wide range of previously characterized lipid-uncoupled receptors, points us to the hypothesis that some element of the lipid bilayer of *Xenopus* oocytes is able to compensate for the absence of Y441. A speculative but attractive option, given the sensitivity of pLGIC activity to the local lipid environment, is that the *Xenopus* "rescue" factor could be cholesterol. Adding cholesterol (and/or negatively charged phospholipids) to reconstituted membranes promotes pLGIC function (Fong and McNamee, 1986; Baenziger et al., 2000). The importance of cholesterol in particular is highlighted by the fact that increasing the percentage of cholesterol in a reconstituted membrane increases the number of nAChRs that open on agonist binding (Rankin et al., 1997).

With regards to our simulations, an inhibitory factor in HEK cells would not necessarily be visible; indeed observed interactions and binding events depend on the local lipid environment around each of the five subunits and can only provide a hint of the role of the involved lipid species. Care must also be taken when comparing experiments and simulations, since the former were carried out at room temperature and the latter at body temperature (in order to keep the modeled membrane above the gel transition temperature of all lipids present).

We cannot as yet, however, rule out an endogenous or exogenous factor which does not act on Y441 itself. From our simulations, we were able to conclude that only sporadic lipid binding events or interactions with D238 occurred at the level of Y441 (**Figure 11**), and that no promotion/inhibition of the radial mechanism was unambiguously observed by lipid molecules. Given that the cholesterol content has been shown to increase the chances of any lipid binding event in the 5-HT_{3A}R receptor (Crnjar and Molteni, 2020), we can speculate that the oocyte membrane, with higher cholesterol content, could promote additional binding events near residue 441.

In conclusion, we have thoroughly investigated the role of a point mutation (Y441A) in the 5-HT_{3A}R M4 helix, using it as a proxy for the role and function of the entire helix in coupling ligand binding to channel opening, using both *in-silico* techniques and experiments in two different expression systems. We showed that Y441-mediated coupling involves D238 on M1 and K255 on M2, creating a radial chain from the channel pore to the lipid-facing M4. No effect is propagated vertically from Y441 toward the M4 tip or the Cys-loop, leading us to conclude that Y441-mediated coupling specifically, and M4-mediated coupling in general does not depend on M4/Cys-loop interactions.

Finally, we speculate that the rescue of uncoupled mutants in *Xenopus* oocytes may be due to the lipid composition of the oocytes, and suggest cholesterol as a potential candidate for rescuing receptors that are non-functional yet expressed in HEK cells.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article is available at: http://doi.org/doi:10.18742/RDM01-701.

AUTHOR CONTRIBUTIONS

AC, SM, SL, and CM participated in the research design and wrote and contributed to the manuscript. AC performed the simulations. SM conducted the experiments. AC, SM, and SL performed data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

AC and CM are grateful for computational support from the UK high performance computing service ARCHER, for which access was obtained via the UKCP consortium and funded by EPSRC grant EP/P022472/1; they also acknowledge the UK Materials and Molecular Modeling Hub for computational resources, which is partially funded by EPSRC grants EP/P020194/1 and EP/T022213/1. For the experimental work, SM was supported by an AstraZeneca studentship. SL was supported by MRC grant MR L021676.

ACKNOWLEDGMENTS

We thank Dr. Alejandro Santana-Bonilla (King's College London) for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.644720/full#supplementary-material

REFERENCES

Alcaino, C., Musgaard, M., Minguez, T., Mazzaferro, S., Faundez, M., Iturriaga-Vasquez, P., et al. (2017). Role of the cys loop and transmembrane domain in the allosteric modulation of $\alpha4\beta2$ nicotinic acetylcholine receptors. *J. Biol. Chem.* 292, 551–562. doi: 10.1074/jbc.M116.751206

Althoff, T., Hibbs, R. E., Banerjee, S., and Gouaux, E. (2014). X-ray structures of GluCl in apo states reveal a gating mechanism of Cys-loop receptors. *Nature* 512, 333–337. doi: 10.1038/nature13669

Aryal, P., Sansom, M. S., and Tucker, S. J. (2015). Hydrophobic gating in ion channels. *J. Mol. Biol.* 427, 121–130. doi: 10.1016/j.jmb.2014.07.030

Baenziger, J. E., Morris, M. L., Darsaut, T. E., and Ryan, S. E. (2000). Effect of membrane lipid composition on the conformational equilibria of the nicotinic acetylcholine receptor. *J. Biol. Chem.* 275, 777–784. doi: 10.1074/jbc.275.2.777

Basak, S., Gicheru, Y., Rao, S., Sansom, M. S. P., and Chakrapani, S. (2018a). Cryo-em reveals two distinct serotonin-bound conformations of full-length 5-HT3A receptor. *Nature* 563, 270–274. doi: 10.1038/s41586-018-0660-7

Basak, S., Gicheru, Y., Samanta, A., Molugu, S. K., Huang, W., la de Fuente, M., et al. (2018b). Cryo-em structure of 5-HT3A receptor in its resting conformation. *Nat. Commun.* 9:514. doi: 10.1038/s41467-018-02997-4

Bocquet, N., Nury, H., Baaden, M., Le Poupon, C., Changeux, J.-P., Delarue, M., et al. (2009). X-ray structure of a pentameric ligand-gated ion channel in an apparently open conformation. *Nature* 457, 111–114. doi: 10.1038/nature07462

Butler, A. S., Lindesay, S. A., Dover, T. J., Kennedy, M. D., Patchell, V. B., Levine, B. A., et al. (2009). Importance of the C-terminus of the human 5-HT3A receptor subunit. *Neuropharmacology* 56, 292–302. doi: 10.1016/j.neuropharm.2008.08.017

Cao, R., Rossetti, G., Bauer, A., and CarIoni, P. (2015). Binding of the antagonist caffeine to the human adenosine receptor hA2AR in nearly physiological conditions. *PLoS ONE* 10:e0126833. doi: 10.1371/journal.pone.0126833

Chan, R. B., Oliveira, T. G., Cortes, E. P., Honig, L. S., Duff, K. E., Small, S. A., et al. (2012). Comparative lipidomic analysis of mouse and human brain with alzheimer disease. *J. Biol. Chem.* 287, 2678–2688. doi: 10.1074/jbc.M111.274142

Comitani, F., Cohen, N., Ashby, J., Botten, D., Lummis, S. C. R., and Molteni, C. (2014). Insights into the binding of GABA to the insect rdl receptor from atomistic simulations: a comparison of models. *J. Comput. Aided Mol. Design* 28, 35–48. doi: 10.1007/s10822-013-9704-0

Comitani, F., Limongelli, V., and Molteni, C. (2016). The free energy landscape of GABA binding to a pentameric ligand-gated ion channel and its disruption by mutations. *J. Chem. Theory Comput.* 12, 3398–3406. doi: 10.1021/acs.jctc.6b00303

Comitani, F., Melis, C., and Molteni, C. (2015). Elucidating ligand binding and channel gating mechanisms in pentameric ligand-gated ion channels by atomistic simulations. *Biochem. Soc. Trans.* 43, 151–156. doi: 10.1042/BST20140259

Cory-Wright, J., Alqazzaz, M., Wroe, F., Jeffreys, J., Zhou, L., and Lummis, S. C. R. (2018). Aromatic residues in the fourth transmembrane-spanning helix M4 are important for GABAρ receptor function. *ACS Chem. Neurosci.* 9, 284–290. doi: 10.1021/acschemneuro.7b00315

Crnjar, A., Comitani, F., Hester, W., and Molteni, C. (2019a). Trans-cis proline switches in a pentameric ligand-gated ion channel: how they are affected by and how they affect the biomolecular environment. *J. Phys. Chem. Lett.* 10, 694–700. doi: 10.1021/acs.jpclett.8b03431

Crnjar, A., Comitani, F., Melis, C., and Molteni, C. (2019b). Mutagenesis computer experiments in pentameric ligand-gated ion channels: the role of simulation tools with different resolution. *Interface Focus* 9:20180067. doi: 10.1098/rsfs.2018.0067

Crnjar, A., and Molteni, C. (2020). Cholesterol content in the membrane promotes 2 key lipid-protein interactions in a pentameric 3 serotonin-gated ion channel. *Biointerphases* 15:161018. doi: 10.1116/6.0000561

da Costa Couto, A. R., Price, K. L., Mesoy, S., Capes, E., and Lummis, S. C. R. (2020). The M4 helix is involved in $\alpha7$ nach receptor function. *ACS Chem. Neurosci.* 11, 1406–1412. doi: 10.1021/acschemneuro.0c00027

DaCosta, C. J. B., and Baenziger, J. E. (2009). A lipid-dependent uncoupled conformation of the acetylcholine receptor. *J. Biol. Chem.* 284, 17819–17825. doi: 10.1074/jbc.M900030200

Dämgen, M. A., and Biggin, P. C. (2020). State-dependent protein-lipid interactions of a pentameric ligand-gated ion channel in a neuronal membrane. *bioRxiv* 28, 130–139.e2. doi: 10.1101/2020.04.07.029603

Dawaliby, R., Trubbia, C., Delporte, C., Noyon, C., Ruysschaert, J.-M., Van Antwerpen, P., et al. (2016). Phosphatidylethanolamine is a key regulator of membrane fluidity in eukaryotic cells. *J. Biol. Chem.* 291, 3658–3667. doi: 10.1074/jbc.M115.706523

Deol, S. S., Bond, P. J., Domene, C., and Sansom, M. S. (2004). Lipid-protein interactions of integral membrane proteins: a comparative simulation study. *Biophys. J.* 87, 3737–3749. doi: 10.1529/biophysj.104.048397

Dickson, C. J., Madej, B. D., Skjevik, Å. A., Betz, R. M., Teigen, K., Gould, I. R., et al. (2014). Lipid14: The amber lipid force field. *J. Chem. Theory Comput.* 10, 865–879. doi: 10.1021/ct4010307

Domville, J. A., and Baenziger, J. E. (2018). An allosteric link connecting the lipid-protein interface to the gating of the nicotinic acetylcholine receptor. *Sci. Rep.* 8:3898. doi: 10.1038/s41598-018-22150-x

Du, J., Lu, W., Wu, S., Cheng, Y., and Gouaux, E. (2015). Glycine receptor mechanism elucidated by electron cryo-microscopy. *Nature* 526, 224–245. doi: 10.1038/nature14853

Elmore, D. E., and Dougherty, D. A. (2003). Investigating lipid composition effects on the mechanosensitive channel of large conductance (MSCL) using molecular dynamics simulations. *Biophys. J.* 85, 1512–1524. doi: 10.1016/S0006-3495(03)74584-6

Fong, T. M., and McNamee, M. G. (1986). Correlation between acetylcholine receptor function and structural properties of membranes. *Biochemistry* 25, 830–840. doi: 10.1021/bi00352a015

Gowers, R. J., Linke, M., Barnoud, J., Reddy, T. J. E., Melo, M. N., Seyler, S. L., et al. (2016). "MDanalysis: a python package for the rapid analysis of molecular dynamics simulations," in *Proceedings of the 15th Python in Science Conference* (Austin, TX), 98–105. doi: 10.25080/Majora-629e541a-00e

Guros, N. B., Balijepalli, A., and Klauda, J. B. (2020). Microsecond-timescale simulations suggest 5-HT-mediated preactivation of the 5-HT_{3A} serotonin receptor. *Proc. Natl. Acad. Sci. U.S.A.* 117, 405–414. doi: 10.1073/pnas.1908848117

Haeger, S., Kuzmin, D., Detro-Dassen, S., Lang, N., Kilb, M., Tsetlin, V., et al. (2010). An intramembrane aromatic network determines pentameric assembly of Cys-loop receptors. *Nat. Struct. Mol. Biol.* 17, 90–99. doi: 10.1038/nsmb.1721

Hassaine, G., Deluz, C., Grasso, L., Wyss, R., Tol, M. B., Hovius, R., et al. (2014). X-ray structure of the mouse serotonin 5-HT_3 receptor. *Nature* 512, 276–281. doi: 10.1038/nature13552

Hénault, C. M., Juranka, P. F., and Baenziger, J. E. (2015). The M4 transmembrane α-helix contributes differently to both the maturation and function of two prokaryotic pentameric ligand-gated ion channels. *J. Biol. Chem.* 290, 25118–25128. doi: 10.1074/jbc.M115.676833

Heusser, S. A., Lycksell, M., Wang, X., McComas, S. E., Howard, R. J., and Lindahl, E. (2018). Allosteric potentiation of a ligand-gated ion channel is mediated by access to a deep membrane-facing cavity. *Proc. Natl. Acad. Sci. U.S.A.* 115, 10672–10677. doi: 10.1073/pnas.1809650115

Hilf, R. J. C., and Dutzler, R. (2008). X-ray structure of a prokaryotic pentameric ligand-gated ion channel. *Nature* 452, 375–379. doi: 10.1038/nature06717

Hovius, R., Tairi, A.-P., Blasey, H., Bernard, A., Lundström, K., and Vogel, H. (2002). Characterization of a mouse serotonin 5-HT3 receptor purified from mammalian cells. *J. Neurochem.* 70, 824–834. doi: 10.1046/j.1471-4159.1998.70020824.x

Huang, X., Chen, H., Michelsen, K., Schneider, S., and Shaffer, P. L. (2015). Crystal structure of human glycine receptor-α3 bound to antagonist strychnine. *Nature* 526, 277–290. doi: 10.1038/nature14972

Hunenberger, P., Mark, A., and van Gunsteren, W. (1995). Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *J. Mol. Biol.* 252, 492–503. doi: 10.1006/jmbi.1995.0514

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

Kandt, C., Ash, W. L., and Tieleman, D. P. (2007). Setting up and running molecular dynamics simulations of membrane proteins. *Methods* 41, 475–488. doi: 10.1016/j.ymeth.2006.08.006

Kraske, W. V., and Mountcastle, D. B. (2001). Effects of cholesterol and temperature on the permeability of dimyristoylphosphatidylcholine bilayers near the chain melting phase transition. *Biochim. Biophys. Acta Biomemb.* 1514, 159–164. doi: 10.1016/S0005-2736(01)00379-0

Kudryashev, M., Castano-Diez, D., Deluz, C., Hassaine, G., Graf-Meyer, A., Vogel, H., et al. (2016). The structure of the mouse serotonin 5-HT_3 receptor in lipid vesicles. *Structure* 24, 165–170. doi: 10.1016/j.str.2015.11.004

Lemoine, D., Jiang, R., Taly, A., Chataigneau, T., Specht, A., and Grutter, T. (2012). Ligand-gated ion channels: new insights into neurological disorders and ligand recognition. *Chem. Rev.* 112, 6285–6318. doi: 10.1021/cr3000829

Lochner, M., and Lummis, S. C. (2010). Agonists and antagonists bind to an A-A interface in the heteromeric 5-HT3AB receptor. *Biophys. J.* 98, 1494–1502. doi: 10.1016/j.bpj.2009.12.4313

Lucas, X., Bauza, A., Frontera, A., and Quinonero, D. (2016). A thorough anion-π interaction study in biomolecules: on the importance of cooperativity effects. *Chem. Sci.* 7, 1038–1050. doi: 10.1039/C5SC01386K

Lummis, S. C. R., McGonigle, I. Ashby, J. A., Dennis A. (2016). Two amino acid residues contribute to a cation-π binding interaction in the binding site of an insect GABA receptor. *J. Neurosci.* 31, 12371–12376. doi: 10.1523/JNEUROSCI.1610-11.2011

Lummis, S. C. R., and Thompson, A. J. (2013). Agonists and antagonists induce different palonosetron dissociation rates in 5-HT3A and 5-HT3AB receptors. *Neuropharmacology* 73, 241–246. doi: 10.1016/j.neuropharm.2013.05.010

Mahmood, M. I., Liu, X., Neya, S., and Hoshino, T. (2013). Influence of lipid composition on the structural stability of g-protein coupled receptor. *Chem. Pharmaceut. Bull.* 61, 426–437. doi: 10.1248/cpb.c12-01059

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* 11, 3696–3713. doi: 10.1021/acs.jctc.5b00255

McCormack, T. J., Melis, C., Colon, J., Gay, E. A., Mike, A., Karoly, R., et al. (2010). Rapid desensitization of the rat α7 nachr is facilitated by the presence of a proline residue in the outer β-sheet. *J. Physiol.* 588, 4415–4429. doi: 10.1113/jphysiol.2010.195495

Melis, C., Lummis, S. C. R., and Molteni, C. (2008). Molecular dynamics simulations of GABA binding to the GABA(C) receptor: the Role of Arg(104). *Biophys. J.* 95, 4115–4123. doi: 10.1529/biophysj.107.127589

Mesoy, S., Jeffreys, J., and Lummis, S. C. R. (2019). Characterization of residues in the 5-HT_3 receptor M4 region that contribute to function. *ACS Chem. Neurosci.* 10, 3167–3172. doi: 10.1021/acschemneuro.8b00603

Mesoy, S. M., and Lummis, S. C. R. (2021). M4, the outermost helix, is extensively involved in opening of the $\alpha 4\beta 2$ nACh receptor. *ACS Chem. Neurosci* 12, 133–139. doi: 10.1021/acschemneuro.0c00618

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). MDanalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* 32, 2319–2327. doi: 10.1002/jcc.21787

Miller, P. S., and Aricescu, A. R. (2014). Crystal structure of a human GABAA receptor. *Nature* 512, 270–275. doi: 10.1038/nature13293

Mitra, A., Bailey, T. D., and Auerbach, A. L. (2004). Structural dynamics of the M4 transmembrane segment during acetylcholine receptor gating. *Structure* 12, 1909–1918. doi: 10.1016/j.str.2004.08.004

Nemecz, Á., Prevost, M. S., Menny, A., and Corringer, P.-J. (2016). Emerging molecular mechanisms of signal transduction in pentameric ligand-gated ion channels. *Neuron* 90, 452–470. doi: 10.1016/j.neuron.2016.03.032

Nys, M., Wijckmans, E., Farinha, A., Yoluk, Ö., Andersson, M., Brams, M., et al. (2016). Allosteric binding site in a cys-loop receptor ligand-binding domain unveiled in the crystal structure of elic in complex with chlorpromazine. *Proc. Natl. Acad. Sci. U.S.A.* 113, E6696–E6703. doi: 10.1073/pnas.1603101113

Oakes, V., and Domene, C. (2019). Influence of cholesterol and its stereoisomers on members of the serotonin receptor family. *J. Mol. Biol.* 431, 1633–1649. doi: 10.1016/j.jmb.2019.02.030

Opekarová, M., and Tanner, W. (2003). "Specific lipid requirements of membrane proteins - a putative bottleneck in heterologous expression," in *Biochimica et Biophysica Acta - Biomembranes* (Elsevier), 11–22. doi: 10.1016/S0005-2736(02)00708-3

Patra, S. M., Chakraborty, S., Shahane, G., Prasanna, X., Sengupta, D., Maiti, P. K., et al. (2015). Differential dynamics of the serotonin1A receptor in membrane bilayers of varying cholesterol content revealed by all atom molecular dynamics simulation. *Mol. Memb. Biol.* 32, 127–137. doi: 10.3109/09687688.2015.1096971

Pfrieger, F. W. (2003). Role of cholesterol in synapse formation and function. *Biochim. Biophys. Acta Biomemb.* 1610, 271–280. doi: 10.1016/S0005-2736(03)00024-5

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

Polovinkin, L., Hassaine, G., Perot, J., Neumann, E., Jensen, A. A., Lefebvre, S. N., et al. (2018). Conformational transitions of the serotonin 5-HT_3 receptor. *Nature* 563, 275–296. doi: 10.1038/s41586-018-0672-3

Pons, S., Sallette, J., Bourgeois, J. P., Taly, A., Changeux, J. P., and Devillers-Thiery, A. (2004). Critical role of the C-terminal segment in the maturation and export to the cell surface of the homopentameric alpha7-5HT3A receptor. *Eur. J. Neurosci.* 20, 2022–2030. doi: 10.1111/j.1460-9568.2004.03673.x

Price, K. L., and Lummis, S. C. (2005). Flexstation examination of 5-HT_3 receptor function using ca2+- and membrane potential-sensitive dyes: advantages and potential problems. *J. Neurosci. Methods* 149, 172–177. doi: 10.1016/j.jneumeth.2005.05.014

Rankin, S. E., Addona, G. H., Kloczewiak, M. A., Bugge, B., and Miller, K. W. (1997). The cholesterol dependence of activation and fast desensitization of the nicotinic acetylcholine receptor. *Biophys. J.* 73, 2446–2455. doi: 10.1016/S0006-3495(97)78273-0

Roe, D. R., and Cheatham, T. E. (2013). Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* 9, 3084–3095. doi: 10.1021/ct400341p

Sauguet, L., Shahsavar, A., Poitevin, F., Huon, C., Menny, A., Nemecz, Á., et al. (2014). Crystal structures of a pentameric ligand-gated ion channel provide a mechanism for activation. *Proc. Natl. Acad. Sci. U.S.A.* 111, 966–971. doi: 10.1073/pnas.1314997111

Schrödinger, L. L. C. (2015). *The PyMOL Molecular Graphics System, Version 1.8.* Available online at: https://pymol.org

Shan, J., Khelashvili, G., Mondal, S., Mehler, E. L., and Weinstein, H. (2012). Ligand-dependent conformations and dynamics of the serotonin 5-HT(2A) receptor determine its activation and membrane-driven oligomerization properties. *PLoS Comput. Biol.* 8:e1002473. doi: 10.1371/journal.pcbi.1002473

Silvius, J. (1982). "Thermotropic phase transitions of pure lipids in model membranes and their modifications by membrane proteins," in *Lipid-Protein Interactions*, eds P. C. Jost and O.H. Griffith (NewYork, NY: John Wiley Sons, Inc.), 239–281.

Sinnokrot, M. O., Valeev, E. F., and Sherrill, C. D. (2002). Estimates of the ab initio limit for π-π interactions: the benzene dimer. *J. Am. Chem. Soc.* 124, 10887–10893. doi: 10.1021/ja025896h

Smith, D. J., Klauda, J. B., and Sodt, A. J. (2018). Simulation best practices for lipid membranes [article v1.0]. *Living J. Comput. Mol. Sci.* 1:5966. doi: 10.33011/livecoms.1.1.5966

Tang, B., Devenish, S. O., and Lummis, S. C. (2018). Identification of novel functionally important aromatic residue interactions in the extracellular domain of the glycine receptor. *Biochemistry* 57, 4029–4035. doi: 10.1021/acs.biochem.8b00425

Thompson, M. J., Domville, J. A., and Baenziger, J. E. (2020). The functional role of the M4 transmembrane helix in the muscle nicotinic acetylcholine receptor probed through mutagenesis and coevolutionary analyses. *J. Biol. Chem.* 295, 11056–11067. doi: 10.1074/jbc.RA120.013751

Tobimatsu, T., Fujita, Y., Fukuda, K., ichi Tanaka, K., Mori, Y., Konno, T., et al. (1987). Effects of substitution of putative transmembrane segments on nicotinic acetylcholine receptor function. *FEBS Lett.* 222, 56–62. doi: 10.1016/0014-5793(87)80191-6

Zhu, S., Noviello, C. M., Teng, J., Walsh, R. M. J., Kim, J. J., and Hibbs, R. E. (2018). Structure of a human synaptic $GABA_A$ receptor. *Nature* 559, 67–72. doi: 10.1038/s41586-018-0255-3

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: 07 May 2021
doi: 10.3389/fmolb.2021.673053

Drug Repurposing on G Protein-Coupled Receptors Using a Computational Profiling Approach

Alessandra de Felice[1†], Simone Aureli[1†] and Vittorio Limongelli[1,2]*

[1] Faculty of Biomedical Sciences, Euler Institute, Università della Svizzera italiana (USI), Lugano, Switzerland, [2] Department of Pharmacy, University of Naples "Federico II", Naples, Italy

G protein-coupled receptors (GPCRs) are the largest human membrane receptor family regulating a wide range of cell signaling. For this reason, GPCRs are highly desirable drug targets, with approximately 40% of prescribed medicines targeting a member of this receptor family. The structural homology of GPCRs and the broad spectrum of applications of GPCR-acting drugs suggest an investigation of the cross-activity of a drug toward different GPCR receptors with the aim of rationalizing drug side effects, designing more selective and less toxic compounds, and possibly proposing off-label therapeutic applications. Herein, we present an original *in silico* approach named "Computational Profiling for GPCRs" (CPG), which is able to represent, in a one-dimensional (1D) string, the physico-chemical properties of a ligand–GPCR binding interaction and, through a tailored alignment algorithm, repurpose the ligand for a different GPCR. We show three case studies where docking calculations and pharmacological data confirm the drug repurposing findings obtained through CPG on 5-hydroxytryptamine receptor 2B, beta-2 adrenergic receptor, and M2 muscarinic acetylcholine receptor. The CPG code is released as a user-friendly graphical user interface with numerous options that make CPG a powerful tool to assist the drug design of GPCR ligands.

Keywords: GPCR, drug repurposing, molecular docking, drug design, drug repositioning, protein sequence profile alignment

OPEN ACCESS

Edited by:
Edina Rosta,
King's College London,
United Kingdom

Reviewed by:
Antonella Di Pizio,
Technical University of Munich,
Germany
Ilpo Vattulainen,
University of Helsinki, Finland
Irina Tikhonova,
Queen's University Belfast,
United Kingdom

***Correspondence:**
Vittorio Limongelli
vittoriolimongelli@gmail.com

[†]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: *26 February 2021*
Accepted: *13 April 2021*
Published: *07 May 2021*

Citation:
de Felice A, Aureli S and Limongelli V (2021) Drug Repurposing on G Protein-Coupled Receptors Using a Computational Profiling Approach. Front. Mol. Biosci. 8:673053. doi: 10.3389/fmolb.2021.673053

INTRODUCTION

G protein-coupled receptors (GPCRs) are integral membrane proteins involved in the transduction of a wide range of signals from outside the cell to the cellular interior. They represent the largest and most pharmacologically relevant protein family—∼4% of the protein-coding genome (Fredriksson et al., 2003; Zhang et al., 2006). From a structural point of view, in spite of low sequence homology, all GPCRs share a common barrel tertiary structure composed of seven *trans*-membrane α-helices (TM1-7). Furthermore, some GPCRs have an additional α-helix (H8) at the C-terminal (Yeagle and Albert, 2007). The orthosteric binding site of endogenous ligands is typically located in the upper, extracellular part of the receptor, underneath the extracellular loop 2 (ECL2). At the intracellular level, GPCRs interact with the G-protein heterotrimer complex (Gαβγ) through a process allosterically modulated by ligand-induced conformational changes that activate a specific signal cascade based on the type of the interacting Gα-protein (Gs, Gi, Go, Gq/11, G12/13)

(Zhang et al., 2006; Katritch et al., 2013). Through such mechanisms, GPCRs respond to a large variety of stimuli, regulating relevant processes including pain, immune response, inflammation, mood regulation, blood pressure regulation, neurotransmission, and many others (Katritch et al., 2013; Venkatakrishnan et al., 2013; Gacasan et al., 2017). As a consequence, GPCRs are the most prominent molecular targets in drug design, targeted by ∼40% of prescribed drugs (25 of the 100 top-selling) (Thomsen et al., 2005; Rask-Andersen et al., 2011).

In this framework, elucidating the cross-activity of a drug toward diverse GPCRs aids in rationalizing its side effects, proposing off-label therapeutic applications (clinical use for a disease different from that the drug was designed for), and designing novel, more selective GPCRs ligands. With this in mind, we have developed an original *in silico* approach named "Computational Profiling for GPCRs" (CPG), which takes into account both the GPCR sequence and the ligand-GPCR binding interactions to repurpose compounds meant to target one specific GPCR as novel ligands for a different GPCR receptor. Drug repurposing is a fast and safe drug discovery approach that has been successfully employed to identify drugs on the market—therefore considered safe—as new ligands for a molecular target different from the original one (Pushpakom et al., 2019). Our approach is made possible due to the conservative nature of the GPCR tertiary structure and the orthosteric binding site location. In particular, our method (i) "translates" the ligand–protein interaction patterns into a one-dimensional (1D) profile; (ii) aligns the 1D strings coming from different GPCR–ligand complexes; and, finally, (iii) selects the most similar ones to identify drug candidates for drug repurposing. The CPG is designed as a graphical user interface (GUI), integrated into the worldwide-used Visual Molecular Dynamics (VMD) software (Humphrey et al., 1996).

Using CPG, the user is able to process ligand–GPCR complexes obtained from the Protein Data Bank (PDB, Berman et al., 2007) or molecular binding simulations and achieve a fast determination of ligand-GPCR binding similarities. While the workflow of CPG can be applied to any ligand in the identification of potential off-targets, it reveals its potency when employed with market-approved drugs. Indeed, it repurposes a drug for a GPCR different from its original one, thus paving the way to possible off-label therapeutic applications, alternative from that originally intended. At the same time, by identifying a novel GPCR target for the drug, CPG may help to rationalize the unexplained side effects of the drug. Finally, data regarding the similarity between different drug–GPCR complexes generated by CPG are useful to guide the development of novel, more selective GPCR ligands. As proof of concept, three case studies are presented.

MATERIALS AND METHODS

Computational Profiling for GPCRs Alignment Tool

The CPG tool is a user-friendly GUI, written in the Tcl/Tk coding language. To improve its ease of use, the software has been released as a plug-in for VMD. CPG has been designed to extract information from PDB files of ligand–GPCR binary complexes. Details of the CPG tool are reported below, where points (A), (B), and (C) refer to the labels given in **Figure 1**.

(A) By taking the primary protein sequence and identifying the ligand-interacting residues (in a range of 4 Å from the ligand), information regarding the binding site of the protein is obtained. The binding site information generated by the CPG tool includes which residues are part of the binding site and their positions in the GPCR primary sequence (in the form of their resID number). A fundamental feature of CPG is its ability to convert the aforementioned data into a protein profile. That is to say, the residue list of the protein can be mutated into two available profiling systems, i.e., the "8 Digit Profile" (8DP) and the "10 Digit Profile" (10DP). Both are based on the physio-chemical properties of the amino acids, grouping them following different approaches. In the 8DP system, the amino acids are divided into four groups, i.e., "hydrophobic," "hydrophilic," "negatively charged," and "positively charged," to which we assigned an integer number (0, 1, 2, and 3, respectively). Consequently, we obtained a 1D array representing the GPCR primary sequence. At the same time, the integer number is increased by a value of 4 for the residues involved in the ligand-binding site (4 for the hydrophobic group, 5 for the hydrophilic, 6 for negatively charged, and 7 for positively charged). As a result, by exploiting only eight symbols, we can easily distinguish the ligand-interacting amino acids and their physio-chemical properties from the rest of the residues not interacting with the ligand. The 10DP system follows a similar fashion; however, we further split the hydrophobic group into "aliphatic" and "aromatic" subgroups, resulting in values ranging from 0 to 4 and 5 to 9 for general and binding site residues, respectively.

(B) The second important property of CPG is its ability to map the protein profile onto the GPCR topology, dividing it according to which helix of the GPCR each profiled residue belongs to. The data can be generated for a "target" and "reference" protein as chosen by the user, obtaining seven 1D arrays for each macromolecule. Once the desired pair of proteins has been selected, a local pairwise Levenshtein algorithm-based alignment is performed, in order to find the best matches between each corresponding helix. As shown in **Figure 1**, it is possible to choose different alignment scoring methods employing values extracted from the BLOSUM62 (Pietrokovski et al., 1996; Choudhuri, 2014) or the GPCRtm (Rios et al., 2015) substitution matrix, both of which have been adapted for the 8DP and 10DP groups. Furthermore, user-defined custom values may also be employed. Finally, the "MISS" and the "GAP" fields should be filled with non-positive values.

Alignments based on 8DP and 10DP can be visualized in **Figure 1B**, where three different outputs are reported for both the target and the reference proteins. In detail, it displays the alignment score for each helix, the total score based on the sum of the scores of the individual helices, and, lastly, a normalized

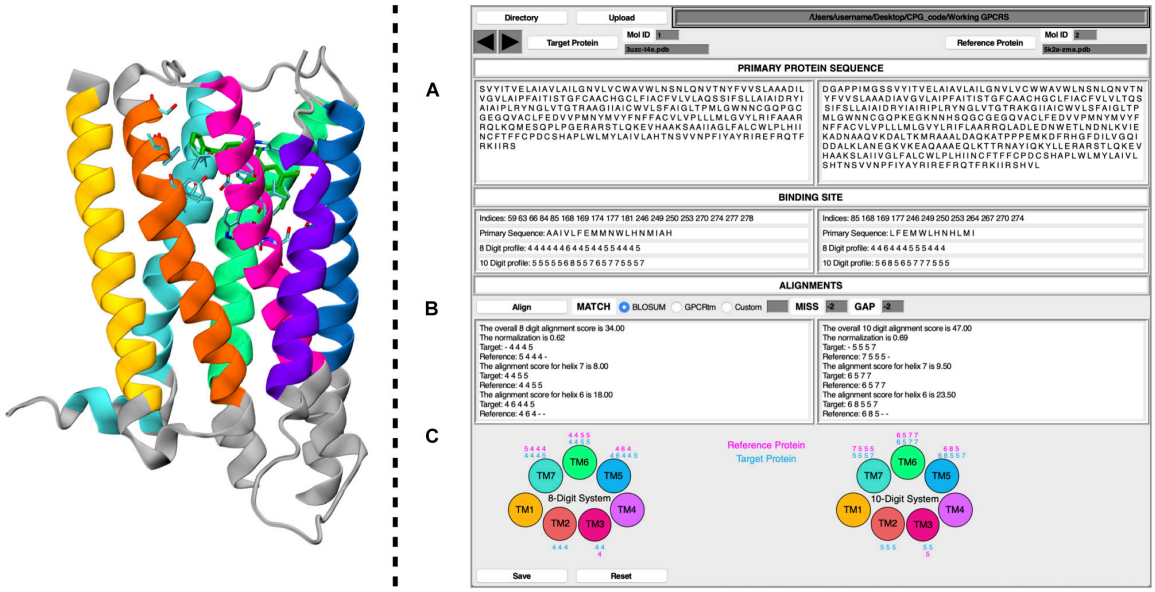


FIGURE 1 | Schematic representation of CPG GUI. On the left, a GPCR is displayed as coded by CPG, where each TM helix is colored according to the following the scheme; TM1, yellow; TM2, orange; TM3, magenta; TM4, purple; TM5, blue; TM6, green; TM7, cyan. Residues unused for the profiling calculation have been colored in gray. The ligand has been colored in dark green, whereas its surrounding amino acids (cut-off 4 Å) are colored according to their atom types. On the right, the main window of the CPG is displayed, where **(A)** includes the primary protein sequence and binding site information, **(B)** the alignment results, and **(C)** the 8 and 10 Digit Profile scores assigned to the binding site residues for each helix of the target and reference proteins.

score which is expressed as:

$$N = T/R$$

where T is the target protein total score (the alignment score with the target protein against the reference protein), and R is the reference protein total score (a self-alignment score value of the reference protein). The normalization of the alignment score is important inasmuch as it considers the length of the aligned strings, thus taking into account different volumes in the binding pocket occupied by the ligands. The value of the normalized alignment score of the pairwise alignment between the target and the reference GPCR indicates a likelihood of repositioning one or both ligands in the reciprocal receptors.
(C) As described in point (B), the profiled binding site residues are divided according to each helix. The CPG tool provides a graphical visualization of the 8DP and 10DP scores attributed to each of the binding site residues of the 7 α-helices of the target and reference GPCRs.

A more detailed explanation of the CGP methodology including the alignment procedure and the scoring functions is provided in **Supplementary Information**, where we also report a tutorial for the use of CPG.

Docking Calculation

We investigated the binding of ligands to repurposed GPCRs by means of molecular docking calculations. This computational technique is widely used to elucidate the ligand-binding mode in various molecular targets, including GPCRs (Anzini et al., 2008, 2011; Nuti et al., 2010; Limongelli, 2020). In particular, we performed cross-docking calculations by docking two ligands in their reciprocal receptor. These calculations were performed on selected pairs of ligand–GPCR complexes that have a CPG score higher than 0.5 and involve seemly pharmacologically unrelated GPCRs.

Molecular docking calculations were carried out using the AutoDock4.2.6 software package (AD4, Morris et al., 2009; Forli et al., 2016). Protonation states of protein residues and ligands were set at pH 7.0. Ligand and receptor structures were prepared and converted to AutoDock format files using AutoDockTools, and the Gesteiger-Marsili partial charges were then assigned. Grid points of 40 × 40 × 40 with a 0.375 Å spacing were calculated around the binding cavity using AD4. Thus, 100 separate docking calculations were performed for each run. Each docking run consisted of 2.5 million energy evaluations using the Lamarckian genetic algorithm local search (GALS) method. Otherwise, default docking parameters were applied. Docking conformations were clustered on the basis of their RMSD (tolerance = 1.5 Å). The analysis on the best binding poses was performed employing the "Drug Discovery Tool" (DDT, Aureli et al., 2019), a GUI recently developed in our group that enables a fast, yet accurate analysis of the docking calculation.

RESULTS

The CPG algorithm is based on protein profiling, a powerful bioinformatics technique that applies a dimensionality reduction process in which multiple properties of amino acid sequences are described by a mono-dimensional information string. By

exploiting such a representation, it is possible to perform fast alignments between diverse proteins based on the chemical similarities of their amino acids. In such a way, it is possible to employ a scoring method based on the conservation of protein residues. For the present study, we set up two scoring functions, namely, 8DP and 10DP, that exploit two well-known scoring matrices: (i) BLOSUM62 (Pietrokovski et al., 1996), which is a generalized scoring method for all proteins; and (ii) GPCRtm (Rios et al., 2015), which has been specifically developed for Class A GPCRs. In detail, we used CPG to generate alignment score tables based on the pairwise alignments of the 55 GPCR pdb files available in the PDB databank. Each score was computed by employing a specific scoring function, reporting a final normalized value. In particular, the 8DP algorithm converts each amino acid into an integer number, following the scheme hydrophobic = 0, hydrophilic = 1, negatively charged = 2, and positively charged = 3. CPG then discriminates the residues interacting with the ligand by increasing their numerical value by 4. 10DP follows a similar rationale, further dividing the hydrophobic group into two subgroups, "aliphatic" and "aromatic." In 10DP, the profiling scheme is aliphatic = 0, aromatic = 1, hydrophilic = 2, negatively charged = 3, and positively charged = 4, while the score of the ligand-interacting residues is increased by 5. Exploiting two different profiling systems allows us to take into account the impact of π–π interactions, which is explicitly accounted for in the 10DP scheme (see "Materials and methods" section and **Supplementary Information** for details). A step-by-step tutorial to guide the reader in the use of CPG is provided in the **Supplementary Information**.

The scoring matrices reported in **Supplementary Tables 1**, **2** were employed to determine the likelihood of drug repositioning considering the alignment between two different ligand–GPCR complexes, where a threshold value of 0.5 for the normalized alignment score was considered. In particular, we found ∼600 complexes that fulfill this condition, most of them obtained from different pdb complexes of the same GPCR, as expected. However, ∼10% of the top-ranked hits regarded complexes of different GPCRs. Among these, three pairs of drug–GPCR complexes, for a total of six systems, were further investigated with the aim of assessing the CPG prediction. In detail, we evaluated the GPCR cross-activity of the drug by means of cross-docking calculations in the newly identified GPCR target and by analyzing the available data on its pharmacological activity. The investigated pairs of complexes are (i) the 5-hydroxytryptamine receptor 2B with the ligand alprenolol and the beta-2 adrenergic receptor with the ligand lisuride; (ii) the 5-hydroxytryptamine receptor 2B with the ligand timolol and the beta-2 adrenergic receptor with the ligand lysergic acid diethylamide (LSD); and (iii) the M2 muscarinic acetylcholine receptor with the ligand ICI-118,551 and the beta-2 adrenergic receptor with the ligand quinuclidinyl benzilate (QNB). The results are discussed in detail in the following paragraphs.

Lisuride–5HT2B and Alprenolol–ADRB2

The first case study regards the 5-hydroxytryptamine receptor 2B (also known as the serotonin receptor 2B, hereafter 5HT2B, Hensler, 2012) bound to lisuride (**Figure 2A**), one of its marketed antagonists, and the beta-2 adrenergic receptor (hereafter ADRB2, Rascol et al., 2007) in complex with its antagonist alprenolol (**Figure 2B**).

Binding Mode in the Native GPCR

In the x-ray structure of the complex 5HT2B–lisuride (PDB ID 6DRX, McCorvy et al., 2018), the ligand forms a salt bridge and a H-bond with Asp135, while its indole ring is placed in a pocket shaped by several hydrophobic/aromatic residues (**Figure 2C**). Here, the ligand engages π–π stacking interactions with Phe217, Phe340, and Phe341, and van der Waals interactions with Val136 and Val366. Finally, the indole ring of the lisuride can also form a π-mediated H-bond with Asn344. In the second complex formed by ADRB2 bound to alprenolol (PDB ID 3NYA, Wacker et al., 2010), protonated amine function of the ligand H-bonds with Asn312, also engaging a salt bridge interaction with Asp113 (**Figure 2D**). In addition, the hydroxy group of the ligand forms H-bonds with Asp113 and Asn312. On the contrary, the aromatic ring of the alprenolol interacts with Val114, Tyr199, Phe289, and Phe290 through π–π stacking and van der Waals interactions. Finally, the ortho-allyl group of the alprenolol is placed in a competent position to form a π-mediated electron transfer interaction with Asn293.

Employing CPG, we found that the lisuride–5HT2B and alprenolol–ADRB2 complexes show a 10DP alignment score of 0.54, higher than the threshold value 0.5 (**Table 1**), suggesting that lisuride and alprenolol could be repurposed as novel ligands for their reciprocal GPCRs.

Binding Mode in the Repurposed GPCR

In order to assess the CPG prediction and validate this hypothesis, we performed cross-docking calculations in which lisuride was studied in the ADRB2 structure, while alprenolol was studied in the 5HT2B structure (see "Materials and methods" section for docking details). The docking results confirmed the ability of these two ligands to cross-bind their reciprocal GPCR, showing binding modes stabilized by a series of favorable interactions (see **Figure 2E** and **Table 2**). In particular, in the most populated binding pose of alprenolol in 5HT2B, the ligand forms a salt bridge interaction with Asp135, which resembles that established with Asp113 in ADRB2. An additional H-bond is formed between the hydroxyl group of the alprenolol and Asp113, while the isopropyl moiety of the ligand engages hydrophobic contacts with Val366 and Trp131. On the contrary, the aromatic ring of the alprenolol is located in a hydrophobic pocket remarkably similar to that present in ADRB2 (**Figures 2D,E**). Here, the ligand forms π–π stacking interactions with Phe217, Phe340, and Phe341, and van der Waals contact with Val136. In addition, the allyl π-electrons of the alprenolol are involved in electron transfer interaction with Asn344 as similarly engaged with Asn293 in ADRB2. In the case of the binding of lisuride in ADRB2, considering the bulkiness of the ligand we performed a flexible docking calculation to allow conformational flexibility of the Asp113 side chain.

In the most populated binding pose, the ligand forms strong interactions with the receptor like the salt bridge and the H-bond

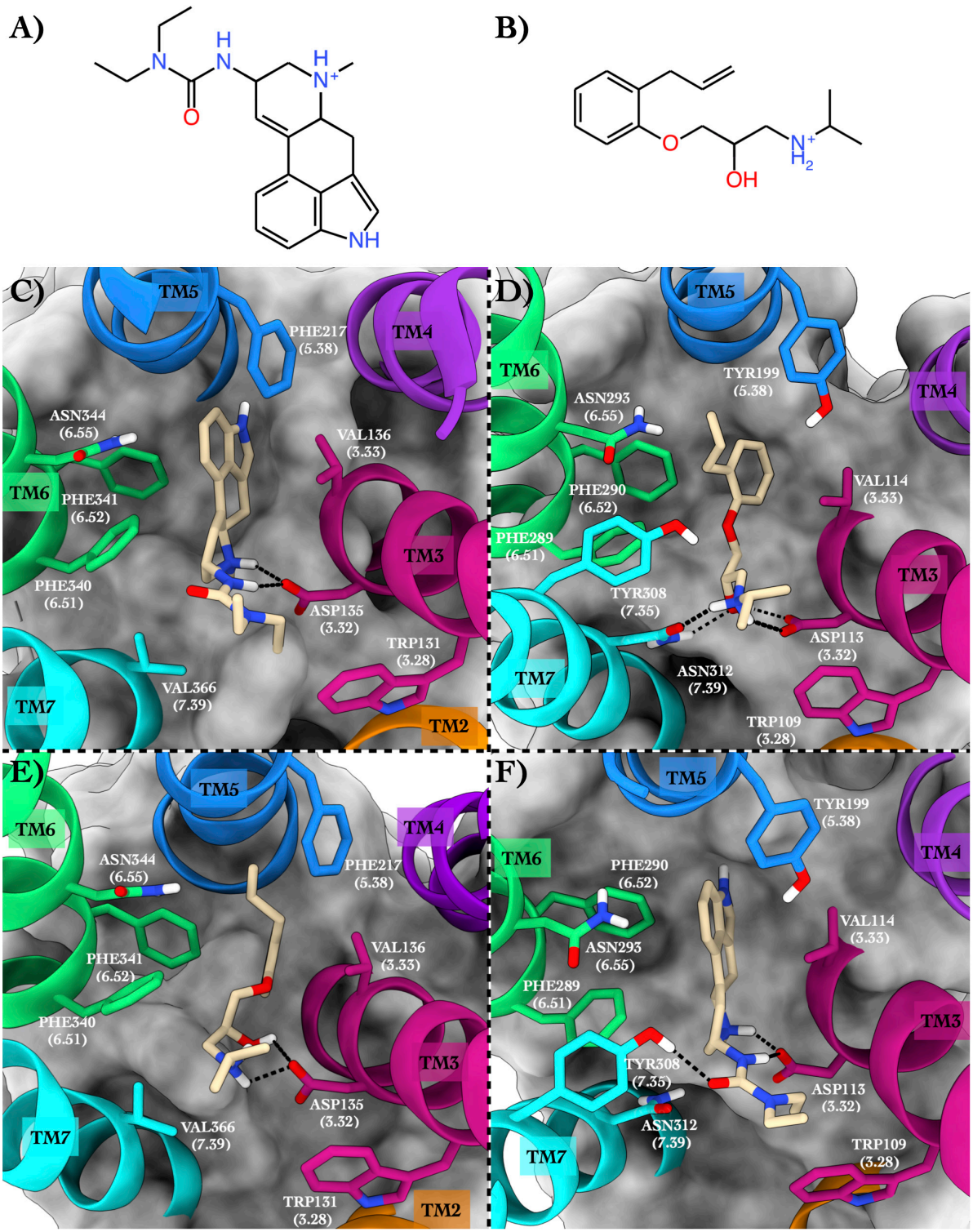


FIGURE 2 | Schematic depictions of lisuride and alprenolol and their respective binding sites inside 5HT2B and ADRB2. **(A)** Chemical structure of lisuride at physiological pH. **(B)** Chemical structure of alprenolol at physiological pH. **(C)** Lisuride–5HT2B crystallographic binding mode. **(D)** Alprenolol–ADRB2 crystallographic binding mode. **(E)** Centroid of the most populated cluster family coming from the docking calculation between 5HT2B 3D structure and alprenolol. **(F)** Centroid of the most populated cluster family coming from the flexible docking calculation between ADRB2 and lisuride. Lisuride and alprenolol have been colored in tan. The surrounding residues are labeled using both primary sequence and Ballesteros–Weinstein numbering. The helices and the marked residues have been depicted according to the CGP color scheme, with TM2 in orange, TM3 in magenta, TM4 in purple, TM5 in blue, TM6 in green, and the TM7 in cyan.

TABLE 1 | The alignment scores computed for the PDB sequence 6DRX and 3NYA using a miss and gap score of −2.

Target protein	Reference protein	8DP BLOSUM min:−0.84 max: 0.73	8DP GPCRtm min:−0.83 max: 0.75	10DP BLOSUM min:−0.77 max: 0.66	10P GPCRtm min: −0.97 max: 0.66
5HT2B/H8G (PDB ID: 6DRX)	ADRB2/JTZ (PDB ID: 3NYA)	0.39	0.4	0.54	0.46
ADRB2/JTZ (PDB ID: 3NYA)	5HT2B/H8G (PDB ID: 6DRX)	0.36	0.37	0.52	0.43

The highest and lowest values for each profiling scoring function obtained by aligning all the diverse GPCRs available in the PDB databank are also reported.

TABLE 2 | The cross-docking calculation scores of the 5HT2B receptor with alprenolol and the ADRB2 receptor with lisuride.

Protein	Ligand	Mean binding energy (docking score)	Runs in cluster	Number of clusters
5HT2B	JTZ	−6.58	62/100	3
ADRB2	H8G	−10.78	100/100	1

with Asp113, reproducing the same interactions established with Asp135 in 5HT2B. A further H-bond formed by the ligand's urea oxygen with Tyr308 stabilizes the binding mode. In addition, while the two ethyl groups form hydrophobic contacts with Trp109, the aromatic moiety engages π-π stacking and Van der Waals interactions with Tyr199, Phe289, Phe290, and Val114. Finally, the indole ring of the lisuride forms a π-mediated H-bond with Asn293 as similarly done with Asn344 in ADRB2. A detailed list of the interactions established by lisuride and alprenolol with 5HT2B and ADRB2 is reported in **Supplementary Table 4**.

Lisuride and Alprenolol Pharmacology

In order to assess the repurposing of lisuride and alprenolol as ligands of ADRB2 and 5HT2B, respectively, we thoroughly studied their pharmacological profiles. Lisuride is an ergot derivative, administered for the treatment of Parkinson's disease, depression, and migraines (Gopinathan et al., 1981; Egan et al., 1998; Hofmann et al., 2006). The mechanism of action of lisuride is due to its agonist activity on several serotonin receptor subtypes (5HT1A, 5HT1B, 5HT1D, 5HT2A, 5HT2B, and 5HT2C) (Egan et al., 1998), as well as on the dopamine receptors D1, D2, D3, D4, and D5 (Hildebrand et al., 1987). It should be underlined that lisuride has already undergone a drug repositioning process where it has been repurposed for the suppression of lactation as it lowers serum prolactin levels (Van Dam and Rolland, 1981).

Alprenolol is a beta-adrenergic antagonist with anti-arrhythmic effects, being able to bind ADRB1, ADRB2, and ADRB3 (Himori et al., 1977). The activity of alprenolol is given by the inhibition of the activity of the beta-adrenergic receptor's natural ligands epinephrine and norepinephrine. As a consequence, alprenolol induces a reduction in heart rate (Wasserman et al., 1970). Alprenolol also has an anti-hypertensive effect by inhibiting the production of renin, thus acting on the renin–angiotensin–aldosterone system (RAAS) by lowering angiotensin II and aldosterone production, which leads to the reduction of vasoconstriction and water retention (Himori et al., 1977). While it has been reported that alprenolol can also bind to the 5HT1A receptor, so far there is no evidence that it is also able to bind the 5HT2B receptor. In particular, the pharmacological activity of alprenolol on the 5-hydroxytryptamine (5-HT)-induced hyperactivity response has been studied as early as 1978 (Costain and Green, 1978); however, the spectrum of its molecular targets is still unexplored. The activity of alprenolol toward 5HT2B might explain the relevant side effects of this drug, such as the gastrointestinal ones (Amjad et al., 2017). This might be due to the presence of adrenergic receptors in the gastrointestinal tract, as well as 5HT2B, which is a ubiquitous GPCR also expressed in the liver and the intestine (Papadimas et al., 2012). On the contrary, 5HT1A is poorly expressed in the gastrointestinal tract, being mostly located on the lymph nodes, the thymus, and the spleen. Elucidating the different GPCRs targeted by alprenolol might lead to a better understanding of the adsorption of the body and the toxicity of this drug. To this end, the results of our study highlight a potential activity of alprenolol on 5HT2B and lisuride on ADRB2, suggesting to further investigate the molecular interaction of these drugs with the two receptors with the scope to rationalize toxicity and propose novel, repurposed clinical applications for these two drugs.

Lysergic Acid Diethylamide–5HT2B and Timolol–ADRB2

The second case study regards 5HT2B and ADRB2 in complex with lysergic acid diethylamide (hereafter LSD) (**Figure 3A**) and timolol (**Figure 3B**), respectively.

Binding Mode in the Native GPCR

The 5HT2B x-ray structure (PDB ID 5TVN, Wacker et al., 2017) in complex with LSD (**Figure 3C**) shows a salt bridge interaction between the charged amine of the LSD and Asp135 and π–π stacking–hydrophobic interactions between the aromatic moiety of the ligand and Phe217, Phe340, Phe341, and Val136. The diethylamide function of the LSD forms additional van der Waals contacts with Trp131 and Val366 that further stabilize the binding pose. The second system is ADRB2 in complex with timolol (PDB ID 3D4S, Hanson et al., 2008; **Figure 3D**). Here, a network of H-bonds stabilizes the binding mode. In detail, the sulphur atom of the timolol's thiadiazole H-bonds with Thr118, while the oxygen of timolol's morpholine ring engages a H-bond with Asn293. On the contrary, the protonated amine group of the LSD forms a salt bridge interaction with Asp113 and a H-bond with Asn312. The same residues also establish H-bonds with the hydroxyl group of the ligand. Finally, π–π stacking and

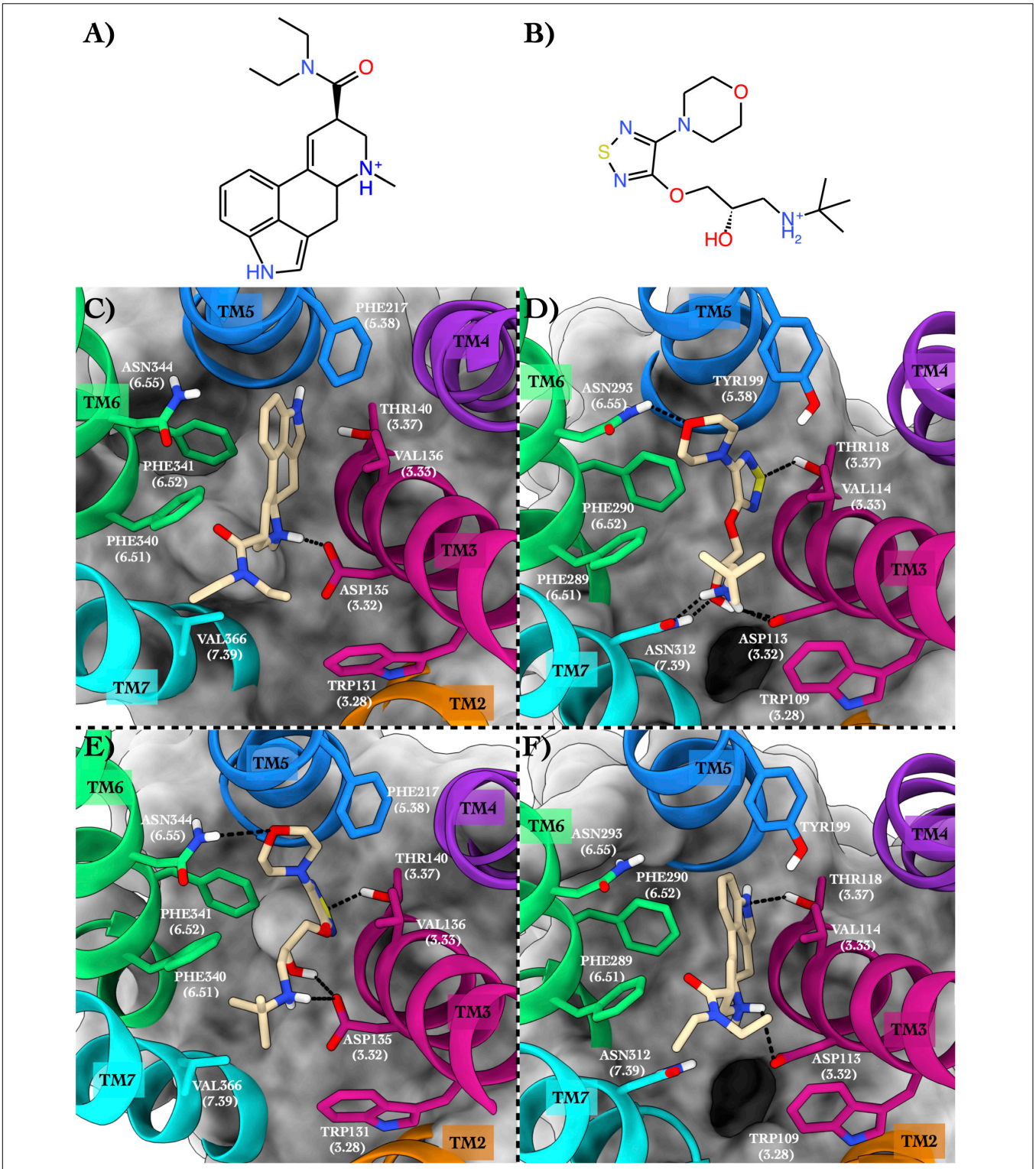


FIGURE 3 | Schematic depictions of LSD and timolol and their binding modes inside the 5HT2B and ADRB2 receptors. **(A)** Chemical structure of LSD at physiological pH. **(B)** Chemical structure of timolol at physiological pH. **(C)** LSD–5HT2B crystallographic binding mode. **(D)** Timolol–ADRB2 crystallographic binding mode. **(E)** Centroid of the most populated cluster family coming from the docking calculation between 5HT2B and timolol. **(F)** Centroid of the most populated cluster family coming from the docking calculation of the ADRB2 3D structure and LSD. LSD and timolol have been colored in tan. The surrounding residues are labeled using both primary sequence and Ballesteros–Weinstein numbering. The helices and the marked residues have been depicted according to the CGP color scheme, with TM2 in orange, TM3 in magenta, TM4 in purple, TM5 in blue, TM6 in green, and TM7 in cyan.

TABLE 3 | The alignment scores computed for the PDB sequence 5TVN and 3D4S using a miss and gap score of −2.

Target protein	Reference protein	8DP BLOSUM min:−0.84 max: 0.73	8DP GPCRtm min:−0.83 max: 0.75	10DP BLOSUM min:−0.77 max: 0.66	10P GPCRtm min: −0.97 max: 0.66
5HT2B/7LD (PDB ID: 5TVN)	ADRB2/TIM (PDB ID: 3D4S)	0.63	0.64	0.47	0.43
ADRB2/TIM (PDB ID: 3D4S)	5HT2B/7LD (PDB ID: 5TVN)	0.61	0.61	0.52	0.43

The highest and lowest values for each profiling scoring function obtained by aligning all the diverse GPCRs available in the PDB databank are also reported.

hydrophobic interactions are made by the thiadiazole moiety of the timolol and the terminal tert-butyl group with Phe290, Phe289, and Trp109.

The above two systems have high CPG alignment scores, especially in the case of the 8DP scoring function (**Table 3**). This scoring function weighs the hydrophilic interactions between the ligand and the GPCR more than the 10DP one, thus assigning a higher score to binding modes characterized by polar contacts—H-bonds and salt bridges—like those present in these two complexes. In order to assess the CPG prediction of cross-affinity of LSD and timolol in their reciprocal GPCR, cross-docking calculations of timolol in 5HT2B and LSD in ADRB2 were performed, and the results are discussed as follows (**Table 4**).

Binding Mode in the Repurposed GPCR

In the most recurring docking pose of timolol in 5HT2B, the charged amine of the ligand forms a salt bridge with Asp135 as similarly done with Asp113 in ADRB2 (**Figure 3E**). Three additional H-bonds further stabilize the timolol binding mode such as those formed by its hydroxyl group with Asp135, its morpholine ring with Asn344, and its thiadiazole sulfur atom with Thr140. Finally, π−π stacking and hydrophobic interactions are formed by the thiadiazole moiety with Phe340 and by the terminal tert-butyl group of the ligand with Phe341 and Val366, respectively.

In **Figure 3F**, we show the cross-docking result of LSD in ADRB2. Here, the ligand occupies the binding pocket establishing π−π stacking and van der Waals interactions with the surrounding residues Phe290 and Val114. The anchor point of the ligand binding is the typical salt bridge made by the charged amine of the LSD with Asp113, whereas the amine of the indole ring of the ligand can form a H-bond with Thr118. Finally, hydrophobic contacts are engaged by the diethylamide group of the ligand with Trp109 and Phe289. It is worth noting that most of these interactions are also present in the timolol-binding mode, showing remarkable strength and similarity in the interaction with ADRB2 for these two drugs. This finding fully agrees with the high binding affinity of LSD to ADRB2 resulted from the docking calculations and reported in **Table 4**. As before, we report the full list of the interactions formed by LSD and timolol with 5HT2B and ADRB2 in **Supplementary Table 5**.

Timolol Pharmacology

When evaluating the possibility of repositioning timolol, it should be noted that timolol is a drug used as eye drops that targets the beta-1 and beta-2 adrenergic receptors which results in a decrease in eye pressure (e.g., caused by glaucoma, Sambhara and Aref, 2014). Furthermore, timolol has also been used for the treatment of hypertension. From a pharmacological point of view, timolol is an antagonist for the beta-adrenergic receptor. One of the most common side effects of timolol is the onset of depression; however, the understanding of such a side effect is yet unknown (Nolan, 1982). Prompted by the CPG results and supported by the cross-docking calculations, we propose timolol as the ligand of the serotonin receptor 5-HT2B similar to LSD. The activity of timolol on 5HT2B, working as off-target, might explain the neurological disorders caused by the use of this drug. This represents an example of how to use CPG in the investigation of drug side effects by evaluating drug off-target activity through its repositioning toward a novel GPCR. This step is valuable, especially in the early stages of drug development, to assess whether the newly designed drug can bind off-targets that might cause undesirable side effects.

TABLE 4 | The cross-docking calculation scores of the 5HT2B receptor with timolol and the ADRB2 receptor with LSD.

Protein	Ligand	Mean binding energy (docking score)	Runs in cluster	Number of clusters
5HT2B	TIM	−7.96	58/100	5
ADRB2	7LD	−10.55	100/100	1

Quinuclidinyl Benzilate–ACM2 and ICI-118,551–ADRB2

As the third case study, we investigated the M2 muscarinic acetylcholine receptor (hereafter ACM2) bound to the antagonist quinuclidinyl benzilate (QNB, **Figure 4A**, Shirakawa et al., 1987) and the ADRB2 receptor in complex with the antagonist ICI-118,551 (JRZ, **Figure 4B**; see **Table 5**).

Binding Mode in the Native GPCR

In the x-ray structure of the QNB–ACM2 complex (PDB ID 3UON, Haga et al., 2012; **Figure 4C**), the ligand engages a salt bridge through the charged amine group with Asp103, whereas its carbonyl and hydroxyl groups form two H-bonds with Asn404. The azabicyclooctan moiety of the ligand is placed in a hydrophobic pocket surrounded by aromatic residues like Trp400, Tyr403, and Tyr426, while one of its two aromatic rings forms π−π stacking interactions with

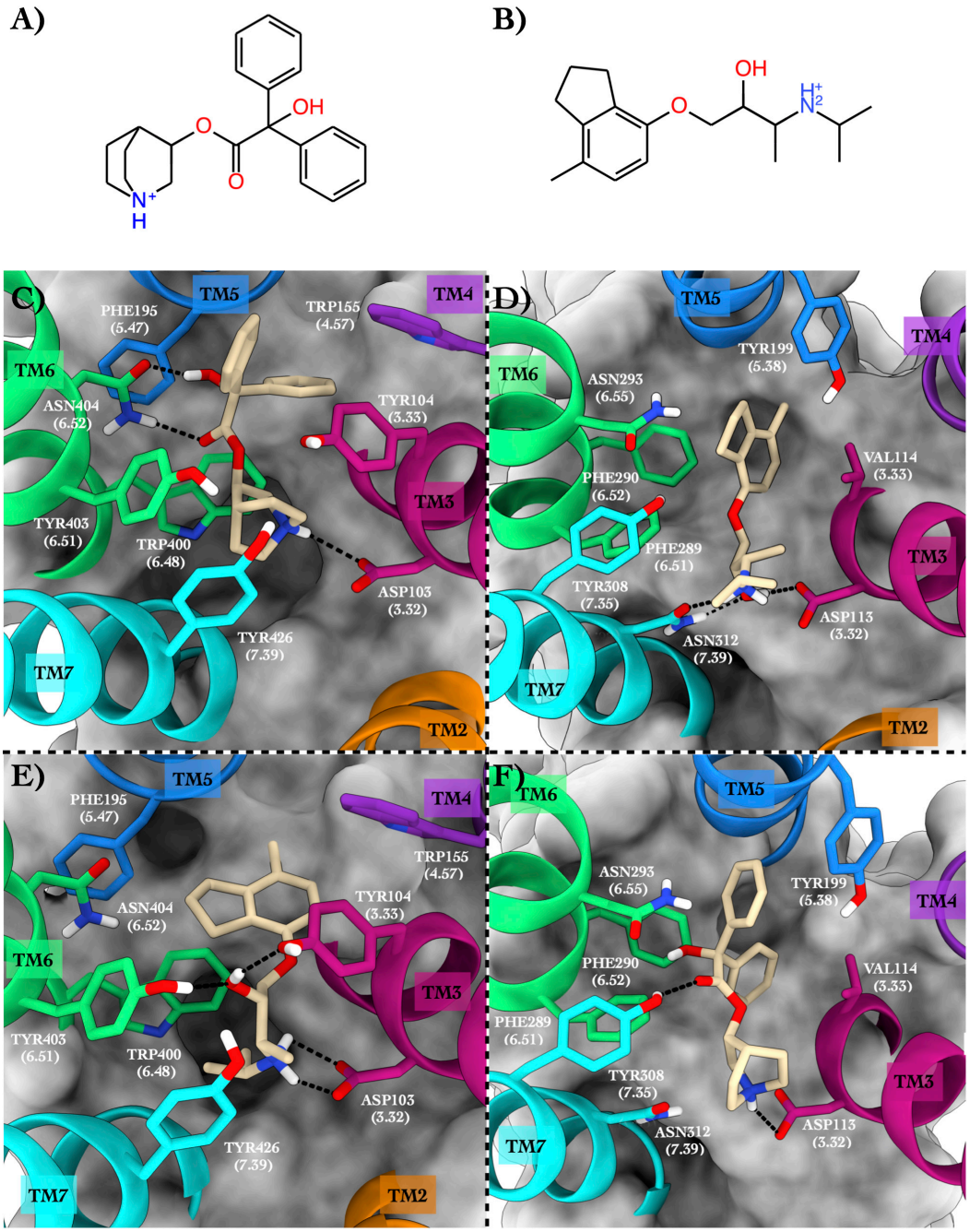


FIGURE 4 | Schematic depictions of the chemical structures of quinuclidinyl benzilate and ICI-118,551 and their binding modes inside the ACM2 and ADRB2 receptors, respectively. **(A)** Chemical structure of quinuclidinyl benzilate at physiological pH. **(B)** Chemical structure of ICI-118,551 at physiological pH. **(C)** Quinuclidinyl benzilate–ACM2 crystallographic binding mode. **(D)** ICI-118,551–ADRB2 crystallographic binding mode. **(E)** Centroid of the most populated cluster family coming from the docking calculation of ICI-118,551 in the ACM2 receptor. **(F)** Centroid of the most populated cluster family coming from the docking calculation between ADRB2 and quinuclidinyl benzilate. Quinuclidinyl benzilate and ICI-118,551 have been colored in tan. The surrounding residues are labeled using both primary sequence and Ballesteros–Weinstein numbering. The helices and the marked residues have been depicted according to the CGP color scheme, with TM2 in orange, TM3 in magenta, TM4 in purple, TM5 in blue, TM6 in green, and TM7 in cyan.

Tyr104 and Trp155. In the crystallographic ADRB2 in complex with JRZ (PDB ID 3NY8, Wacker et al., 2010; **Figure 4D**), the charged amine and hydroxyl groups of the ligand form three H-bonds with Asn312 and Asp113, while the indanyl moiety engages hydrophobic interactions with Val114, Tyr199, and Phe290. The high CPG alignment score for the above two drug–GPCR complexes prompted us to further assess through docking calculations the capability of JRZ and QNB

TABLE 5 | The alignment scores computed for the PDB sequence 3UON against 3NY8 using a miss and gap score of −2.

Target protein	Reference protein	8DP BLOSUM min:−0.84 max: 0.73	8DP GPCRtm min:−0.83 max: 0.75	10DP BLOSUM min:−0.77 max: 0.66	10P GPCRtm min: −0.97 max: 0.66
ACM2/QNB (PDB ID: 3UON)	ADRB2/JRZ (PDB ID: 3NY8)	0.61	0.62	0.59	0.54
ADRB2/JRZ (PDB ID: 3NY8)	ACM2/QNB (PDB ID: 3UON)	0.47	0.49	0.43	0.42

The highest and lowest values for each profiling scoring function obtained by aligning all the diverse GPCRs available in the PDB databank are also reported.

to interact with their reciprocal receptors ACM2 and ADRB2, respectively (**Table 5**).

Binding Mode in the Repurposed GPCR

The ligand JRZ shows a strong affinity for ACM2 with a remarkable docking score of −9.43 for the most populated binding mode (**Table 6**). In this pose (**Figure 4E**), the charged amine group of JRZ forms a salt bridge with Asp103, mimicking that made by QNB (**Figure 4C**). The hydroxyl group of the ligand engages two H-bonds with Tyr104 and Tyr403, while the indanyl moiety establishes π−π interactions with aromatic residues like Trp155, Phe195, and Trp400.

Regarding QNB in ADRB2 (**Figure 4F**), docking calculations show a strong interaction between the ligand and this GPCR with a low docking score (**Table 6**). The best and most populated docking pose shows the ligand interacting with the typical salt bridge interaction with Asp113, as seen in the case of the JRZ–ADRB2 binary complex (**Figure 4D**). In addition, while the ligand interaction with Asn312 is lost if compared with JRZ, a new H-bond is formed between the hydroxyl group of the QNB and Tyr308. Interestingly, the two aromatic rings of QNB contribute to further stabilize the binding mode through hydrophobic and π−π stacking interactions with Val114, Tyr199, Phe289, Phe290, and *via* a π-mediated H-bond with Asn293. As done for the previously discussed systems, the full list of the interactions established by QNB and JRZ with ACM2 and ADRB2 is reported in **Supplementary Table 6**.

ICI-118,551 Pharmacology

ICI-118,551 is an ADRB2 antagonist widely used in research for its 100-fold higher selective inhibition of ADRB2 with respect to ADRB1 and ADRB3. A recent work (Kashihara et al., 2014) has reported that administrating to mice an ADRB agonist, isoproterenol, together with ICI-118,551 gives similar pharmacological effects compared to mice administrated with a combination of isoproterenol and atropine, a well-known ACM2 antagonist. The authors explained this finding based on the common intracellular pathways shared by adrenergic and cholinergic signaling. However, our results pave the way to a new hypothesis that the similar pharmacological outcome of ICI-118,551 and atropine might be due to their common affinity toward the ACM2 receptor, an intriguing perspective that is worthy of further investigations.

TABLE 6 | The cross-docking calculation scores of the ACM2 receptor with JRZ and the ADRB2 receptor with QNB.

Protein	Ligand	Mean binding energy (docking score)	Runs in cluster	Number of clusters
ACM2	JRZ	−9.43	50/100	3
ADRB2	QNB	−9.91	72/100	3

DISCUSSION

The pharmacological relevance of GPCRs is highlighted by the fact that almost 40% of prescribed drugs target this receptor family (Rask-Andersen et al., 2011). The structural conservation in these membrane proteins allows for the relatively systematic profiling of their binding sites because the helices of GPCRs form a "barrel" structure composed of seven helices connecting the extracellular and intracellular spaces. The GPCRs do this by binding to a variety of ligands (small molecules, peptides, and even other proteins), which can be either exogenous or endogenous. A profiling methodology called Computational Profiling GPCRs (CPG) has been proposed here, which combines the primary structure of a GPCR with three-dimensional (3D) information when the receptor is complexed with a ligand, thus making the extraction of valuable data relating to the ligand–GPCR binding affinity possible. In particular, by converting the protein sequence into a 1D string of values representing the chemico-physical properties of the amino acids and the ligand–receptor binding interactions, a pairwise alignment of the GPCR-binding sites can be done in a simplified manner. A proper alignment driven by scoring methods based on the conservation of protein residues enables the detection of possible drug repositioning with important consequences in our understanding of drug pharmacology and side effects.

The profiling and aligning of ligand–GPCRs complexes were carried out using CPG on the available crystal structures. Our results show that there are promiscuous ligands that might be able to bind different GPCRs. As proof of concept, we have reported and discussed in detail three case studies that are: (i) lisuride–5HT2B and alprenolol–ADRB2; (ii) LSD–5HT2B and timolol–ADRB2; and (iii) quinuclidinyl benzilate–ACM2 and ICI-118,551–ADRB2. The CPG algorithm reported these systems among the top-scored ones, thus suggesting the repurposing of these drugs for their reciprocal receptor. We validated the CPG results by molecular docking calculations and provided a pharmacological basis with the data available in the literature. We showed that CPG can be useful to propose

novel, repurposed clinical applications for the investigated drugs and for the rationalization of drug side effects by evaluating their off-target activity through repositioning toward a novel GPCR. The latter is a valuable process, especially in the early stages of drug development, to assess whether the newly designed drug can bind off-targets that might cause undesirable side effects. Certainly, further investigations, for instance, using binding free-energy calculations (Limongelli et al., 2013; Comitani et al., 2016; Moraca et al., 2017; Brotzakis et al., 2018; Yuan et al., 2018; D'Annessa et al., 2019; Raniolo and Limongelli, 2020) and *in vitro* experiments are necessary to properly assess the binding affinity and the pharmacological activities of the investigated ligands. Particularly, in GPCRs where ligand binding involves parts of the receptor endowed with conformational flexibility like the extracellular loops, molecular-binding simulations should be performed using methodologies as molecular dynamics that are more efficacious than docking in taking into account receptor flexibility and ligand-induced fit effects, thus providing a reliable validation of the CPG predictions. We point out that when preparing the ligand–GPCR complex for the validation simulations, receptor and ligand properties like the protonation state of specific residues or ligand functional groups, might be not immediately apparent from the sequence and structural data and need to be carefully considered by the investigator.

We note that CPG results rely on the type and quality of input data including the class of the GPCR, the chemical structure of the ligand, the similarity of the ligand-binding site, and the resolution of the ligand–GPCR complex structure. In this regard, one might observe that aminergic class A GPCRs are typically reported among the top-scored systems. This finding is not surprising considering that they are the most representative GPCR subgroup in the PDB databank, which is used as data source of ligand–GPCR complexes in CPG. Furthermore, one of the substitution matrices used in our study, GPCRtm, was developed based on sequences of class A GPCRs, thus performing better in scoring alignments of this GPCR subgroup. However, CPG is designed to work with any GPCR, and we expect that it will provide useful results even for GPCRs of the other classes as more receptor structures will be resolved, and alignment scoring functions optimized for the other classes of GPCRs will be available.

In addition, there is still room for improvement of the methodology. Namely, due to the employment of aligning procedures based only on generalized physio-chemical properties, the spatial information on the ligand–receptor interaction is lost as well as the binding cavity accessible volume. This means that in some instances, the alignment score may seem promising; however, the residues at the binding site might not be in a proper position to allow ligand binding. In such a case, a practical solution is to compute how the alignment score changes as a function of the gap penalty applied. Based on our experience, the less the score changes using different gap penalty values, the more the size of the ligands under examination are similar. Therefore, by looking at the global alignment of all the helices, not merely at each individual helix separately, a greater understanding of binding site similarities is achieved. This procedure can improve the accuracy and the specificity of the methodology (fewer false positives).

CONCLUSION

In conclusion, CPG has proved to be an appealing tool to rapidly investigate drug repurposing for GPCRs. Our tool performs particularly well with aminergic class A GPCRs since they are the most representative GPCR structures in the PDB databank, and they were also employed to develop one of the alignment scoring functions used in our study. We report the full list of GPCRs repurposed ligand candidates identified in our study in **Supplementary Table 3**. This represents a useful data source for investigations on the pharmacological activities of these compounds. A future extension of our methodology, including profiling of binding sites for apo GPCRs, is desirable as it would pave the way for applications of CPG not only in GPCR drug repurposing but also in *de novo* drug discovery pipelines. More than 800 GPCRs have been identified by sequence analysis on the human genome; however, only a comparatively low number of them have been targeted (Sriram and Insel, 2018). Due to their pharmacological relevance, there is clearly the urgency of finding methods that are able to speed up the identification of lead compounds, which then can finally undergo a lead optimization process. In addition, having a reliable dimensionality-reduced description of the drug–GPCR molecular interaction, especially in 1D string, represents a precious tool in the employment of machine learning approaches in drug development as expected in the near future. Our CPG is a promising methodology that points exactly in this direction.

DATA AVAILABILITY STATEMENT

The code of CPG is available at this link https://sites.google.com/site/vittoriolimongelli/downloads, together with a directory containing the GPCRs employed in the present manuscript. An easy-to-use tutorial has been reported in **Supplementary Material**.

AUTHOR CONTRIBUTIONS

VL designed the research. SA and VL conceived the algorithm. AD and SA wrote the code and performed the docking calculations. AD, SA, and VL analyzed the results and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

VL acknowledges the support from the European Research Council (ERC Consolidator Grant "CoMMBi"), the Swiss National Science Foundation (Project No. 200021_163281), the Italian MIUR-PRIN 2017 (2017FJZZRC), and the Swiss National Supercomputing Centre (CSCS) (project ID s1013).

ACKNOWLEDGMENTS

We especially thank Daniele Di Marino for the inspiring and fruitful discussions during the design of the research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.673053/full#supplementary-material

REFERENCES

Amjad, W., Qureshi, W., Farooq, A., Sohail, U., Khatoon, S., Pervaiz, S., et al. (2017). Gastrointestinal side effects of antiarrhythmic medications: a review of current literature. *Cureus* 9:e1646.

Anzini, M., Braile, C., Valenti, S., Cappelli, A., Vomero, S., Marinelli, L., et al. (2008). Ethyl 8-fluoro-6-(3-nitrophenyl)-4H-imidazo[1,5-a][1,4]benzodiazepine-3-carboxylate as novel, highly potent, and safe antianxiety agent. *J. Med. Chem.* 51, 4730–4743. doi: 10.1021/jm8002944

Anzini, M., Valenti, S., Braile, C., Cappelli, A., Vomero, S., Alcaro, S., et al. (2011). New Insight into the Central Benzodiazepine Receptor-Ligand Interactions: design, Synthesis, Biological Evaluation, and Molecular Modeling of 3-Substituted 6-Phenyl-4H-imidazo[1,5-a][1,4]benzodiazepines and Related Compounds. *J. Med. Chem.* 54, 5694–5711. doi: 10.1021/jm2001597

Aureli, S., Di Marino, D., Raniolo, S., and Limongelli, V. (2019). DDT-Drug Discovery Tool: a fast and intuitive graphics user interface for docking and molecular dynamics analysis. *Bioinformatics* 35, 5328–5330.

Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* 35, D301–D303.

Brotzakis, Z. F., Limongelli, V., and Parrinello, M. (2018). Accelerating the Calculation of Protein-Ligand Binding Free Energy and Residence Times using Dynamically Optimized Collective Variables. *J. Chem. Theory Comput.* 15, 743–750. doi: 10.1021/acs.jctc.8b00934

Choudhuri, S. (2014). *Chapter 6 - Sequence Alignment and Similarity Searching in Genomic Databases: BLAST and FASTA, in Bioinformatics for Beginners, Sequence Alignment and Similarity Searching in Genomic Databases*, (Cambridge: Elsevier Academic Press), 133–155.

Comitani, F., Limongelli, V., and Molteni, C. (2016). The Free Energy Landscape of GABA Binding to a Pentameric Ligand-Gated Ion Channel and Its Disruption by Mutations. *J. Chem. Theory Comput.* 12, 3398–3406. doi: 10.1021/acs.jctc.6b00303

Costain, D. W., and Green, A. R. (1978). ß-adrenoceptor antagonists inhibit the behavioural responses of rats to increased brain 5-hydroxytryptamine. *Br. J. Pharmacol.* 64, 193–200. doi: 10.1111/j.1476-5381.1978.tb17289.x

D'Annessa, I., Raniolo, S., Limongelli, V., Di Marino, D., and Colombo, G. (2019). Ligand Binding, Unbinding, and Allosteric Effects: deciphering Small-Molecule Modulation of HSP90. *J. Chem. Theory Comput.* 15, 6368–6381. doi: 10.1021/acs.jctc.9b00319

Egan, C. T., Herrick-Davis, K., Miller, K., Glennon, R. A., and Teitler, M. (1998). Agonist activity of LSD and lisuride at cloned 5HT2A and 5HT2C receptors. *Psychopharmacology* 136, 409–414. doi: 10.1007/s002130050585

Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., and Olson, A. J. (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.* 11, 905–919. doi: 10.1038/nprot.2016.051

Fredriksson, R., Lagerström, M. C., Lundin, L. G., and Schiöth, H. B. (2003). The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* 63, 1256–1272. doi: 10.1124/mol.63.6.1256

Gacasan, S. B., Baker, D. L., and Parrill, A. L. (2017). G protein-coupled receptors: the evolution of structural insight. *AIMS Biophys.* 4, 491–527.

Gopinathan, G., Teravainen, H., Dambrosia, J. M., Ward, C. D., Sanes, J. N., Stuart, W. K., et al. (1981). Lisuride in parkinsonism. *Neurology* 31, 371–371.

Haga, K., Kruse, A. C., Asada, H., Yurugi-Kobayashi, T., Shiroishi, M., Zhang, C., et al. (2012). Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* 482, 547–551.

Hanson, M. A., Cherezov, V., Griffith, M. T., Roth, C. B., Jaakola, V. P., Chien, E. Y., et al. (2008). A specific cholesterol binding site is established by the 2.8 A structure of the human beta2-adrenergic receptor. *Structure* 16, 897–905. doi: 10.1016/j.str.2008.05.001

Hensler, J. G. (2012). Serotonin. *Basic Neurochem.* 15, 300–322.

Hildebrand, M., Hümpel, M., Krause, W., and Täuber, U. (1987). Pharmacokinetics of bromerguride, a new dopamineantagonistic ergot derivative in rat and dog. *Eur. J. Drug Metab. Pharmacokinet.* 12, 31–40. doi: 10.1007/bf03189859

Himori, N., Ishimori, T., Izumi, A., and Hiramatsu, Y. (1977). Effects of ß-adrenoceptor blocking agents, pindolol, alprenolol and practolol on blood pressure and heart rate in conscious renal hypertensive dogs. *Arch Int. Pharmacodyn. Ther.* 225, 152–165.

Hofmann, C., Penner, U., Dorow, R., Pertz, H. H., Jähnichen, S., Horowski, R., et al. (2006). Lisuride, a dopamine receptor agonist with 5-HT2B receptor antagonist properties: absence of cardiac valvulopathy adverse drug reaction reports supports the concept of a crucial role for 5-HT2B receptor agonism in cardiac valvular fibrosis. *Clin. Neuropharmacol.* 29, 80–86. doi: 10.1097/00002826-200603000-00005

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

Kashihara, T., Hirose, M., Shimojo, H., Nakada, T., Gomi, S., Hongo, M., et al. (2014). β2-Adrenergic and M2-muscarinic receptors decrease basal t-tubular L-type Ca2+ channel activity and suppress ventricular contractility in heart failure. *Eur. J. Pharmacol.* 724, 122–131. doi: 10.1016/j.ejphar.2013.12.037

Katritch, V., Cherezov, V., and Stevens, R. C. (2013). Structure-function of the G protein–coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* 53, 531–555. doi: 10.1146/annurev-pharmtox-032112-135923

Limongelli, V. (2020). Ligand Binding Free Energy and Kinetics Calculation in 2020. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 10:e1455.

Limongelli, V., Bonomi, M., and Parrinello, M. (2013). Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci. U. S. A.* 110, 6358–6363. doi: 10.1073/pnas.1303186110

McCorvy, J. D., Wacker, D., Wang, S., Agegnehu, B., Liu, J., Lansu, K., et al. (2018). Structural determinants of 5-HT2B receptor activation and biased agonism. *Nat. Struct. Mol. Biol.* 25, 787–796. doi: 10.1038/s41594-018-0116-7

Moraca, F., Amato, J., Ortuso, F., Artese, A., Pagano, B., Novellino, E., et al. (2017). Ligand binding to telomeric G-quadruplex DNA investigated by funnel-metadynamics simulations. *Proc. Natl. Acad. Sci. U. S. A* 114, E2136–E2145.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc.21256

Nolan, B. T. (1982). Acute suicidal depression associated with use of timolol. *Jama* 247, 1567–1567. doi: 10.1001/jama.1982.03320360019022

Nuti, E., Casalini, F., Avramova, S. I., Santamaria, S., Fabbi, M., Ferrini, S., et al. (2010). Potent arylsulfonamide inhibitors of tumor necrosis factor-alpha converting enzyme able to reduce activated leukocyte cell adhesion molecule shedding in cancer cell models. *J. Med. Chem.* 53, 2622–2635. doi: 10.1021/jm901868z

Papadimas, G. K., Tzirogiannis, K. N., Mykoniatis, M. G., Grypioti, A. D., Manta, G. A., and Panoutsopoulos, G. I. (2012). The emerging role of serotonin in liver regeneration. *Swiss Med. Wkly* 142:w13548.

Pietrokovski, S., Henikoff, J. G., and Henikoff, S. (1996). The Blocks database-a system for protein classification. *Nucleic Acids Res.* 24, 197–200. doi: 10.1093/nar/24.1.197

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* 18, 41–58. doi: 10.1038/nrd.2018.168

Raniolo, S., and Limongelli, V. (2020). FMAP: the Funnel-Metadynamics Advanced Protocol for ligand binding free energy calculations. *Nat. Protoc.* 15, 2837–2866. doi: 10.1038/s41596-020-0342-4

Rascol, O., Slaoui, T., Regragui, W., Ory−Magne, F., Brefel−Courbon, C., and Montastruc, J. L. (2007). *Dopamine Agonists. In Handbook of Clinical Neurology*. Amsterdam: Elsevier, 73–92.

Rask-Andersen, M., Almén, M. S., and Schiöth, H. B. (2011). Trends in the exploitation of novel drug targets. *Nat. Rev. Drug Discov.* 10, 579–590. doi: 10.1038/nrd3478

Rios, S., Fernandez, M. F., Caltabiano, G., Campillo, M., Pardo, L., and Gonzalez, A. (2015). GPCRtm: an amino acid substitution matrix for the transmembrane region of class A G Protein-Coupled Receptors. *BMC bioinformatics* 16:206. doi: 10.1186/s12859-015-0639-4

Sambhara, D., and Aref, A. A. (2014). Glaucoma management: relative value and place in therapy of available drug treatments. *Ther. Adv. Chronic Dis.* 5, 30–43. doi: 10.1177/2040622313511286

Shirakawa, O., Takayoshi, K., and Tanaka, C. (1987). Antimuscarinic effects of antihistamines: quantitative evaluation by receptor-binding assay. *Jpn. J. Pharmacol.* 43, 277–282. doi: 10.1254/jjp.43.277

Sriram, K., and Insel, P. A. (2018). G protein-coupled receptors as targets for approved drugs: how many targets and how many drugs? *Mol. Pharmacol.* 93, 251–258. doi: 10.1124/mol.117.111062

Thomsen, W., Frazer, J., and Unett, D. (2005). Functional assays for screening GPCR targets. *Curr. Opin. Biotechnol.* 16, 655–665.

Van Dam, L. J., and Rolland, R. (1981). Lactation-inhibiting and prolactin-lowering effect of lisuride and bromocriptine: a comparative study. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 12, 323–330. doi: 10.1016/0028-2243(81)90055-1

Venkatakrishnan, A. J., Deupi, X., Lebon, G., Tate, C. G., Schertler, G. F., and Babu, M. M. (2013). Molecular signatures of G-protein-coupled receptors. *Nature* 494, 185–194.

Wacker, D., Fenalti, G., Brown, M. A., Katritch, V., Abagyan, R., Cherezov, V., et al. (2010). Conserved binding mode of human β2 adrenergic receptor inverse agonists and antagonist revealed by X-ray crystallography. *J. Am. Chem. Soc.* 132, 11443–11445. doi: 10.1021/ja105108q

Wacker, D., Wang, S., McCorvy, J. D., Betz, R. M., Venkatakrishnan, A. J., Levit, A., et al. (2017). Crystal Structure of an LSD-Bound Human Serotonin Receptor. *Cell* 168, 377–389. doi: 10.1016/j.cell.2016.12.033

Wasserman, A. J., Proctor, J. D., Allen, F. J., and Kemp, V. E. (1970). Human Cardiovascular Effects of Alprenolol, A ß−Adrenergic Blocker: hemodynamic, Antiarrhythmic, and Antianginal. *J. Clin. Pharmacol. J. New Drugs* 10, 37–49.

Yeagle, P. L., and Albert, A. D. (2007). G-protein coupled receptor structure. *Biochim. Biophys. Acta Biomembr.* 1768, 808–824.

Yuan, X., Raniolo, S., Limongelli, V., and Xu, Y. (2018). The Molecular Mechanism Underlying Ligand Binding to the Membrane-Embedded Site of a G-Protein-Coupled Receptor. *J. Chem. Theory Comput.* 14, 2761–2770. doi: 10.1021/acs.jctc.8b00046

Zhang, Y., DeVries, M. E., and Skolnick, J. (2006). Structure modeling of all identified G protein–coupled receptors in the human genome. *PLoS Comput. Biol.* 2:e13. doi: 10.1371/journal.pcbi.0020013

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: 11 May 2021
doi: 10.3389/fmolb.2021.661520

Reconciling Simulations and Experiments With BICePs: A Review

Vincent A. Voelz[1]*, *Yunhui Ge*[2] and *Robert M. Raddi*[1]

[1] Department of Chemistry, Temple University, Philadelphia, PA, United States, [2] Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, CA, United States

Bayesian Inference of Conformational Populations (BICePs) is an algorithm developed to reconcile simulated ensembles with sparse experimental measurements. The Bayesian framework of BICePs enables population reweighting as a post-simulation processing step, with several advantages over existing methods, including the proper use of reference potentials, and the estimation of a Bayes factor-like quantity called the BICePs score for model selection. Here, we summarize the theory underlying this method in context with related algorithms, review the history of BICePs applications to date, and discuss current shortcomings along with future plans for improvement.

Keywords: Bayesian inference, conformational populations, MCMC, cyclic peptides, peptoids, peptidomimetics, HDX protection factors, molecular simulation

OPEN ACCESS

Edited by:
Maya Topf, Birkbeck, University of London, United Kingdom

Reviewed by:
Kresten Lindorff-Larsen, University of Copenhagen, Denmark
Shruthi Viswanath, National Centre for Biological Sciences, India

***Correspondence:**
Vincent A. Voelz
voelz@temple.edu

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

Received: *30 January 2021*
Accepted: *12 April 2021*
Published: *11 May 2021*

Citation:
Voelz VA, Ge Y and Raddi RM (2021) Reconciling Simulations and Experiments With BICePs: A Review. Front. Mol. Biosci. 8:661520. doi: 10.3389/fmolb.2021.661520

1. INTRODUCTION

Bayesian Inference of Conformational Populations (BICePs) is an algorithm developed to reconcile simulated ensembles with sparse experimental measurements. The inputs to BICePs are: (1) a set of discrete conformational states and corresponding populations predicted from a theoretical prior, and (2) a set of experimental observables. The primary output of BICePs are estimates of reweighted conformational populations that balances the information from theory and experiment using a Bayesian framework.

The Bayesian Inference of Conformational Populations (BICePs) algorithm arose from a need to predict conformational ensembles of organic molecules with significant structural heterogeneity in solution, such as natural product macrocycles and peptidomimetics. Specifically, we aimed to develop an approach that leaned more heavily on high-quality theory/simulation-based conformational ensembles, to be later reconciled with potentially sparse experimental observables.

Existing methods for this purpose, such as NAMFIS (NMR Analysis of Molecular Flexibility in Solution, Cicero et al., 1995) and DISCON (Distribution of in-solution conformations, Atasoylu et al., 2010) were used primarily by the organic chemistry community in the context of NMR refinement. While these methods do estimate populations of conformational states, they are essentially a kind of "maximum parsimony" method, where all possible solution-state conformations are enumerated in order to find a small number of structures compatible with NMR-based constraints. Such approaches are less useful for simulated structural ensembles, for which ensemble-averaged observables should be restrained, in a way that can sufficiently account for uncertainties in experimental measurements.

Another class of algorithms, categorized as "maximum entropy" approaches (Pitera and Chodera, 2012; Bonomi et al., 2017; Orioli et al., 2020) focus primarily on using bias potentials to enforce constraints on an experimental observable throughout the course of a molecular simulation. While this can be approximated efficiently in practice by restraining replica-averaged observables (Vendruscolo et al., 2003; Best and Vendruscolo, 2006) it must be modified to account

for experimental uncertainty, a problem more recently addressed by the Metainference algorithm of Bonomi and Vendruscolo (Bonomi et al., 2016a,b) which employs Bayesian inference.

In contrast to this approach, we sought a method that could reweight a *discrete* set of conformational populations as a "post-processing" step, after a simulation was performed. Such *post-hoc* reweighting would nicely complement Markov State Model approaches for biomolecular simulation, which require partitioning of trajectory data into discrete conformational states. Another reason to develop a reweighting approach was the growing problem of bespoke force field parameterization for peptidomimetics. A reweighting approach could enable a sufficiently accurate general force field [e.g., GAFF Wang et al., 2004] to generate an initial model of conformational populations that could then be further refined against experimental data.

BICePs was modeled closely after the Inferential Structural Determination (ISD) algorithm of Rieping, Habeck, and Nilges (Rieping et al., 2005). Like BICePs, ISD is a Bayesian approach where simulated conformational populations are used as the Bayesian prior, and experimental restraints form the likelihood function. The full posterior distribution of conformational states and model parameters is then sampled using a Monte Carlo algorithm (Habeck et al., 2006). But as we soon discovered when developing BICePs, not all experimental restraints impart the same amount of information, and BICePs makes critical improvements over ISD by accounting for this fact.

The information gained upon obtaining a measurement is relative to the prior information we possess. For example, suppose we want to use Bayesian inference to refine the conformational distribution of a linear peptide, given an experimental distance measurement between two residues. The measurement is highly informative if the the residues are distant in sequence, but non-informative if the residues are close in sequence. To account for a more diverse range of informative experimental restraints, BICePs implements *reference potentials*, which are discussed more fully in the Theory section.

As a consequence of better accounting for the information content of experimental restraints, BICePs is able to calculate a Bayes factor-like quantity, that we call the *BICePs score*, that can be used for model selection. The BICePs score is highly useful: it is a number that can report the extent to which a conformational ensemble is consistent with experimental data. Not only can this be used to show that reweighted populations are more consistent with experimental data, it can also be used to rank different simulated ensembles by their accuracy in reproducing experimental observables (Ge and Voelz, 2018). While there are still some improvements to BICePs needed to use this for automated force field validation (see Discussion), the BICePs score is highly useful, and we expect it will continue to provide attractive incentive to use this algorithm.

In the remainder of this review, we will first discuss the theory underlying the BICePs algorithm, and describe some of the ways we implement this theory to sample the Bayesian posterior distribution of conformational state and model parameters. We then provide a case-by-case review of past examples where BICePs has been applied to model conformational distributions. Finally, we discuss some of the shortcomings of BICePs and ongoing challenges we hope to address with future improvements to BICePs.

2. THEORY

2.1. Bayesian Inference

The goal of BICePs is to model a *posterior* distribution $P(X|D)$ of conformational states X, given some experimental data D. This posterior probability is proportional to a product of (1) a *likelihood* function $Q(D|X)$ representing experimental restraints, and (2) a *prior* distribution $P(X)$.

$$P(X|D) \propto Q(D|X)P(X) \tag{1}$$

The prior distribution, $P(X)$, represents prior knowledge about the populations of conformational states X derived from theoretical modeling. This distribution can computed directly from a molecular simulation, or come from any number of theoretical models of the conformational free energy landscape (e.g., from QM calculations).

The likelihood function, $Q(D|X)$, reflects how well a given conformation X agrees with experimental measurements. It is assumed to obey a normally-distributed error model of the form:

$$Q(D|X,\sigma) = \prod_j \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\left[r_j(X) - r_j^{\text{exp}}\right]^2 / 2\sigma^2\right). \tag{2}$$

Here, the data D comprise a set of experimental observables indexed by $j = 1, ..., N_j$. The $r_j(X)$ represent observables back-calculated from the theoretical model (ensemble-averaged over states within X), and r_j^{exp} represent the experimental values of each observable. In Equation (2), we assume that each experimental observable has the same uncertainty σ. In practice, different types of observables r_j can be assigned specific uncertainties σ_j, although this is usually done in groups (different values of σ_j for sets of NOE distances, J-coupling constants, etc.) for the sake of computational efficiency. There are of course many situations where experimental uncertainty can vary even within different sets measurements, which can be addressed by defining custom restraint groups.

The likelihood function $Q(D|X)$ can be thought of as the quantity that reweights the prior estimate of the population $P(X)$. Conformational states X that better agree with the experimental measurements get higher weights. An important distinction to note: as BICePs is currently formulated, the likelihood function $Q(D|X)$ compares the experimental value r_j^{exp} to the back-calculated observable $r_j(X)$ of a *single* conformational state X, rather than an ensemble-averaged back-calculated observable $\langle r_j \rangle = \sum_X r_j(X)P(X)$. Consequently, the error model parameter σ reflects both uncertainty in the experimental measurements and heterogeneity in the conformational ensemble. Errors in the forward model $r_j(X)$ are included in σ, and in many cases may dominate the experimental errors (chemical shifts being the most dramatic example).

As for choosing values of the uncertainty parameter(s) σ, these uncertainties are usually not known *a priori*, and must be treated

as a so-called *nuisance parameter*, which can be modeled using some prior model $P(\sigma)$:

$$P(X,\sigma|D) \propto Q(D|X,\sigma)P(X)P(\sigma) \tag{3}$$

While we don't know the exact value of σ, we treat $P(\sigma)$ as non-informative Jeffreys prior $[P(\sigma) \sim 1/\sigma]$, and include this parameter in the posterior in order to sample the joint distribution of (X,σ). Then $P(X|D)$ can then be obtained as the marginal distribution $\int P(X,\sigma|D)d\sigma$. In the case where estimates of the errors from experiments *are* known, a limited range of possible σ values can be imposed.

Because the likelihood function enforces restraints on individual conformational states (not ensemble-averages), $P(X|D)$ represents an "uncertainty ensemble" rather than a "statistical ensemble" of conformational populations, to use the language of Bonomi et al. (2017). However, it is quite useful think of $P(X|D)$ as conformational populations, as we show in the examples below. For example, if $P(X|D)$ gives equal values for two conformational states, then BICePs predicts they are equally consistent with the experimental data. While BICePs doesn't explicitly predict a mixture of two conformations, by maximum ignorance (i.e., MaxEnt) we would infer equal populations in the conformational ensemble. Future improvements to BICePs will address this by constraining ensemble averages across multiple replicas (see Discussion).

2.2. Reference Potentials

While the likelihood function $Q(D|X)$ weights the conformational space X, the actual restraints exist in some restraint space $\mathbf{r}$, a low-dimensional projection of the state space of X. As a result, we need to introduce a reference potential $Q_{\text{ref}}(\mathbf{r})$ that reflects the distribution of observables $\mathbf{r}$ in the *absence* of any experimental measurements. With this modification, Equation (1) becomes:

$$P(X|D) \propto \left[\frac{Q(\mathbf{r}(X)|D)}{Q_{\text{ref}}(\mathbf{r}(X))}\right] P(X). \tag{4}$$

The negative logarithm of the bracketed weighting function, $-\ln[Q(\mathbf{r}|D)/Q_{\text{ref}}(\mathbf{r})]$, can be thought of as equivalent to a potential of mean force (Hamelryck et al., 2010; Olsson et al., 2011, 2013). With a proper reference potential, the relative information content of each restraint becomes meaningful. In our previous work, we have shown that using BICePs with proper reference potentials can be essential for obtaining accurate results (Voelz and Zhou, 2014; Ge and Voelz, 2018).

As an example of why reference potentials are needed, consider experimental measurements of interresidue distances in a protein. A distance measurement of 5 Å for a pair of residue indices $(i, i+2)$ is essentially non-informative, since we already know these residues are close in sequence along the polypeptide chain, while a distance measurement of 5 Å for $(i, i+50)$ is highly informative. The ratio $Q(\mathbf{r}(X)|D)/Q_{\text{ref}}(\mathbf{r}(X))$ is needed to correctly characterize the change in our state of knowledge.

The choice of what reference potential to use in a particular situation is subject to some interpretation. Since BICePs is designed to be used with sparse and/or noisy experimental data, the likelihood function $Q(D|X)$ typically enforces experimental restraints smoothly over broad ranges of values. Similarly, reference potentials should be sufficiently smooth and broad, to avoid regions of restraint space with unrealistically small values of $Q_{\text{ref}}(\mathbf{r})$, which may in turn produce artificially inflated weights for certain conformations. For this reason, we advocate the use of conservative reference potentials, which do not make unnecessary assumptions about the underlying distribution of a given observable in the absence of experimental information.

We currently support three kinds of reference potentials in our software implementation of BICePs: (1) uniform (non-informative), (2) exponential, and (3) Gaussian. An exponential reference potential is the least-informative distribution if only the first moment of Q_{ref} (the mean, $\langle r\rangle$) is known. A Gaussian distribution is the least-informative distribution if only the first and second moments are known ($\langle r\rangle$ and $\langle r^2\rangle$).

As an interesting example, consider the reference potential used for a set of interproton distances measured in an NMR study of a 14-membered macrocycle, a situation we considered in Voelz and Zhou (2014). In the absence of all other information, our reference information is that the space of molecular conformations are 14-membered rings. At the very least, the the distribution of interproton distances must be non-negative, and bounded from above. To get an idea of the empirical distribution of possible interproton distances, we examined all input conformations to BICePs, regardless of their energy, and found no clear pattern other than a well-defined mean. Therefore, we chose an exponential function as the reference potential. In practice, the reference potential was fairly flat, since the average interproton distance had a mean near 4 Å, and a maximum near 5 Å.

2.3. Sampling the Posterior Using MCMC

Markov Chain Monte Carlo (MCMC) is used to sample the posterior distribution of X and σ, with $-\ln P(X,\sigma|D)$ used an the effective energy function. The energy function can be obtained as the negative logarithm of the posterior probability given in Equation (3):

$$\begin{aligned}-\ln P(X,\sigma|D) \;=\; &\big(N_j+1\big)\ln\sigma + \chi^2(X)/2\sigma^2 - \ln Q_{\text{ref}} \\ &+ \big(N_j/2\big)\ln 2\pi - \ln P(X).\end{aligned} \tag{5}$$

The quantity $\chi^2(X)$ is the sum of squared errors, computed as

$$\chi^2(X) = \sum_j w_j \left(r_j(X) - r_j^{\text{exp}}\right)^2 \tag{6}$$

where w_j is a weight parameter designated for equivalent observables (For example: $w_j = 1/3$ is used for hydrogens in a methyl group).

The Metropolis-Hastings algorithm is used to perform MCMC sampling of the energy function defined in Equation (5), yielding an estimate of the full posterior distribution $P(X,\sigma|D)$. The most probable values of σ can be obtained by the marginal distribution $P(\sigma|D) = \int P(X,\sigma|D)dX$, and the state populations are estimated as $P(X|D) = \int P(X,\sigma|D)d\sigma$ (**Figure 1**).

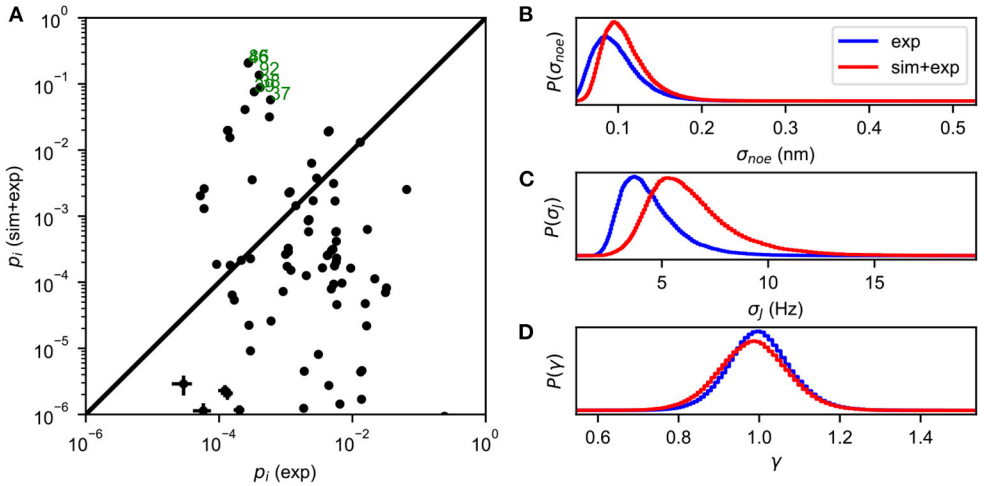


FIGURE 1 | An example of BICePs output for albocycline (Liang et al., 2018). **(A)** A comparison of conformational state populations p_i (exp) inferred using only experimental restraints, vs. BICePs populations p_i (sim + exp) inferred using a combination of the simulation-based prior and experimental restraints. States on the lower right are highly compatible with experimental restraints, but are conformationally strained according the simulation model. Conformational states near the top of the graph are both reasonably compatible with experimental restraints, and highly-populated according to the simulation model. States labeled in green correspond closely to the two crystal isoforms of albocycline. **(B)** The marginal posterior distribution of σ_{noe}, the uncertainty parameter for NOE distance restraints. **(C)** The marginal posterior distribution of σ_J, the uncertainty parameter for *J*-coupling constants. **(D)** The marginal posterior distribution of γ, the scaling parameter for the NOE distances, remains near 1.0 throughout the MCMC sampling.

2.3.1. Enhanced Sampling of the Posterior

Accurate BICePs results require sufficiently converged sampling of the entire (X, σ) landscape. To achieve enhanced sampling of $P(X, \sigma|D)$, we use a free energy perturbation (FEP) method, where posterior sampling for a series of models with priors $P_\lambda(X) \sim [P(X)]^\lambda$, where $0 \leq \lambda \leq 1$. The λ value serves to linearly scale the $-\ln P(X)$ term in Equation (5). The expanded ensemble of posterior distributions $P_\lambda(X, \sigma|D)$ thus spans a range of prior information: When $\lambda = 0$, the prior $P_\lambda(X)$ prior is uniform, and there is no information from theoretical modeling included in the sampling. When $\lambda = 1$, all the information from theoretical modeling is included in the sampling.

In the current implementation of BICePs, MCMC is performed in parallel for a fixed number of λ values ranging from 0 to 1. The multistate Bennett acceptance ratio (MBAR) free energy estimator (Shirts and Chodera, 2008) is then used to integrate samples from each ensemble to make statistically optimal estimates of all $P_\lambda(X|D)$.

2.4. The BICePs Score

The quality of a model k that uses a prior $P^{(k)}(X)$ from theoretical modeling can be assessed by the posterior likelihood $Z^{(k)}$ of model k:

$$Z^{(k)} = \int P^{(k)}(X, \sigma|D) dX d\sigma = \int P^{(k)}(X) Q(X) dX. \quad (7)$$

One way to think of $Z^{(k)}$ is as an integral over the entire input space (including nuisance parameters) of the model. Another way, however, is to think of $Z^{(k)}$ as an *overlap* integral between the prior $P^{(k)}(X)$ and a likelihood function $Q(X) = \int [Q(\mathbf{r}(X)|D, \sigma)/Q_{\text{ref}}(\mathbf{r}(X))] P(\sigma) d\sigma$. This integral reaches the maximum when $P^{(k)}(X)$ most closely matches the likelihood distribution $Q(X)$ specified by the experimental restraints.

Suppose we have two models (1) and (2) with priors $P^{(1)}$ and $P^{(2)}$, and we want to know which one is more consistent with experimental measurements. In Bayesian statistics, the comparison is often made using the ratio of posterior model probabilities, $Z^{(1)}/Z^{(2)}$, also called the Bayes factor.

In BICePs, we consider a free energy-like quantity, called the BICePs score:

$$f^{(k)} = -\ln \frac{Z^{(k)}}{Z_0}, \quad (8)$$

which compares a model probability $Z^{(k)}$ to a standard reference Z_0 where no theoretical information is used (i.e., a model using the prior $P_\lambda(X)$ where $\lambda = 0$). The use of this standard reference is useful in several ways. For one, if the BICePs score $f^{(k)}$ is *positive* for a given model k, it means that the theoretical model is *worse* than a totally uninformative prior–the theoretical model is somehow inconsistent with experiment. More importantly, since the BICePs score $f^{(k)}$ is always computed against an absolute reference, it is a scalar quantity that can be used to perform model selection. The BICePs score therefore can be very useful for automatic model selection; for example, molecular simulation force field validation and parameterization (Ge and Voelz, 2018).

Unlike maximum-likelihood approaches, The BICePs score has the advantage of avoiding overfitting to a particular set of experimental observable values, especially when the data are sparse and/or noisy. Consider an alternative approach where the values of σ_j that maximize the posterior are identified for two different models and used to compute χ^2 values for model selection. The χ^2 values only compare the models at particular

points in parameter space, while the BICePs score considers the total evidence integrated over all of parameter space.

3. APPLICATIONS OF BICEPS

Applications of BICePs to date fall into two main categories: studies of small molecules like peptides and peptidomimetics, and studies of larger proteins like apomyoglobin (**Figure 2**).

3.1. Modeling Macrolide Antibiotics

The first application of BICePs, described in the seminal article that first introduced the algorithm, was to determine the solution-state conformational populations of the 14-membered macrolide antibiotic cineromycin B (Voelz and Zhou, 2014). Knowledge of solution-state structure is essential to identify potential targets of natural products, and to rationally design new kinds macrolide antibiotics.

A combination of theoretical modeling and sparse experimental NMR observables were used as input to BICePs. The theoretical modeling was performed in two steps. First, implicit solvent replica-exchange molecular dynamics (REMD) simulation in GAFF was performed to exhaustively sample the conformational landscape. The resulting sampling was then clustered into 100 discrete states. Next, each cluster center was subjected to QM minimization and single-point energy calculation at the B3LYP/6-3111G(2d,p)//HF/6-31G(d) level of theory. State populations were considered to be Boltzmann-distributed according to the computed QM energies. The sparse experimental constraints consisted of 13 interproton NOEs and five vicinal $^3J_{HH}$ coupling constants.

For this system, BICePs predicted a nearly equal mixture of two main conformational populations, each closely similar in structure to the two crystal isoforms found for albocycline, the O-methylated analog of cineromycin B. This work also showed the importance of the reference potentials in producing more correct posterior models.

In subsequent work, BICePs predicted a similar (nearly equal) mixture of solution-state populations for albocycline, using 12 NOE distance restraints and seven dihedral restraints from vicinal $^3J_{HH}$ coupling constants (Chatare and Andrade, 2017). This information helped inform molecular simulation and computational docking studies of albocycline binding to MurA, an enzyme involved in peptidoglycan biosynthesis, a potential new target for Methicillin-resistant *Staphylococcus aureus* (MRSA) infection (Liang et al., 2018).

3.2. Modeling Peptoid Foldamers

Peptoids (N-substituted oligoglycines) are a class of sequence-specific peptidomimetics that can be easily synthesized, and fold into unique structural scaffolds (Sun and Zuckermann, 2013). While the peptoid backbone is achiral and lacks hydrogen bond donors, rational design of N-substituents can control the amide *cis*/*trans* populations and secondary structure. An important goal for molecular modeling and simulation of these systems is to accurately predict solution-state conformational populations. Reliable methods to do this would enable the computational design of preorganized peptoid structural scaffolds to function as new bio-inspired materials or therapeutics (Voelz et al., 2011; Butterfoss et al., 2012; Kang et al., 2017; Schneider et al., 2018; Gimenez et al., 2019).

A particular challenge in simulating peptoids is the lack of accurate force fields. Unlike peptides, the chemical diversity of N-substituents is practically limitless, with each new peptoid residue requiring custom parameterization. BICePs can help avoid this by using a general-purpose force field to generate a prior conformational distribution, to be further refined against experimental data.

An example of this approach was pursued by Mukherjee et al. to model the solution-state conformational populations of an

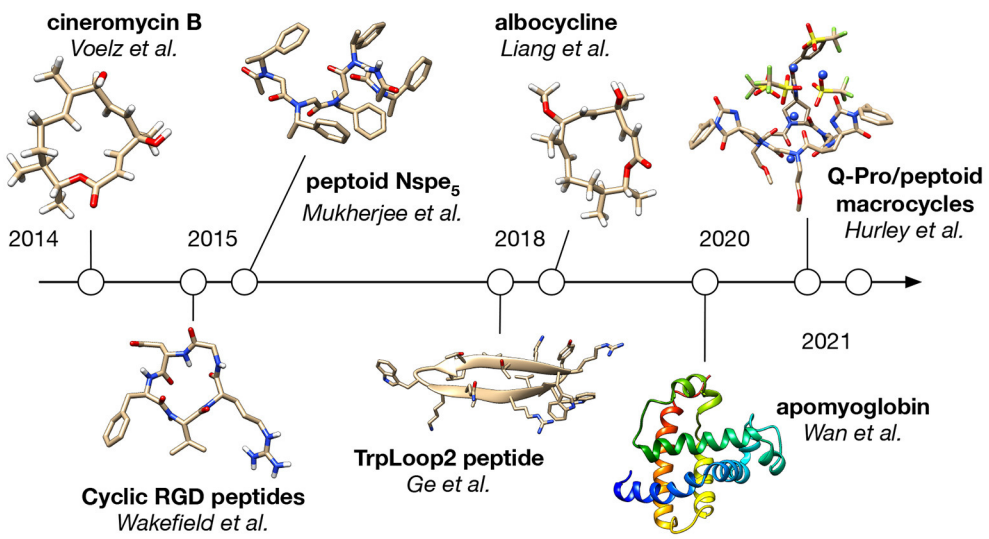


FIGURE 2 | A timeline of BICePs application.

(S)-*N*-(1-phenylethyl) glycine pentamer, $(Nspe)_5$, whose bulky chiral N-subtituents help this sequence fold into a right-handed *cis*-amide helix (Mukherjee et al., 2015). Disagreement between *ab initio* dihedral scans of the Nspe residue and the results of GAFF simulations motivated the development of an improved backbone potential, GAFF-ϕ, to better model the right-handed (negative ϕ-angle) preference of Nspe oligomers in solution.

BICePs was used to reweight GAFF and GAFF-ϕ predictions using sparse experimental restraints derived from previously published NMR studies: NOE distances (Armand et al., 1998) and *cis*/*trans* amide equilibria ($K_{ct} \sim 2.5$). BICePs scores for both GAFF and GAFF-ϕ were negative, suggesting the models are compatible with experiment. However, the GAFF-ϕ model was found to have a likelihood of 1.5 times that of the GAFF model, indicating it to be superior. Indeed, GAFF-ϕ predicted a much higher *cis*-amide helix population for $(Nspe)_5$, consistent with previous NMR refinement and circular dichroism measurements.

By reweighting pre-defined conformational states, BICePs also provides a convenient methodology to avoid costly sampling. Unlike peptides, peptoids can populate both *cis* and *trans* amide conformations. Amides have large isomerization barriers in most force fields, typically requiring enhanced sampling methods like REMD to sample the full conformational landscape of peptoids. Thus, the "post-processing" aspect of BICePs can help to avoid the costly alternative of having to perform slow-to-converge simulations in the presence of restraints.

More recently, this approach was used to determine the solution-state structure and ion-binding mechanism of cyclic peptoid-spiroligomer hexamer macrocycles (Hurley et al., 2021). Northrup et al. found that particular sequences of alternating Q-proline and peptoid residues are able to bind metal cations, forming highly preorganized structures in the process (Northrup et al., 2020). To model this process, the BICePs algorithm was used to reconcile conformational populations from implicit-solvent REMD simulations in GAFF, against sparse experimental ROESY correlations. While GAFF simulations predict a range of macrocycle conformations with an overall preference for *cis*-amide backbones, the reweighted populations had a preference for *trans* amides, with the most populated conformation having five of six amides in the *trans* state. This conformation was then used to initiate more accurate explicit-solvent simulations of macrocycles in the presence of K^+ and Na^+ cations, in which several direct-binding events–coupled with a transition to an all-*trans* state–were observed. In validation of this model, the authors were able to correctly rank the ion-, solvent-, and sequence-dependence of cation-binding in agreement with experiment. Interestingly, a racemic crystal structure obtained for a peptoid-spiroligomer macrocycle in the absence of bound cation contains a mixture of *cis* and *trans* backbone amide, underscoring the need for an integrated modeling approach using BICePs to determine cation bound macrocycle conformations in solution.

3.3. Modeling Linear and Cyclic Peptides

Like peptoid foldamers, both cyclic and linear peptides can form preorganized structures in solution, and BICePs can be a valuable tool to help computationally model and design sequences with desirable solution-state properties. Wakefield et al. (2015) simulated 18 cyclic RGD peptides studied extensively by the Kessler group using NMR, including the anticancer drug candidate cilengitide, cyclo(RGDf-[*N*-Me]V), which targets integrin $\alpha_V\beta_3$ (Dechantsreiter et al., 1999; Mas-Moruno et al., 2010). BICePs was used to validate excellent agreement between simulations and experimental NOE distances. The results reproduce the highly preorganized solution conformation of cilengitide, which has the highest affinity to integrin. Estimated differences in conformational entropy suggested that N-methylation provided about 0.5 kcal mol^{-1} of stabilization, and rigid non-natural amino acid mimics can provide even more preorganzation.

Ge and Voelz (2018) explored how the BICePs score could be used for force field validation and parameterization. Using a 2D lattice model as a toy system, they first demonstrated that BICePs was able to select the correct value of an interaction energy parameter given ensemble-averaged experimental distance measurements. The toy model was used to study the sensitivity of the results to the choice of reference potential, the number of conformational clusters used in the calculations, and the robustness of the calculation to experimental noise and measurement sparsity. In this work, the authors introduce support for chemical shift modeling in BICePs, which they use as experimental restraints to refine conformational populations of designed β-hairpin TrpLoop2 peptides in a number of force fields (Ge et al., 2017). BICePs results show unambiguously that explicit-solvent simulations in AMBER ff99-ildn-nmr (Li and Brüschweiler, 2010; Lindorff-Larsen et al., 2010) yield models most consistent with the experimental data. While this work suggests that BICePs is up to the task of model selection in the context of all-atom simulations, it also reveals several challenges that need to be overcome to perform these calculations reliably (see Discussion).

3.4. Reconciling Models of Globular Proteins With Experimental HDX Data

Recent work by Wan et al. expands the scope of BICePs—both in terms of system size and sampling complexity—by introducing support for yet another experimental observable: hydrogen/deuterium exchange (HDX) protection factors (Wan et al., 2020). HDX protection factors are challenging to enforce in molecular simulations, because they are *dynamical* restraints, corresponding to the relative rates of local unfolding events, where solvent exposure of backbone amides leads to exchange. For BICePs to refine structural ensembles using HDX protection factors, it requires a structural *proxy* that correlates with local unfolding, which the authors capture using the simple model:

$$\ln \mathrm{PF}_i = \beta_c \langle N_c \rangle_i + \beta_h \langle N_h \rangle_i + \beta_0. \tag{9}$$

In this model, the logarithm of the protection factor for residue i is predicted by the ensemble average number of heavy-atom contacts $\langle N_c \rangle_i$ and hydrogen bonds $\langle N_h \rangle_i$.

The free parameters in this model, λ (the β parameters and others defining how contacts and hydrogen bonds are tallied), are first determined using Bayesian inference, by training on two ultralong simulation trajectories of ubiquitin and bovine pancreatic trypsin inhibitor (BPTI), each well-studied systems

with published experimental protection factors. The result is not a set of optimal (maximum-likelihood) parameters λ^*, but rather the full posterior distribution of parameters $P(\lambda)$, which is imported into the likelihood model for BICePs (More details can be found at https://github.com/vvoelz/HDX-forward-model). All parameters are then treated as nuisance parameters that are sampled in the BICePs posterior distribution.

To test this approach, Wan et al. applied the modified BICePs method to apomyoglobin, which has a disordered helix F and C-terminal portion of helix H in the absence of heme at pH 6. NMR studies provide no structural information for these regions, but HDX protection factors and chemical shifts are available. To model the structural ensemble of these regions, a series of simulations were performed at different temperatures and different bias potentials to encourage local unfolding. The resulting trajectory data was used to construct several competing multi-ensemble Markov Models (MEMMs) (Wu et al., 2016), where each could be evaluated using the BICePs score. The best-scoring model predicts a mixture of two predominant conformational states, one with a partially disordered yet compact helix F and other having a more disordered and exposed helix F, consistent with slow chemical exchange for helix F. Using the populations of these states predicted by BICePs, back-calculation of the HDX protection factors results in values that correlate well the experimental values ($R^2 = 0.72$).

4. DISCUSSION

In the future, we expect that BICePs will play an increasingly important role in molecular simulation-based prediction and design, for several reasons. First, unlike many similar algorithms for Bayesian inference, which enforce restraints during the course of a molecular simulation, BICePs can be implemented as a post-processing step. This means the algorithm should be considerably easier to implement and utilize across many applications.

Second, the ability to "tune" predictions of force fields using sparse experimental restraints can eliminate the need for custom parameterization, which can widen the scope of applications that can be addressed by general-purpose force fields. This is evidenced by the many examples of peptidomimetic and peptoid modeling we have described above. A further avenue, made possible by Markov state models (Prinz et al., 2011; Bowman et al., 2013), is to obtain reweighted predictions of equilibrium populations from BICePs to infer improved kinetic properties, through maximum caliber (MaxCal) approaches, for instance (Dixit et al., 2015; Wan et al., 2016; Ghosh et al., 2020).

Third, the BICePs score provides an unambiguous metric to rank model quality and perform model selection. As discussed above, this makes objective force field evaluation possible. Given a standard test set of molecular systems and associated corpus of experimental observables, BICePs could be a uniquely suitable Bayesian approach for systematically benchmarking and/or parameterizing new potentials. Similarly, the BICePs score could help quantify the progress toward an objective in adaptive sampling.

For BICePs to achieve the status of indispensable tool, there are several practical shortcomings and improvements that we are working to address.

4.1. Future Algorithmic Improvements

4.1.1. Replica Averaging

One conceptual problem with BICePs and related methods like ISD is that the likelihood function compares individual conformational states to ensemble-averaged experimental observables. As result, the uncertainty parameter σ reflects a combination of both agreement with the experimental measurements and heterogeneity in the conformational ensemble (Bonomi et al., 2016a). A better comparison–and one that will result in lower uncertainty in most cases–is a likelihood function that compares a predicted ensemble-average to experimental observables. A simple way to achieve this, implemented currently in algorithms, such as Metainference (Löhr et al., 2019), is to use a forward model that incorporates the average of multiple MCMC *replicas*. In the limit of large numbers of replicas, such a likelihood function results in the least-biased, maximum entropy (MaxEnt) posterior distribution given ensemble-averaged experimental constraints (Pitera and Chodera, 2012; Cavalli et al., 2013; Roux and Weare, 2013; Hummer and Köfinger, 2015; Bonomi et al., 2016a; Xu, 2019).

One issue we believe replica averaging will improve is the performance of BICePs when used with many experimental restraints. This will increase the impact of BICePs by enabling its application to larger systems with many structural measurements. When modeling peptides with many NOE distance restraints (as in Ge et al., 2017; Ge and Voelz, 2018), we have noticed that while BICePs is able to correctly predict solution-state structures, it can overestimate the posterior populations of folded states. This occurs because particular conformational states that satisfy multiple restraints are highly rewarded by the likelihood function. This behavior is akin to the many constraint-based NMR structural refinement algorithms which seek to generate ensembles of structures that satisfy all or most distance constraints. A similar artifact was found by Ge et al. (2020) when evaluating MSM models of a series of cyclic β-hairpin peptides against structural NMR observables measured by Danelius et al. (2016).

In the replica-averaging section of the Discussion, we discuss this fairly extensively. The issue is not the system *size* *per se* (we have successfully applied BICePs to apomyglobin, a large globular protein, for example) but large numbers of experimental restraints, which become problematic because the likelihood function currently uses a forward model for individual states rather than ensemble-averages. In light of the reviewer's comments, we have added to this in our revised manuscript:

With replica averaging, direct comparison (via the BICePs score) between predictions from BICePs and constraint-based algorithms like NAMFIS (Cicero et al., 1995) should yield more favorable results.

4.1.2. Hamiltonian Replica Exchange

As mentioned in the Theory section, better estimations of conformational populations and more accurate BICePs scores are achieved by implementing a free energy perturbation-like framework, in which parallel MCMC trajectories are perfomed for a series of theoretical priors scaled by $\lambda \in [0, 1]$. An issue that arises from this approach is the inability to sample all states in a reasonably low number of iterations, especially when $\lambda = 1$.

To enhance the sampling of all the states (across all the λ-ensembles), we aim to implement Hamiltonian replica exchange in future versions of BICePs, an approach previously pioneered with ISD (Habeck et al., 2005). In this approach, parallel MCMC trajectories are coupled so that exchanges of conformational states across λ-ensembles are attempted at regular intervals and accepted according to the Metropolis criterion.

4.2. Support for More Experimental Observables and Reference Potentials

Another area of improvement we are working on is the incorporation of more experimental observables, and support for users to be able to extend BICePs by adding custom experimental restraints and reference potentials with relative ease. Our most recent addition to the roster of supported experimental observables is HDX protection factors, $\ln \mathrm{PF}_i$. Custom experimental restraints will require a user to write a derived class and a few simple methods to parse input data files, compute a sum of squared errors, and specify the posterior $-\ln P$ (i.e., the energy function).

Small angle X-ray scattering (SAXS) has proven to be very useful for determining molecular shape and resolving structural dynamics over large range of biomolecular sizes (Bonomi et al., 2017). In the future, we hope to support SAXS observables as experimental restraints, joining the ranks of other Bayesian inference algorithms that can utilize such data (Antonov et al., 2016; Bonomi and Camilloni, 2017; Shevchuk and Hub, 2017; Potrzebowski et al., 2018). One issue to consider is how best to enforce uncertainties when mixed with other types of data, since SAXS experiments typically have a large number of not fully independent measurements. Here a Bayesian approach that can automatically "tune" uncertainties might be particularly powerful.

5. CONCLUSION

We have reviewed the theory and application of BICePs, an algorithm for Bayesian Inference of Conformational Populations, that has several advantages over similar methods. In BICePs, reweighting of populations can be performed as a post-processing step, with proper reference potentials. A review of previous applications demonstrates the utility of BICePs for improving the predictions of general-purpose force fields for modeling and designing peptidomimetics. A unique feature of the algorithm is the BICePs score, which can be used for objective, systematic model selection.

Since the first inception of the BICePs algorithm (Voelz and Zhou, 2014) (which we call "BICePs 1.0") many modifications have been implemented, including support for more experimental observables, such as chemical shifts and HDX protection factors, and improved analysis and visualization. We have officially released the improved algorithm (BICePs 2.0) at https://github.com/vvoelz/biceps. This latest version is designed to lower the barriers for researchers to use and extended the BICePs algorithm.

AUTHOR CONTRIBUTIONS

VV, YG, and RR contributed to the conception, writing, and graphical figures in this work. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported in part by National Institutes of Health grant 1R01GM123296. Calculations were performed on HPC resources supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189. Research was also performed on the CB2RR cluster made possible through NIH Research Resource computer instrumentation grant S10-OD020095.

ACKNOWLEDGMENTS

We thank the editors for the invitation to contribute to this special article collection.

REFERENCES

Antonov, L. D., Olsson, S., Boomsma, W., and Hamelryck, T. (2016). Bayesian inference of protein ensembles from SAXS data. *Phys. Chem. Chem. Phys.* 18, 5832–5838. doi: 10.1039/C5CP04886A

Armand, P., Kirshenbaum, K., Goldsmith, R. A., Farr-Jones, S., Barron, A. E., Truong, K. T., et al. (1998). Nmr determination of the major solution conformation of a peptoid pentamer with chiral side chains. *Proc. Natl. Acad. Sci. U.S.A.* 95, 4309–4314. doi: 10.1073/pnas.95.8.4309

Atasoylu, O., Furst, G., Risatti, C., and Smith III, A. B. (2010). The solution structure of (+)-spongistatin 1 in DMSO. *Organ. Lett.* 12, 1788–1791. doi: 10.1021/ol100417d

Best, R. B., and Vendruscolo, M. (2006). Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of Ci2. *Structure* 14, 97–106. doi: 10.1016/j.str.2005.09.012

Bonomi, M., and Camilloni, C. (2017). Integrative structural and dynamical biology with PLUMED-ISDB. *Bioinformatics* 33, 3999–4000. doi: 10.1093/bioinformatics/btx529

Bonomi, M., Camilloni, C., Cavalli, A., and Vendruscolo, M. (2016a). Metainference: a Bayesian inference method for heterogeneous systems. *Sci. Adv.* 2:e1501177. doi: 10.1126/sciadv.1501177

Bonomi, M., Camilloni, C., and Vendruscolo, M. (2016b). Metadynamic metainference: enhanced sampling of the metainference ensemble using metadynamics. *Sci. Rep.* 6:31232. doi: 10.1038/srep31232

Bonomi, M., Heller, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* 42, 106–116. doi: 10.1016/j.sbi.2016.12.004

Bowman, G. R., Pande, V. S., and Noé, F. (2013). *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Vol. 797. Dordrecht: Springer Science & Business Media.

Butterfoss, G. L., Yoo, B., Jaworski, J. N., Chorny, I., Dill, K. A., Zuckermann, R. N., et al. (2012). *De novo* structure prediction and experimental characterization of folded peptoid oligomers. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14320–14325. doi: 10.1073/pnas.1209945109

Cavalli, A., Camilloni, C., and Vendruscolo, M. (2013). Molecular dynamics simulations with replica-averaged structural restraints generate structural

ensembles according to the maximum entropy principle. *J. Chem. Phys.* 138:03B603. doi: 10.1063/1.4793625

Chatare, V. K., and Andrade, R. B. (2017). Total synthesis of (-)-albocycline. *Angew. Chem. Int. Ed.* 56, 5909–5911. doi: 10.1002/anie.201702530

Cicero, D., Barbato, G., and Bazzo, R. (1995). NMR analysis of molecular flexibility in solution: a new method for the study of complex distributions of rapidly exchanging conformations. Application to a 13-residue peptide with an 8-residue loop. *J. Am. Chem. Soc.* 117, 1027–1033. doi: 10.1021/ja00108a019

Danelius, E., Pettersson, M., Bred, M., Min, J., Waddell, M. B., Guy, R. K., et al. (2016). Flexibility is important for inhibition of the MDM2/p53 protein-protein interaction by cyclic β-hairpins. *Organ. Biomol. Chem.* 14, 10386–10393. doi: 10.1039/C6OB01510G

Dechantsreiter, M. A., Planker, E., Mathä, B., Lohof, E., Hölzemann, G., Jonczyk, A., et al. (1999). N-methylated cyclic RGD peptides as highly active and selective $\alpha v\beta 3$ integrin antagonists. *J. Med. Chem.* 42, 3033–3040. doi: 10.1021/jm970832g

Dixit, P. D., Jain, A., Stock, G., and Dill, K. A. (2015). Inferring transition rates of networks from populations in continuous-time Markov processes. *J. Chem. Theory Comput.* 11, 5464–5472. doi: 10.1021/acs.jctc.5b00537

Ge, Y., Kier, B. L., Andersen, N. H., and Voelz, V. A. (2017). Computational and experimental evaluation of designed β-cap hairpins using molecular simulations and kinetic network models. *J. Chem. Inform. Model.* 57, 1609–1620. doi: 10.1021/acs.jcim.7b00132

Ge, Y., and Voelz, V. A. (2018). Model selection using BICePs: a Bayesian approach for force field validation and parameterization. *J. Phys. Chem. B* 122, 5610–5622. doi: 10.1021/acs.jpcb.7b11871

Ge, Y., Zhang, S., Erdelyi, M., and Voelz, V. (2020). Solution-state preorganization of cyclic-hairpin ligands determines binding mechanism and affinities for MDM2. *ChemRxiv*. doi: 10.26434/chemrxiv.13500765.v1

Ghosh, K., Dixit, P. D., Agozzino, L., and Dill, K. A. (2020). The maximum caliber variational principle for nonequilibria. *Annu. Rev. Phys. Chem.* 71, 213–238. doi: 10.1146/annurev-physchem-071119-040206

Gimenez, D., Zhou, G., Hurley, M. F., Aguilar, J. A., Voelz, V. A., and Cobb, S. L. (2019). Fluorinated aromatic monomers as building blocks to control α-peptoid conformation and structure. *J. Am. Chem. Soc.* 141, 3430–3434. doi: 10.1021/jacs.8b13498

Habeck, M., Nilges, M., and Rieping, W. (2005). Replica-exchange monte carlo scheme for Bayesian data analysis. *Phys. Rev. Lett.* 94:018105. doi: 10.1103/PhysRevLett.94.018105

Habeck, M., Rieping, W., and Nilges, M. (2006). Weighting of experimental evidence in macromolecular structure determination. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1756–1761. doi: 10.1073/pnas.0506412103

Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., Andreetta, C., et al. (2010). Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS ONE* 5:e13714. doi: 10.1371/journal.pone.0013714

Hummer, G., and Köfinger, J. (2015). Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* 143:12B634_1. doi: 10.1063/1.4937786

Hurley, M. F. D., Northrup, J. D., Ge, Y., Schafmeister, C. E., and Voelz, V. A. (2021). Metal cation-binding mechanisms of q-proline peptoid macrocycles in solution. *ChemRxiv*. doi: 10.26434/chemrxiv.13567853.v1

Kang, B., Yang, W., Lee, S., Mukherjee, S., Forstater, J., Kim, H., et al. (2017). Precisely tuneable energy transfer system using peptoid helix-based molecular scaffold. *Sci. Rep.* 7:4786. doi: 10.1038/s41598-017-04727-0

Li, D. W., and Brüschweiler, R. (2010). NMR-based protein potentials. *Angew. Chem. Int. Ed.* 49, 6778–6780. doi: 10.1002/anie.201001898

Liang, H., Zhou, G., Ge, Y., D'Ambrosio, E. A., Eidem, T. M., Blanchard, C., et al. (2018). Elucidating the inhibition of peptidoglycan biosynthesis in staphylococcus aureus by albocycline, a macrolactone isolated from streptomyces maizeus. *Bioorg. Med. Chem.* 26, 3453–3460. doi: 10.1016/j.bmc.2018.05.017

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

Löhr, T., Camilloni, C., Bonomi, M., and Vendruscolo, M. (2019). "A practical guide to the simultaneous determination of protein structure and dynamics using metainference," in *Biomolecular Simulations*, eds Bonomi M., and Camilloni, C. (Humana, New York, NY: Springer), 313–340. doi: 10.1007/978-1-4939-9608-7_13

Mas-Moruno, C., Rechenmacher, F., and Kessler, H. (2010). Cilengitide: the first anti-angiogenic small molecule drug candidate. design, synthesis and clinical evaluation. *Anti-Cancer Agents Med. Chem.* 10, 753–768. doi: 10.2174/187152010794728639

Mukherjee, S., Zhou, G., Michel, C., and Voelz, V. A. (2015). Insights into peptoid helix folding cooperativity from an improved backbone potential. *J. Phys. Chem. B* 119, 15407–15417. doi: 10.1021/acs.jpcb.5b09625

Northrup, J. D., Wiener, J., Hurley, M. F. D., Hou, C.-F. D., Baxter, R. H. G., Zdilla, M. J., et al. (2020). Metal-binding q-proline macrocycles. *ChemRxiv*. doi: 10.26434/chemrxiv.13554731

Olsson, S., Boomsma, W., Frellsen, J., Bottaro, S., Harder, T., Ferkinghoff-Borg, J., et al. (2011). Generative probabilistic models extend the scope of inferential structure determination. *J. Magn. Reson.* 213, 182–186. doi: 10.1016/j.jmr.2011.08.039

Olsson, S., Frellsen, J., Boomsma, W., Mardia, K. V., and Hamelryck, T. (2013). Inference of structure ensembles of flexible biomolecules from sparse, averaged data. *PLoS ONE* 8:e79439. doi: 10.1371/journal.pone.0079439

Orioli, S., Larsen, A. H., Bottaro, S., and Lindorff-Larsen, K. (2020). How to learn from inconsistencies: integrating molecular simulations with experimental data. *Prog. Mol. Biol. Transl. Sci.* 170, 123–176. doi: 10.1016/bs.pmbts.2019.12.006

Pitera, J. W., and Chodera, J. D. (2012). On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* 8, 3445–3451. doi: 10.1021/ct300112v

Potrzebowski, W., Trewhella, J., and Andre, I. (2018). Bayesian inference of protein conformational ensembles from limited structural data. *PLoS Comput. Biol.* 14:e1006641. doi: 10.1371/journal.pcbi.1006641

Prinz, J. H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., et al. (2011). Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* 134:174105. doi: 10.1063/1.3565032

Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science* 309, 303–306. doi: 10.1126/science.1110428

Roux, B., and Weare, J. (2013). On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* 138:02B616. doi: 10.1063/1.4792208

Schneider, J. A., Craven, T. W., Kasper, A. C., Yun, C., Haugbro, M., Briggs, E. M., et al. (2018). Design of peptoid-peptide macrocycles to inhibit the β-catenin tcf interaction in prostate cancer. *Nat. Commun.* 9:4396. doi: 10.1038/s41467-018-06845-3

Shevchuk, R., and Hub, J. S. (2017). Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics. *PLoS Comput. Biol.* 13:e1005800. doi: 10.1371/journal.pcbi.1005800

Shirts, M. R., and Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* 129, 124105–11. doi: 10.1063/1.2978177

Sun, J., and Zuckermann, R. N. (2013). Peptoid polymers: a highly designable bioinspired material. *ACS Nano* 7, 4715–4732. doi: 10.1021/nn4015714

Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M. (2003). Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. *J. Am. Chem. Soc.* 125, 15686–15687. doi: 10.1021/ja036523z

Voelz, V. A., Dill, K. A., and Chorny, I. (2011). Peptoid conformational free energy landscapes from implicit-solvent molecular simulations in amber. *Peptide Sci.* 96, 639–650. doi: 10.1002/bip.21575

Voelz, V. A., and Zhou, G. (2014). Bayesian inference of conformational state populations from computational models and sparse experimental observables. *J. Comput. Chem.* 35, 2215–2224. doi: 10.1002/jcc.23738

Wakefield, A. E., Wuest, W. M., and Voelz, V. A. (2015). Molecular simulation of conformational pre-organization in cyclic RGD peptides. *J. Chem. Inform. Model.* 55, 806–813. doi: 10.1021/ci500768u

Wan, H., Ge, Y., Razavi, A., and Voelz, V. (2020). Reconciling simulated ensembles of apomyoglobin with experimental hydrogen/deuterium exchange data using Bayesian inference and multiensemble Markov state models. *J. Chem. Theory Comput.* 16, 1333–1348. doi: 10.1021/acs.jctc.9b01240

Wan, H., Zhou, G., and Voelz, V. A. (2016). A maximum-caliber approach to predicting perturbed folding kinetics due to mutations. *J. Chem. Theory Comput*. 12, 5768–5776. doi: 10.1021/acs.jctc.6b00938

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

Wu, H., Paul, F., Wehmeyer, C., and Noé, F. (2016). Multiensemble markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci. U.S.A*. 113, E3221–E3230. doi: 10.1073/pnas.1525092113

Xu, H. (2019). Molecular simulations minimally restrained by experimental data. *J. Chem. Phys*. 150:154121. doi: 10.1063/1.5089924

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: 21 May 2021
doi: 10.3389/fmolb.2021.658269

Bayesian Random Tomography of Particle Systems

Nima Vakili[1] *and Michael Habeck*[1,2]*

[1]Microscopic Image Analysis Group, Jena University Hospital, Jena, Germany, [2]Statistical Inverse Problems in Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Random tomography is a common problem in imaging science and refers to the task of reconstructing a three-dimensional volume from two-dimensional projection images acquired in unknown random directions. We present a Bayesian approach to random tomography. At the center of our approach is a meshless representation of the unknown volume as a mixture of spherical Gaussians. Each Gaussian can be interpreted as a particle such that the unknown volume is represented by a particle cloud. The particle representation allows us to speed up the computation of projection images and to represent a large variety of structures accurately and efficiently. We develop Markov chain Monte Carlo algorithms to infer the particle positions as well as the unknown orientations. Posterior sampling is challenging due to the high dimensionality and multimodality of the posterior distribution. We tackle these challenges by using Hamiltonian Monte Carlo and a global rotational sampling strategy. We test the approach on various simulated and real datasets.

Keywords: 3D Reconstruction, random tomography, cryo-EM, bayesian inference, coarse-grained modeling, markov chain Monte Carlo, inferential structure determination

OPEN ACCESS

Edited by:
Edina Rosta,
King's College London,
United Kingdom

Reviewed by:
Takanori Nakane,
MRC Laboratory of Molecular Biology (LMB), United Kingdom
Slavica Jonic,
UMR7590 Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie (IMPMC), France

***Correspondence:**
Michael Habeck
michael.habeck@uni-jena.de

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

Received: *25 January 2021*
Accepted: *26 April 2021*
Published: *21 May 2021*

Citation:
Vakili N and Habeck M (2021) Bayesian Random Tomography of Particle Systems. Front. Mol. Biosci. 8:658269. doi: 10.3389/fmolb.2021.658269

1 INTRODUCTION

Many different imaging techniques acquire two-dimensional (2D) projection data of an unknown three-dimensional (3D) object. If the projection directions are known, tomographic reconstruction methods can be used to recover the 3D structure of the object (Natterer, 2001). An additional complication arises, if the projection directions are unknown. This imaging modality is of particular relevance to single-particle cryo-electron microscopy (cryo-EM). In recent years, cryo-EM has emerged as a powerful technique to determine the structure of large biomolecular assemblies at near atomic resolution (Frank, 2006). In cryo-EM, many copies of the particle of interest are first applied to a carbon grid and then plunge-frozen to prevent the formation of ice crystals. The frozen randomly orientated particles are imaged with electrons resulting in thousands to millions of noisy projection images. Similar reconstruction problems arise in cryo-electron tomography as well as single-particle diffraction experiments at free-electron lasers (von Ardenne et al., 2018). A completely different field of application is *in situ* microscopy of various specimens such as mesoscopic organisms (Levis et al., 2018).

The reconstruction problem common to all of these imaging methods is to recover a 3D volume from 2D images acquired in random projection directions and has been termed random tomography (Panaretos, 2009). Since the projection directions are unknown, we have to estimate them in the course of the reconstruction. Moreover, to avoid model bias, the desired reconstruction method should not rely on an initial guess of the volume (ab initio reconstruction).

Various ab initio reconstruction methods have been proposed (Bendory et al., 2020) including maximum likelihood via expectation maximization (Scheres et al., 2007) and maximum a posteriori (MAP) estimation (Jaitly et al., 2010; Scheres, 2010, 2012a), regularized maximum likelihood (Scheres, 2012b), stochastic gradient descent (Punjani et al., 2017), common lines (Vainshtein and Goncharov, 1986; Van Heel, 1987; Penczek et al., 1996; Elmlund et al., 2008; Singer and Shkolnisky, 2011; Elmlund and Elmlund, 2012; Lyumkis et al., 2013), the method of moments (Kam, 1980; Levin et al., 2018), random-model methods (Yan et al., 2007; Sanz-Garcia et al., 2010), methods using stochastic hill climbing (Elmlund et al., 2013) or nonlinear dimensionality reduction (Vargas et al., 2014) and frequency marching (Barnett et al., 2017).

These approaches typically reconstruct the unknown volume by solving an optimization problem. However, optimization approaches do not offer any uncertainty quantification. Another drawback is that many reconstruction algorithms are iterative procedures that critically depend on the initialization, which counteracts the idea of achieving an unbiased ab initio reconstruction. Moreover, most algorithms employ a number of ad hoc parameters that need to be tuned by the user and impact the final result in a way that is not always obvious.

Our goal is to develop a fully Bayesian approach to 3D reconstruction using a meaningful model of the unknown structure (including a physically realistic prior) and utilizing sampling algorithms for parameter estimation and uncertainty quantification. In our previous work (Joubert and Habeck, 2015), we already took the first step towards this goal. We considered the reconstruction problem in random tomography as a density estimation problem utilizing a mixture of Gaussians. With the help of conjugate priors and the introduction of latent assignment variables, we could derive analytical updates for a Gibbs sampler that infers the unknown rotations and component means.

However, there are various problems with our previous Gibbs sampling approach. First, Gibbs sampling suffers from slow convergence and depends strongly on the initial conditions. Therefore, to locate the posterior mode many restarts of the Gibbs sampler from varying initial conditions are necessary. Second, our Gibbs sampling algorithm is restricted to a Poissonian likelihood. The Poisson model is limited in that it ignores the effect of the point spread function and correlations in the noise. Third, the prior over the component means (particle positions) is chosen to be a conjugate, zero-centered Gaussian distribution, which is not realistic for biomolecular structures, because it ignores excluded-volume effects.

Here, we overcome these limitations by developing a more general probabilistic model for particle systems and their projection images. We no longer aim to develop analytical updates for the Gibbs sampler, but use of Markov chain Monte Carlo (MCMC) algorithms to infer both the particle positions as well as the unknown rotations. Sampling conformations of the particle system for fixed rotations can be achieved with Hamiltonian Monte Carlo (HMC). To sample the rotations, we use a Metropolis-Hastings algorithm that explores the unit quaternions parameterizing the unknown projection directions. Since Metropolis-Hastings samples a probability distribution only locally, we occasionally run a global sampling step that is computationally more expensive. Using simulated and real experimental data, we demonstrate that our Bayesian approach to random tomography is capable of estimating physically plausible coarse-grained models.

2 PROBABILISTIC MODEL AND POSTERIOR SAMPLING

We aim to reconstruct a 3D volume $f(\boldsymbol{r})$ for $\boldsymbol{r} \in \mathbb{R}^3$ and $f : \mathbb{R}^3 \mapsto \mathbb{R}_+$. We do not observe $f(\boldsymbol{r})$ directly but only projection images

$$g(\boldsymbol{u}) = \int f(\boldsymbol{R}^T \boldsymbol{r})\, \mathrm{d}z = \int f(\boldsymbol{\theta}^\perp \boldsymbol{u} + \boldsymbol{\theta} z)\, \mathrm{d}z =: \mathcal{X}_{\boldsymbol{\theta}}[f](\boldsymbol{u}) \tag{1}$$

where $\boldsymbol{R} \in SO(3)$ is a 3D rotation matrix whose last row $\boldsymbol{\theta} \in \mathbb{R}^3$ is a unit vector pointing into the projection direction, and $\boldsymbol{\theta}^\perp \in \mathbb{R}^{3\times 2}$ is the matrix whose columns span the plane orthogonal to $\boldsymbol{\theta}$ such that $\boldsymbol{R}^T = [\boldsymbol{\theta}^\perp, \boldsymbol{\theta}]$. Throughout this article, $\boldsymbol{u} \in \mathbb{R}^2$ denotes a position in the projection image, and $\boldsymbol{r} \in \mathbb{R}^3$ a position in the volume. The integral transform $\mathcal{X}_{\boldsymbol{R}}[f]$ (**Eq. 1**) is known as the X-ray transform or John transform (Natterer, 2001). In 2D, the X-ray transform is identical to the Radon transform. The reconstruction problem in random tomography is to estimate $f(\boldsymbol{r})$ from N random projection directions $\boldsymbol{\theta}_n$, or equivalently $\boldsymbol{R}_n$, such that

$$g_n(\boldsymbol{u}) = \mathcal{X}_{\boldsymbol{\theta}_n}[f](\boldsymbol{u}) + n(\boldsymbol{u}), \quad n = 1, \ldots, N \tag{2}$$

where $n(\boldsymbol{u})$ is the noise.

2.1 Kernel Expansion of Images and Volumes

The standard discretization of images and volumes is based on pixels and voxels placed on regular 2D and 3D grids. Instead, we expand images and volumes into sums of basis functions that can be centered at irregular positions (as in meshless methods). We use a radial basis function (RBF) kernel ϕ such that the kernel expansion of the volume becomes

$$f(\boldsymbol{r}) = \sum_{k=1}^{K} w_k\, \boldsymbol{\phi}(\boldsymbol{r} - \boldsymbol{x}_k) \tag{3}$$

where K is the number of basis functions, $\|\cdot\|$ is the Euclidean norm, w_k a coefficient or weight (if $w_k > 0$) and $\boldsymbol{x}_k \in \mathbb{R}^3$ a position vector that determines the center of the kth kernel. We can represent members of a reproducing kernel Hilbert space using this expansion. RBF representations are widely used in machine learning (Schölkopf and Smola, 2002), image processing (Takeda et al., 2007) and numerical applications (Schaback and Wendland, 2006).

A physical interpretation of the kernel representation is that we model the object as a collection of K particles at positions $\boldsymbol{x}_k$ with mass $w_k > 0$. The model **(3)** can then be interpreted as the blurred version of a particle system:

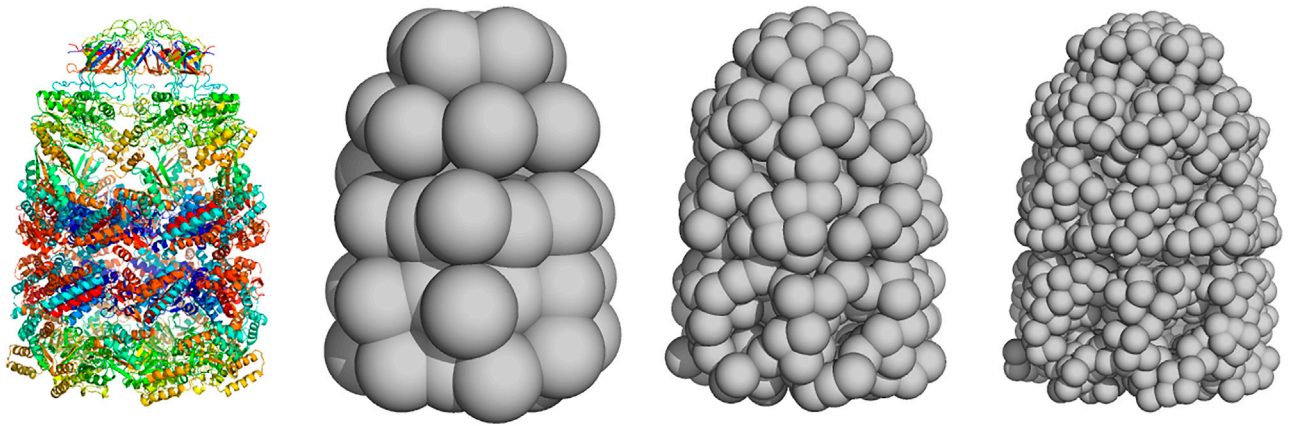


FIGURE 1 | Coarse-grained representation of GroEL/GroES using a varying number of particles **(left)** atomic structure (PDB code 1aon). Panels **(middle left)** to **(right)** show coarse-grained models using $K = 50, 300$, and 1000 particles.

$$f(\boldsymbol{r}) = \left(\phi * \sum_{k=1}^{K} w_k \delta_{\boldsymbol{x}_k} \right)(\boldsymbol{r}) \tag{4}$$

where $\delta_{\boldsymbol{x}_k}$ is the delta function centered a $\boldsymbol{x}_k$ and the particle density, $\sum_k w_k \delta_{\boldsymbol{x}_k}$, is blurred by a convolution (denoted by $*$) with the RBF kernel. The particle locations and weights $\{(\boldsymbol{x}_k, w_k); k = 1, \ldots, K\}$ can also be viewed as a weighted point cloud. The component means $\boldsymbol{x}_k$ could be fixed to a regular 3D grid. But we will consider particle systems that are not tied to a grid and can be distributed in an irregular fashion (similar to meshless or meshfree methods used in numerical analysis). Typically, the particle system is a coarse-grained representation of the unknown structure rather than an atomic-resolution representation. Therefore, 3D reconstruction from 2D projection data provides a pseudo-atomic representation whose resolution depends on the number of particles K (**Figure 1** for an illustration).

One motivation for our choice of the volume representation (**Eq. 3**) are its efficient transformation properties. Rigid transformations of $f(\boldsymbol{r})$ involve a shift by the translation vector $\boldsymbol{t}$ and a reorientation brought about by the rotation matrix $\boldsymbol{R}$. Under the RBF expansion these transformations reduce to rigid transformations of the particle positions:

$$f(\boldsymbol{r}) \overset{R,t}{\mapsto} f(\boldsymbol{R}^T(\boldsymbol{r} - \boldsymbol{t})) = \sum_k w_k \phi(\boldsymbol{r} - \boldsymbol{R}\boldsymbol{x}_k - \boldsymbol{t}) = \sum_k w_k \phi(\boldsymbol{r} - \boldsymbol{x}'_k) \tag{5}$$

where $\boldsymbol{x}'_k = \boldsymbol{R}\boldsymbol{x}_k + \boldsymbol{t}$.

There are many options for $\phi(r)$. We will restrict ourselves to Gaussian RBF kernels. The d-dimensional spherical Gaussian is defined by

$$\phi_d(\boldsymbol{r}; \boldsymbol{x}, \sigma^2) := \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{ -\frac{1}{2\sigma^2} \|\boldsymbol{r} - \boldsymbol{x}\|^2 \right\} \tag{6}$$

where $\sigma > 0$ is the bandwidth of the kernel. The volume representation that we will use throughout this paper is a mixture of K spherical Gaussians:

$$f(\boldsymbol{r}) = \sum_{k=1}^{K} w_k \phi_3(\boldsymbol{r}; \boldsymbol{x}_k, \sigma^2) \tag{7}$$

This representation is very common in statistics, in particular in density estimation where $\boldsymbol{x}_k$ are observed samples resulting in a kernel density estimate of an unknown probability density function. Indeed, our original motivation (Joubert and Habeck, 2015) to choose this representation of $f(\boldsymbol{r})$ was mainly driven by viewing 3D reconstruction from random projections as an instance of a density estimation problem. Other examples for uses of (Gaussian) particle representations in cryo-EM data analysis such as denoising or the analysis of continuous conformational changes have been proposed by Jin et al. (2014); Jonić et al. (2016); Jonić and Sorzano (2016).

A convenient property of the spherical Gaussian kernel is its behavior under the X-ray transform (**Eq. 1**):

$$\mathcal{X}_{\boldsymbol{\theta}}[\phi_d](\boldsymbol{u}) = \int \phi_d(\boldsymbol{\theta}^\perp \boldsymbol{u} + \boldsymbol{\theta} z; \boldsymbol{x}, \sigma^2)\, \mathrm{d}z = \phi_{d-1}(\boldsymbol{u}; \boldsymbol{P}\boldsymbol{R}\boldsymbol{x}, \sigma^2) \tag{8}$$

where again $\boldsymbol{R} = [\boldsymbol{\theta}^\perp, \boldsymbol{\theta}]^T \in SO(3)$ and the 2×3 projection matrix $\boldsymbol{P}$ is

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{9}$$

Spherical Gaussians are closed under the X-ray transform, and the projected volume **(7)** is again a K component mixture of spherical Gaussians

$$\mathcal{X}_{\boldsymbol{\theta}}[f](\boldsymbol{u}) = \sum_{k=1}^{K} w_k \phi_2(\boldsymbol{u}; \boldsymbol{P}\boldsymbol{R}\boldsymbol{x}_k, \sigma^2) \tag{10}$$

with centers $\boldsymbol{x}'_k = \boldsymbol{P}\boldsymbol{R}\boldsymbol{x}_k \in \mathbb{R}^2$. This fact motivates us to also represent the input images as mixtures of spherical Gaussians in 2D (see ***Representation of Projection Images by Point Clouds*** for a concrete application).

2.2 Probabilistic Model

The unknown parameters of our model are the particle positions $\boldsymbol{x}_k$ and weights w_k as well as the unknown rotation matrices $\boldsymbol{R}_n$. Since we interpret the Gaussian components as particles of equal mass, we fix the weights: $w_k = K^{-1}$, such that the main inference parameters are $\boldsymbol{x}_k$ and $\boldsymbol{R}_n$.

2.2.1 Likelihoods

We tested two probabilistic models for the input data. The first model uses the input images $\{g_n; n = 1, \ldots, N\}$ directly. For each image, the intensities are $g_{nm} = g_n(\boldsymbol{u}_{nm})$ at pixel positions $\boldsymbol{u}_{nm}$ where $m = 1, \ldots, M_n$ with M_n being the number of pixels in the nth image. Typically, the number of pixels M_n is identical for all projection images.

A simple image model is to assume pixelwise identically and independently distributed Gaussian noise in the image formation **(2)**, such that the likelihood of the nth image is

$$\Pr\left(g_n \middle| \boldsymbol{x}, \boldsymbol{R}_n, \boldsymbol{t}_n, \gamma_n, \alpha_n, \tau_n\right) = \left(\frac{\tau_n}{2\pi}\right)^{M_n/2} \exp\left\{ -\frac{\tau_n}{2} \sum_{m=1}^{M_n} \left[g_{nm} - \alpha_n - \gamma_n \sum_k \phi_2\left(\boldsymbol{u}_{nm}; \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n, \sigma^2\right)\right]^2 \right\} \tag{11}$$

where $\tau_n > 0$ is the precision of the image, and α_n, γ_n are an offset and a scaling factor (the constant weight $w_k = 1/K$ has been absorbed by the scaling factor γ_n). The two-dimensional translation $\boldsymbol{t}_n$ accounts for a shift of the image. These three to five nuisance parameters per image (depending on whether shifts $\boldsymbol{t}_n$ are fitted or not) need to be estimated in addition to the particle positions $\boldsymbol{x} = \{\boldsymbol{x}_k; k = 1, \ldots, K\}$ and the rotations $\boldsymbol{R} = \{\boldsymbol{R}_n; n = 1, \ldots, N\}$. Model **(11)** is an idealized image formation model. It ignores important effects such as the CTF or correlated noise that are highly relevant for cryo-EM applications.

The second model also uses a kernel expansion of the input image motivated by the fact that ideally, according to our image model, the projection image should also be a mixture of spherical Gaussians (**Eq. 10**). In a preprocessing step, we fit a point cloud $Y_n = \{\boldsymbol{y}_{nm} \in \mathbb{R}^2; m = 1, \ldots, M_n\}$ to the nth input image g_n such that

$$g_n(\boldsymbol{u}) \approx \alpha_n + \gamma_n \sum_{m=1}^{M_n} \phi_2\left(\boldsymbol{u}; y_{nm}, \sigma_n^2\right) \tag{12}$$

Typically, we choose $M_n = M$ but this is not a requirement. Again, model **(12)** does not account for the CTF or other important effects in cryo-EM image formation. In each projection direction, the 2D point cloud can be blurred to a different degree captured by the width σ_n. The Supplementary Material details how projection images can be converted to point clouds; ***Representation of projection images by point clouds*** in Results shows a practical example for further illustration.

As in Joubert and Habeck (2015), we model the 2D point clouds as samples from the projected 3D volume:

$$\Pr\left(Y_n \middle| \boldsymbol{x}, \boldsymbol{R}_n, \boldsymbol{t}_n, \sigma_n\right) = \prod_{m=1}^{M_n} \frac{1}{K} \sum_{k=1}^{K} \phi_2\left(\boldsymbol{y}_{nm}; \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n, \sigma_n^2\right) \tag{13}$$

In the following, we will denote all nuisance parameters, i.e. all parameters except particle positions and rotations, collectively by ξ. In case of the image likelihood **(11)**, we have $\xi = \{(\alpha_n, \gamma_n, \tau_n, \boldsymbol{t}_n); n = 1, \ldots, N\}$. In case of the point cloud likelihood (**Eq. 13**), we have $\xi = \{(\sigma_n, \boldsymbol{t}_n); n = 1, \ldots, N\}$. Moreover, we will denote both likelihoods as $\Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi)$ where D are the data (projection images or 2D point clouds).

2.2.2 Priors

After incorporating our prior beliefs about the model parameters, we are able to derive the posterior distribution by invoking Bayes' theorem:

$$\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi|D) = \frac{\Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi)\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi)}{\Pr(D)} \tag{14}$$

where $\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi)$ is the prior which we assume to factor into

$$\Pr(\boldsymbol{x}, \boldsymbol{R}, \xi) = \Pr(\boldsymbol{x})\Pr(\boldsymbol{R})\Pr(\xi) \tag{15}$$

The normalization factor $\Pr(D)$ is the model evidence, which can be ignored if we are only interested in parameter estimation.

We use standard priors for the nuisance parameters: Jeffreys priors for precisions τ_n and $1/\sigma_n^2$. The prior for the scaling factors and offsets are flat. Note that these priors are improper (i.e., not normalizable). Since we are only interested in parameter estimation, this does not pose a problem. The priors for the scaling factor and offset could be improved. For example, cryo-EM images are often normalized such that the mean intensity is zero and the standard deviation is one. It is possible to express this information as a prior on the offset and scaling factor. The Supplementary Material provides more details about these priors. For the image shifts $\boldsymbol{t}_n$, a zero-centered two-dimensional Gaussian distribution is a reasonable choice.

Typically, biomolecules orient themselves randomly in the ice layer that is imaged by cryo-EM. Therefore, we choose a uniform distribution over $SO(3)$:

$$\Pr(\boldsymbol{R}) = \prod_{n=1}^{N} \Pr(\boldsymbol{R}_n) \propto 1 \tag{16}$$

These priors are proper, because the rotation group is compact.

In our previous work (Joubert and Habeck, 2015), we used a zero-centered Gaussian prior for all particle positions $\boldsymbol{x}_k$ to ensure that prior and likelihood are conjugate, which enabled the derivation of closed-form updates for the component means. However, this prior is very unrealistic, if we think of the Gaussian basis functions as massive particles that should not occupy the same region in space (excluded volume), but rather repel each other. Since the packing of biomolecular structures is reminiscent of fluids (Liang and Dill, 2001), the prior should favor particle configurations that show similar packing characteristics. To model repulsive interactions between particles, we use a

Boltzmann distribution over the positions $\boldsymbol{x}_k$ involving a soft repulsive interaction potential $E(\boldsymbol{x})$:

$$\Pr(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) \propto \exp\{-\beta E(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)\} \tag{17}$$

Furthermore, the particles are confined to a box with soft boundaries (Habeck, 2017). Pairs of particles repel each other if the distance is smaller than the particle diameter $2R$ where R is the effective particle radius. We choose a quartic repulsion which is commonly used in NMR structure calculation:

$$E(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) = \sum_{k<k'} \left[\left\|\boldsymbol{x}_k - \boldsymbol{x}_{k'}\right\| \le 2R\right] \left(1 - \frac{\left\|\boldsymbol{x}_k - \boldsymbol{x}_{k'}\right\|}{2R}\right)^4 \tag{18}$$

where $[\cdot]$ is the Iverson bracket. Given the total number of atoms L of the system, the particle radius can be predicted for a desired number of particles K by using the relation

$$R \approx 0.92\ (L/K)^{0.42}\ \text{Å}. \tag{19}$$

Using a configurational temperature estimator (Mechelke and Habeck, 2013), the inverse temperature is estimated to $\beta \approx 175$. The estimates for R and β are based on an analysis of several biomolecular structures at different levels of coarse graining. See Supplementary Material for details.

Since the excluded-volume term (**Eq. 18**) is purely repulsive, we add a radius of gyration term such that the overall prior for particle positions is

$$\Pr(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) \propto \exp\{-\beta E(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)\} \exp\{-\alpha R_g(\boldsymbol{x})\} \tag{20}$$

where $R_g(\boldsymbol{x})$ is the radius of gyration of the coarse-grained structure $\boldsymbol{x}$ and α a positive constant. The radius of gyration term imposes a weak preference for compact structures and prevents configurations with isolated particles that do not contact another particle. In our experiments, we set $\alpha = 10$ Å; in principle, we could estimate α by using techniques similar to those used in the estimation of β. But since α does not have a strong impact on the final structure, we restricted ourselves to a single fixed value for α.

2.3 Inference

Bayesian random tomography employs MCMC sampling from the posterior distribution **(14)**. We use a Gibbs sampling strategy (Geman and Geman, 1984) where each group of parameters, the particle positions $\boldsymbol{x}$, the rotations $\boldsymbol{R}$ and the nuisance parameters ξ, is updated separately while clamping the other parameters to their current values. To update the nuisance parameters, we use standard samplers for generating Gamma variates and normally distributed random variables (more details can be found in the Supplementary Material). However, the conditional posteriors of the particle positions $\boldsymbol{x}$ and the rotations $\boldsymbol{R}$ are not of a standard form and need to be updated with more sophisticated algorithms.

2.3.1 Sampling Particle Positions With Hamiltonian Monte Carlo

To sample the particle positions, we use Hamiltonian Monte Carlo (HMC) (Neal, 2011). The conditional posterior distribution over particle positions is

$$\Pr(\boldsymbol{x}|\boldsymbol{R}, \xi, D) \propto \Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi) \Pr(\boldsymbol{x})$$

In HMC, $-\log\Pr(\boldsymbol{x}|\boldsymbol{R}, \xi, D)$ defines a potential energy over configuration space that is composed of an attractive term $-\log\Pr(D|\boldsymbol{x}, \boldsymbol{R}, \xi)$ matching particle positions to the projection data, and a repulsive contribution $-\log\Pr(\boldsymbol{x})$ stemming from the excluded-volume term **(18)**. For fixed rotations and nuisance parameters, the particle positions undergo Hamiltonian dynamics following the gradient of $-\Pr(\boldsymbol{x}|\boldsymbol{R}, \xi, D)$ during a short leapfrog integration. The resulting configuration is accepted or rejected according to the Metropolis criterion.

2.3.2 Sampling Rotational Parameters With Metropolis-Hastings

A challenging problem is to estimate the rotations. Because the projection images are statistically independent of each other, the problem decomposes into N subproblems:

$$\Pr(\boldsymbol{R}_n|\boldsymbol{x}, \xi, D) \propto \exp\left\{-\frac{\tau_n}{2} \sum_{m=1}^{M_n} \left[g_{nm} - \alpha_n - \gamma_n \sum_{k=1}^{K} \phi_2\left(\boldsymbol{u}_{nm}; \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n, \sigma^2\right)\right]^2\right\} \tag{21}$$

if projection images g_n are fitted directly, or

$$\Pr(\boldsymbol{R}_n|\boldsymbol{x}, \xi, D) \propto \prod_{m=1}^{M_n} \sum_{k=1}^{K} \phi_2\left(\boldsymbol{y}_{nm}; \boldsymbol{P}\boldsymbol{R}_n\boldsymbol{x}_k + \boldsymbol{t}_n, \sigma_n^2\right) \tag{22}$$

if we fit 2D point clouds. In Joubert and Habeck (2015), we introduced assignment variables such that the conditional posterior **(22)** is replaced by the matrix von Mises-Fisher distribution, which can be simulated in a straightforward fashion (Habeck, 2009). However, because the assignment variables are highly coupled to the other parameters, this strategy converges only slowly to the next local minimum. Moreover, there is no flexibility regarding the likelihood function.

We use the Metropolis-Hastings (MH) algorithm (Liu, 2001) to estimate the rotation matrices. We parameterize rotation matrices using unit quaternions (Horn, 1987) and propose new quaternions by adding a random perturbation that is sampled from a uniform distribution. We run 10 MH steps to update the quaternions representing each projection direction in every Gibbs sampling iteration and adapt the step-size automatically: Upon acceptance, the step-size increases by multiplying it with a factor of 1.02; in case of rejection, the step-sizes decreases by a factor of 0.98. This rule results in an acceptance rate of approximately 50%. We use this sampling algorithm to simulate both types of conditional posteriors **(21)** and **(22)**.

2.3.3 Global Sampling of Rotational Parameters

Since the MH algorithm achieves only local sampling of probability distributions, we occasionally scan all rotations systematically. The unit quaternions are elements of the 3-sphere, the unit sphere embedded in the four-dimensional

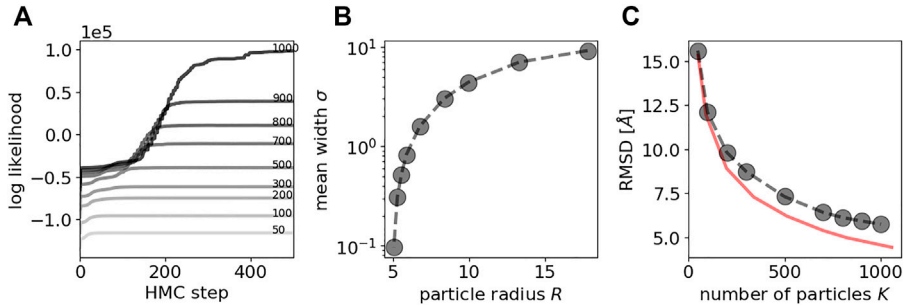


FIGURE 2 | HMC sampling of particle positions with fixed rotations for a simulated data set of GroEL/ES. A Evolution of the log likelihood during HMC sampling. The larger the number of particles K, the higher is the final log likelihood. Increasing darkness indicates larger number of particles. Line annotations also indicate the number of particles. B Average standard deviation (computed over all 35 input point clouds) vs. the size of the particle R. C RMSD between Carbon-alpha positions of the crystal structure and the coarse-grained models inferred with HMC. As a reference, the RMSD between the Carbon-alpha positions and the coarse-grained versions of the crystal structures is shown as red curve.

space. To evenly cover rotation space, we discretize the 3-sphere using the 600-cell (Coxeter, 1973). The 600-cell is composed of even sized tetrahedra whose corners lie on the unit sphere. By projecting the center of a tetrahedron onto the unit sphere we obtain a unit quaternion parameterizing a valid rotation matrix. Due to the degeneracy of the quaternions we only have to consider the upper half of the 4D sphere that is covered by 330 tetrahedra at the coarsest level of discretization. To obtain a finer tessellation of $SO(3)$, we can split each tetrahedron into eight tetrahedra whose corners again lie on the 4D unit sphere. By default, we use a frequency of 0.1 to run a global rotation scan. The conditional posterior is evaluated for all rotations and then sampled from the discrete distribution.

The source code and scripts for reproducing the tests are available at github.com/michaelhabeck/bayesian-random-tomography.

3 RESULTS

3.1 Sampling Tests

To test MCMC strategies for inferring particle positions and rotations, we use the structure of the GroEL/GroES complex. This system has been studied extensively with cryo-EM. Since our focus is mainly on algorithmic aspects, we first use simulated data that exactly follow our probabilistic model. To generate input point clouds in 2D, we use the crystal structure of GroEL/GroES (PDB code 1aon; 58,674 atom coordinates in total). The 2D point clouds are generated by projecting the 3D positions of every 10th Carbon-alpha atom (802 points in total) along 35 random directions into 2D. We also generated corresponding projection images by blurring the point clouds with a Gaussian filter of width 5 Å.

3.1.1 Sampling Particle Positions and Precisions With Fixed Rotations

We first studied the performance of sampling particle positions by fixing the rotations to the correct values and sampling only the particle positions and the precisions of the projection data. HMC sampling of particle positions started from a random initial configuration for K ranging between 50 and 1,000 particles. In all of our HMC experiments, the number of leapfrog steps was set to 10, whereas the step-size was adjusted automatically. The precisions $1/\sigma_n^2$ follow Gamma distributions and can be sampled directly.

Figure 2A shows the evolution of the log likelihood achieved by the particle system during HMC. After roughly 200 to 500 HMC steps (depending on K), the particle cloud reproduces the input data well, which is reflected in high values of the log likelihood. The sampled particle configurations are very similar to the true structure at the same level of coarse graining. Successful sampling of $\Pr(\boldsymbol{x}|\boldsymbol{R}, \xi, D)$ with HMC is observed reliably for many different initial particle configurations.

It is clear that an increasing number of particles K results in a higher goodness of fit, which is obvious from **Figures 2A,B** showing the average standard deviation σ_n of the point cloud likelihood (**Eq. 13**) as a function of particle radius: A higher number of particles K results in more flexible models that result in a better goodness of fit and higher precision. These findings indicate that HMC is highly suited to sample particle configurations.

Figure 2C shows the accuracy of the coarse-grained models inferred from the projection data with HMC. The accuracy is quantified by the root mean square deviation (RMSD) between corresponding positions in a reference structure and a coarse-grained model. Here, our reference structure is the atomic structure of GroEL/ES reduced to the positions of 8,015 Carbon-alpha atoms listed in the PDB entry 1aon. To compare this structure with a coarse-grained model, positions in the atomic structure are assigned to positions in the coarse-grained model that are closest in 3D space. There are two factors that contribute to this measure of accuracy: the level of coarse graining as well as the performance of posterior sampling based on the 2D projection data. To disentangle both contributions, we also show the accuracy between the crystal structure and its coarse-grained versions (obtained with the DP-means algorithm by Kulis and Jordan (2012); also see the Supplementary Material).

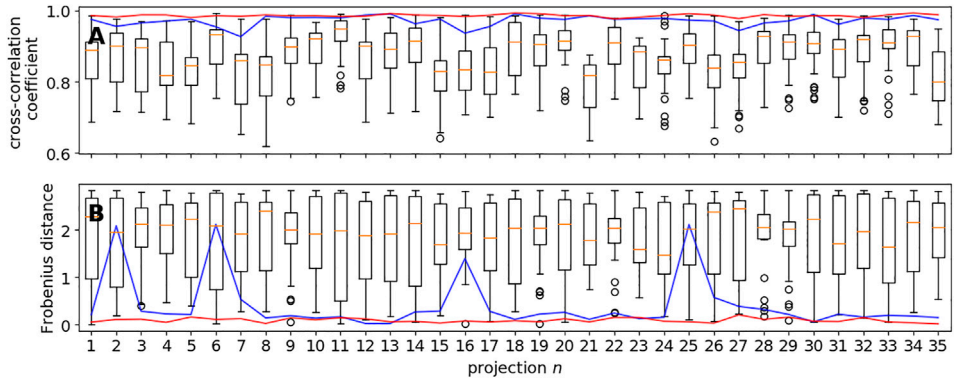


FIGURE 3 | Global vs. local sampling of orientational parameters. Shown are the cross-correlation coefficients **(panel A)** and Frobenius distances **(panel B)** for each of the 35 input directions achieved with local sampling based on the MH algorithm and global sampling using a regular discretization of the 3-hemisphere. The blue curve shows the results obtained with the coarsest covering based on 330 unit quaternions; the red curve shows the results obtained with a finer covering (2,460 quaternions). The box plots show the variability within 30 trials of MH starting from random rotations.

This curve shows that coarse-grained models of GroEL/ES using 1,000 particles achieve an accuracy of about 4.6 Å, whereas an ultra coarse-grained model based on only 50 particles is on average 15.5 Å away from any Carbon-alpha atom in the crystal structure. For very high levels of coarse graining (small K), the models inferred with HMC reach the maximum accuracy that is possible at this level of coarse graining. With increasing number of particles K, the gap in accuracy widens but is still similar to the maximum attainable value. For example, with K = 1000 the model obtained with HMC achieves an RMSD of 5.7 Å, whereas the coarse grained model obtained directly from the crystal structure achieves an accuracy of 4.6 Å.

If we estimate particle configurations from projection images instead of point clouds, we obtain similar results. **Supplementary Figure S4** shows the log likelihood and cross-correlation coefficients obtained with different numbers of particles, again ranging between 50 and 1,000. The evolution of the log likelihood indicates that the HMC sampler seems to converge even faster compared to a simulation based on point cloud data: within 20–150 HMC steps the log likelihood plateaus. The accuracy of the structure after 500 HMC steps is similar to or better than the accuracy of the particle models fitted against 2D point clouds and almost reaches the accuracy of the coarse-grained models derived from the crystal structure. **Supplementary Figure S5** shows FSC curves for all 3D models. For the same number of particles, the FSC curves are similar with a slight preference for the image-based models when using larger numbers of particles. The resolution ranges from 12.2 Å (50 particles) to 4.5 Å (1,000 particles). **Supplementary Table S1** shows resolution estimates for all models.

3.1.2 Sampling Rotational Parameters and Precisions With Fixed Particle Positions

To test our rotational sampling approach, we fixed the particle positions to an ultra coarse-grained structure (K = 200) of GroEL/ES. Although each rotation can be updated independently of the other rotations, and each conditional posterior (given either by **Eqs. 21** or **22**) is only a four-dimensional probability distribution over the quaternions, the sampling problem is still challenging due to its multimodality. Since Metropolis-Hastings (MH) is a local sampling algorithm, it tends to become trapped in subordinate modes of the conditional posterior, which are typical for rigid registration problems. As a result, running MH on the conditional posteriors is not sufficient to reliably recover the rotation matrices.

Figure 3A shows the cross-correlation coefficients for the 35 projection images obtained with global rotational sampling in comparison with MH runs starting from 30 random rotations. Global rotational sampling was based on the first two discretizations of the 3-hemisphere using 330 and 2,640 quaternions, respectively. The number of local sampling attempts was set to 30 so as to match the speed of global sampling at the finer level. That is, the coarse sampling based on 330 quaternions is approximately 8 times faster than the 30 local sampling trials. As evidenced by **Figure 3A**, global sampling is capable of finding rotation matrices that yield high cross-correlation coefficients, whereas MH alone fails to do so in a systematic fashion. **Figure 3B** shows the Frobenius distances (ranging from 0 to a maximum of $2\sqrt{2}$) between the true rotation matrix and the estimated rotation matrices. Again, global rotational sampling achieves more accurate rotations, whereas the distances scatter largely for the local MH trials. These findings suggest that global rotational sampling is indispensable for Bayesian random tomography in agreement with our previous findings (Joubert and Habeck, 2015) where we had to resort to repeated Gibbs sampling runs.

Before we study sampling of the full posterior distribution (all parameters $\boldsymbol{R}$, $\boldsymbol{x}$ and ξ are unknown), we will first outline how experimental projection images can be converted to 2D point clouds that are suitable for our approach to random tomography.

3.2 Representation of Projection Images by Point Clouds

Experimental projection data are typically presented as projection images rather than point clouds. In this subsection, we discuss

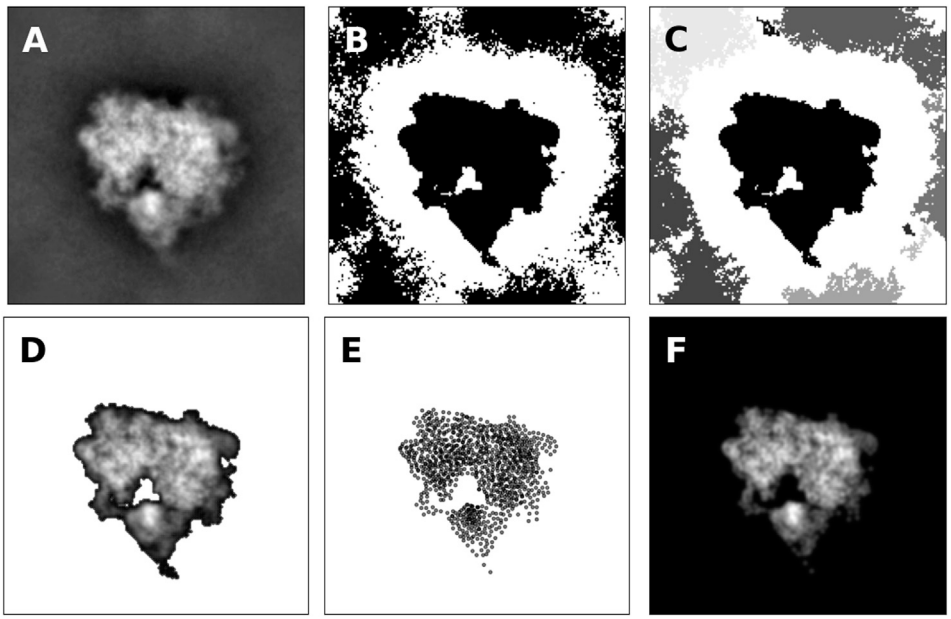


FIGURE 4 | Representation of projection images by 2D point clouds **(A)** Class average of the 80S ribosome **(B)** Mask obtained by thresholding image intensities greater than the median intensity. Black pixels are part of the mask **(C)** Clustering of pixels that are part of the mask. Pixels that form a connected component are grouped together and shown in different grayscale colors **(D)** Pixels that form the most central connected components with shifted image intensities **(E)** 2D point cloud composed of 1,000 particles obtained by running the Expectation Maximization algorithm **(F)** Model image according to **Eq. 10**. The cross-correlation coefficient between the model and the original image is 95.8%. If only pixels are considered that are part of the mask indicating the central connected component, the cross-correlation coefficient increases to 99.6%.

how to convert 2D projection images to 2D point clouds that are suitable for our Bayesian random tomography approach. We discuss this for a cryo-EM data set, but similar techniques are also applicable to other data, as we will demonstrate later.

The projection properties of mixtures of spherical Gaussians (**Eq. 10**) suggest to also represent the projection image as a mixture of Gaussians. Our model can only capture nonnegative intensities. Therefore, we first have to choose a suitable threshold θ above which image intensities are considered real signal. The threshold will be used to construct a binary mask: the intensities of pixels that are part of the mask will be shifted by θ such that their shifted intensities are nonnegative; the intensities of pixels that are not part of the mask will be set to zero (i.e., they will be ignored in the construction of the point cloud). A simple choice of θ for class averages from cryo-EM is the median intensity, but a different choice might be more suitable for other types of images.

An example of the thresholding procedure is shown in **Figure 4B** for a class average showing the projection of the 80S ribosome (shown in **Figure 4A**). Black pixels indicate pixels with intensity above the median. By looking at the mask, it is clear that only the central pixels forming a connected component carry signal.

Next, we identify pixels that form connected components. Again this applies to cryo-EM images; other types of images might require a different treatment to construct a suitable mask. To identify signal pixels that form a connected component, we convert the thresholded image to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where the pixels with intensities above the threshold are the vertices $\mathcal{V} = \{\boldsymbol{u}_m; g(\boldsymbol{u}_m) > \theta, m = 1, \ldots, M\}$. Edges are introduced between all pairs of pixels that are nearest neighbors on the 2D square lattice, i.e. their Euclidean distance is smaller than or equal to one pixel:

$$\mathcal{E} = \left\{ (i,j) \in \{1, \ldots, |\mathcal{V}|\}^2; \parallel \boldsymbol{u}_i - \boldsymbol{u}_j \parallel \leq 1 \right\}.$$

As shown in **Figure 4C**, multiple connected components are typically found in the masked pixels. Since cryo-EM class averages are often centered, we pick the connected component whose center of mass is closest to the image center. The selected pixels including their intensity (shifted by θ) are shown in **Figure 4D**.

To obtain a particle-based representation of the central connected component, we run the Expectation Maximization algorithm (details in Supplementary Material). **Figure 4E** shows the estimated point cloud using 1,000 particles. The estimated standard deviation of the Gaussian is 1.34 pixels. The density generated by the 2D particles is shown in **Figure 4** and correlates highly with the original image and the masked image. **Supplementary Figure S1** shows more examples of class averages represented as 2D point clouds.

3.3 3D Reconstruction by Sampling the Full Posterior Distribution

We applied Bayesian random tomography to three real datasets, two cryo-EM datasets and one dataset from stochastic microscopy experiments visualizing marine microorganisms.

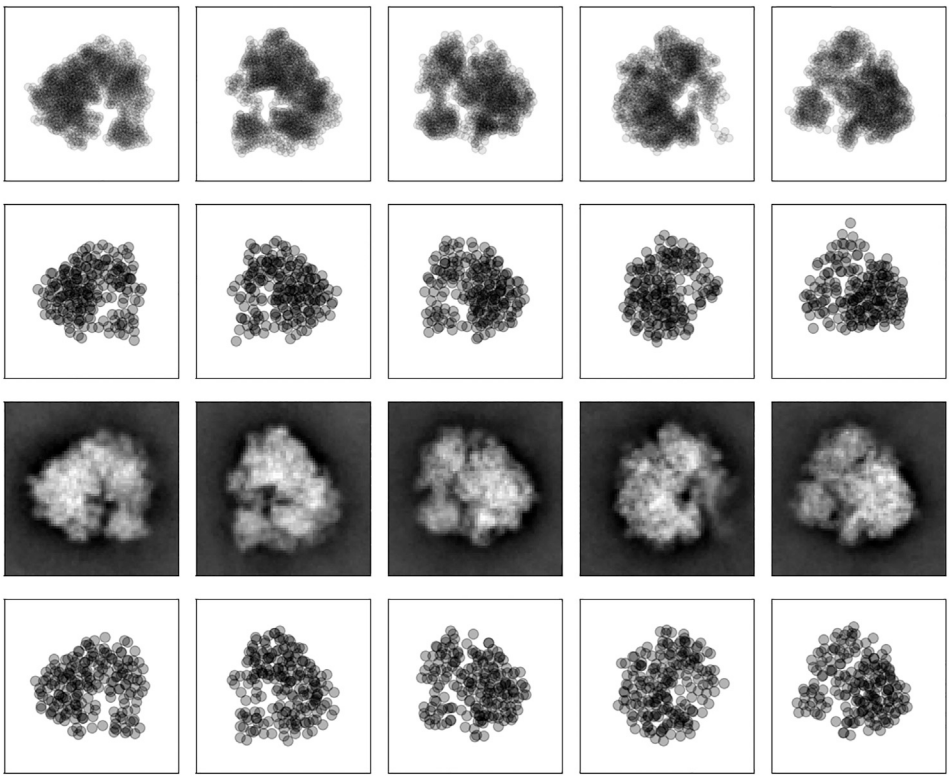


FIGURE 5 | 2D projections of the 80S ribosome. First row: point clouds derived from class averages. Each projection image is represented by 1,000 points. Second row: 2D projections of the coarse-grained model calculated with Bayesian random tomography based on 2D point clouds. Third row: Class averages. Bottom row: 2D projections of the coarse-grained model calculated with Bayesian random tomography based on class averages.

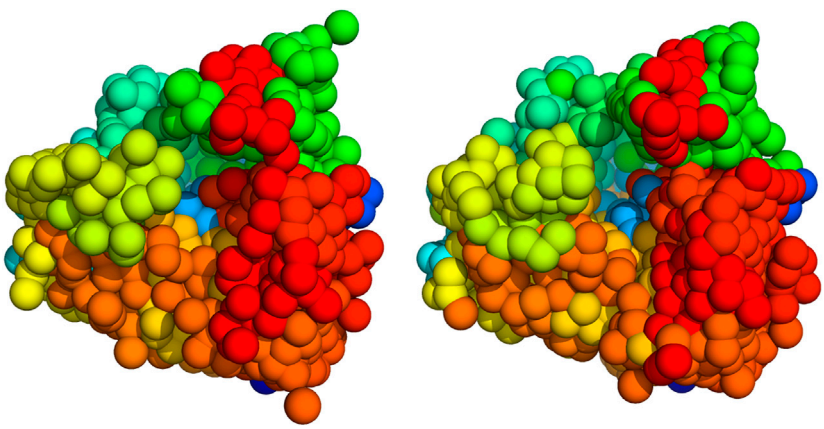


FIGURE 6 | 3D models of the 80S ribosome **(Left)** 1,000 particle model inferred with Bayesian random tomography **(Right)** Initial model computed with PRIME. The particles are sorted such that spatially close particles have similar indices. By using Pymol's chainbow command, we can then visualize the particle models such that substructures are better visible.

In these applications, we sampled the joint posterior distribution of all unknown parameters, particle positions $\boldsymbol{x}_k$, rotations $\boldsymbol{R}_n$ and nuisance parameters ξ, with the MCMC techniques discussed above. We started our reconstruction simulations from spherical random structures and random rotations and did not observe any dependence on the initial values.

The first dataset is comprised of 400 2D class averages of the 80S ribosome computed with SIMPLE2 (Elmlund and Elmlund, 2012) from cryo-EM micrographs (EMPIAR-10028); the size of the images is 80×80 pixels, the pixel size is 2.68 Å. The class averages are part of a SIMPLE2 tutorial and publicly available at https://simplecryoem.com/SIMPLE3.0/old_pages/2.5/data/

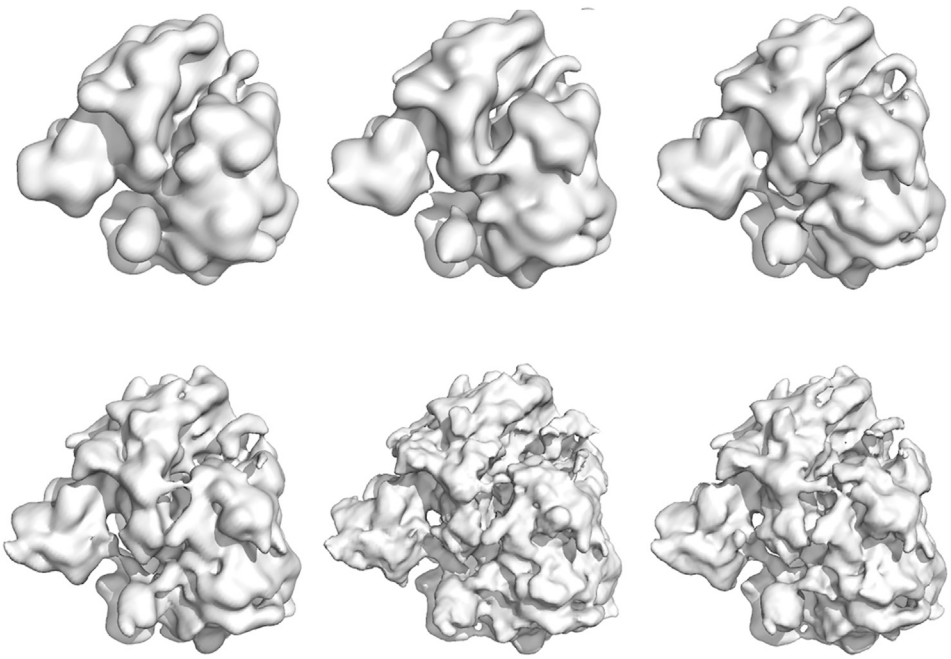


FIGURE 7 | Density maps of the 80S ribosome obtained with Bayesian random tomography using 50 class averages as input. Top row: 200, 1,000, 2000 particles **(left to right)**. Bottom row: 4,000, 8,000, 12,000 particles **(left to right)**.

simple2.5tutorials.tgz. **Figure 4** and **Supplementary Figure S1** show some example images and the 2D point clouds that were generated with the procedure outlined in **subsection 3.2**. Class averages were converted to 2D point clouds each composed of 1,000 points. Because the dataset is highly redundant, we only used the first 50 class averages and point clouds in the posterior simulations.

We used $K = 200$ and $K = 1000$ particles with a radius of $R = 16.4$ and $R = 8.4$ Å, respectively to fit the ribosome point clouds. We ran 500 iterations of Gibbs sampling with the global strategy for the rotational parameters and HMC for the particle positions. **Figure 5** shows five input point clouds and the projected model after convergence. We observe a good agreement between the experimental point clouds and the model point clouds with an RMSD ranging between 6.4 Å and 9.8 Å and an average of 7.7 ± 0.7 Å.

We also compared our 3D coarse-grained model of the 80S ribosome with a structure obtained with PRIME (Elmlund et al., 2008). To simplify the comparison, we converted the density map obtained with PRIME to a structure made up of 1,000 particles. The indices of the particle models were ordered such that spatially close particles have similar particle indices (which can be achieved, for example, by solving a traveling salesman problem using the matrix of inter-particle distances as input). Both structures show similar features (**Figure 6**); an FSC analysis reveals a resolution of 15.5 Å using the 0.143 criterion (**Supplementary Figure S6**).

We also ran simulations based on the first 50 class averages rather than 2D point clouds using 200 up to 12,000 particles. Again, we ran 500 steps of Gibbs sampling where the rotational parameters were updated globally with a frequency of 0.1. Projections of the 200 particle model are shown in the bottom rows of **Figure 5**. The cross-correlation coefficient between the class averages and the model images ranges between a minimum and maximum value of 90%–96% with an average of 94 ± 1%. For comparison, we also report the RMSDs to the particle clouds which range between 6.1 Å and 13.1 Å and an average of 8.3 ± 3.0 Å.

Using the last 100 particle configurations, we also generated density maps for each simulation and compared them to the high-resolution reconstruction EMD-2660 (Wong et al., 2014). The density maps are shown in **Figure 7**. To assess the quality of the particle models, we computed the FSC between the high-resolution map and the model maps (**Supplementary Figure S6**). Based on the 0.143 criterion, the resolution of the particle models ranges from 23.6 Å (200 particles) to 10.6 Å (12,000 particles). For comparison, the reconstruction obtained with SIMPLE reaches a resolution of 6.2 Å based on 200 class averages. More details about the quality of the reconstruction and computation times can be found in the Supplementary Material (**Supplementary Tables S2, S3**).

The posterior samples can be also used to assess the uncertainty of the particle models in the form of structural error bars. To carry out uncertainty quantification, the particle models first need to be superimposed and a correspondence between particles across different samples has to be established. We solve these two tasks by using the Iterative Closed Point (ICP) method followed by a linear assignment step where particle distances between superimpose clouds are used as a cost. **Supplementary Figure S7** shows an example for

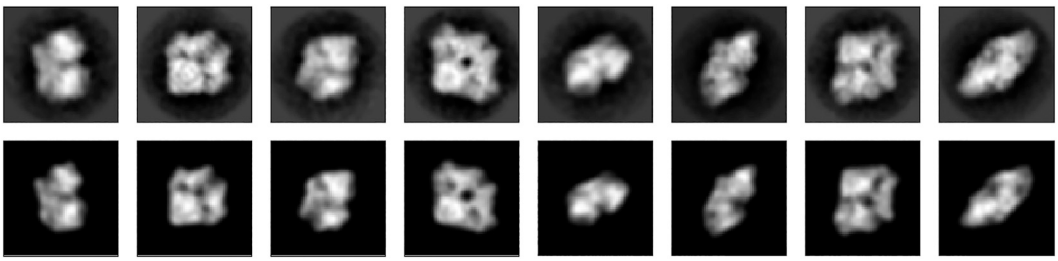


FIGURE 8 | 2D projections of beta-galactosidase. **Top row:** eight (out of 16) projection images (RELION class averages). **Bottom row:** Projection images calculated with Bayesian random tomography using 500 particles.

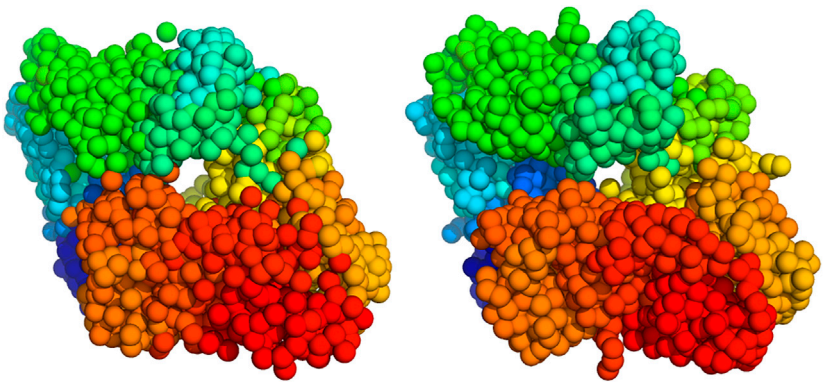


FIGURE 9 | 3D models of beta-galactosidase **(Left)** 2000 particle model inferred with Bayesian random tomography **(Right)** Coarse-grained model of the atomic structure (PDB code 1jz8).

structures based on 200 and 2000 particles. The distribution of uncertainties is inhomogeneous. Highly uncertain particles tend localize on the surface of the 200-particle model. The 2000-particle model shows smaller variations in the uncertainty of particle positions. So the large variations in the uncertainties of the 200-particle model might also be caused by the small number of particles.

The second cryo-EM dataset comprises 16 class averages of beta-galactosidase. These images are part of a RELION tutorial and available at ftp://ftp.mrc-lmb.cam.ac.uk/pub/scheres/relion31_tutorial_precalculated_results.tar.gz. The class average based on the data from EMPIAR-10204. The size of the images is 60×60 pixels, the pixel size is 3.54 Å. In this test, we inferred the structure from the images directly using likelihood **(11)** without converting the class averages to 2D point clouds.

Similar to the ribosome simulations we used 500 steps of Gibbs sampling with occasional global sampling of the rotational parameters to infer the coarse-grained structure of beta-galactosidase. We inferred structural models for systems with 100 up to 2000 particles.

The top row of **Figure 8** shows the first eight class averages that were used as an input for particle-based random tomography. The bottom row shows the projection images of a model composed of 500 particles that was obtained with sampling the full posterior distribution. Starting from a random initial structure and rotations, our sampling algorithm estimates a model structure and orientations that reproduce the experimental images closely with cross-correlation coefficients ranging between 94.7% and 97.5% and an average of $95.9 \pm 0.01\%$.

We compared the structure inferred with Bayesian random tomography against a high-resolution crystal structure (PDB code 1jz8) and a near-atomic cryo-EM reconstruction (EMD-5995). To enable this comparison, we converted the PDB structure to a 3D point cloud composed of 2000 particles. Correspondences between particles in our model and the model based on the crystal structure were established as in the calculation of the RMSD. **Figure 9** shows both models. The RMSD between our particle model and the Carbon-alpha atoms of the high-resolution structure 1jz8 is 3.4 Å. For comparison, we also report the RMSD between 1jz8 and its coarse-grained version (shown on the right of **Figure 9**) which is 2.4 Å. Bayesian random tomography achieves a similar accuracy by inferring a 3D model from the class averages as direct coarse graining of the high-resolution structure. **Supplementary Figure S8** shows density maps for all of the five simulations. By comparison with the high-resolution reconstruction (EMD-5995) we assess the resolution of the models to range between 25 Å (100 particles) and 11.5 Å (2000 particles). For comparison, the initial model from

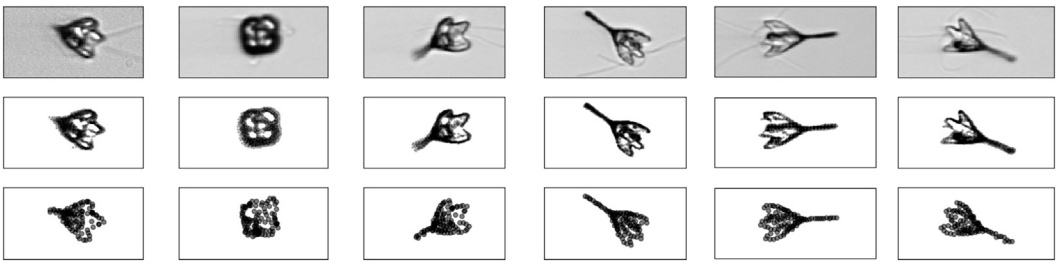


FIGURE 10 | Stochastic microscopy images of a plankton species. **Top row:** six (out of 16) projection images. **Middle row:** 2D point clouds representing the image data. **Bottom row:** 2D projections of the particle model calculated with Bayesian random tomography.

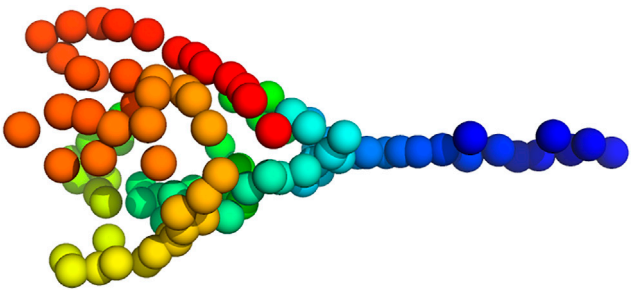


FIGURE 11 | 3D model of Pyramimonas Longicauda using 100 particles inferred from the point clouds shown in **Figure 10**.

RELION achieves a resolution of 9.8 Å (**Supplementary Figure S9** shows the corresponding FSC curves).

To assess the impact of the Boltzmann prior (**Eq. 17**), we ran two posterior simulations using 200 and 1,000 particles with the inverse temperature set to zero (i.e. the repulsive inter-particle energy is switched off). The quality of the reconstructed density map is largely unaffected by this change. For the 200 particles model, the average cross-correlation with Boltzmann prior is 94.7 ± 1.1%; without the Boltzmann prior we have 95.7 ± 0.9%. For the 1,000 particles model, these averages are 95.5 ± 1.5% (with Boltzmann prior) and 95.9 ± 1.5% (without Boltzmann prior). A comparison of the FSC curves obtained with and without Boltzmann prior confirms this finding (**Supplementary Figure S11**). The estimated resolution of the 200-particle model is 20.5 (19.4) Å with (without) Boltzmann prior; the 1000-particle model achieves a resolution of 12.0 (11.6) Å with (without) Boltzmann prior.

However, the Boltzmann prior has a strong effect on the packing of particles as assessed by the radial distribution functions (**Supplementary Figure S11**). With Boltzmann prior, the radial distribution shows a prominent peak close to the particle diameter, which is indicative of local order similar to a fluid. Without the Boltzmann prior, this peak disappears and we observe an enrichment of very short distances indicating a physically unrealistic particle packing. If our goal is to reconstruct a single 3D density from a homogeneous dataset, introducing the Boltzmann prior is not harmful, but dispensable. Turning the argument around, we find that the Boltzmann prior is compatible with the data and does not result in a severe loss of fitting quality. We expect that the prior will become essential in more advanced 3D reconstruction tasks, in particular when facing conformational heterogeneity.

Finally, we applied our random tomography approach to a dataset that shows structures on length scales that are much larger than the length scales imaged in cryo-EM. Following the work by Levis et al. (2018), we downloaded *in situ* microscopy images of the marine plankton species Pyramimonas Longicauda; the data are available at https://darchive.mblwhoilibrary.org/handle/1912/7341. These mesoscopic organisms are transparent and therefore allow for 3D reconstruction from 2D microscopic images. Since the organism seems to be quasi symmetric, we selected out of the 121 projection images recorded in 2013, 16 representative images. The selected images cover most of the views that are present in the dataset.

The intensity of microscopic images g_n is proportional to the transmissivity, which is related to the optical density of the object via an exponential transform. Therefore, to convert the images to 2D point clouds, we use the expectation maximization approach (see Supplementary Material) with weights proportional to $-\log g_n > 0$, since $g_n \in (0, 1)$. The six out of the 16 selected images and their point cloud representations are shown in **Figure 10**. Each microscopic image was converted to 2D cloud composed of 1,000 points.

The fact that the magnification can vary from image to image requires that we extend the likelihood for 2D point clouds **(13)** (also **Supplementary Equations S1, S2** in the Supplementary Material). These variations are accounted for by an additional factor that scales the coordinates of the projected model so as to match the 2D point cloud derived from the microscopic image. Moreover, we need to account for shifts in the image plane. These extensions increase the number of unknown parameters per image from four to eight: four quaternions parameterizing the unknown orientation, two translation parameters accounting for a shift, a scaling factor compensating variations in the magnification and a precision.

Inference of a 3D particle model proceeded as before. We estimated a model composed of 100 particles from the 16 2D point clouds starting from a random structure and random rotations (the initial values for the scaling factors and

translations were one and zero, respectively). **Figure 11** shows a 3D model of the plankton species inferred with Bayesian random tomography.

4 DISCUSSION

We outlined a Bayesian approach to random tomography, the problem of reconstructing a 3D structure from 2D views along unknown random directions. At the core of our approach is a representation of 3D volumes using a radial basis function kernel whose centers are our main inference parameters. We interpret the kernel centers as particle positions and use an excluded-volume prior to ensure that estimated particle configurations show a physically plausible packing. We demonstrated that coarse-grained models can be inferred from projection data (images or point clouds) with MCMC algorithms such as HMC and global sampling of the rotations.

In cryo-EM applications, our approach can be used to generate an initial model that can be refined further. So far, we tested the method only an class averages that displayed a high SNR. In future applications, we plan to explore the use of Bayesian random tomography from raw cryo-EM images and include the effect of the CTF into our model. Another route for extending the approach is account for conformational heterogeneity, which is one of the major bottlenecks in cryo-EM data processing. An interesting approach to characterize conformational variability in the presence of continuous flexibility has been proposed recently by Chen and Ludtke (2021) who use an autoencoder network with a Gaussian mixture model to represent conformational changes in a low dimensional latent space.

In all applications discussed in this paper, the number of particles K was fixed. An interesting question for future research is to estimate the number of particles based on the projection data. This might also provide a new way of measuring the resolution of the input data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://darchive.mblwhoilibrary.org/handle/1912/7341 ftp://ftp.mrc-lmb.cam.ac.uk/pub/scheres/relion31_tutorial_precalculated_results.tar.gz https://simplecryoem.com/SIMPLE3.0/old_pages/2.5/data/simple2.5tutorials.tgz.

AUTHOR CONTRIBUTIONS

MH designed research. NV and MH performed research. NV and MH contributed new analytic tools. NV and MH analyzed data. NV and MH wrote the paper.

ACKNOWLEDGMENTS

MH acknowledges funding from the German Research Foundation (DFG) under project SFB 860, TP B09 as well as funding from the Carl-Zeiss foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.658269/full#supplementary-material

REFERENCES

Barnett, A., Greengard, L., Pataki, A., and Spivak, M. (2017). Rapid Solution of the Cryo-Em Reconstruction Problem by Frequency Marching. *SIAM J. Imaging Sci.* 10, 1170–1195. doi:10.1137/16m1097171

Bendory, T., Bartesaghi, A., and Singer, A. (2020). Single-particle Cryo-Electron Microscopy: Mathematical Theory, Computational Challenges, and Opportunities. *IEEE Signal. Process. Mag.* 37, 58–76. doi:10.1109/msp.2019.2957822

Chen, M., and Ludtke, S. (2021). Deep Learning Based Mixed-Dimensional Gmm for Characterizing Variability in Cryoem.arXiv preprint arXiv:2101.10356

Coxeter, H. S. M. (1973). *Regular Polytopes*. New York, NY: Courier Corporation.

Elmlund, D., and Elmlund, H. (2012). SIMPLE: Software for ab initio Reconstruction of Heterogeneous Single-Particles. *J. Struct. Biol.* 180, 420–427. doi:10.1016/j.jsb.2012.07.010

Elmlund, H., Elmlund, D., and Bengio, S. (2013). PRIME: Probabilistic Initial 3D Model Generation for Single-Particle Cryo-Electron Microscopy. *Structure* 21, 1299–1306. doi:10.1016/j.str.2013.07.002

Elmlund, H., Lundqvist, J., Al-Karadaghi, S., Hansson, M., Hebert, H., and Lindahl, M. (2008). A New Cryo-EM Single-Particle ab initio Reconstruction Method Visualizes Secondary Structure Elements in an ATP-Fueled AAA+ Motor. *J. Mol. Biol.* 375, 934–947. doi:10.1016/j.jmb.2007.11.028

Frank, J. (2006). *Three-dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*. Oxford University Press. doi:10.1093/acprof:oso/9780195182187.001.0001

Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-6, 721–741. doi:10.1109/tpami.1984.4767596

Habeck, M. (2017). Bayesian Modeling of Biomolecular Assemblies with Cryo-EM Maps. *Front. Mol. Biosci.* 4, 15. doi:10.3389/fmolb.2017.00015

Habeck, M. (2009). Generation of Three-Dimensional Random Rotations in Fitting and Matching Problems. *Comput. Stat.* 24, 719–731. doi:10.1007/s00180-009-0156-x

Horn, B. K. P. (1987). Closed-form Solution of Absolute Orientation Using Unit Quaternions. *J. Opt. Soc. Am. A.* 4, 629–642. doi:10.1364/josaa.4.000629

Jaitly, N., Brubaker, M. A., Rubinstein, J. L., and Lilien, R. H. (2010). A Bayesian Method for 3D Macromolecular Structure Inference Using Class Average Images from Single Particle Electron Microscopy. *Bioinformatics* 26, 2406–2415. doi:10.1093/bioinformatics/btq456

Jin, Q., Sorzano, C. O. S., de la Rosa-Trevín, J. M., Bilbao-Castro, J. R., Núñez-Ramírez, R., Llorca, O., et al. (2014). Iterative Elastic 3d-To-2d Alignment Method Using Normal Modes for Studying Structural Dynamics of Large Macromolecular Complexes. *Structure* 22, 496–506. doi:10.1016/j.str.2014.01.004

Jonić, S., Vargas, J., Melero, R., Gómez-Blanco, J., Carazo, J. M., and Sorzano, C. O. (2016). Denoising of High-Resolution Single-Particle Electron-Microscopy Density Maps by Their Approximation Using Three-Dimensional Gaussian Functions. *J. Struct. Biol.* 194, 423–433. doi:10.1016/j.jsb.2016.04.007

Jonic, S., and Sanchez Sorzano, C. O. (2016). Coarse-graining of Volumes for Modeling of Structure and Dynamics in Electron Microscopy: Algorithm to

Automatically Control Accuracy of Approximation. *IEEE J. Sel. Top. Signal. Process.* 10, 161–173. doi:10.1109/JSTSP.2015.2489186

Joubert, P., and Habeck, M. (2015). Bayesian Inference of Initial Models in Cryo-Electron Microscopy Using Pseudo-atoms. *Biophysical J.* 108, 1165–1175. doi:10.1016/j.bpj.2014.12.054

Kam, Z. (1980). The Reconstruction of Structure from Electron Micrographs of Randomly Oriented Particles. *J. Theor. Biol.* 82, 15–39. doi:10.1016/0022-5193(80)90088-0

Kulis, B., and Jordan, M. I. (2012). "Revisiting K-Means: New Algorithms via Bayesian Nonparametrics," in Proceedings of the 29th International Conference on Machine Learning (ICML-12). Editors J. Langford and J. Pineau (New York, NY, USA), 513–520.

Levin, E., Bendory, T., Boumal, N., Kileel, J., and Singer, A. (2018). "3d ab initio Modeling in Cryo-Em by Autocorrelation Analysis,"in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 1569–1573.

Levis, A., Schechner, Y. Y., and Talmon, R. (2018). Statistical Tomography of Microscopic Life. *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 6411–6420.

Liang, J., and Dill, K. A. (2001). Are Proteins Well-Packed? *Biophys. J.* 81, 751–766. doi:10.1016/s0006-3495(01)75739-6

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer.

Lyumkis, D., Vinterbo, S., Potter, C. S., and Carragher, B. (2013). Optimod - an Automated Approach for Constructing and Optimizing Initial Models for Single-Particle Electron Microscopy. *J. Struct. Biol.* 184, 417–426. doi:10.1016/j.jsb.2013.10.009

Mechelke, M., and Habeck, M. (2013). Estimation of Interaction Potentials through the Configurational Temperature Formalism. *J. Chem. Theor. Comput.* 9, 5685–5692. doi:10.1021/ct400580p

Natterer, F. (2001). *The Mathematics of Computerized Tomography*. Philadelphia, Pa: SIAM. doi:10.1137/1.9780898719284

Neal, R. M. (2011). *Handbook of Markov Chain Monte Carlo*, 113–162.Mcmc Using Hamiltonian Dynamics

Panaretos, V. M. (2009). On Random Tomography with Unobservable Projection Angles. *Ann. Stat.* 37, 3272–3306. doi:10.1214/08-aos673

Penczek, P. A., Zhu, J., and Frank, J. (1996). A Common-Lines Based Method for Determining Orientations for N > 3 Particle Projections Simultaneously. *Ultramicroscopy* 63, 205–218. doi:10.1016/0304-3991(96)00037-x

Punjani, A., Rubinstein, J. L., Fleet, D. J., and Brubaker, M. A. (2017). cryoSPARC: Algorithms for Rapid Unsupervised Cryo-EM Structure Determination. *Nat. Methods* 14, 290–296. doi:10.1038/nmeth.4169

Sanz-García, E., Stewart, A. B., and Belnap, D. M. (2010). The Random-Model Method Enables ab initio 3D Reconstruction of Asymmetric Particles and Determination of Particle Symmetry. *J. Struct. Biol.* 171, 216–222. doi:10.1016/j.jsb.2010.03.017

Schaback, R., and Wendland, H. (2006). Kernel Techniques: from Machine Learning to Meshless Methods. *Acta numerica* 15, 543–639. doi:10.1017/s0962492906270016

Scheres, S. H. W. (2012a). A Bayesian View on Cryo-EM Structure Determination. *J. Mol. Biol.* 415, 406–418. doi:10.1016/j.jmb.2011.11.010

Scheres, S. H. W., Gao, H., Valle, M., Herman, G. T., Eggermont, P. P. B., Frank, J., et al. (2007). Disentangling Conformational States of Macromolecules in 3D-EM through Likelihood Optimization. *Nat. Methods* 4, 27–29. doi:10.1038/nmeth992

Scheres, S. H. W. (2010). Maximum-likelihood Methods in Cryo-EM. Part II: Application to Experimental Data. *Methods Enzymol.* 482, 295–320. doi:10.1016/s0076-6879(10)82012-9

Scheres, S. H. W. (2012b). RELION: Implementation of a Bayesian Approach to Cryo-EM Structure Determination. *J. Struct. Biol.* 180, 519–530. doi:10.1016/j.jsb.2012.09.006

Schölkopf, B., and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*. MIT.

Singer, A., and Shkolnisky, Y. (2011). Three-Dimensional Structure Determination from Common Lines in Cryo-EM by Eigenvectors and Semidefinite Programming. *SIAM J. Imaging Sci.* 4, 543–572. doi:10.1137/090767777

Takeda, H., Farsiu, S., and Milanfar, P. (2007). Kernel Regression for Image Processing and Reconstruction. *IEEE Trans. Image Process.* 16, 349–366. doi:10.1109/tip.2006.888330

Vainshtein, B. K., and Goncharov, A. B. (1986). Determination of the Spatial Orientation of Arbitrarily Arranged Identical Particles of Unknown Structure from Their Projections. *Soviet Phys. Doklady* 31, 278.

Van Heel, M. (1987). Angular Reconstitution: A Posteriori Assignment of Projection Directions for 3D Reconstruction. *Ultramicroscopy* 21, 111–123. doi:10.1016/0304-3991(87)90078-7

Vargas, J., Álvarez-Cabrera, A.-L., Marabini, R., Carazo, J. M., and Sorzano, C. O. S. (2014). Efficient Initial Volume Determination from Electron Microscopy Images of Single Particles. *Bioinformatics* 30, 2891–2898. doi:10.1093/bioinformatics/btu404

von Ardenne, B., Mechelke, M., and Grubmüller, H. (2018). Structure Determination from Single Molecule X-Ray Scattering with Three Photons Per Image. *Nat. Commun.* 9, 1–9. doi:10.1038/s41467-018-04830-4

Wong, W., Bai, Xc., Brown, A., Fernandez, I. S., Hanssen, E., Condron, M., et al. (2014). Cryo-em Structure of the Plasmodium Falciparum 80s Ribosome Bound to the Anti-protozoan Drug Emetine. *Elife* 3, e03080. doi:10.7554/eLife.03080

Yan, X., Dryden, K. A., Tang, J., and Baker, T. S. (2007). Ab Initio random Model Method Facilitates 3D Reconstruction of Icosahedral Particles. *J. Struct. Biol.* 157, 211–225. doi:10.1016/j.jsb.2006.07.013

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: 26 May 2021
doi: 10.3389/fmolb.2021.694130

How to Determine Accurate Conformational Ensembles by Metadynamics Metainference: A Chignolin Study Case

Cristina Paissoni and Carlo Camilloni**

Dipartimento di Bioscienze, Università degli Studi di Milano, Milan, Italy

The reliability and usefulness of molecular dynamics simulations of equilibrium processes rests on their statistical precision and their capability to generate conformational ensembles in agreement with available experimental knowledge. Metadynamics Metainference (M&M), coupling molecular dynamics with the enhanced sampling ability of Metadynamics and with the ability to integrate experimental information of Metainference, can in principle achieve both goals. Here we show that three different Metadynamics setups provide converged estimate of the populations of the three-states populated by a model peptide. Errors are estimated correctly by block averaging, but higher precision is obtained by performing independent replicates. One effect of Metadynamics is that of dramatically decreasing the number of effective frames resulting from the simulations and this is relevant for M&M where the number of replicas should be large enough to capture the conformational heterogeneity behind the experimental data. Our simulations allow also us to propose that monitoring the relative error associated with conformational averaging can help to determine the minimum number of replicas to be simulated in the context of M&M simulations. Altogether our data provides useful indication on how to generate sound conformational ensemble in agreement with experimental data.

Keywords: molecular dynamics, metadynamics, metainference, statistical error, conformational ensembles

OPEN ACCESS

Edited by:
Gregory Bowman, Washington University School of Medicine in St. Louis, United States

Reviewed by:
Luca Bellucci, Istituto Nanoscienze, Consiglio Nazionale delle Ricerche, Italy
Sandro Bottaro, University of Copenhagen, Denmark

****Correspondence:***
Cristina Paissoni
cristina.paissoni@unimi.it
Carlo Camilloni
carlo.camilloni@unimi.it

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

Received: *12 April 2021*
Accepted: *14 May 2021*
Published: *26 May 2021*

Citation:
Paissoni C and Camilloni C (2021) How to Determine Accurate Conformational Ensembles by Metadynamics Metainference: A Chignolin Study Case. Front. Mol. Biosci. 8:694130. doi: 10.3389/fmolb.2021.694130

INTRODUCTION

Molecular dynamics simulations (MD) are a powerful tool to study at high resolution the dynamics of biomolecules in solution, yet they rely on the quality of the physical model used to describe molecules (i.e., the force field) as well as on the computing power needed to acquire longer and longer trajectories that is better and better statistics (Bottaro and Lindorff-Larsen, 2018; Grossfield et al., 2019). Force fields have been dramatically improving in the last years and computing power is always increasing allowing to study more and more complex systems (Best et al., 2014; Huang et al., 2017; Robustelli et al., 2018). To further improve the extent of the sampling and the accuracy of the physical model, enhanced sampling techniques (Sugita and Okamoto, 1999; Laio and Parrinello, 2002) as well as techniques to integrate experimental data in MD have been developed (Fennen et al., 1995; Bonomi et al., 2016; Köfinger et al., 2019). Reviewing the vast literature on both topics is outside the scope and space of the present work and excellent reviews are available (Spiwok et al., 2015; Allison, 2017; Bonomi et al., 2017; Bottaro and Lindorff-Larsen, 2018; Camilloni and Pietrucci, 2018; Bernetti

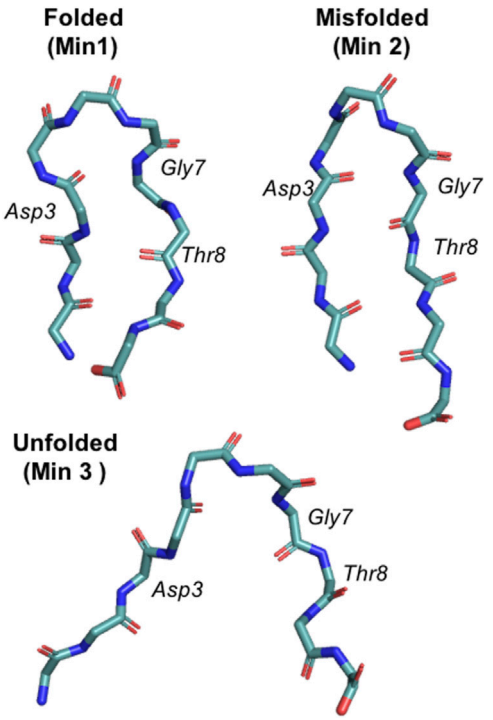


FIGURE 1 | Representation of the three main chignolin minima corresponding to the folded (Min 1), misfolded (Min 2) and unfolded (Min 3) states.

et al., 2020). Among these methods we have contributed to develop Metadynamics Metainference (M&M) (Bonomi et al., 2016a) that is a combination of Metadynamics (Laio and Parrinello, 2002), a popular enhanced sampling technique, and Metainference (Bonomi et al., 2016), a Bayesian scheme that allows for the integration of equilibrium experimental observables as restraints over multiple replicas of a simulation. M&M has been applied to combine different experimental observables and to work on a large variety of systems (Löhr et al., 2017; Eshun-Wilson et al., 2019; Heller et al., 2020; Jussupow et al., 2020).

In this work we aim to understand how Metadynamics should be ideally coupled to Metainference in order to guarantee optimal statistical precision and experimental accuracy. Multiple MetaD variants are available and M&M has always been coupled with Parallel Bias Metadynamics (PBMetaD), a variant specifically designed to enhance the sampling along many one-dimensional collective variables (CVs) (Pfaendtner and Bonomi, 2015). In particular we identified three key questions: 1) how reliable are the error estimates resulting from Metadynamics simulations when using a standard technique as block averaging (Flyvbjerg and Petersen, 1989); 2) how does multiple-walkers PBMetaD compare to conventional multiple-walkers MetaD and what are their pros-and-cons; 3) how do the two approaches combine with Metainference to achieve at the same time an optimal sampling and an optimal integration of experimental data? Of note, the first two questions apply not only to M&M but to the sound application of enhanced sampling techniques. To answer these questions, we investigated thoroughly the conformational space of chignolin (**Figure 1**), a 10 residues peptide that can populate three states and whose complexity, while not comparable to that of full-length proteins, is definitely greater than the widely used alanine dipeptide in vacuum (Kührová et al., 2012). In doing so we introduced a scheme to combine simple CVs into more complex ones with the aim of discriminating some identified reference states. By performing PBMetaD simulations with many simple CVs (PB20), PBMetaD simulations with less, optimally combined, CVs (PB4); as well as MetaD simulations with the same optimally combined CVs (ME2), all in triplicate (**Table 1**), we show that, 1) block-averaging provides a robust estimate of statistical errors; 2) PBMetaD and MetaD dramatically decrease the effective number of frames collected by MD and this effect is worse in MetaD. This second effect is very relevant in combining Metadynamics with Metainference because it decreases the number of effective replicas that can actually contribute to the estimation of the conformational heterogeneity associated with experimental observables. To test this effect, we then performed (**Table 1**) M&M simulations using either PBMetaD or MetaD and 10 or 100 replicas. To avoid effects related to the quality of the experimental data and the forward model, synthetic SAXS data have been obtained using as a reference a 40 μs long simulated tempering simulation of chignolin by Piana et al., 2020. Our results indicate that the minimum number of replicas in M&M simulations can be set by monitoring the relative error associated with the averaging of back calculated observables, and that this number is affected not only by the system and the calculated observable but also by the details of the Metadynamics setup.

MATERIALS AND METHODS

Molecular Dynamics Simulations of Chignolin

Simulations of chignolin were performed using GROMACS 2019 (Abraham et al., 2015) and PLUMED 2 (Tribello et al., 2014). In the first round of simulations the DES-amber force field (Piana et al., 2020) was used in combination with the tip4p water model with increased dispersion (Piana et al., 2015). A starting model of CLN025 chignolin was taken from PDB 5AWL (Honda et al., 2008) and solvated with 2,553 water molecules in a dodecahedron box initially 1.4 nm larger than the protein in each direction. The system was neutralized with a salt concentration of 100 mM NaCl. After an initial energy minimization to a maximum force of 100 kJ/mol/nm, the solute was equilibrated under NVT condition at the temperature of 340 K for 50 ps using the Berendsen thermostat (Berendsen et al., 1984); then Berendsen barostat was used to equilibrate the system in the NPT ensemble to the target pressure of 1 atm for 200 ps, maintaining the temperature at 340 K with the Bussi thermostat (Bussi et al., 2007). The equilibration phase was followed by an initial MD simulation of 250 ns, from which a pool of conformations was extracted to be used as starting models for the subsequent runs (run 1). Starting points for replicates run

TABLE 1 | Summary of the simulations performed or analyzed in this work.

Name	Replicates	#Replicas	Enhanced sampling technique	#CV	Force field	Replica length (total) μs	Color code
Reference[a]	1	1	Simulated tempering	NA	DES-amber	40	Dark grey
PB20	3	10	PBMetaD	20	DES-amber	1 (30)	Blues
PB4	3	10	PBMetaD	4	DES-amber	1 (30)	Greens
ME2	3	10	MetaD	2	DES-amber	1 (30)	Violets
Prior	1	10	PBMetaD	4	99sb-ildn	1 (10)	Light grey
PB4(10r)	1	10	PBMetaD	4	99sb-ildn + M&M	0.5 (5)	Cyan
ME2 (10r)	1	10	MetaD	2	99sb-ildn + M&M	0.5 (5)	Yellow
PB4(100r)	1	100	PBMetaD	4	99sb-ildn + M&M	0.5 (50)	Blue
ME2 (100r)	1	100	MetaD	2	99sb-ildn + M&M	0.5 (50)	Orange

For each simulation are reported: the number of replicates, the replicas (or walkers), the enhanced sampling technique employed, the number of CVs, the force field, the length of each replica (and the total simulation time) and the color code associated to the simulation in the figures.
[a]*This simulation was performed by Piana et al. (2020).*

2 and run 3, where instead extracted from run 1 thus resulting in very different initial conditions. The production runs were all performed in the NPT ensemble, maintaining temperature and pressure at the values of 340 K and 1 atm respectively, using the Bussi thermostat (Bussi et al., 2007) and the Parrinello-Rahman barostat (Parrinello and Rahman, 1981). Electrostatic was treated by using the particle mesh Ewald scheme (Essmann et al., 1995) with a short-range cutoff of 0.9 nm and a Fourier grid spacing of 0.12 nm; van der Waals interaction cutoff was set to 0.9 nm. For these simulations the hydrogen mass repartitioning scheme (Hopkins et al., 2015) was used to reduce the computational cost: the mass of heavy atoms was repartitioned into the bonded hydrogen atoms using the *heavyh* flag in the pdb2gmx tool; the LINCS algorithm (Hess et al., 1997) was used to constraint all bonds, eventually allowing to use a time step of 5 fs.

Using this set-up, we ran three different Metadynamics simulations, each performed in triplicates (named run 1, run 2, run 3, starting from different set of conformations). These are:

1. PB20: in which PBMetaD was employed and 20 CVs were biased. These include the phi/psi dihedral angles of the 10 amino acids composing chignolin (18 CVs), the gyration radius and the antiparallel beta sheet-content.
2. PB4: in which PBMetaD was employed biasing 4 CVs, comprising the gyration radius, the antiparallel beta sheet-content and 2 CVs optimized based on the knowledge of the folded, misfolded and unfolded chignolin conformations (named *back* and *cmap*, and based on a combination of backbone dihedral angle and of contacts between groups of atoms, see next section).
3. ME2: in which MetaD was employed using 2 CVs, the gyration radius and the optimized *cmap* collective variable.

All the simulations were performed adopting the multiple-walker scheme (Raiteri et al., 2006), simulating 10 replicas (or walkers): each replica was evolved for 1 μs, resulting in a 10 μs sampling per each simulation. Metadynamics was used in its well-tempered version (Barducci et al., 2008), where Gaussians with an initial height of 0.5 kJ/mol were deposited every 1 ps using a bias factor of 10. For all the CVs, the width of the Gaussians was determined with the dynamically adapted geometry-based Gaussian approach (Branduardi et al., 2012), using 0.015 nm as the extent of Cartesian space covered by a variable to estimate CVs fluctuations, and setting a minimum value for the width specific for each CV (0.03 rad for the dihedral angles, 0.004 nm for the gyration radius, 0.02 for the antiparallel beta sheet-content, 0.01 and 0.001 for the *back* and *cmap* optimized CVs).

Each simulation was analyzed by creating a concatenated trajectory and reweighting each frame by using the final Metadynamics bias potential, assuming a constant bias during the entire course of the simulation (Branduardi et al., 2012). To assess the convergence of the simulations and the associated statistical errors we used block-average analysis (Flyvbjerg and Petersen, 1989; Bussi and Tribello, 2019). According to this technique, the trajectory is split into a set of *NB* blocks of equal length. By comparing the averages of a given quantity from each block we can calculate the error bar on our estimate of that quantity: for large enough blocks the averages should not be time correlated so that the estimate of the error converges. As our blocks could be characterized by different weights, this must be taken into account in the estimation of the error as described in (Invernizzi et al., 2020). Given W_b the weight of the block b, obtained as the sum of the weights of the frames composing the block, the statistical error on the observable O is: $err_O = \sqrt{\frac{1}{(NB_{eff}-1)}\frac{\sum_{b=1}^{NB} W_b\,[\widehat{O_b}-\widehat{O}]^2}{\sum_{b=1}^{NB} W_b}}$, where $NB_{eff} = (\sum_{b=1}^{NB} W_b)^2/\sum_{b=1}^{NB} W_b^2$ is the effective block size, the sums run on the number of blocks *NB*, $\widehat{O}_b$ is the average computed over the frames of block b and $\widehat{O}$ is the average computed over all the frames, which corresponds to the average computed over the block averages, i.e. $\widehat{O} = \sum_{b=1}^{NB} W_b\,\widehat{O}_b / \sum_{b=1}^{NB} W_b$. As pointed out in (Invernizzi et al., 2020) when the weights of the blocks are unbalanced, using *NB* instead of NB_{eff} can significantly underestimate the uncertainty.

Optimized Collective Variables

PBMetaD can in principle bias many CVs using one-dimensional Gaussians (Pfaendtner and Bonomi, 2015), but often these CVs are simple in nature (like dihedrals or distances) thus losing the complex correlations that may be at play in slow reaction coordinates. Finding optimal CVs is a complex problem that requires the previous knowledge not only of the different states

but also of the pathways connecting them. Example of methods using reactive pathways to estimate optimal CVs include TICA, SGOOP and machine learning approaches (Tiwary and Berne, 2016; McCarty and Parrinello, 2017; Sultan and Pande, 2017, 2018; Wang and Tiwary, 2020). Instead of learning from reactive pathways one can instead try to only maximize the discrimination of the different states as implemented in HLDA (Mendels et al., 2018). One possible limitation of this latter approach, which has the clear advantage of being more affordable for large and complex systems, is that a CV that optimally discriminate states may not correspond to an efficient reaction coordinate. Here we propose a simple method to generate a novel $CV(\boldsymbol{a},\boldsymbol{\varphi}) = \sum_{i=1}^{N} a_i\varphi_i$, where $\boldsymbol{a}$ is a normalized vector of size N, starting from N input simple collective variables $\boldsymbol{\varphi}$ (e.g., these could be the backbone dihedral angles, or the Cα-Cα contacts). $CV(\boldsymbol{a},\boldsymbol{\varphi})$ while trying to discriminate two or more states, tries also to 1) discard as few of the input CVs $\boldsymbol{\varphi}$ as possible by keeping the weights $\boldsymbol{a}$ of the combined CVs as uniform as possible; and 2) keep the width of the minima comparable. This latter property is relevant for methods like Metadynamics that uses Gaussians. To achieve these properties the optimal value $\boldsymbol{a}$ is obtained by minimizing the following scoring function (here given for two states indicated as 1 and 2):

$$\psi(\boldsymbol{a}) = -\frac{< CV_1 > - < CV_2 >}{2\left(\sigma_{CV_1}^2 + \sigma_{CV_2}^2\right)} + \frac{\max\left(\sigma_{CV_1}, \sigma_{CV_2}\right)}{\min\left(\sigma_{CV_1}, \sigma_{CV_2}\right)} + \sum_{i=1}^{N} a_i^2 \ln\frac{a_i^2}{1/N}$$

where the first term maximizes the discrimination among states, the second keeps the width of the minima comparable, the last keeps the parameters as uniform as possible.

This approach is then applied to optimize two CVs, *back* and *cmap*, as the combination of chignolin backbone dihedral angles and the contacts among the center of the backbone of i − i + 3 aminoacids, respectively. The CVs are first calculated for the three states as observed in the preliminary 250 ns long simulation (**Supplementary Figure S1**) and then their combination is obtained as described above. The distribution of the values for the *cmap* CV before and after optimization is reported in **Supplementary Figure S1**.

Metainference

Metainference is a technique based on Bayesian inference and replica-averaging modeling (Rieping, 2005; Cavalli et al., 2013; Bonomi et al., 2016). Following the replica-averaging modeling strategies, multiple replicas of the system are simulated in parallel and the quantities to be restrained against experimental data are back-calculated as averages over the replicas, thus taking into account the effects of conformational averaging. Bayesian inference allows to modulate the strength of the restraints estimating, along with the model, statistical errors, which include random and systematic errors as well as inaccuracies of the forward model.

In the case of Gaussian noise, the Metainference energy is described by (Löhr et al., 2017): $E_{MI} = E_{FF} + \frac{k_B T}{2}\sum_{i=1}^{N_d}\sum_{r=1}^{NR}[d_i - \lambda\langle f_i(X)\rangle]^2/(\sigma_{r,i}^B)^2 + (\sigma_i^{SEM})^2 + E_\sigma$, where E_{FF} is the force field energy, k_B is the Boltzmann constant, T the temperature, d the set of N_d experimental data, $f(\boldsymbol{X})$ is the forward model used to back-calculate the observable from conformation $\boldsymbol{X}$, $f_i(X)$ indicates the average over the NR replicas for observable i, $\sigma_{r,i}^B$ is an uncertainty parameter that describes random and systematic errors, σ_i^{SEM} is the standard error of the mean related to conformational averaging, λ is an optional scaling parameter and E_σ is an energy term that accounts for normalization of the data likelihood and error priors. In Metainference Monte Carlo sampling is used to sample both the uncertainty $\sigma_{r,i}^B$ (which depends on both the replica and the observable) and optionally the scaling parameter λ.

Metainference can be combined with Metadynamics (M&M) to accelerate the exploration of the conformational space (Bonomi et al., 2016; Löhr et al., 2017). In M&M the replicas share the Metadynamics bias potential as in the case of multiple-walkers method (Raiteri et al., 2006). Depending on the bias potential V_G each replica r has a different weight that can be approximated on the fly as $w_r \sim e^{V_G(CV(X_r))/k_B T}$, with $CV(\boldsymbol{X}_r)$ representing the set of selected CVs, functions of the microscopic coordinates $\boldsymbol{X}$. Therefore, these weights must be taken into account when calculating the experimental averages and the standard error of the mean σ_i^{SEM}, that are computed as: $f_i(\boldsymbol{X}) = \sum_{r=1}^{NR} w_r f_i(\boldsymbol{X}_r)/\sum_{r=1}^{NR} w_r$ and $\sigma_i^{SEM} = \sqrt{\frac{1}{(NR_{eff}-1)}\frac{\sum_{r=1}^{NR} w_r\,[f_i(\boldsymbol{X}_r)-\langle f_i(\boldsymbol{X})\rangle]^2}{\sum_{r=1}^{NR} w_r}}$, with $NR_{eff} = (\sum_{r=1}^{NR} w_r)^2/\sum_{r=1}^{NR} w_r^2$ representing the number of effective replicas. In order to reduce the noise resulting from the instantaneous fluctuations of the bias, the weight of each replica is calculated *via* a moving average of the bias over a given number of MD steps (set by the keyword AVERAGING). Also, to reduce the oscillations of σ_i^{SEM} we used the maximum value of σ_i^{SEM} over the same time window defined by AVERAGING keyword. Finally, we automatically determined the maximum values that can be sampled for $\sigma_{r,i}^B$ as $\max(\sigma_{r,i}^B) = \sigma_i^{SEM}\sqrt{NR}$, with NR being the number of replicas (this option can be set in plumed using the keyword OPTSIGMAMEAN = SEM_MAX).

Small-Angle X-Ray Scattering (SAXS)-Driven Molecular Dynamics Simulations

Synthetic SAXS intensities, to be used as target for the restraints in our simulations, were calculated from a reference 40 μs long MD trajectory, performed with the DES-amber forcefield and provided by Piana et al. 2020. From this simulation a set of 24 representative SAXS intensities at different scattering angles, ranging between 0.01 and 1.39 $Å^{-1}$ and equally spaced, were calculated with PLUMED using atomistic structure factors and considering only the trajectory frames with temperature close to 340 K (Paissoni et al., 2019; Paissoni et al., 2020). While we know that this range is not representative of a realistic SAXS experiment, considering the small dimension of the protein we decided to use such a large range to include higher resolution details. SAXS restraints were applied every 2 MD steps and atomic scattering factors were used to back-calculate the 24 SAXS intensities. The SEM_MAX option was used to automatically estimate both the σ_i^{SEM} as well as the maximum

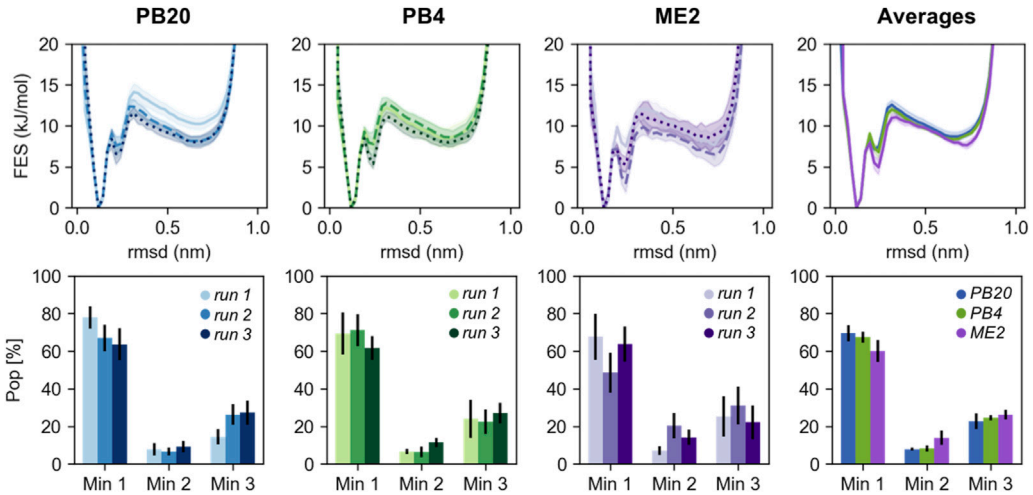


FIGURE 2 | The RMSD free energies **(top panels)** and the population of the three main chignolin minima **(bottom panels)** are represented for different sets of simulations (PB20 in blues, PB4 in greens, ME2 in violets shades). The different shades indicate each of the runs of the triplicates and the errors are estimated *via* block-average analysis. In the rightest panels are reported the averages computed over the triplicates for each set of simulations; here the errors are determined as standard errors over the triplicates as described in the main text. In all the pictures the free energies are shifted to set their minimum to 0.

value of $\sigma^B_{r,i}$ for the M&M simulations; the window averaging for the estimation of the weights was performed on a time window of 1 ps to match the frequency of deposition of Metadynamics hills.

For the set of SAXS-driven simulations we used as prior the amber99sb-ildn (Lindorff-Larsen et al., 2010) force field with the tip3p water model (Mark and Nilsson, 2001). The system was prepared and equilibrated as described above and a set of starting conformations was generated from a 1 μs long plain MD simulation. We performed five Metadynamics simulations (**Table 1**): one unrestrained, prior, amber99sb-ildn simulation, using the PB4 setup with 10 replicas; two simulations, PB4 (10r) and PB4 (100r), with the PB4 setup plus SAXS restraints using either 10 or 100 replicas; two simulations, ME2 (10r) and ME2 (100r), with the ME2 setup plus SAXS restraints, using either 10 or 100 replicas. The unrestrained prior simulation evolved for 1 μs per replica, while the SAXS driven simulations evolved for 500 ns per replica.

The input files for all the simulations of this work are deposited in PLUMED-NEST (The PLUMED Consortium, 2019) as plumID:21.014.

RESULTS

Metadynamics and M&M simulations, using either PBMetaD or conventional MetaD, were performed to understand: 1) the statistical precision achievable by different Metadynamics setups; 2) the role played by enhanced sampling in the integration of experimental information in MD simulation by Metainference.

Assessing the Statistical Precision of Metadynamics Simulations

PBMetaD or conventional MetaD, was used to simulate the folding and unfolding of chignolin close to the transition temperature and to compute the free energy and the equilibrium population related to its three main conformational states (**Figure 1**). In particular we focused our attention on the ability to correctly estimate the errors associated to these calculations. Estimating statistical errors in enhanced sampling MD of large systems is a relevant problem because of their high computational cost. Previous works have already noted the importance of running multiple replicates, alternatively block-averaging can be used to estimate errors taking into accounts the time-correlated nature of MD. Here we compare statistical errors estimated from replicates with those resulting from block-averaging. In **Figure 2** we rebuilt a free-energy profile as function of an unbiased collective variable, the RMSD (computed over the main chain plus the Cβ atoms) with respect to a reference folded state of chignolin, and we estimated the population of three minima: folded (Min 1, RMSD ≤1.9 Å), misfolded (Min 2, 1.9 Å ≤ RMSD ≤3.0 Å) and unfolded (Min 3, RMSD >3.0 Å, see **Figure 1**). The error of each simulation is estimated using block-averaging. Furthermore, averages and errors are obtained by the triplicates, where the average free energy of bin b is computed as $F_b = -k_B T \log(p_b)$ and the associated errors are estimated as $err_{F_b} = \frac{1}{\sqrt{3}} k_B T \frac{\sigma_{p_b}}{p_b}$, with p_b being the average probability of the bin computed over the triplicates and σ_{p_b} its standard deviation.

Qualitatively, the resulting free energies display a good overlap both within the triplicates and when comparing the three simulation setups (**Figure 2**). Major deviations are mainly located in the high energy regions ($>2k_BT$). Nevertheless, we note that the variability among simulations strongly affect the population of the three minima, leading to differences for the folded minimum from less than 10% for PB4 simulations to ~20% for ME2 simulations. The populations estimated by averaging over the replicates are more precise and in quantitative agreement among the three simulation setups stressing once again the

TABLE 2 | For each replicate of the PB20, PB4, and ME2 simulations are reported: 1) the number of transitions per microsecond from the folded (F) to the unfolded (U) state and vice versa; 2) the percentage of the effective frames, NF_{eff}, over the total number of frames (NF).

		Transition per μs		NF_{eff}/NF (%)
		U -> F	F -> U	
PB20	Run 1	2.2 ± 0.4	2.8 ± 0.3	37%
	Run 2	2.2 ± 0.4	1.9 ± 0.3	39%
	Run 3	1.9 ± 0.3	1.8 ± 0.3	39%
	Average	2.1 ± 0.1	2.2 ± 0.3	38 ± 0 1%
PB4	Run 1	2.0 ± 0.4	2.8 ± 0.5	22%
	Run 2	2.1 ± 0.4	1.8 ± 0.4	20%
	Run 3	2.2 ± 0.4	2.5 ± 0.5	26%
	Average	2.1 ± 0.1	2.4 ± 0.3	23 ± 0 2%
ME2	Run 1	1.6 ± 0.6	2.4 ± 0.5	2.7
	Run 2	1.2 ± 0.4	1.3 ± 0.4	4.1
	Run 3	1.1 ± 0.3	1.4 ± 0.4	3.0
	Average	1.3 ± 0.2	1.7 ± 0.4	3.3 ± 0.4%

NF_{eff} *is computed as:* $NF_{eff} = (\sum_{i=1}^{NF} w_i)^2 / \sum_{i=1}^{NF} w_i^2$, *where* w_i *is the weight associated to each frame. The average and the standard error over the triplicates are also reported.*

importance of running independent simulations. Reassuringly, errors calculated by block-averages (comparing the free-energy obtained from blocks of lengths in the range 30 ns–1 μs), correctly estimate the variability observed within the triplicates (**Figure 2**), with ME2 simulations showing the largest error in the set. Free-energies and errors estimated as a function of the other biased and unbiased CVs (**Supplementary Figures S2, S3**) display a consistent behavior. Furthermore, we compared our results with a reference 40 μs long simulated tempering simulation published in Piana et al. (2020), showing that the populations of the minima are quantitatively in agreement with those obtained averaging over our replicates (**Supplementary Figure S4**).

To rationalize the higher variability observed in ME2 simulations with respect to the PB20 and PB4 simulations, we calculated the number of transitions between the folded and unfolded state as well as the effective statistics, i.e., the fraction of frames actually contributing to our statistical observations (cf. **Table 2**). While the number of transitions per microsecond is slightly lower in ME2 with respect to PB20 and PB4, the effective number of frames is surprisingly low for all simulations and dramatically so for ME2 (**Table 2**). This is likely due to the wider exploration of the conformational space by MetaD, that spends more time in high free-energy regions, thus reducing the fraction of frames that actually populate the most relevant conformations (see also **Supplementary Figure S5**). This reminds us that enhanced sampling is not a free lunch: indeed, while favoring the exploration of a wider conformational space, it reduces the statistical precision of the low free-energy regions reconstruction. A similar observation can explain the difference in the effective frames observed between PB20 and PB4. To improve the statistics one possibility is to fine tune and decrease the bias factor employed for well-tempered Metadynamics (here it was 10 for all setups, a very common value for simulations of biological molecules) and thus focus the sampling only within regions of interest.

Metadynamics Metainference: Enhanced Sampling and Conformational Averaging

The poor statistics characterizing our Metadynamics simulations, and ME2 in particular, raises issues about their combination with Metainference, in particular when the experimental data to be integrated represent averages over multiple conformational states. To test this effect, we performed 4 M&M simulations with the amber99sb-ildn force field, using as restraints synthetic SAXS data derived from reference 40 μs long DES-amber trajectory. The choice of SAXS is due to the ability of this technique to capture the overall size and shape of the molecules, thus being particularly sensitive to the equilibrium between the different conformational states (see **Supplementary Figure S6**); herein the use of synthetic data allows to avoid experimental and forward model errors and to focus on the effect of Metadynamics on the number of effective replicas.

We firstly performed a prior 10 replicas PB4 simulation, with the amber99sb-ildn force field, verifying that the resulting conformational ensemble and the back-calculated SAXS profiles are significantly far from the reference DES-amber simulation (**Supplementary Figures S7, S8**). Then we tested four different SAXS-restrained M&M setup, either using PB4 or ME2 with 10 or 100 replicas (**Table 1**). The inclusion of SAXS restraints improve, as expected, the agreement with the input scattering profile I_{ref}. We found that the relative error of the calculated SAXS intensity, defined as $R_{factor} = \left|\frac{I - I_{ref}}{I_{ref}}\right| \times 100$, is in the range 0.4–1.0% (**Supplementary Figure S9**), representing a significant improvement with respect to the prior amber99sb-ildn simulations (R_{factor} = 6.7%, **Supplementary Figure S8**). Also, we observe that in all the cases the input profile is well in agreement with the one back calculated from the simulations within the error estimated by Metainference (**Supplementary Figure S9**). Nevertheless, it is worth noting that the estimated errors differ in the four simulations as it will be discuss later, thus slightly impacting the extent of the agreement with the input data: i.e., larger errors result in slightly worse agreement as in ME2 (10r), while smaller errors lead to better agreement as in PB4 (100r).

Next, for each of the four simulations we monitored the number of effective replicas as a function of the simulations time. With the same number of actual replicas, the PB4 setup displayed more effective replicas than ME2 (**Figure 3**): the average NR_{eff} in PB4 (10r) was two times larger, 4.3, than in ME2 (10r), 2.0, and it was more than three times larger in PB4 (100r) than in the ME2 (100r) setup (NR_{eff} of 35 vs. 10). This difference impacts on the resulting conformational ensemble (**Figure 4**; **Supplementary Figure S10**). A striking effect is seen for the ME2 (10r) simulation (NR_{eff} = 2.0), in which the inclusion of SAXS data caused a strong distortion of the original ensemble leading to the formation of a new main minimum and a clear deviation from the target also in the low free-energy regions. This is consequence of the fact that, in time, we are forcing approximately NR_{eff} = 2.0 conformations to fit SAXS data that can only be explained by larger conformational ensembles. Importantly, the reconstructed ensembles become increasingly close to the target for larger values of NR_{eff}, with the best

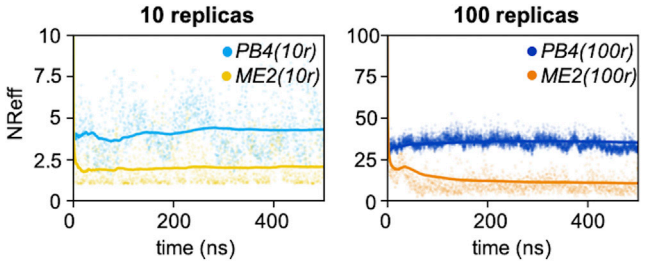


FIGURE 3 | Number of effective replicas as a function of the simulation time for PB4 (10r), ME2 (10r), PB4 (100r) and ME2 (100r) simulations. The dots represent the value at the exact time, while the straight lines indicate the cumulative averages.

agreement obtained for PB4 (100r) (NR_{eff} = 35). We observe that this does not mean that PB4 allows a better agreement than ME2 in general, but it suggests that to obtain a comparable agreement more replicas are needed when using the ME2 setup.

The number of effective replicas also affect the model errors sampled by Metainference. As expected, we observed a direct effect on σ_i^{SEM} (**Supplementary Figure S11**), where less effective replicas resulted in larger errors: indeed σ_i^{SEM} represents the uncertainty related to conformational averaging, which is consequence of the fact that we are using a small number of conformations (NR_{eff}) to back calculate experimental data, that are ideally obtained as averages over an infinite number of conformations. We also noted an indirect effect on $\sigma_{r,i}^{B}$, where again fewer effective replicas resulted in larger errors (**Supplementary Figure S11**). This is likely due to a better agreement with input data for larger NR_{eff} allowed by the larger number of conformations on which averaging is performed. Overall, this implies smaller model errors for simulations with higher number of effective replicas, where the total model error is computed as $\sigma_{r,i}^2 = \sigma_i^{SEM2} + \sigma_{r,i}^{B2}$. σ_i^{SEM} sets the lower limit for the model error and measures the impact of conformational averaging for the ith data point. We suggest that the relative error σ_i^{SEM}/d_i, where d_i is the ith experimental data, can indicate whether the number of effective replicas (and consequently the number of simulated replicas) is sufficient to capture the conformational variability needed to correctly interpret the corresponding data (**Supplementary Figure S11**). Our results suggest that a relative error lower than 5–10% could be sufficient to achieve a reasonable agreement with the target ensemble. We also note that the relative errors provide information about the sensitivity of different data points to conformational averaging.

These results underlie the importance of using a sufficient large number of replicas in M&M simulations, taking particular care of the number of effective frames in time, which depends on the enhanced sampling technique used, including the employed CVs, the investigated system as well as the specific experimental observable.

CONCLUSION

Reliability of MD simulations depends on their statistical precision and experimental accuracy. M&M aims to achieve both by coupling enhanced sampling and Bayesian Inference. Here, we assessed the performance of different MetaD setups, optionally coupled with Metainference, using as test system chignolin. Chignolin is a 10 residues peptide that is able to populate three different conformational states with diverse degrees of compactness and folding thus representing a simple but realistic test case.

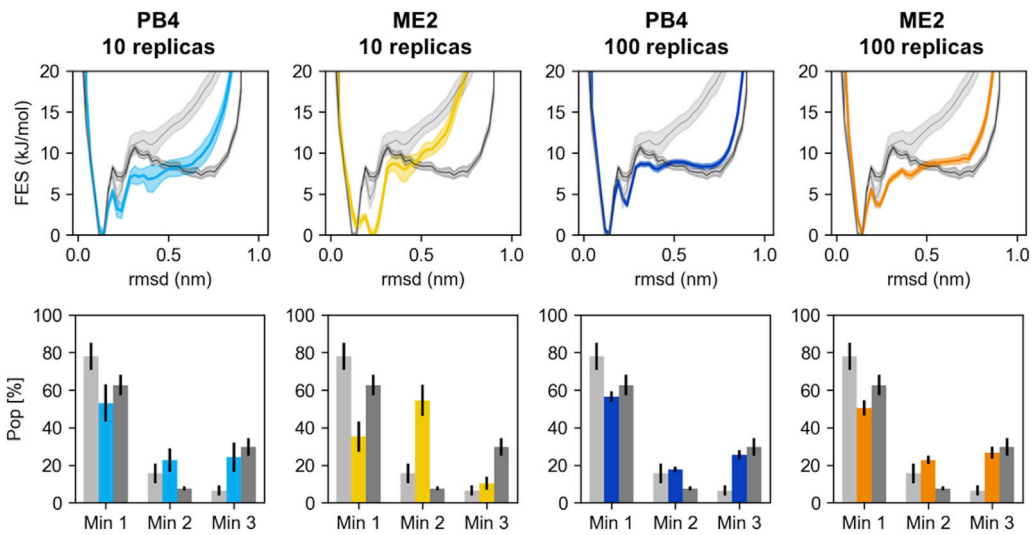


FIGURE 4 | The RMSD free energies **(top panels)** and the population of the three main chignolin minima **(bottom panels)** are represented for different simulations (PB4 or ME2, with either 10 or 100 replicas). The results from the prior simulation, performed with the amber 99sb-ildn force field and no SAXS restraints, are represented in light grey. The results from the reference simulation, performed with the DES-amber force field and whose back-calculated SAXS intensities were used as target, are represented in dark grey. The errors are estimated via block-average analysis; the free energies are shifted to set their minimum to 0.

In order to assess the statistical precision achievable by diverse enhanced sampling simulations we run three independent replicates for three different Metadynamics setups, either employing PBMetaD or traditional MetaD coupled with multiple walkers and using different combinations of CVs. We showed that block averaging is a robust technique to estimate statistical error, being always a slight overestimation of the standard error computed from the comparison of the triplicates. Still, we observed quite strong deviations in the population values when compared among replicates, suggesting that quantitative conclusions should be drawn with care from a single simulation. Importantly, when using averages calculated over the triplicates, we found an optimal agreement among the different setups, both concerning the free-energies and the population estimates. This quantitative agreement is maintained also with an independent reference simulation (Piana et al., 2020), performed with simulated annealing. Thus, as long as the simulations are well converged and possibly properties are evaluated as averages over independent copies of the simulations, the choice of the enhanced sampling technique does not influence the overall results. These observations support the idea that performing replicates, even if expensive, should become a more common practice, in particular when statistical precision is a core message.

Experimental accuracy can be obtained by Metainference *via* the introduction of restraints toward a set of experimental data. Different issues could affect the success of Metainference simulations, including the quality and quantity of experimental data (Löhr et al., 2017) and the quality of the forward model, as also discussed in this special issue (Ahmed et al., 2021). Here, we highlighted how the combination of MetaD and Metainference (M&M) could create an additional issue related to the number of effective replicas. In Metainference, to restrain the simulation, the experimental data are compared with the same back-calculated observables, averaged over the replicas: this is done to account for the conformational heterogeneity of the system. Nevertheless, the coupling with MetaD, while helping in accelerating the sampling and achieving better statistical precision, could reduce the number of effective replicas (NR_{eff}) on which this averaging is performed. Indeed, MetaD modulates the relative weights of the replicas, where some of them are found in low-energy areas (high probability) and other are in high energy regions (low relative weight). In this work, by performing M&M simulations with either PBMetaD or traditional MetaD setup and using 10 or 100 replicas, we showed how the number of effective replicas is extremely relevant for the reconstruction of conformational ensembles. A too small NR_{eff} leads to distortions of the prior ensemble that are very far from the desired target. To keep this effect under control we suggest monitoring the relative error caused by σ_i^{SEM}. The latter represents the statistical error we introduce when trying to capture the conformational heterogeneity underlying an experimental observable with a finite number of replicas. Also, we showed that in the context of M&M, PBMetaD could be preferred to traditional MetaD, as it results in a milder reduction in the number of actual replicas. Indeed, the number of replicas should be high enough to capture the conformational heterogeneity of the system as detected by an experimental observable while also compensating to the loss of effective frames resulting from the combination of Metainference with Metadynamics.

Concluding, enhanced sampling techniques and integrative techniques can generate precise and accurate conformational ensembles. Here we showed that well established enhanced sampling techniques provide robust results in particular when performing multiple independent simulations. Moreover, we improve our understanding of Metainference by suggesting how to optimally chose the number of simulated replicas needed to describe correctly the conformational heterogeneity of an ensemble.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

CP and CC designed the work, performed the simulations, analyzed the data and wrote the paper.

ACKNOWLEDGMENTS

We thank D. E. Shaw Research for sharing their molecular dynamics trajectory. We acknowledge PRACE for awarding us access to Piz Daint, at the Swiss National Supercomputing Centre (CSCS), Switzerland.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.694130/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

Ahmed, M. C., Skaanning, L. K., Jussupow, A., Newcombe, E. A., Kragelund, B. B., Camilloni, C., et al. (2021). Refinement of α-synuclein Ensembles against SAXS Data: Comparison of Force fields and Methods. *bioRxiv* 8, 654333. doi:10.1101/2021.01.15.426794

Allison, J. R. (2017). Using Simulation to Interpret Experimental Data in Terms of Protein Conformational Ensembles. *Curr. Opin. Struct. Biol.* 43, 79–87. doi:10.1016/j.sbi.2016.11.018

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* 100 (2), 020603. doi:10.1103/PhysRevLett.100.020603

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* 81 (8), 3684–3690. doi:10.1063/1.448118

Bernetti, M., Bertazzo, M., and Masetti, M. (2020). Data-Driven Molecular Dynamics: A Multifaceted Challenge. *Pharmaceuticals* 13 (9), 253. doi:10.3390/ph13090253

Best, R. B., Zheng, W., and Mittal., J. (2014). Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-specific Protein Association. *J. Chem. Theor. Comput.* 10 (11), 5113–5124. doi:10.1021/ct500569b

Bonomi, M., Camilloni, C., Cavalli, A., and Vendruscolo, M. (2016). Metainference: A Bayesian Inference Method for Heterogeneous Systems. *Sci. Adv.* 2 (1), e1501177. doi:10.1126/sciadv.1501177

Bonomi, M., Camilloni, C., and Vendruscolo, M. (2016a). Metadynamic Metainference: Enhanced Sampling of the Metainference Ensemble Using Metadynamics. *Sci. Rep.* 6 (1), 31232. doi:10.1038/srep31232

Bonomi, M., Heller, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of Protein Structural Ensemble Determination. *Curr. Opin. Struct. Biol.* 42, 106–116. doi:10.1016/j.sbi.2016.12.004

Bottaro, S., and Lindorff-Larsen, K. (2018). Biophysical Experiments and Biomolecular Simulations: A Perfect Match? *Science* 361 (6400), 355–360. doi:10.1126/science.aat4010

Branduardi, D., Bussi, G., and Parrinello, M. (2012). Metadynamics with Adaptive Gaussians. *J. Chem. Theor. Comput.* 8 (7), 2247–2254. doi:10.1021/ct3002464

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* 126 (1), 014101. doi:10.1063/1.2408420

Bussi, G., and Tribello, G. A. (2019). Analyzing and Biasing Simulations with PLUMED. *Biomol. Simul.* 2020, 529–578. doi:10.1007/978-1-4939-9608-7_21

Camilloni, C., and Pietrucci, F. (2018). Advanced Simulation Techniques for the Thermodynamic and Kinetic Characterization of Biological Systems. *Adv. Phys. X* 3 (1), 1477531. doi:10.1080/23746149.2018.1477531

Cavalli, A., Camilloni, C., and Vendruscolo, M. (2013). Molecular Dynamics Simulations with Replica-Averaged Structural Restraints Generate Structural Ensembles According to the Maximum Entropy Principle. *J. Chem. Phys.* 138 (9), 094112. doi:10.1063/1.4793625

Eshun-Wilson, L., Zhang, R., Portran, D., Nachury, M. V., Toso, D. B., Löhr, T., et al. (2019). Effects of α-tubulin Acetylation on Microtubule Structure and Stability. *Proc. Natl. Acad. Sci. USA* 116 (21), 10366–10371. doi:10.1073/pnas.1900441116

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* 103 (19), 8577–8593. doi:10.1063/1.470117

Fennen, J., Torda, A. E., and van Gunsteren, W. F. (1995). Structure Refinement with Molecular Dynamics and a Boltzmann-Weighted Ensemble. *J. Biomol. NMR* 6 (2), 163–170. doi:10.1007/BF00211780

Flyvbjerg, H., and Petersen, H. G. (1989). Error Estimates on Averages of Correlated Data. *J. Chem. Phys.* 91 (1), 461–466. doi:10.1063/1.457480

Grossfield, A., Patrone, P. N., Roe, D. R., Schultz, A. J., Siderius, D., and Zuckerman, D. M. (2019). Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *LiveCoMS* 1 (1), 5067. doi:10.33011/livecoms.1.1.5067

Heller, G. T., Aprile, F. A., Michaels, T. C. T., Limbocker, R., Perni, M., Ruggeri, F. S., et al. (2020). Small-molecule Sequestration of Amyloid-β as a Drug Discovery Strategy for Alzheimer's Disease. *Sci. Adv.* 6 (45), eabb5924. doi:10.1126/sciadv.abb5924

Hess, B., Bekker, H., Berendsen, H. J. C., Fraaije, J. G. E. M., and JohannesFraaije, G. E. M. (1997). LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* 18 (12), 1463–1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

Honda, S., Akiba, T., Kato, Y. S., Sawada, Y., Sekijima, M., Ishimura, M., et al. (2008). Crystal Structure of a Ten-Amino Acid Protein. *J. Am. Chem. Soc.* 130 (46), 15327–15331. doi:10.1021/ja8030533

Hopkins, C. W., Le Grand, S., Walker, R. C., and Roitberg, A. E. (2015). Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theor. Comput.* 11 (4), 1864–1874. doi:10.1021/ct5010406

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., et al. (2017). CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* 14 (1), 71–73. doi:10.1038/nmeth.4067

Invernizzi, M., Piaggi, P. M., and Parrinello, M. (2020). Unified Approach to Enhanced Sampling. *Phys. Rev. X* 10 (4), 041034. doi:10.1103/PhysRevX.10.041034

Jussupow, A., Messias, A. C., Stehle, R., Geerlof, A., Solbak, S. M. Ø., Paissoni, C., et al. (2020). The Dynamics of Linear Polyubiquitin. *Sci. Adv.* 6 (42), eabc3786. doi:10.1126/sciadv.abc3786

Köfinger, J., Różycki, B., and Hummer, G. (2019). Inferring Structural Ensembles of Flexible and Dynamic Macromolecules Using Bayesian, Maximum Entropy, and Minimal-Ensemble Refinement Methods. *Methods Mol. Biol.* 2022, 341–352. doi:10.1007/978-1-4939-9608-7_14

Kührová, P., De Simone, A., Otyepka, M., and Best, R. B. (2012). Force-Field Dependence of Chignolin Folding and Misfolding: Comparison with Experiment and Redesign. *Biophysical J.* 102 (8), 1897–1906. doi:10.1016/j.bpj.2012.03.024

Laio, A., and Parrinello, M. (2002). Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci.* 99 (20), 12562–12566. doi:10.1073/pnas.202427399

Lindorff-Larsen, K. S. P., 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 (8), 1950–1958. doi:10.1002/prot.22711

Löhr, T., Jussupow, A., and Camilloni, C. (2017). Metadynamic Metainference: Convergence towards Force Field Independent Structural Ensembles of a Disordered Peptide. *J. Chem. Phys.* 146 (16), 165102. doi:10.1063/1.4981211

Mark, P., and Nilsson, L. (2001). Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A.* 105 (43), 9954–9960. doi:10.1021/jp003020w

McCarty, J., and Parrinello, M. (2017). A Variational Conformational Dynamics Approach to the Selection of Collective Variables in Metadynamics. *J. Chem. Phys.* 147 (20), 204109. doi:10.1063/1.4998598

Mendels, D., Piccini, G., and Parrinello, M. (2018). Collective Variables from Local Fluctuations. *J. Phys. Chem. Lett.* 9 (11), 2776–2781. doi:10.1021/acs.jpclett.8b00733

Paissoni, C., Jussupow, A., and Camilloni, C. (2020). Determination of Protein Structural Ensembles by Hybrid-Resolution SAXS Restrained Molecular Dynamics. *J. Chem. Theor. Comput.* 16 (4), 2825–2834. doi:10.1021/acs.jctc.9b01181

Paissoni, C., Jussupow, A., and Camilloni, C. (2019). Martini Bead Form Factors for Nucleic Acids and Their Application in the Refinement of Protein-Nucleic Acid Complexes against SAXS Data. *J. Appl. Cryst.* 52 (2), 394–402. doi:10.1107/S1600576719002450

Parrinello, M., and Rahman, A. (1981). Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* 52 (12), 7182–7190. doi:10.1063/1.328693

Pfaendtner, J., and Bonomi, M. (2015). Efficient Sampling of High-Dimensional Free-Energy Landscapes with Parallel Bias Metadynamics. *J. Chem. Theor. Comput.* 11 (11), 5062–5067. doi:10.1021/acs.jctc.5b00846

Piana, S., Donchev, A. G., Robustelli, P., and Shaw, D. E. (2015). Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* 119 (16), 5113–5123. doi:10.1021/jp508971m

Piana, S., Robustelli, P., Tan, D., Chen, S., and Shaw, D. E. (2020). Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes. *J. Chem. Theor. Comput.* 16 (4), 2494–2507. doi:10.1021/acs.jctc.9b00251

Raiteri, P., Laio, A., Gervasio, F. L., Micheletti, C., and Parrinello, M. (2006). Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics†. *J. Phys. Chem. B* 110 (8), 3533–3539. doi:10.1021/jp054359r

Rieping, W. (2005). Inferential Structure Determination. *Science* 309 (5732), 303–306. doi:10.1126/science.1110428

Robustelli, P., Piana, S., Shaw, D. E., and Shaw, P. (2018). Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. USA* 115 (21), E4758–E4766. doi:10.1073/pnas.1800690115

Spiwok, V., Sucur, Z., and Hosek, P. (2015). Enhanced Sampling Techniques in Biomolecular Simulations. *Biotechnol. Adv.* 33 (6), 1130–1140. doi:10.1016/j.biotechadv.2014.11.011

Sugita, Y., and Okamoto, Y. (1999). Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* 314 (1–2), 141–151. doi:10.1016/S0009-2614(99)01123-9

Sultan, M. M., and Pande, V. S. (2018). Automated Design of Collective Variables Using Supervised Machine Learning. *J. Chem. Phys.* 149 (9), 094106. doi:10.1063/1.5029972

Sultan, M. M., and Pande, V. S. (2017). TICA-metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables. *J. Chem. Theor. Comput.* 13 (6), 2440–2447. doi:10.1021/acs.jctc.7b00182

The PLUMED Consortium (2019). Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* 16 (8), 670–673. doi:10.1038/s41592-019-0506-8

Tiwary, P., and Berne, B. J. (2016). Spectral Gap Optimization of Order Parameters for Sampling Complex Molecular Systems. *Proc. Natl. Acad. Sci. USA* 113 (11), 2839–2844. doi:10.1073/pnas.1600917113

Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. (2014). PLUMED 2: New Feathers for an Old Bird. *Comp. Phys. Commun.* 185 (2), 604–613. doi:10.1016/j.cpc.2013.09.018

Wang, Y., Lamim Ribeiro, J. M., and Tiwary, P. (2020). Machine Learning Approaches for Analyzing and Enhancing Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* 61, 139–145. doi:10.1016/j.sbi.2019.12.016

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: 04 June 2021
doi: 10.3389/fmolb.2021.671011

Automatic Bayesian Weighting for SAXS Data

Yannick G. Spill[1†], *Yasaman Karami*[1†], *Pierre Maisonneuve*[2,3,4], *Nicolas Wolff*[3,5] *and Michael Nilges*[1*]

[1]Department of Structural Biology and Chemistry, Structural Bioinformatics Unit, CNRS UMR 3528, Institute Pasteur, Paris, France, [2]Université Pierre et Marie Curie, Cellule Pasteur UPMC, Paris, France, [3]Department of Structural Biology and Chemistry, NMR of Biomolecules Unit, CNRS UMR 3528, Institute Pasteur, Paris, France, [4]Center for Molecular, Cell and Systems Biology, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada, [5]Department of Neuroscience, Current Address, Channel-Receptors Unit, CNRS UMR 3571, Institut Pasteur, Paris, France

Small-angle X-ray scattering (SAXS) experiments are important in structural biology because they are solution methods, and do not require crystallization of protein complexes. Structure determination from SAXS data, however, poses some difficulties. Computation of a SAXS profile from a protein model is expensive in CPU time. Hence, rather than directly refining against the data, most computational methods generate a large number of conformers and then filter the structures based on how well they satisfy the SAXS data. To address this issue in an efficient manner, we propose here a Bayesian model for SAXS data and use it to directly drive a Monte Carlo simulation. We show that the automatic weighting of SAXS data is the key to finding optimal structures efficiently. Another key problem with obtaining structures from SAXS data is that proteins are often flexible and the data represents an average over a structural ensemble. To address this issue, we first characterize the stability of the best model with extensive molecular dynamics simulations. We analyse the resulting trajectories further to characterize a dynamic structural ensemble satisfying the SAXS data. The combination of methods is applied to a tandem of domains from the protein PTPN4, which are connected by an unstructured linker. We show that the SAXS data contain information that supports and extends other experimental findings. We also show that the conformation obtained by the Bayesian analysis is stable, but that a minor conformation is present. We propose a mechanism in which the linker may maintain PTPN4 in an inhibited enzymatic state.

Keywords: SAXS, bayesian scoring, automatic weighting, inferential structure determination, PTPN4, allosteric regulation, conformational dynamics

OPEN ACCESS

Edited by:
Gregory Bowman, Washington University School of Medicine in St. Louis, United States

Reviewed by:
Kresten Lindorff-Larsen, University of Copenhagen, Denmark
Jochen Hub, Saarland University, Germany

***Correspondence:**
Michael Nilges
michael.nilges@pasteur.fr

†These authors have contributed equally to this work and share first authorship.

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

Received: *22 February 2021*
Accepted: *14 May 2021*
Published: *04 June 2021*

Citation:
Spill YG, Karami Y, Maisonneuve P, Wolff N and Nilges M (2021) Automatic Bayesian Weighting for SAXS Data. Front. Mol. Biosci. 8:671011. doi: 10.3389/fmolb.2021.671011

1 INTRODUCTION

Integrative structural biology uses multiple techniques to determine three-dimensional structures of large, potentially flexible complexes of biological macromolecules. Typically, structures of the individual components (e.g., individual domains or proteins) are known but the overall arrangement of the components is to be determined. Despite their relatively low information content, Small Angle Scattering [Small Angle X-ray Scattering (SAXS), or Small Angle Neutron Scattering (SANS)] experiments play an important role, since they are performed in solution, and can provide crucial conformational information on the arrangement of individual components.

In order to incorporate SAXS data, many approaches generate poses of the components and then use the experimental data to filter solutions by means of a χ^2 criterion [e.g., Mareuil et al. (2007); Yang et al. (2010); Rozycki et al. (2011)]. For a larger number of degrees of freedom, or when a large conformational space needs to be covered, this becomes computationally intensive, and one might miss structures that satisfy the data. Preferentially, one would like to employ methods that can use the data directly as restraints to drive the structure calculation, since they should converge faster to conformations satisfying the data. In methods that refine directly against the data, definite choices on unmeasurable model parameters must be made before the minimization. Examples for such parameters are the scale factor between the experimental and the back-calculated data, and the quality or consistency of the data, which has a relationship to the weight on the data employed during the calculation (data with lower quality should get a lower weight). Yet, the optimal weight one should put on the data is never known beforehand. These parameter choices have important consequences, and even more so if SAXS data are to be used together with other data, for which similar problems exist.

When modeling structures from experimental data, appropriate relative weighting is of particular importance. In crystallography, for example, the free R-value Brünger (1992) is often used to find suitable values for unknown parameters such as the weight on the experimental data. This becomes rapidly cumbersome if more than one value needs to be optimized, and it is hardly an option for data with low information content such as SAXS or SANS.

A more powerful and statistically more accurate solution to this problem can be obtained in the context of a Bayesian treatment of the structure determination problem. We previously developed the Bayesian framework we called "Inferential Structure Determination" (ISD) and applied it to Nuclear Magnetic Resonance (NMR) data Rieping et al. (2005). We showed that the Bayesian formalism converges better than standard minimization strategies Rieping et al. (2005). We also showed that an optimal weight on a χ^2 type experimental term can be obtained from a 3D structure and the data Habeck et al. (2006), and that this weight can be optimized simultaneously with the structure Nilges et al. (2008), Bernard et al. (2011). More recently, we extended the concept of ISD and Bayesian weight optimization to the treatment of cross-linking mass spectrometry data Ferber et al. (2016) and electron microscopy Bonomi et al. (2019).

In this paper, we develop a Bayesian framework for the analysis of SAXS data. This model allows us to automatically weight the SAXS data based on its agreement with other structural modeling terms. The modeling is performed in several stages, adding additional detail at each stage, starting with rigid body motions of protein domains, and subsequently adding and sampling conformations of the linker and the termini. This is followed by extensive unbiased molecular dynamics (MD) simulation starting from the optimal structure. We apply the new formalism and modelling strategy to the determination of the structure of the tandem domain of the protein PTPN4. This is a good test case since, due to its flexible linker, several conformations may be simultaneously present and influence the measured SAXS data, which hampered previous attempts to obtain useful insights with more standard approaches to interpret SAXS data obtained for this protein.

The protein PTPN4 belongs to the non-receptor protein tyrosine phosphatase (PTP) family. It is involved in various biological processes such as T-cell signalling, learning, spatial memory and cerebellar synaptic plasticity Kina et al. (2007), Kohda et al. (2013), Young et al. (2008). PTPN4 also regulates cell proliferation and presents an anti-apoptotic function Gu et al. (1996), Préhaud et al. (2010), Zhou et al. (2013), Zhang et al. (2019). PTPN4 is a large modular protein containing a N-terminal FERM (Band 4.1, Ezrin, radixin, and Moesin) domain, a PDZ (PSD-95/Dlg/ZO-1) domain and a C-terminal catalytic tyrosine phosphatase domain. The phosphatase is cleaved in the cell, leading to enzyme activation and its active form consists of the PDZ and PTP domains connected by a linker Gu and Majerus (1996). We previously demonstrated that the catalytic activity of the PTP domain is inhibited by the PDZ domain, while the binding of a ligand to the PDZ releases this auto-inhibition and activates the phosphatase Maisonneuve et al. (2014). A biochemical study suggests that this mechanism of regulation of PTPN4 allows for the specific dephosphorylation of cellular partners such as the mitogen-activated protein kinase (MAPK) $p38\gamma$ recruited through the PDZ domain of the phosphatase Maisonneuve et al. (2016). The importance of the PDZ domain for PTPN4 is further supported by the fact that the G protein of an attenuated rabies virus strain target this domain to deregulates PTPN4 phosphatase function and ultimately causes neuronal cell death Préhaud et al. (2010), Babault et al. (2011), Caillet-Saguy et al. (2015).

However, the structural mechanism by which the PDZ domain modulates the activity of the phosphatase domain remains elusive. We showed that a conserved hydrophobic patch in the linker connecting the PDZ and the PTP domains is involved in the communication between the two domains and participates in the phosphatase's regulation Caillet-Saguy et al. (2017). NMR and SAXS characterization of the PDZ-PTP domains of PTPN4 showed that the tandem adopts a compact conformation compatible with inter-domain interactions. However, no interaction was detected by NMR between the phosphatase domain and either the PDZ domain or the unstructured and flexible linker Maisonneuve et al. (2014). This suggests that the compact conformation of the PDZ-PTP domains is stabilized by fuzzy intramolecular interactions. Interestingly, ligand binding to the PDZ domain disrupts the transient interactions of the PDZ domain and the linker with the phosphatase domain. Ligand binding to the PDZ induces dynamic rearrangements of the two domains, resulting in the activation of the phosphatase domain Maisonneuve et al. (2014).

The Bayesian SAXS treatment generates a model of the conformations adopted by the PDZ, linker and phosphatase of PTPN4. This model allows us to propose a mechanism by which the linker can regulate the PTPN4 activity. The structure we obtain is based on the implicit assumption that an ensemble covering a small volume of conformational space can explain the SAXS data. We therefore used the MD simulations to investigate the conformational dynamics of PTPN4 and showed that the proposed preferential relative orientation of the two domains and

the linker is stable and corresponds best to the SAXS data. However, the simulations sample other orientations of two domains and the linker, albeit with a worse fit to the SAXS data. By using machine learning and a genetic algorithm we test combinations of structures from the MD trajectories and obtain a dynamic model of PTPN4 that optimally fits the SAXS data.

2 RESULTS

2.1 Bayesian Small Angle X-ray Scattering Restraint Term

In Bayesian modeling Rieping et al. (2005), one directly evaluates Bayes' equation

$$p(X,\sigma,\xi|B,D) \propto p(X|B)p(\sigma)p(\xi)p(D|X,\sigma,\xi) \quad (1)$$

where X is the 3D structure, σ is a parameter quantifying the deviation of the back-calculated data from the experimental data, and ξ stands for any other unknown parameters that one needs to model the data from the structure. B is the background information that we have on the structure, which allows us to evaluate the probability of a structure in absence of experimental data, for example, a molecular dynamics force field. To evaluate the discrepancy of the calculated data from the experimental data, we need a forward model $m(X)$ to calculate the intensities $\mathcal{I} = m(X)$ from a structure X. We used the FoXS model Schneidman-Duhovny et al. (2013), which has, in addition to a scale factor γ, two parameters c_1 and c_2, where c_1 is the scaling of the atomic radius used to adjust the total excluded volume of the atoms, and c_2 is used to adjust the difference between the density of the hydration layer and the bulk water.

As derived in detail in the Appendix, the negative log likelihood is

$$-\log p(\mathbf{I}|X,\gamma,c_1,c_2,\sigma^2) = \frac{M}{2\sigma^2}\chi^2 + M\log(\sigma) \quad (2)$$

$$\chi^2 \equiv \frac{1}{M}\sum_{i=1}^{M}\left(\frac{I(q_i) - \gamma m(X,q_i,c_1,c_2)}{s(q_i)}\right)^2 \quad (3)$$

where $\mathbf{I}$ is the experimental intensity, M is the number of points in the SAXS profile, q_i is the momentum transfer $q = (4\pi\sin(\theta))/\lambda$, with scattering angle θ and X-ray beam wavelength λ. $s(q_i)$ is the experimental uncertainty of the SAXS profile at q_i estimated from merging multiple experimental profiles.

2.2 Application to Protein Tyrosine Phosphatase Non-Receptor 4

To illustrate the Bayesian SAXS score, we perform exhaustive sampling of the conformational space of the PDZ and PTP domains of PTPN4, which for simplicity we call PTPN4. The PDZ (92 residues) and PTP (275 residues) domains are connected by a linker of 34 residues, and flanked by N-terminal (13 residues) and C-terminal (13 residues) sequences. The structures of individual domains are known Babault et al. (2011), Barr et al. (2009). However, the linker and the termini are highly flexible as monitored by NMR Maisonneuve et al. (2014). They thus prevented the determination by X-ray crystallography of the overall organization of the two domains of PTPN4 tethered by the linker.

To efficiently characterize the structural conformation of PTPN4 by a Bayesian SAXS score, we subdivided the problem into three subsequent stages (**Figure 1**). First, the linker and the termini were removed and the conformational space was explored with rigid body movements of the folded domains. Second, linker and termini were added, while keeping the domains fixed. Third, the whole structure was further refined with rigid body movements for the two domains and flexible backbones for the linker and the termini. In all three stages, we used **Eqs. 2**,**3** to incorporate the SAXS profile of PTPN4. Volume exclusion was used to produce physically realistic structures.

2.2.1 Rigid Body Docking

We started with 64 parallel simulations by placing the PDZ domain randomly around the PTP domain (without the linker and termini), avoiding physical contact between the two proteins (**Figures 2A,B**). The simulations rapidly converged to two distinct sets of conformations in which the PDZ domain (**Figure 2C**) is located on either of the two most distant points of the phosphatase domain, each subdivided in two further conformations (**Figures 2D,E**). In these conformations, the α2-helix of the PDZ domain is roughly aligned with the main axis of the phosphatase domain. This indicates a preferred orientation of the PDZ domain relative to the PTP domain.

To analyse the trajectories, we trained a self-organizing map (SOM) Bouvier et al. (2015). The subdivision of the two distinct sets of conformations into two further sets is clearly visible in the SOM, making it possible to define a total of four clusters (**Figure 3A**). Each cluster corresponds to one of the four possible combinations of position of the PDZ domain, and orientation of the α2-helix of the PDZ domain, with respect to the main axis of the phosphatase domain.

2.2.2 Linker Construction

We then extracted a clash-free conformation displaying the lowest χ^2 for each of the neurons of the SOM (**Figure 3A**). For every selected structure, we generated an average of 1,224 conformations for the linker and the termini sequences (see Methods). A Bayesian SAXS score was calculated for each of these structures. Depending on the pose, the linker raised or lowered the Bayesian SAXS score (**Figure 3B**). For each neuron we retained the structure with linker and termini displaying the lowest χ^2 (**Figure 3B**). Interestingly, the models with the lowest Bayesian SAXS scores are located in the two left clusters of the SOM corresponding to PDZ domains exclusively located on the side of the PTP β-sheet (**Figure 2E**). These clusters differ in a rotation of the PDZ by 180°. In these conformations, the attachment points of the linker to the PTP domain are located on the opposite side from where the PDZ domain is positioned. This implies that the linker passes over the surface of the phosphatase to reach the PDZ domain.

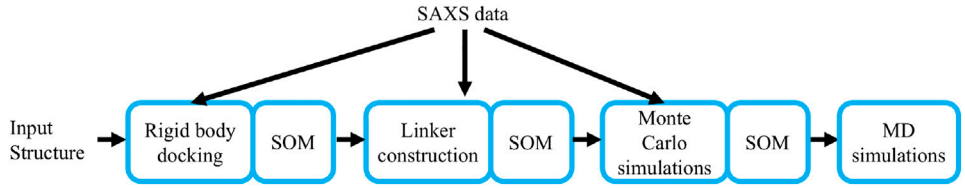


FIGURE 1 | The workflow of the method. The main steps of the algorithm are depicted: rigid body docking, linker construction, Monte Carlo simulations, and Molecular Dynamics (MD) simulations. The Small-angle X-ray scattering (SAXS) data is used to derive the first three steps.

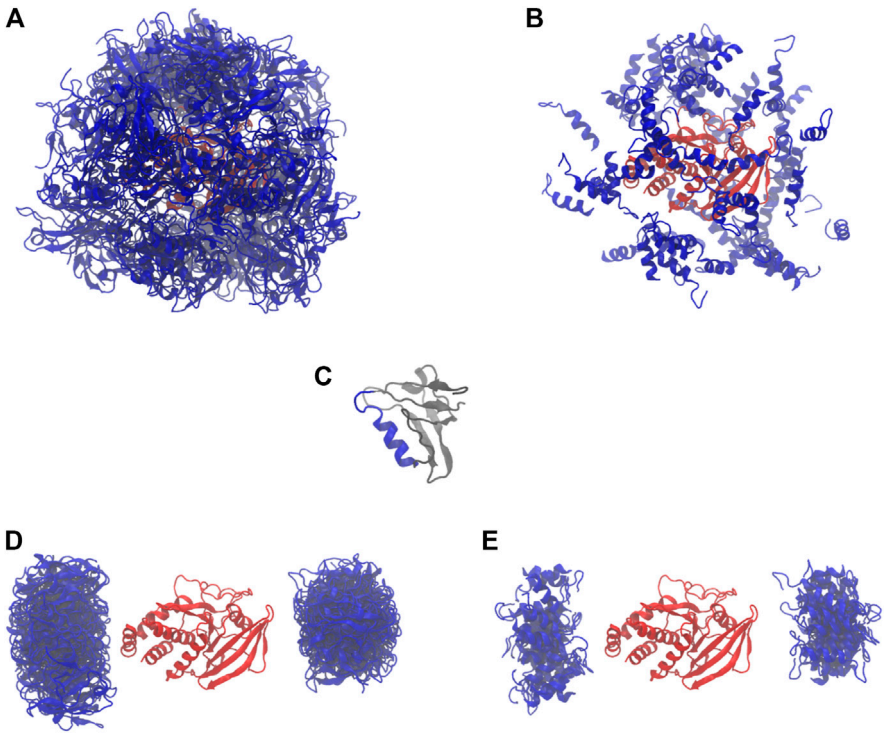


FIGURE 2 | Starting and final conformations of the 64 rigid body simulations. PDZ in blue, PTP in red. **(A)** Starting conformations (full PDZ). **(B)** Starting conformations (only α2-helix for PDZ). **(C)** PDZ, with α2-helix in blue. **(D)** Final conformations (full PDZ). **(E)** Final conformations (only α2-helix for PDZ).

2.2.3 Monte-Carlo Refinement

To further improve the sampling of the conformational space of the linker and termini, we performed an exhaustive refinement of the best structures of each neuron of the SOM map. We used a Monte-Carlo algorithm to sample the linker conformations in the dihedral angles of the linker and termini. As previously, we used only the Bayesian SAXS scoring term and volume exclusion to calculate the energy. This approach allowed the added residues and the domains to adjust jointly to the SAXS profile. The χ^2 significantly improved compared to the previous step for all clusters (**Figure 3C**), and values lowered by 36% on average to a range between 3 and 18. However, the trend in the four clusters remained the same. The structures with the lowest χ^2 scores after Monte-Carlo simulations belong to the cluster in the upper left corner of the self-organizing map as previously observed in the step of linker construction (**Figure 3C**). This indicates that the linker passes over the surface of the phosphatase for the structures which are in best agreement with the SAXS data.

The 10 conformations with the best final χ^2 after Monte-Carlo simulations, ranging from 2.5 to 2.9 are presented in **Figure 4**. In these 10 conformations, the linker is wrapped around the phosphatase domain and passes in close proximity to the catalytic site of the phosphatase domain. Interestingly, a conserved sequence in the linker (shown in green), involved in the allosteric regulation of PTPN4 Caillet-Saguy et al. (2017), is facing both the β5-loop-β6 region and the WPD loop, a conserved catalytic motif. This observation suggests a possible effect of the linker on these two regions.

2.2.4 Influence of the Weight Adjustment

During the calculations, the weight of the Bayesian SAXS score adjusted substantially (**Figure 5**). From the initial rigid body

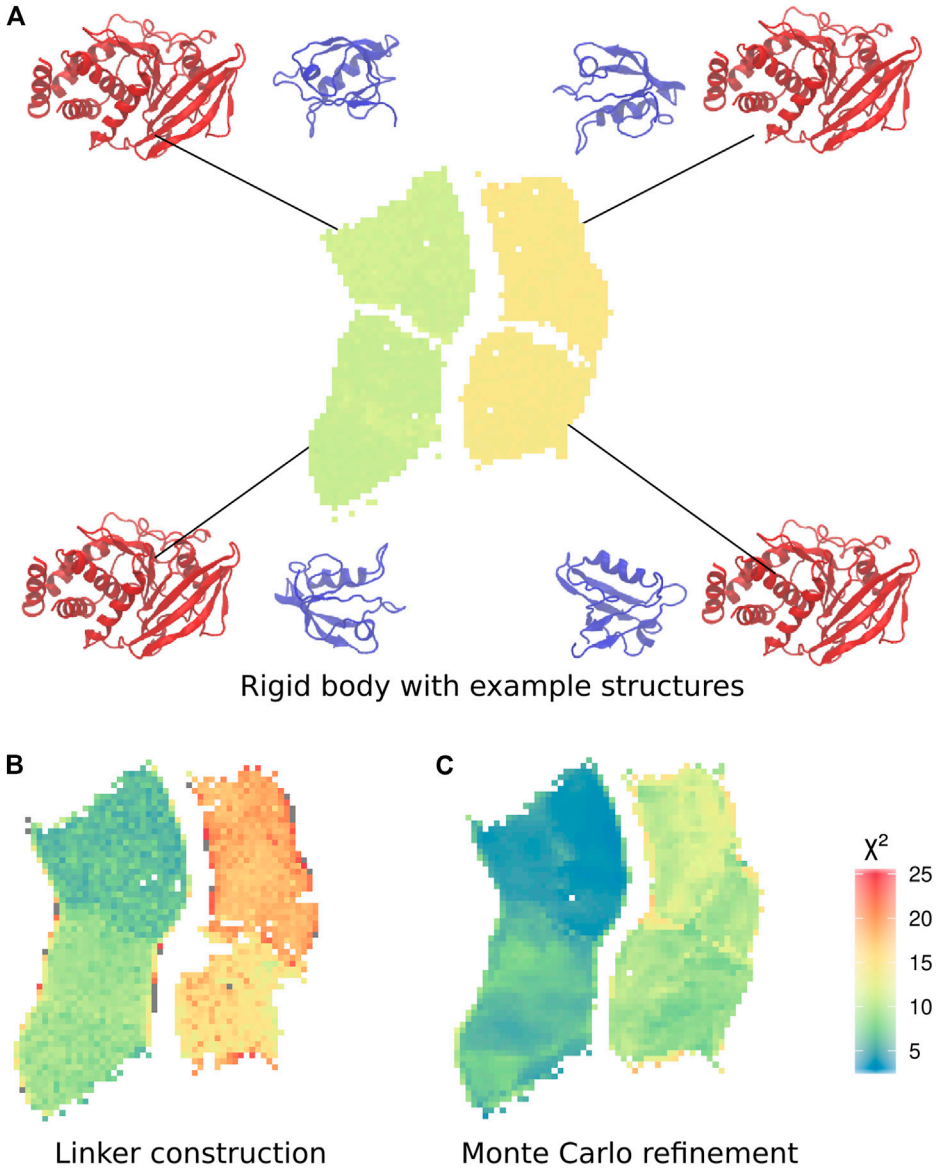


FIGURE 3 | Self-organizing maps (SOMs) of the three calculation stages. **(A)** Final conformations of the rigid body docking stage, coloured by χ^2. Only SOM neurons with at least one structure are shown throughout. For each cluster, an example structure shows the relative orientation of the PDZ with respect to the phosphatase domain. **(B)** Best χ^2 for each docking pose with linker added. **(C)** Average χ^2 of minimized linker conformations.

docking to the best structure after refinement, the weight was multiplied by 17. This means that the SAXS data was given 17 times more importance at the end of the procedure compared to the beginning. To see why this matters, we performed 20 linker refinement simulations with a fixed weight for the SAXS restraint, varying from 10^{-4}–10^{2} and compared it to 10 simulations using the Bayesian SAXS restraint. We then examined the χ^2 along the simulation step, for all replicates (**Figure 6**). All Bayesian SAXS simulations consistently reach low χ^2 values. In contrast, two limiting cases emerge in the fixed-weight simulations. When the weight is very large, agreement to the SAXS data is substantial, and the simulation quickly finds a local SAXS restraint minimum. Sometimes, conformers with can be obtained, but more often less optimal basins are targeted, with $\chi^2 \sim 10$ in this example. The Monte Carlo acceptance rate then drops to zero, and the simulation stops exploring new conformations. In contrast, when the weight is very small, the SAXS score has little influence. The simulation can scan conformational space easily, but it has no chance of finding structures in good agreement with the SAXS data.

2.3 Stability of the Optimal Conformation

2.3.1 Molecular Dynamics Simulations and Conformational Clustering

To further assess the stability of the optimal conformation obtained from the Bayesian analysis, we performed three MD

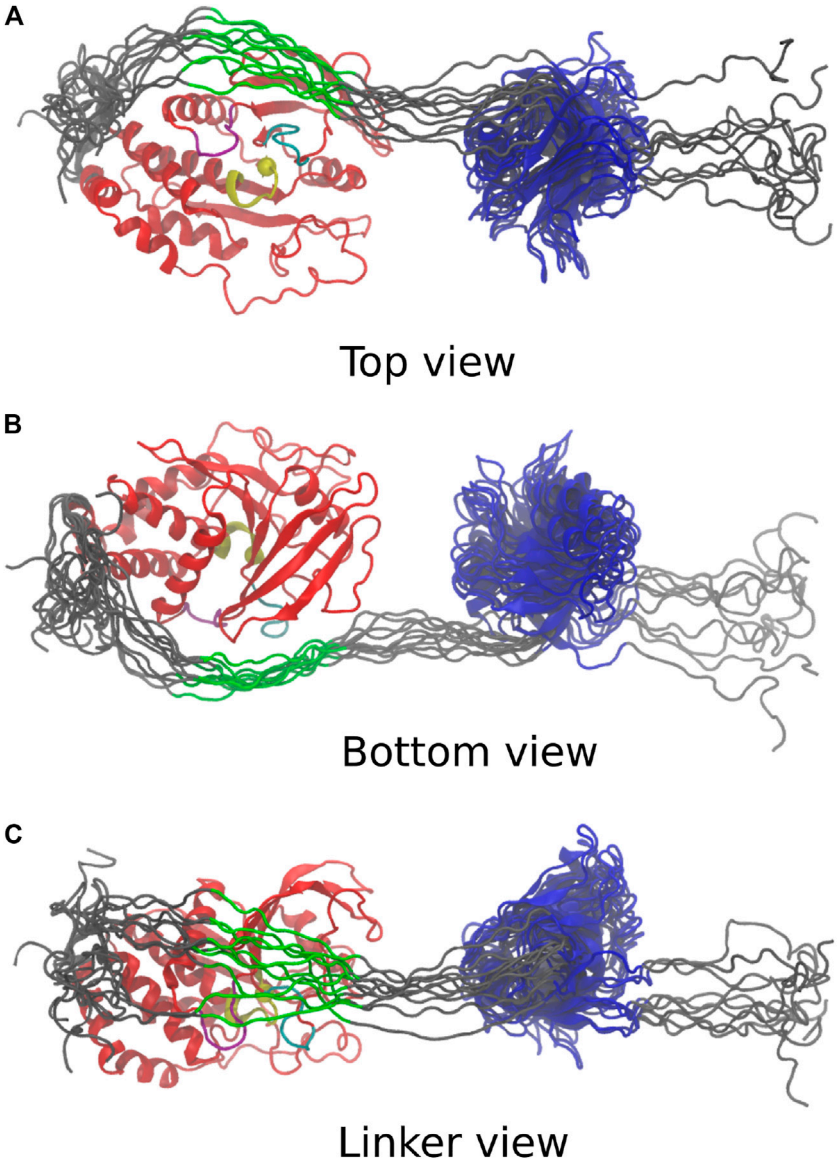


FIGURE 4 | Last frame of the top 10 simulations, aligned on the PTP domain. PTP: red; PTP loop and catalytic cytosine [H (851)CSAGIGRT (859)]: yellow; WPD loop [W (818)PDHGVP(824)]: purple; β5-loop-β6 region [T (754)QVERGRV (761)]: cyan; C-terminus, N-terminus and linker: grey; highly conserved linker region [E (617) PDFQYIP(624)]: green; PDZ: blue. **(A)** Top view depicts catalytic site in vicinity to the linker. **(B)** Bottom view adopts same orientation as **Figure 3**. **(C)** Linker view shows conformational variability of linker.

simulations of 200 ns starting from the model with lowest χ^2 = 2.48 (**Figure 7**). Initially, the relative position of the two domains fluctuates, but it converges in each case to a more compact structure with direct and stable interactions between the two domains after a maximum of 75 ns. This behaviour is reflected in the analysis of the distances between the two domains (**Figure 8**), showing an initial increase of the distances (~ 9–18 Å) followed by gradual reduction of distances (~ −10 Å), with respect to the initial conformation.

To better characterize the observed conformational transitions along the MD simulations of PTPN4, we clustered the set of conformations with the Self Organizing Maps (SOM) method already used above Bouvier et al. (2015). A total of 60 clusters were retrieved from a pool of 60,000 conformations (**Figure 9**). We then projected the χ^2 values, the changes of distances between the two domains and the simulation time on the SOM map (**Figures 9A–C**). The analysis of the two maps suggested four groups of clusters, where G2 had the highest χ^2 and maximum increase of distances, and G4 the lowest χ^2 and minimum changes of distances. **Figures 9D–G** shows one representative conformation per cluster, clearly indicating four distinct relative positions of the PDZ with respect to the PTP. These four conformations satisfy the SAXS data to a very different degree, indicated by the color in the SOM maps and in the

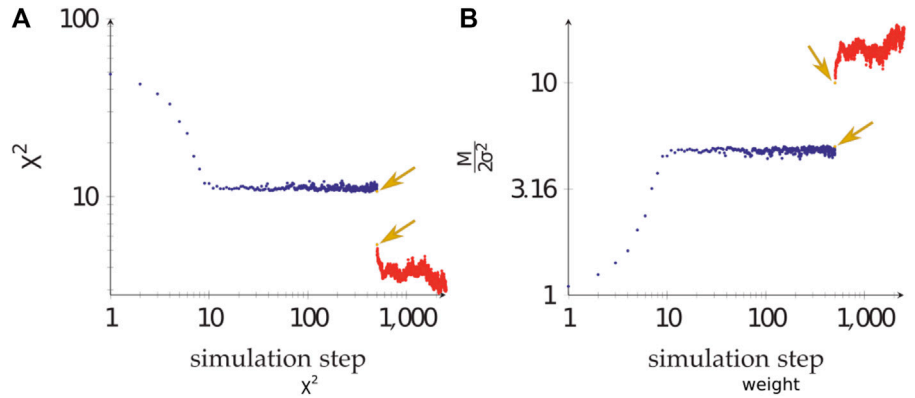


FIGURE 5 | The adjustment of the Bayesian SAXS score. **(A)** χ^2 and **(B)** SAXS restraint weight of a composite simulation, starting from rigid body steps (blue), followed by linker modelling (yellow, pointed to by arrows), and ending with Monte Carlo flexible refinement (red).

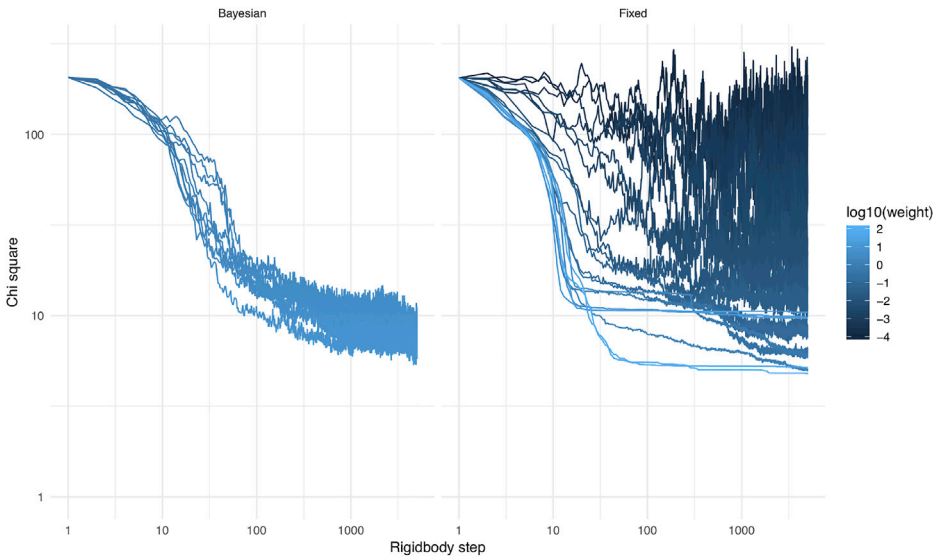


FIGURE 6 | χ^2 score as a function of simulation step, for the 64 rigid body simulations with Bayesian SAXS score (left), and for 20 simulations with a fixed weight (different in each simulation) and the same, random, starting structure (right).

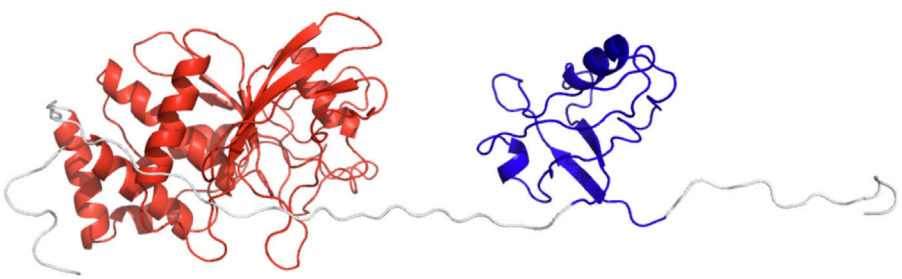


FIGURE 7 | The cartoon representation of starting conformation for the MD simulations. PDZ is colored in blue and PTP in red.

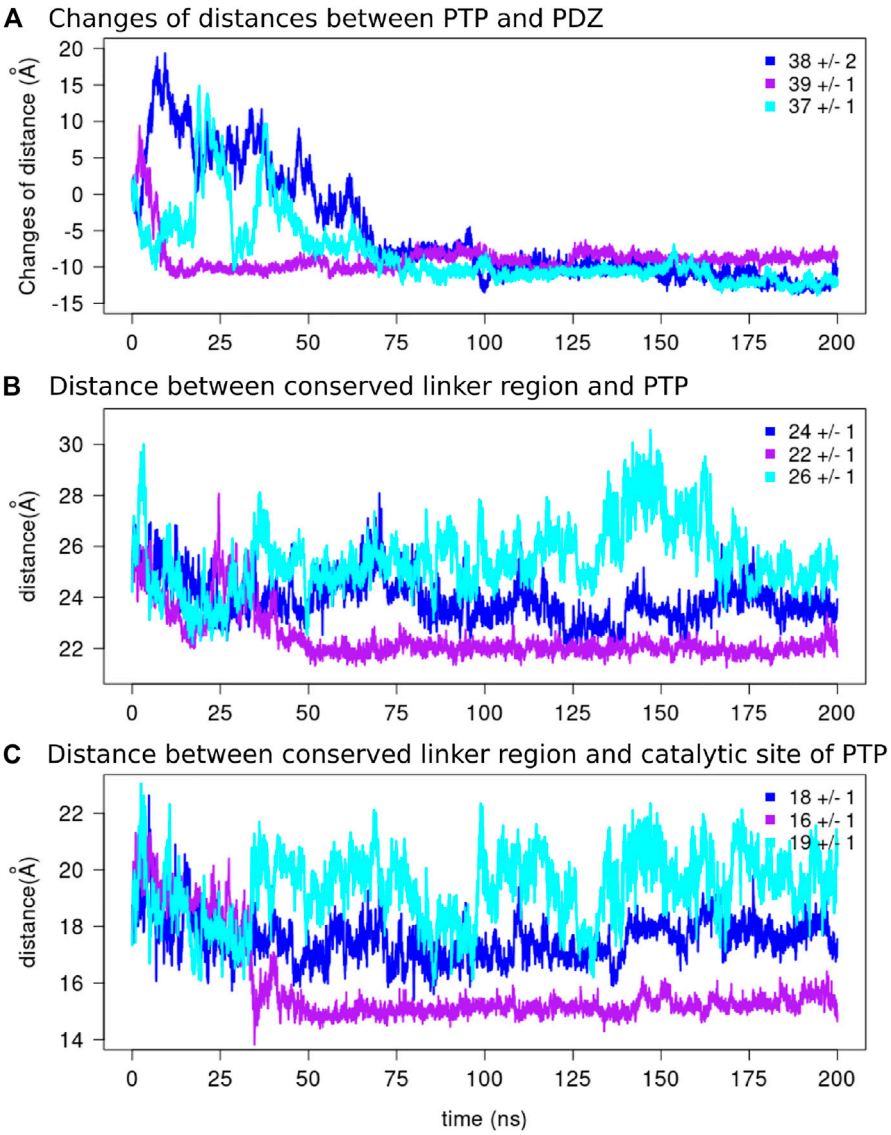


FIGURE 8 | The distances along the MD simulations. **(A)** The changes of distances between the two domains and the average values over the final 125 ns of the simulations are reported for each replicate. The distance between the center of mass of the conserved linker region [E (617)PDFQYIP(624)] and the center of mass of **(B)** the PTP domain and **(C)** the catalytic site of the PTP domain are depicted for each replicate.

conformations shown (each group is colored according to their average χ^2 value from dark violet for the minimum values to dark green for the maximum values). The analysis of the PTPN4 conformational changes revealed the existence of four distinct conformational states for the PDZ with respect to the PTP, one of which is close to the Bayesian SAXS restraint model and has a low χ^2.

To investigate overall convergence of the simulations, we analyzed the number of conformations from different replicates in each group (**Figure 9H**). The three replicates cover rather different conformational space. The groups G1, G2, and G3 contain conformations from only one replicate. Interestingly, only G4, which is the closest one to the starting conformation and has the lowest χ^2 scores, contains conformations from all the three replicates. The further analysis of the clusters along the simulation time (**Figure 9C**) showed that G4 contains trajectories appearing at the beginning of the simulations, and the G1-G3 are visited subsequently. Interestingly, the position of the linker with respect to the PTP, remained unchanged in all the clusters as can be seen in **Figures 9D–G**. In order to further investigate the conformational changes of the linker, we measured the distances between the center of mass of the conserved linker region (E617-P624) and i) the PTP domain and ii) the catalytic site of the PTP domain (the β5-loop-β6 region and the WPD loop) along the three replicates of the MD simulations (**Figures 8B,C**). The variation of distances are within the range of 1 Å, therefore suggesting the rather stable position of the linker with respect to the PTP domain.

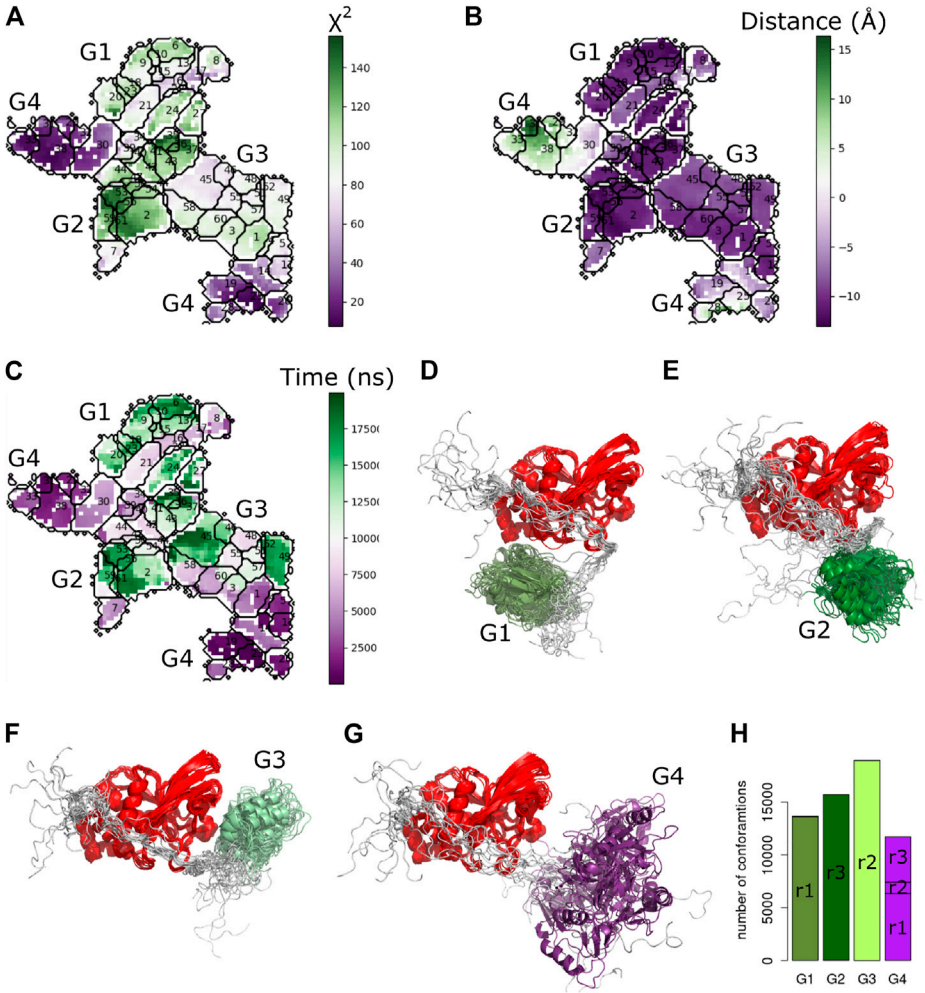


FIGURE 9 | Cluster analysis of PTPN4 from the MD simulations. The self-organizing map of the PTPN4 conformations colored by **(A)** χ^2, **(B)** changes of distances between the center of mass of the PDZ and PTP domains, and **(C)** simulations time (ns) of the replicates. The clusters are numbered on the maps from 1 to 60, and divided into four groups (G1-G4). One representative conformation is shown for the clusters that are forming the four groups G1, G2, G3 and G4 in **(D–G)**, respectively. The PTP is colored in red and the four identified groups of PDZ clusters in difference shades of green and purple, reflecting their average χ^2 values. **(H)** The number of conformations from different replicates (r1, r2, and r3) are reported for each group.

2.3.2 Selection of Minimal Small Angle X-ray Scattering Ensemble

The above analysis assumes that a single structure or an ensemble covering a small part of conformational space represents the SAXS data. The sampling of conformational space by the free MD trajectories enabled us to try to investigate if more disperse ensembles fit the SAXS data better. For this, we used a method based on the genetic algorithm (GA) that was developed for a similar problem Delhommel et al. (2017). This method searches for the minimal subset of conformations minimizing the error between the experimental data and computed data from the MD simulations. The χ^2 values obtained after fitting were reduced from 6.03 to 2.79 for an ensemble size of three. Increasing the ensemble beyond three did not reduce the χ^2 further (**Figure 10A**). We clustered the weighted conformations obtained in all the ensembles according to the four conformational groups identified by the SOM analysis (G1, G2, G3, and G4). **Figure 10B** shows the ratio of conformations that belong to each group for each ensemble size, averaged over the 5 GA runs. The ratios of the conformations belonging to the four groups are similar for different ensemble sizes, where G4 is always most represented with a weight of about 70%, while G1 has about 30% of the weight. The experimental and fitted profiles (for the ensemble size of three) are compared (**Figure 10C**, shown in black and cyan, respectively), and the conformations obtained for the ensemble size three are shown in **Figure 10D**. We conclude that the SAXS data are best represented by two major conformations, an open and a closed states. The open state has the highest weight (70%) and is similar to the initial conformation obtained by the Bayesian method.

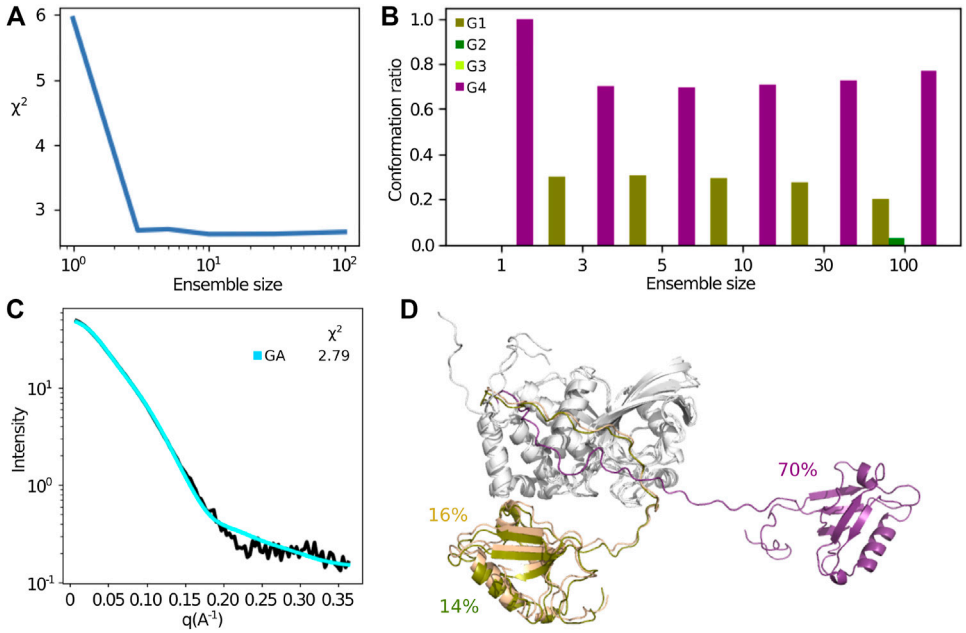


FIGURE 10 | Extracting a minimum subset of conformations from the MD simulations, that describes best the SAXS data using a genetic algorithm. **(A)** Improvement of the χ^2 with respect to the ensemble size in the genetic algorithm selection. **(B)** The conformations are clustered in four categories according to the clustering obtained using SOM; conformations in the groups G1, G2, G3, and G4 are colored in olive, green, yellow-green and purple bars, respectively. The proportion of each group is represented for each ensemble size. **(C)** Back-calculated SAXS profile using the genetic algorithm derived from the ensemble size of three (in blue) and the experimental profile (in black), with the $\chi^2 = 2.79$. **(D)** Representative ensemble of PTPN4 superimposed on PTP showing the PDZ in G1 conformation (olive and khaki) or in G4 conformation (purple) with the proportion of each conformer.

3 DISCUSSION

3.1 Automatic Weight Adjustment

In general, and also in the Bayesian formalism, the SAXS scoring term is based on χ^2 (**Eq. 3**), here multiplied by a weight $M/2\sigma^2$. Commonly, the weight on the scoring term is based on some heuristics, for example the number of independent data points Shevchuk and Hub (2017). Experience shows that this weight is not easy to set and can require adjustment during the simulation, in particular when χ^2 is expressed with SAXS intensities (as opposed to their logarithms) Chen and Hub (2015). In the context of the Bayesian formalism, the weight is set by changing σ. This parameter does not only depend on the quality and consistency of the experimental data but also on the forward model used. The nuisance parameter σ evidently scales the experimental errors with a constant factor, and it is unknown before the calculation. It is the hallmark of the Bayesian formalism that this parameter is treated as an unknown, at the same level as the coordinates. σ, and in consequence the weight, is adjusted during the calculation, without making any additional assumptions on the values it can take. To do this, we use the second term on the right hand side of **Eq. 2**, $M\log(\sigma)$. In absence of this term proportional to the logarithm of σ, the trivial minimum of the score would be reached when σ diverges and the weight becomes zero. This automatic weighting modulates the effect of χ^2 on the final scoring term. This treatment is analogous to what we introduced for NMR data, electron microscopy data Habeck et al. (2006), Nilges et al. (2008), Bernard et al. (2011) and cross-linking mass spectrometry data Ferber et al. (2016).

3.2 Influence of the Weight Adjustment

As an illustration, suppose structure determination is performed with a bad guess for the initial structure. In this case, χ^2 will be large. Adjustment of the weight will drive σ towards larger values, and the weight becomes smaller. σ acts to reset the scale of the restraint. Notice however that its update is less frequent than that of χ^2. That way, structures are sampled with χ^2 values around σ^2, which is then slowly lowered to increase stringency on the restraint. σ^2 acts as an annealing parameter. As long as the structure is in strong disagreement with SAXS data, the weight of the Bayesian SAXS score will be small. This behaviour allows other terms of the force field to dominate, and conformational exploration can happen unhindered by an irrelevant SAXS term. If exploration leads to a structure with a smaller χ^2, the weight will increase. The SAXS term therefore becomes more discriminant, guiding the calculation to propose structures which match the SAXS profile more closely. Bayesian formulation of SAXS structure determination therefore transforms a rugged energy landscape into a funnel-shaped landscape Dill and Chan, 1997.

Note that, the σ is being adjusted on the fly, and the maximum likelihood estimate of σ is approximately χ^2 (**Supplementary Equation S4**). Therefore, the proper quantity to look at is $M/2\sigma^2$

(see **Figure. 5B**), which is a function of the degrees of freedom in the curve (Spill, 2013) (**Section 2**.4.8.3, pp 171). In case of multiple datasets, it is therefore crucial that each has their own σ.

3.3 Fixed Weight vs. Bayesian Automatic Weighting

The optimal weight, at which the simulation has reasonable acceptance rates and makes good use of SAXS information, is *a priori* unknown. It is the purpose of the Bayesian SAXS restraint to determine this optimal weight. As shown in **Supplementary Equation S4** (see Supplementary Material), the number of SAXS data points and the overall agreement of data and structures will greatly influence the optimal weight. Therefore, it is expected that it will be different for each SAXS dataset, but also for each simulation setting, for example depending on which force field is used.

3.4 Log Score vs. Linear Score

An equivalent form for the Bayesian score without any additional parameter σ can be derived by an operation called marginalization (**Supplementary Equation S5**, Supplementary Material). As shown for NMR data Habeck et al. (2006), this form is equivalent to the weighted χ^2 term, but does not need automatic weight adjustment, because it incorporates the behavior described above. Its form is simply the logarithm of the traditional χ^2. Using the logarithm of the χ^2 lowers the score penalty for large values of χ^2, while keeping its effect similar to the standard χ^2 formulation when it is close to one. Interestingly, it has been observed by Chen and Hub (2015) that a χ^2 formulation using the logarithm of the intensities does not require much adjustments of the weight. While they apply the logarithm on the individual intensities and not the χ^2 as a whole, the effect of lowering the impact of large discrepancies remains. When using a χ^2 on linear scale (as proposed here), the authors observe the need to adjust this weight specifically in the beginning of the simulation. That is, when discrepancies in the low-q and high diffusion intensity region of the SAXS curve are likely to occur, and contribute most to the scoring. Applying a logarithm on the first part of the SAXS curve is therefore what probably alleviates the need to adjust the weight. In contrast, we have employed a χ^2 on a linear scale (including error bars, **Eq. 3**) because the SAXS measurements and noise scale linearly. The logarithm is applied afterwards, for scoring purposes.

3.5 A Point on Exhaustivity

The calculations presented here attempted to sample a large part of the conformational space of this two-domain system, since the energy landscape can be expected to be rugged. We showed that the energy surface is less rugged when using automatically adapted weights. The strength of this Bayesian restraint is that, regardless of the initial conformation, the calculations converge to low χ^2 structures. This is particularly beneficial when computer resources are limited. In our PTPN4 example, one in every four simulations ends up in the basin with the lowest χ^2 conformers.

3.6 Protein Tyrosine Phosphatase Non-Receptor 4

Using the novel Bayesian SAXS restraint, we have shown a conserved sequence in the linker of PTPN4, involved in the allosteric regulation of PTPN4 Caillet-Saguy et al. (2017), is facing both the β5-loop-β6 region and the WPD loop. The β5-loop-β6 region is thought to participate in defining substrate specificity Andersen et al. (2001). The WPD loop is well-known to be important for the phosphatase catalysis. The WPD loop switches from an open to close position upon substrate binding and adopts a catalytically active close conformation Barr et al. (2009). Previous experimental evidence showed that the linker participates in the control of the catalytic activity of the phosphatase domain Maisonneuve et al. (2014).

Mutations of a conserved hydrophobic patch in the linker suggested that the linker modulates the WPD loop open/closed conformations Caillet-Saguy et al. (2017). The close proximity of the linker with the β5-loop-β6 region and the WPD loop observed in our simulations further supports and reinforces the current model in which the linker of PTPN4 could regulate the phosphatase activity of PTPN4 by modulating the WPD loop closure.

3.7 Ensemble Modelling

The focus of this study is to illustrate the power and utility of the Bayesian SAXS score. The setup was deliberately simple, to emphasize to what degree the final conformations were driven by the data. Emphasis was also on calculation efficiency, and the molecule was deliberately described in the simplest terms by excluding volume, rigid bodies for the two domains, and rigid covalent geometry. The experimental data was limited to SAXS data up to $q < 0.37\text{Å}^{-1}$. The SAXS data do not contain any information on specific interactions between the linker and the surface of the PTP domain. In our models the linker wraps around the PTP domain but does not directly contact the domain. This is consistent with the fact that there is no experimental NMR data that indicates a specific contact, but does not explain the sequence conservation in the linker and on the surface of the PTP domain. The tandem of PDZ-PTP domains in PTPN4 may be the location of continuous conformational changes due to the fuzzy nature of the intramolecular interactions that stabilize the spatial organization of the two domains Maisonneuve et al. (2014, 2016), Caillet-Saguy et al. (2017). This is further confirmed by the analysis of conformations generated by MD simulations starting from the top model, where four distinct groups of conformations are identified (**Figure 9**). The flexible and unstructured linker is most likely in transient interactions with the PTP domain as monitored by NMR (R2 relaxation rate, Maisonneuve et al. (2014) **Figure 5B**). In our calculations, the models with low-χ^2 (the upper left cluster in **Figure 3C**) present conformations of the linker that covers almost half of the PTP domain. This conformation of the linker with respect to the PTP domain remains rather stable along the MD simulations.

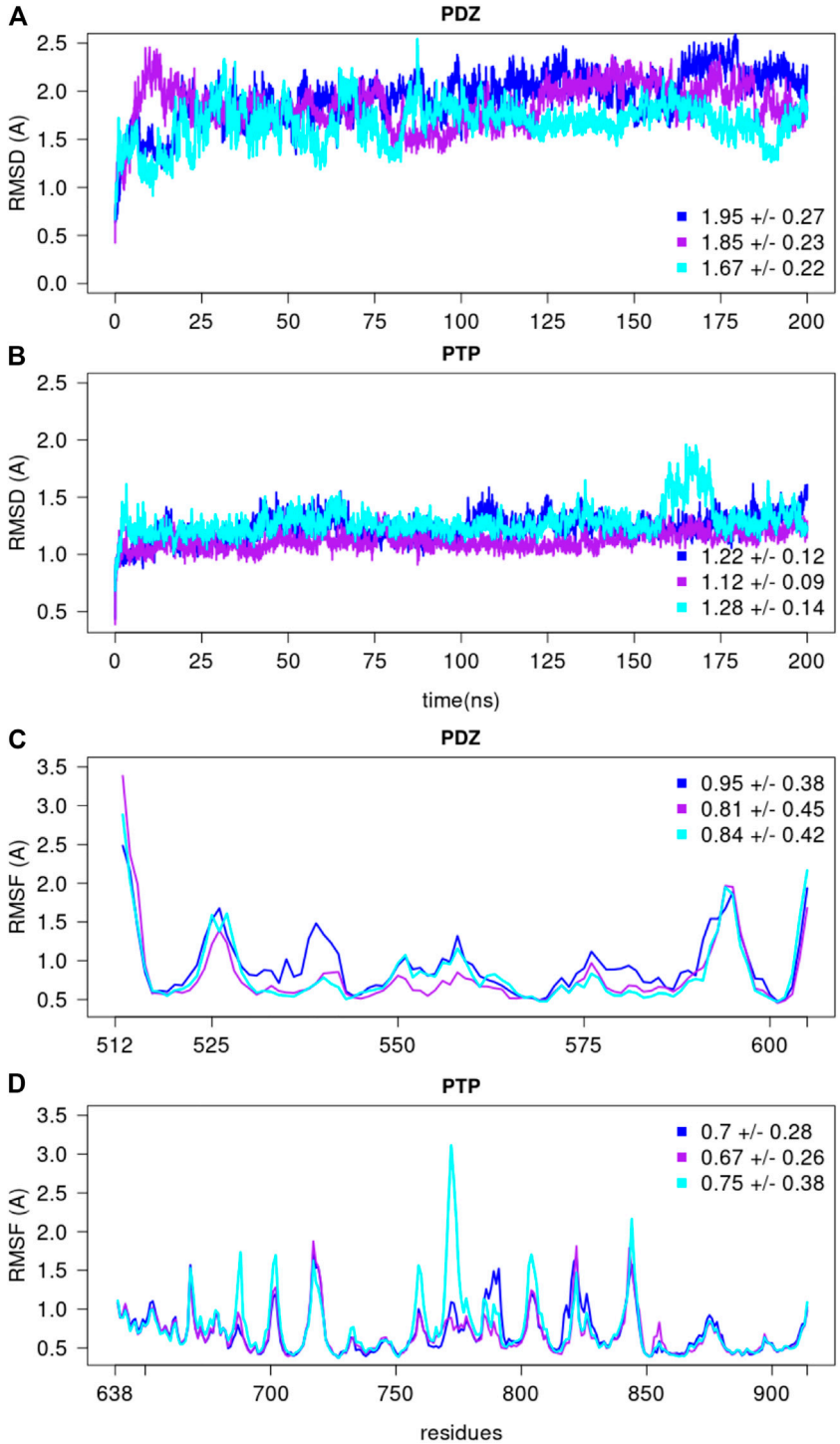


FIGURE 11 | The root mean square deviations and fluctuations of each MD simulation. The RMSD over backbone atoms (Cα, C, N, O) measured from the initial structure are shown for the **(A)** PDZ and **(B)** PTP domains. The residue RMSF over backbone atoms (Cα, C, N, O) measured with respect to the average conformation are depicted for the **(C)** PDZ and **(D)** PTP domains, over the last 150 ns of each replicate. The average and standard deviation values are reported for every replicate.

The conformations we obtain can serve as the basis of more detailed simulations with state of the art ensemble methods Potrzebowski et al. (2018), Shrestha et al. (2019), Paissoni et al. (2020). For a system of rather moderate size as the PTPN4 tandem (52 kDa), one could obviously directly refine against the data in a complete force field Shevchuk and Hub

(2017). This would not allow for as extensive searching of conformational space as it was performed in this work. The aim of the current calculation protocol is to sample large relevant parts of conformational space efficiently, a task that is difficult to perform for large fully solvated molecules. An adaptation of the Bayesian SAXS restraint with automated weighting as described here could be useful also in this context. We note that such adaptation however, would not address the issue of multiple conformations representing the SAXS data. In this study we proposed a method to overcome this problem by first concentrating on obtaining the dominate conformational ensemble in a largely simplified force field without explicit solvent, and then further exploring a larger ensemble by a free, fully solvated simulation, and finally obtaining an optimal, small ensemble by combining different conformations from these simulations. While the best conformer obtained by Bayesian SAXS restraint has $\chi^2 = 2.48$, our approach allowed us to reveal an ensemble of three structures capturing two different states of PTPN4 with a fitted χ^2 of 2.79. Interestingly, while for several of the proteins studied with the CHARMM36m force field, the resulting structures are more compact than indicated by experiment (unless protein-water interactions are increased) Huang et al. (2017), our analysis highlights both compact and open states for PTPN4.

4 MATERIALS AND METHODS

4.1 Protein Production and Data Collection

The PDZ-PTP$^{C/S}$ construct, harboring the mutant C852S, hereafter referred to as PTPN4, was expressed and purified as previously described Maisonneuve et al. (2014). SAXS experiments were carried out as previously described except that the protein concentration used for SAXS experiments was 75 μM Maisonneuve et al. (2014).

4.2 Rigid Body Docking

In the first stage, we used IMP Russel et al. (2012) to perform rigid body docking of the PDB structures of PTP (PDB code 2I75; residues 638–913) and PDZ (PDB code 3NFK chain B; residues 512–604). 64 different simulations were performed with 500 steps each. Initial orientations of PDZ with respect to PTP cover a wide range of orientations both around the PTP and of the PDZ itself (see **Figure 2**). Energy terms were the SAXS restraint (**Supplementary Equation S7**) and a quadratic volume exclusion term. The FoXS model was used on heavy atoms Schneidman-Duhovny et al. (2013). Each step consisted in alternating 100 Monte Carlo rotation/translation moves (510^{-2}rad/Å) of PDZ with respect to PTP, and optimizing c_1, c_2, σ and γ. σ and γ were optimized by setting them to their maximum posterior (**Supplementary Equation S4** and Spill et al. (2014)). c_1 is constrained to be between 0.95 and 1.05, while c_2 is constrained between −2 and 4. c_1 and c_2 are jointly optimized by a two-dimensional grid search, as follows. First, a 11 × 11 grid of values is tried on the admissible range of c_1 and c_2. Then, the pair with the lowest score is used as the center of a new 11 × 11 grid, whose total size covers that of four cells of the previous grid. The same procedure yields a refined estimate of c_1 and c_2. This pair is in turn used in a second round of refinement, for which another 11 × 11 grid is generated with half the gridsize of the previous round, yielding the final estimate of c_1 and c_2. Importantly, before each evaluation of the score at a given c_1 and c_2 pair, σ and γ are set to their maximum posterior estimates.

4.3 Rigid Body Self Organizing Map

A 50 × 50 SOM Bouvier et al. (2015), Spill et al. (2013) was trained on the last 200 frames of each of the 64 simulations. Specifically, we used descriptors with seven dimensions, extracted from the structures as follows. The coordinates of all 12,800 structures were recalculated in a reference frame in which the center of mass of PTP is at the origin, and its orientation is constant across the structures. The first three dimensions of the descriptors are the center of mass of PDZ in this reference frame, while the last four are the quaternions of the rotation of PDZ with respect to PTP. The metric used to compare a neuron n and a descriptor m is a weighted sum between euclidean distance between the center of masses and geodesic distance between the quaternions Huynh (2009).

$$d(n,m) = \sqrt{\sum_{i=1}^{3}(n_i - m_i)^2 + \frac{2d_{\max}}{\pi}\arccos\left|\sum_{i=4}^{7} n_i m_i\right|} \quad (4)$$

where $d_{\max}$ is the length of the largest space diagonal of the bounding box of the descriptor's first three coordinates. Neurons were updated by interpolation either in Cartesian space (first three coordinates) or in quaternion space, e.g., on the unit 4-sphere (last four coordinates).

4.4 Linker Modeling

In the second stage, we added linkers to our models. Due to the particular choice of the format of the SOM descriptors, a 3D structure can be reconstructed from the coordinates of the trained neurons. 1,999 clash-free structures could be extracted from the SOM neurons in such a way.

Missing residues were added with IMP Russel et al. (2012) so that the modeled part of the protein spanned residues 496–926. The termini were assigned random ϕ/ψ dihedral angles in such a way that no clash was caused.

The linker was generated in two steps. First, C_α atoms were placed on a path that connects the two endpoints without passing through either PTP or PDZ. The C_α linker was then minimized with a harmonic distance restraint between consecutive C_α atoms (target distance D = 3.86Å) and an excluded volume restraint to avoid interpenetration. C_α atoms within the linker had a normal diameter D while other atoms had diameter $2D$ to push the linker outside of the protein during initial minimization. 1,000 steps of steepest descent were followed by 1,000 steps of conjugate gradient.

Second, all atoms were placed around their corresponding C_α at random in a sphere of diameter D. CHARMM bonded restraints were enforced MacKerell et al. (1998), and 250 steps of steepest descent were performed, followed by 1,000 steps of conjugate gradient. Then, volume exclusion was turned on, with

standard CHARMM radii, and followed by the same 250 + 1,000 steps of minimization.

On average, this step resulted in 1,224 structures per pose, or a total of 2,461,844 structures.

4.5 Monte Carlo Refinement

For each of the 1,999 rigid body poses, the structure with linkers which had the best Bayesian SAXS score was used as starting conformation for a Monte Carlo refinement simulation. Each simulation consisted of 2,000 steps, each of which was an alternation between 10 Monte Carlo moves and optimization of σ and γ. Each Monte Carlo move was made in internal coordinates, and consisted in a Gaussian perturbation of the backbone dihedrals of residues 496–511, 606–636, and 914–926. The standard deviation of the Gaussian was $5 \times 10^{-2} rad$ for the termini and $5 \times 10^{-3} rad$ for the linker.

4.6 Fixed-Weight Small-angle X-ray scattering Simulations

To compare fixed-weight and self-adjusting simulations, we used a similar setup. 20 fixed-weight simulations were performed with a SAXS restraint with a weight spaced logarithmically from 10^{-4} to 10^{2}. 10 simulations using the Bayesian SAXS score described here were performed for comparison. The starting structure was always identical, and consisted of a random orientation of PDZ with respect to PTP, with linkers and termini added. Each simulation was performed for 5,000 steps.

4.7 Molecular Dynamics Simulations

We selected the top PTPN4 conformation determined using the Bayesian SAXS score, i.e., the one with the lowest χ^2 score (2.48). This conformation was used as the starting structure for the molecular dynamics simulations (7). MD simulations were performed with NAMD2.13 Phillips et al. (2005) using CHARMM36m force field parameter set Huang et al. (2017): i) hydrogen atoms were added, ii) the solute was hydrated with a cuboid box of explicit TIP3P water molecules Jorgensen et al. (1983) with a buffering distance up to 10 Å, iii) 10 Na+counter-ions were added to neutralise the system, leading to a total system size of 150,730 atoms. The following minimization procedure was applied: i) 10,000 steps of minimization of the water molecules keeping protein atoms fixed, ii) 10,000 steps of minimization keeping only protein backbone fixed to allow protein side chains to relax, iii) 10,000 steps of minimization without any constraint on the system. Heating of the system to the target temperature of 310 K was performed at constant volume using the Berendsen thermostat Berendsen et al. (1984). Thereafter, the system was equilibrated for 100 ps at constant volume (NVT) and for further 100 ps using a Langevin piston (NPT) Loncharich et al. (1992) to maintain the pressure. The production was realised in the NPT ensemble. The time step was set to 2.0 fs. The temperature was kept at 310 K and pressure at 1 bar using the Langevin piston coupling algorithm. The SHAKE algorithm was used to freeze bonds involving hydrogen atoms, allowing for an integration time step of 2.0 fs. The Particle Mesh Ewald method Darden et al. (1993) was employed to treat long-range electrostatics. The coordinates of the system were written every 10 ps. We performed three replicates of 200 ns, with different initial velocities. To assess the stability of each replicate, the root mean square deviation (RMSD) and fluctuation (RMSF) were recorded along each MD simulation (**Figure 11**). We also measured the distances along the simulations between the center of mass of the two domains in each replicate (**Figure 8**).

4.8 Back Calculated Small-angle X-ray scattering Profiles

For every conformation of the MD simulations, the theoretical scattering profiles were calculated using CRYSOL from the ATSAS 2.8.3 software suite Svergun et al. (1995), with the default parameters. Their corresponding χ^2 values were measured using the following equation:

$$\chi^2 = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{I_{calc}(i) - I_{exp}(i)}{\sigma_{exp}(i)} \right)^2 \quad (5)$$

where M is the number of points in SAXS profile, I_{calc} is the back calculated intensity, I_{exp} and σ_{exp} are the experimental intensity and error, respectively.

4.9 Genetic Algorithm

We followed a similar procedure as in Delhommel et al. (2017), in which 1,000 steps of GA were performed, the number of generated ensemble was set to 1,000 with a cross over frequency of 0.8 and a mutation frequency of one. We performed the GA for different ensemble sizes: 1, 3, 5, 30, and 100. In addition, the GA was repeated five times for every ensemble size and average values were reported.

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: The SAXS data, refined structures, and MD simulation trajectories generated for the PTPN4 for this study are deposited in the Zenodo. org database (accession doi: 10.5281/zenodo.4739101). Direct link: https://zenodo.org/record/4739101.

AUTHOR CONTRIBUTIONS

All authors wrote and reviewed the article.

FUNDING

This work was supported by The Fondation pour la Recherche Medicale (Equipe FRM 2017M.DEQ20170839114) to YK and MN. PM was supported by grants from the Ministère de l'Enseignement Supérieur et de la Recherche and the Fondation pour la Recherche Médicale (FDT20130927999).

ACKNOWLEDGMENTS

The SAXS data, refined structure, and MD simulation trajectories generated for the PTPN4 for this study are deposited in the Zenodo.org database (accession doi: 10.5281/zenodo.4739101).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.671011/full#supplementary-material

REFERENCES

Andersen, J. N., Mortensen, O. H., Peters, G. H., Drake, P. G., Iversen, L. F., Olsen, O. H., et al. (2001). Structural and Evolutionary Relationships Among Protein Tyrosine Phosphatase Domains. *Mol. Cel. Biol.* 21, 7117–7136. doi:10.1128/mcb.21.21.7117-7136.2001

Babault, N., Cordier, F., Lafage, M., Cockburn, J., Haouz, A., Prehaud, C., et al. (2011). Peptides Targeting the PDZ Domain of PTPN4 Are Efficient Inducers of Glioblastoma Cell Death. *Structure* 19, 1518–1524. doi:10.1016/j.str.2011.07.007

Barr, A. J., Ugochukwu, E., Lee, W. H., King, O. N. F., Filippakopoulos, P., Alfano, I., et al. (2009). Large-scale Structural Analysis of the Classical Human Protein Tyrosine Phosphatome. *Cell* 136, 352–363. doi:10.1016/j.cell.2008.11.038

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular Dynamics with Coupling to an External bath. *J. Chem. Phys.* 81, 3684–3690. doi:10.1063/1.448118

Bernard, A., Vranken, W. F., Bardiaux, B., Nilges, M., and Malliavin, T. E. (2011). Bayesian Estimation of NMR Restraint Potential and Weight: a Validation on a Representative Set of Protein Structures. *Proteins* 79, 1525–1537. doi:10.1002/prot.22980

Bonomi, M., Hanot, S., Greenberg, C. H., Sali, A., Nilges, M., Vendruscolo, M., et al. (2019). Bayesian Weighing of Electron Cryo-Microscopy Data for Integrative Structural Modeling. *Structure* 27, 175–188. doi:10.1016/j.str.2018.09.011

Bouvier, G., Desdouits, N., Ferber, M., Blondel, A., and Nilges, M. (2015). An Automatic Tool to Analyze and Cluster Macromolecular Conformations Based on Self-Organizing Maps. *Bioinformatics* 31, 1490–1492. doi:10.1093/bioinformatics/btu849

Brünger, A. T. (1992). Free R Value: a Novel Statistical Quantity for Assessing the Accuracy of crystal Structures. *Nature* 355, 472–475. doi:10.1038/355472a0

Caillet-Saguy, C., Maisonneuve, P., Delhommel, F., Terrien, E., Babault, N., Lafon, M., et al. (2015). Strategies to Interfere with PDZ-Mediated Interactions in Neurons: What We Can Learn from the Rabies Virus. *Prog. Biophys. Mol. Biol.* 119, 53–59. doi:10.1016/j.pbiomolbio.2015.02.007

Caillet-Saguy, C., Toto, A., Guerois, R., Maisonneuve, P., Di Silvio, E., Sawyer, K., et al. (2017). Regulation of the Human Phosphatase PTPN4 by the Interdomain Linker Connecting the PDZ and the Phosphatase Domains. *Scientific Rep.* 7, 2–10. doi:10.1038/s41598-017-08193-6

Chen, P.-c., and Hub, J. S. (2015). Interpretation of Solution X-ray Scattering by Explicit-Solvent Molecular Dynamics. *Biophysical J.* 108, 2573–2584. doi:10.1016/j.bpj.2015.03.062

Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald: AnN·Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397

Delhommel, F., Cordier, F., Bardiaux, B., Bouvier, G., Colcombet-Cazenave, B., Brier, S., et al. (2017). Structural Characterization of Whirlin Reveals an Unexpected and Dynamic Supramodule Conformation of its PDZ Tandem. *Structure* 25, 1645–1656. doi:10.1016/j.str.2017.08.013

Dill, K. A., and Chan, H. S. (1997). From Levinthal to Pathways to Funnels: The "New View" of Protein Folding Kinetics. *Nat. Struct. Biol.* 4, 10.

Ferber, M., Kosinski, J., Ori, A., Rashid, U. J., Moreno-Morcillo, M., Simon, B., et al. (2016). Automated Structure Modeling of Large Protein Assemblies Using Crosslinks as Distance Restraints. *Nat. Methods* 13, 515–520. doi:10.1038/nmeth.3838

Gu, M., and Majerus, P. W. (1996). The Properties of the Protein Tyrosine Phosphatase PTPMEG. *J. Biol. Chem.* 271, 27751–27759. doi:10.1074/jbc.271.44.27751

Gu, M., Meng, K., and Majerus, P. W. (1996). The Effect of Overexpression of the Protein Tyrosine Phosphatase PTPMEG on Cell Growth and on colony Formation in Soft agar in COS-7 Cells. *Proc. Natl. Acad. Sci.* 93, 12980–12985. doi:10.1073/pnas.93.23.12980

Habeck, M., Rieping, W., and Nilges, M. (2006). Weighting of Experimental Evidence in Macromolecular Structure Determination. *Proc. Natl. Acad. Sci.* 103, 1756–1761. doi:10.1073/pnas.0506412103

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

Huynh, D. Q. (2009). Metrics for 3d Rotations: Comparison and Analysis. *J. Math. Imaging Vis.* 35, 155–164. doi:10.1007/s10851-009-0161-2

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

Kina, S.-i., Tezuka, T., Kusakawa, S., Kishimoto, Y., Kakizawa, S., Hashimoto, K., et al. (2007). Involvement of Protein-Tyrosine Phosphatase PTPMEG in Motor Learning and Cerebellar Long-Term Depression. *Eur. J. Neurosci.* 26, 2269–2278. doi:10.1111/j.1460-9568.2007.05829.x

Kohda, K., Kakegawa, W., Matsuda, S., Yamamoto, T., Hirano, H., and Yuzaki, M. (2013). The 2 Glutamate Receptor gates Long-Term Depression by Coordinating Interactions between Two AMPA Receptor Phosphorylation Sites. *Proc. Natl. Acad. Sci.* 110, E948–E957. doi:10.1073/pnas.1218380110

Loncharich, R. J., Brooks, B. R., and Pastor, R. W. (1992). Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates ofN-Acetylalanyl-N?-Methylamide. *Biopolymers* 32, 523–535. doi:10.1002/bip.360320508

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., et al. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. *J. Phys. Chem. B* 102, 3586–3616. doi:10.1021/jp973084f

Maisonneuve, P., Caillet-Saguy, C., Raynal, B., Gilquin, B., Chaffotte, A., Pérez, J., et al. (2014). Regulation of the Catalytic Activity of the Human Phosphatase Ptpn4 by its Pdz Domain. *Febs J.* 281, 4852–4865. doi:10.1111/febs.13024

Maisonneuve, P., Caillet-Saguy, C., Vaney, M.-C., Bibi-Zainab, E., Sawyer, K., Raynal, B., et al. (2016). Molecular Basis of the Interaction of the Human Protein Tyrosine Phosphatase Non-receptor Type 4 (PTPN4) with the Mitogen-Activated Protein Kinase P38γ. *J. Biol. Chem.* 291, 16699–16708. doi:10.1074/jbc.m115.707208

Mareuil, F., Sizun, C., Perez, J., Schoenauer, M., Lallemand, J.-Y., and Bontems, F. (2007). A Simple Genetic Algorithm for the Optimization of Multidomain Protein Homology Models Driven by NMR Residual Dipolar Coupling and Small Angle X-ray Scattering Data. *Eur. Biophys. J.* 37, 95–104. doi:10.1007/s00249-007-0170-2

Nilges, M., Bernard, A., Bardiaux, B., Malliavin, T., Habeck, M., and Rieping, W. (2008). Accurate NMR Structures through Minimization of an Extended Hybrid Energy. *Structure* 16, 1305–1312. doi:10.1016/j.str.2008.07.008

Paissoni, C., Jussupow, A., and Camilloni, C. (2020). Determination of Protein Structural Ensembles by Hybrid-Resolution SAXS Restrained Molecular Dynamics. *J. Chem. Theor. Comput.* 16, 2825–2834. doi:10.1021/acs.jctc.9b01181

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

Potrzebowski, W., Trewhella, J., and Andre, I. (2018). Bayesian Inference of Protein Conformational Ensembles from Limited Structural Data. *Plos Comput. Biol.* 14, e1006641. doi:10.1371/journal.pcbi.1006641

Préhaud, C., Wolff, N., Terrien, E., Lafage, M., Mégret, F., Babault, N., et al. (2010). Attenuation of Rabies Virulence: Takeover by the Cytoplasmic Domain of its Envelope Protein. *Sci. Signaling* 3, ra5. doi:10.1126/scisignal.2000510

Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential Structure Determination. *Science* 309, 303–306. doi:10.1126/science.1110428

Rozycki, B., Kim, Y. C., and Hummer, G. (2011). SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* 19, 109–116.

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, e1001244. doi:10.1371/journal.pbio.1001244

Schneidman-Duhovny, D., Hammel, M., Tainer, J. A., and Sali, A. (2013). Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. *Biophysical J.* 105, 962–974. doi:10.1016/j.bpj.2013.07.020

Shevchuk, R., and Hub, J. S. (2017). Bayesian Refinement of Protein Structures and Ensembles against SAXS Data Using Molecular Dynamics. *Plos Comput. Biol.* 13, e1005800. doi:10.1371/journal.pcbi.1005800

Shrestha, U. R., Juneja, P., Zhang, Q., Gurumoorthy, V., Borreguero, J. M., Urban, V., et al. (2019). Generation of the Configurational Ensemble of an Intrinsically Disordered Protein from Unbiased Molecular Dynamics Simulation. *Proc. Natl. Acad. Sci. USA* 116, 20446–20452. doi:10.1073/pnas.1907251116

Spill, Y. (2013). Développement de méthodes d'échantillonnage et traitement bayésien de données continues: nouvelle méthode d'échange de répliques et modélisation de données SAXS. *Ph.D. Thesis, Paris* 7.

Spill, Y. G., Bouvier, G., and Nilges, M. (2013). A Convective Replica-Exchange Method for Sampling New Energy Basins. *J. Comput. Chem.* 34, 132–140. doi:10.1002/jcc.23113

Spill, Y. G., Kim, S. J., Schneidman-Duhovny, D., Russel, D., Webb, B., Sali, A., et al. (2014). Saxs Merge: an Automated Statistical Method to Merge Saxs Profiles Using Gaussian Processes. *J. Synchrotron Radiat.* 21, 203–208. doi:10.1107/s1600577513030117

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

Yang, S., Blachowicz, L., Makowski, L., and Roux, B. (2010). Multidomain Assembled States of Hck Tyrosine Kinase in Solution. *Proc. Natl. Acad. Sci.* 107, 15757–15762. doi:10.1073/pnas.1004569107

Young, J. A., Becker, A. M., Medeiros, J. J., Shapiro, V. S., Wang, A., Farrar, J. D., et al. (2008). The Protein Tyrosine Phosphatase PTPN4/PTP-MEG1, an Enzyme Capable of Dephosphorylating the TCR ITAMs and Regulating NF-Kb, Is Dispensable for T Cell Development And/or T Cell Effector Functions. *Mol. Immunol.* 45, 3756–3766. doi:10.1016/j.molimm.2008.05.023

Zhang, B. D., Li, Y. R., Ding, L. D., Wang, Y. Y., Liu, H. Y., and Jia, B. Q. (2019). Loss of PTPN4 Activates STAT3 to Promote the Tumor Growth in Rectal Cancer. *Cancer Sci.* 110, 2258–2272. doi:10.1111/cas.14031

Zhou, J., Wan, B., Shan, J., Shi, H., Li, Y., and Huo, K. (2013). PTPN4 Negatively Regulates CrkI in Human Cell Lines. *Cell Mol Biol Lett* 18, 297–314. doi:10.2478/s11658-013-0090-3

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: 07 June 2021
doi: 10.3389/fmolb.2021.682858

Structural Basis of the Function of Yariv Reagent—An Important Tool to Study Arabinogalactan Proteins

*Tereza Přerovská[1], Anna Pavlů[1,2], Dzianis Hancharyk[2], Anna Rodionova[2], Anna Vavříková[2] and Vojtěch Spiwok[1]**

[1]Department of Biochemistry and Microbiology, University of Chemistry and Technology, Prague, Czechia, [2]Department of Informatics and Chemistry, University of Chemistry and Technology, Prague, Czechia

Arabinogalactan proteins are very abundant, heavily glycosylated plant cell wall proteins. They are intensively studied because of their crucial role in plant development as well as their function in plant defence. Research of these biomacromolecules is complicated by the lack of tools for their analysis and characterisation due to their extreme heterogeneity. One of the few available tools for detection, isolation, characterisation, and functional studies of arabinogalactan proteins is Yariv reagents. Yariv reagent is a synthetic aromatic glycoconjugate originally prepared as an antigen for immunization. Later, it was found that this compound can precipitate arabinogalactan proteins, namely, their ß-D-(1→3)-galactan structures. Even though this compound has been intensively used for decades, the structural basis of arabinogalactan protein precipitation by Yariv is not known. Multiple biophysical studies have been published, but none of them attempted to elucidate the three-dimensional structure of the Yariv-galactan complex. Here we use a series of molecular dynamics simulations of systems containing one or multiple molecules of ß-D-galactosyl Yariv reagent with or without oligo ß-D-(1→3)-galactan to predict the structure of the complex. According to our model of Yariv-galactan complexes, Yariv reagent forms stacked oligomers stabilized by π-π and CH/π interactions. These oligomers may contain irregularities. Galactan structures crosslink these Yariv oligomers. The results were compared with studies in literature.

Keywords: arabinogalactan proteins (AGPs), Yariv phenylglycoside, molecular dynamics simulation, noncovalent interactions, glycochemistry

OPEN ACCESS

Edited by:
Massimiliano Bonomi, Institut Pasteur, France

Reviewed by:
Jim Pfaendtner, University of Washington, United States
Matteo Salvalaglio, University College London, United Kingdom
Giovanni Grazioso, University of Milan, Italy

****Correspondence:***
Vojtěch Spiwok
spiwokv@vscht.cz

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

Received: *19 March 2021*
Accepted: *21 May 2021*
Published: *07 June 2021*

Citation:
Přerovská T, Pavlů A, Hancharyk D, Rodionova A, Vavříková A and Spiwok V (2021) Structural Basis of the Function of Yariv Reagent—An Important Tool to Study Arabinogalactan Proteins. Front. Mol. Biosci. 8:682858. doi: 10.3389/fmolb.2021.682858

INTRODUCTION

Arabinogalactan proteins (AGPs) represent an extremely heterogeneous group of plant cell wall proteoglycans, which together with moderately glycosylated extensins and minimally glycosylated proline-rich proteins belong to the superfamily of hydroxyproline-rich glycoproteins (HRGPs, Showalter et al., 2010). A general feature of all HRGP members is the presence of hydroxylated proline residues, which is a prerequisite for their further glycosylation (Gorres and Raines, 2010; Nguema-Ona et al., 2014). Despite the increasing amount of discovered chimeric or hybrid AGPs, the general characteristics of AGPs were defined over the years. Among them belong the high amounts of Pro, Ala, Ser, and Thr (altogether known as PAST) regularly arranged in Ala-Pro, Ser-Pro and Thr-Pro dipeptide motifs, which governs AGP specific *O*-glycosylation (Tan et al., 2003; Ma et al., 2017). Commonly, the carbohydrate moiety consists of β-D-(1,3)-galactan backbone with β-D-

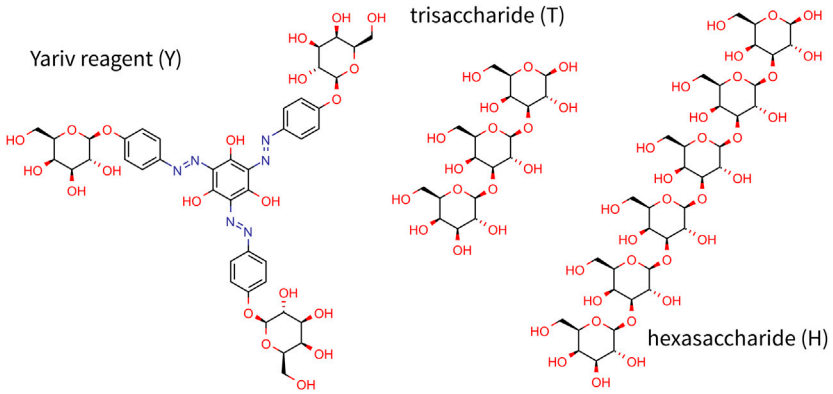


FIGURE 1 | Chemical structure of Yariv reagent and oligosaccharides studied in this study.

(1,6)-galactan side chains, which are often further substituted by arabinose, rhamnose, fucose, or glucuronic acid (Ellis et al., 2010; Knoch et al., 2014; Ma et al., 2018). Moreover, AGPs contain signal sequences directing them to the extracellular location and often can be found anchored to the plasma membrane by the glycosylphosphatidylinositol (GPI) anchor (Seifert and Roberts, 2007). AGPs are being extensively studied especially in higher plants because they have been shown to play an essential role in plant growth, development, reproduction, signaling, and stress responses (Nguema-Ona et al., 2012, Nguema-Ona et al., 2013; Lamport and Várnai 2013; Olmos et al., 2017; Ma et al., 2018; Su and Higashiyama 2018; Seifert 2020).

There are various tools to study these proteins including monoclonal antibodies, β-Yariv reagents, specific degradation of AGP sugar chains, chemical synthesis, and bioinformatics approach (Su and Higashiyama, 2018). While bioinformatics offers quick and high-throughput analysis, the results reflect only the differences in the protein backbone and no information regarding their glycosylation, the most important part in terms of their function, can be obtained in such a way (Showalter et al., 2010; Johnson et al., 2017; Ma et al., 2017). From the experimental point of view, immunolabeling with antibodies or the use of β-Yariv reagents is the most commonly used (Nguema-Ona et al., 2012). Compared to the monoclonal antibodies, the β-Yariv reagents are able (in addition to the visualization of AGPs) also to perturb their function, which is widely exploited in AGP functional studies (Willats and Knox, 1996; Tang et al., 2006; Nguema-Ona et al., 2007; Yu and Zhao, 2012; Olmos et al., 2017; Su and Higashiyama, 2018; Castilleux et al., 2020).

Yariv reagents (**Figure 1**) are synthetic phenylglycosides, which were formerly developed as protein-free precipitatory antigens for determining the content of sugar-binding proteins and their purification (Yariv et al., 1962). Nevertheless, later, certain types of β-Yariv reagents (β-D-glucosyl and β-D-galactosyl) were demonstrated to selectively bind to AGPs (Yariv et al., 1967; Nothnagel 1997). Thus, the ability to bind β-Yariv reagents is also considered a characteristic feature of AGPs. Despite their wide use and attempts to resolve their mode of action, their target structure as well as the mechanism remained elusive for decades (Nothnagel, 1997). Only recently, the target structure has been at least partially clarified. Kitazawa et al. (2013) proved that β-D-galactosyl Yariv reagent interacts with the β-(1,3)-galactan backbone, which has to be longer than five residues for the interaction to occur. Moreover, Sato et al. (2018) showed that the extent of β-(1,6)-galactan substitution affects the Yariv reagent binding ability. Interestingly, the Yariv reagent self-aggregates in the aqueous solution up to approximately 305 units (Nothnagel, 1997; Paulsen et al., 2014). The size of aggregates influences the interaction with AGPs, when the AGP precipitation is known to take place in a solution with ionic strength corresponding to 1% NaCl in which the number of aggregated molecules is approximately 185. On the other hand, 10% NaCl inhibited the precipitation, and the aggregate comprised approximately 125 units (Nothnagel, 1997; Paulsen et al., 2014).

Unfortunately, the mechanism of action still remains largely understudied and more studies are needed to fully understand the nature and functionality of Yariv reagents. To elucidate the structure of Yariv-galactan complexes, we carried out a series of molecular dynamics simulations of systems containing Yariv reagent (namely β-D-galactosyl Yariv, further referred to as Yariv) and/or galactan oligosaccharides in explicitly modeled water. The protein part of AGP was not modeled because it has been shown that Yariv recognizes the carbohydrate part of AGPs (Kitazawa et al., 2013). Galactan trisaccharide was chosen as a minimal model oligosaccharide (**Figure 1**). This choice was driven by the fact that Yariv is selective for 1→3 linked oligomers, for which a trisaccharide is the smallest representative. Another oligosaccharide studied in this study is hexassacharide (**Figure 1**) because it was shown that oligosaccharides with more than five units form stable complexes with the Yariv reagent (Kitazawa et al., 2013).

MATERIALS AND METHODS

All simulations were done in Gromacs package version 2018 (Abraham et al., 2015). Galactooligosaccharides were modeled using the Glycam 06j-1 force field (Kirschner et al., 2008). Yariv

compound was modeled by a manually combined Glycam and General Amber Force Field (Wang et al., 2004). Acpype (Sousa da Silva and Vranken, 2012) was used to convert AMBER files to Gromacs files. Galactose units were modeled by Glycam nonbonded and bonded parameters, except for partial atomic charges. Non-saccharidic part of Yariv compound was modeled by General Amber Force Field in AMBER tools version 16 (Salomon-Ferrer et al., 2013) non-bonded and bonded parameters, except for partial atomic charges. Parameters for the connection between both parts were taken from the analogous parameters of glycoside linkage in Glycam. Partial atomic charges were calculated by an Antechamber routine utilizing the semiempirical AM1-BCC method (Jakalian et al., 2000). Gromacs topology of all studied molecules is available *via* Zenodo (see below).

Systems containing different molecular assemblies were solvated by TIP3P water molecules (Jorgensen et al., 1983). Next, it was minimized by the steepest descent algorithm and equilibrated by 1 ns simulation in a NPT ensemble and 1 ns simulation in a NVT ensemble. This was followed by a 100 ns production simulation. Simulation step was set to 2 fs and all bonds to hydrogens were constrained by the LINCS algorithm (Hess et al., 1997). Electrostatic interactions were modeled by the Particle-Mesh Ewald (PME) (Darden et al., 1993) method with the cutoff set to 1 nm. Temperature and pressure was maintained by Parrinello-Bussi (Bussi et al., 2007) and Parrinello-Raman (Parrinello and Rahman, 1981) algorithm, respectively.

Simulations were analysed using in-house scripts in Python with MDTraj library (McGibbon et al., 2015). Simulation inputs and results (input files for simulations, trajectories without water molecules) are available *via* Zenodo (DOI: 10.5281/zenodo.4767970).

RESULTS AND DISCUSSION

To elucidate the structural organisation of complexes of Yariv reagent with β-D-(1→3)-galactan molecules, we carried out simulations of, in total, 48 systems containing various numbers of Yariv reagent molecules (Y) and carbohydrate molecules. Carbohydrate molecules included trisaccharide (T, β-D-Gal-β-D-(1→3)-Gal-β-D-(1→3)-Gal), or hexasaccharide (H, β-D-Gal-β-D-(1→3)-Gal-β-D-(1→3)-β-D-Gal-β-D-(1→3)-Gal-β-D-(1→3)-β-D-Gal-β-D-(1→3)-Gal). They are further referred to as Y2 for a system with two molecules of Yariv reagent, YT for a system with Yariv reagent and trisaccharide, etc. Initial structures of the systems were assembled manually. In summary, simulated systems included Y2, YT, YH, Y4, Y4T, and Y4H, all in eight replicas.

Our initial simulations (data not shown) showed that noncovalent interactions between a Yariv molecule and a carbohydrate and especially between two Yariv molecules are relatively strong. It would be necessary to run very long simulations to observe relevant structural transitions. Therefore, to map possible modes of interactions between Yariv and carbohydrates we carried out a series of short (100 ns) simulations starting from different initial structures. We believe such simulations are more representative than few long simulations due to long lifetimes of complexes. Initial coordinates were built manually to represent wide diversity in terms of initial distances and orientations of molecules.

First, we were interested in the interactions between two Yariv molecules. The simulated systems of Yariv dimers contained 2073–2083 water molecules. This corresponds to a concentration of Yariv reagent equal to 53–54 mmol/L (approx 28 g/L). This is approximately 30 times higher than concentrations used in precipitation experiments of arabinogalactans, however, in simulations the interactions between Yariv reagent molecules are limited by periodic boundary conditions.

The first Yariv molecule of all snapshots was fitted onto the first snapshot to eliminate its translational and rotational motions. This was possible due to relatively high rigidity of the Yariv molecule (conjugated diazenyl groups). Coordinates of the second molecule were analysed in terms of free-energy-like function. 3D histograms of all carbon atoms of the second Yariv molecule were calculated with 1 Å × 1 Å × 1 Å bins. Next, these values were converted to free-energy-like functions as:

$$A_i = -kT\log P_i$$

where P_i is the histogram count, k is Boltzmann constant, and T is the temperature in Kelvin. Finally, the value of the global minimum was subtracted. The difference between a free energy and the free-energy-like function used in this study is in the fact that the free energy depicts probability of finding a molecule at a certain point, whereas the free-energy-like function depicts the probability of finding any carbon atom at a certain point. The advantage of the free-energy-like function is in its higher resolution. The resulting free-energy-like functions are depicted in **Figure 2**.

Yariv reagent formed stable and almost perfectly parallel dimers (**Figure 2**). Two Yariv reagent molecules interact by π-π stacking *via* all four aromatic moieties, despite different initial monomer orientations. These interactions are relatively strong, but still reversible as indicated by replica seven. In this simulation replica, we observed that the second Yariv molecule (initially at the bottom, depicted as a free-energy-like function in **Figure 2**) migrated to the top face of the first Yariv molecule (depicted as atoms in **Figure 2**). Interestingly, assemblies stabilized by CH/π interactions between the carbohydrate part of one Yariv molecule and the aromatic ring in the second molecule were rare. This can be explained by the fact that galactose forms aromatic CH/π interactions in carbohydrate-protein complexes *via* its C-H bonds on carbons C3, C4, C5, and C6. Such complexes are not parallel. In contrast, glucose forms parallel aromatic CH/π complexes in carbohydrate-protein complexes *via* C-H bonds on carbon atoms C1, C3, and C5 or C2, C4, and C6. In conclusion, Yariv forms dimers that are parallel and stabilized predominantly by π-π stacking between its aromatic moieties. This assembly is relatively stable, nevertheless rearrangements of the assembly are possible in sub-microsecond time scales.

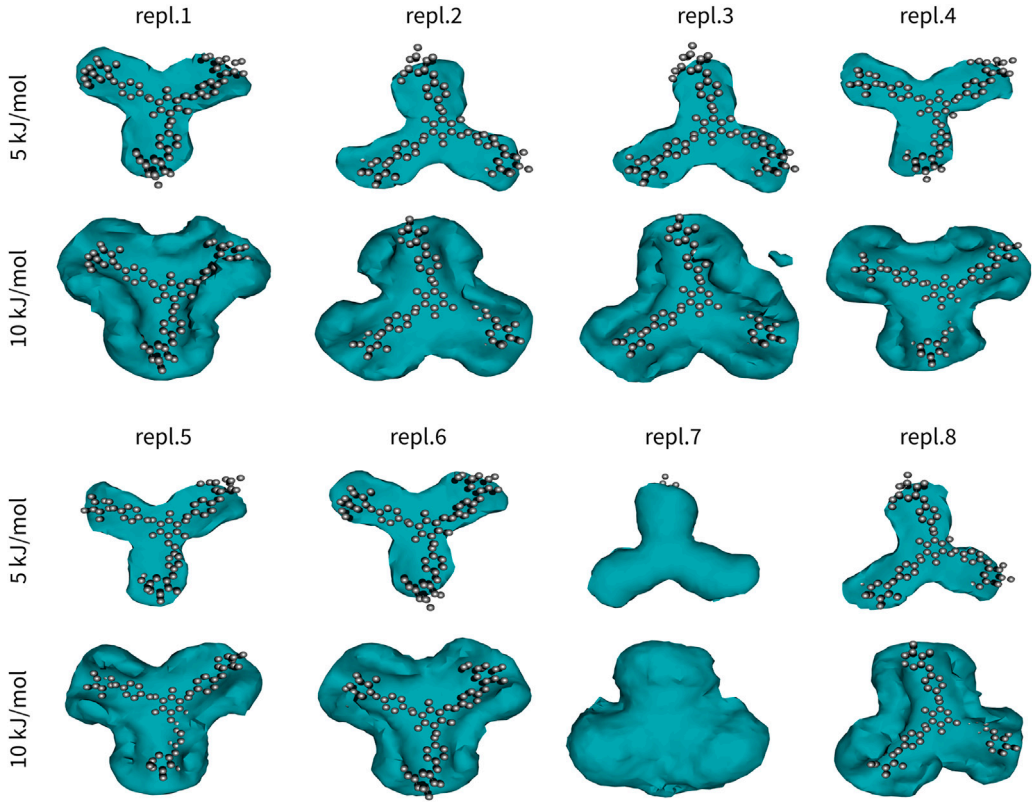


FIGURE 2 | Interactions in Yariv reagent dimer (Y2). One molecule of Yariv was fitted to the initial coordinates. Coordinates of the second molecule were analysed in terms of a free-energy-like function. This function is depicted as isosurfaces at 5 or 10 kJ/mol (relative to the global minimum).

The second series of simulations studied assemblies of a single Yariv reagent with a single trisaccharide or hexasaccharide. Similarly to Y2, also these results were analysed in terms of free-energy-like functions (**Figure 3**). Similarly to Yariv dimers, the complexes of Yariv reagent with oligosaccharides were stable. We observed the migration of trisaccharide from the bottom to the top face of Yariv in five of eight 100-ns-long simulations. Galactooligosaccharides with β-(1→3) linkage are characterized by approximately 120° angle between three adjacent monosaccharide units. They are perfectly aligned with the orientation of aromatic rings (*peripheral-central-peripheral*) in a Yariv reagent molecule. Yariv-galactooligosaccharide complexes were stabilized by carbohydrate-aromatic CH/π interactions (Spiwok, 2017). Free-energy-like functions of YT complexes were triangular. This can be explained by the formation of three possible complexes in which the three adjacent monosaccharide units interact either with *peripheral1-central-peripheral2*, *peripheral2-central-peripheral3*, or *peripheral3-central-peripheral1* aromatic rings. Fast interconversion between these complexes determines the triangular free-energy-like functions of YT complexes.

Complexes of Yariv reagent with hexasaccharide (YH) were comparably strong as YT. Free-energy-like functions were mostly boomerang-shaped (five of eight simulations). This can be explained by the fact that the interconversion between complexes was much slower compared to YT and one assembly was predominant.

Assemblies with more than two molecules included Y4, Y4T, and Y4H. The complexes formed in these systems were visualized as structures after 100 ns (**Figures 4**–**6**). These complexes were formed very quickly (<20 ns) and they were stable in terms of topology along 100 ns simulations. These figures show that Yariv reagent molecules were arranged in parallel, however, these assemblies were not perfectly parallel and contained numerous irregularities. Yariv molecules interacted predominantly *via* π-π stacking of their aromatic molecules. There were also CH/π interactions between the carbohydrate part of one Yariv molecule and an aromatic ring of another molecule. Furthermore, there were CH/π interactions in which aromatic rings were playing both roles—donors and acceptors.

Complexes of multiple Yariv molecules with a saccharide (**Figures 5**, **6**) combined the properties of Yariv tetramers and binary complexes described above. The complexes were formed by a parallel assembly of Yariv molecules with irregularities from a perfectly parallel shape. A molecule of tri- or hexasaccharide was sitting on top of the Yariv molecule which was most exposed to the solvent. Binding of saccharides onto Yariv was slightly different from that in binary complexes. This can be explained by the fact that there are irregularities in the parallel shape of Yariv

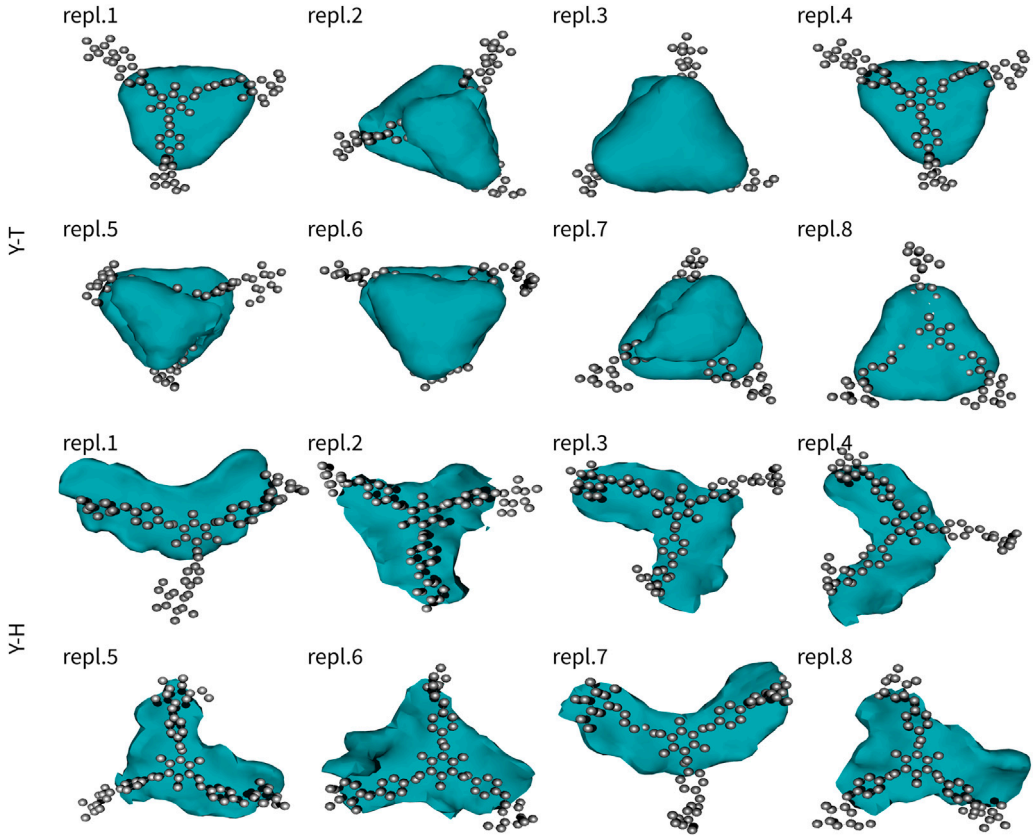


FIGURE 3 | Interactions of Yariv reagent with trisaccharide (YT) and hexasaccharide (YH). The molecule of Yariv was fitted to the initial coordinates. Coordinates of the saccharide molecule were analysed in terms of a free-energy-like function. This function is depicted as isosurfaces at 5 kJ/mol (relative to the global minimum).

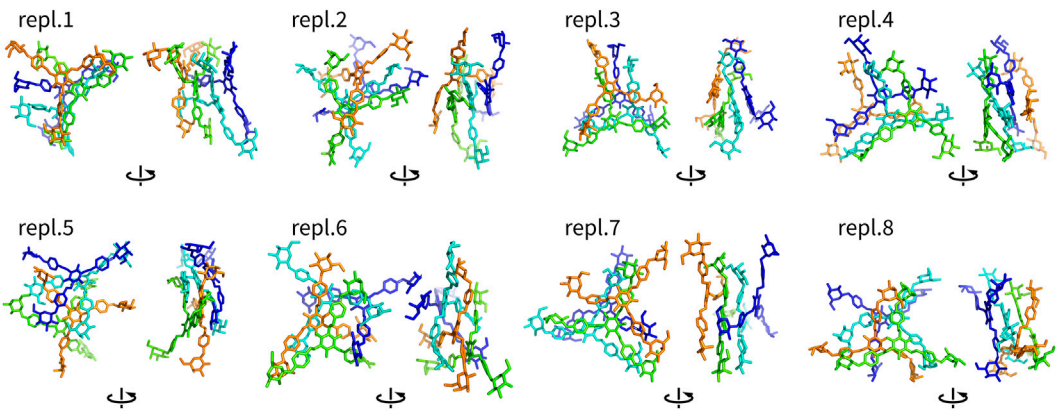


FIGURE 4 | Assemblies Y4 (non-hydrogen atoms) after 100 ns shown in two different orientations rotated by 90°. They are colored by atom ID in the system.

tetramers. These irregularities expose more aromatic groups for interaction with a saccharide and give saccharides more freedom.

One exception was the Y4H complex formed in the replica eight. In this case the oligosaccharide molecule docked into the groove formed by two peripheral and one central aromatic moieties of parallely stacked Yariv tetramer. This complex is stabilized by numerous hydrogen bonds. This assembly can be seen as an alternative model of Yariv-oligosaccharide interactions.

It is important to assess the accuracy of the simulations in this study. This accuracy is determined by the accuracy of molecular mechanics potentials (force fields) and by the completeness of

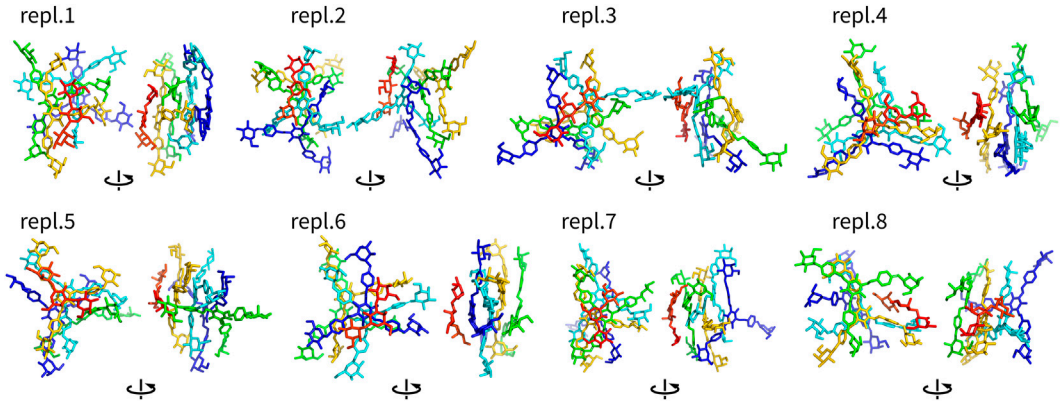


FIGURE 5 | Assemblies Y4T (non-hydrogen atoms) after 100 ns shown in two different orientations rotated by 90°. They are colored by atom ID in the system (Yariv molecules are blue to orange, saccharides are in red).

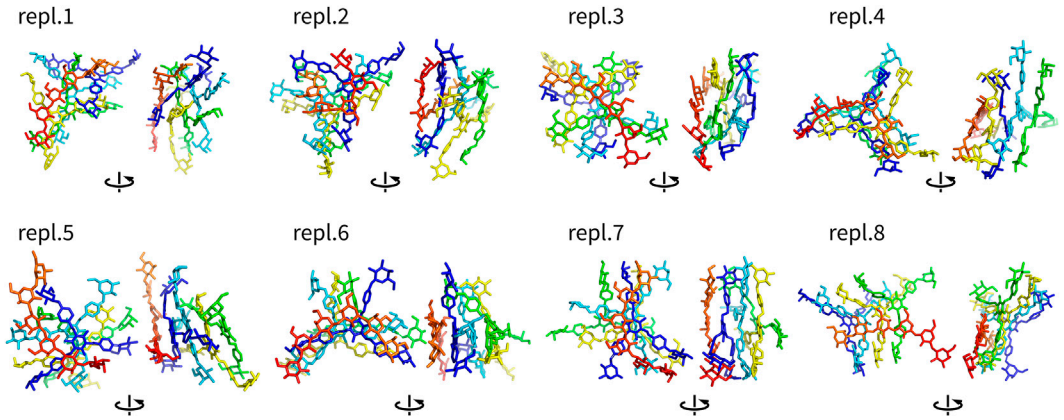


FIGURE 6 | Assemblies Y4H (non-hydrogen atoms) after 100 ns shown in two different orientations rotated by 90°. They are colored by atom ID in the system (Yariv molecules are blue to orange, saccharides are in red).

sampling. The studied complexes were dominated by π-π and CH/π interactions. Comparison of quantum chemical and molecular mechanical energies of sample π-π (Sponer et al., 2006) or CH/π (Spiwok et al., 2005) complexes has shown that these interactions are relatively accurately modeled by the available molecular mechanics force fields. Another issue in carbohydrate modeling is ring puckering. Hexopyranoses may exist in the chair as well as in the boat or skew-boat conformers. The chair structure is predominant for β-D-galactose units. Visual inspection of trajectories revealed that carbohydrates stayed in the chair conformation as expected.

Finally, molecular dynamics simulation suffers limited time scales due to its computational complexity. Here we used multiple replicas of simulated systems differing in the initial structure of the systems rather than running a few long simulations. This was motivated by the necessity to map possible interaction patterns.

Our model of complexes of Yariv reagent with β-(1→3)-galactans is depicted in **Figure 7A** as a schematic view. It is our speculation of the structure of large Yarive-polysaccharide complexes based on the results of our simulations. Yariv forms parallel stacked oligomers. The sizes of these oligomers may vary, but we expect their size in tens or hundreds of units. In simulations of Yariv dimers, we observed a trend of rotation of its units, i.e., one unit is rotated by a few degrees. This rotation seems to be asymmetric (right handed). We speculate that this may explain the helical chirality of Yariv aggregates that has been observed by circular dichroism (Hoshing et al., 2020).

These stacked oligomers are not perfect and contain irregularities. This is probably the reason Yariv reagent and its complexes resisted the application of conventional experimental methods. Some experimental methods of structure elucidation, such as crystallography, require strong periodicity of the studied system. Irregularities in oligomeric structures provide more accessible aromatic rings as platforms for interaction with carbohydrates.

An assembly that cannot be ruled out as a model of Yariv-oligosaccharide interaction is the one formed in the eighth replica

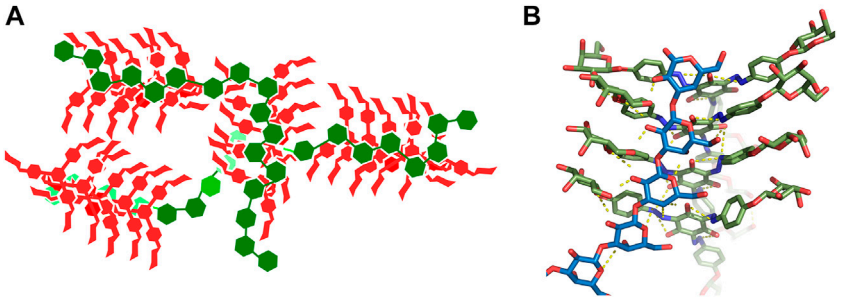


FIGURE 7 | (A)-Schematic view of a model of a complex of Yariv reagent with *β*-(1→3)-galactans. Yariv is in red, saccharides are in green. **(B)**-Detailed view of the Y4H complex in replica eight.

of Y4H simulation (**Figure 6** and **Figure 7B** in detail). We plan to study this assembly in future.

The selectivity of Yariv reagent towards 1→3-linked oligo- and polysaccharides may be explained by the fact that the three adjacent aromatic rings in Yariv and three adjacent monosaccharide units in 1→3-linked oligosaccharide are bent by 120°. For example, in 1→4-linked glycans, the orientation is linear (not bent) and such oligo- or polysaccharide would bind to the Yariv reagent very weakly.

The fact that Yariv complexes are stabilized mostly by π-π and CH/π interactions is in good agreement with the fact that Yariv assemblies are resistant to high ionic strength (Nothnagel, 1997; Paulsen et al., 2014). These interactions are mediated by the hydrophobic parts of both molecules, namely, by aromatic rings and C-H-rich patches of carbohydrates. Due to this, many researchers present these interactions as hydrophobic. It was not a subject of this work to determine whether these interactions are physical attractive interactions or a result of solvation and desolvation. The nature of, for example, CH/π interactions remains a question of debate (physical van der Waals vs. hydrophobic) (Spiwok, 2017).

In conclusion, simulations of systems containing the Yariv reagent with model oligosaccharides provide predictions of main interaction types and structural arrangements in these complexes. We understand that our models are based on simplified systems and short time scales, nevertheless, we believe they can inspire other researchers studying the Yariv reagent to design new biophysical experiments or Yariv derivatives to complete our picture of the function of this useful reagent.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://zenodo.org, DOI: 10.5281/zenodo.4767970.

AUTHOR CONTRIBUTIONS

TP and VS. designed the study, AP, DH, AR, and AV. carried out the simulations and analyzed their results, TP and VS. wrote the manuscript.

ACKNOWLEDGMENTS

Authors would like to thank Czech Science Foundation (project 19-16857S). Computational resources were provided by e-Infrastruktura CZ (LM2018140), the CERIT Scientific Cloud (LM2015085) and ELIXIR-CZ project (LM2018131), provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures.”

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

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* 126, 014101. doi:10.1063/1.2408420

Castilleux, R., Ropitaux, M., Manasfi, Y., Bernard, S., Vicré-Gibouin, M., and Driouich, A. (2020). “Contributions to Arabinogalactan Protein Analysis,” in *The Plant Cell wall: Methods and Protocols*. Editor Z. A. Popper (New York, NY: Springer), 383–402. doi:10.1007/978-1-0716-0621-6_22

Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald: AnN·Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397

Ellis, M., Egelund, J., Schultz, C. J., and Bacic, A. (2010). Arabinogalactan-Proteins: Key Regulators at the Cell Surface?. *Plant Physiol.* 153, 403–419. doi:10.1104/pp.110.156000

Gorres, K. L., and Raines, R. T. (2010). Prolyl 4-hydroxylase. *Crit. Rev. Biochem. Mol. Biol.* 45, 106–124. doi:10.3109/10409231003627991

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* 18, 1463–1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

Hoshing, R., Leeber III, B. W., III, Kuhn, H., Caianiello, D., Dale, B., Saladino, M., et al. (2020). The Chirality of Yariv Reagent Aggregates Correlates with AGP-Binding Ability. *ChemrXiv (preprint)*, 1–25. doi:10.26434/chemrxiv.13154261.v1

Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. (2000). Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* 21, 132–146. doi:10.1002/(sici)1096-987x(20000130)21:2<132::aid-jcc5>3.0.co;2-p

Johnson, K. L., Cassin, A. M., Lonsdale, A., Wong, G. K.-S., Soltis, D. E., Miles, N. W., et al. (2017). Insights into the Evolution of Hydroxyproline-Rich Glycoproteins from 1000 Plant Transcriptomes. *Plant Physiol.* 174, 904–921. doi:10.1104/pp.17.00295

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

Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., et al. (2008). GLYCAM06: A Generalizable Biomolecular Force Field. Carbohydrates. *J. Comput. Chem.* 29, 622–655. doi:10.1002/jcc.20820

Kitazawa, K., Tryfona, T., Yoshimi, Y., Hayashi, Y., Kawauchi, S., Antonov, L., et al. (2013). β-Galactosyl Yariv Reagent Binds to the β-1,3-Galactan of Arabinogalactan Proteins. *Plant Physiol.* 161, 1117–1126. doi:10.1104/pp.112.211722

Knoch, E., Dilokpimol, A., and Geshi, N. (2014). Arabinogalactan Proteins: Focus on Carbohydrate Active Enzymes. *Front. Plant Sci.* 5, 198. doi:10.3389/fpls.2014.00198

Lamport, D. T. A., and Várnai, P. (2013). Periplasmic Arabinogalactan Glycoproteins Act as a Calcium Capacitor that Regulates Plant Growth and Development. *New Phytol.* 197, 58–64. doi:10.1111/nph.12005

Ma, Y., Yan, C., Li, H., Wu, W., Liu, Y., Wang, Y., et al. (2017). Bioinformatics Prediction and Evolution Analysis of Arabinogalactan Proteins in the Plant Kingdom. *Front. Plant Sci.* 8, 66. doi:10.3389/fpls.2017.00066

Ma, Y., Zeng, W., Bacic, A., and Johnson, K. (2018). "AGPs through Time and Space," in *Annual Plant Reviews Online* (Chichester, United Kingdom: American Cancer Society), 767–804. doi:10.1002/9781119312994.apr0608

McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., et al. (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical J.* 109, 1528–1532. doi:10.1016/j.bpj.2015.08.015

Nguema-Ona, E., Bannigan, A., Chevalier, L., Baskin, T. I., and Driouich, A. (2007). Disruption of Arabinogalactan Proteins Disorganizes Cortical Microtubules in the Root of *Arabidopsis thaliana*. *Plant J.* 52, 240–251. doi:10.1111/j.1365-313X.2007.03224.x

Nguema-Ona, E., Coimbra, S., Vicré-Gibouin, M., Mollet, J.-C., and Driouich, A. (2012). Arabinogalactan Proteins in Root and Pollen-Tube Cells: Distribution and Functional Aspects. *Ann. Bot.* 110, 383–404. doi:10.1093/aob/mcs143

Nguema-Ona, E., Vicré-Gibouin, M., Cannesan, M.-A., and Driouich, A. (2013). Arabinogalactan Proteins in Root-Microbe Interactions. *Trends Plant Sci.* 18, 440–449. doi:10.1016/j.tplants.2013.03.006

Nguema-Ona, E., Vicré-Gibouin, M., Gotté, M., Plancot, B., Lerouge, P., Bardor, M., et al. (2014). Cell wall O-Glycoproteins and N-Glycoproteins: Aspects of Biosynthesis and Function. *Front. Plant Sci.* 5, 499. doi:10.3389/fpls.2014.00499

Nothnagel, E. A. (1997). "Proteoglycans and Related Components in Plant Cells," in *International Review of Cytology*. Editor K. W. Jeon (San Diego, CA: Academic Press), 195–291. doi:10.1016/s0074-7696(08)62118-x

Olmos, E., García De La Garma, J., Gomez-Jimenez, M. C., and Fernandez-Garcia, N. (2017). Arabinogalactan Proteins Are Involved in Salt-Adaptation and Vesicle Trafficking in Tobacco By-2 Cell Cultures. *Front. Plant Sci.* 8, 1092. doi:10.3389/fpls.2017.01092

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

Paulsen, B. S., Craik, D. J., Dunstan, D. E., Stone, B. A., and Bacic, A. (2014). The Yariv Reagent: Behaviour in Different Solvents and Interaction with a Gum Arabic Arabinogalactanprotein. *Carbohydr. Polym.* 106, 460–468. doi:10.1016/j.carbpol.2014.01.009

Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013). An Overview of the Amber Biomolecular Simulation Package. *Wires Comput. Mol. Sci.* 3, 198–210. doi:10.1002/wcms.1121

Sato, K., Hara, K., Yoshimi, Y., Kitazawa, K., Ito, H., Tsumuraya, Y., et al. (2018). Yariv Reactivity of Type II Arabinogalactan from Larch wood. *Carbohydr. Res.* 467, 8–13. doi:10.1016/j.carres.2018.07.004

Seifert, G. J. (2020). On the Potential Function of Type II Arabinogalactan O-Glycosylation in Regulating the Fate of Plant Secretory Proteins. *Front. Plant Sci.* 11, 563735. doi:10.3389/fpls.2020.563735

Seifert, G. J., and Roberts, K. (2007). The Biology of Arabinogalactan Proteins. *Annu. Rev. Plant Biol.* 58, 137–161. doi:10.1146/annurev.arplant.58.032806.103801

Showalter, A. M., Keppler, B., Lichtenberg, J., Gu, D., and Welch, L. R. (2010). A Bioinformatics Approach to the Identification, Classification, and Analysis of Hydroxyproline-Rich Glycoproteins. *Plant Physiol.* 153, 485–513. doi:10.1104/pp.110.156554

Sousa da Silva, A. W., and Vranken, W. F. (2012). ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. Notes* 5, 367. doi:10.1186/1756-0500-5-367

Spiwok, V., Lipovová, P., Skálová, T., Vondrácková, E., Dohnálek, J., Hašek, J., et al. (2005). Modelling of Carbohydrate-Aromatic Interactions: Ab Initio Energetics and Force Field Performance. *J. Comput. Aided Mol. Des.* 19, 887–901. doi:10.1007/s10822-005-9033-z

Spiwok, V. (2017). CH/π Interactions in Carbohydrate Recognition. *Molecules* 22, 1038. doi:10.3390/molecules22071038

Sponer, J., Jurecka, P., Marchan, I., Luque, F. J., Orozco, M., and Hobza, P. (2006). Nature of Base Stacking: Reference Quantum-Chemical Stacking Energies in Ten Unique B-DNA Base-Pair Steps. *Chemistry* 12, 2854–2865. doi:10.1002/chem.200501239

Su, S., and Higashiyama, T. (2018). Arabinogalactan Proteins and Their Sugar Chains: Functions in Plant Reproduction, Research Methods, and Biosynthesis. *Plant Reprod.* 31, 67–75. doi:10.1007/s00497-018-0329-2

Tan, L., Leykam, J. F., and Kieliszewski, M. J. (2003). Glycosylation Motifs that Direct Arabinogalactan Addition to Arabinogalactan-Proteins. *Plant Physiol.* 132, 1362–1369. doi:10.1104/pp.103.021766

Tang, X.-C., He, Y.-Q., Wang, Y., and Sun, M.-X. (2006). The Role of Arabinogalactan Proteins Binding to Yariv Reagents in the Initiation, Cell Developmental Fate, and Maintenance of Microspore Embryogenesis in *Brassica Napus* L. Cv. Topas. *J. Exp. Bot.* 57, 2639–2650. doi:10.1093/jxb/erl027

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and Testing of a General Amber Force Field. *J. Comput. Chem.* 25, 1157–1174. doi:10.1002/jcc.20035

Willats, W. G. T., and Knox, J. P. (1996). A Role for Arabinogalactan-Proteins in Plant Cell Expansion: Evidence from Studies on the Interaction of Beta-Glucosyl Yariv Reagent with Seedlings of *Arabidopsis thaliana*. *Plant J.* 9, 919–925. doi:10.1046/j.1365-313x.1996.9060919.x

Yariv, J., Lis, H., and Katchalski, E. (1967). Precipitation of Arabic Acid and Some Seed Polysaccharides by Glycosylphenylazo Dyes. *Biochem. J.* 105, 1C–2C. doi:10.1042/bj1050001c

Yariv, J., Rapport, M., and Graf, L. (1962). The Interaction of Glycosides and Saccharides with Antibody to the Corresponding Phenylazo Glycosides. *Biochem. J.* 85, 383–388. doi:10.1042/bj0850383

Yu, M., and Zhao, J. (2012). The Cytological Changes of Tobacco Zygote and Proembryo Cells Induced by Beta-Glucosyl Yariv Reagent Suggest the Involvement of Arabinogalactan Proteins in Cell Division and Cell Plate Formation. *BMC Plant Biol.* 12, 126. doi:10.1186/1471-2229-12-126

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: 28 June 2021
doi: 10.3389/fmolb.2021.676235

Ubiquitin Interacting Motifs: Duality Between Structured and Disordered Motifs

Matteo Lambrughi[1,2], *Emiliano Maiani*[1], *Burcu Aykac Fas*[1], *Gary S. Shaw*[3], *Birthe B. Kragelund*[4], *Kresten Lindorff-Larsen*[4], *Kaare Teilum*[4], *Gaetano Invernizzi*[4†] *and Elena Papaleo*[1,5*]

[1]Computational Biology Laboratory, Danish Cancer Society Research Center, Copenhagen, Denmark, [2]Department of Biotechnology and Bioscience, University of Milano-Bicocca, Milano, Italy, [3]Department of Biochemistry, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, ON, Canada, [4]Structural Biology and NMR Laboratory and The Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark, [5]Cancer Systems Biology, Section for Bioinformatics, Department of Health and Technology, Technical University of Denmark, Lyngby, Denmark

OPEN ACCESS

Edited by:
Gregory Bowman,
Washington University School of Medicine in St. Louis, United States

Reviewed by:
Sophie Sacquin-Mora,
UPR9080 Laboratoire de Biochimie Théorique (LBT), France
Igor N. Berezovsky,
Bioinformatics Institute (A*STAR), Singapore

***Correspondence:**
Elena Papaleo
elenap@cancer.dk
elpap@dtu.dk

†Present address:
Department of Proteins and Peptides Biophysics, Novo Nordisk A/S, Måløv, Denmark

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

Received: 04 March 2021
Accepted: 14 May 2021
Published: 28 June 2021

Citation:
Lambrughi M, Maiani E, Aykac Fas B, Shaw GS, Kragelund BB, Lindorff-Larsen K, Teilum K, Invernizzi G and Papaleo E (2021) Ubiquitin Interacting Motifs: Duality Between Structured and Disordered Motifs. Front. Mol. Biosci. 8:676235. doi: 10.3389/fmolb.2021.676235

Ubiquitin is a small protein at the heart of many cellular processes, and several different protein domains are known to recognize and bind ubiquitin. A common motif for interaction with ubiquitin is the Ubiquitin Interacting Motif (UIM), characterized by a conserved sequence signature and often found in multi-domain proteins. Multi-domain proteins with intrinsically disordered regions mediate interactions with multiple partners, orchestrating diverse pathways. Short linear motifs for binding are often embedded in these disordered regions and play crucial roles in modulating protein function. In this work, we investigated the structural propensities of UIMs using molecular dynamics simulations and NMR chemical shifts. Despite the structural portrait depicted by X-crystallography of stable helical structures, we show that UIMs feature both helical and intrinsically disordered conformations. Our results shed light on a new class of disordered UIMs. This group is here exemplified by the C-terminal domain of one isoform of ataxin-3 and a group of ubiquitin-specific proteases. Intriguingly, UIMs not only bind ubiquitin. They can be a recruitment point for other interactors, such as parkin and the heat shock protein Hsc70-4. Disordered UIMs can provide versatility and new functions to the client proteins, opening new directions for research on their interactome.

Keywords: molecular dynamics, peptide arrays, ubiquitin, short linear motifs, moonlight functions, intrinsic disorder

INTRODUCTION

Protein biochemistry relied for a long time on the paradigm that a protein's function is tied to its three-dimensional structure. Over the past 20 years, several proteins or regions in proteins that do not fit within the structure-function paradigm have been reported (Wright and Dyson, 1999; Chen and Kriwacki, 2018; Milles et al., 2018). They are known as intrinsically disordered proteins (IDPs) or regions (IDRs). IDPs and IDRs lack stable tertiary contacts, are highly dynamic, pliable, and typically do not exhibit stable secondary structures. Proteins containing IDRs constitute 30–44% of eukaryotic proteomes (Perdigão et al., 2015). They attain multiple and chameleon conformations for interactions with different partners (Wright and Dyson, 2014; Bugge et al., 2020). Consequently, the modulation of the structural landscape of an IDP can result in opposing actions on different — or

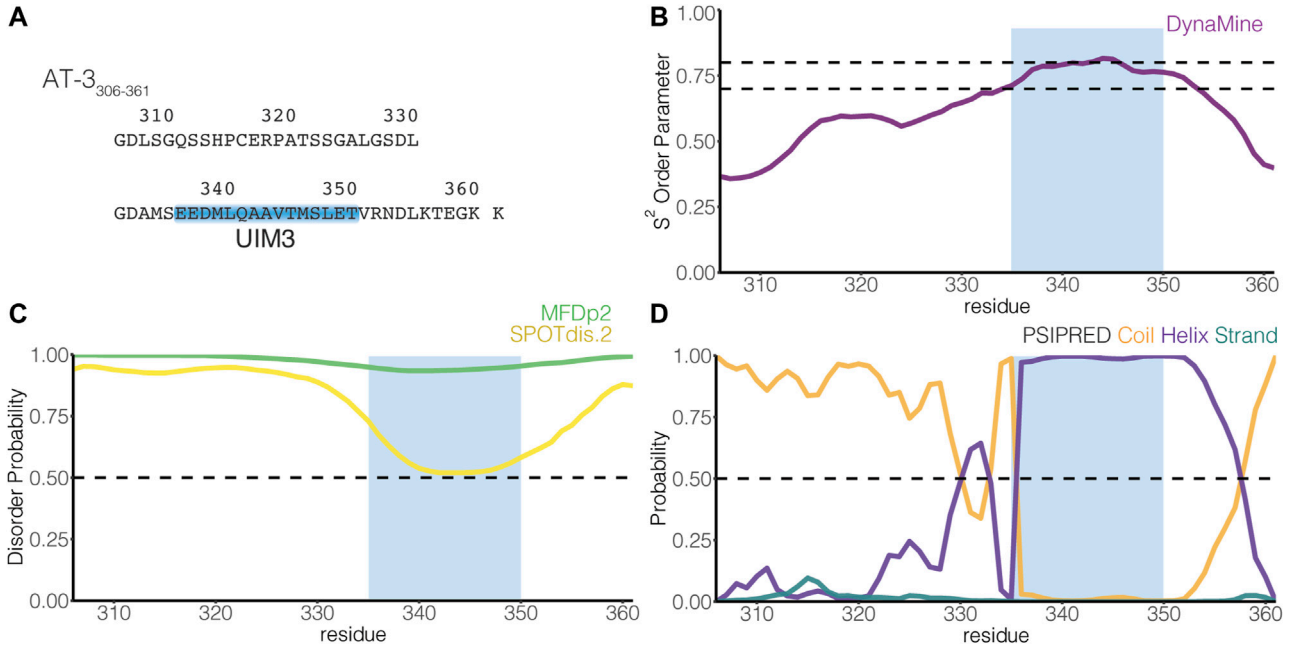


FIGURE 1 | The C-terminus of the UIM3 isoform of ataxin-3 is predicted disordered and includes a UIM. **(A)** The figure illustrates the amino acid sequence of AT-$3_{306-361}$. The blue box indicates UIM3 of AT-3 (residues E336-T350). **(B)** The panel shows the residue-wise prediction of the S^2 order parameter in AT-$3_{306-361}$ from *DynaMine* (purple). The dashed lines show the thresholds of S^2 predicted scores to classify residues as ordered ($S^2 \geq 0.75$), context-dependent ($0.65 < S^2 < 0.75$), and disordered ($S^2 \leq 0.65$). **(C)** The panel shows the residue-wise prediction of disorder propensity in AT-$3_{306-361}$ from the *Multilayered Fusion-based Disorder predictor v. 2.00* (MFDp2, green), and *SPOT-Disorder2* (SPOTdis.2, yellow) predictors. The dashed line shows the threshold of disorder probability to classify the residues as disordered (≥ 0.5) and ordered (< 0.5). **(D)** The plot shows the residue-wise prediction of secondary structure propensity (Coil orange, Helix violet, Strand teal) from *PSI-PRED*. The dashed line shows the threshold of secondary structure probability for each structural class to classify the residues as coil, helix or strand (≥ 0.5). The light blue box indicates UIM3. The predictors report AT-$3_{306-361}$ as a mainly disordered tract with propensity to order and helical structures in proximity of UIM3.

even the same — binding partners, making them elusive, but attractive targets to study (Metallo, 2010; Flock et al., 2014). IDPs and IDRs can also be involved in allosteric mechanisms with key roles in many processes, including modulation of protein-protein interactions and catalytic activities of enzymes (Ma et al., 2011; Li et al., 2017; Berlow et al., 2018; Guarnera and Berezovsky, 2019; Tee et al., 2020).

IDPs and IDRs often interact with binding partners through short stretches of conserved residues, called short linear motifs (SLiMs), embedded in otherwise non-conserved regions (Davey et al., 2012; Van Der Lee et al., 2014). The occurrence of two or more SLiMs in the same IDP/IDR can increase the interaction strength via avidity by multivalent interactions (Van Roey et al., 2014; Fung et al., 2018). Although individual SLiMs are short and mostly participate in transient interactions, they are essential to protein binding specificity and function (Bugge et al., 2020; Kumar et al., 2020).

Some functional motifs of proteins that were traditionally defined as helical elements have been recently reclassified as disordered SLiMs, such as the Bcl-2 Homology 3 motifs (Hinds et al., 2007; Aouacheria et al., 2015). Another well-known functional motif traditionally considered to have a high helical propensity (Scott et al., 2015) is the so-called Ubiquitin Interacting Motif (UIM) or 'LALAL-motif'. UIMs are motifs of approximately 20 residues and were described for the first time in the 26S proteasome subunit PSD4/RPN-10 to bind ubiquitin (Young et al., 1998; Hofmann and Falquet, 2001), now representing the archetypal UIM in the families of ubiquitin binding domains (Scott et al., 2015). UIMs can be found, often in tandem or triplets, in a multitude of proteins involved in ubiquitination, ubiquitin metabolism, or that interact with ubiquitin-like modifiers (Buchberger, 2002). UIM binding partners are not limited to ubiquitin. As an example, ubiquitin-like proteins involved in autophagy feature an interface to recruit UIMs (Marshall et al., 2019; Sora et al., 2020). The UIM consensus motif is X-Ac-Ac-Ac-X-Φ-X-X-Ala-Φ-X-X-Ser-X-X-Ac-X, where Φ represents any hydrophobic residues (often Leu or Ile), Ac represents an acidic residue (Glu, Asp), and X loosely conserved positions (Hofmann and Falquet, 2001; Scott et al., 2015).

Among different UIMs, we focused our attention on the poorly characterized UIM within the C-terminus (residues 306–361) of the human ataxin-3 (AT-3). AT-3 is a multi-domain polyglutamine deubiquitinating enzyme used as a model system to study polyglutamine neurodegenerative diseases (Burnett et al., 2003; Carvalho et al., 2018; Invernizzi et al., 2012). AT-3 contains two UIM regions (UIM1 and UIM2) in the central part of the protein, surrounded by disordered regions (Burnett et al., 2003; Invernizzi et al., 2013; Masino et al., 2003; Sicorello et al., 2018, 2021). AT-3 also undergoes alternative splicing, and its isoforms differ in the C-terminus (Harris et al., 2010). Among the main isoforms, one isoform

contains a third UIM, called UIM3 (**Figure 1A** (Goto et al., 1997; Bettencourt et al., 2010). The UIM3-containing isoform is widely expressed and appears to be the predominant form in the human brain (Ichikawa et al., 2001; Harris et al., 2010). Furthermore, AT-3 UIMs are involved in multivalent binding to the Ubl domain of the E3 ubiquitin ligase parkin (Bai et al., 2013; Aguirre et al., 2018). It has also been suggested that the three UIMs of AT-3 interact with the heat shock protein Hsc70-4 in *Drosophila melanogaster* (Johnson et al., 2020).

Recent advances in all-atom molecular dynamics (MD) simulations in terms of enhanced sampling (Abrams and Bussi, 2013; Spiwok et al., 2015; Bonomi et al., 2017; Sugita et al., 2019; Bussi and Laio, 2020) and physical models for disordered proteins (Best, 2017; Huang and MacKerell, 2018) offer a possibility to unveil heterogeneous conformational ensembles at the atomic level. The presence of multiple UIMs in the disordered C-terminus of AT-3 that are involved in the binding of different interaction partners makes this protein a good model to investigate the structural propensities of UIM using molecular dynamics simulations and chemical shifts from NMR.

We here report a study on the structural propensity and dynamics of the C-terminus of the UIM3-containing isoform of AT-3 (residues 306–361, AT-$3_{306\text{-}361}$). We used two different methods to enhance the sampling of the MD simulations based on temperature exchange or bias along with selected collective variables. We also employed three different force fields (available at the time we performed the simulations) suitable to study disordered/unfolded states of proteins (Best and Mittal, 2010; Knott and Best, 2012; Lindorff-Larsen et al., 2012; Best et al., 2014). The simulation results for AT-$3_{306\text{-}361}$ were then been compared to NMR data for other UIMs in solution (Sgourakis et al., 2010; Lim et al., 2011; Lange et al., 2012; Anamika et al., 2014; Shi et al., 2014; Wen et al., 2014; Sicorello et al., 2018) or to NMR data recorded in this work. In addition, we validated the simulations against previously published NMR chemical shifts of a construct of AT-3 including UIM3 (Bai et al., 2013).

We find that UIM-containing regions can account for both stable helical conformations and more disordered ones, which, in turn, are the more pliable toward a wider range of interactors beyond ubiquitin itself. Thus, our study provides a broader view on the ubiquinome through uncovering an enhanced structural heterogeneity within the groups of UIMs.

MATERIALS AND METHODS

Bioinformatic Analysis

For the sequence-based prediction of secondary structure propensity, we used the PSIPRED predictor (Jones, 1999). We performed disorder prediction from the amino acid sequence, using DynaMine (Cilia et al., 2013), Multilayered Fusion-based Disorder predictor v. 2.00 (MFDp2, (Mizianty et al., 2013), and SPOT-Disorder2 (Hanson et al., 2019). MFDp2 is a meta-method that combines disorder probabilities predicted at residue- and sequence-level by MFDp and DisCon, respectively, and uses post-processing filters and sequence alignment. SPOT-Disorder2 combines long short-term memory with deep bidirectional neural networks to capture non-local and long-range interactions, integrating information from evolutionary profiles of aligned sequences. DynaMine allows high-quality predictions of protein backbone dynamics using an accurate NMR data set for training.

Replica-Exchange Molecular Dynamics Simulations

REMD simulations were performed by GROMACS (Groninger MAchine for Chemical Simulation) using a conformation of the C-terminus of AT-3 (56 residues, 306–361, AT-$3_{306\text{-}361}$) initially generated with Crystallography and NMR System version 1.3 (Brunger, 2007) as the starting structure. We further imposed a helical structure for the region E336-T357, according to the secondary structure prediction by PSIPRED, using MODELLER 9.14 (Eswar et al., 2007). In particular, we selected the model that lacked intermolecular side-chain contacts (defined as intramolecular contacts at a distance in sequence over three residues).

The models were soaked in a dodecahedron box of water molecules with periodic boundary conditions, with a minimal distance for the protein atoms from the box edges of at least 14 Å. We applied the Particle-Mesh Ewald method (Darden et al., 1993) with a 1.2 Å grid spacing. Van der Waals and Coulomb interactions were truncated at 12 Å. Na^+ and Cl^- counterions were added to the system to neutralize the overall charge and to simulate a physiological ionic strength (i.e., 150 mM).

Each system was initially relaxed by 10,000 steps of energy minimization by the steepest descent method. The optimization step was followed by 50 ps of solvent equilibration at 300 K, while restraining the protein atomic positions using a harmonic potential. The systems were subsequently simulated for five ns at 300 K at a constant pressure of 1 bar (NPT ensemble) with coupling constants of 5 and 10 ps, respectively. From the NPT trajectories, we selected a conformation with the volume close to the average volume of the trajectory and used as the starting point for the subsequent NVT preparatory step at 300 K for 20 ns. The 64 initial conformations for REMD simulations were selected from different points (between 10 and 20 ns) along the NVT trajectory using the v-rescale thermostat (Bussi et al., 2007). Other details are reported in the parameter files in the GitHub repository.

In the temperature REMD scheme a number of different copies (replicas) of the system were simulated in parallel at different temperatures and exchanges of configurations are attempted periodically between pairs of replicas. The advantage of this method is that if the trajectory is temporarily trapped in a local minimum can exchange with a replica at a higher temperature and cross high-energy barriers. We carried out REMD simulations using 64 replicas, each replica for 50 ns for a collective simulation time of 3.2 μs. Each replica was run at a different temperature in the range 299–360 K. We selected the temperature spacing between each neighboring replica to ensure an exchange probability higher than 0.2. The replica-exchanges were attempted every ten ps.

Well-Tempered-Metadynamics Simulations

The WT-metaD (Barducci et al., 2008) simulations were performed using GROMACS and the open-source, community-developed PLUMED library (Bonomi et al., 2009; PLUMED Consortium, 2019). In the WT-metaD simulations, the sampling of the free energy surface is enhanced by adding a history-dependent potential to a set of collective variables (CVs). Similar approaches have been applied to simulations of other intrinsically disordered proteins and peptides (Do et al., 2014; Palazzesi et al., 2015). We employed two CVs in our simulations, i.e., 1) the Cα radius of gyration, and 2) *alphabeta*, a CV that measures the similarity of each ψ dihedral angle of AT-$3_{306\text{-}361}$ to a reference value of 0.7854 rad, which corresponds to *a*-helix. Gaussian potentials with an initial height of 0.12 kcal/mol were added to the time-dependent potential every two ps. We used an initial bias factor of four for rescaling the Gaussian height following the WT-metaD scheme. In addition, we used Gaussian widths of 0.2 and one for each CV, respectively. We collected one-μs WT-metaD simulations. We used an extended and disordered conformation of the peptide generated by Profasi (Irbäck and Mohanty, 2006) as the initial structure for the WT-metaD simulations.

Force Fields and Water Models Employed in the REMD and WT-metaD Simulations

For the REMD simulations, we employed four different combinations of protein force fields and water models in our simulations: 1) Amber ff03w [ff03w (Best and Mittal, 2010)] with TIP4P/2005 (Abascal and Vega, 2005), 2) Amber ff03ws [ff03ws (Best et al., 2014)] with TIP4P/2005, in which the protein–water pair interactions have been modified to improve the description of disordered proteins, 3) CHARMM22* (Piana et al., 2011) with TIP3P (CHARMM22*$_1$) (Jorgensen et al., 1983) or 4) TIPS3P (CHARMM22*$_2$) (MacKerell, et al., 1998). WT-metaD was carried out only for ff03w, ff03ws, and CHARMM22*$_2$.

Analyses of the Simulations

The replica at 304 K was used for the analysis. To study the temperature distributions, we converted each replica to be continuous to the simulation time to follow each replica through the temperature space. We used DSSP (Kabsch and Sander, 1983) to estimate the helical content. We used MDAnalysis (Michaud-Agrawal et al., 2011) to calculate the root mean square deviation (RMSD) of UIM3 of AT-$3_{306\text{-}361}$ with respect to the starting helical conformation. We considered the Cβ atom of A343 and the backbone (Cα, C, O, N) atoms of the residues E336-T350 of UIM3 for rigid body superposition and the RMSD calculation.

For the WT-metaD simulations, we reconstructed the one-dimensional free energy landscape from the deposited bias during the simulation with a stride value of 10,000. We extracted four ensembles of structures of AT-$3_{306\text{-}361}$ from the CHARMM22*$_2$ metadynamics trajectory with alphabeta values in the ranges of 1) 9–17, 2) 18–23, 3) 24–30, and 4) 31–34, respectively. On these ensembles, we estimated the propensity to helical structures using the DSSP dictionary (Kabsch and Sander, 1983) and including *a*-helix, *π*-helix and 3.10 helix in the analyses. We applied the MDplot R/CRAN package (Margreitter and Oostenbrink, 2017) to calculate a residue-wise persistence degree of helical secondary structures. On the ensembles selected from the CHARMM22*$_2$ metadynamics trajectory, we used MDAnalysis to calculate the RMSD of UIM3 of AT-$3_{306\text{-}361}$ (residues 336–350) with respect to: 1) the starting structure of AT-$3_{306\text{-}361}$ used for the REMD simulations, 2) the experimental structure of yeast vps27 UIM1 [residue E259-E273, PDB entry 1Q0W (Swanson et al., 2003)], human proteasome subunit S5a UIM1 [residue A212-E226, PDB entry 1YX5 (Wang et al., 2005)] and UIM2 [residue E283-G297, PDB entry 1YX6 (Wang et al., 2005)], and mouse RAP80 UIM1 [residues E81-E95, PDB entry 3A1Q (Sato et al., 2009)] in complex with ubiquitin. We used the same subset of atoms for structural alignment and RMSD calculations, i.e., the Cβ atom of A343 and the backbone (Cα, C, O, N) atoms of residues E336-T350 of AT-$3_{306\text{-}361}$.

Comparison to the Available Chemical Shifts of AT-$3_{306\text{-}361}$

To evaluate the REMD ensembles, we calculated the backbone chemical shifts as a function of the simulation time using PPM (Li and Brüschweiler, 2012) and compared them to the available NMR backbone chemical shifts for a construct of AT-3 including UIM3 [residues 194–361 (Bai et al., 2013)]. To compare the calculated backbone chemical shifts with the experimental ones, we used a reduced χ^2 metric as previously described (Papaleo et al., 2018), using the Python package delta_cs (Sora et al., 2021). The reduced χ^2 relates the squared deviation between the predicted and experimental value and normalized by the variance of the chemical shift predictor for each type of chemical shift and the total number of chemical shifts. Lower values of χ^2_{red} metric indicate a better agreement between experimental and calculated chemical shifts.

Protein Purification

We produced recombinant yeast ubiquitin in E. coli strain BL21 using a pMCSG7vector. Ubiquitin was expressed as a 6X histidine (6His)-TEV N-tagged fusion protein by the addition of 1 mM IPTG and incubation 5 h at 37°C. Cells were harvested, resuspended in a lysis buffer (50 mM Tris pH 8.0, 150 mM NaCl, 10 mM imidazole) plus protease inhibitor mixture (Roche), and disrupted by sonication. 6His-TEV-Ubiquitin was affinity purified with Ni Sepharose 6Fast Flow (GE Healthcare) and eluted with 20 mM Na2HPO4.2H2O, 0.5M NaCl, 500 mM imidazole, pH7.4.

For the construct of human AT-3 including residues 182–291 (AT-$3_{182-291}$) we cloned it in frame with glutathione S-transferase (GST) in a pGEX-6P-1 (GE Healthcare LifeSciences, Little Chalfont, England) plasmid and expressed in E. coli BL21 Codon Plus strain (Stratagene, La Jolla, CA, United States) in auto-inducing growth minimal medium (Tyler et al., 2005). For the production of ^{15}N labeled proteins, we included $^{15}NH_4Cl$ or $(^{15}NH_4)_4SO_4$ 1 g/l as the sole nitrogen source. For $^{15}N^{13}C$ labeled proteins, we added $^{15}NH_4Cl$ 1 g/l or $(^{15}NH_4)_4SO_4$ 1 g/l and substituted the carbon source with a solution of 0.4% ^{13}C-glucose. Cells were harvested, resuspended in a lysis buffer (50 mM KH_2PO_4, 50 mM Na_2HPO_4, 300 mM NaCl, pH 7.4) to

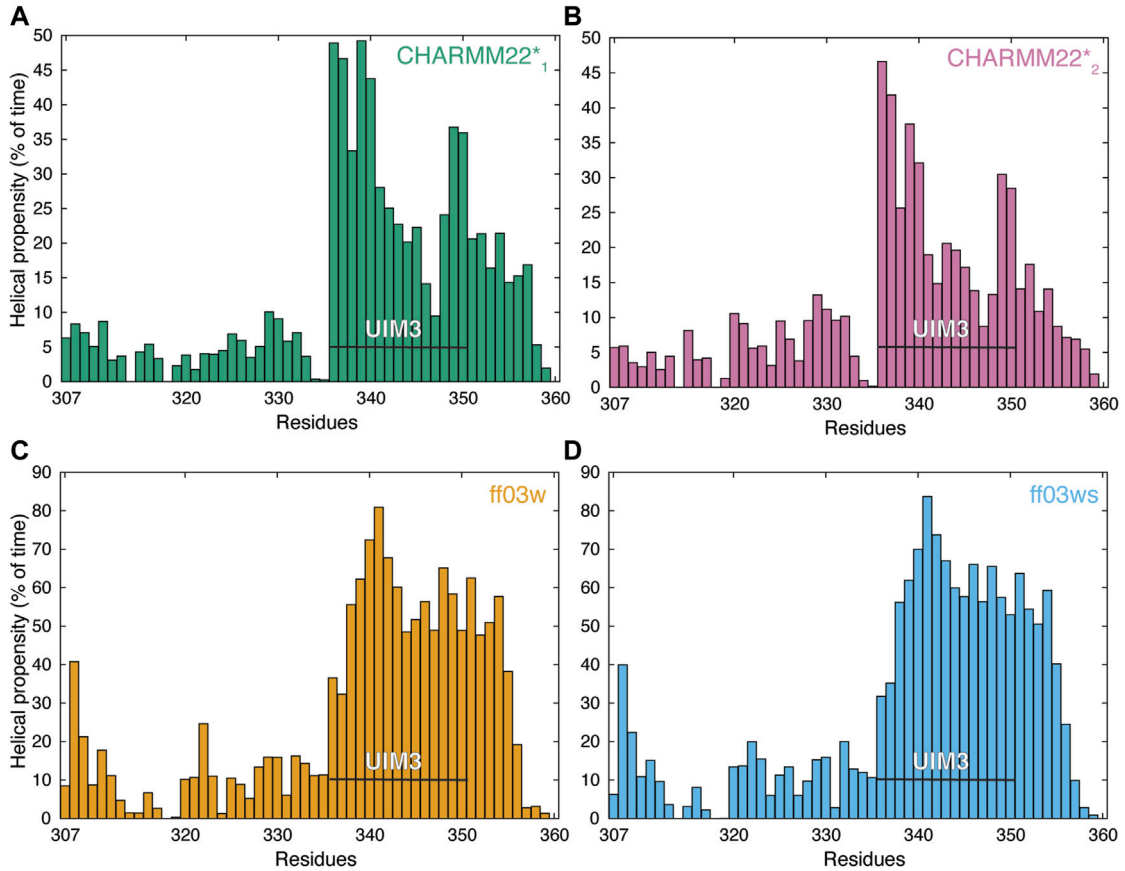


FIGURE 2 | AT-$3_{306\text{-}361}$ in the free state assumes both helical and non-helical conformations. The panels show the per-residue helical content of each replica at 304 K from the REMD simulations of the AT-$3_{306\text{-}361}$ for each combination of protein force fields and water models: **(A)** CHARMM22*-TIP3P (CHARMM22*$_1$, green), **(B)** CHARMM22*-TIPS3P (CHARMM22*$_2$, pink), **(C)** Amber ff03w-TIP4P/2005 (ff03w, orange), and **(D)** Amber ff03ws-TIP4P/2005 (ff03ws, blue). The residues of UIM3 (residues 336–350) are highlighted by the black bars. The REMD simulations with ff03w and ff03ws show high helical content for UIM3 while the CHARMM22*$_2$ simulation reports more disordered and heterogeneous conformations of UIM3.

which DNAse (10 μg/ml, Sigma-Aldrich, St. Louis, MO, United States) and PMSF (1 mM) were added and then disrupted by sonication. We purified the soluble protein fractions by affinity chromatography with Glutathione Sepharose four Fast Flow resin (GE Healthcare Bio-Sciences, Uppsala, Sweden) and subsequently in-site cleaved with 60 units of PreScission Protease (HRV 3C Protease Sino Biological inc., Beijing, P.R.China) per ml of resin. We then further purified the eluted samples by size-exclusion chromatography on a Superdex 75 10/300 GL column (GE Healthcare LifeSciences, Little Chalfont, England) in PBS buffer, pH 7.4, 150 mM NaCl.

Peptide Array

We purchased peptide arrays from Intavis and modified the procedures for blocking and probing the arrays from (Frank and Dubel, 2006). Briefly, the peptide array was re-hydrated through incubation in 100% ethanol and transferred in TBS (137 mM NaCl, 2.7 mM KCl, and 50 mM Tris, pH 7.0) for 5 min at room temperature. The blocking was performed by incubating the membrane 4°C overnight in TBS with 5% nonfat dry milk (MBS). Membranes were then incubated with 10 ml MBS with 2 mg/ml of 6His-TEV-Ubiquitin for 3 h at room temperature. The peptide array was then rinsed with a blocking buffer and then incubated with anti-6His antibody (Sigma Aldrich C6594) diluted 1:1,000 in the blocking buffer for 2 h at room temperature. The membrane was washed in Tween TBS 3 times and then incubated 1 h at room temperature with the secondary antibody (anti-mouse AP from Immunstar kit 170–5010).

NMR Spectroscopy of AT-$3_{182\text{-}291}$

NMR samples were prepared by dissolving the purified protein in 90% PBS buffer, pH 7.4, 150 mM NaCl and 10% D_2O with4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) added as internal calibration standard. Protein concentrations were from 0.5 to 1 mM in a volume of 400 μl. Assignment of backbone chemical shifts was performed on a 0.5 mM ^{13}C, ^{15}N AT-$3_{182\text{-}291}$ sample and ^{1}H, ^{15}N-HSQC spectrum and the following triple resonance spectra were recorded, HNCA, HN(CO)CA, HNCO, HN(CA)

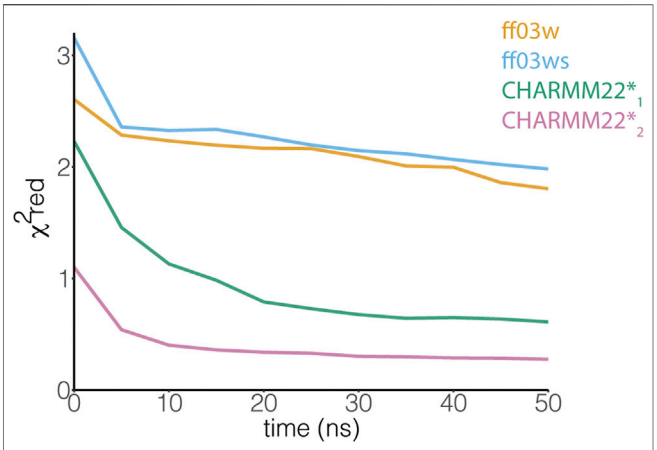


FIGURE 3 | The ensemble of AT-$3_{306-361}$ from the CHARMM22*$_2$ REMD simulation better resembles the experimental chemical shifts. The plot shows the comparison between experimental Cα chemical shifts and calculated Cα chemical shifts from the REMD simulations. Similar results have been achieved using the other backbone (N, HN, C, O, Hα) and Cβ chemical shifts. Among the protein force fields and water models tested in this study, the CHARMM22*$_2$ REMD simulation shows a better agreement with the experimental NMR measurements.

CO, CBCA(CO)NH, CBCANH, CC(CO)NH and H(CCO)NH (all pulse programs from Agilent BioPack) at 25 °C on a Varian Unity Inova 750 and 800 Mhz instruments. NMR data were processed by NMRPipe (Delaglio et al., 1995) and analyzed using CCPNMR (Skinner et al., 2016). The chemical shift assignment for AT-$3_{182-291}$ is deposited in the Biological Magnetic Resonance Bank (BMRB) with entry 50888.

Prediction of Secondary Structural Propensity From NMR Chemical Shifts

We downloaded NMR chemical shift data from the Biological Magnetic Resonance Bank (BMRB) for STAM1 [BMRB entry 17065 (Lim et al., 2011)], STAM2 [BMRB entry 18403 (Lange et al., 2012)], vps27 [BMRB entry 16114 (Sgourakis et al., 2010)], USP25 [including UIM1 and UIM2, BMRB entry 19111 (Shi et al., 2014)], RAP80 ([including UIM1 and UIM2, BMRB entry 19774 (Anamika et al., 2014)], USP28 [BMRB entries 18560 and 19077 (Wen et al., 2014)], and AT-3 [including UIM1, UIM2, and UIM3, BMRB entry 27380 (Sicorello et al., 2018)]. Furthermore, we included in the analyses the chemical shifts for AT-$3_{182-291}$ (including UIM1 and UIM2 of AT-3) from experiments performed in this study, along with previously published data for an AT-3 construct including the UIM3 residues 194–361 (Bai et al., 2013). We used the backbone chemical shifts from these NMR sets to predict the secondary structure propensity by δ2D (Camilloni et al., 2012).

Helical Wheel Projections

We calculated the helical wheel projections of UIMs of the selected proteins by the freely available NetWheels web-based application (Mól et al., 2018).

RESULTS

Conformational Ensemble of AT-$3_{306-361}$ in Solution

We used NMR data for AT-3 UIM3 from a previous publication (including residues 194–361) (Bai et al., 2013). MD simulations of such a long and disordered region are challenging, due to several conformational transitions to sample and a large number of degrees of freedom involved. We thus focused on a shorter construct for MD simulations, i.e., AT-$3_{306-361}$.

We employed two different methods to characterize the conformational ensemble of AT-$3_{306-361}$ in solution, i.e., REMD and WT-metaD. These methods provide the possibility to enhance the sampling of the conformational space in MD simulations while keeping a description of both the protein and the solvent at the atomic level. We also evaluated the influence of different force-field descriptions for both the protein and the solvent: Amber ff03w-TIP4P/2005, Amber ff03ws-TIP4P/2005, CHARMM22*-TIP3P, CHARMM22*-TIPS3P (indicated as ff03w, ff03ws, CHARMM22*$_1$, and CHARMM22*$_2$, respectively) to assess the reproducibility of the result and identify force-field dependent properties. These approaches enabled us 1) to address if AT-$3_{306-361}$ is stable or not in a helical conformation in solution, 2) to estimate the population of the helical conformations and compare them to the available experimental information on a variant of AT-3 (residues 194–361) characterized by NMR and on other known UIMs that have been similarly studied by solution NMR (see *Materials and Methods*) or recorded by us in this work (AT-$3_{182-291}$), 3) to identify conformations that resemble ubiquitin-bound states in the ensemble of the free AT-$3_{306-361}$ region through the comparison of our ensembles to the experimentally known ubiquitin-bound UIM structures of other proteins.

Low Structural Propensity and Heterogeneous Helical Formation in the Free State of AT-$3_{306-361}$ Domain in Solution

UIMs are thought to assume an α-helical structure also in the absence of ubiquitin binding (Hofmann and Falquet, 2001; Scott et al., 2015). Nevertheless, many investigations on UIMs focus on characterizing the binding with ubiquitin, making it unclear if UIMs present transient propensity to disordered conformations in their free state, a typical trait of SLiMs (Davey et al., 2012; Van Roey et al., 2012, 2014). In AT-$3_{306-361}$, UIM3 spans residues E336-T350 [(Donaldson et al., 2003), **Figure 1A**]. To identify inherent structural properties, we used four sequence-based methods to predict disorder or secondary structure propensity (Nielsen and Mulder, 2019). Overall, the predictors showed a disordered state for the AT-$3_{306-361}$ region with propensity to order and helicity around UIM3 (**Figures 1B–D**).

We subsequently modeled this region as an α-helix in the starting structure for the REMD simulations. In the REMD simulations, the UIM3 region assumed both helical and non-helical conformations (**Figure 2**). The REMD simulations with ff03w and ff03ws showed higher helical content for UIM3 (~60%

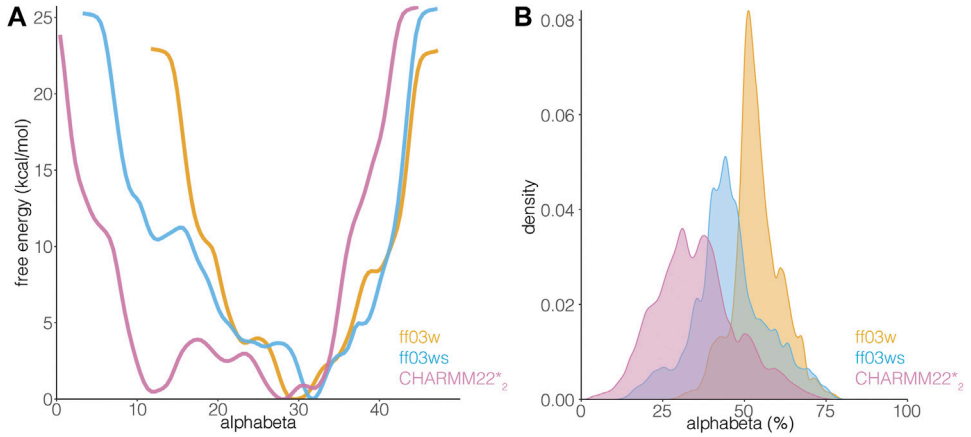


FIGURE 4 | AT-$3_{306\text{-}361}$ is characterized by a disordered ensemble with a low structural propensity and heterogeneous helical states. **(A)** The plot shows the one-dimensional free energy profile associated with the collective variable alphabeta for the ff03w (orange), ff03ws (blue), and CHARMM22*$_2$ (pink) metadynamics simulations. **(B)** The plot shows the distribution of alphabeta values expressed as a percentage, i.e., alphabeta values divided by the total number of torsional angles considered. Ff03w and ff03ws show overstabilization of helical conformations while CHARMM22*$_2$ better describes the disordered ensemble of AT-$3_{306\text{-}361}$ in the free state.

and 56%, respectively) than with CHARMM22*$_1$, and CHARMM22*$_2$ (~31% and 25%, respectively) (**Figure 2**). In the case of CHARMM22*$_2$ simulation, we observed a more disordered ensemble for UIM3, with helical content < 20% in the region Q341-L348.

An NMR backbone chemical shift assignment for AT-$3_{306\text{-}361}$is available (Bai et al., 2013). We used this set of experimental data to evaluate the REMD structural ensembles. In particular, we calculated backbone chemical shifts as a function of the simulation time and compared these to the experimental values. The calculated chemical shifts from our simulations converged after only 5 ns of REMD simulations (**Figure 3**). They are in agreement with the experimental values with low χ^2_{red} values of CHARMM22* simulations, but not in the ff03w/ff03ws simulations (**Figure 3**). In ff03w/ff03ws, the simulations converged to structures that are unlikely to resemble the ones observed by solution NMR, probably due to the high helicity sampled by these trajectories.

The differences in the sampling of helical structures in the REMD simulations with different force fields could be ascribed to either force field differences or limitation of the conformational sampling. Since we started from an α-helical conformation the simulation time might not have been sufficient, even with the temperature replica-exchange, to allow the protein to exhaustively explore the conformational space in the different force-field simulations. Thus, to be able to discriminate between these two scenarios, we applied another method for enhanced sampling, based on metadynamics. In particular, we carried out simulations with WT-metaD, which should allow a more extensive exploration of the conformational space by using the Cα radius of gyration and *alphabeta* as collective variables (CVs) to bias the systems. Alphabeta is a collective variable in which we measured similarity for all ψ dihedral angles of the peptide to the ψ dihedral angles of an ideal α-helix (**Figure 4**). It is a suitable CV to enhance the sampling of disordered regions which might have local propensity for helical structures (Granata et al., 2015). The alphabeta estimated by the three different force fields were different with the CHARMM22*$_2$ simulation providing more disordered conformations (i.e., alphabeta between 8 and 15 residues in **Figure 4**). As also observed in the REMD simulations, the ff03w ensemble was characterized by a higher helical content, suggesting that the difference observed is not necessarily related to limitations in the sampling or initial conformation, but to differences in force field parameters. In this context, overstabilization of helical conformations with ff03w has been observed also in other studies (Huang and Mackerell, 2014). The modification of the ff03ws force field with more balanced interactions between the protein and the solvent (Best et al., 2014) partially mitigates this effect, providing an ensemble of structures with a lower helical propensity, including also disordered states corresponding to the ones observed for CHARMM22*$_2$ (**Figure 4**).

The transition between more ordered and disordered states is favored in the description provided by CHARMM22*$_2$ (with difference in free energy of 1.5 kcal/mol). In the ff03ws simulation, the two states were separated by a barrier of more than 8 kcal/mol. The intrinsic preference for helical conformations of ff03w/ff03ws is likely to make the sampling of disordered states more challenging even with an enhanced sampling approach. The high energy barriers observed are thus likely to be due to limitations of the sampling. Longer simulations or other enhanced sampling approaches could help to obtain free energy profiles with a larger number of order to disorder transitions for this peptide and ff0ws (Bussi and Laio, 2020).

In summary, AT-$3_{306\text{-}361}$ is characterized by a disordered ensemble, which is better described by CHARMM22*$_2$ among the force fields tested in this study. The UIM3 region of AT-$3_{306\text{-}}$

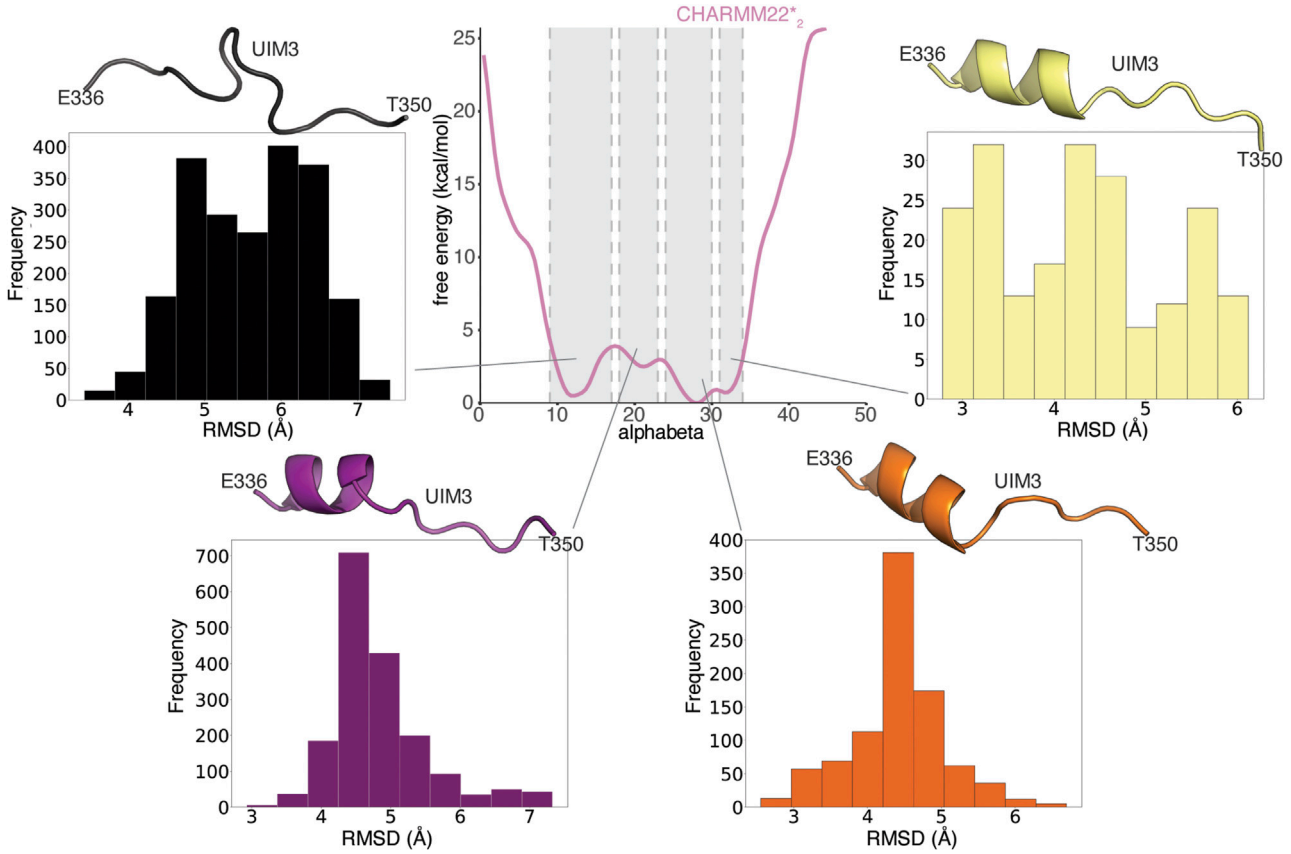


FIGURE 5 | UIM3 of AT-$3_{306-361}$ interconverts between disordered conformations and partially helical states in the free ensemble. The central panel shows the mono-dimensional free energy profile associated with alphabeta for CHARMM22*$_2$ (pink) metadynamics simulation. The four side panels show the RMSD calculations of UIM3 structures from the CHARMM22*$_2$ metadynamics trajectory, using as a reference the starting structure of AT-$3_{306-361}$ for the REMD simulations, in which UIM3 is modeled as an α-helix. We used the Cβ atom of A343 and the backbone (Cα, C, O, N) atoms of the residues E336-T350 of UIM3 for rigid body superposition and the RMSD calculation. We calculated the RMSD of four ensembles of structures of UIM3 with alphabeta values in the ranges of: i) 9–17 (black), ii) 18–23 (purple), iii) 24–30 (red), and iv) 31–34 (yellow), respectively. The cartoon representations show the structures of UIM3 with the lowest RMSD in each of the four subsets. We identify partially helical states of UIM3 associated with alphabeta values in the range of 24-30 and 31–34.

$_{361}$ can interconvert between more disordered and partially helical states.

Bound-Like Conformations in the Unbound AT-$3_{306-361}$ Ensemble in Solution

Both ordered and disordered proteins often sample bound-like states that could be important for their binding, which may sometimes occur via a mechanism known as conformational selection (Davey, 2019). We, therefore, asked if this was also the case for UIM3 of AT-$3_{306-361}$. To this end, we compared conformations from the CHARMM22*$_2$ WT-metaD simulation with the starting structure of AT-$3_{306-361}$ for the REMD simulations, in which UIM3 is modeled as a well-folded *a*-helix (**Figure 5**). We identify partially folded states of UIM3 (around 3 Å of RMSD), characterized by helical conformations in the N-terminal region of the motif (residues 336–344) (**Figure 5**) and alphabeta values in the range of 24-30 and 31–34 residues (**Supplementary Figure S1**). We observed that the region with the highest propensity to fold to helix corresponds to the UIM3 region (residues 336–344). This accounts for approximately 20% of the structures from the entire WT-MetaD. We also observed a minor helical propensity in other regions of the peptide, especially around residues 320–334 (less than 10% of the structures from the metaD).

We performed the same RMSD analysis on the replicas at 304 K from the REMD simulations (**Supplementary Figure S2**). In contrast with the results from WT-metaD, the REMD simulations tend to show a group of fully helical conformations of UIM3 (which are a minority of the frames in the CHARMM22*$_2$ simulations, i.e. ~ 3% of the frames) (**Supplementary Figure S2**). These analyses suggest that the REMD simulations provide a limited sampling and they are still biased by the initial helical conformation of UIM3. We thus relied on the WT-metaD results for the following analyses.

To identify the presence of bound-like states, we then compared the partially helical conformations of UIM3 of AT-$3_{306-361}$ from the CHARMM22*$_2$ WT-metaD simulation with the experimental structures of ubiquitin in complex with UIMs from other proteins (**Figure 6**).

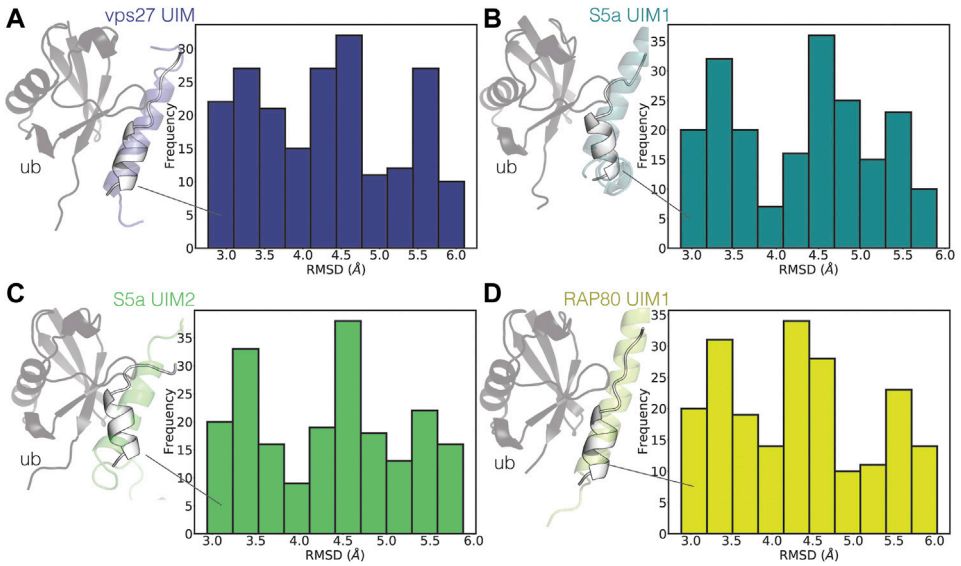


FIGURE 6 | UIM3 of AT-$3_{306\text{-}361}$ samples states partially resembling ubiquitin-bound conformations. The plots show the RMSD calculations of UIM3 of AT-$3_{306\text{-}361}$ from the CHARMM22*$_2$ metadynamics trajectory, using as a reference the experimental structure of the UIMs of other proteins in complex with ubiquitin **(A)** VPS27 UIM (PDB entry 1Q0W, blue), **(B)** S5a UIM1 (PDB entry 1YX5, teal), **(C)** S5a UIM2 (PDB entry 1YX6, green), **(D)** RAP80 UIM1 (PDB entry 3A1Q, yellow). We calculated RMSD for the ensembles of structures of UIM3 with alphabeta values in the range of 31-34. The cartoon representations show the structures of the experimental complexes, with the ubiquitin monomers shown as gray cartoons. The white cartoon representation shows the conformations of UIM3 (residues 336–350) from the CHARMM22*$_2$ metadynamics simulation with the lowest RMSD to each experimental structure.

UIMs are generally in folded helical conformations when bound to ubiquitin (Fisher et al., 2003; Swanson et al., 2003). We identified states of UIM3 partially resembling the bound conformations of other UIMs, characterized by RMSD around 3 Å with respect to the experimental complexes (**Figure 6**).

Disordered UIMs With Low Helical Propensity: A More General Class of UIMs

To discriminate if the low occurrence of a helical conformation in solution is a distinctive trait of UIM3 or a more common property of other UIMs, we searched the NMR database BMRB for chemical shift data on other UIMs in solution. We identified nine sets of released chemical shifts for AT-3 (including UIM1, UIM2, and UIM3), STAM1, STAM2, USP28, USP25, and RAP80 (holding two UIMs each) (**Supplementary Figure S3**). We also used a set of chemical shifts of VPS27 UIM1 in fusion with ubiquitin (**Supplementary Figure S3**). In addition, we recorded NMR experiments to collect backbone and side-chain chemical shifts for UIM1 and UIM2 of AT-3 in solution, using AT-$3_{182\text{-}291}$. From the chemical shifts, we predicted the secondary structural propensity by δ2D (**Figure 7A** and **Supplementary Figure S3**). We observed UIMs with high helical content, such as UIM1 and UIM2 of AT-3, UIM1 and UIM2 of RAP80, and UIM1 of USP25 (average δ2D helix population higher than 0.3), and low helical content, as in the case of UIM3 of AT-3 and UIM2 of USP25 (average δ2D helix population lower than 0.1). USP28 has also a lower helical content compared to other UIMs suggesting a heterogeneous ensemble of conformations. We observe a lower helical content in the case of VPS27 UIM in fusion with ubiquitin, possibly suggesting that in the bound state some UIMs could retain disorder. Our NMR data of AT-$3_{182\text{-}291}$ are in agreement with previously published sets of chemical shifts of AT-3, showing high helical content for UIM1 and UIM2 (average δ2D helix population above 0.3 for UIM1 and 0.4 for UIM2 in all datasets) (**Supplementary Figure S3**). Furthermore, our analysis on the two sets of chemical shifts of UIM3 shows low helical content for both of them (average δ2D helix population under 0.1 for each set) (**Figure 7A** and **Supplementary Figure S3**), confirming the presence of disordered conformations.

Our results overall indicate that UIMs in the unbound state can span not only fully-formed helical conformations but also rather disordered counterparts. Moreover, a peptide SPOT arrays in which we studied the interaction of some representative UIMs with recombinant yeast ubiquitin shows that both disordered UIMs (AT-3 UIM3 and USP25 UIM1 and 2) and helical UIMs (STAM1, STAM2 and, AT-3 UIM1) interact with ubiquitin (**Supplementary Figure S4**). Thus, as all other UIMs tested are folded in the unbound state, these data suggest that a disordered UIM is not a barrier to bind ubiquitin.

To address if different classes of UIMs can be derived based on sequence and disorder, we compared the UIM3 sequence to other known UIMs and with the consensus sequence deposited in the Pfam database (entry PF02809) (**Figure 7B**). The 336–350 region of the C-terminus of AT-3 presents the typical signature of a UIM with conserved residues, such as L340, A343, S347, acidic residues in the N-terminal part of the motif (E336-D338), and the pattern of hydrophobic residues (**Figure 7B**). Moreover, in comparison to other UIMs, we should notice that suboptimal residues for helical formation are observed in the UIM3 sequence in

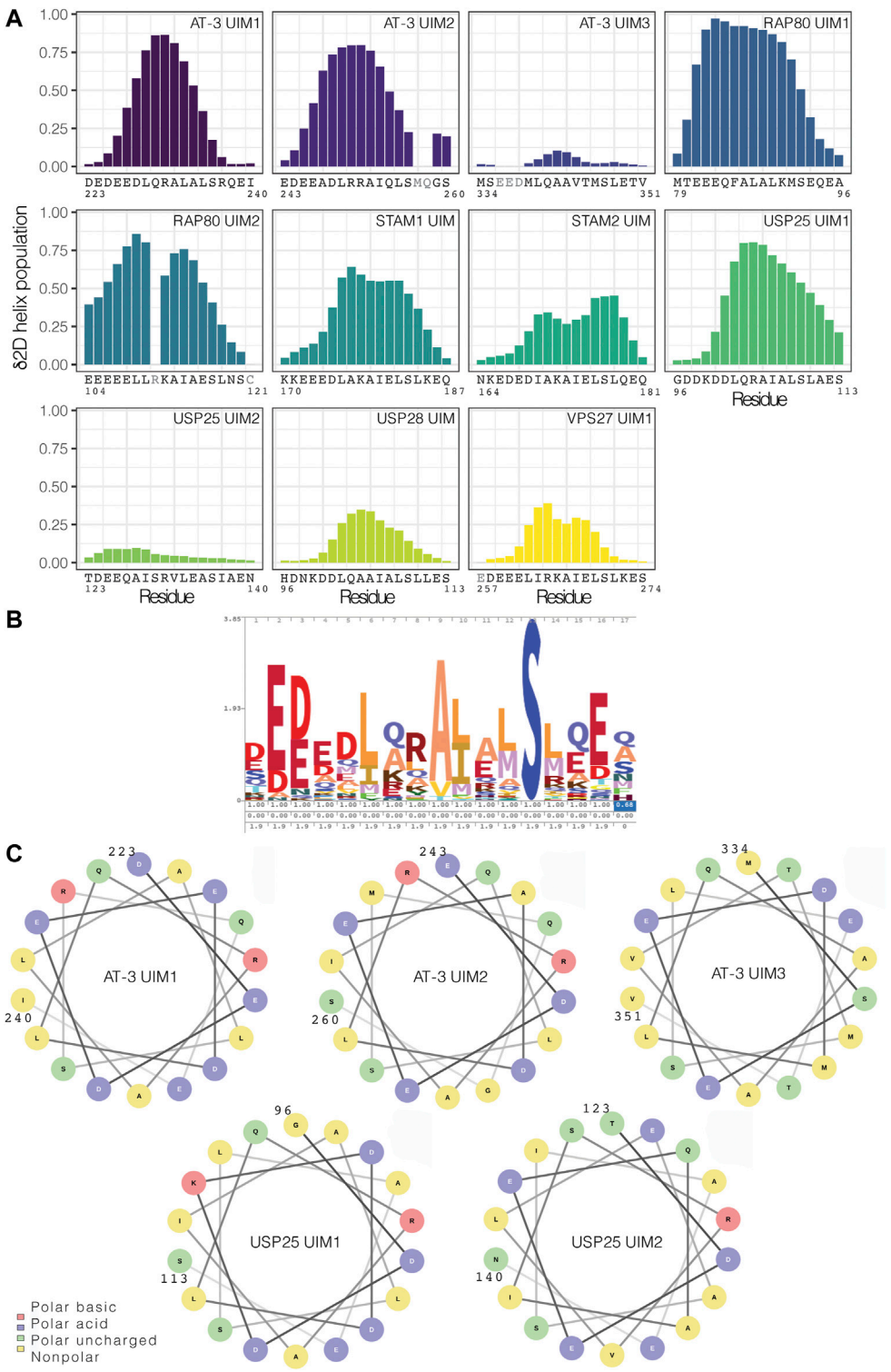


FIGURE 7 | UIMs in the free state can vary from highly helical conformations to disordered counterparts. **(A)** Helical content for the UIMs predicted from chemical shifts by δ2D. We used nine sets of released chemical shifts of UIMs in the free state in solution from the BMRB database, including AT-3 UIM3, STAM1 UIM, STAM2 UIM, USP28 UIM, USP25 UIM1, and UIM2, and RAP80 UIM1 and UIM2. We also used a set of released chemical shifts of VPS27 UIM1 in fusion with the ubiquitin. In addition, we used the NMR chemical shifts that we recorded for AT-$3_{182-291}$ UIM1 and UIM2. We highlighted in gray the residues for which there are not enough chemical shifts to run the prediction with δ2D. UIMs have a wide range of predicted helical content. **(B)** Consensus sequence for UIMs in the PFAM database. **(C)** Helical wheel representation of AT-3 UIM1, UIM2, and UIM3, and of USP25 UIM1 and UIM2.

comparison to other UIMs. For example, V344 and T345 are both at low helix propensity (Nick Pace and Martin Scholtz, 1998) and are localized in the region of UIM3 where the helix tend to break in some of the simulation frames (see above). Furthermore, USP25 has a valine replacing the invariant alanine of the motif and an insertion of an arginine in the N-terminal region of the motif which might alter the helical pattern.

To further identify if this is a common signature to other disordered UIMs, we carried out a helical wheel analysis of AT-3 UIM1, UIM2, UIM3, and of UIMs previously investigated (USP25 UIM1 and UIM2 **Figure 7C**). The analysis shows that when UIM3 assumes a helical conformation T345 is located on the face of the helix with one of the acidic residues (i.e., E337) that is conserved in UIMs. Moreover, T350 is located at the same face of the helix as A343 and S347; two residues that are strictly conserved in all UIMs since they are involved in the interaction with ubiquitin (Fisher et al., 2003). For the disordered UIM2 of USP25, a similar valine and isoleucine, two beta-branched amino acids, break the helicity. This means that disordered UIMs may carry similar sequence properties that allow for their identification. Our analysis and simulations overall suggest that the location of suboptimal residues, especially threonine and valine, coul be related to the low propensity to populate stable helical conformations in solution. A search based on regular expression with the motif x-[ED]-[ED]-[ED]-x-[AILVFWMP]-x-x(1,2)-[AVP]-[VPL]-[EDVNTCGPH]-x-S-x-x-[EDTVNCGPH]-x against the sequences associated with the Pfam entry PF02809 highlights other 1,614 hits in 626 sequences of UIMs with likelihood to be (partially) disordered in the unbound state (against 172 hits found in a randomized dataset from Uniprot). Among the disordered UIM candidates we find UBP37 from different species (residues 704–720), which feature patterns similar to USP25. The motif search suggests that disordered UIM could be a common class of SLiMs (see **Supplementary Text S5**).

DISCUSSION

We focused on the structural characterization of the conformational ensemble of a functional motif that has been classically defined for its helical conformation and originally associated with the binding of ubiquitin; the Ubiquitin Interacting Motif (UIM). We showed that the motifs could be more degenerate and account for both helical and more disordered members, a diversity that has functional implications. With an approach integrating simulations and experimental biophysical data, we showed that a C-terminal UIM of AT-3 is embedded in an intrinsically disordered region, bearing a predominantly disordered UIM of which a small fraction of the ensemble has helical propensity in the N-terminal region. An unbound ensemble as the one depicted by WT-metaD might suggest that a combination of conformational selection (i.e., pre-formed regions of the UIM in helical conformation) and folding upon binding could be in place for UIM3. The occurrence of one or the other mechanism might also depend on the nature of the client protein and help to confer UIM3 promiscuity toward different partners of interaction, an effect that can be further tuned by post-translational modifications. These mechanisms will require future investigations in which kinetics can be accounted for.

We also discovered that the disordered nature of UIM3, and low helical propensity in the free state, is not an isolated example, as shown by the analysis of the NMR data of USP25 UIM2, USP28 UIM, and VPS27 UIM1.

In our work, all the UIMs tested for binding with ubiquitin are either folded or disordered in the unbound state. These data suggest that a disordered UIM is not a barrier to bind ubiquitin. NMR measurements on UIM3 still suggest that the binding could be of lower affinity than what observed for helical UIMs (Bai et al., 2013), supporting a more pliable partner toward a different range of client proteins, at the cost of larger entropy loss in binding ubiquitin.

UIMs not only bind ubiquitin but can also be interfaces to recruit other proteins, such as the case of UIM and parkin. Proteins including disordered UIMs can have additional diversity in their protein-protein interactions and cellular functions. For example, it has been suggested by mass-spectrometry and co-immunoprecipitation assays that AT-3 isoforms differ in their interaction with other proteins (Weishäupl et al., 2019). Post-translational modifications are likely to add an extra level of regulation, and they could modulate the helical propensity of disordered UIMs and their preferences for binding partners, as seen for other IDRs (Mylona et al., 2016; Hendus-Altenburger et al., 2017; Csizmok and Forman-Kay, 2018; Marceau et al., 2019). For example, UIM3 is sumoylated at K356 and this enhances affinity for the binding to ATPase p97 to transfer proteins for proteasomal degradation (Almeida et al., 2015). Further experimental and computational studies of these disordered UIMs here identified and their post-translational regulation or the study of UIM from other proteins could contribute to clarify the structural and sequence features of disordered and folded UIMs, along with their connection with certain binding partners and biological functions.

Our analysis and simulations overall suggest that the location of suboptimal residues for helix formation, especially beta-branched residues as threonine, valine and isoleucine could play a role in gearing the low propensity of UIMs to populate stable helical conformations in solution and provide a gateway to multispecificity.

UIMs are not the only example of such structural duality. Some regions of proteins that were traditionally defined as helical elements, due to their conformation in the bound states, have been reclassified as disordered SLiMs, as in the case of the Bcl-2 Homology 3 motifs (Hinds et al., 2007; Aouacheria et al., 2015). The presence of this emerging higher structural variability into different classes of SLiMs is a shift in our view of functional protein regions that can account for both helical and more disordered counterparts. A better understanding of the structural diversity within each class of functional motifs could open new directions to understand biomolecular interactions and their specificity or flexibility toward multiple partners of interaction. SLiMs with disorder propensity and a more versatile interface could enhance the pool of functions of a certain protein, for example increasing the number of potential binding partners, allowing the protein to act at the

cross-road among different biological processes, or allow for a fine regulation by post-translational modifications.

DATA AVAILABILITY STATEMENT

All the scripts, raw data, and outputs associated with this publication are available at the repository on GitHub https://github.com/ELELAB/disoUIM and on OSF https://osf.io/zfy9s/.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: EP, GI, and ML. Performed the experiments: ML, EP, GI, EM, and BA. Analyzed the data: EP, GI, ML, EM, and KT. Discussion of the data: ML, EP, GI, BK, KT, KL-L, and GS. Contributed reagents/materials/analysis tools: ML, EP, BK, GS, and KT. Wrote the paper: ML, EP. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

EP group was supported by Carlsberg fondet Distinguished Fellowship (CF18-0314), Danmarks Grundforskningsfond (DNRF125) and NovoNordisk Fonden Bioscience and Basic Biomedicine NNF20OC0065262. BA was supported by COST-STSM-BM1405-34558. GI was supported by a Marie Curie IEF Fellowship. BK was supported by Novo Nordisk Foundation Challenge grant REPIN – rethinking protein interactions. GS was supported by Canadian Institutes of Health Research (PJT – 166019). The calculations described in this paper were performed using the ISCRA-CINECA grant diso-UIMs HP10C4LACQ.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.676235/full#supplementary-material

REFERENCES

Abascal, J. L. F., and Vega, C. (2005). A General Purpose Model for the Condensed Phases of Water: TIP4P/2005. *J. Chem. Phys.* 123, 234505. doi:10.1063/1.2121687

Abrams, C., and Bussi, G. (2013). Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy*. 16, 163–199. doi:10.3390/e16010163

Aguirre, J. D., Dunkerley, K. M., Lam, R., Rusal, M., and Shaw, G. S. (2018). Impact of Altered Phosphorylation on Loss of Function of Juvenile Parkinsonism-Associated Genetic Variants of the E3 Ligase Parkin. *J. Biol. Chem.* 293, 6337–6348. doi:10.1074/jbc.RA117.000605

Almeida, B., Abreu, I. A., Matos, C. A., Fraga, J. S., Fernandes, S., Macedo, M. G., et al. (2015). SUMOylation of the Brain-Predominant Ataxin-3 Isoform Modulates its Interaction with P97. *Biochim. Biophys. Acta.* 1852, 1950–1959. doi:10.1016/j.bbadis.2015.06.010

Anamika, S., Markin, C. J., Rout, M. K., and Spyracopoulos, L. (2014). Molecular Basis for Impaired DNA Damage Response Function Associated with the RAP80 ΔE81 Defect. *J. Biol. Chem.* 289, 12852–12862. doi:10.1074/jbc.M113.538280

Aouacheria, A., Combet, C., Tompa, P., and Hardwick, J. M. (2015). Redefining the BH3 Death Domain as a 'Short Linear Motif'. *Trends Biochem. Sci.* 40, 736–748. doi:10.1016/j.tibs.2015.09.007

Bai, J. J., Safadi, S. S., Mercier, P., Barber, K. R., and Shaw, G. S. (2013). Ataxin-3 Is a Multivalent Ligand for the Parkin Ubl Domain. *Biochemistry*. 52, 7369–7376. doi:10.1021/bi400780v

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* 100, 1–4. doi:10.1103/PhysRevLett.100.020603

Berlow, R. B., Dyson, H. J., and Wright, P. E. (2018). Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation. *J. Mol. Biol.* 430, 2309–2320. doi:10.1016/j.jmb.2018.04.003

Best, R. B. (2017). Computational and Theoretical Advances in Studies of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* 42, 147–154. doi:10.1016/j.sbi.2017.01.006

Best, R. B., and Mittal, J. (2010). Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. *J. Phys. Chem. B.* 114, 14916–14923. doi:10.1021/jp108618d

Best, R. B., Zheng, W., and Mittal, J. (2014). Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-specific Protein Association. *J. Chem. Theor. Comput.* 10, 5113–5124. doi:10.1021/ct500569b

Bettencourt, C., Santos, C., Montiel, R., Costa, M. d. C., Cruz-Morales, P., Santos, L. R., et al. (2010). Increased Transcript Diversity: Novel Splicing Variants of Machado-Joseph Disease Gene (ATXN3). *Neurogenetics* 11, 193–202. doi:10.1007/s10048-009-0216-y

Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., et al. (2009). PLUMED: A Portable Plugin for Free-Energy Calculations with Molecular Dynamics. *Computer Phys. Commun.* 180, 1961–1972. doi:10.1016/j.cpc.2009.05.011

Bonomi, M., Heller, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of Protein Structural Ensemble Determination. *Curr. Opin. Struct. Biol.* 42, 106–116. doi:10.1016/j.sbi.2016.12.004

Brunger, A. T. (2007). Version 1.2 of the Crystallography and Nmr System. *Nat. Protoc.* 2, 2728–2733. doi:10.1038/nprot.2007.406

Buchberger, A. (2002). From UBA to UBX: New Words in the Ubiquitin Vocabulary. *Trends Cel Biol* 12, 216–221. doi:10.1016/S0962-8924(02)02269-9

Bugge, K., Brakti, I., Fernandes, C. B., Dreier, J. E., Lundsgaard, J. E., Olsen, J. G., et al. (2020). Interactions by Disorder - A Matter of Context. *Front. Mol. Biosci.* 7, 110. doi:10.3389/fmolb.2020.00110

Burnett, B., Li, F., and Pittman, R. N. (2003). The Polyglutamine Neurodegenerative Protein Ataxin-3 Binds Polyubiquitylated Proteins and Has Ubiquitin Protease Activity. *Hum. Mol. Genet.* 12, 3195–3205. doi:10.1093/hmg/ddg344

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* 126, 014101. doi:10.1063/1.2408420

Bussi, G., and Laio, A. (2020). Using Metadynamics to Explore Complex Free-Energy Landscapes. *Nat. Rev. Phys.* 2, 200–212. doi:10.1038/s42254-020-0153-0

Camilloni, C., De Simone, A., Vranken, W. F., and Vendruscolo, M. (2012). Determination of Secondary Structure Populations in Disordered States of Proteins Using Nuclear Magnetic Resonance Chemical Shifts. *Biochemistry*. 51, 2224–2231. doi:10.1021/bi3001825

Carvalho, A. L., Silva, A., and Macedo-Ribeiro, S. (2018). Polyglutamine-independent Features in Ataxin-3 Aggregation and Pathogenesis of Machado-Joseph Disease. *Adv. Exp. Med. Biol.* 1049, 275–288. doi:10.1007/978-3-319-71779-1_14

Chen, J., and Kriwacki, R. W. (2018). Intrinsically Disordered Proteins: Structure, Function and Therapeutics. *J. Mol. Biol.* 430, 2275–2277. doi:10.1016/j.jmb.2018.06.012

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. (2013). From Protein Sequence to Dynamics and Disorder with DynaMine. *Nat. Commun.* 4, 2741. doi:10.1038/ncomms3741

Csizmok, V., and Forman-Kay, J. D. (2018). Complex Regulatory Mechanisms Mediated by the Interplay of Multiple post-translational Modifications. *Curr. Opin. Struct. Biol.* 48, 58–67. doi:10.1016/j.sbi.2017.10.013

Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald: AnN·Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 98, 10089–10092. doi:10.1063/1.464397

Davey, N. E. (2019). The Functional Importance of Structure in Unstructured Protein Regions. *Curr. Opin. Struct. Biol.* 56, 155–163. doi:10.1016/j.sbi.2019.03.009

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., et al. (2012). Attributes of Short Linear Motifs. *Mol. Biosyst.* 8, 268–281. doi:10.1039/C1MB05231D

Delaglio, F., Grzesiek, S., Vuister, G., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. *J. Biomol. NMR.* 6, 277–293. doi:10.1007/BF00197809

Do, T. N., Choy, W.-Y., and Karttunen, M. (2014). Accelerating the Conformational Sampling of Intrinsically Disordered Proteins. *J. Chem. Theor. Comput.* 10, 5081–5094. doi:10.1021/ct5004803

Donaldson, K. M., Li, W., Ching, K. A., Batalov, S., Tsai, C.-C., and Joazeiro, C. A. P. (2003). Ubiquitin-mediated Sequestration of normal Cellular Proteins into Polyglutamine Aggregates. *Proc. Natl. Acad. Sci.* 100, 8892–8897. doi:10.1073/pnas.1530212100

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. y., et al. (2007). Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Protein Sci.* 50. doi:10.1002/0471140864.ps0209s50

Fisher, R. D., Wang, B., Alam, S. L., Higginson, D. S., Robinson, H., Sundquist, W. I., et al. (2003). Structure and Ubiquitin Binding of the Ubiquitin-Interacting Motif. *J. Biol. Chem.* 278, 28976–28984. doi:10.1074/jbc.M302596200

Flock, T., Weatheritt, R. J., Latysheva, N. S., and Babu, M. M. (2014). Controlling Entropy to Tune the Functions of Intrinsically Disordered Regions. *Curr. Opin. Struct. Biol.* 26, 62–72. doi:10.1016/j.sbi.2014.05.007

Frank, R., and Dubel, S. (2006). Analysis of Protein Interactions with Immobilized Peptide Arrays Synthesized on Membrane Supports. *Cold Spring Harbor Protoc.* 2006, prot4566. doi:10.1101/pdb.prot4566

Fung, H. Y. J., Birol, M., and Rhoades, E. (2018). IDPs in Macromolecular Complexes: the Roles of Multivalent Interactions in Diverse Assemblies. *Curr. Opin. Struct. Biol.* 49, 36–43. doi:10.1016/j.sbi.2017.12.007

Goto, J., Watanabe, M., Ichikawa, Y., Yee, S.-B., Ihara, N., Endo, K., et al. (1997). Machado-Joseph Disease Gene Products Carrying Different Carboxyl Termini. *Neurosci. Res.* 28, 373–377. doi:10.1016/S0168-0102(97)00056-4

Granata, D., Baftizadeh, F., Habchi, J., Galvagnion, C., De Simone, A., Camilloni, C., et al. (2015). The Inverted Free Energy Landscape of an Intrinsically Disordered Peptide by Simulations and Experiments. *Sci. Rep.* 5, 1–15. doi:10.1038/srep15449

Guarnera, E., and Berezovsky, I. N. (2019). On the Perturbation Nature of Allostery: Sites, Mutations, and Signal Modulation. *Curr. Opin. Struct. Biol.* 56, 18–27. doi:10.1016/j.sbi.2018.10.008

Hanson, J., Paliwal, K. K., Litfin, T., and Zhou, Y. (2019). SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. *Genomics, Proteomics & Bioinformatics*. 17, 645–656. doi:10.1016/j.gpb.2019.01.004

Harris, G. M., Dodelzon, K., Gong, L., Gonzalez-Alegre, P., and Paulson, H. L. (2010). Splice Isoforms of the Polyglutamine Disease Protein Ataxin-3 Exhibit Similar Enzymatic yet Different Aggregation Properties. *PLoS One*. 5, e13695. doi:10.1371/journal.pone.0013695

Hendus-Altenburger, R., Lambrughi, M., Terkelsen, T., Pedersen, S. F., Papaleo, E., Lindorff-Larsen, K., et al. (2017). A Phosphorylation-Motif for Tuneable helix Stabilisation in Intrinsically Disordered Proteins - Lessons from the Sodium Proton Exchanger 1 (NHE1). *Cell Signal.* 37, 40–51. doi:10.1016/j.cellsig.2017.05.015

Hinds, M. G., Smits, C., Fredericks-Short, R., Risk, J. M., Bailey, M., Huang, D. C. S., et al. (2007). Bim, Bad and Bmf: Intrinsically Unstructured BH3-Only Proteins that Undergo a Localized Conformational Change upon Binding to Prosurvival Bcl-2 Targets. *Cell Death Differ*. 14, 128–136. doi:10.1038/sj.cdd.4401934

Hofmann, K., and Falquet, L. (2001). A Ubiquitin-Interacting Motif Conserved in Components of the Proteasomal and Lysosomal Protein Degradation Systems. *Trends Biochem. Sci.* 26, 347–350. doi:10.1016/S0968-0004(01)01835-7

Huang, J., and Mackerell, A. D. (2014). Induction of Peptide Bond Dipoles Drives Cooperative helix Formation in the (AAQAA)3 Peptide. *Biophysical J.* 107, 991–997. doi:10.1016/j.bpj.2014.06.038

Huang, J., and MacKerell, A. D. (2018). Force Field Development and Simulations of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* 48, 40–48. doi:10.1016/j.sbi.2017.10.008

Ichikawa, Y., Goto, J., Hattori, M., Toyoda, A., Ishii, K., Jeong, S.-Y., et al. (2001). The Genomic Structure and Expression of MJD, the Machado-Joseph Disease Gene. *J. Hum. Genet.* 46, 413–422. doi:10.1007/s100380170060

Invernizzi, G., Lambrughi, M., Regonesi, M. E., Tortora, P., and Papaleo, E. (2013). The Conformational Ensemble of the Disordered and Aggregation-Protective 182-291 Region of Ataxin-3. *Biochim. Biophys. Acta*. 1830, 5236–5247. doi:10.1016/j.bbagen.2013.07.007

Invernizzi, G., Papaleo, E., Sabate, R., and Ventura, S. (2012). Protein Aggregation: Mechanisms and Functional Consequences. *Int. J. Biochem. Cel Biol.* 44, 1541–1554. doi:10.1016/j.biocel.2012.05.023

Irbäck, A., and Mohanty, S. (2006). PROFASI: A Monte Carlo Simulation Package for Protein Folding and Aggregation. *J. Comput. Chem.* 27, 1548–1555. doi:10.1002/jcc.20452

Johnson, S. L., Ranxhi, B., Libohova, K., Tsou, W.-L., and Todi, S. V. (2020). Ubiquitin-interacting Motifs of Ataxin-3 Regulate its Polyglutamine Toxicity through Hsc70-4-dependent Aggregation. *Elife*. 9, 1–22. doi:10.7554/ELIFE.60742

Jones, D. T. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices 1 1Edited by G. Von Heijne. *J. Mol. Biol.* 292, 195–202. doi:10.1006/jmbi.1999.3091

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

Kabsch, W., and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*. 22, 2577–2637. doi:10.1002/bip.360221211

Knott, M., and Best, R. B. (2012). A Preformed Binding Interface in the Unbound Ensemble of an Intrinsically Disordered Protein: Evidence from Molecular Simulations. *Plos Comput. Biol.* 8, e1002605. doi:10.1371/journal.pcbi.1002605

Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., et al. (2020). ELM-the Eukaryotic Linear Motif Resource in 2020. *Nucleic Acids Res.* 48, D296–D306. doi:10.1093/nar/gkz1030

Lange, A., Ismail, M.-B., Rivière, G., Hologne, M., Lacabanne, D., Guillière, F., et al. (2012). Competitive Binding of UBPY and Ubiquitin to the STAM2 SH3 Domain Revealed by NMR. *FEBS Lett.* 586, 3379–3384. doi:10.1016/j.febslet.2012.07.047

Li, D.-W., and Brüschweiler, R. (2012). PPM: A Side-Chain and Backbone Chemical Shift Predictor for the Assessment of Protein Conformational Ensembles. *J. Biomol. NMR*. 54, 257–265. doi:10.1007/s10858-012-9668-8

Li, J., White, J. T., Saavedra, H., Wrabl, J. O., Motlagh, H. N., Liu, K., et al. (2017). Genetically Tunable Frustration Controls Allostery in an Intrinsically Disordered Transcription Factor. *Elife*. 6, e30688. doi:10.7554/eLife.30688

Lim, J., Hong, Y.-H., Lee, B.-J., and Ahn, H.-C. (2011). Backbone 1H, 13C, and 15N Assignments for the Tandem Ubiquitin Binding Domains of Signal Transducing Adapter Molecule 1. *Biomol. NMR Assign.* 5, 51–54. doi:10.1007/s12104-010-9265-2

Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., and Shaw, D. E. (2012). Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *J. Am. Chem. Soc.* 134, 3787–3791. doi:10.1021/ja209931w

Ma, B., Tsai, C.-J., Haliloğlu, T., and Nussinov, R. (2011). Dynamic Allostery: Linkers Are Not Merely Flexible. *Structure*. 19, 907–917. doi:10.1016/j.str.2011.06.002

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., et al. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* 102, 3586–3616. doi:10.1021/jp973084f

Marceau, A. H., Brison, C. M., Nerli, S., Arsenault, H. E., McShan, A. C., Chen, E., et al. (2019). An Order-To-Disorder Structural Switch Activates the Foxm1 Transcription Factor. *Elife*. 8, e46131. doi:10.7554/eLife.46131

Margreitter, C., and Oostenbrink, C. (2017). MDplot: Visualise Molecular Dynamics. *R. J.* 9, 164–186. doi:10.32614/rj-2017-007

Marshall, R. S., Hua, Z., Mali, S., McLoughlin, F., and Vierstra, R. D. (2019). ATG8-Binding UIM Proteins Define a New Class of Autophagy Adaptors and Receptors. *Cell*. 177, 766–781. doi:10.1016/j.cell.2019.02.009

Masino, L., Musi, V., Menon, R. P., Fusi, P., Kelly, G., Frenkiel, T. A., et al. (2003). Domain Architecture of the Polyglutamine Protein Ataxin-3: A Globular Domain Followed by a Flexible Tail. *FEBS Lett.* 549, 21–25. doi:10.1016/S0014-5793(03)00748-8

Metallo, S. J. (2010). Intrinsically Disordered Proteins Are Potential Drug Targets. *Curr. Opin. Chem. Biol.* 14, 481–488. doi:10.1016/j.cbpa.2010.06.169

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* 32, 2319–2327. doi:10.1002/jcc.21787

Milles, S., Salvi, N., Blackledge, M., and Jensen, M. R. (2018). Characterization of Intrinsically Disordered Proteins and Their Dynamic Complexes: From *In Vitro* to Cell-like Environments. *Prog. Nucl. Magn. Reson. Spectrosc.* 109, 79–100. doi:10.1016/j.pnmrs.2018.07.001

Mizianty, M. J., Peng, Z., and Kurgan, L. (2013). MFDp2. *Intrinsically Disordered ProteinsProteins*. 1, e24428. doi:10.4161/idp.24428

Mól, A. R., Castro, M. S., and Fontes, W. (2018). NetWheels: A Web Application to Create High Quality Peptide Helical Wheel and Net Projections. *bioRxiv*. doi:10.1101/416347

Mylona, A., Theillet, F.-X., Foster, C., Cheng, T. M., Miralles, F., Bates, P. A., et al. (2016). Opposing Effects of Elk-1 Multisite Phosphorylation Shape its Response to ERK Activation. *Science*. 354, 233–237. doi:10.1126/science.aad1872

Nick Pace, C., and Martin Scholtz, J. (1998). A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophys. J.* 75 (1), 422–427. doi:10.1016/s0006-3495(98)77529-0

Nielsen, J. T., and Mulder, F. A. A. (2019). Quality and Bias of Protein Disorder Predictors. *Sci. Rep.* 9, 5137. doi:10.1038/s41598-019-41644-w

Palazzesi, F., Prakash, M. K., Bonomi, M., and Barducci, A. (2015). Accuracy of Current All-Atom Force-fields in Modeling Protein Disordered States. *J. Chem. Theor. Comput.* 11, 2–7. doi:10.1021/ct500718s

Papaleo, E., Camilloni, C., Teilum, K., Vendruscolo, M., and Lindorff-Larsen, K. (2018). Molecular Dynamics Ensemble Refinement of the Heterogeneous Native State of NCBD Using Chemical Shifts and NOEs. *PeerJ.* 6, e5125. doi:10.7717/peerj.5125

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., et al. (2015). Unexpected Features of the Dark Proteome. *Proc. Natl. Acad. Sci. USA*. 112, 15898–15903. doi:10.1073/pnas.1508380112

Piana, S., Lindorff-Larsen, K., and Shaw, D. E. (2011). How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization?. *Biophysical J.* 100, L47–L49. doi:10.1016/j.bpj.2011.03.051

PLUMED Consortium (2019). Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods*. 16, 670–673. doi:10.1057/palcomms.2015.1310.1038/s41592-019-0506-8

Sato, Y., Yoshikawa, A., Mimura, H., Yamashita, M., Yamagata, A., and Fukai, S. (2009). Structural Basis for Specific Recognition of Lys 63-linked Polyubiquitin Chains by Tandem UIMs of RAP80. *EMBO J.* 28, 2461–2468. doi:10.1038/emboj.2009.160

Scott, D., Oldham, N. J., Strachan, J., Searle, M. S., and Layfield, R. (2015). Ubiquitin-binding Domains: Mechanisms of Ubiquitin Recognition and Use as Tools to Investigate Ubiquitin-Modified Proteomes. *Proteomics*. 15, 844–861. doi:10.1002/pmic.201400341

Sgourakis, N. G., Patel, M. M., Garcia, A. E., Makhatadze, G. I., and McCallum, S. A. (2010). Conformational Dynamics and Structural Plasticity Play Critical Roles in the Ubiquitin Recognition of a UIM Domain. *J. Mol. Biol.* 396, 1128–1144. doi:10.1016/j.jmb.2009.12.052

Shi, L., Wen, Y., and Zhang, N. (2014). 1H, 13C and 15N Backbone and Side-Chain Resonance Assignments of the N-Terminal Ubiquitin-Binding Domains of USP25. *Biomol. NMR Assign.* 8, 255–258. doi:10.1007/s12104-013-9495-1

Sicorello, A., Kelly, G., Oregioni, A., Nováček, J., Sklenář, V., and Pastore, A. (2018). The Structural Properties in Solution of the Intrinsically Mixed Folded Protein Ataxin-3. *Biophysical J.* 115, 59–71. doi:10.1016/j.bpj.2018.05.029

Sicorello, A., Różycki, B., Konarev, P. V., Svergun, D. I., and Pastore, A. (2021). Capturing the Conformational Ensemble of the Mixed Folded Polyglutamine Protein Ataxin-3. *Structure*. 29, 70–81. doi:10.1016/j.str.2020.09.010

Skinner, S. P., Fogh, R. H., Boucher, W., Ragan, T. J., Mureddu, L. G., and Vuister, G. W. (2016). CcpNmr AnalysisAssign: a Flexible Platform for Integrated NMR Analysis. *J. Biomol. NMR*. 66, 111–124. doi:10.1007/s10858-016-0060-y

Sora, V., Kumar, M., Maiani, E., Lambrughi, M., Tiberti, M., and Papaleo, E. (2020). Structure and Dynamics in the ATG8 Family from Experimental to Computational Techniques. *Front. Cel Dev. Biol.* 8, 1–28. doi:10.3389/fcell.2020.00420

Sora, V., Sanchez, D., and Papaleo, E. (2021). Bcl-xL Dynamics under the Lens of Protein Structure Networks. *J. Phys. Chem. B*. 125, 4308–4320. doi:10.1021/acs.jpcb.0c11562

Spiwok, V., Sucur, Z., and Hosek, P. (2015). Enhanced Sampling Techniques in Biomolecular Simulations. *Biotechnol. Adv.* 33, 1130–1140. doi:10.1016/j.biotechadv.2014.11.011

Sugita, Y., Kamiya, M., Oshima, H., and Re, S. (2019). Replica-Exchange Methods for Biomolecular Simulations. *Methods Mol. Biol.* 2022, 155–177. doi:10.1007/978-1-4939-9608-7_7

Swanson, K. A., Kang, R. S., Stamenova, S. D., Hicke, L., and Radhakrishnan, I. (2003). Solution Structure of Vps27 UIM-Ubiquitin Complex Important for Endosomal Sorting and Receptor Downregulation. *EMBO J.* 22, 4597–4606. doi:10.1093/emboj/cdg471

Tee, W.-V., Guarnera, E., and Berezovsky, I. N. (2020). Disorder Driven Allosteric Control of Protein Activity. *Curr. Res. Struct. Biol.* 2, 191–203. doi:10.1016/j.crstbi.2020.09.001

Tyler, R. C., Sreenath, H. K., Singh, S., Aceti, D. J., Bingman, C. A., Markley, J. L., et al. (2005). Auto-induction Medium for the Production of [U-15N]- and [U-13C, U-15n]-Labeled Proteins for NMR Screening and Structure Determination. *Protein Expr. Purif.* 40, 268–278. doi:10.1016/j.pep.2004.12.024

Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., et al. (2014). Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* 114, 6589–6631. doi:10.1021/cr400525m

Van Roey, K., Gibson, T. J., and Davey, N. E. (2012). Motif Switches: Decision-Making in Cell Regulation. *Curr. Opin. Struct. Biol.* 22, 378–385. doi:10.1016/j.sbi.2012.03.004

Van Roey, K., Uyar, B., Weatheritt, R. J., Dinkel, H., Seiler, M., Budd, A., et al. (2014). Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. *Chem. Rev.* 114, 6733–6778. doi:10.1021/cr400585q

Wang, Q., Young, P., and Walters, K. J. (2005). Structure of S5a Bound to Monoubiquitin Provides a Model for Polyubiquitin Recognition. *J. Mol. Biol.* 348, 727–739. doi:10.1016/j.jmb.2005.03.007

Weishäupl, D., Schneider, J., Peixoto Pinheiro, B., Ruess, C., Dold, S. M., von Zweydorf, F., et al. (2019). Physiological and Pathophysiological Characteristics of Ataxin-3 Isoforms. *J. Biol. Chem.* 294, 644–661. doi:10.1074/jbc.RA118.005801

Wen, Y., Cui, R., Zhang, H., and Zhang, N. (2014). 1H, 13C and 15N Backbone and Side-Chain Resonance Assignments of the N-Terminal Ubiquitin-Binding Domains of the Human Deubiquitinase Usp28. *Biomol. NMR Assign.* 8, 251–254. doi:10.1007/s12104-013-9494-2

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

Wright, P. E., and Dyson, H. J. (1999). Intrinsically Unstructured Proteins: Re-assessing the Protein Structure-Function Paradigm. *J. Mol. Biol.* 293, 321–331. doi:10.1006/jmbi.1999.3110

Young, P., Deveraux, Q., Beal, R. E., Pickart, C. M., and Rechsteiner, M. (1998). Characterization of Two Polyubiquitin Binding Sites in the 26 S Protease Subunit 5a. *J. Biol. Chem.* 273, 5461–5467. doi:10.1074/jbc.273.10.5461

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: 12 August 2021
doi: 10.3389/fmolb.2021.676268

An Integrative Approach to Determine 3D Protein Structures Using Sparse Paramagnetic NMR Data and Physical Modeling

Kari Gaalswyk[1†], *Zhihong Liu*[2†], *Hans J. Vogel*[2] *and Justin L. MacCallum*[1*]

[1]Department of Chemistry, University of Calgary, Calgary, AB, Canada, [2]Department of Biological Sciences, University of Calgary, Calgary, AB, Canada

Paramagnetic nuclear magnetic resonance (NMR) methods have emerged as powerful tools for structure determination of large, sparsely protonated proteins. However traditional applications face several challenges, including a need for large datasets to offset the sparsity of restraints, the difficulty in accounting for the conformational heterogeneity of the spin-label, and noisy experimental data. Here we propose an integrative approach to structure determination combining sparse paramagnetic NMR with physical modelling to infer approximate protein structural ensembles. We use calmodulin in complex with the smooth muscle myosin light chain kinase peptide as a model system. Despite acquiring data from samples labeled only at the backbone amide positions, we are able to produce an ensemble with an average RMSD of ~2.8 Å from a reference X-ray crystal structure. Our approach requires only backbone chemical shifts and measurements of the paramagnetic relaxation enhancement and residual dipolar couplings that can be obtained from sparsely labeled samples.

Keywords: paramagnetic relaxation enhancement, NMR, modeling, protein structure, integrative structural biology, calmodulin

OPEN ACCESS

Edited by:
Massimiliano Bonomi, Institut Pasteur, France

Reviewed by:
Enrico Ravera, University of Florence, Italy
Alexandre M. J. J. Bonvin, Utrecht University, Netherlands

***Correspondence:**
Justin L. MacCallum
justin.maccallum@ucalgary.ca

[†]*These authors have contributed equally to this work*

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

Received: *04 March 2021*
Accepted: *29 July 2021*
Published: *12 August 2021*

Citation:
Gaalswyk K, Liu Z, Vogel HJ and MacCallum JL (2021) An Integrative Approach to Determine 3D Protein Structures Using Sparse Paramagnetic NMR Data and Physical Modeling. Front. Mol. Biosci. 8:676268. doi: 10.3389/fmolb.2021.676268

INTRODUCTION

Protein nuclear magnetic resonance (NMR) spectroscopy has played an important role in biomolecular structure determination. To date more than 13,000 NMR structures have been deposited in the Protein Data Bank [PDB (Berman et al., 2000)], accounting for about 7.5 percent of all available protein structures (PDB Statistics, 2021). The vast majority of the deposited NMR solution structures are determined for smaller proteins or independently-folded isolated protein domains (Tugarinov et al., 2004). Without special stable isotopic labelling techniques, NMR methods struggle with structure determination of proteins larger than ~25 kDa as the slow molecular tumbling results in rapid relaxation, leading to poor resolution and spectral quality (Kay, 2016). The most commonly used method to overcome this challenge is to combine transverse relaxation optimized spectroscopy (TROSY) (Pervushin et al., 1997) with site-specific protonation in an otherwise perdeuterated background (Tugarinov et al., 2006), although there are many alternatives, e.g., Tugarinov et al. (2004) and Ruschak and Kay (2010). While such site-specific isotope labelling can dramatically increase the spectral quality and interpretability, the overall perdeuteration results in protons being sparsely distributed within the structure.

The overwhelming majority of solution NMR structures in the PDB are based around Nuclear Overhauser Effect Spectroscopy (NOESY), which provides information about through-space interactions between protons that can be used to derive distance restraints for 3D structure determination. For the homonuclear NOE to be detectable, however, the protons must be within about 6 Å or closer, which can pose a substantial challenge for sparsely-labelled samples due to the lack of proton pairs that are in close proximity within the folded protein structure (Gardner and Kay, 1998).

This phenomenon is particularly acute for samples that are labelled with protons only on the exchangeable backbone amide positions, as the amide protons within an alpha helix are typically too far away from amide protons in other secondary structure elements to produce a detectable NOE. Consequently, labelling of only the amide protons of alpha-helical proteins leads to a restraint network that is too sparse to calculate a 3D structure. Other site-specific labelling schemes can supplement amide labelling, leading to a denser restraint network (Goto and Kay, 2000). Site-specific labelling of the terminal methyl groups of isoleucine, leucine, and valine (ILV-labeling) is particularly common (Tugarinov et al., 2006), but several alternatives and complementary labeling methods exist, e.g., Kainosho and Güntert (2009), Otten et al. (2010), and Gifford et al. (2011). While these additional labelling schemes can increase the density of the restraint network, they often come at the cost of increased complexity and the need to synthesize or purchase expensive precursors that are required to generate the isotope-labelled samples.

Paramagnetic NMR methods have emerged as potentially viable alternatives, capable of providing valuable information about electron-nucleus distances up to ~20–30 Å (Koehler and Meiler, 2011; Pilla et al., 2017). Paramagnetic relaxation enhancement (PRE) experiments have been performed with native metalloproteins and proteins modified with covalent paramagnetic tags such as nitroxide spin labels and metal chelates (Bertini et al., 2005; Clore and Iwahara, 2009; Keizers and Ubbink, 2011). These techniques can be used to extend the scope of NMR methods to larger, more complex systems by providing long-range distances when short-range NOEs are unavailable or limited. Due to the long-range nature of paramagnetic relaxation enhancement, PRE experiments can provide valuable distance restraints even in sparsely labelled perdeuterated protein samples.

The utility of a distance restraint generally depends on two factors (Sullivan et al., 2003; Sullivan and Kuntz, 2004). First, restraints with short spatial distances are more valuable than those with long spatial distances because there are many more ways for two particles to be far apart than close together. Thus, a short-distance restraint provides more information than a long one. Second, this effect is more substantial for restraints involving residues that are more distant in the sequence. Thus, the most valuable restraints involve residues distant in sequence but close together in space. NOESY experiments provide powerful short spatial distance restraints (<6Å) but can miss many crucial long sequence distance restraints due to distribution of the isotope labels. In contrast, PRE experiments will yield more distance restraints due to the paramagnetic relaxation enhancement effect's long-range nature. Many of these restraints will be of limited utility due to their long spatial distances; however, the collective effect of all of these long spatial distance restraints with the remaining short spatial distance (<12 Å) restraints can still be potent. As an aside, PRE methods have also become a popular approach to explore lowly populated transient protein states (Iwahara and Clore, 2006).

Using PRE data for 3D protein structure determination presents several challenges. First, each experiment only provides information about the spatial proximity of a given proton to a single site labelled with a paramagnetic tag. Adequately determining the 3D structure requires multiple experiments with different tag locations, increasing both experimental time and cost. Second, each experiment provides only a limited amount of information (Battiste and Wagner, 2000; Gottstein et al., 2012). Although a single experiment provides information about the spatial proximity of each residue to the paramagnetic tag, much of this information is redundant. For example, if a residue is close to the tag, then neighbouring residues in the sequence are also likely to be close. Furthermore, information that a residue is close to the tag provides a far more powerful structural constraint than information that a residue is distant from the tag, but the latter occurrence is far more frequent. Third, the derived distances can be imprecise due to intermolecular interactions, secondary metal-binding sites, and diamagnetic contamination (Clore and Iwahara, 2009). Fourth, heterogeneity and dynamics (Clore et al., 1990; Ryabov and Fushman, 2007; Bertini et al., 2012) can complicate the interpretation of PRE data. Relaxation can be strongly affected by conformational heterogeneity due to the inverse sixth power relationship between the PRE and the electron-nucleus distance; i.e., a minor structural population with a strong paramagnetic effect can have a significant impact on the measured data (Clore and Iwahara, 2009). Finally, the effects of spin diffusion due to dipole-dipole coupling can limit the accuracy of the measured PRE data (Vlasie et al., 2007; Vlasie et al., 2008; Bellomo et al., 2021). These challenges have slowed the widespread adoption of PRE-based methods for structure determination in favour of traditional NOE-based approaches.

Residual dipolar coupling (RDC) measurements are a common supplemental data source to PRE and NOE-based experiments. RDC measurements are carried out on systems where the protein is weakly aligned relative to the external magnetic field. Rather than reporting on distances, RDCs report on the angles between bonded atoms (typically backbone N-H bond vectors) and the external magnetic field, which provides valuable orientational information that complements distance information from PRE or NOE experiments (Prestegard et al., 2004). RDCs have been used for structure refinement and as restraints in *de novo* structure prediction software (Banci et al., 1998; Raman et al., 2010; Prestegard et al., 2014). While many protein structures based on RDC measurements have been reported, molecular modeling and low temperature annealing procedures are often used to derive and refine the 3D structures (Chou et al., 2000; Lipsitz and Tjandra, 2004; Huang and Vogel, 2012). Clearly there is room for

more unbiased approaches to incorporate such RDC data into protein structure calculations.

Integrative approaches to structure determination (Ward et al., 2013) have emerged as practical tools for converting NMR and other experimental data into useful structural models. For example, PRE and RDC measurements have been used to drive molecular docking studies (van Dijk et al., 2005; Gelis et al., 2007; Gochin et al., 2011), as restraints in molecular dynamics simulations to generate ensembles of conformers (Dedmon et al., 2005; Asciutto et al., 2011), or they have been incorporated into Rosetta scoring functions (Lange et al., 2012; Kuenze et al., 2019). We recently demonstrated the structure determination of a small protein using PRE measurements in solid-state NMR (Perez et al., 2019). However, integrative methods are not without their own set of challenges. Even the most sophisticated methods can still struggle as the data becomes sparse, ambiguous, or unreliable, and considerable method development is often required to treat a new type of experimental data in order to correctly account for its characteristics, e.g., the conformational heterogeneity of spin-labels in PRE measurements (Iwahara et al., 2004; Anthis et al., 2011; Andrałojć et al., 2017).

Here, we show that these challenges can be overcome by using a sophisticated integrative structural biology approach called Modeling Employing Limited Data (MELD) (MacCallum et al., 2015). MELD combines experimental data from multiple sources with physical modelling to overcome the challenges of sparse, ambiguous, and difficult to interpret experimental data to infer accurate protein structural ensembles. We combine PRE and RDC measurements with secondary structure predictions based on backbone chemical shifts. We use MELD to infer the structure of Calmodulin in complex with the 20-residue smooth muscle myosin light chain kinase peptide (169 residues total). Calmodulin was selected for this exploratory work as it has an almost completely helical structure where the absence of inter-helical close contacts between amide protons makes 3D structure determination by NOE-based approaches difficult. Calmodulin-peptide complex have previously been used as models for integrative approaches using sparse NMR data (Andrałojć et al., 2014; Carlon et al., 2019). We show that MELD can identify conformations within 3 Å of a reference X-ray crystal structure using only sparse paramagnetic NMR restraints and RDCs from amide protons in combination with backbone chemical shifts, while successfully addressing conformational heterogeneity and noise in the NMR data.

EXPERIMENTAL METHODS

Calmodulin–Peptide Complex as a Model System

In this work, we illustrate our approach for the protein calmodulin in complex with the smooth muscle Myosin Light Chain Kinase (smMLCK) peptide. Throughout, we will use a previously solved crystal structure of this complex [PDB ID: 1cdl (Meador et al., 1992)] as a reference.

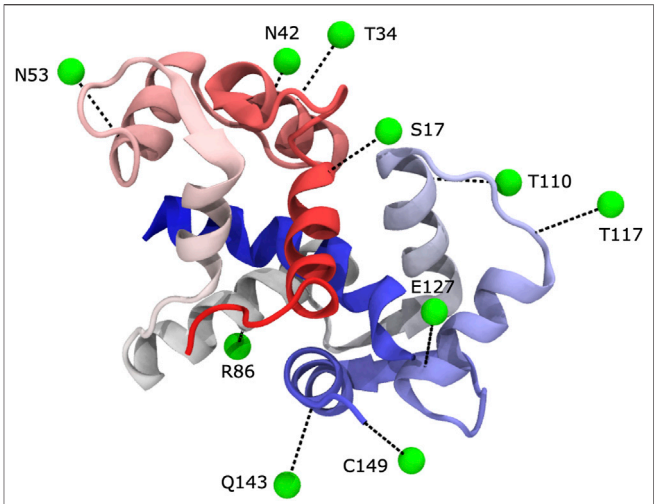


FIGURE 1 | We carried out PRE experiments with ten different label sites. In each experiment, Calmodulin was MTSL-labeled at a different position. Spin labels were generally located in predicted surface-exposed sites within secondary structures. Spin labels are shown in green as virtual sites (see text), and their corresponding cysteine mutation linkage site is shown in black (PDB 1cdl).

Overview of Labeling Strategy

In previous studies, specific nitroxide spin-labeled target peptides that bind to calmodulin were used; in this manner it was possible to map out the orientation of the peptide with respect to the protein (Zhang et al., 1995; Yuan et al., 2004). In this work, we collected PRE data for a total of ten spin-labelled protein sites (**Figure 1**). Nine of the sites on the protein were chosen to be solvent-exposed and within secondary structure elements by manual inspection of predicted secondary structure and solvent exposure. The remaining site, C149, was a single-residue extension of the C-terminus. To better simulate the process for a system without a previously determined structure, the protein's known structure **was not used** in choosing the spin-labelled sites. Indeed, we learned later that several of the selected sites provided little useful information because they are either distant from the rest of the protein or they provided information that is mostly redundant with that obtained from other labeling sites.

To simulate the limited availability of isotopically labelled peptide, either due to cost or difficulty of production, only four of the ten spin-label data sets (chosen randomly) were collected with isotopically labelled peptide. The remaining six data sets were collected without labelled peptide, which results in the peptide being present but unlabeled and undetectable in the ^{1}H, ^{15}N HSQC NMR experiments.

Protein Production

The ten single-cysteine point calmodulin (CaM) mutants (S17C, T34C, N42C, N53C, R86C, T110C, T117C, E127C, Q143C, 149C) were made by standard site-directed mutagenesis methods for attachment of the thiol-specific nitroxide spin label (1-oxyl-2,2,5,5-tetramethyl-δ-3-pyrroline-3-methyl)

methanethiosulfonate (MTSL, Toronto Research Chemicals). Correctness of the mutations was confirmed by DNA sequencing. Calmodulin contains no Cys residues, so highly site-specific labeling can be obtained in this manner. MTSL is a relatively compact, yet highly reactive molecule compared to other commercially available nitroxide spin-labels that have been used to modify Cys residues; its shorter more rigid structure would be an advantage for the PRE studies [for discussion see for example (Guo et al., 2008; Fawzi et al., 2011)]. ^{13}C and ^{15}N-labeled CaM was expressed in M9 minimal medium with 99.9% $^{15}NH_4Cl$ and $^{13}C_6$-glucose (0.5 gr/L and 3 gr/L, respectively; Cambridge Isotopes Laboratories) as isotope sources. Proteins were expressed and purified as described previously (Liu and Vogel, 2012; Ishida et al., 2016).

We followed a standard protocol for attaching the nitroxide spin-label to each single-cysteine CaM mutants with the spin-labelling reagent MTSL (Battiste and Wagner, 2000). To prepare the CaM/smMLCK complex sample, a 1.2-fold excess of either unlabeled or labelled peptide was mixed with each CaM mutant protein. All preparations were divided into two NMR samples. One sample was reduced to inactivate the spin-label by adding a 3-fold excess of ascorbic acid.

Peptide Production

A construct with a 6xHis-KSI (D38A) fusion-protein tag was generated for smMLCK peptide expression in *Escherichia coli* (Jaroniec et al., 2005; Ishida and Vogel, 2010). The ketosteroid isomerase (KSI) coding sequence generates an insoluble protein, and this directs the protein-peptide fusion directly into inclusion bodies, where they are protected from proteolytic cleavage (Hwang et al., 2014). A linkage sequence "GGGGSSDP" with the Asp-Pro acid cleavage site was designed between the KSI protein and the sequence of the smMLCK peptide. The entire 6xHis-KSI-GGGGSSDP-smMLCKp gene sequence was inserted between the NdeI and XhoI sites of the pET15b(+) plasmid (Novagen), which was subsequently transferred into BL21(DE3) *E. coli* cells for protein expression. The cells were grown in either LB media (for unlabeled peptide) or minimal M9 media (containing $^{13}C_6$-glucose and $^{15}NH_4Cl$ isotope to produce isotope-labeled peptide) and they were induced at OD_{600} = 0.6 with 1 mM IPTG for 4 h at 37°C. A cell lysate was prepared as previously described. The insoluble fusion protein was separated after one hour of centrifugation (18,000 rpm) and then resuspended in 6 M guanidine hydrochloride. Impurities were removed before the insoluble proteins can be extracted with metal chelate chromatography on a nickel affinity column. After extensive dialysis with double distilled H_2O, the precipitated insoluble protein was collected and the Asp-Pro bond was cleaved in 10% formic acid at 80°C for 90 min (Hwang et al., 2014). The protein-peptide mixture was flash frozen with liquid nitrogen and lyophilized. Insoluble proteins and other impurities were removed after the lyophilized mixture was resuspended in a 20 mM Tris-HCl buffer (pH = 8.0). Finally, the unlabeled and isotope-labeled smMLCK peptides were purified with reverse-phase HPLC (COSMOSIL $5C_{18}$-AR-300, Nacalai United States). All purified peptides were lyophilized and stored at −20°C for further use. The final peptide sequence after cleaving is PARRKWQKTGHAVRAIGRLSS. The N-terminal proline is not observable in the NMR experiments and was not included in modeling with MELD.

Chemical Shift Assignments

All NMR experiments were carried out on a 600 MHz Bruker AVANCE spectrometer with a field strength of 14.1 T. Backbone resonance assignments for the protein and the bound peptide were confirmed with the following 3D experiments: HNCO, HNCA, HNCOCA, HNCACB, and CBCA(CO)NH, as described previously (Liu and Vogel, 2012). All data were processed using NMRPipe (Delaglio et al., 1995) and analyzed with the program NMRView (Johnson and Blevins, 1994). All chemical shifts from these experiments were used to obtain backbone torsion angles from the program TALOS+ (Cornilescu et al., 1998; Shen et al., 2009). Secondary structure elements as identified through the assigned chemical shifts were as expected based on the known structure.

Paramagnetic Relaxation Enhancement Measurements

Two ^1H, ^{15}N HSQC spectra were obtained for each spin-label construct. Each system contained each ^{15}N-labeled protein and either unlabeled or ^{15}N-labeled peptide, depending on the spin-label site (S17C, N53C, T127C, and 149C had isotopically labeled peptide). One HSQC was collected with active spin-label, whereas the other HSQC was collected with reduced, inactivated spin-label. The distances between the spin-label and the affected nuclei were calculated using the two-time point method (Iwahara et al., 2007).

Residual Dipolar Coupling Measurements

Finally, to supplement the PRE experiments, we obtained RDC measurements for the amide groups in the complex with a sample where both the protein and peptide are isotopically labelled. Residual dipolar couplings (RDC) were measured for the CaM/smMLCK complex sample in a partially aligned media, which contains 2 mM bis-Tris (pH = 7.0), 300 mM KCl and 16 mg/ml Pf1 bacteriophage (Asla Biotech Ltd.). The IPAP-HSQC experiment was used for the RDC measurements (Ottiger et al., 1998). In these experiments the effects of dipole-dipole cross-correlated relaxation can impact the accuracy of $^1J_{NH}$ splitting measured from the spectra introducing a small residual bias in the RDCs. While these systematic errors can be eliminated by using a selectively-decoupled sequence (Yao et al., 2009), the errors are small relative to the magnitude of the measured RDCs and are expected to have a minimal effect on structure determination (Yao et al., 2009). Our work uses only a single RDC alignment. Notably, a mutant of Calmodulin is capable of selective binding to lanthanides, which provides a strategy for the measurement of multiple RDCs (Bertini et al., 2009). A quantitative assessment of protein mobility/heterogeneity by RDC would require the use of multiple alignments (Barbieri et al., 2002; Tolman, 2002; Bouvignies et al., 2006; Higman et al., 2011; Guerry et al., 2013; Andrałojć et al., 2015). However, as discussed further

below, it is not currently possible to conduct such an analysis with the MELD approach, as MELD compares individual structures, rather than ensembles of structures, to the experimental data.

COMPUTATIONAL APPROACH

Overview of Modeling Employing Limited Data Approach

Here, we employ MELD, a physics-based Bayesian approach for structural determination to infer the ensemble of structures most consistent with the known physics of protein structure and experimental data (MacCallum et al., 2015; Perez et al., 2015). MELD uses a Bayesian framework to combine a physics-based prior distribution with a data likelihood function to make statistically consistent inferences about conformations that explain the experimental data.

MELD uses Bayes' theorem:

$$p(x|D) \propto p(x)p(D|x), \tag{1}$$

where x represents the atomic coordinates and D represents the data. The physics-based prior, $p(x)$, specifies which structures are more likely *a priori* and determines the distribution of structures in the absence of data. In the present study the physics-based prior is given by the Amber ff14SB force field (Maier et al., 2015) with a grid-based torsion potential (Perez et al., 2015) and the OBC generalized-Born implicit solvent model (Onufriev et al., 2004). The likelihood function, $p(D|x)$, captures the compatibility between the data and some structure x. In MELD, the likelihood function takes the form of a unique restraint function (MacCallum et al., 2015), explained in more detail below. The goal of Bayesian inference is to compute the posterior distribution, $p(x|D)$, which is the most statistically consistent inference given the prior, likelihood, and data.

As discussed in the Results section below, the term *ensemble* is highly overloaded in structural biology and care is required in interpretation. MELD belongs to the class of methods where a single structure, rather than entire ensemble of structures, is considered in the likelihood function, such that each member of the ensemble individually agrees with the experimental data. Any conformational heterogeneity (e.g., flexible loops) may represent true intrinsic heterogeneity, but may also simply reflect a lack of data. As such, MELD produces a form of *uncertainty ensemble* in the terminology of Bonomi et al., 2017.

Overview of Experimental Data

The input to our approach is: 1) the protein sequence, 2) TALOS+ secondary structure predictions derived from backbone chemical shifts (Cornilescu et al., 1998; Shen et al., 2009; MacCallum et al., 2015), 3) distance restraints derived from PRE measurements, and 4) orientational restraints derived from RDC measurements. We have recently demonstrated the success of a similar approach for PRE measurements in solid-state NMR (Perez et al., 2019).

PRE data is often both noisy and sparse (Kim et al., 2014), which makes structural inference challenging. Despite collecting data for ten spin-label positions, we can derive only a few distance restraints that are short in spatial distance (in this case, we define short as <12 Å) (**Figure 2**). Of these short spatial distances, only a small number correspond to residues that are distant in sequence, which would provide the most information about folding (Gottstein et al., 2012). Furthermore, as stated previously, to simulate the limited availability of isotopically labelled peptide, only four of ten datasets (S17C, N53C, T127C, and 149C) had labelled peptide, and there are no short distance PREs between the peptide and the protein. This leaves the peptide's correct placement to be dictated by longer, less informative spatial distance restraints and the physical model, which makes accurate inference more challenging.

Deriving Distances From Paramagnetic Relaxation Enhancement Data

Our first step was to develop a consistent method to convert ensemble-averaged PRE measurements into distance restraints. PRE data were turned into approximate distances using the Solomon-Bloembergen equations following the standard approach (Battiste and Wagner, 2000; Iwahara et al., 2007).

For nitroxides, Curie-spin relaxation is negligible and the transverse relaxation enhancement, Γ_2, is dominated by direct dipole-dipole interactions (Clore and Iwahara, 2009). In this case, Γ_2 is related to the distance between the paramagnetic center and the observed nucleus, r:

$$\Gamma_2 = \frac{K}{r^6}\left[4\tau_c + \frac{3\tau_c}{1 + \omega_H^2\tau_c^2}\right]$$

where ω_H is the Larmor frequency of the proton, and τ_c is the correlation time for the electron-nuclear interaction defined as $\tau_c = (\tau_r^{-1} + \tau_s^{-1})^{-1}$ where τ_r is the rotational correlation time, and τ_s is the electron relaxation time. Previous experiments (Lee et al., 2002) have determined that $\tau_c \approx 9.5\,ns$ for Calmodulin in complex with smMLCK peptide at 25°C. For nitroxides, the electron relaxation time is long (Iwahara et al., 2007) ($\tau_s > 10^{-7}s$), so the rotational correlation time dominates and $\tau_c \approx \tau_r \approx 9.5\,ns$. K is given by:

$$K = \frac{1}{15}\left(\frac{\mu_0}{4\pi}\right)^2 \gamma_1^2 g^2 \mu_B^2 S(S+1)$$

where μ_0 is the permeability of vacuum, γ_1 is the nuclear gyromagnetic ratio, g is the electronic g factor, μ_B is the Bohr magneton, and S is the electron spin quantum number. Γ_2 can be estimated using a two-point time measurement (Iwahara et al., 2007):

$$\Gamma_2 = R_{2,para} - R_{2,dia} = \frac{1}{\Delta T}\ln\frac{I_{dia}(T_b)I_{para}(T_a)}{I_{dia}(T_a)I_{para}(T_b)}$$

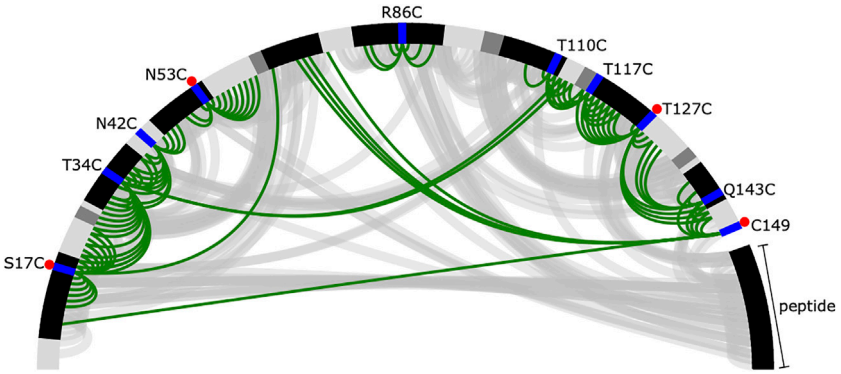


FIGURE 2 | Summary of short distances from the reference crystal structure and inferred from the PRE data. The protein and peptide backbones are represented in a half-circle where the colour depends on the secondary structural element (black—helix, dark grey—extended, light grey—loop). The spin-label locations are shown in blue. A red dot indicates the experiment was performed with labelled peptide. Short distances [(i, j) pairs where $|i - j| > 4$ and $r_{ij}^{C\alpha} < 7.6$ Å] derived from the crystal structure are shown as grey arcs. Short distances ($r_{ij}^{label-NH} < 12$ Å) derived from the PRE data are shown in green. Note that C149 is a single-residue extension of calmodulin, which is 148 residues long.

TABLE 1 | Distance bounds for calculated PRE distances, r.

	PRE distance range (Å)	Restraint upper bound (Å)	Restraint lower bound (Å)
Short	$r \leq 12$	17	0.0
Medium	$12 < r < 20$	$r + 5$	$r - 5$
Long	$r \geq 20$	∞	15

Ranges are chosen based on the nature of the PRE and include a +/− 5 Å buffer to account for heterogeneity and flexibility.

where ΔT is a time delay chosen to minimize the error in Γ_2, and I_{dia} and I_{para} are the peak intensities for the diamagnetic and paramagnetic samples, respectively (Iwahara et al., 2007). In this work, we use $\Delta T = 20\,ms$.

Incorporating Paramagnetic Relaxation Enhancement Information Into Modeling Employing Limited Data Calculations

The distances derived from PRE data correspond to ensemble averages with an r^{-6} weighting, but in MELD (and most other structure determination software), restraints are applied to single structures rather than ensembles. To account for conformational heterogeneity of both the protein and the flexible spin-label, the PRE-derived distances are turned into flat bottomed harmonic restraints that allow for a range of distances without penalty. This approach is a tradeoff that ensures that individual structures are not erroneously over-restrained but this can allow discrepancies between the measured and modelled ensemble averages. Our aim is to produce an approximately correct ensemble starting from an extended chain. If desired, the resulting ensemble can be further refined using a variety of ensemble approaches (Boomsma et al., 2014; Hummer and Köfinger, 2015; Bonomi et al., 2016; Gaalswyk et al., 2018).

We divided the data into *short*, *medium*, and *long* distances with corresponding upper and lower bounds (**Table 1**). *Short* and *long* distances are difficult to quantify with precision because the peak is either completely broadened for residues close to the spin-label or barely changes intensity for those that are far away. These distances are turned into broad restraints that either start from zero or extend to infinity for *short* and *long*, respectively. *Medium* distances correspond to peak intensity changes that can be quantified more precisely and are turned into restraints centered around the predicted value. All distances include a buffer of ±5 Å of the measured distance to account for the flexibility of the spin-label and noise in the experimental data (Perez et al., 2019).

Due to noise in the experimental data, partially overlapping peaks, and instantaneous fluctuations in both the protein structure and the position of the spin label, we observed that restraints are sometimes violated even with a ±5 Å buffer. To mitigate this issue, we used MELD's unique ability to require that only a certain fraction of the restraints must be satisfied by each structure. We set this *active fraction* to 0.9. Essentially, as long as 90 percent of the restraints are satisfied, the resulting restraint energy will be zero. We treat the remaining restraints as being derived from spurious data, so they are entirely ignored. Every timestep, MELD decides which restraints are active based on the current structure. Further details can be found in the SI and in MacCallum et al. (2015).

In our approach, the various hyperparameters (boundaries between *short*/*medium*/*long*, size of buffer, active fraction) are fixed. One potential improvement would be to place a hyperprior on these values and infer them using an extended Bayesian approach like Inferential Structure Determination (Rieping et al., 2005). This would allow the data and physical model to determine the most likely values of these hyperparameters, rather

than requiring their specification *a priori*. As MELD does not currently support inference of hyperparameters, we chose the simpler approach of setting a wide buffer and lower active fraction, which potentially sacrifices a small amount of information.

The spin-label was modeled using virtual sites (Banci et al., 1996) following the approach of Islam and Roux (Islam et al., 2013; Islam and Roux, 2015). These virtual sites represent the spin-label as a non-interacting dummy particle to simplify the simulation without losing relevant information for structural refinement. These simplified dummy nitroxide spin-labels are parameterized to match the spin-labels' 3D spatial distribution and dynamics in all-atom simulations. The virtual sites are non-interacting, allowing us to account for all ten spin labels in a single simulation without the risk of interactions between them.

Secondary structure restraints were derived from TALOS+ (Cornilescu et al., 1998; Shen et al., 2009) and used to restraint MELD simulations as previously described (MacCallum et al., 2015). Our approach works by first breaking the protein into overlapping 5-residue fragments. If 4/5 of the residues in the fragment are predicted to be helical or extended, then the fragment is restrained using a combination of torsion and distance restraints (MacCallum et al., 2015). All secondary structure restraints are then combined into a collection with an active fraction of 0.95, which allows 5 percent of fragments to differ from their predicted secondary structure.

Incorporation of Residual Dipolar Coupling Information Into Modeling Employing Limited Data

The traditional approach to incorporate RDCs into simulations is based on solving for the optimal alignment tensor, which requires solving a system of equations every time step using singular value decomposition (SVD) or related methods, which can be computationally intensive (Losonczi et al., 1999). We found this to be particularly problematic in the GPU-accelerated framework of MELD, where this traditional approach led to a 300 percent increase in run time (data not shown), primarily due to the extreme speed of the rest of the force/energy calculations and the challenge of efficiently parallelizing SVDs for small systems of equations on a GPU. To mitigate this issue, we instead followed the approach in Habeck, Nilges, and Riepling (Habeck et al., 2008), which we implemented using an OpenMM CustomCentroidBondForce (Eastman et al., 2017). In our implementation, the alignment tensor elements are encoded in two non-interacting dummy particles coupled to the rest of the system through an additional energy term. This approach has two benefits. First, it is dramatically faster than the standard approach on GPUs, with negligible cost compared to the calculation of the non-bonded forces. Second, this approach accounts for uncertainty and produces a joint distribution of alignment tensors and structures, providing a Bayesian posterior estimate of the conformational ensemble that better reflects uncertainty. A full explanation of our implementation can be found in the SI. To account for uncertainty in the experimental data and to avoid erroneously over-restraining individual structures to the ensemble average data, we use a flat-bottomed restraint where the energy is zero if the computed RDC is within 1.5 Hz of the measured value. Another approach that avoids the need to solve for the alignment tensor is given in Camilloni and Vendruscolo (2015).

RESULTS AND DISCUSSION

Interpretation of Modeling Employing Limited Data-Computed Ensembles

The term "ensemble" is highly overloaded in structural biology, with a variety of meanings in different contexts (Bonomi et al., 2017; Andralojc and Ravera, 2018; Gaalswyk et al., 2018). Care must be taken to ensure correct interpretation.

MELD samples from a well-defined conformational ensemble (Gaalswyk et al., 2018), specifically the Bayesian posterior distribution given by **Eq. 1**. Interpretation of this ensemble is straightforward: it is the statistically consistent posterior distribution inferred from the prior, likelihood, and data. How should one select or report structures from this ensemble? The approach we take here is simply to report all structures, as this fully captures the heterogeneity of the distribution. If there is a limit to the number of structures reported, one simple, correct approach is to select a subset of structures at random. Alternatively, one could cluster the structures and report the cluster medoids and populations along with some measure of the variance of structures within the cluster. A variety of approaches are supported by the PDB-Dev archival system which is being developed for structural models obtained using integrative modeling (Vallat et al., 2018). However, we note that since MELD samples structures with the correct posterior probabilities, it is incorrect to further select structures based on other criteria, such as selecting the lowest energy structures.

A second consideration in ensemble interpretation is the nature of the likelihood function. The experimental measurements are averages over a thermodynamic ensemble. The most correct modeling approach is to use a likelihood function that considers an entire ensemble of models, ensuring that the predicted average quantities match their corresponding experimental measurements (Bonomi et al., 2017; Andralojc and Ravera, 2018; Gaalswyk et al., 2018). This is an ill-defined inverse problem (Bonomi et al., 2017; Andralojc and Ravera, 2018), so regularization is required, typically in the form of physical modelling and entropy maximization (Bonomi et al., 2017; Andralojc and Ravera, 2018; Gaalswyk et al., 2018). While conceptually appealing, ensemble likelihood methods are complex with high computational requirements. Alternatively, most methods in structural biology, including the MELD approach described here, use single-structure likelihoods (Boomsma et al., 2014). These methods are overly restrictive, as they require each member of the ensemble to be consistent with the data to within some tolerance. In the current approach, we use relatively wide tolerances, but this still does not guarantee that that the computed ensemble accurately models the true distribution.

The primary issue is that for a given set of experimental measurements, there are many possible ensembles that could produce it. The ensemble that MELD generates ensures that each

TABLE 2 | Grouping of restraints for simulations and description of individual trials.

Trial	Number of collections	Description
Trial 1	1	All PRE restraints in a single collection
Trial 2	3 × 10 = 30	PRE restraints are combined by spin-label position and distance (Short, Medium Long) into 30 collections.
Trial 3	3 × 10 = 30	PRE restraints are combined by spin-label position and distance (short, medium long) into 30 collections. RDC restraints are included.
Trial 4	3 × 10 = 30	PRE restraints are combined by spin-label position and distance (short, medium long) into 30 collections. RDC restraints are included. Simulation starts from native crystal structure.

PRE restraints have a force constant of 250 kJ mol^{-1} nm^{-2}. RDC restraints have a force constant of 0.5 kJ mol^{-1} Hz^{-2}.

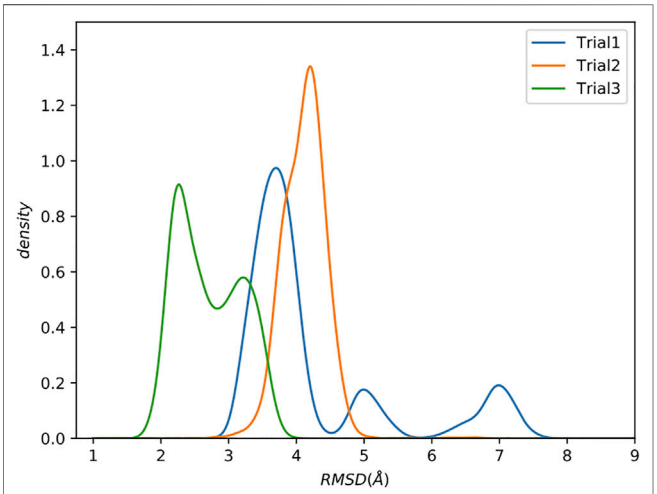


FIGURE 3 | Increasing complexity of how restraints are incorporated results in better sampling. Kernel Density Estimate (KDE) plot of backbone RMSD to the reference structure. The last 0.5 μs were analyzed. Flexible tails at the N- and C-termini of the protein were excluded.

structure is in reasonable agreement with the data and allows for a reasonable degree of flexibility. However, the MELD average might not precisely match the experimental measurement due to the use of wide tolerances. Furthermore, the true ensemble could be "broader" than the one generated by MELD—the true ensemble could have many structures that are individually in poor agreement with the data, while still having the same ensemble average, see **Figure 1** of Gaalswyk et al. (2018) for a simple illustration. In the terminology of Bonomi et al. (2017), MELD produces an *uncertainty ensemble*, where heterogeneity in the calculated ensemble could represent true heterogeneity in the system or could simply reflect a lack of data for some part of the protein. The single-structure approach is likely reasonable when the true ensemble has only a modest amount of heterogeneity, e.g., small fluctuations around an average structure, but could be expected to break down for highly heterogenous systems, e.g., systems containing intrinsically disordered regions.

Although we do not pursue it here, a promising approach would be to use a method like MELD to compute an initial approximate ensemble that could then be used as a starting point for ensemble approaches (Boomsma et al., 2014; Hummer and Köfinger, 2015; Bonomi et al., 2016; Gaalswyk et al., 2018).

The Accuracy of Inference Depends on the Protocol Used

To determine how the experimental data should be incorporated, we performed several simulations varying in their set up (Trial1–Trial4). We explored various ways of combining the restraints into *collections* (**Table 2**). In MELD, at every timestep, the restraints in a collection are sorted by energy, and the *active fraction* with the lowest energy are "active" and contribute their forces and energy to the system, while the remainder are "inactive" and ignored. The division of restraints into collections matters because it determines how MELD decides which restraints are active and which are ignored. For example, Trial1 combines all of the restraints into a single collection. In this case with an active fraction of 0.9, MELD can freely ignore any 10 percent of the restraints, which could be, for example, ignoring one of the ten spin labels entirely. Trial2 separates the restraints by spin-label and into *short*, *medium*, and *long-distance* ranges, resulting in 30 collections. Now MELD can only ignore 10 percent from each spin label/distance combination, while the remaining 90% will be active. Trial3 extends Trial2 by adding the RDC restraints. Trial1–Trail3 start from an extended conformation generated by the tleap tool from the AmberTools suite (Case et al., 2021). Trial4 follows the same protocol as Trial3 but starts from the reference crystal structure as a control.

For each trial, we ran a 2.5 μs replica exchange simulation using 48 replicas. The temperature and the force constant for each restraint collection varied across replicas (see SI for details). The last 0.5 μs of the lowest replica was used for analysis.

Using Only Paramagnetic Relaxation Enhancement-Derived Information Leads to Modest Structural Quality

We compare the trials using kernel density estimation plots (KDE; see Supporting Information for details) of the backbone root mean square deviation (RMSD) to the reference structure [PDB: 1cdl (Meador et al., 1992)], excluding the flexible tails at the N and C terminals of the protein which are not present in the reference (**Figure 3**).

Trial1 is the most straightforward approach and combines all of the data into a single collection. Many of the structures have relatively low RMSD to the reference (<4Å), which is

promising considering the rather limited experimental data, but there are also structures with much higher RMSDs of ~5 Å and ~7.5 Å RMSD. There are various explanations for these high RMSD conformations, but perhaps the simplest is that this way of grouping all restraints into a single collection allows MELD to ignore short spatial distance restraints that would otherwise eliminate these conformations. As previously stated, the utility of a restraint depends on its spatial distance. Shorter distances provide highly constraining information. However, this highly constraining nature means that these restraints are more difficult to form, leading MELD ignore them in favour of more easily satisfied restraints.

To test this hypothesis, in Trial2, we further subdivided the restraints by separating the *short*, *medium*, and *long* restraints from each dataset into separate collections, resulting in 30 total collections.

The resulting RMSD distribution is centered at a modest RMSD of ~4 Å, which is slightly worse than the mode from Trial1. However, this method of combining restraints into collections has wholly eliminated the high RMSD conformations.

Although the RMSDs obtained are only modest (~4 Å), these results were obtained with a very sparse dataset with only one spin-label per 17 amino acids. This equates to 6.4 total restraints per residue, and only 0.8 short-distance restraints per residue. For context, NOE-based structures from fully protonated samples typically have >15 NOE restraints per residue, all with short distances.

Based on visual examination, several of our spin-label sites appear to give restraints that are largely uninformative, either because they are far from the remainder of the protein (e.g., R86C) or because they are mostly redundant with other spin-label positions (e.g., T110C), see **Figures 1**, **2**. Our results could be improved with a more judicious choice of the 10 label sites, but it is unclear how to do this without pre-existing knowledge of the structure. The results could likely be improved further by adding additional spin-label sites using the calculated structural ensemble to optimize probe location, although we do not pursue this here. Such an iterative strategy could be a viable approach to improve model accuracy but comes at an additional experimental cost. More rigid spin-labels (Fawzi et al., 2011) could also improve results, as MTSL still displays significant conformational heterogeneity that results in less precise distance restraints.

Residual Dipolar Couplings Provide Complementary Information That Improves Accuracy

Despite collecting data for 10 different spin-label sites, few yielded informative short spatial distance, high sequence distance restraints (**Figure 2**), limiting the models' achievable accuracy to relatively modest RMSDs of around 4 Å. Rather than collect additional PRE data, we instead chose to explore the utility of combining PRE information with residual dipolar couplings (RDCs) measured for the amide groups.

FIGURE 4 | Superposition of a typical model (green) from the Trial3 ensemble with the reference structure (white). Peptide is shown in dark green and grey respectively. The superposition was over residues 4–146 of calmodulin plus the peptide. The backbone RMSD of this structure is 2.8 Å, which is near the mean of the ensemble. Structures with RMSDs as low as 1.6 Å are sampled.

Residual dipolar couplings provide information about the orientation of amide NH bonds complementary to the distance information from PRE experiments. In Trial3, we combined PRE information (using the same strategy as Trial2) with RDC data. The inclusion of RDC data led to a substantial improvement in the RMSD (**Figure 3**). The RMSD ranges from approximately 1.6–4.0 Å with an average RMSD of 2.8 Å, including both Calmodulin and the smMLCK peptide (**Figure 4**). This improvement of RMSD upon inclusion of RDC data is consistent with previous studies showing that RDC data provides valuable information on the relative orientation of the two lobes of calmodulin (Mal et al., 2002; Gifford et al., 2011).

RDCs provide information about how the amides are oriented, which, when combined with secondary structure restraints (derived from the measured backbone chemical shifts) and distance restraints (derived from PREs), serves to dramatically limit the possible structures that simultaneously agree with the experimental data and the physical model.

To assess the potential quality of sidechain packing, we examined the single best structure obtained during our simulations, which has an RMSD of 1.6 Å (**Supplementary Figure S1**). For this "best" structure, the RMSD for sidechain heavy atoms is 2.1 Å for all sidechains and 1.4 Å for core sidechains. This is notable as there are no restraints on the side chains themselves, only between the spin labels and the backbone amide protons. This packing phenomenon with MELD has been noted previously and can be attributed to the accuracy of the physical model (Perez et al., 2019). However, we note that the

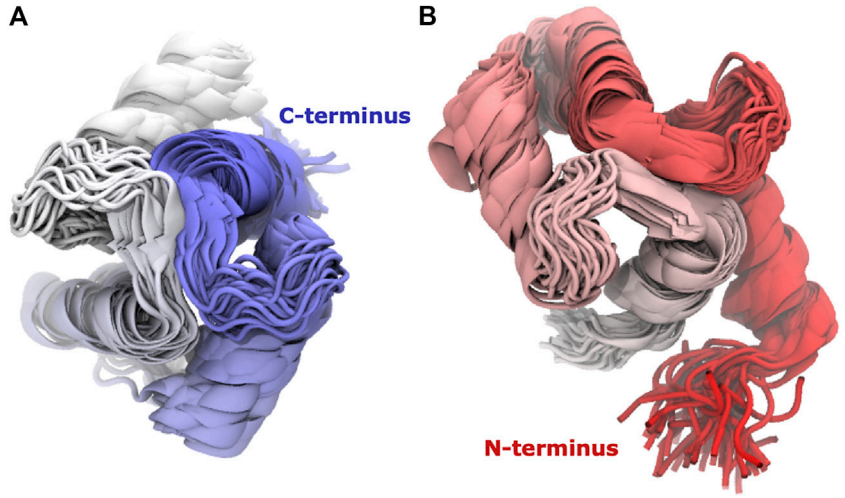


FIGURE 5 | The domains have tightly clustered ensembles. Superpositions of **(A)** C-lobe (residues 82–149) and **(B)** N-lobe (residues 1–76) of Calmodulin for Trial3. Every 100th frame from the last 0.5 microseconds is shown, coloured from N-terminus (red) to C-terminus (blue).

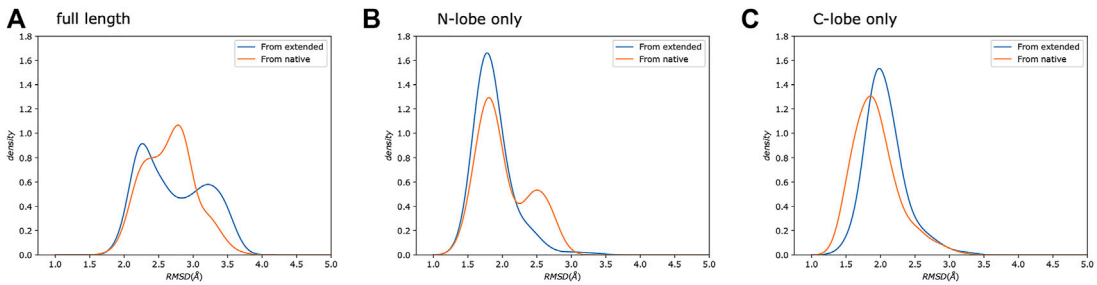


FIGURE 6 | Simulations from extended and native show similar distributions. We show a comparison of the same protocol started from either an extended chain (Trial3, blue) or from the reference crystal structure (Trial4, orange). Each panel shows the backbone RMSD compared to the reference crystal structure. We compare: **(A)** the full-length protein, as well as the **(B)** N-, and **(C)** C-lobes.

sidechain and backbone RMSDs are generally correlated, and this structure has a lower backbone RMSD than average, so the average side chain RMSDs will be higher than these figures.

Despite Limited Data, the Peptide Is Routinely Placed Correctly

As noted previously, the experimental data contained no short distance PREs to the peptide, so placement of the peptide is dependent on a combination of *medium* and *long* restraints with the physical model. Furthermore, only 4 of 10 experiments contained labelled peptide, with the peptide undetected in the remaining experiments. Nevertheless, the combination of available data and the physical model was still able to routinely position the peptide correctly (the peptide is included in the RMSD calculations shown in **Figure 3**). The structure of calmodulin depends on the peptide and its binding (Barbato et al., 1992; Ikura et al., 1992), so correct placement of the peptide is critical.

The Individual Lobes Are Better Defined Than the Complex

Calmodulin consists of two lobes connected by a flexible linker that becomes structured upon peptide binding (Barbato et al., 1992). Examination of each lobe individually shows that our modeled ensembles are tightly clustered (**Figure 5**), indicating that most of the heterogeneity in our calculated ensemble arises from the relative motion of the two lobes. If we consider the RMSD of each lobe to the reference individually, the results are consistently lower than for the whole protein (**Figure 6**). The RMSD of the C-lobe to the reference structure is ~2.1 Å (**Figure 6B**), which is consistent with typical RMSDs for small globular proteins seen in MD simulations. The results for the N-lobe are similar. The resulting heterogeneity in the relative orientation of the two domains should be interpreted with caution, due to the use of a single-structure likelihood, as discussed above.

A previous study (Carlon et al., 2019) examining the joint X-ray/NMR refinement of Calmodulin in complex with the

Death-Associated Protein kinase (DAPk) peptide revealed poor agreement between X-ray and NMR data due to large interprotein contacts in the crystal that stabilize a conformation that is in poor agreement with the solution NMR data. The crystal structure of the full-length DAPk protein in complex with Calmodulin lacks these contacts and is in much better agreement with the NMR data. These results highlight the need for caution when comparing structures determined by X-ray crystallography and NMR, particularly in cases where flexibility can be expected.

A study of 109 pairs of NMR and crystal structures (Sikic et al., 2010) showed that typical Cα RMSDs range from ~0.5 to 4 Å with a mean of ~2.0 Å when using the DALI (Holm and Sander, 1993) alignment. The typical variability of models within a given NMR ensemble was similar (Sikic et al., 2010). Our results for the individual lobes of calmodulin give similar average RMSDs, indicating that our approach is producing results comparable to typical NMR structures using NOEs and fully protonated samples despite the substantial sparsity in our data. Our results for the full-length complex produce a slightly higher average RMSD, which reflects heterogeneity in the exact relative placement of the two lobes.

To further test our predictions' quality, we also ran calculations using the same protocol as Trial3 but starting from the reference crystal structure rather than from an extended chain (**Figure 6**), which sets a bound on the possible accuracy that could be obtained. The resulting RMSD distributions are similar to our predictions. This indicates that given: 1) the available experimental data, 2) potential limitations of the physical model used, 3) the use of a single-structure rather than ensemble likelihood, and 4) the challenges of comparing with a static crystal structure, the results obtained using MELD are essentially as good as they could be.

Computational Requirements

Each calculation was over 48 replicas for 2.5 μs, which required approximately 6 days on 48 GTX 1080Ti GPUs. However, examination of the RMSD over time (**Supplementary Figure S2**) shows that the simulations appear to be converged after ~500 ns. In hindsight, the simulation length could have been reduced to 1 μs without a loss in quality, which would reduce simulation time to 2.5 days. While computationally expensive, our approach is readily feasible with access to advanced research computing or cloud computing resources.

CONCLUSION

Our approach can generate accurate protein structures starting from an extended chain using backbone chemical shifts in combination with PRE and RDC measurements from backbone amide labeled samples. We demonstrate this on a relatively large, complex system with only one spin label per 17 residues. This gives an average of 6.4 PRE restraints per residue of which less than 0.8 per residue are short-distance, compared to the >15 short-distance restraints per residue that are typical in NOE-based structure determination. Our approach is able to routinely identify dominant conformations within 3 Å of the reference crystal structure for calmodulin in complex with a peptide and correctly places the peptide despite a lack of information relating the peptide to the protein. These results approach the quality of gold-standard, fully protonated NMR structures based on NOEs, but were obtained from a far sparser dataset using methods that are more applicable to large proteins. The inclusion of RDCs highlights their value in structure determination with minimal PRE-derived distance restraints. These results showcase the importance of spin label location and the effect it has on the value of the resulting restraints. We show that MELD can accurately account for challenges related to conformational heterogeneity and noise and achieve moderate side chain packing. These results also highlight the capabilities of integrative approaches when experimental information is limited.

DATA AVAILABILITY STATEMENT

Restraint files, run scripts, and analysis scripts can be found on our github repository (https://github.com/maccallumlab/calmodulin_pre_paper) and are archived with Zenodo (DOI: 10.5281/zenodo.5071079).

AUTHOR CONTRIBUTIONS

KG, ZL, HV, and JM contributed to the design of the study, analysis of the data, and wrote the manuscript. ZL performed the NMR experiments. KG performed the MELD calculations.

FUNDING

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada, Canada Foundation for Innovation, and Compute Canada. JM is a Canada Research Chair.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.676268/full#supplementary-material

REFERENCES

Andralojc, W., and Ravera, E. (2018). "Chapter 4. Treating Biomacromolecular Conformational Variability," in Paramagnetism in Experimental Biomolecular NMR, 107–133. doi:10.1039/9781788013291-00107

Andrałojć, W., Berlin, K., Fushman, D., Luchinat, C., Parigi, G., Ravera, E., et al. (2015). Information Content of Long-Range NMR Data for the Characterization of Conformational Heterogeneity. *J. Biomol. NMR* 62 (3), 353–371. doi:10.1007/s10858-015-9951-6

Andrałojć, W., Hiruma, Y., Liu, W.-M., Ravera, E., Nojiri, M., Parigi, G., et al. (2017). Identification of Productive and Futile Encounters in an Electron Transfer Protein Complex. *Proc. Natl. Acad. Sci. USA* 114 (10), E1840–E1847. doi:10.1073/pnas.1616813114

Andrałojć, W., Luchinat, C., Parigi, G., and Ravera, E. (2014). Exploring Regions of Conformational Space Occupied by Two-Domain Proteins. *J. Phys. Chem. B* 118 (36), 10576–10587. doi:10.1021/jp504820w

Anthis, N. J., Doucleff, M., and Clore, G. M. (2011). Transient, Sparsely Populated Compact States of Apo and Calcium-Loaded Calmodulin Probed by Paramagnetic Relaxation Enhancement: Interplay of Conformational Selection and Induced Fit. *J. Am. Chem. Soc.* 133 (46), 18966–18974. doi:10.1021/ja2082813

Asciutto, E. K., Dang, M., Pochapsky, S. S., Madura, J. D., and Pochapsky, T. C. (2011). Experimentally Restrained Molecular Dynamics Simulations for Characterizing the Open States of Cytochrome P450cam. *Biochemistry* 50 (10), 1664–1671. doi:10.1021/bi101820d

Banci, L., Bertini, I., Bren, K. L., Cremonini, M. A., Gray, H. B., Luchinat, C., et al. (1996). The Use of Pseudocontact Shifts to Refine Solution Structures of Paramagnetic Metalloproteins: Met80Ala Cyano-Cytochrome C as an Example. *J. Biol. Inorg. Chem.* 1 (2), 117–126. doi:10.1007/s007750050030

Banci, L., Bertini, I., Huber, J. G., Luchinat, C., and Rosato, A. (1998). Partial Orientation of Oxidized and Reduced Cytochromeb5at High Magnetic Fields: Magnetic Susceptibility Anisotropy Contributions and Consequences for Protein Solution Structure Determination. *J. Am. Chem. Soc.* 120 (49), 12903–12909. doi:10.1021/ja981791w

Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W., and Bax, A. (1992). Backbone Dynamics of Calmodulin Studied by Nitrogen-15 Relaxation Using Inverse Detected Two-Dimensional NMR Spectroscopy: The Central Helix Is Flexible. *Biochemistry* 31 (23), 5269–5278. doi:10.1021/bi00138a005

Barbieri, R., Bertini, I., Cavallaro, G., Lee, Y.-M., Luchinat, C., and Rosato, A. (2002). Paramagnetically Induced Residual Dipolar Couplings for Solution Structure Determination of Lanthanide Binding Proteins. *J. Am. Chem. Soc.* 124 (19), 5581–5587. doi:10.1021/ja025528d

Battiste, J. L., and Wagner, G. (2000). Utilization of Site-Directed Spin Labeling and High-Resolution Heteronuclear Nuclear Magnetic Resonance for Global Fold Determination of Large Proteins with Limited Nuclear Overhauser Effect Data†. *Biochemistry* 39 (18), 5355–5365. doi:10.1021/bi000060h

Bellomo, G., Ravera, E., Calderone, V., Botta, M., Fragai, M., Parigi, G., et al. (2021). Revisiting Paramagnetic Relaxation Enhancements in Slowly Rotating Systems: How Long Is the Long Range? *Magn. Reson.* 2 (1), 25–31. doi:10.5194/mr-2-25-2021

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1), 235–242. doi:10.1093/nar/28.1.235

Bertini, I., Kursula, P., Luchinat, C., Parigi, G., Vahokoski, J., Wilmanns, M., et al. (2009). Accurate Solution Structures of Proteins from X-ray Data and a Minimal Set of NMR Data: Calmodulin–Peptide Complexes as Examples. *J. Am. Chem. Soc.* 131 (14), 5134–5144. doi:10.1021/ja8080764

Bertini, I., Luchinat, C., Nagulapalli, M., Parigi, G., and Ravera, E. (2012). Paramagnetic Relaxation Enhancement for the Characterization of the Conformational Heterogeneity in Two-Domain Proteins. *Phys. Chem. Chem. Phys.* 14 (25), 9149–9156. doi:10.1039/C2CP40139H

Bertini, I., Luchinat, C., Parigi, G., and Pierattelli, R. (2005). NMR Spectroscopy of Paramagnetic Metalloproteins. *ChemBioChem* 6 (9), 1536–1549. doi:10.1002/cbic.200500124

Bonomi, M., Camilloni, C., and Vendruscolo, M. (2016). Metadynamic Metainference: Enhanced Sampling of the Metainference Ensemble Using Metadynamics. *Sci. Rep.* 6, 31232. doi:10.1038/srep31232

Bonomi, M., Heller, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of Protein Structural Ensemble Determination. *Curr. Opin. Struct. Biol.* 42, 106–116. doi:10.1016/j.sbi.2016.12.004

Boomsma, W., Ferkinghoff-Borg, J., and Lindorff-Larsen, K. (2014). Combining Experiments and Simulations Using the Maximum Entropy Principle. *PLOS Comput. Biol.* 10 (2), e1003406. doi:10.1371/journal.pcbi.1003406

Bouvignies, G., Markwick, P., Brüschweiler, R., and Blackledge, M. (2006). Simultaneous Determination of Protein Backbone Structure and Dynamics from Residual Dipolar Couplings. *J. Am. Chem. Soc.* 128 (47), 15100–15101. doi:10.1021/ja066704b

Camilloni, C., and Vendruscolo, M. (2015). A Tensor-free Method for the Structural and Dynamical Refinement of Proteins Using Residual Dipolar Couplings. *J. Phys. Chem. B* 119 (3), 653–661. doi:10.1021/jp5021824

Carlon, A., Ravera, E., Parigi, G., Murshudov, G. N., and Luchinat, C. (2019). Joint X-Ray/NMR Structure Refinement of Multidomain/Multisubunit Systems. *J. Biomol. NMR* 73 (6), 265–278. doi:10.1007/s10858-018-0212-3

Case, D. A., Aktulga, H. M., Belfon, K., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., et al. (2021). *Amber 2021*. San Francisco: University of California.

Chou, J. J., Li, S., and Bax, A. (2000). Study of Conformational Rearrangement and Refinement of Structural Homology Models by the Use of Heteronuclear Dipolar Couplings. *J. Biomol. NMR* 18 (3), 217–227. doi:10.1023/A:1026563923774

Clore, G. M., and Iwahara, J. (2009). Theory, Practice, and Applications of Paramagnetic Relaxation Enhancement for the Characterization of Transient Low-Population States of Biological Macromolecules and Their Complexes. *Chem. Rev.* 109 (9), 4108–4139. doi:10.1021/cr900033p

Clore, G. M., Szabo, A., Bax, A., Kay, L. E., Driscoll, P. C., and Gronenborn, A. M. (1990). Deviations from the Simple Two-Parameter Model-free Approach to the Interpretation of Nitrogen-15 Nuclear Magnetic Relaxation of Proteins. *J. Am. Chem. Soc.* 112 (12), 4989–4991. doi:10.1021/ja00168a070

Cornilescu, G., Marquardt, J. L., Ottiger, M., and Bax, A. (1998). Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* 120 (27), 6836–6837. doi:10.1021/ja9812610

Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M., and Dobson, C. M. (2005). Mapping Long-Range Interactions in α-Synuclein Using Spin-Label NMR and Ensemble Molecular Dynamics Simulations. *J. Am. Chem. Soc.* 127 (2), 476–477. doi:10.1021/ja044834j

Delaglio, F., Grzesiek, S., Vuister, G., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: A Multidimensional Spectral Processing System Based on UNIX Pipes. *J. Biomol. NMR* 6 (3), 277–293. doi:10.1007/BF00197809

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., et al. (2017). OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLOS Comput. Biol.* 13 (7), e1005659. doi:10.1371/journal.pcbi.1005659

Fawzi, N. L., Fleissner, M. R., Anthis, N. J., Kálai, T., Hideg, K., Hubbell, W. L., et al. (2011). A Rigid Disulfide-Linked Nitroxide Side Chain Simplifies the Quantitative Analysis of PRE Data. *J. Biomol. NMR* 51 (1), 105–114. doi:10.1007/s10858-011-9545-x

Gaalswyk, K., Muniyat, M. I., and MacCallum, J. L. (2018). The Emerging Role of Physical Modeling in the Future of Structure Determination. *Curr. Opin. Struct. Biol.* 49, 145–153. doi:10.1016/j.sbi.2018.03.005

Gardner, K. H., and Kay, L. E. (1998). The Use of2H,13C,15N Multidimensional Nmr Gto Study the Structure and Dynamics of Proteins. *Annu. Rev. Biophys. Biomol. Struct.* 27, 357–406. doi:10.1146/annurev.biophys.27.1.357

Gelis, I., Bonvin, A. M. J. J., Keramisanou, D., Koukaki, M., Gouridis, G., Karamanou, S., et al. (2007). Structural Basis for Signal-Sequence Recognition by the Translocase Motor SecA as Determined by NMR. *Cell* 131 (4), 756–769. doi:10.1016/j.cell.2007.09.039

Gifford, J. L., Ishida, H., and Vogel, H. J. (2011). Fast Methionine-Based Solution Structure Determination of Calcium-Calmodulin Complexes. *J. Biomol. NMR* 50 (1), 71–81. doi:10.1007/s10858-011-9495-3

Gochin, M., Zhou, G., and Phillips, A. H. (2011). Paramagnetic Relaxation Assisted Docking of a Small Indole Compound in the HIV-1 Gp41 Hydrophobic Pocket. *ACS Chem. Biol.* 6 (3), 267–274. doi:10.1021/cb100368d

Goto, N. K., and Kay, L. E. (2000). New Developments in Isotope Labeling Strategies for Protein Solution NMR Spectroscopy. *Curr. Opin. Struct. Biol.* 10 (5), 585–592. doi:10.1016/S0959-440X(00)00135-4

Gottstein, D., Reckel, S., Dötsch, V., and Güntert, P. (2012). Requirements on Paramagnetic Relaxation Enhancement Data for Membrane Protein Structure Determination by NMR. *Structure* 20 (6), 1019–1027. doi:10.1016/j.str.2012.03.010

Guerry, P., Salmon, L., Mollica, L., Ortega Roldan, J. L., Markwick, P., van Nuland, N. A. J., et al. (2013). Mapping the Population of Protein Conformational Energy Sub-States from NMR Dipolar Couplings. *Angew. Chem. Int. Ed.* 52 (11), 3181–3185. doi:10.1002/anie.201209669

Guo, Z., Cascio, D., Hideg, K., and Hubbell, W. L. (2008). Structural Determinants of Nitroxide Motion in Spin-Labeled Proteins: Solvent-Exposed Sites in Helix B of T4 Lysozyme. *Protein Sci.* 17 (2), 228–239. doi:10.1110/ps.073174008

Habeck, M., Nilges, M., and Rieping, W. (2008). A Unifying Probabilistic Framework for Analyzing Residual Dipolar Couplings. *J. Biomol. NMR* 40 (2), 135–144. doi:10.1007/s10858-007-9215-1

Higman, V. A., Boyd, J., Smith, L. J., and Redfield, C. (2011). Residual Dipolar Couplings: Are Multiple Independent Alignments Always Possible? *J. Biomol. Nmr* 49 (1), 53–60. doi:10.1007/s10858-010-9457-1

Holm, L., and Sander, C. (1993). Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* 233 (1), 123–138. doi:10.1006/jmbi.1993.1489

Huang, H., and Vogel, H. J. (2012). Structural Basis for the Activation of Platelet Integrin αIIbβ3 by Calcium- and Integrin-Binding Protein 1. *J. Am. Chem. Soc.* 134 (8), 3864–3872. doi:10.1021/ja2111306

Hummer, G., and Köfinger, J. (2015). Bayesian Ensemble Refinement by Replica Simulations and Reweighting. *J. Chem. Phys.* 143 (24), 243150. doi:10.1063/1.4937786

Hwang, P. M., Pan, J. S., and Sykes, B. D. (2014). Targeted Expression, Purification, and Cleavage of Fusion Proteins from Inclusion Bodies inEscherichia Coli. *FEBS Lett.* 588 (2), 247–252. doi:10.1016/j.febslet.2013.09.028

Ikura, M., Clore, G., Gronenborn, A., Zhu, G., Klee, C., and Bax, A. (1992). Solution Structure of a Calmodulin-Target Peptide Complex by Multidimensional NMR. *Science* 256 (5057), 632–638. doi:10.1126/science.1585175

Ishida, H., Nguyen, L. T., Gopal, R., Aizawa, T., and Vogel, H. J. (2016). Overexpression of Antimicrobial, Anticancer, and Transmembrane Peptides inEscherichia Colithrough a Calmodulin-Peptide Fusion System. *J. Am. Chem. Soc.* 138 (35), 11318–11326. doi:10.1021/jacs.6b06781

Ishida, H., and Vogel, H. J. (2010). The Solution Structure of a Plant Calmodulin and the CaM-Binding Domain of the Vacuolar Calcium-ATPase BCA1 Reveals a New Binding and Activation Mechanism. *J. Biol. Chem.* 285 (49), 38502–38510. doi:10.1074/jbc.M110.131201

Islam, S. M., and Roux, B. (2015). Simulating the Distance Distribution between Spin-Labels Attached to Proteins. *J. Phys. Chem. B* 119 (10), 3901–3911. doi:10.1021/jp510745d

Islam, S. M., Stein, R. A., Mchaourab, H. S., and Roux, B. (2013). Structural Refinement from Restrained-Ensemble Simulations Based on EPR/DEER Data: Application to T4 Lysozyme. *J. Phys. Chem. B* 117 (17), 4740–4754. doi:10.1021/jp311723a

Iwahara, J., and Clore, G. M. (2006). Detecting Transient Intermediates in Macromolecular Binding by Paramagnetic NMR. *Nature* 440 (7088), 1227–1230. doi:10.1038/nature04673

Iwahara, J., Schwieters, C. D., and Clore, G. M. (2004). Ensemble Approach for NMR Structure Refinement against1H Paramagnetic Relaxation Enhancement Data Arising from a Flexible Paramagnetic Group Attached to a Macromolecule. *J. Am. Chem. Soc.* 126 (18), 5879–5896. doi:10.1021/ja031580d

Iwahara, J., Tang, C., and Marius Clore, G. (2007). Practical Aspects of 1H Transverse Paramagnetic Relaxation Enhancement Measurements on Macromolecules. *J. Magn. Reson.* 184 (2), 185–195. doi:10.1016/j.jmr.2006.10.003

Jaroniec, C. P., Kaufman, J. D., Stahl, S. J., Viard, M., Blumenthal, R., Wingfield, P. T., et al. (2005). Structure and Dynamics of Micelle-Associated Human Immunodeficiency Virus Gp41 Fusion Domain†,‡. *Biochemistry* 44 (49), 16167–16180. doi:10.1021/bi051672a

Johnson, B. A., and Blevins, R. A. (1994). NMR View: A Computer Program for the Visualization and Analysis of NMR Data. *J. Biomol. NMR* 4 (5), 603–614. doi:10.1007/BF00404272

Kainosho, M., and Güntert, P. (2009). SAIL - Stereo-Array Isotope Labeling. *Quart. Rev. Biophys.* 42 (4), 247–300. doi:10.1017/S0033583510000016

Kay, L. E. (2016). New Views of Functionally Dynamic Proteins by Solution NMR Spectroscopy. *J. Mol. Biol.* 428 (2, Part A), 323–331. doi:10.1016/j.jmb.2015.11.028

Keizers, P. H. J., and Ubbink, M. (2011). Paramagnetic Tagging for Protein Structure and Dynamics Analysis. *Prog. Nucl. Magn. Reson. Spectrosc.* 58 (1), 88–96. doi:10.1016/j.pnmrs.2010.08.001

Kim, D. E., DiMaio, F., Yu-Ruei Wang, R., Song, Y., and Baker, D. (2014). One Contact for Every Twelve Residues Allows Robust and Accurate Topology-Level Protein Structure Modeling. *Proteins* 82 (0 2), 208–218. doi:10.1002/prot.24374

Koehler, J., and Meiler, J. (2011). Expanding the Utility of NMR Restraints with Paramagnetic Compounds: Background and Practical Aspects. *Prog. Nucl. Magn. Reson. Spectrosc.* 59 (4), 360–389. doi:10.1016/j.pnmrs.2011.05.001

Kuenze, G., Bonneau, R., Leman, J. K., and Meiler, J. (2019). Integrative Protein Modeling in RosettaNMR from Sparse Paramagnetic Restraints. *Structure* 27 (11), 1721–1734.e5. doi:10.1016/j.str.2019.08.012

Lange, O. F., Rossi, P., Sgourakis, N. G., Song, Y., Lee, H.-W., Aramini, J. M., et al. (2012). Determination of Solution Structures of Proteins up to 40 KDa Using CS-Rosetta with Sparse NMR Data from Deuterated Samples. *Proc. Natl. Acad. Sci.* 109 (27), 10873–10878. doi:10.1073/pnas.1203013109

Lee, A. L., Sharp, K. A., Kranz, J. K., Song, X.-J., and Wand, A. J. (2002). Temperature Dependence of the Internal Dynamics of a Calmodulin–Peptide Complex. *Biochemistry* 41 (46), 13814–13825. doi:10.1021/bi026380d

Lipsitz, R. S., and Tjandra, N. (2004). Residual Dipolar Couplings in NMR Structure Analysis. *Annu. Rev. Biophys. Biomol. Struct.* 33, 387–413. doi:10.1146/annurev.biophys.33.110502.140306

Liu, Z., and Vogel, H. J. (2012). Structural Basis for the Regulation of L-type Voltage-Gated Calcium Channels: Interactions between the N-Terminal Cytoplasmic Domain and Ca2+-Calmodulin. *Front. Mol. Neurosci.* 5. doi:10.3389/fnmol.2012.00038

Losonczi, J. A., Andrec, M., Fischer, M. W. F., and Prestegard, J. H. (1999). Order Matrix Analysis of Residual Dipolar Couplings Using Singular Value Decomposition. *J. Magn. Reson.* 138 (2), 334–342. doi:10.1006/jmre.1999.1754

MacCallum, J. L., Perez, A., and Dill, K. A. (2015). Determining Protein Structures by Combining Semireliable Data with Atomistic Physical Models by Bayesian Inference. *Proc. Natl. Acad. Sci. USA* 112 (22), 6985–6990. doi:10.1073/pnas.1506788112

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theor. Comput.* 11 (8), 3696–3713. doi:10.1021/acs.jctc.5b00255

Mal, T. K., Skrynnikov, N. R., Yap, K. L., Kay, L. E., and Ikura, M. (2002). Detecting Protein Kinase Recognition Modes of Calmodulin by Residual Dipolar Couplings in Solution NMR. *Biochemistry* 41 (43), 12899–12906. doi:10.1021/bi0264162

Meador, W., Means, A., and Quiocho, F. (1992). Target Enzyme Recognition by Calmodulin: 2.4 A Structure of a Calmodulin-Peptide Complex. *Science* 257 (5074), 1251–1255. doi:10.1126/science.1519061

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 (2), 383–394. doi:10.1002/prot.20033

Otten, R., Chu, B., Krewulak, K. D., Vogel, H. J., and Mulder, F. A. A. (2010). Comprehensive and Cost-Effective NMR Spectroscopy of Methyl Groups in Large Proteins. *J. Am. Chem. Soc.* 132 (9), 2952–2960. doi:10.1021/ja907706a

Ottiger, M., Delaglio, F., and Bax, A. (1998). Measurement ofJand Dipolar Couplings from Simplified Two-Dimensional NMR Spectra. *J. Magn. Reson.* 131 (2), 373–378. doi:10.1006/jmre.1998.1361

PDB Statistics (2021). PDB Statistics: Growth of Structures from NMR Experiments Released Per Year. Available at: https://www.rcsb.org/stats/growth/growth-nmr (Accessed February 28, 2021).

Perez, A., Gaalswyk, K., Jaroniec, C. P., and MacCallum, J. L. (2019). High Accuracy Protein Structures from Minimal Sparse Paramagnetic Solid-State NMR Restraints. *Angew. Chem. Int. Ed.* 58 (20), 6564–6568. doi:10.1002/anie.201811895

Perez, A., MacCallum, J. L., Brini, E., Simmerling, C., and Dill, K. A. (2015). Grid-Based Backbone Correction to the ff12SB Protein Force Field for Implicit-Solvent Simulations. *J. Chem. Theor. Comput.* 11 (10), 4770–4779. doi:10.1021/acs.jctc.5b00662

Perez, A., MacCallum, J. L., and Dill, K. A. (2015). Accelerating Molecular Simulations of Proteins Using Bayesian Inference on Weak Information.

Proc. Natl. Acad. Sci. USA 112 (38), 11846–11851. doi:10.1073/pnas.1515561112

Pervushin, K., Riek, R., Wider, G., and Wüthrich, K. (1997). Attenuated T2 Relaxation by Mutual Cancellation of Dipole-Dipole Coupling and Chemical Shift Anisotropy Indicates an Avenue to NMR Structures of Very Large Biological Macromolecules in Solution. *Proc. Natl. Acad. Sci.* 94 (23), 12366–12371. doi:10.1073/pnas.94.23.12366

Pilla, K. B., Gaalswyk, K., and MacCallum, J. L. (2017). Molecular Modeling of Biomolecules by Paramagnetic NMR and Computational Hybrid Methods. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1865 (11, Part B), 1654–1663. doi:10.1016/j.bbapap.2017.06.016

Prestegard, J. H., Agard, D. A., Moremen, K. W., Lavery, L. A., Morris, L. C., and Pederson, K. (2014). Sparse Labeling of Proteins: Structural Characterization from Long Range Constraints. *J. Magn. Reson.* 241, 32–40. doi:10.1016/j.jmr.2013.12.012

Prestegard, J. H., Bougault, C. M., and Kishore, A. I. (2004). Residual Dipolar Couplings in Structure Determination of Biomolecules. *Chem. Rev.* 104 (8), 3519–3540. doi:10.1021/cr030419i

Raman, S., Lange, O. F., Rossi, P., Tyka, M., Wang, X., Aramini, J., et al. (2010). NMR Structure Determination for Larger Proteins Using Backbone-Only Data. *Science* 327 (5968), 1014–1018. doi:10.1126/science.1183649

Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential Structure Determination. *Science* 309 (5732), 303–306. doi:10.1126/science.1110428

Ruschak, A. M., and Kay, L. E. (2010). Methyl Groups as Probes of Supramolecular Structure, Dynamics and Function. *J. Biomol. NMR* 46 (1), 75–87. doi:10.1007/s10858-009-9376-1

Ryabov, Y. E., and Fushman, D. (2007). A Model of Interdomain Mobility in a Multidomain Protein. *J. Am. Chem. Soc.* 129 (11), 3315–3327. doi:10.1021/ja067667r

Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009). TALOS+: A Hybrid Method for Predicting Protein Backbone Torsion Angles from NMR Chemical Shifts. *J. Biomol. NMR* 44 (4), 213–223. doi:10.1007/s10858-009-9333-z

Sikic, K., Tomic, S., and Carugo, O. (2010). Systematic Comparison of Crystal and NMR Protein Structures Deposited in the Protein Data Bank. *Open Biochem. J.* 4, 83–95. doi:10.2174/1874091X01004010083

Sullivan, D. C., Aynechi, T., Voelz, V. A., and Kuntz, I. D. (2003). Information Content of Molecular Structures. *Biophysical J.* 85 (1), 174–190. doi:10.1016/S0006-3495(03)74464-6

Sullivan, D. C., and Kuntz, I. D. (2004). Distributions in Protein Conformation Space: Implications for Structure Prediction and Entropy. *Biophysical J.* 87 (1), 113–120. doi:10.1529/biophysj.104.041723

Tolman, J. R. (2002). A Novel Approach to the Retrieval of Structural and Dynamic Information from Residual Dipolar Couplings Using Several Oriented Media in Biomolecular NMR Spectroscopy. *J. Am. Chem. Soc.* 124 (40), 12020–12030. doi:10.1021/ja0261123

Tugarinov, V., Hwang, P. M., and Kay, L. E. (2004). Nuclear Magnetic Resonance Spectroscopy of High-Molecular-Weight Proteins. *Annu. Rev. Biochem.* 73, 107–146. doi:10.1146/annurev.biochem.73.011303.074004

Tugarinov, V., Kanelis, V., and Kay, L. E. (2006). Isotope Labeling Strategies for the Study of High-Molecular-Weight Proteins by Solution NMR Spectroscopy. *Nat. Protoc.* 1 (2), 749–754. doi:10.1038/nprot.2006.101

Vallat, B., Webb, B., Westbrook, J. D., Sali, A., and Berman, H. M. (2018). Development of a Prototype System for Archiving Integrative/Hybrid Structure Models of Biological Macromolecules. *Structure* 26 (6), 894–904.e2. doi:10.1016/j.str.2018.03.011

van Dijk, A. D. J., Fushman, D., and Bonvin, A. M. J. J. (2005). Various Strategies of Using Residual Dipolar Couplings in NMR-Driven Protein Docking: Application to Lys48-Linked Di-ubiquitin and Validation against 15N-Relaxation Data. *Proteins* 60 (3), 367–381. doi:10.1002/prot.20476

Vlasie, M. D., Comuzzi, C., van den Nieuwendijk, A. M. C. H., Prudêncio, M., Overhand, M., and Ubbink, M. (2007). Long-Range-Distance NMR Effects in a Protein Labeled with a Lanthanide-DOTA Chelate. *Chem. Eur. J.* 13 (6), 1715–1723. doi:10.1002/chem.200600916

Vlasie, M. D., Fernández-Busnadiego, R., Prudêncio, M., and Ubbink, M. (2008). Conformation of Pseudoazurin in the 152 KDa Electron Transfer Complex with Nitrite Reductase Determined by Paramagnetic NMR. *J. Mol. Biol.* 375 (5), 1405–1415. doi:10.1016/j.jmb.2007.11.056

Ward, A. B., Sali, A., and Wilson, I. A. (2013). Integrative Structural Biology. *Science* 339 (6122), 913–915. doi:10.1126/science.1228565

Yao, L., Ying, J., and Bax, A. (2009). Improved Accuracy of 15N-1H Scalar and Residual Dipolar Couplings from Gradient-Enhanced IPAP-HSQC Experiments on Protonated Proteins. *J. Biomol. NMR* 43 (3), 161–170. doi:10.1007/s10858-009-9299-x

Yuan, T., Gomes, A. V., Barnes, J. A., Hunter, H. N., and Vogel, H. J. (2004). Spectroscopic Characterization of the Calmodulin-Binding and Autoinhibitory Domains of Calcium/Calmodulin-dependent Protein Kinase I. *Arch. Biochem. Biophys.* 421 (2), 192–206. doi:10.1016/j.abb.2003.11.012

Zhang, M., Yuan, T., Aramini, J. M., and Vogel, H. J. (1995). Interaction of Calmodulin with its Binding Domain of Rat Cerebellar Nitric Oxide Synthase. *J. Biol. Chem.* 270 (36), 20901–20907. doi:10.1074/jbc.270.36.20901

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.

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