COMPUTATIONAL MODELING, PREDICTION AND DESIGN OF PROTEIN-PROTEIN INTERACTIONS

by

Ameya Harmalkar

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

July, 2023

Abstract

Protein-protein interactions (PPIs) govern nearly all biological processes in human health and diseases, ranging from enzyme catalysis and inhibition, to signaling and gene regulation. Understanding the dynamics of protein interactions and the structure of protein complexes at an atomic level is key in delineating disease mechanisms, such as Huntington's, Alzheimer's, and cancer, and developing intervention strategies. Investigation of these structural complexes by experimental techniques is often expensive, laborious, and limited. Computational modeling provides an alternative route to elucidate structures and guide molecular engineering based on PPIs. A long-standing challenge limiting the accuracy of computational methods is the ability to predict binding-induced conformational changes during protein-protein association.

In my thesis, I address this challenge by creating new tools to predict atomistic models of flexible protein complexes. First, I develop a protein docking protocol that incorporates temperature replica exchange Monte Carlo (T-REMC) and backbone flexibility to mimic induced-fit approach of protein interactions. On a benchmark of 88 protein complexes with varying degrees of flexibility, this protocol, ReplicaDock 2.0, is the first method to successfully dock 62% of complexes with conformational changes up to 2.2 Å. Building on this success of ReplicaDock2.0, I extend it to develop a novel sampling approach, namely resolution exchange. In this approach, exchanges

are performed between the full-atom and the centroid configurations to improve backbone sampling and escape entrapment in non-native minima. Finally, I conclude my docking methods development work by creating a pipeline that fuses AlphaFold (a deep-learning tool for protein sequence-to-structure prediction) with aforementioned docking techniques to develop a method for improved complex structure prediction.

In conjunction with development of foundational protein structure prediction tools, I equip docking tools to make contributions to human health and disease. First, to extend the functionality of MC approaches for capturing dynamics, I model the interactions between an outer membrane nutrient transporter (on bacterial cells) and a bacteriocin (Colicin B), and deduce the translocation pathway for Colicin B through the transporter. Next, I apply my knowledge of PPIs to create novel complex designs. I demonstrate a computational approach to create orthogonal interfaces with experimental validation for the PDGF signaling system. This technology has tremendous potential in regenerative medicine and therapeutic discovery as an orthogonal signaling system eliminates off-target risks (e.g., cancer) and promotes exclusivity.

In sum, my work has advanced our understanding and our ability to model and design flexible protein-protein interactions.

Thesis Committee

Jeffrey J. Gray (Primary Advisor)
Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins Whiting School of Engineering

Paulette Clancy (Reader)
Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins Whiting School of Engineering

Jamie B. Spangler
Assistant Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins Whiting School of Engineering

Margaret E. Johnson (Reader)
Associate Professor
T.C. Jenkins Department of Biophysics
Johns Hopkins Krieger School of Arts and Sciences

Albert Y. Lau
Associate Professor
Department of Biophysics and Biophysical Chemistry
Johns Hopkins School of Medicine

Alternate Readers

Thi Vo
Assistant Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins Whiting School of Engineering

Stephen Fried
Assistant Professor
Department of Chemistry
Johns Hopkins Krieger School of Arts and Sciences

Acknowledgments

I am deeply grateful for the many people who have enriched my life and supported me throughout my time in graduate school. I have had an incredible time this past few years moving to a new country (ideally two if I count my small detour in Germany), making new (and hopefully long-lasting) friendships, exchanging culture, and learning, not just as a scientist but as a person.

First and foremost, to my advisor, Jeff Gray, thank you so much for making this possible. I still recall the emails I sent you as an undergraduate senior, figuring out whether I want to come to Hopkins for my PhD. Your infectious optimism, empathy, and kindness came through with that little exchange of ours, and I consider myself lucky to be a part of your group. You made science an adventure, your enthusiasm boosted me through downfalls and rejections, and I am in awe of your ability to connect ideas and people. One of the best attributes of the Gray lab, and I think it trickles down from Jeff himself, is an atmosphere of belonging and feeling of worthiness. Criticism is always constructive, and achievements are celebrated together. Thank you so much for mentoring me, Jeff. You have helped me grow as a scientist and I can undoubtedly say that I am a better person today than I was almost five years ago.

Another reason what makes the Gray lab an amazing work place are the amazing

people that constitute the Gray lab, past and present. I can hardly sum up these wonderful people in a sentence but I will try. My tenure in the Gray lab started working with Pooja on replica exchange and protein docking, and that eventually became a major chunk of my thesis. Pooja has been a wonderful mentor, a great co-worker, and the best office mate (Go MD303!) one could ask for. Shourya and Jeliazko, who graduated within an year since I joined were my go-to folks for Rosetta questions and I have bothered them even after they left. Jeliazko, especially was the other MD 303 office mate, who made working fun with his hot-takes. Morgan has been a constant support throughout, and she set an example for rigorous and reproducible benchmarking. Additionally, she has been a wonderful friend and I have thoroughly enjoyed our conference outings. Jeff Ruffolo and I had little scientific overlap, however, I have always been inspired by his creativity and the wave of deep learning that he brought along when he joined. Rahel, Jing, and Sudhanshu (Sid) were incredible postdocs and even better friends. I have admired Rahel and the vast experimental+computational knowledge that she brought to our group. More than that, she has been a great friend to discuss life decisions with, and I am glad to have her as a friend. Hanging out with her, Malte, and Max, made my stay in Germany during DAAD a pleasant and heart-warming experience. Sid and I unfortunately had no scientific overlap, however, owing to our journey in the Gray lab starting almost a month apart, we went through the Rosetta struggles together and figured out life in Baltimore.

As the pandemic turned the world upside down, so did the composition of Gray lab, and I am happy to have these wonderful, new folks that I call lab mates. Rituparna has been a wonderful friend and colleague, and I have been lucky to

have worked with her, not just scientifically but also for our DEI objectives. She is unwavering and steadfast and is a patient mentor, all the qualities I hope to have. Lee-Shin and Sam are both incredibly smart and hard-working, and more importantly, excellent athletes. Lee-Shin is a volleyball champ and have aptly replaced Jeliazko's seat in MD303. Mikey and Fatima are the youngest of the lot and their energy and enthusiasm is inspiring and addictive. I am excited for all the cool work they would be doing. Fatima, Lee-Shin, and I had a coffee tradition on-going since I returned from Munich, and I will treasure those small walks with cultural debates and an exchange of unpopular opinions. I also had the opportunity to mentor three wonderful undergraduates: Priyamvada Prathima, Ranjani Ramasubramaniam, and Brandon Ameglio. I am obliged to have mentored them and they have inspired me to be a better mentor and an even better scientist every single day. Apart from these folks, I was blessed to have interacted with Wenhao Gao, Rebecca Alford, Brittany Lasher, Denis Akpinaroglu, Sen Wei, Isobel Garett, Richard Shaui, Jason Labonte, Jacky Chen, and Ivan Riveira.

Next, I want to thank my thesis committee for providing critical feedback on my research. Thank you Dr. Paulette Clancy, Dr. Jamie Spangler, Dr. Margaret Johnson, Dr. Albert Lau, Dr. Thi Vo, and Dr. Stephen Fried for serving on my thesis committee. I would like to extend special thanks to Dr. Clancy and Dr. Kevrekidis, who have always asked me fundamental questions in our annual review meets, not just testing my knowledge but also shaping research direction. Prof. Clancy also provided me invaluable feedback as I was considering future options and I am indebted to her for her advise. Dr. Spangler has been a collaborator on an exciting protein design project and our conversations have made me better at conveying data to experimentalists. I

also want to convey huge thanks to Prof. Martin Zacharias, who hosted me in Munich, and with whom I developed the foundations of resolution exchange. Your positive outlook and scientific curiosity was refreshing and I am indebted by your hospitality for all the time I was in Munich. I hope to collaborate on more projects in the future. I also want to go convey my deepest thanks to the JHU ChemBE staff: Beth Rannie, Kourtney Roussey, Sharon Punte, Alisha Wells, and Marcellas Preston.

And then there is the group that enriched my being: *my friends*. Ankita, Saumil, and Shashank, you have known me since undergrad, and have willingly tolerated me all these years. I admire your tolerance, your loyalty, and your unwavering support towards me. Thank you for encouraging me, for helping me celebrating my success, and for asking the right questions whenever I required a honest feedback. To Harnish, Pranay, and Nikita: thanks for being great friends these past four years. The banter, the hiking trips, and all the good times. I am grateful for your company and for making Baltimore feel like home. I can't imagine a better circle to have spent my time in Baltimore with. Thank you Revathi Reddy, Saina Prabhu, Pujita Pathak, Haonan Xu, Rajiv Nair, Nicolas Evangelou, Mukund Goyal, Gayatri Dhara, Swetha Kumar, Aprameya Prasad, Adrian Johnston, Bargeen Turzo, Isaiah Chen, and all my classmates for making this journey pleasant and worth accomplishing. Luis Vollmers, Brianda Santini, Christian Sustay, Shu-Yu Chen, Simone Goppert, Patrick Quoika, thanks so much for making my time in TU München so wonderful. Although it was a short trip of 4 months, working with you was a pleasure. Thanks for making me feel a part of the group every single day. I will forever remember the game night, the barbeques, the trip to Neuschwanstein, picnics in the Englischer Garten, and the amazing farewell that you folks hosted for me. Whoever says Germans can't party, I

am sure they haven't met you all!

Lastly, I am extremely grateful to have an amazing family. Thank you for years of support and nurturing and for always encouraging me to pursue my dreams. Thank you to my late paternal grandparents, *Baba* and *Aaji*, and my late aunt *Shraddha Ragaji*, who dreamt of seeing me be a doctor (a medical doctor to be precise, but I am positive they would be happy with the PhD too). My maternal grandparents who have been my biggest cheerleaders, especially my grandmother. Thank you to my Mum and all my aunts. You have been the strong women in my life since childhood, and I am thankful for making me the person I am today. To Mum, you have always asked me to be the best version of myself, from sitting besides me to do math problems together for olympaids, to getting up early morning to make me tiffins for a busy day in college, I have seen how you have shaped your life around mine and I am grateful for this unconditional love and support. Thank you mom and dad for making me independent, self-reliant, and for sowing the seeds of hard-work in me. This thesis is for you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Protein-protein interactions govern biological functions

Proteins are ubiquitous in most, if not all, biological mechanisms. Composed of linear chains of amino acid sequences that fold into compact three-dimensional structures, proteins encode the machinery of life by intimately linking sequence and structure to biological function. The structure of a protein, mediated by its environment and by non-covalent interactions between the chemically diverse amino acid side-chains, determines its folds and in-turn affects its functionality. To annotate the functional role of proteins and tune the protein interaction networks for specific engineering tasks, learning the nuances of protein-protein interactions (PPIs) and predicting protein structures is paramount.

For the second half of twentieth century, much of the contemporary biology and chemistry was concerned with the structural characteristics of these unbranched

biopolymers, we now know as proteins. In 1951, Linus Pauling deduced the α-helix[1] and the β-sheet[2], sparking a seminal discovery in molecular biology. This was followed by the discovery of myoglobin[3], the first protein structure to be crystallized in 1957, to the creation of the Protein Data Bank (PDB) almost two decades later. The advances in experimental methods for structure determination have exploded in the past few decades with up to 203,607 experimentally-determined 3D structures deposited in the PDB till date (*as of 20 April, 2023*).[4,5] Currently, experimental methods for protein structure determination include X-ray crystallography[6], nuclear magnetic resonance (NMR) spectroscopy[7], and cryo-electron microscopy[8]. While these methods are accurate and provide static snapshots of protein structures and PPIs, they are laborious, resource-intensive, and often infeasible for higher-order protein assemblies. When experimental approaches are recalcitrant, computational modeling provides an alternative to elucidate structures and decipher the structure-function relationship in PPIs.[9] In spite of the limitations in accuracy, computational models are faster, cheaper, and generalizable to a broad variety of biomolecular moieties. In this dissertation, I focus on modeling these interactions and predicting protein complex structures with downstream applications for human health and disease.

1.2 Protein-protein docking: an overview

Proteins dynamically interact with each other forming transient or permanent complexes mediating function: insulin binds with its receptor inducing a conformational change that activates tyrosine kinases[10], hemoglobin undergoes shape changes in interacting protein chains improving its oxygenation ability[11,12], or the motion of the receptor-binding domain (RBD) of the SARS-CoV-2 spike protein to bind with

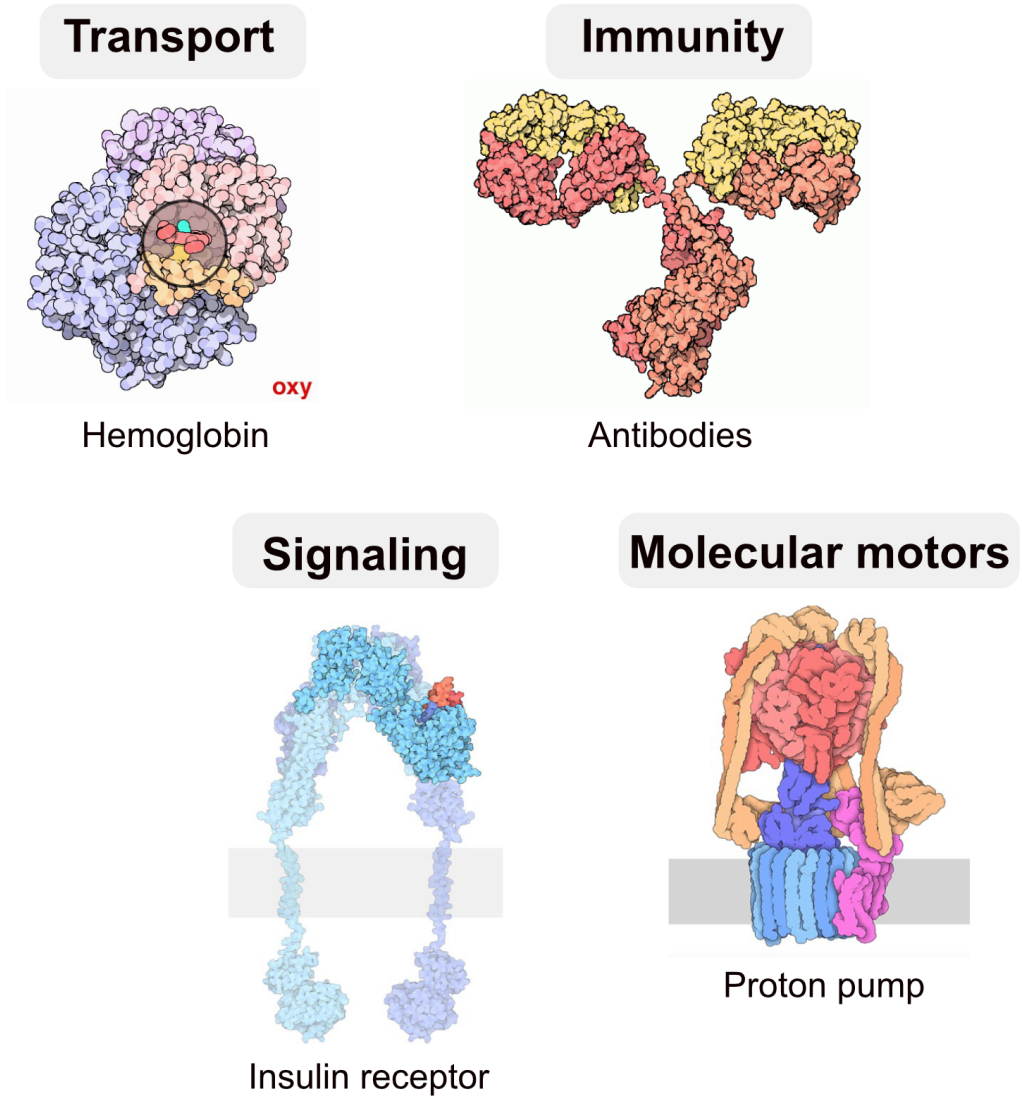


Figure 1.1: Protein-protein interactions govern biological mechanisms. (A) Transport: Hemoglobin in oxy state. (B) Immunity: antibody-antigen interactions in the adaptive immune system. (C) Signaling: Insulin receptor activates receptor tyrosine kinase (RTK) pathways. (D) Molecular motors: Proton pumps on cellular outer-membranes. (*figures inspired by David Goodsell, PDB 101, and created via ccsb.scripps.edu/illustrate/*)

the ACE2 cellular receptor.[13,14] Computational approaches have attempted to model these interactions underpinning most biological processes. From the early ideas of shape and charge complementarity[Janin2003, 15], to the discretized search in roto-translational space with reduced-representation of proteins[16], computational algorithms have vastly aimed at capturing PPIs. Almost four decades prior (in 1986), Michael Connolly described these approaches as the protein docking problem: "Given the three-dimensional structures of any two proteins, is it possible to predict whether they will associate, and if so, in what way?"[17]

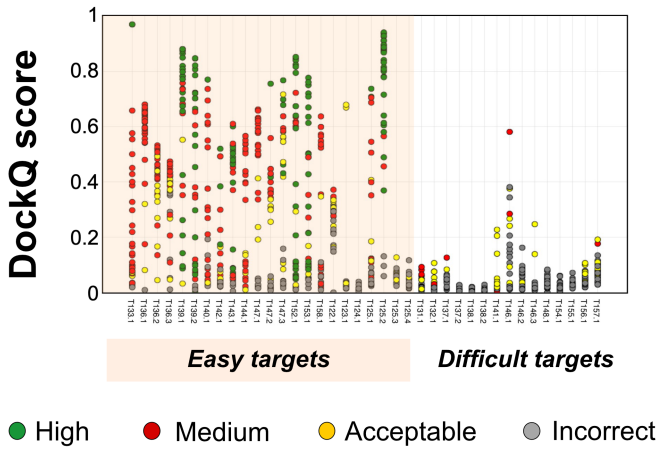


Figure 1.2: Performance of protein docking approaches on blind targets in CAPRI Rounds 38–46. Distribution of DockQ scores for the best model submitted by each predictor group (points) for each individual target (x-axis). DockQ measures a combination of intermolecular residue-residue contacts, interface RMSD, and ligand RMSD on a scale of 0 (incorrect) to 1 (matching the experimental structure). Targets are labelled by their CAPRI target number and, when needed, interface number (after the decimal). The targets are classified into rigid (easy) targets (high-homology monomer templates and under 1.5 Å unbound–bound backbone motion, and flexible targets (poor template availability and/or over 1.5 Å $RMSD_{BU}$). DockQ scores are color-coded by CAPRI model quality ranking: blue, high; green, medium; yellow, acceptable; gray, incorrect. Data graciously provided by Lensink *et al.*[18]

With this underlying theme, docking approaches have provided a route to predict the three-dimensional structures of protein assemblies from structures of known monomeric proteins. This approaches included template-based searches, fast Fourier

transform (FFT)-based docking[16,19], geometric hashing[20], and even exhaustive Monte Carlo (MC)[21,22] or molecular dynamics (MD) simulations.[23,24] However, as the description by Connolly fails to address, one of the most challenging aspects of protein association is binding-induced conformational changes in protein backbones. The intrinsic flexibility of proteins still confounds the protein docking community at large.[25] Protein docking no longer remains a problem of just predicting whether proteins will associate, but rather has extended to a problem of predicting what conformational changes are necessary for proteins to bind.

1.3 The Rosetta Software Suite

The Rosetta modeling suite is a software for biomolecular structure prediction and design.[26] Initially developed exclusively for protein structure prediction and design in the 1990s, the software has expanded to incorporate challenging and diverse modeling tasks, including small molecules, peptides, carbohydrates, and nucleic acids. Over 20 years, a global community of developers, scientists and trainees have contributed to Rosetta, resulting in a software suite with a versatile and modular interface for broad applications in structural biology.

1.3.1 Sampling and Scoring

Following the premise of Anfinsen's hypothesis[27] that natural sequences fold into conformations lying at the global free-energy minimum, Rosetta equips a heuristic sampling-and-scoring approach to find native-like states. To sample putative conformations in the free-energy landscape, Rosetta employs a Monte Carlo-plus minimization (MCM) routine. The degrees of freedom of a biomolecular system (proteins, carbohydrates, nucleic acids) are randomly perturbed by transformations

(chosen from a defined set of moves), energy minimized, and the transformations are accepted or rejected based on the Metropolis criterion[28]:

$$\alpha = \begin{cases} accept, & \text{if } P \geq U(0,1); \text{ for } P = min\left(1, e^{[-\Delta E/k_B T]}\right) \\ reject, & \text{otherwise} \end{cases} \tag{1.1}$$

Here, P represents the probability of acceptance sampled from a Boltzmann distribution, such that $\Delta E = E_j - E_i$, where E_j and E_i represent the energies of the initial(i) and final(j) states respectively, k_B is the Boltzmann constant, and T is the temperature. $U(0,1)$ represents a random number selected from a uniform distribution between zero and one.

To estimate the energies, Rosetta uses an energy function comprising of physics-based energy terms, empirically-derived terms, and knowledge-based terms.[29] A weighted linear combination of these energy terms (E_i) as a function of the degrees of freedom (Θ) sums up the energy function:

$$\Delta E_{net} = \sum_i w_i E_i(\Theta_i) \tag{1.2}$$

The energy terms constituting the standard Rosetta scoring function (`REF2015`) are thoroughly described in a comprehensive review by Alford *et al.*[29] and briefly include the following terms:

- Physics-based energy terms: Lennard-Jones potential (`fa_rep`,`fa_atr`, and `fa_intra_rep`), Gaussian exclusion implicit solvation (`fa_sol`, `fa_intra_sol`) and orientation-dependent solvation (`lk_ball_wtd`), coloumbic electrostatic potential (`fa_elec`)

- Empirical terms: an orientation dependent hydrogen-binding potential (`hbond`

terms) and disulfide potential (`dslf_fa13`)

- Knowledge-based terms: statistical potentials for backbone dihedrals (`rama_prepro`), side-chain torsion (`fa_dun`), and amino acid identity (`p_aa_pp`).

1.3.2 Architectural overview

Rosetta is written in an object-oriented fashion, with the `Pose`, the `Movers`, and the `Scorefunctions` constituting the trifecta at the core of any Rosetta protocol.[26] The `Pose` object stores the conformation of the biomolecules in internal coordinate system (ϕ, ψ, ω, and χ). Each `Pose` object could be manipulated by a `Mover` object. Movers are capable of remodeling (e.g. implementing conformational changes, mutating residues) and analyzing (e.g. reporting geometric/energetic information) the pose, or even selecting a subset of the pose object for specific operations (e.g., selecting the complementary-determining regions of an antibody). The `ScoreFunction` evaluates the energetics of the pose to decide whether the new state is accepted or rejected.

Further, Rosetta code architecture is organized into libraries: (a)`core` libraries that contain structural and scoring information; (b) `protocols` libraries that contains code for structural manipulation; and (c) `utility` libraries with code for common data structures (e.g., containers, owing pointers, etc). New classes or methods are declared and defined in files and categorized to specific libraries based on their functionality. The object-oriented design enables new code to inherit from existing code to develop broadly applicable protocols. Typically, Rosetta protocols generate independent MC trajectories and each pose generated within the trajectory is evaluated for specific objective, for e.g, docking protocols are evaluated for interface scores.

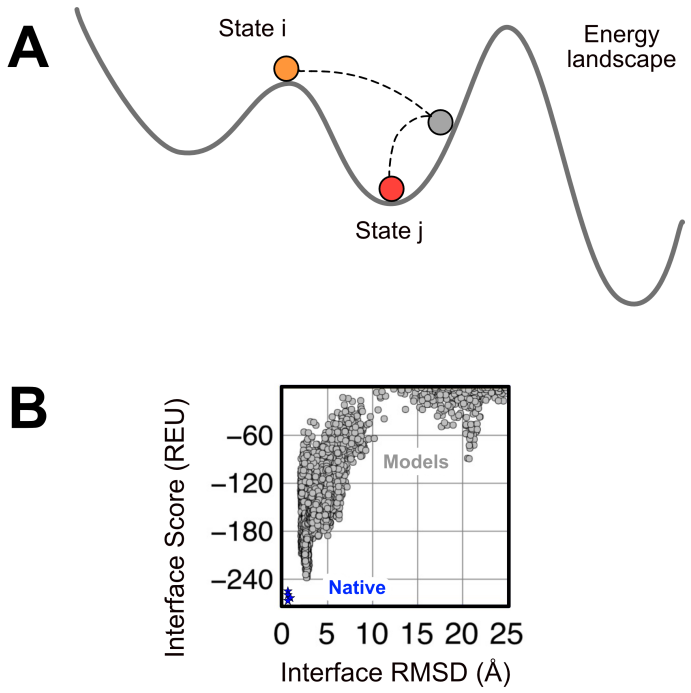


Figure 1.3: Rosetta strategies and metrics. (A) Schematic example of the Monte Carlo minimization strategy in Rosetta. Implementing a random move from state i (*red*) in the energy landscape, the minimization inevitably leads to state j(*orange*) even if the move is made to an arbitrary intermediate state (*gray*). (B) A sample funnel plot of a protein target demonstrating the interface scores (REU) against interface root-mean-square-deviation (RMSD) of the modeled decoys with respect to the native structure. The interface scores (*y-axis*) are analogous to binding free energy, and estimate the score of the complex minus that of individual binding partners. Sampled models show in *gray*. Native crystal structure was relaxed to generate a native funnel (*blue*).

1.4 Dissertation Outline

The association and dissociation of protein complexes is ubiquitous in almost all biological processes and understanding these protein-protein interactions is key towards delineating biological mechanisms and modifying their function. With an aim to elucidate these interactions, this thesis will primarily focus on modeling and design of protein complexes and PPIs. Protein interactions are transient and protein-protein association often induces conformational changes. The first two chapters will describe the development of computational tools for capturing conformational changes upon association, *i.e.*, protein docking. Chapter 2 details the development of ReplicaDock 2.0[30], a temperature-replica exchange MC protocol that performs on-the-fly backbone motions to mimic induced-fit mechanisms of protein binding. In Chapter 3, I describe the extension of this replica exchange strategy for the development of Resolution exchange - an approach to swap configurations (for e.g., all-atom and centroid) for better sampling. Next (Chapter 4), I highlight challenging targets during my tenure as a participant in the Criticial Assessment of PRotein Interactions (CAPRI) over the past 4 years (Rounds 47-54)[18], and discuss the impact of AlphaFold (DeepMind's AI for protein structure prediction) on the field[31]. In Chapter 5, I fuse our docking models with AlphaFold's structural module to develop a pipeline that transforms protein amino acid sequences to accurate protein complex structures. This strategy demonstrates better performance than existing models and shows how AI models could be integrated with biophysics to improve accuracy of structural predictions. The next two chapter (Chapter 6 and Chapter 7) describe the application of computational modeling approaches to real-life examples of protein interactions and design. In Chapter 6, I extend my computational modeling expertise to characterize

the translocation of a bacteriocin through a outer-membrane receptor.[32] Chapter 7 discusses the development of a protein design pipeline for the generation of orthogonal protein interfaces. Orthogonal proteins are mutually exclusive to each other and do not interact with wildtype, endogenous proteins, thereby imparting high specificity and no off-target activity. I describe our efforts in successfully creating an orthogonal interface for the platelet-derived growth factor (PDGF) signaling system with applications to bone growth and tissue engineering. In Chapter 8, I detail a few other miscellaneous contributions and research avenues. Finally, in Chapter 9, I map my contributions to the field and detail directions for future explorers.

References

1. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences* **37,** 205–211. https://doi.org/10.1073/pnas.37.4.205 (4 1951).
2. Pauling, L. & Corey, R. B. The Pleated Sheet, A New Layer Configuration of Polypeptide Chains. *Proceedings of the National Academy of Sciences* **37,** 251–256. https://doi.org/10.1073/pnas.37.5.251 (5 1951).
3. Kendrew, J. C., Bodo, G, Dintzis, H. M., Parrish, R. G., Wyckoff, H & Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181,** 662–666. ISSN: 1476-4687. https://doi.org/10.1038/181662a0 (4610 1958).
4. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242. ISSN: 0305-1048. https://doi.org/10.1093/nar/28.1.235 (1 2000).
5. WwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* **47,** D520–D528. ISSN: 0305-1048. https://doi.org/10.1093/nar/gky949 (D1 2019).
6. Smyth, M. S. & Martin, J. H. J. x Ray crystallography. *Molecular Pathology* **53,** 8–14. ISSN: 1366-8714. https://mp.bmj.com/content/53/1/8 (1 2000).
7. Howard, M. J. Protein NMR spectroscopy. *Current Biology* **8,** R331–R333. ISSN: 0960-9822. https://doi.org/10.1016/S0960-9822(98)70214-3 (10 1998).
8. Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., Tran, E. E. H., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A. & Subramaniam, S. Cryo-electron microscopy – a primer for the non-microscopist. *The FEBS Journal* **280,** 28–45. ISSN: 1742-464X. https://doi.org/10.1111/febs.12078 (1 2013).
9. Mosca, R., Céol, A. & Aloy, P. Interactome3D: Adding structural details to protein networks. *Nature Methods* **10,** 47–53. ISSN: 15487091 (1 2013).

10. Menting, J. G., Whittaker, J., Margetts, M. B., Whittaker, L. J., Kong, G. K.-W., Smith, B. J., Watson, C. J., Žáková, L., Kletvíková, E., Jiráček, J., Chan, S. J., Steiner, D. F., Dodson, G. G., Brzozowski, A. M., Weiss, M. A., Ward, C. W. & Lawrence, M. C. How insulin engages its primary binding site on the insulin receptor. *Nature* **493,** 241–245. ISSN: 1476-4687. https://doi.org/10.1038/nature11781 (7431 2013).

11. Harrington, D. J., Adachi, K. & Royer, W. E. The high resolution crystal structure of deoxyhemoglobin S11Edited by K. Nagai. *Journal of Molecular Biology* **272,** 398–407. ISSN: 0022-2836. https://www.sciencedirect.com/science/article/pii/S0022283697912535 (3 1997).

12. Shaanan, B. Structure of human oxyhaemoglobin at 2·1resolution. *Journal of Molecular Biology* **171,** 31–59. ISSN: 0022-2836. https://www.sciencedirect.com/science/article/pii/S0022283683803131 (1 1983).

13. Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T. & Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181,** 281–292.e6. ISSN: 10974172 (2 2020).

14. Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., Graham, B. S. & McLellan, J. S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (New York, N.Y.)* **367,** 1260–1263. ISSN: 1095-9203 (Electronic) (6483 2020).

15. Greer, J & Bush, B. L. Macromolecular shape and surface maps by solvent exclusion. *Proceedings of the National Academy of Sciences* **75,** 303–307. https://doi.org/10.1073/pnas.75.1.303 (1 1978).

16. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics* **65,** 392–406. ISSN: 08873585. https://onlinelibrary.wiley.com/doi/epdf/10.1002/prot.21117http://doi.wiley.com/10.1002/prot.21117 (2 2006).

17. Connolly, M. L. Shape complementarity at the hemoglobin $\alpha 1\beta 1$ subunit interface. *Biopolymers* **25,** 1229–1247. ISSN: 0006-3525. https://doi.org/10.1002/bip.360250705 (7 1986).

18. Lensink, M. F. *et al.* Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics* **89,** 1800–1823. ISSN: 0887-3585. https://doi.org/10.1002/prot.26222 (12 2021).

19. Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D. & Vajda, S. The ClusPro web server for protein-protein docking. *Nature Protocols* **12,** 255–278. ISSN: 17502799 (2 2017).

20. Fischer, D., Lin, S. L., Wolfson, H. L. & Nussinov, R. A geometry-based suite of moleculardocking processes. *Journal of Molecular Biology* **248,** 459–477. ISSN: 0022-2836. https://www.sciencedirect.com/science/article/pii/S0022283695800638 (2 1995).

21. Wang, C., Schueler-Furman, O. & Baker, D. Improved side-chain modeling for protein-protein docking. *Protein science : a publication of the Protein Society* **14,** 1328–39. ISSN: 0961-8368. http://www.ncbi.nlm.nih.gov/pubmed/15802647http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2253276 (5 2005).

22. Wang, C., Bradley, P. & Baker, D. Protein–Protein Docking with Backbone Flexibility. *Journal of Molecular Biology* **373,** 503–519. ISSN: 00222836. http://www.ncbi.nlm.nih.gov/pubmed/17825317http://linkinghub.elsevier.com/retrieve/pii/S0022283607010030 (2 2007).

23. Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y. & Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **330,** 341–346. ISSN: 0036-8075. https://science.sciencemag.org/content/330/6002/341 (6002 2010).

24. Christoffer, C., Terashi, G., Shin, W. H., Aderinwale, T., Subramaniya, S. R. M. V., Peterson, L., Verburgt, J. & Kihara, D. Performance and enhancement of the LZerD protein assembly pipeline in CAPRI 38-46. *Proteins: Structure, Function and Bioinformatics,* 1–14. ISSN: 10970134 (November 2019).

25. Harmalkar, A. & Gray, J. J. Advances to tackle backbone flexibility in protein docking. *Current Opinion in Structural Biology* **67,** 178–186. ISSN: 0959-440X. http://arxiv.org/abs/2010.07455 (2020).

26. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology* **487,** 545–574. ISSN: 00766879 (C 2011).

27. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (New York, N.Y.)* **181,** 223–230. ISSN: 0036-8075 (Print) (4096 1973).

28. Metropolis, N. The beginning of the Monte Carlo method. *Los Alamos Science (1987 Special Issue dedicated to Stanis?aw Ulam),* 125??130 (1987).

29. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T. & Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13,** 3031–3048. ISSN: 15499626 (6 2017).

30. Harmalkar, A., Mahajan, S. P. & Gray, J. J. Induced fit with replica exchange improves protein complex structure prediction. *PLOS Computational Biology* **18,** 1–21. https://doi.org/10.1371/journal.pcbi.1010124 (6 2022).

31. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589. ISSN: 14764687. http://dx.doi.org/10.1038/s41586-021-03819-2 (7873 2021).

32. Cohen-Khait, R., Harmalkar, A., Pham, P., Webby, M. N., Housden, N. G., Elliston, E., Hopper, J. T., Mohammed, S., Robinson, C. V., Gray, J. J. & Kleanthous, C. Colicin-Mediated Transport of DNA through the Iron Transporter FepA. *mBio* **12.** ISSN: 21507511 (5 2021).

Chapter 2

Capturing conformational changes during protein association

2.1 Overview

Despite the progress in prediction of protein complexes over the last decade, recent blind protein complex structure prediction challenges revealed limited success rates (less than 20% models with DockQ score > 0.4) on targets that exhibit significant conformational change upon binding. To overcome limitations in capturing backbone motions, I developed a new, aggressive sampling method, ReplicaDock 2.0, that incorporates temperature replica exchange Monte Carlo (T-REMC) and conformational sampling techniques within docking protocols in Rosetta. ReplicaDock 2.0 mimics induced-fit mechanism of protein binding to sample backbone motions across putative interface residues on-the-fly, thereby recapitulating binding-partner induced

conformational changes. Furthermore, ReplicaDock 2.0 clocks in at 150-500 CPU hours per target (protein-size dependent); a runtime that is significantly faster than molecular dynamics based approaches. For a benchmark set of 88 proteins with moderate to high flexibility (unbound-to-bound iRMSD over 1.2 Å), ReplicaDock 2.0 successfully docks 61% of moderately flexible complexes and 35% of highly flexible complexes. Additionally, I demonstrate that by biasing backbone sampling particularly towards residues comprising flexible loops or hinge domains, highly flexible targets can be predicted to under 2 Å accuracy. This indicates that additional gains are possible when mobile protein segments are known.

2.2 Introduction

Protein-protein interactions (PPIs) mediate most molecular processes in human health and disease, ranging from enzyme catalysis and inhibition to signaling and gene regulation. Predicting protein complex structures can aid in the systematic mapping of PPI networks in the cell, thereby revealing biological mechanisms and providing insights in protein structure-function relationships[1]. Experimental techniques can determine high-resolution protein structures, however, they can be expensive, laborious, and limited. Computational modeling of protein complexes, *i.e.*, protein-protein docking, provides an alternative to elucidate structures and to identify putative interfaces. The accuracy of most docking methods is hampered by binding-induced conformational rearrangements between protein partners[2]. The recent rounds of the community-wide blind docking experiment, Critical Assessment of PRediction of Interactions (CAPRI)[3,4], showed that capturing large-scale conformational changes between protein partners (unbound to bound C_α root mean square deviation (RMSD_{BU}) > 1.2

Å) remains a longstanding challenge: Less than 20% of models submitted for these targets achieved a DockQ score[5] > 0.4 (see first figure in Harmalkar and Gray, 2020[2]).

To improve docking performance, extensive sampling of the protein's backbone conformations is critical. Earlier studies have incorporated backbone motions either by docking a small ensemble (10-20) of backbone conformations of two proteins[6,7] or by moving a restricted set of coordinates[8–10], but they obtained limited success, underscoring the need of better backbone sampling[11]. To push towards larger conformational changes, algorithms broadly emulate two kinetic binding models: (1) conformer selection (CS), and (2) induced-fit (IF)[12–14]. In CS, unbound protein monomers exist in an ensemble of diverse conformations, and the monomer conformations corresponding to the thermodynamically stable minima are selected upon binding[13]. This mechanism motivated a prior method, RosettaDock 4.0[11], a Monte-carlo (MC) minimization protocol that was efficient enough to use 100 pre-generated backbone structures of each unbound protein. RosettaDock 4.0 improved docking to highest reported success rates on flexible targets (49% of moderate, $\text{RMSD}_{BU} > 1.2$ Å, and 31% of difficult targets, $\text{RMSD}_{BU} > 2.2$ Å, successful predictions). However, since the performance of CS-based approaches depends on having native-like backbone conformations in the monomer ensembles, to capture binding-induced conformational changes, it is desirable to sample backbones in a partner-dependent fashion.

Induced-fit (IF) approaches incorporate partner-specific, localized conformational rearrangements. In IF, proteins 'induce' conformational changes upon molecular encounter[15,16]. Since simulating backbone changes throughout the entire protein concomitantly with rigid-body perturbations is computationally expensive ($\mathcal{O}(2 \times 3^{N+1})$ as opposed to $\mathcal{O}(6)$ for N atoms), IF docking approaches have typically been

restricted to small backbone perturbations and side-chain movements[9,17,18]. Molecular dynamics (MD) simulations follow the IF-approach for all atoms, however, they are bound by time and length scales[19,20]. Thus, expensive molecular dynamics (MD) simulations are accelerated with alternative sampling techniques such as steered MD[21], replica-exchange[22], or metadynamics[23] to refine rigid-body poses of docked proteins or dock small, rigid proteins[24,25].

Replica exchange methods, in particular, have been employed for protein docking to perform an unprecedented sampling of putative protein complex structures[26] and association pathways[21]. Temperature replica exchange methods modulate temperature across parallel replicas, with periodic exchanges between the high temperature replicas and the low temperature ones[22]. While temperature affects all atoms, Hamiltonian replica exchange methods update the energy function between the replicas and focus on a relevant degree of freedom of the system[25,27,28]. To date, however, none of these methods incorporate larger conformational rearrangements between protein partners upon docking. Moreover, most of the modeling examples have been limited to rigid-proteins with little flexibility ($\text{RMSD}_{BU} < 1.2$ Å).

Here, I couple the sampling prowess of replica exchange algorithms with the induced-fit binding mechanism to develop a new, aggressive, flexible backbone protein docking method, ReplicaDock 2.0. This method builds on Zhang *et al.*'s prior work on replica-exchange MC-based rigid-docking (ReplicaDock[22,27]) and adds backbone motions along with a fast-scoring, low-resolution energy function to tackle moderate and highly flexible targets. I test ReplicaDock 2.0 on a diverse set of protein targets from the Dockground benchmark[29] that spans rigid, moderately flexible, and highly flexible targets. Despite the power of REMC, it is still unfeasible to explore all

backbone conformational degrees of freedom, therefore I test the efficacy of choosing different flexible subsets. Finally, I examine whether biasing the sampling choices can generate sub-angstrom quality predictions.

2.3 Results

Protein-protein docking studies with T-REMC by Zhang *et al.*[22,27] demonstrated significant improvement in sampling docking orientations, albeit with two important limitations. (1) No backbone degrees of freedom were sampled, restricting the search to rigid-body moves and thus precluding success on medium and highly-flexible docking targets. (2) The low-resolution energy function was inaccurate[22], so the improved sampling often led to incorrect complex structures. In this work, I address these limitations and improve protein-protein docking for previously intractable flexible targets.

2.3.1 ReplicaDock 2.0 protocol selectively samples backbone degrees of freedom while docking

To address the backbone sampling limitation, I created ReplicaDock 2.0, an induced-fit (IF) inspired, T-REMC plus minimization algorithm that samples backbone conformations on-the-fly while docking. ReplicaDock 2.0 consists of two stages, low-resolution sampling and high-resolution refinement (**Fig 2.1**). To capture backbone degrees of freedom, the low-resolution stage performs replica-exchange and samples both backbone conformations and rigid-body orientations. For each docking pose sampled, backbone moves are sampled via *Rosetta Backrub*[30] over the interface residues. My hypothesis was that by narrowing the search to the putative interface, the protocol would capture realistic conformational changes while maintaining feasible compute

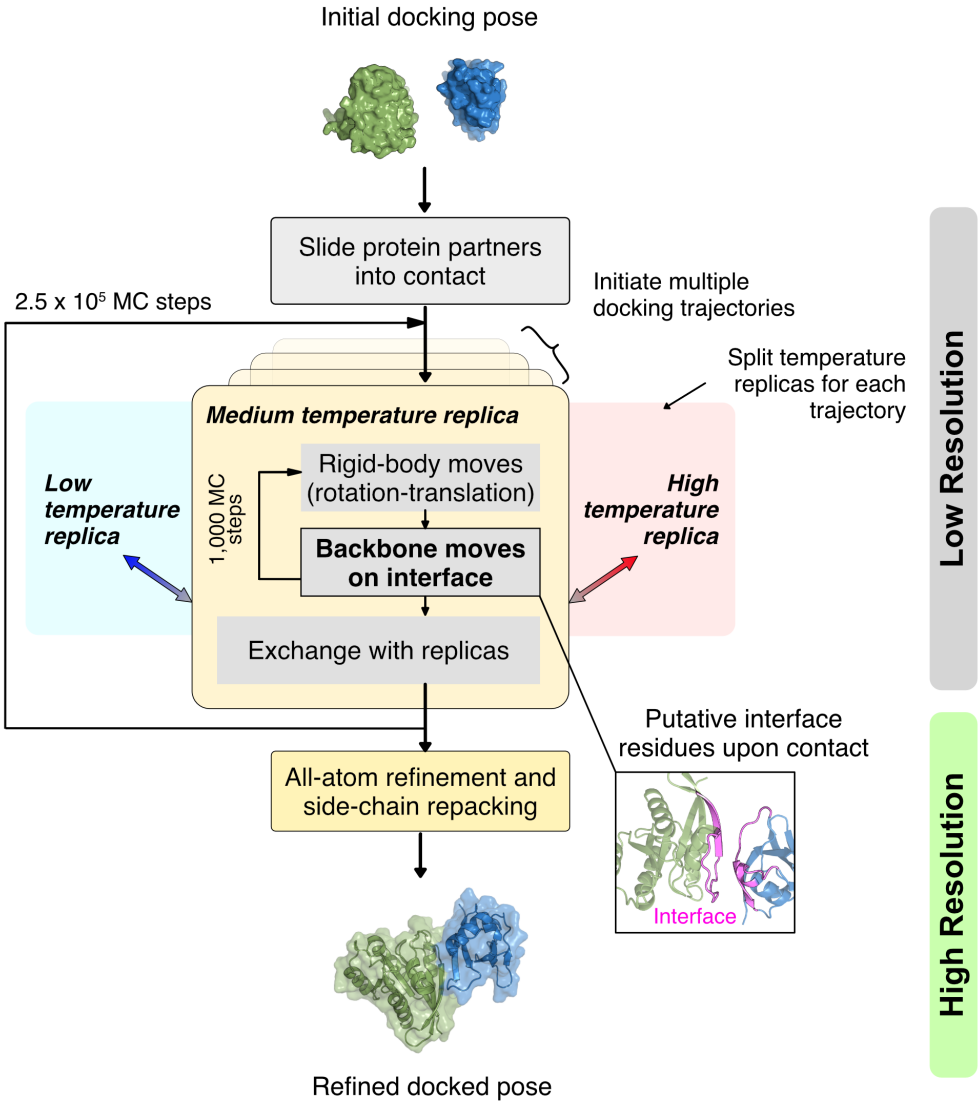


Fig 2.1: Overview of the ReplicaDock2 protocol. *Caption follows on next page*

Fig 2.1: Starting from an initial docking pose *i.e.* a structural model with randomly oriented protein partners, the protocol perturbs the protein partners and slides them into contact. This creates an initial docking pose for the low-resolution stage. Here, the pose object is copied to three parallel replicas per trajectory, and each replica performs rigid body moves (rotation-translation) and backbone moves for each MC trial, followed by exchange between replicas after every $1{,}000^{th}$ trial. Each exchange obeys the Metropolis acceptance criterion and if accepted, the low resolution structure is output. Each trajectory completes 2.5×10^5 MC trial steps, and produces $\sim$5,000 candidate structures. Lastly, all produced structures undergo an all-atom refinement comprising of side-chain packing, small rigid-body motions, and energy minimization to output final docked structural models.

times. The low-resolution IF-based method samples the six rigid-body degrees of freedom along with the 3^N backbone degrees of freedom (ϕ, ψ, ω) for N interface residues. By extending this sampling procedure over three replicas with inverse temperatures, β, of 1.5^{-1} kcal$^{-1}\cdot$mol, 3^{-1} kcal$^{-1}\cdot$mol and 5^{-1} kcal$^{-1}\cdot$mol, a range of backbone conformations sampled. I chose the number of replicas and replica-temperatures such that the energy distribution at any replica overlaps sufficiently with adjacent replicas, allowing efficient exchanges (Figure 2.A.1). After every $1,000$ MC trials of rigid-body and backbone motions, an MC-swap is attempted between neighboring replicas as per the Metropolis criterion[31] (**Methods**). Higher temperature replicas accept backbone moves that would be otherwise rejected at lower temperatures. To expand the diversity of sampled structures, up to 8 independent trajectories are initiated from the starting docking pose. After generation of candidate docking poses in the low-resolution stage, the high-resolution stage performs an all-atom refinement which employs finer rigid-body motions (random rigid-body perturbations in a Gaussian distribution of 0.1 Å to 3°) with side-chain rearrangements followed by energy minimization in the torsional space. This stage does not explicitly move the backbone of the docked proteins but resolves any side-chain clashes and forms a compact, low-energy, high-resolution interface. After evaluating the refined structures' all-atom

scores, the lowest scoring structure is the complex prediction.

2.3.2 ReplicaDock 2.0 uses a residue-transform based scorefunction

As ReplicaDock 2.0 performs backbone sampling and generates docking poses during the low-resolution stage, it is crucial to have a score function that favors native-like interfaces. Thus, the next step in improving docking performance was to tackle the limitation of the inaccurate low-resolution centroid score function as observed by Zhang *et al.*[22]. In their recent CS-based approach, RosettaDock 4.0, Marze *et al.*[11] created the Motif Dock Score (MDS), a pre-tabulated score based on the residue-pair transforms approach[32] where energy of the interacting residues is defined by the 6-dimensional translation and rotation coordinates specifying their relative backbone locations. This simple scorefunction accurately estimated the well-tested all-atom score-function with a faster compute time[32]. MDS is restricted to inter-chain energies, which worked well for pre-generated monomer ensembles with fixed backbones in RosettaDock 4.0. For IF-based ReplicaDock 2.0, however, intra-chain energies must be included as the backbone moves, especially clashes. Therefore, I incorporated knowledge-based backbone torsion statistics terms and Van der Waals interaction terms[33] to create the Motif Updated Dock Score (MUDS). I optimized the relative weights of the MUDS energy terms based on the number of CAPRI acceptable quality structures in the top-scoring 10% of sampled structures (enrichment). With this updated scoring and sampling schemes, I tested the performance of the ReplicaDock 2.0 protocol for global and local docking tasks.

2.3.3 Rigid global docking with ReplicaDock2.0 can identify local binding patches

Docking challenges can be categorized as either global (without any prior knowledge of binding interface) or local (using knowledge of putative binding patches). Conventionally, predictors search with a rigid protein backbone to identify putative binding interfaces (e.g., with ClusPro[34] or ZDOCK[35]), and then each binding interface is refined, often with backbone conformational change. This strategy breaks down docking hierarchically. Global docking has been performed with a T-REMC approach (ReplicaDock[22]), but low-scoring structures were often far from the experimental structure owing to the inaccurate centroid score function. With an updated score function MUDS, I hypothesized that its discriminative power would enable a rigid-body global docking simulation to better identify native-like interfaces. To test this hypothesis, I ran ReplicaDock 2.0 without backbone conformational sampling (only rigid-body rotational and translational moves) on 10 protein targets starting from random orientations of the protein partners.

To illustrate the rigid global docking performance, **Fig 2.2A** plots the low-resolution score (MUDS) versus the RMSD from the native structure for all generated candidate structures for two representative, medium-flexibility protein targets (2CFH, trafficking protein particle complex subunits, 1.55 Å $RMSD_{BU}$[36] and 1XQS, HspB1 core domain complexed with Hsp70 ATPase domain, 1.77 Å $RMSD_{BU}$)[37]. As a reference, I relaxed the experimental bound structure with relatively small rigid-body moves (rotations and translations of 0.5° and 0.1 Å, respectively) to generate near-native structures (blue in **Fig 2.2A**).

ReplicaDock 2.0 generates low-scoring near-native orientations (under 5 Å RMSD)

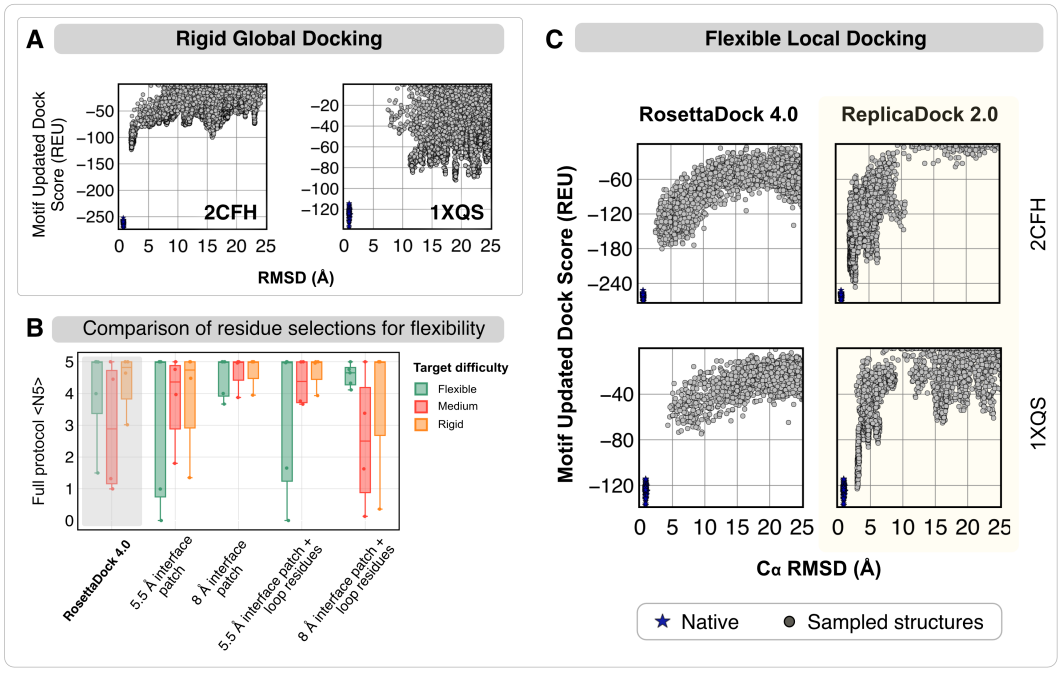


Fig 2.2: T-REMC improves low-resolution performance in global rigid-body and local flexible docking for two representative protein targets. ***(A) Global rigid-body docking performance*** for protein targets 2CFH (trafficking protein particle complex subunits)[36] and 1XQS (HspB1 core domain complexed with Hsp70 ATPase domain)[37]. Plots show the Motif Updated Dock Score (REU) vs all-atom Cα rmsd (Å). Blue points denote the refined native structures. ***(B) Comparison of different residue selections for performing backbone moves.*** Performance of ReplicaDock 2.0 with four conditions: (1) 5.5 Å interface patch, (2) 8 Å interface patch (3) 5.5 Å interface patch + loops, (4) 8 Å interface patch + loops. The metric is $\langle$N5$\rangle$, the average number of near-native models in the five top-scoring structures. For reference, RosettaDock 4.0 performance is highlighted in gray. ***(C) Local flexible backbone docking performance.*** Motif Updated Dock Score (REU) vs Cα rmsd (Å) for two targets, 2CFH[36] and 1XQS[37]. Panels show $\sim$5,000 decoys generated by RosettaDock 4.0 (left) and ReplicaDock 2.0 (right, this work).

for 2CFH, however, for 1XQS, sampling is limited to RMSD values above 6 Å, with the lowest scoring structures about 20 Å away from the experimental structure. On 10 protein targets (Figure 2.A.2), ReplicaDock 2.0 produced models within 5 Å of the native-bound structure for 8 of 10 targets. For comparison, ClusPro[38] successfully predicts 6 of 10 targets. Thus, ReplicaDock 2.0 can perform exhaustive global sampling on the protein energy landscape with better near-native discrimination. One limitation is that global docking with ReplicaDock 2.0 requires 600-800 CPU hours, compared with 35 CPU hours (as reported by Varela *et al.*[39]) for ClusPro[34]. Rigid-backbone global docking results from either ClusPro or ReplicaDock 2.0 can serve as the input to a local, flexible-backbone docking search. (AlphaFold[40] or AlphaFoldMultimer[41] could also be used to generate starting structures[42,43] for refinement, if the multiple sequence alignments are sufficient for the target. I discuss some comparisons for past CASP14-CAPRI targets[44] in the supplementary of the published article.**Harmalkar2022**

2.3.4 Flexible local docking with ReplicaDock2.0 samples deeper energy funnels

When given a putative, broadly-defined binding patch, local docking approaches strive to obtain the biological complex structure by capturing conformational changes in protein partners. ReplicaDock 2.0 explores conformational changes by restrictively sampling backbone moves at putative interfaces. To evaluate the extent of flexibility that can be incorporated while docking for optimum performance, I tested ReplicaDock 2.0 protocol (low resolution sampling with high resolution refinement) with different selections of residues for backbone sampling (Figure 2.A.3). First, I performed backbone moves conservatively over only the set of residues with atoms lying within 5.5 Å of the binding partner (Set 1: 5.5 Å interface patch). Then, I expanded the

selection to residues with atoms lying within 8 Å of the binding partner (Set 2: 8 Å interface patch). As loops are the most flexible secondary structural element in a protein structure, I incorporated residues belonging to all the loop regions from the unbound protein monomers, and added them to prior residue sets to obtain Set 3 (5.5 Å interface patch + loop residues) and Set 4 (8 Å interface patch + loops residues) respectively. For local docking on 12 test targets, I generated ∼5,000 structures and sub-sampled sets of 1,000 structures to calculate the expected number of near-native structures (defined as CAPRI acceptable quality or better) in the 5 top-scoring structures ($\langle$N5$\rangle$). $\langle$N5$\rangle$ evaluates the ability of a protocol to sample near-native conformations and discriminate them from false-positive structures (see Methods). Higher $\langle$N5$\rangle$ indicates that in blind predictions, top-scoring structures are more likely to be correct. **Fig** 2.2**B** compares traditional CS-based RosettaDock 4.0 performance with IF-based ReplicaDock 2.0 using each of the four flexibility scopes. Extending the backbone moves to 8 Å interface patch increased $\langle$N5$\rangle$ across all targets, and offered enough flexibility to capture the binding-induced conformational changes. Incorporating loops reduced performance for medium-flexible and rigid targets (average performance for medium-flexible targets dropped from $\langle$N5$\rangle$=5 in Case 2 to $\langle$N5$\rangle$=2.5 in Case 4), possibly due to over-sampling of backbone moves in relatively rigid regions of the protein structure. Adding flexibility to all loops, the scorefunction misdirects sampling in non-native, spurious minimas, resulting in alternate binding modes with large buried surface area or distorted protein tertiary structures (as shown by the false positive minimas in Figure 2.A.4). To capture realistic backbone conformations, I therefore restrict backbone moves to an 8 Å local interface. Unfortunately, this selection precludes longer range, off-site conformational changes.

With the 8 Å selection chosen as the mobile residue set, I next evaluated the local docking performance of ReplicaDock 2.0 against RosettaDock 4.0. This also served as a head-to-head comparison between two kinetic mechanisms of binding *i.e.* IF versus CS. As an example, **Fig 2.2C** shows the generated candidate structures for two representative protein targets 2CFH and 1XQS with the two docking methods. The low-resolution score (MUDS) versus C_α RMSD plots for the targets 2CFH and 1XQS show that ReplicaDock 2.0 samples structures that score lower than RosettaDock 4.0. Further, in contrast with RosettaDock 4.0 funnels, ReplicaDock 2.0 produces deeper funnels, suggesting that as induced-fit enables the protocol to capture better backbone conformations, replica exchange improves the docking orientations of the encounter complexes generated, thereby allowing us to reach lower, native-like energies (bound-derived funnel in blue).

2.3.5 Induced-fit recapitulates native contacts but fails to push backbone sampling towards bound conformations

With low-resolution sampling, ReplicaDock 2.0 explores larger conformational space in a rapid fashion and avoids entrapment in local minima. However, the structures generated are limited with their accuracy and often require all-atom refinement to penalize side-chain clashes or spurious interfaces and yield realistic structures. The all-atom refinement can further lead to smaller motions and side-chain rearrangements that can result in compact binding between protein partners. Hence, I refined the candidate structures generated in low-resolution stage with the Rosetta all-atom `ref2015` energy function[33]. Supplementary[i] in Harmalkar *et al.* represents the low-resolution candidate structures colored with their final high-resolution CAPRI quality.

[i]Figures for the entire benchmark set (88 targets) are included in the supplementary of the manuscript

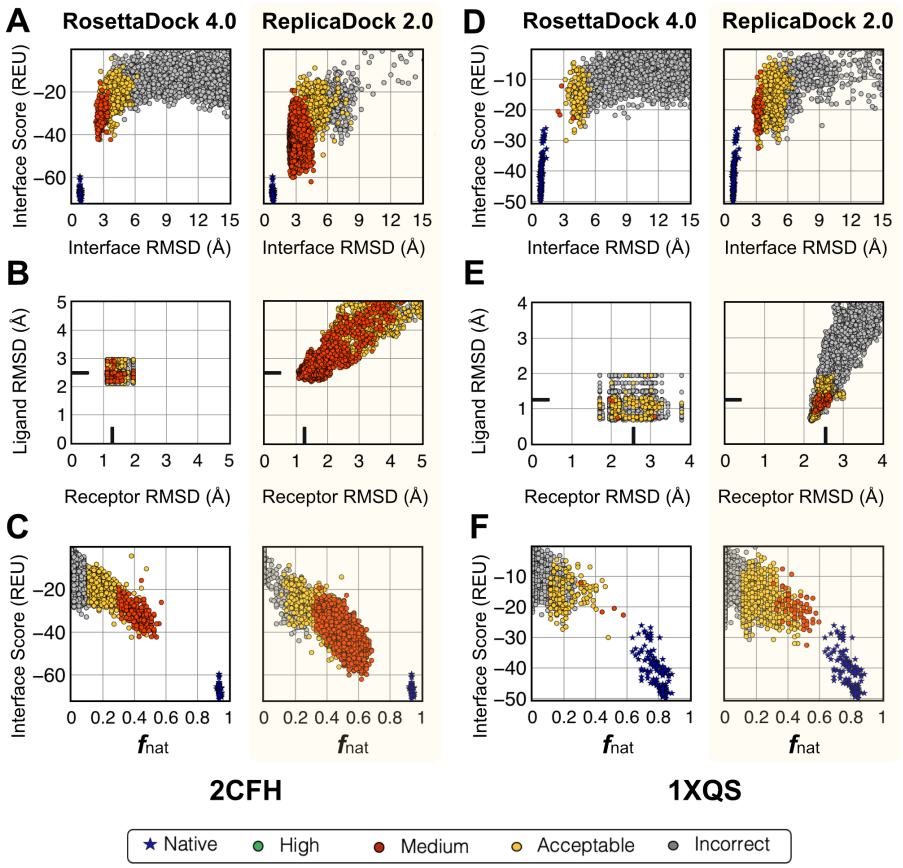


Fig 2.3: Improvement in docking performance after full protocol for two representative targets. (A,D) Interface score (REU) vs I-rmsd (Å), (B,E) Ligand-RMSD(Å) versus Receptor-RMSD(Å), and (C,F) Interface score (REU) vs fraction of native-like contacts post all-atom refinement for RosettaDock 4.0[11] and ReplicaDock 2.0(this work) for two targets 2CFH and 1XQS. Relative to RosettaDock 4.0, ReplicaDock 2.0 samples decoys that score better, are closer to the native, have higher native-like contacts(f_{nat}) and better CAPRI quality. However, backbone RMSDs (B,E) have not moved closer to the native but rather diverged away from it.

In multiple cases (e.g., medium targets 1MQS and 2HRK), the high-resolution stage penalizes poor, false-positive structures and refines near-native structures to improve their quality, showing that best results are achieved by combining the MUDS T-REMC stage with all-atom refinement.

Fig 2.3A and **2.3D** highlight the high-resolution performance for the same two protein targets (2CFH and 1XQS) by comparing the interface energies (equivalent to thermodynamic binding energies) versus the interface-RMSD. For both protein targets, ReplicaDock 2.0 retains the better-scoring structures from the low-resolution stage (**Fig 2.2C**). Relative to RosettaDock 4.0, ReplicaDock 2.0 structures have better all-atom scores and an improved CAPRI quality as evident by the greater number of medium-quality decoys. Despite this improvement, there remains a gap in interface-RMSD between the lowest scoring docked structures and the refined native structures (blue in **Fig 2.3A** and **2.3D**). To determine how induced-fit affects the backbones, I calculated the monomer component backbone RMSDs from the bound backbone conformations (**Fig 2.3B** and **2.3E**). Although ReplicaDock 2.0 generates a much more diverse set of backbone conformations than RosettaDock 4.0, the best RMSDs attained by both the methods are comparable. Note that RosettaDock 4.0 uses pre-generated ensembles resulting in all candidate docking structures being limited in the backbone conformation space (all RMSDs within a rectangular region), whereas ReplicaDock 2.0 generates more diversity. These docking metrics for the entire benchmark set are illustrated in the supplementary[ii].**Harmalkar2022** Further, I calculated the native-like interactions made by the interface residues with the fraction of native residue-residue contacts, f_{nat} (**Fig 2.3C** and **2.3F**). With the induced-fit strategy, ReplicaDock 2.0 increases the f_{nat} over RosettaDock 4.0 by $\sim$0.2. By sampling protein conformations

[ii]Figures for the entire benchmark set (88 targets) are included in the supplementary of the manuscript

in the vicinity of its binding partner, ReplicaDock2.0 is able to orient more interface residues to a native-like state, thereby recapitulating a larger fraction of bound contacts.

2.3.6 Benchmark evaluation demonstrates improved performance over conformer-selection methods

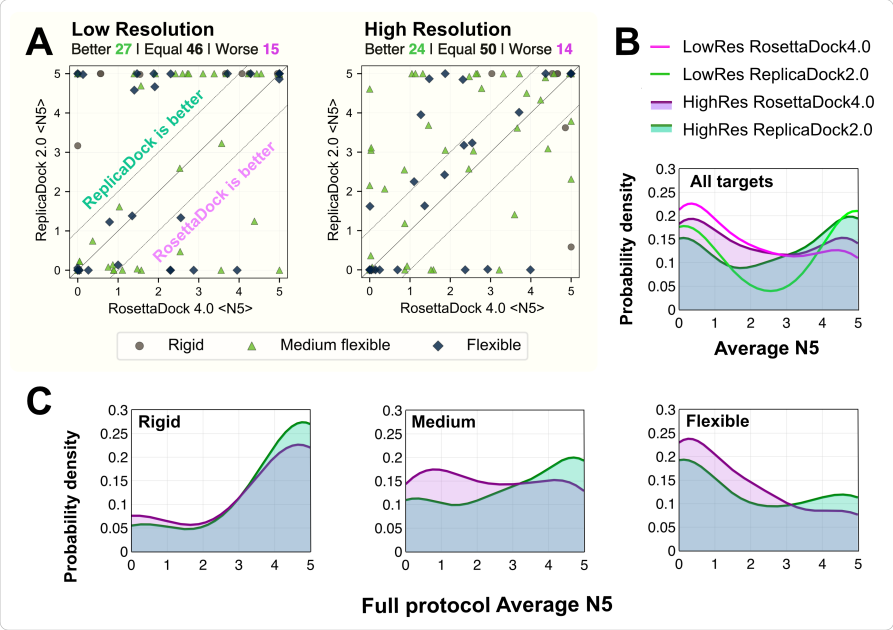


Fig 2.4: Comparison of performance metrics between RosettaDock 4.0 and ReplicaDock 2.0 for individual complexes in a benchmark set of 88 docking targets. (A) Comparison of ⟨N5⟩ values after low-resolution and high-resolution stages (full protocol), respectively. Dashed lines highlight the region in which the two protocols differ significantly, *i.e.* by more than one point in their ⟨N5⟩ values. Different symbols correspond to each target's difficulty category (circle: rigid; triangle: medium; diamond: flexible). Points above the solid line represent better performance in ReplicaDock 2.0, while points below the line represent better performance in RosettaDock 4.0. After the full protocol, 24 targets are modeled significantly better and 14 complexes are modeled significantly worse. (B) Probability density curves versus ⟨N5⟩ for all targets for ReplicaDock 2.0 (green) and RosettaDock 4.0 (purple). Low-resolution performance is indicated by lines (bright pink and bright green), and high-resolution performance is denoted by shaded area (purple and green). (C) Probability density curves versus full-protocol average N5 for rigid, medium and flexible targets respectively.

To evaluate the accuracy of local docking with ReplicaDock 2.0, I benchmarked

the protocol on 88 protein targets from Docking Benchmark DB5.5[29], constituting 10 rigid targets along with all 44 moderately-flexible (medium) and 34 highly flexible (difficult) targets. For each target, I generated ∼5,000 candidate structures with ReplicaDock 2.0 and, for comparison, RosettaDock 4.0. The ensemble generation and pre-packing for the RosettaDock 4.0 protocol was performed as described in Marze *et al.*[11]. For ReplicaDock 2.0, I docked protein targets as summarized in **Methods**.

To compare the performance, I measured $\langle N5 \rangle$ after the low-resolution and high-resolution stage for the full benchmark set of 88 targets. I define a structure as near-native if the C_α RMSD $\leq$ 5 Å for the low-resolution stage, and if the CAPRI rank is acceptable or better for the high resolution stage. **Fig 2.4A** shows the $\langle N5 \rangle$ scores of the benchmark targets for the two protocols. The dashed lines demarcate the region of no improvement *i.e.*, the two protocols differ by less than one point in their $\langle N5 \rangle$ scores. For targets above the dashed line (upper diagonal region), ReplicaDock 2.0 performs better, while for those below the dashed line (lower diagonal region), RosettaDock 4.0 performs better. In the low-resolution stage, ReplicaDock 2.0 outperforms RosettaDock 4.0 with nearly a third of the targets having better $\langle N5 \rangle$ (27 out of 88). After the high-resolution stage, ReplicaDock 2.0 outperforms RosettaDock 4.0 on 24 targets.

To better illustrate the trend, I plotted the probability density of $\langle N5 \rangle$ across all targets (**Fig 2.4B**). With the probability density curves, since the criterion to determine a near-native structure is different for high-resolution and low-resolution stages, the area under curve (AUC) differs. However, it can capture some overarching trends in performance: ReplicaDock 2.0 shifts the curve towards higher $\langle N5 \rangle$, particularly for moderately-flexible targets (**Fig 2.4C**). For 37 out of 44 moderately-flexible targets,

ReplicaDock 2.0 performance is either equivalent or better than RosettaDock 4.0. However, for highly flexible targets, the improvement is modest; docking proteins with higher conformational changes ($RMSD_{BU} > 2.2$ Å) is still a challenge. On an absolute basis with $\langle N5 \rangle \geq 3$ as a success criteria, ReplicaDock 2.0 correctly predicts near-native docked structures in 80% of rigid, 61% of medium-flexible and 35% of highly-flexible docking targets.

Finally, I also compared the run time of ReplicaDock 2.0 with RosettaDock 4.0 for local docking across the benchmark targets. For all benchmark targets, I could generate the ReplicaDock 2.0 trajectories on our[iii] current hardware (24 processors) in a compute time of 8-72 CPU-hrs. The scaling of the ReplicaDock 2.0 and RosettaDock 4.0 protocols with the number of residues in the complex, N_{res}, is illustrated in Figure 2.A.5.

2.3.7 Sampling of known mobile residues captures near-bound conformations of highly flexible protein targets

While ReplicaDock 2.0 generates better quality structures, it fails to reach sub-angstrom interface accuracy for many flexible targets, as shown in **Fig 2.3A** and **2.3C** for 2CFH and 1XQS. Upon inspection of the bound and unbound structures of medium and difficult targets, I observed that the backbone conformational changes were diverse, ranging from motion of loops and changes in the secondary structure to hinge-like motion between intra-protein domains. The residue sets for backbone sampling in ReplicaDock 2.0 were not broad enough to capture these conformational changes. To push towards these larger backbone motions, I wondered whether ReplicaDock 2.0 might attain native-like backbones if it used the information of the residue set

[iii] Production runs were simulated on XSEDE's Rockfish cluster

that results in the conformational change. To test this claim, I identified the mobile residues on the unbound protein partners of Ras:RALGDS domain complex (1LFD, 1.79 Å $RMSD_{BU}$[45]) that showed more than 0.5 Å RMSD when superimposed over the bound structure. Next, instead of automating the selection of interface residues on-the-fly in the baseline protocol, I fed the ReplicaDock 2.0 protocol the identity of these mobile regions. In this version, I restricted the replica-exchange backbone sampling strategy towards pre-selected mobile residues, thereby implementing a *directed* induced-fit mechanism for protein docking.

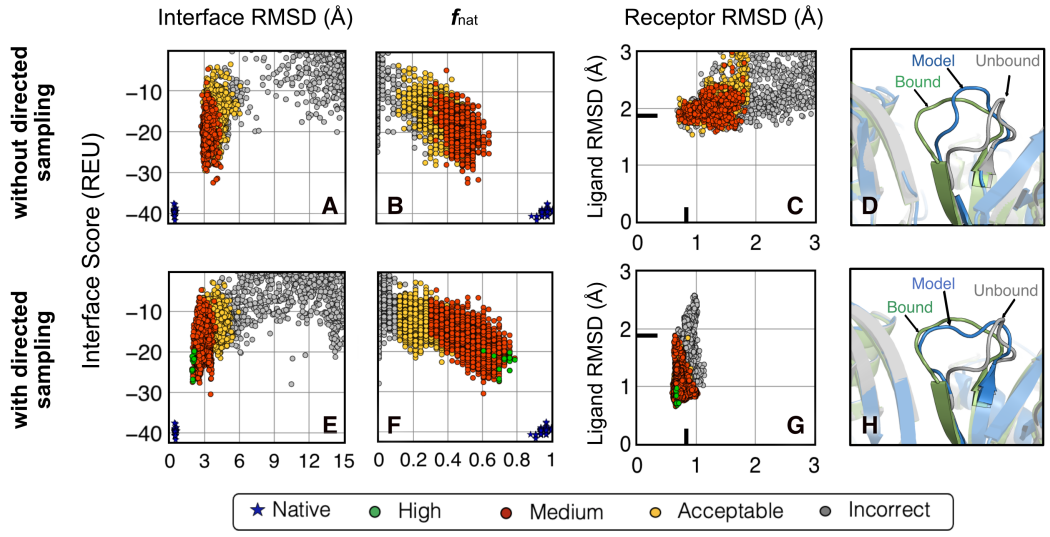


Fig 2.5: Directed induced-fit improves flexible protein docking performance. (top) (a,b,c) ReplicaDock 2.0 without directed backbone sampling of putative interfaces *i.e.* unbiased moves, finds medium-quality structures (colors : green = high quality, red = moderate quality, yellow = acceptable quality, gray = incorrect) (bottom) (e,f,g) ReplicaDock 2.0 with directed backbone sampling of mobile residues improves protein docking and obtains high-quality structures. (d,h) Comparing with the Ras' unbound structure (grey) superimposed over the bound (green), the docked structure loop (blue) has moved closer to the bound state (green) for the two cases respectively. With directed sampling, it is able to capture the backbone structure to sub-angstrom accuracy.

To investigate whether directed induced-fit improves the docking performance, I

evaluated the interface scores, native-like contacts and near-bound backbone conformations. **Fig** 2.5 compares the directed IF approach (*bottom*) with the vanilla version (*top*), which performs unbiased backbone sampling over putative interface residues. The results from **Fig** 2.5**A** and 2.5**E** suggest that with directed IF, the protocol is now able to generate sub-angstrom structures with high-quality CAPRI ranks. In addition, **Fig** 2.5**B** and 2.5**F** show that it also increases the fraction of native-like contacts at the interface from an f_{nat} score of roughly 0.6 to 0.8. The most significant difference is illustrated by the backbone RMSDs of the ligand and receptor chains relative to the bound structure. With directed sampling, the backbone RMSD values do not go higher than ∼0.5 Å away from the bound, starting from the unbound (**Fig** 2.5**G**), whereas the unbiased case samples extensive conformation space away from both bound and unbound (**Fig** 2.5**C**).

Finally, to give a structural perspective, **Fig** 2.5**D** and 2.5**H** show a cartoon-model representation of the unbound and model structure superimposed over the bound structure. With directed induced-fit, the flexible loop retraces an orientation similar to the bound structure. The protocol similarly identified high or medium-quality structures for 15 flexible test targets. Thus, if the flexible residue set could be better identified from the unbound structure, ReplicaDock 2.0 could improve docking further for flexible targets.

2.4 Discussion and conclusions

In this work, I built on advances in T-REMC methods to develop a docking protocol that mimics induced-fit motion and effectively predicts protein complex structures

upon binding. I determined that our IF-based docking protocol, ReplicaDock 2.0, generates more native-like structures than the state-of-the-art CS-based docking method, RosettaDock 4.0, on a benchmark set of moderate and difficult targets. This work highlights two key advances. First, the updated scoring function (MUDS) recognizes native-like interfaces better and penalizes candidate structures with intra/inter-residue clashes, less frequent conformations or low thermodynamic stability. Second, ReplicaDock 2.0 augments the conventional REMC approach with backbone sampling. The protocol explores the ability of an induced-fit approach to manipulate backbone flexibility on docking by flexing interface residues with Rosetta Backrub[30]. This work demonstrates that instead of pre-configuring backbones for protein-docking (*i.e.*, conformer selection), partner-dependent conformational changes (*i.e.*,induced-fit) can result in better molecular recognition.

ReplicaDock 2.0 can be employed for both global and local docking simulations. I demonstrated that the global docking performance of ReplicaDock 2.0 is often better or at par with one leading global docking method (ClusPro), albeit requiring considerably more compute time. With local docking, ReplicaDock 2.0 consistently produced higher success rates using a stringent success criteria. I expand the table created by Marze *et al.*, to compare our results with six other leading docking methods: HADDOCK[46], ClusPro[34], iATTRACT[9], ZDOCK[35,47], RosettaDock 3.2[6] and RosettaDock 4.0[11]. **S1 Table** compares these docking methods, their results and success metrics as well as the size of their benchmark set. Analogous to recent blind prediction challenges, the predictor methods perform with acceptable accuracy for rigid targets, however, the accuracy exceedingly drops as flexibility increases. ReplicaDock 2.0 improves the accuracy for docking moderate flexible targets to 61%, a significant

increase over RosettaDock 4.0 (49%). On difficult targets, the improvement is still limited at 35%, a meager increase over RosettaDock 4.0 (31%). To the best of my knowledge, I present the first instance of a protein docking algorithm attaining ∼60% success rate on moderately flexible targets (1.2 Å $<$ RMSD$_{BU}$ $<$ 2.2 Å).

As ReplicaDock 2.0 is restricted to backbone sampling at putative interfaces, it fails to accommodate off-site conformational changes, for e.g., co-evolutionary residues triggering off-site domain motions or maintaining a fold in the tertiary protein structure. By directing backbone torsional sampling over known mobile residues, I observed that ReplicaDock 2.0 protocol substantially improves the quality and accuracy of docking predictions. Thus, for blind targets, if I could identify potentially flexible residues from homologous structures or from AlphaFold's confidence metrics, such as predicted LDDT Cα score (pLDDT) or predicted TM Score (pTM)[40,41], ReplicaDock 2.0 could be guided to allow targeted flexibility. I anticipate that by improving the ability to predict intrinsic flexibility of residues, T-REMC docking with ReplicaDock 2.0 has potential to make even larger strides in flexible-backbone protein docking.

Despite the improvement in docking performance, ReplicaDock 2.0 brings limited computational speed-up. Currently, ReplicaDock 2.0 generates local docking structures in 30-60% less time than RosettaDock 4.0. Yet, the high compute time is a caveat of the protocol, particularly for larger complexes, or complexes with higher interacting residues. As opposed to embarrassingly parallel approaches that can utilize higher compute power for a similar time-frame, ReplicaDock 2.0 requires 8-72 hrs on 24 processors for a docking simulation. Although this is computational efficient over conformer-selection methods, such as RosettaDock 4.0 or MD simulations, by

increasing sampling trajectories and utilizing multiple processors, I could improve run times without compromising on backbone sampling.

Binding-induced conformational changes, and backbone flexibility at large, has long confounded protein-docking algorithms[2,48]. Protein-protein docking with backbone flexibility via induced-fit for moderate to large-scale motions has not yet been reported. ReplicaDock 2.0 mimics induced fit by moving the backbone in conjunction with docking - for the first time - to consistently reach motions beyond 1 Å in the backbone during protein binding. By improving the understanding of protein interactions and the molecular recognition process, I could determine structures that are yet to be experimentally validated e.g., SERCA-PLB transmembrane complex critical for cardiac function[49], and explore potential association pathways, such as the translocation of protein antibiotics (e.g., colicins) through cellular nutrient transporters[2]. Insights into protein docking and binding interfaces have enabled successful computational designs such as symmetrical oligomers for self-assembling nanocages[50,51] and orthogonal designs of cytokine-receptor complexes[52]. Capturing larger conformational changes will eventually impact the ability to design proteins with complex functions. Looking ahead, I anticipate that capturing the dynamic behaviour of proteins in docking will guide molecular engineering and de novo interface modelling to develop functional protein interfaces for biology, medicine and engineering.

2.5 Methods

2.5.1 Energy Function

2.5.1.1 Low-Resolution energy function

The low-resolution mode of the docking protocol utilizes score function built upon the existing six-dimensional, residue-pair transform dependent energy function, called the `motif_dock_score`[11]. To evaluate backbone sampling and penalize poor backbone conformations, I combine the `motif_dock_score` with energy terms that account for protein backbone dihedral conformations and torsion angles, such as `rama_prepro` (to evaluate backbone Φ and Ψ angles), `omega` (to account for omega torsion corresponding to rotation about the C-N atoms), `p_aa_pp` (a knowledge-based score term that observes the propensity of an amino acid relative to the other amino acids)[33]. To account for inter- and intra-molecular clashes owing to on-the-fly backbone sampling, I also utilize a clash penalty based on atom-pair interactions (*i.e.* Van der Waals attractive and repulsive interactions). The updated score function, called Motif Updated Dock Score (MUDS), serves as the energy function for the low-resolution docking stage in ReplicaDock 2.0.

$$E_{\text{MUDS}} = E_{\text{motif-dock}} + E_{\text{LJrepulsive}} + E_{\text{LJattractive}} + E_{\text{backbone-statistics}}$$

2.5.1.2 All-atom energy function

To refine the docked outputs obtained from low-resolution docking, I use the standard all-atomistic energy function in Rosetta, called `ref_2015` an energy function based on physical, empirical, statistical and knowledge-based score terms[33].

2.5.2 Generation of initial conformations

The docking challenge can be categorized dependent on two scopes, namely, (1) Global docking, where there is no *a priori* knowledge about protein binding, and (2) Local docking, where I have limited information about the binding regions. Global docking challenges are blind protein docking challenges involving prediction of the potential binding sites or orientations. After identifying potential binding regions, local docking aims to narrow down the scope to a localized region of protein to predict the conformations of complexes with better confidence. ReplicaDock2.0 can be applied for both global docking and local docking. For a global docking search, the initial unbound conformations of the binding partners constitute the starting pose (structural model). To generate this starting pose, I randomize the initial orientation of the protein partners (unbound monomers prepacked with Rosetta FastRelax) with the Rosetta option `randomize1, randomize2 and spin` (details in the sample XML script). This orients a binding partner (say ligand), at a random orientation around the other binding partner (say receptor), resulting in a blind global docking set-up.

For local docking simulations, wherein the binding site or patch on the binding partners is known, I start by superimposing the unbound monomer structures over the bound structure. Then, I move the unbound monomers 15 Å away from each other with a 45°) rotation to the ligand (smaller monomer) with respect to the receptor. This serves as the input structure to the ReplicaDock 2.0 protocol. For each trajectory, a Gaussian random 1 Å and 1° perturbation provides slightly different starting states. I have observed that higher temperature replicas often result in much broader exploration of the protein surface. The experimental bound structure is passed to the protocol as the native structure, and is employed as the reference for calculating the

RMSDs. Further details about the protocols, command lines and scripts are reported in the **S1 Text**.

2.5.3 ReplicaDock 2.0 protocol

To sample binding-induced conformational changes during docking, I employed a temperature Replica-exchange MC protocol with backbone conformational sampling in ReplicaDock 2.0. Backbone conformations are sampled with Rosetta Backrub[30]. Amongst the putative interface residues, two terminal residues for each contiguous fragment on the interface are chosen as pivots and backbone dihedral angles are sampled for the residues in between, thereby providing a restrictive IF-like motion.

We scale temperature across three replicas with inverse temperatures set to 1.5^{-1} kcal^{-1}.mol, 3^{-1} kcal^{-1}.mol and 5^{-1} kcal^{-1}.mol, respectively. Replica exchange swaps are attempted every 1,000 MC steps and candidate structures are stored after every successful swap. An exchange attempt is successful if the Metropolis criterion is obeyed as stated below:

$$P_i < \min\left\{1, \frac{\exp\left(\frac{-E_j}{k_B T_i} - \frac{-E_i}{k_B T_j}\right)}{\exp\left(\frac{-E_i}{k_B T_i} - \frac{-E_j}{k_B T_j}\right)}\right\}$$

Here, i and j are the replica-levels across which the swap is performed, E is the MUDS energy, k_B is the Boltzmann's constant, T is temperature and P_i is the probability set in Metropolis criterion that needs to be obeyed for acceptance (generally set to 0.5). Thus, ReplicaDock 2.0 simulations scale the temperatures to modulate the acceptance of backbone and docking moves, so motions that are penalized heavily at lower replicas can be accepted at higher replicas, thereby allowing more diversity in capturing backbone conformations as well as docking orientations. The generated

candidate structures are further passed to the high-resolution stage for all-atom refinement. The all-atom, high-resolution refinement resolves any side-chain clashes and penalizes false-positive orientations that the low-resolution score function failed to penalize. This ensures that the generated output structures are at the lowest possible energetic state achievable for the attained conformations. For each local docking simulation, I initiate 8 trajectories, each trajectory spanning over 3 replicas, run for 2.5 x 10^5 MC steps generating ~5,000 candidate structures. For global docking, I run 10^6-10^8 MC steps, generating roughly 24,000 candidate structures.

2.5.4 Benchmarking, evaluation and success metrics

Four interface residue selection tests were performed on 12 unbound targets from the DockGround Benchmark Set[29] to optimize the flexibility scope over interface residues, number of trajectories and MC trials. Dockground Benchmark Set[29] classifies protein targets as rigid ($RMSD_{unbound\text{-}bound}$ < 1.2 Å), medium (1.2 Å $\leq$ $RMSD_{unbound\text{-}bound}$ $\leq$ 2.2 Å) and difficult targets ($RMSD_{unbound\text{-}bound}$ $\geq$ 2.2 Å), depending upon the conformational change between unbound and bound structures. ReplicaDock 2.0 docking runs were performed on the entire Dockground benchmark set of 44 medium and 34 difficult targets. I added 10 rigid targets for a final set with 88 targets. As defined in CAPRI[53], I calculated the interface RMSD (I-rms), ligand RMSD (L-rms), all-atom RMSD(RMSD), C_α RMSD and fraction of native-like contacts (f_{nat}) against the bound complex. Further, the results of the docking simulations were evaluated with the expected N5 metric. N5 denotes the number of near-native decoys in the five top-scoring structures. A structure is deemed as near-native if the C_α RMSD $\leq$ 5 Å for the low-resolution stage, and if CAPRI rank $\geq$ 1 for the high resolution stage[6]. First, I bootstrapped 1,000 structures, *i.e.* randomly selected 1,000 structures with

replacement from the generated candidate structures. Then, by evaluating whether the five top-scoring structures were near-native, I determined the N5 value. This procedure was repeated 1,000 times for robustness to obtain the expected value ($\langle \mathrm{N5} \rangle$). Successful docking for a target is defined as $\langle \mathrm{N5} \rangle \geq 3$.

2.A Appendix

Here, I show some of the supplementary figures and tables for this work. Additional figures and table illustrating the entire benchmark set is available online

2.A.1 Supplemental Figures

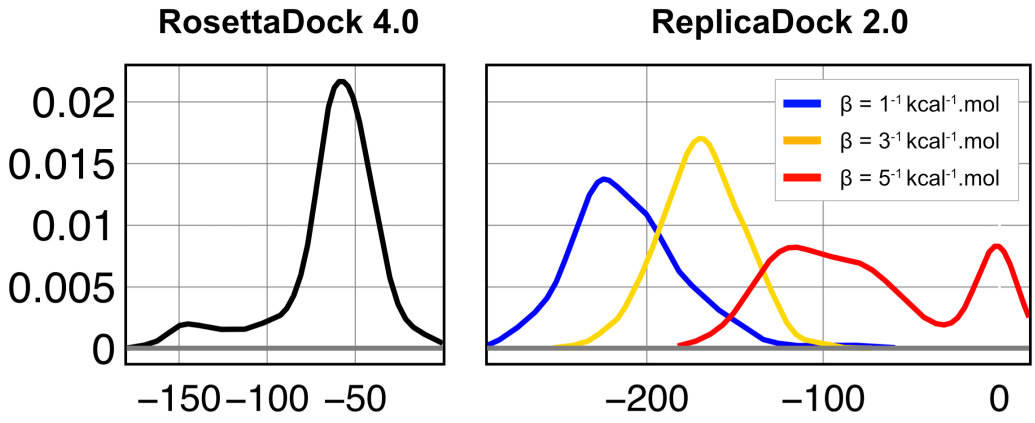


Fig 2.A.1: Energy distribution of conformations sampled with RosettaDock 4.0 and ReplicaDock 2.0 (at respective inverse temperatures) for protein target 2CFH

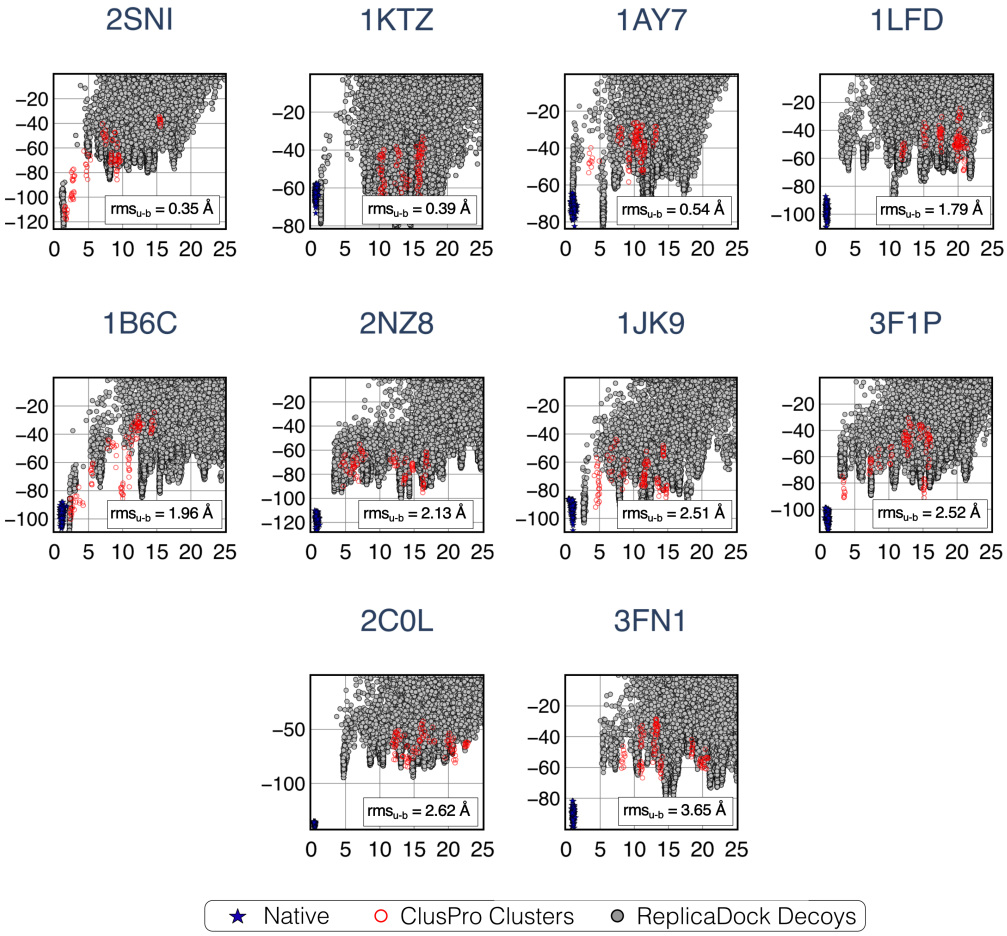
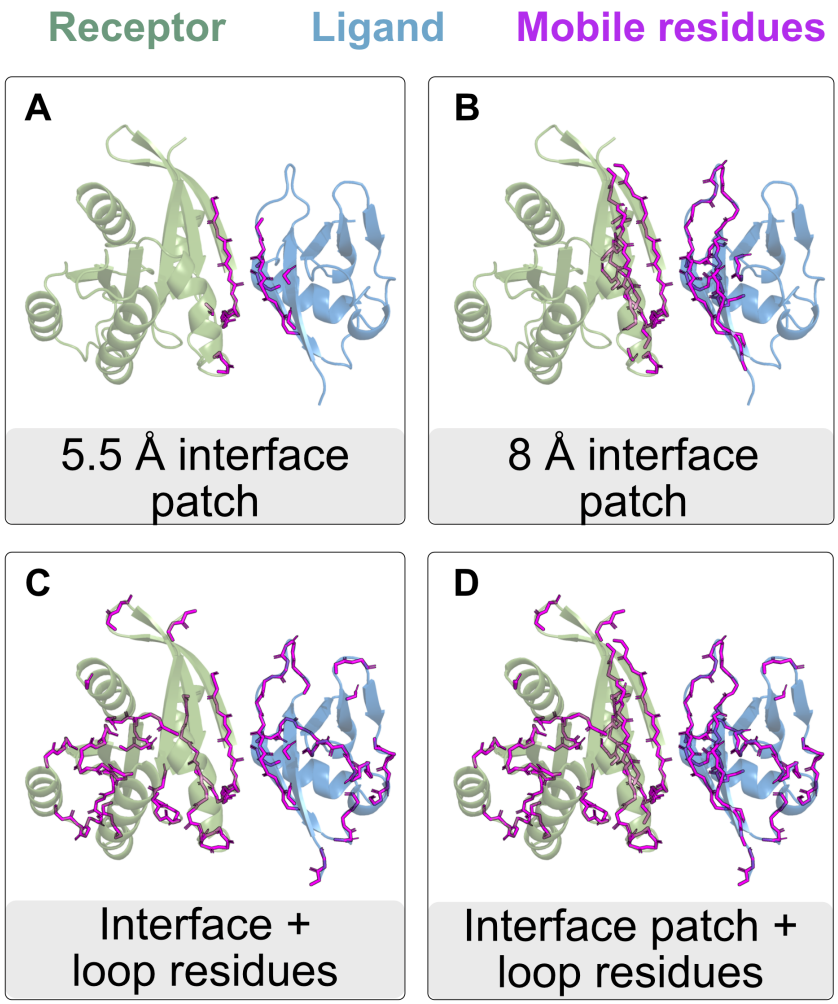


Fig 2.A.2: Global docking performance. Interface score (REU) vs I-rmsd (Å) for each of the 10 benchmark targets, arranged by target difficulty. ReplicaDock 2.0 decoys colored in gray and ClusPro models, relaxed with Rosetta and scored with MUDS, highlighted in red.



Benchmarking targets

Difficult targets	Medium targets	Rigid targets
1FQ1 1JK9 3F1P 3FN1	1GRN 1IJK 3DAW 4FZA	1AY7 1MAH 2PCC 2SNI

Fig 2.A.3: Interface residue selections (*in magenta*) highlighted over a protein target (Receptor, *in green* and ligand, *in blue*). The four residues selections are as follows: (1) 5.5 Å interface patch, (2) 8 Å interface patch (3) 5.5 Å interface patch + loops, (4) 8 Å interface patch + loops. Note that, we also performed a test set by including all the residues of the protein for backbone sampling, however, with T-REMC, such simulations resulted in distortion of the protein quaternary structure (*i.e.* resulted in protein unfolding). Therefore, we chose to exclude that test.

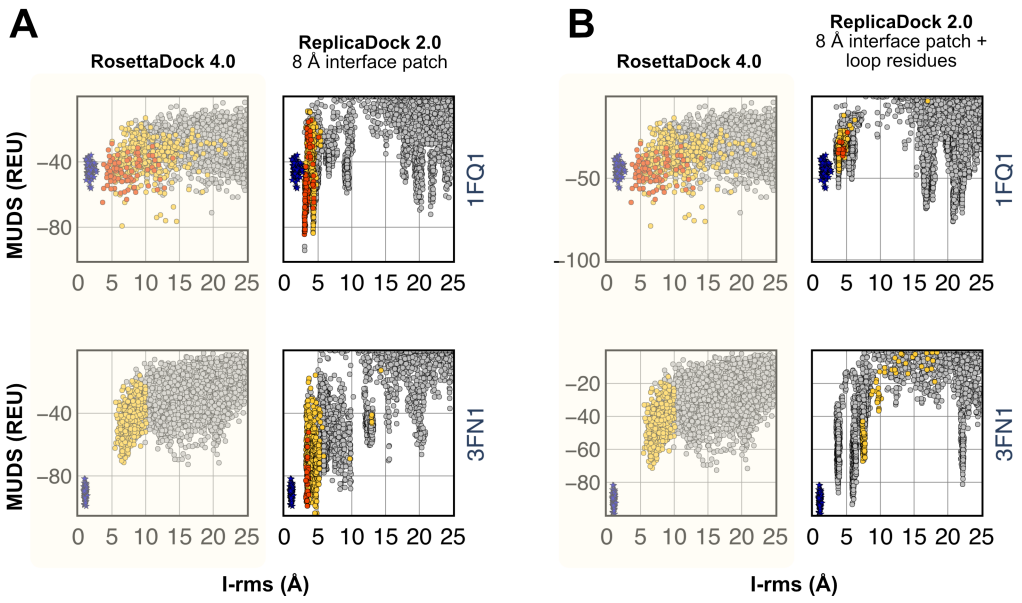


Fig 2.A.4: Performance of updated motif_dock_score MUDS versus Cα-RMSD(Å) plots for RosettaDock 4.0 and ReplicaDock 2.0 for two sets of residue selections in the low-resolution stage. Mobile residues sets are as follows: (left) 8 Å interface patch, and (right) 8 Å interface patch + loops. Candidate structures are colored by the CAPRI quality post all-atom refinement (colors : green = high quality, red = moderate quality, yellow = acceptable quality, gray = incorrect).

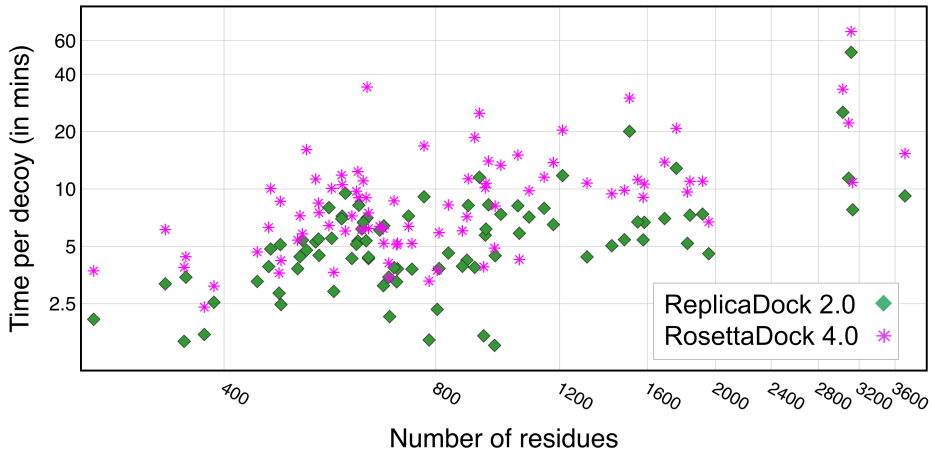


Fig 2.A.5: Compute time comparison between ReplicaDock2.0 and RosettaDock4.0. Scaling of docking simulations on Rockfish Cluster for protein docking targets from the DB5.5 with respect to the number of residues (*log*-scale).

2.A.2 Supplemental Tables

Table 2.A.1: Performance of RosettaDock 4.0 vs. ReplicaDock 2.0 across an 88-target benchmark set. 5,000 decoys were generated by each protocol for each target. Bootstrapped N5 values (plus standard deviations), both after the low-resolution phase and after the full protocol, are listed for each target. Success is defined as $\langle \text{N5} \rangle \geq 3$ for the N5 metrics

Target	i-RMSD$_{u\text{-}b}$ (Å)	Difficulty	RosettaDock (LowRes)	RosettaDock	ReplicaDock (LowRes)	ReplicaDock
1AY7	0.54	Rigid	0.6 ± 0.8	5 ± 0	5 ± 0	5 ± 0
1BVK	1.24	Rigid	0 ± 0	4.9 ± 0.5	3.2 ± 1.1	3.6 ± 1.1
1KTZ	0.39	Rigid	4.3 ± 0.9	4.7 ± 0.6	5 ± 0	5 ± 0
1MAH	0.61	Rigid	1.5 ± 1.2	4.7 ± 0.6	5 ± 0.2	5 ± 0
1MLC	0.6	Rigid	0 ± 0.2	0.1 ± 0.3	0 ± 0	0 ± 0
2BTF	0.75	Rigid	5 ± 0.3	5 ± 0	5 ± 0	0.6 ± 0.9
2JEL	0.17	Rigid	3.7 ± 1.3	4.5 ± 0.8	5 ± 0	5 ± 0
2PCC	0.39	Rigid	0 ± 0	3 ± 1.4	5 ± 0	5 ± 0
2SIC	0.36	Rigid	0.6 ± 0.8	5 ± 0	5 ± 0	5 ± 0
2SNI	0.35	Rigid	4.1 ± 0.9	5 ± 0	5 ± 0	5 ± 0
1B6C	1.96	Medium	5 ± 0	5 ± 0	5 ± 0.1	5 ± 0
1CGI	2.02	Medium	1.9 ± 1.3	1.9 ± 1.4	5 ± 0	3 ± 1.1
1FC2	1.69	Medium	0 ± 0.2	0 ± 0.2	0.2 ± 0.5	3.1 ± 1.3
1GP2	1.65	Medium	1.6 ± 1.1	1.6 ± 1.1	0 ± 0	0 ± 0
1GRN	1.22	Medium	1.4 ± 1.1	1.3 ± 1	5 ± 0	5 ± 0
1HE8	0.92	Medium	4.5 ± 0.7	4.6 ± 0.7	5 ± 0	5 ± 0
1I2M	2.12	Medium	0 ± 0	0 ± 0	5 ± 0	4.6 ± 0.8
1IB1	2.09	Medium	0 ± 0	0 ± 0	0 ± 0	0 ± 0
1IJK	0.68	Medium	3.7 ± 1.1	3.7 ± 1	5 ± 0	5 ± 0
1JIW	2.07	Medium	0.8 ± 1.1	0.9 ± 1.1	0.1 ± 0.3	2.5 ± 1.2
1K5D	1.19	Medium	0.9 ± 0.9	0.9 ± 0.9	0 ± 0.1	1.2 ± 1
1KKL	2.2	Medium	3.1 ± 1.2	3.2 ± 1.2	5 ± 0	0 ± 0
1LFD	1.79	Medium	5 ± 0	5 ± 0	5 ± 0	5 ± 0
1M10	2.1	Medium	0 ± 0.2	0 ± 0.2	0.2 ± 0.5	3 ± 1.2
1MQ8	1.76	Medium	5 ± 0	5 ± 0	0 ± 0	2.3 ± 1.3
1N2C	2.13	Medium	1.3 ± 0.8	0 ± 0	5 ± 0	0.1 ± 0.3
1NW9	1.97	Medium	3.6 ± 1.1	3.6 ± 1.1	0 ± 0	1.4 ± 1.1
1R6Q	1.67	Medium	2.7 ± 1.3	2.7 ± 1.3	5 ± 0	5 ± 0
1SYX	1.64	Medium	5 ± 0	5 ± 0	5 ± 0	5 ± 0
1WQ1	1.16	Medium	2.8 ± 1.3	2.8 ± 1.3	5 ± 0	4.6 ± 0.6
1XQS	1.77	Medium	4.2 ± 0.9	4.2 ± 0.8	5 ± 0	4.3 ± 0.8
1ZM4	2.11	Medium	2.5 ± 1.1	2.5 ± 1.2	0.5 ± 0.9	4.9 ± 0.3
2CFH	1.55	Medium	5 ± 0	5 ± 0	5 ± 0	5 ± 0
2H7V	1.63	Medium	2.6 ± 1.1	2.6 ± 1.2	5 ± 0	3 ± 1.1
2HRK	2.03	Medium	4.4 ± 0.8	4.4 ± 0.9	1.2 ± 1	4.9 ± 0.4
2NZ8	2.13	Medium	3.6 ± 1.1	3.7 ± 1.1	3.2 ± 1.1	3.6 ± 1.1

Target	i-RMSD$_{u-b}$ (Å)	Difficulty	RosettaDock (LowRes)	RosettaDock	ReplicaDock (LowRes)	ReplicaDock
2OZA	1.89	Medium	0 ± 0	0 ± 0	0 ± 0	0 ± 0
2Z0E	2.15	Medium	0.9 ± 0.9	0.9 ± 0.9	0.1 ± 0.5	0.1 ± 0.4
3AAA	1.78	Medium	1.5 ± 1.1	1.5 ± 1.1	0 ± 0.2	0 ± 0
3AAD	2	Medium	0.9 ± 0.9	0.9 ± 0.9	0 ± 0	0 ± 0
3BX7	1.63	Medium	2.4 ± 1.4	2.5 ± 1.3	5 ± 0	2.6 ± 1.2
3CPH	2.12	Medium	0.4 ± 0.7	0.4 ± 0.7	0.7 ± 0.9	2.1 ± 1.2
3DAW	1.49	Medium	4.4 ± 0.8	4.4 ± 0.8	5 ± 0	3.1 ± 1.1
3EO1	1.37	Medium	5 ± 0.2	5 ± 0.2	5 ± 0	5 ± 0
3G6D	1.86	Medium	1.1 ± 1	1.1 ± 1.1	0 ± 0.1	5 ± 0
3HI6	1.65	Medium	0 ± 0.1	0 ± 0.2	0 ± 0	0.4 ± 0.7
3L5W	0.48	Medium	5 ± 0	5 ± 0	5 ± 0	3.8 ± 1
3S9D	1.69	Medium	0.8 ± 0.9	3.9 ± 1.1	0.1 ± 0	4.5 ± 0.8
3SZK	2.1	Medium	3.4 ± 1.2	3.3 ± 1.2	5 ± 0	5 ± 0
3V6Z	1.83	Medium	0 ± 0	0 ± 0	5 ± 0	2.1 ± 1.2
4FZA	2.04	Medium	1 ± 1	1 ± 1	1.6 ± 1.2	5 ± 0
4IZ7	1.56	Medium	0 ± 0.1	0 ± 0.1	0 ± 0	0 ± 0
4JCV	1.62	Medium	2.5 ± 1.1	2.6 ± 1.2	2.6 ± 1.3	5 ± 0
4LW4	1.6	Medium	1.6 ± 1.1	1.5 ± 1.1	4.7 ± 0.8	3.7 ± 1.3
1ACB	2.26	Difficult	2.3 ± 1.2	2.3 ± 1.2	5 ± 0.1	3.2 ± 1.2
1ATN	3.28	Difficult	1.9 ± 1.1	1.9 ± 1.1	4.7 ± 0.6	2.5 ± 1.2
1BGX	6.91	Difficult	0 ± 0	0 ± 0	0 ± 0	0 ± 0
1BKD	2.86	Difficult	0 ± 0	0 ± 0	0 ± 0	0 ± 0
1DE4	2.59	Difficult	2.3 ± 1.2	2.3 ± 1.2	0 ± 0	4.8 ± 0.5
1E4K	2.6	Difficult	0 ± 0.2	0 ± 0.2	0 ± 0	0 ± 0
1EER	2.44	Difficult	0.1 ± 0.4	0.1 ± 0.4	5 ± 0.2	0 ± 0
1F6M	4.9	Difficult	0.3 ± 0.7	0.2 ± 0.6	0 ± 0	0 ± 0
1FAK	6.18	Difficult	0 ± 0.2	0 ± 0.2	0 ± 0	0 ± 0
1FQ1	3.41	Difficult	5 ± 0.1	5 ± 0.1	4.9 ± 0.4	5 ± 0
1H1V	6.62	Difficult	0 ± 0	0 ± 0	0 ± 0	0 ± 0
1IBR	2.54	Difficult	0 ± 0	0 ± 0	0 ± 0	0 ± 0
1IRA	8.38	Difficult	0 ± 0	0 ± 0	0 ± 0	0 ± 0
1JK9	2.51	Difficult	5 ± 0	5 ± 0	5 ± 0	5 ± 0
1JMO	3.21	Difficult	1.4 ± 1	1.4 ± 1	4.6 ± 0.8	1.6 ± 1.2
1JZD	2.71	Difficult	1.9 ± 1.3	1.9 ± 1.2	5 ± 0	5 ± 0
1PXV	2.63	Difficult	2.6 ± 1.3	2.5 ± 1.3	1.3 ± 1.1	3.2 ± 0.6
1R8S	3.73	Difficult	0 ± 0	0 ± 0	0.1 ± 0.3	0 ± 0.1

Target	i-RMSD$_{u-b}$ (Å)	Difficulty	RosettaDock (LowRes)	RosettaDock	ReplicaDock (LowRes)	ReplicaDock
1RKE	4.25	Difficult	0 ± 0.1	0 ± 0.1	0 ± 0	1.6 ± 1.2
1Y64	4.69	Difficult	0 ± 0	0 ± 0	0 ± 0	0 ± 0
1ZLI	2.53	Difficult	0.8 ± 0.9	0.7 ± 0.8	1.2 ± 1	0 ± 0
2C0L	2.62	Difficult	4 ± 1	4 ± 1	0 ± 0	0 ± 0
2HMI	2.26	Difficult	2.9 ± 1.3	2.9 ± 1.3	0 ± 0	0 ± 0.1
2I9B	3.79	Difficult	1 ± 1	1.1 ± 1	0.1 ± 0.4	2.2 ± 1.2
2IDO	2.79	Difficult	3.7 ± 1.1	3.7 ± 1	5 ± 0	4 ± 0.9
2J7P	2.67	Difficult	0 ± 0	0 ± 0	0 ± 0	0 ± 0
2O3B	3.13	Difficult	1.4 ± 1.1	1.3 ± 1.1	1.4 ± 1.4	4 ± 1
2OT3	2.79	Difficult	0 ± 0	0 ± 0	5 ± 0	0 ± 0
3AAD	4.37	Difficult	0 ± 0	0 ± 0	0 ± 0	0 ± 0
3F1P	2.52	Difficult	5 ± 0.1	5 ± 0.1	5 ± 0	5 ± 0
3FN1	3.65	Difficult	1.5 ± 1.2	1.5 ± 1.1	5 ± 0	4.9 ± 0.5
3H11	3.79	Difficult	5 ± 0	5 ± 0	5 ± 0	5 ± 0
3L89	2.51	Difficult	4.3 ± 0.8	4.4 ± 0.8	5 ± 0	5 ± 0
4GAM	5.79	Difficult	2.3 ± 1.2	2.4 ± 1.2	0 ± 0	0 ± 0

Table 2.A.2: Comparison of leading docking methods with ReplicaDock 2.0(derived from Marze *et al.*[11]). (1) Nearest-native structures from rigid-body docking selected for refinement. (2) Half successes awarded for targets with multiple binding sites evaluated, where at least one but not all binding sites are captured. (3) 2.5 cutoff for near-native structures. (4) Cases where bootstrapping gives ≥50% chance of N5≥3 are considered successfully docked. (5) For CAPRI sets, medium and difficult targets are combined, comprising all targets without at least one high-quality prediction by any predictor. (6) Lensink *et al.*[54] (7) Hwang *et al.*[55] (8) Vreven *et al.*[29]. The ReplicaDock 2.0 and RosettaDock 4.0 test sets differ slightly because we omitted some easy targets and we added flexible targets that had been too large for the prior ensemble methods.

Methods						Performance		
Method	Description	Flexibility?	Benchmark Set	Docking Search	Success Metric	Easy Targets	Medium Targets[5]	Difficult Targets[5]
HADDOCK (2017)	Restraint-based docking, minimization	Yes	CASP-CAPRI[6]	Mixed global/local	N10 = 1	12/12 (100%)	4/13 (31%)	
ClusPro (2017)	FFT docking, cluster evaluation	No	CAPRI Rds. 13−35	Mixed global/local	N10 = 1	12.5[2]/16 (78%)	6.5[2]/26 (25%)	
iATTRACT (2015)	Rigid-body docking, interface refinement	Yes	Docking Benchmark 4.0[7]	Global[1]	N200 = 30	55/119 (46%)	9/28 (32%)	0/19 (0%)
ZDOCK (2011)	FFT docking, model evaluation	No	Docking Benchmark 4.0[7]	Global	N100 = 1[3]	58/121 (48%)	7/30 (23%)	0/25 (0%)
Rosetta Dock 3.2 (2011)	Monte Carlo docking, model evaluation	Yes	Docking Benchmark 4.0[7]	Local	N5 = 3	49/84 (58%)	5/17 (29%)	2/14 (14%)
RosettaDock 4.0 (2018)	Monte Carlo docking, model evaluation	Yes	Docking Benchmark 5.0[8]	Local	N5 = 3[4]	10/13 (77%)	21/43 (49%)	10/32 (31%)
ReplicaDock 2.0 (2021)	Replica Exchange Monte Carlo docking	Yes	Docking Benchmark 5.0[8]	Local	N5 > 3[4]	8/10 (80%)	27/44 (61%)	12/34 (35%)

References

1. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems* **12,** 969–982.e6. ISSN: 2405-4712. https://www.sciencedirect.com/science/article/pii/S2405471221003331 (2021).

2. Harmalkar, A. & Gray, J. J. Advances to tackle backbone flexibility in protein docking. *Current Opinion in Structural Biology* **67,** 178–186. ISSN: 0959-440X. arXiv: 2010.07455. http://arxiv.org/abs/2010.07455 (2020).

3. Lensink, M. F., Nadzirin, N., Velankar, S. & Wodak, S. J. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins: Structure, Function and Bioinformatics* **-,** 1–23. ISSN: 10970134 (2019).

4. Lensink, M. F. *et al.* Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins: Structure, Function and Bioinformatics* **87,** 1200–1221. ISSN: 10970134 (2019).

5. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE* **11,** 1–9. https://doi.org/10.1371/journal.pone.0161879 (2016).

6. Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H. & Gray, J. J. Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. *PLoS ONE* **6** (ed Uversky, V. N.) e22477. ISSN: 1932-6203. https://dx.plos.org/10.1371/journal.pone.0022477 (2011).

7. Zhang, Z., Ehmann, U. & Zacharias, M. Monte Carlo replica-exchange based ensemble docking of protein conformations. *Proteins: Structure, Function, and Bioinformatics* **85,** 924–937. ISSN: 08873585. http://doi.wiley.com/10.1002/prot.25262 (2017).

8. Moal, I. H. & Bates, P. A. SwarmDock and the use of normal modes in protein-protein docking. *International journal of molecular sciences* **11,** 3623–48. ISSN: 1422-0067. http://www.ncbi.nlm.nih.gov/pubmed/21152290http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2996808 (2010).

9. Schindler, C. E. M., de Vries, S. J. & Zacharias, M. iATTRACT: Simultaneous global and local interface optimization for protein-protein docking refinement. *Proteins: Structure, Function, and Bioinformatics* **83,** 248–258. ISSN: 08873585. http://doi.wiley.com/10.1002/prot.24728 (2015).

10. Venkatraman, V. & Ritchie, D. W. Flexible protein docking refinement using pose-dependent normal mode analysis. *Proteins: Structure, Function and Bioinformatics* **80,** 2262–2274. ISSN: 08873585 (2012).

11. Marze, N. A., Roy Burman, S. S., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics* **34,** 3461–3469. ISSN: 14602059 (2018).

12. Chaudhury, S. & Gray, J. J. Conformer Selection and Induced Fit in Flexible Backbone Protein-Protein Docking Using Computational and NMR Ensembles. *Journal of Molecular Biology* **381,** 1068–1087. ISSN: 00222836. arXiv: NIHMS150003 (2008).

13. Changeux, J.-P. & Edelstein, S. Conformational selection or induced fit? 50 years of debate resolved. eng. *F1000 biology reports* **3,** 19. ISSN: 1757-594X (Electronic) (2011).

14. Vogt, A. D. & Di Cera, E. Conformational Selection or Induced Fit? A Critical Appraisal of the Kinetic Mechanism. *Biochemistry* **51,** 5894–5902. https://doi.org/10.1021/bi3006913 (2012).

15. Mashiach, E., Nussinov, R. & Wolfson, H. J. FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins: Structure, Function and Bioinformatics* **78,** 1503–1519. ISSN: 08873585 (2010).

16. Kuroda, D. & Gray, J. J. Pushing the Backbone in Protein-Protein Docking. *Structure* **24,** 1821–1829. ISSN: 18784186. http://dx.doi.org/10.1016/j.str.2016.06.025 (2016).

17. Wang, C. H. U. & Schueler-furman, O. R. A. Improved side-chain modeling for protein – protein docking, 1328–1339 (2005).

18. Wang, C., Bradley, P. & Baker, D. Protein–Protein Docking with Backbone Flexibility. *Journal of Molecular Biology* **373,** 503–519. ISSN: 00222836. http://www.ncbi.nlm.nih.gov/pubmed/17825317http://linkinghub.elsevier.com/retrieve/pii/S0022283607010030 (2007).

19. Abrams, C. & Bussi, G. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy* **16,** 163–199. ISSN: 10994300. arXiv: 1401.0387 (2014).

20. Luitz, M., Bomblies, R., Ostermeir, K. & Zacharias, M. Exploring biomolecular dynamics and interactions using advanced sampling methods. *Journal of Physics Condensed Matter* **27.** ISSN: 1361648X (2015).

21. Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature Chemistry* **9,** 1005–1011. ISSN: 17554349 (2017).

22. Zhang, Z. & Lange, O. F. Replica Exchange Improves Sampling in Low-Resolution Docking Stage of RosettaDock. *PLoS ONE* **8** (ed Zhang, Y.) e72096. ISSN: 1932-6203. http://dx.plos.org/10.1371/journal.pone.0072096 (2013).

23. Basciu, A., Malloci, G., Pietrucci, F., Bonvin, A. M. J. J. & Vargiu, A. V. Holo-like and Druggable Protein Conformations from Enhanced Sampling of Binding Pocket Volume and Shape. *Journal of Chemical Information and Modeling* **59,** 1515–1528. ISSN: 1549-9596. https://doi.org/10.1021/acs.jcim.8b00730 (2019).

24. Pfeiffenberger, E. & Bates, P. A. Refinement of protein-protein complexes in contact map space with metadynamics simulations. *Proteins: Structure, Function and Bioinformatics* **87,** 12–22. ISSN: 10970134 (2019).

25. Pan, A. C., Jacobson, D., Yatsenko, K., Sritharan, D., Weinreich, T. M. & Shaw, D. E. Atomic-level characterization of protein–protein association. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 4244–4249. ISSN: 10916490 (2019).

26. Siebenmorgen, T., Engelhard, M. & Zacharias, M. Prediction of protein–protein complexes using replica exchange with repulsive scaling. *Journal of Computational Chemistry* **-,** 1436–1447. ISSN: 1096987X (2020).

27. Zhang, Z., Schindler, C. E. M., Lange, O. F. & Zacharias, M. Application of Enhanced Sampling Monte Carlo Methods for High-Resolution Protein-Protein Docking in Rosetta. *PLOS ONE* **10** (ed Colombo, G.) e0125941. ISSN: 1932-6203. http://dx.plos.org/10.1371/journal.pone.0125941 (2015).

28. Ostermeir, K. & Zacharias, M. Accelerated flexible protein-ligand docking using Hamiltonian replica exchange with a repulsive biasing potential. *PLoS ONE* **12.** ISSN: 19326203 (2017).

29. Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., Chaleil, R., Jiménez-García, B., Bates, P. A., Fernandez-Recio, J., Bonvin, A. M. & Weng, Z. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology* **427,** 3031–3041. ISSN: 10898638 (2015).

30. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. eng. *Journal of molecular biology* **380,** 742–756. ISSN: 1089-8638 (Electronic) (2008).

31. Chib, S. & Greenberg, E. Understanding the Metropolis-Hastings Algorithm. *The American Statistician* **49,** 327–335. eprint: https://www.tandfonline.com/doi/pdf/10.1080/00031305.1995.10476177. https://www.tandfonline.com/doi/abs/10.1080/00031305.1995.10476177 (1995).

32. Fallas, J. A., Ueda, G., Sheffler, W., Nguyen, V., McNamara, D. E., Sankaran, B., Pereira, J. H., Parmeggiani, F., Brunette, T. J., Cascio, D., Yeates, T. R., Zwart, P. & Baker, D. Computational design of self-assembling cyclic protein homo-oligomers. *Nature Chemistry* **9,** 353–360. ISSN: 17554349 (2017).

33. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T. & Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13,** 3031–3048. ISSN: 15499626 (2017).

34. Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D. & Vajda, S. The ClusPro web server for protein-protein docking. *Nature Protocols* **12,** 255–278. ISSN: 17502799 (2017).

35. Pierce, B. G., Wiehe, K., Hwang, H., Kim, B. H., Vreven, T. & Weng, Z. ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30,** 1771–1773. ISSN: 14602059 (2014).

36. Kümmel, D., Müller, J. J., Roske, Y., Henke, N. & Heinemann, U. Structure of the Bet3-Tpc6B Core of TRAPP: Two Tpc6 Paralogs Form Trimeric Complexes with Bet3 and Mum2. *Journal of Molecular Biology* **361,** 22–32. ISSN: 00222836 (2006).

37. Shomura, Y., Dragovic, Z., Chang, H. C., Tzvetkov, N., Young, J. C., Brodsky, J. L., Guerriero, V., Hartl, F. U. & Bracher, A. Regulation of Hsp70 function by HspBP1: Structural analysis reveals an alternate mechanism for Hsp70 nucleotide exchange. *Molecular Cell* **17,** 367–379. ISSN: 10972765 (2005).

38. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics* **65,** 392–406. ISSN: 08873585. https://onlinelibrary.wiley.com/doi/epdf/10.1002/prot.21117http://doi.wiley.com/10.1002/prot.21117 (2006).

39. Varela, D. & André, I. a Memetic Algorithm Enables Global All - Atom Protein - Protein Docking With Sidechain Flexibility. *bioRxiv* (2021).

40. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O.,

Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589. ISSN: 14764687. http://dx.doi.org/10.1038/s41586-021-03819-2 (2021).

41. Evans, R., Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Ž\'\idek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J. & Hassabis, D. Protein complex prediction with AlphaFold-Multimer. *bioRxiv.* https://www.biorxiv.org/content/early/2021/10/04/2021.10.04.463034 (2021).
42. *AlphaFold open source code* https://github.com/deepmind/alphafold. 2021.
43. *ColabFold open source code* https://github.com/sokrypton/ColabFold. 2021.
44. Lensink, M. F. *et al.* Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics,* 1–24. ISSN: 0887-3585 (2021).
45. Lan, H., Franz, H., Steven, M. & Sung-Hou, K. Structural basis for the interaction of Ras with RalGDS. *Nature Structural Biology* **5,** 422–426. http://link.springer.com/10.1007/978-1-62703-429-6 (1998).
46. Vangone, A, Rodrigues, J. P. G. L. M., Xue, L. C., van Zundert, G. C. P., Geng, C, Kurkcuoglu, Z, Nellen, M, Narasimhan, S, Karaca, E, van Dijk, M, Melquiond, A. S. J., Visscher, K. M., Trellet, M, Kastritis, P. L. & Bonvin, A. M. J. J. Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. *Proteins: Structure, Function, and Bioinformatics* **85,** 417–423. https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25198 (2017).
47. Pierce, B. G., Hourai, Y. & Weng, Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS ONE* **6,** 0–5. ISSN: 19326203 (2011).
48. Mandell, D. J. & Kortemme, T. Backbone flexibility in computational protein design. *Current Opinion in Biotechnology* **20,** 420–428. ISSN: 0958-1669. https://www.sciencedirect.com/science/article/pii/S0958166909000913 (2009).
49. Alford, R. F., Smolin, N., Young, H. S., Gray, J. J. & Robia, S. L. Protein docking and steered molecular dynamics suggest alternative phospholamban-binding sites on the SERCA calcium transporter. eng. *The Journal of biological chemistry* **295,** 11262–11274. ISSN: 1083-351X (Electronic) (2020).
50. King, N. P., Sheffler, W., Sawaya, M. R., Vollmar, B. S., Sumida, J. P., André, I., Gonen, T., Yeates, T. O. & Baker, D. Level Accuracy. *Science* **828,** 1171–1175 (2012).

51. King, N. P., Bale, J. B., Sheffler, W., McNamara, D. E., Gonen, S., Gonen, T., Yeates, T. O. & Baker, D. Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510,** 103–108. ISSN: 1476-4687. https://doi.org/10.1038/nature13404 (2014).

52. Sockolosky, J. T., Trotta, E., Parisi, G., Picton, L., Su, L. L., Le, A. C., Chhabra, A., Silveria, S. L., George, B. M., King, I. C., Tiffany, M. R., Jude, K., Sibener, L. V., Baker, D., Shizuru, J. A., Ribas, A., Bluestone, J. A. & Garcia, K. C. Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes. eng. *Science (New York, N.Y.)* **359,** 1037–1042. ISSN: 1095-9203 (Electronic) (2018).

53. Méndez, R., Leplae, R., De Maria, L. & Wodak, S. J. Assessment of blind predictions of protein-protein interactions: current status of docking methods. eng. *Proteins* **52,** 51–67. ISSN: 1097-0134 (Electronic) (2003).

54. Lensink, M. F. *et al.* Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins: Structure, Function and Bioinformatics* **84,** 323–348. ISSN: 10970134. http://www.ncbi.nlm.nih.gov/pubmed/27122118http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5030136 (2016).

55. Hwang, H., Vreven, T., Janin, J. & Weng, Z. Protein-protein docking benchmark version 4.0. eng. *Proteins* **78,** 3111–3114. ISSN: 1097-0134 (Electronic) (2010).

Chapter 3

Coupling resolutions for enhanced sampling

This work was performed in collaboration with Prof. Dr. Martin Zacharias, Technical University of Munich, and was partly funded by the Deutscher Akademischer Austauschdienst (DAAD) research fellowship.

3.1 Overview

Protein-protein interactions (PPIs) are involved in almost all biological processes in human health and disease. Understanding the structure of a protein complex and the associated dynamics can reveal biological mechanisms and suggest intervention strategies. However, modeling biologically relevant protein association at feasible time scales is challenging. To overcome these limitations, I introduce a new enhanced sampling algorithm, called 'resolution replica exchange' (ResEx). The ResEx algorithm improves canonical sampling over rugged, atomistic energy landscapes by swapping conformations between the coarse-grained (CG) and all-atom (AA) states. I apply this algorithm to improve sampling in flexible backbone protein docking. To capture large-scale conformational changes, my ResEx docking strategy mimics the

induced-fit mechanism of protein binding and samples backbone moves on-the-fly. I demonstrate the performance of my method on a small benchmark set of nine protein targets with moderate to high flexibility (unbound to bound RMSD over 0.8 Å up to 4.2 Å). Moreover, this advanced simulation approach leverages the computational advantages of coarse-graining and requires 200-250 CPU hours for a docking simulation (depending on protein sizes), which is efficient compared to molecular dynamics-based approaches. With this work, we show that a CG/AA exchange scheme with conformational sampling shows substantial promise towards quantitative and qualitative modeling of protein complex structures and their dynamic interactions. The proposed algorithm paves the way towards challenging applications in enhanced sampling of biomolecular systems, employing coarse-grained models for capturing conformations without compromising on the accuracy of the all-atomistic models.

3.2 Introduction

The protein conformational energy landscape is extensive and rugged. Modeling protein landscapes and understanding the dynamics of interacting proteins over physically relevant timescales and with feasible computational resources has been a major hurdle in biomolecular simulations.[1] As the conformational space grows exponentially with the system size, ergodic sampling of probable conformational ensembles is challenging. To address this limitation, approaches such as simulated annealing, replica exchange[2], umbrella-sampling[3], and metadynamics[4] have aimed at enhancing sampling while preserving detailed balance. Alternatively, coarse-grained approaches circumvent the exploration challenge by smoothening the protein energy landscape by reducing the number of degrees of freedom (DOFs) of the biomolecular

system.[5] As opposed to all-atom (AA) simulations, coarse-grained (CG) simulations introduce approximations both structurally (centroid representation for side-chains) and physically (simplified energy functions), speeding up simulations by several orders of magnitude. However, these approximations and simplifications of protein biophysics can reduce accuracy and skewed searches in false-positive local minima.

Prior work has demonstrated the benefits of combining CG and AA systems for faster, robust exploration of protein surfaces. Utilizing reduced-protein and centroid models, softwares such as ATTRACT[6] and RosettaDock[7,8] have equipped CG modes with AA refinement for improved protein docking. and van Gunsteren[9] and Kar and Feig[10] have demonstrated the utility of combining force-fields for Hamiltonian replica exchanges (for e.g., by combining CHARMM and coarse-grained PRIMO force-fields). On similar lines, Lyman *et al.* have coupled the AA and CG resolutions for replica exchange and tested this strategy on smaller peptide and butane systems.[11,12] This approach, named as "resolution exchange"[11], extends the replica exchange principle and swaps coordinate representations instead of temperatures or parameters of a hamiltonian. The simplicity of the ResEx strategy lies in the application of a faster, less-accurate scoring mode to accelerate sampling to improve conformational diversity and exploration. Early work with ResEx focused on two replicas only, CG and AA, with later studies introducing incremental coarsening to model intermediate replicas.[13,14] Extrapolating these studies to large protein systems is challenging owing to two primary reasons:

1. With increasing system size, the distinction between CG and AA population distributions increases rapidly, nullifying the exchange acceptance rate to 0

2. Adding more replicas to bridge between the CG and AA modes is difficult. Studies have attempted incremental coarsening (*i.e.* coarse-graining domains and sections of a protein system) to generated intermediate resolutions. However, such resolutions necessitate the development of energy potentials to characterize them.

In this work, I build on prior knowledge of ResEx methods and tackle their limitations to develop a resolution replica exchange strategy for protein docking. First, I discuss our hypothesis and methodological approach to implement resolution exchange. I develop a mixed resolution approach to emulate the intermediate resolutions in a way that overcomes prior challenges with incremental coarsening. Then, I discuss the applicability of these methods into a protein docking strategy that mimics induced-fit mechanism of protein binding. Finally, I test our strategy on a small set of flexible protein targets from the Dockground Benchmark Set 5.5.[15] With this work, I provide a proof-of-principle to the resolution exchange strategy for protein docking and extending its application towards multi-scale modeling of biomolecular systems.

3.3 Theory

3.3.1 Resolutions and score-functions in biomolecular modeling

Conventionally, two types of representations (resolutions) are employed to simulate biomolecules: all-atom (AA) and coarse-grained (CG).[16] All-atom models represent biomolecules in atomistic detail (with explicit or implicit solvent) and have demonstrated their utility in evaluating thermodynamic and kinetic properties.[17] However, explicit all-atom simulations are limited by their longer simulation times and remain computationally intractable for simulating biologically relevant events or large

assemblies. Coarse-grained models, on the other hand, circumvent this issue by reducing the number of degrees of freedom to represent a biomolecular system and by smoothening the rugged energy landscape. For proteins, coarse-graining often involves replacing the side-chain atoms (or fraction of side-chain atoms) with a singular atom capturing the characteristics of the amino acid side-chain while simplifying the backbone heavy atom representation. Some examples of CG representations include the UNRES (united residue) model[18], CABS model[19] (representing only Cα and Cβ atoms with the side-chain as a singular atom), the ATTRACT reduced protein model[6] (three pseudo-atoms representing each amino acid residue), and the Rosetta *centroid* model[20] (representing the amino-acid side-chain except the Cβ with a CEN atom with all backbone heavy atoms). In all, CG models have improved computational efficiency, however the approximations of coarse-graining come at the cost of accuracy.[1]

Within *Rosetta*[21], the high-resolution (AA model) and the low-resolution (CG model) stage are often equipped for biomolecular simulations, especially protein docking.[22] The low-resolution (CG, denoted as Rosetta *centroid*) samples in the smoothed protein energy landscape to obtain putative encounter complexes and the high-resolution (AA) stage refines these structures to sample in the local minima. With different resolutions, one needs calibrated energy functions for each stage to evaluate each sampled decoy. Similar to CHARMM[23] and MARTINI[24] force-fields for all-atom and coarse-grained representations in molecular dynamics simulations, Rosetta simulations employ force-fields for scalability throughout the different resolutions. For all-atom simulations, the Rosetta energy function (`ref2015`[25]) is the gold-standard. It comprises of physic-based, empirical, statistical and knowledge-based terms. For coarse-grained simulations, multiple versions of the centroid representations are

available. Earlier versions employed a score-term that captured sequence-dependent one-body and two-body interactions (solvation, electrostatics) and also sequence-independent terms to capture steric repulsions and excluded volume. Protein docking in low-resolution employed a subset of these terms, with the sequence-dependent terms. As conventional docking protocols were rigid-body (no backbone motions, only 6 degrees of freedom to sample), the interface energy (analogous to binding free energy) was evaluated over interchain contacts *i.e.* interface residues. By evaluating energies on interfaces, *centroid* docking in Rosetta was reportedly faster and efficient for sampling in the protein landscape as opposed to CG versions in molecular dynamics (or alternative explicit solvent approaches).[7,22] Recent updates to the low-resolution score-function involved development of `motif_dock_score`[26], a six-dimensional motif-based residue transform score for fast, coarse-grained searches. Our recent work[27] demonstrated its application for induced-fit docking (rigid body with backbone moves) in Rosetta.

3.3.2 Resolution replica exchange method

Similar to replica exchange that employs a ladder of replicas with variable temperatures or Hamiltonians and swaps their parameters, one can also perform exchanges of configurational coordinates. Resolution replica exchange (ResEx) stems from this hypothesis as a method to exchange low-resolution states with high-resolution states.[11] The general idea in temperature or Hamiltonian replica exchanges is that higher replicas (say higher temperature or a lower-penalty potential function) have lower energy barriers to overcome between local energy minima. Extending the same principle to resolutions, a low-resolution (CG) energy landscape of a protein system approximates to a smoothed version of the rugged high-resolution (AA) landscape. Exchanges

between these replicas would follow the principle of conventional replica exchange strategies and lead to better spatial and conformational sampling. In the upcoming sections, I will validate this hypothesis and set the stage for our ResEx strategy.

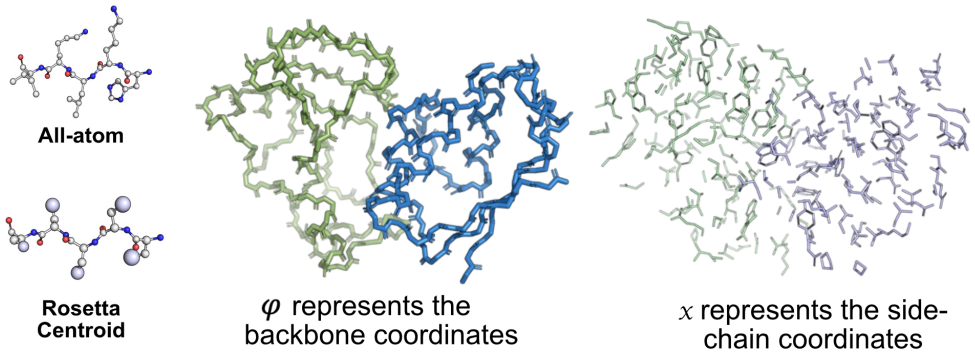


Figure 3.1: Protein representations in biomolecular simulations. All-atom and centroid representations in Rosetta. Each protein pose can be represented by a set of configurational coordinates, say functions ϕ and x, such that ϕ represents the backbone coordinates for a protein pose and x represents the side-chain coordinates.

With temperature replica exchange, backbone conformations are sampled, however many irrelevant states are sampled in higher temperature replicas. Extensively capturing backbone orientations in higher temperature replicas often skews the search in false positive funnels and evades native-like conformations. Hamiltonian replica exchange modifies the energy function across replicas and focuses on a relevant degree of freedom of the system (whereas temperature affects all atoms) and can serve as an efficient strategy to escape entrapment in irrelevant binding states. This, however, limits its utility in capturing larger conformational changes. ResEx is a promising strategy in this regard as the configurational coordinates affect the entirety of the system and can lead to better conformational sampling in low resolution; while the high-resolution would penalize entrapment in false-positive minima.

Prior studies have attempted to employ resolution exchange approaches for

biomolecular simulations, particularly on toy systems with peptides.[12,14] This highlights the utility of this strategy, however, brings forth challenges that have limited the extension of ResEx to large-scale biomolecular simulations. An important limitation of ResEx is that to swap resolutions between two replicas, there should be an overlap between sampled distributions. As system size increases, the energy landscape for AA and CG modes change drastically and can be distinct enough to prevent any exchange between replicas. Some studies have reported incremental coarsening strategies where regions of a molecule are coarsened.[13] This brings us to the next challenge, the incremental coarsening or intermediate resolutions would require robust energy functions to evaluate the generated distributions. So, for every intermediate replica, along with the configurational state (resolution), elucidating an energy function is a must. With this work, we address these challenges and extrapolate the resolution exchange strategy for protein-protein docking.

3.4 Results

3.4.1 Exchanging configurations between CG and AA modes: a feasible strategy for better sampling and efficient scoring

The hypothesis of resolution exchange is that it allows better conformational sampling while escaping false-positive mimima. To validate this hypothesis, demonstrating the benefits of low and high resolution stages is paramount. Here, I asked two questions, specifically:

1. Which resolution successfully penalizes false-positive (non-native) interfaces?
2. Which resolution samples more diverse backbone structures?

To answer the first question, I performed global rigid-body docking on protein

target[28] with the all-atom (`ref2015`[25]) and centroid (`motif_dock_score`[26]) energy functions respectively. This task aimed at sampling orientations across the protein partners and identifying which mode can capture native-like funnels. Figure 3.2.A shows the initial configuration for the global rigid-body docking with the results. In this representative case, the all-atom score function efficiently discriminates false-positive funnels and samples in the near-native region, whereas the centroid sampling is led askew to a non-native site. The all-atom score function, thus, is a better evaluator for discriminating native structures. Next, to answer which phase samples diverse backbone structures, I simplified the degrees of freedom of the problem to focus on backbone sampling. Instead of initiating docking from a random spatial orientation, I superimposed the unbound structures over the native, and initiated docking with backbone moves. The initial configuration is illustrated in 3.2.B and highlighted in *red* in the funnel plots. This narrows down the search space and allows the protocol to focus exclusively on sampling backbone structures. In the *right panel*, I compare interface RMSD with the ligand Cα RMSD (for comparison, the native ligand*blue* has a flexible loop region). All-atom mode samples backbones closer to the unbound template and the diversity of backbones captured is fairly minimal (under 1 Å). On the other hand, centroid sampling of the backbone samples sub-Angstrom RMSDs (both interface and ligand Cα) with a range over 5 Å Lig Cα-RMSD. The lower energy barriers in the centroid phase allows the protocol to accept backbone moves which would otherwise be rejected in all-atom mode owing to backbone/sidechain clashes. CG mode, thus, captures better backbone moves. This validates our initial hypothesis and provides a proof-of-concept for the ResEx strategy.

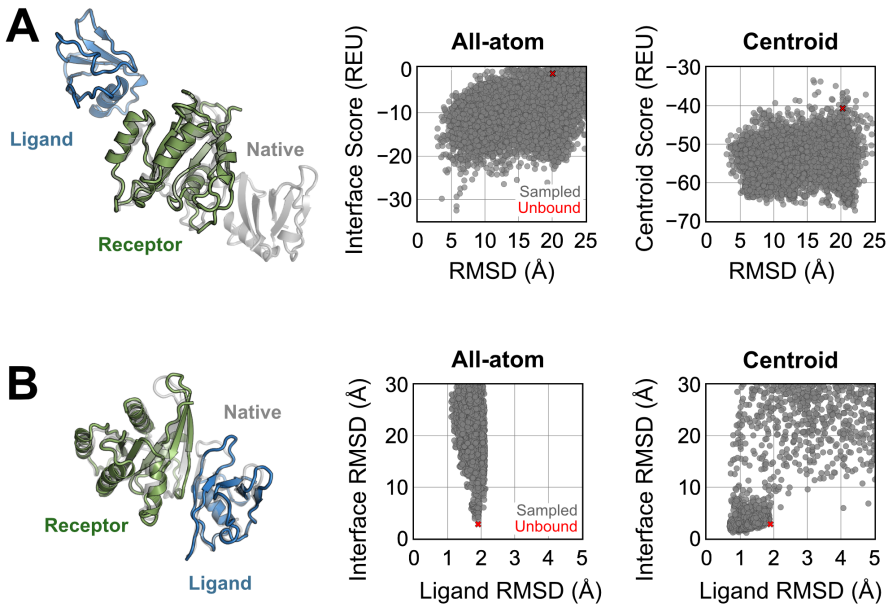


Figure 3.2: Features of low- and high-resolution score-functions. (A) Which energy function discriminates native-like structures? Starting from the docking decoy (highlighted in *green-blue*), global rigid-body docking of the protein partners was initiated with all-atom and coarse-grained score-functions (`ref2015` and updated `motif_dock_score` respectively). Interface score (REU) is compared against RMSD (Å) for sampled decoys with the starting unbound decoy highlighted in *red*. (B) Which energy function samples diverse backbones? Starting from the unbound protein partners superimposed over the bound, I initiate docking simulations with rigid-body and backbone moves. Interface RMSD (Å) v/s ligand Cα RMSD (Å) for the sampled decoys shows that centroid score-function allows for diverse backbone sampling.

3.4.2 Mathematical foundations of resolution exchange

By choosing a subset of coordinates from the all-atom model that could be transformed to make the entire subset of coordinates for the coarse-grained model, ResEx could swap between resolutions (Figure 3.1). Say we have two independent and parallel simulations of a protein system (protein-protein docking simulations), one in low-resolution and other in high-resolution represented by distributions (of conformations) π_{CG} and π_{AA} respectively. The distributions for both the phases are dependent on temperature (T), parameter(k) of the potential function (U), a set of coordinates representing the backbone heavy atoms(ϕ), and a set of coordinates representing the side-chain atoms (x). So, any sampled distribution π_i can be represented as follows:

$$\pi_i(\phi_i,\ x_i;\ T_i;\ ki) = \frac{1}{Z_i} \times \exp\left[\frac{-U(\phi_i,\ x_i;\ ki)}{k_B T_i}\right] \tag{3.1}$$

Here, Z is the partition function and $k_B T$ is the product of Boltzmann constant with temperature. For temperature replica exchange, the set of configuration coordinates ($\phi,\ x$) and potential function parameter (k) is constant and T is swapped; whereas a swap between the parameters (k) results in Hamiltonian replica exchange.

In ResEx, for independent simulations π_{CG} and π_{AA}, with a potential function U_{CG} and U_{AA} defined for the CG and AA phases respectively, one performs exchanges by swapping the backbone coordinate subset (ϕ). Swapping from AA to CG is easier as side-chains could be easily mapped to a single atom representation. However, CG to AA involves swapping ϕ and building the side-chain subset (x) from the new backbone coordinates and that might require side-chain packing/minimization disobeying detailed balance. For independent simulations with observed detailed balance, the exchange acceptance probability (P) would be defined by the Metropolis

criterion:

$$P < min\left[1, \frac{\pi_{\mathsf{AA}}(\phi_b,\ x_b)\ \pi_{\mathsf{CG}}(\phi_a)}{\pi_{\mathsf{AA}}(\phi_a,\ x_a)\ \pi_{\mathsf{CG}}(\phi_b)}\right] \tag{3.2}$$

This demonstrates how conventional replica exchange could be extended to perform resolution exchange simulations by splitting configurational coordinates into all-atom and coarse-grained resolutions.

3.4.2.1 Coupling resolutions

One of the major hurdles in implementing a successful ResEx protocol is to generate replicas at intermediate levels of resolutions. This would ensure that the sampled distributions by these intermediate replicas would overlap, allowing a significant probability of exchanges while sampling from AA to CG stages. Unlike prior work that employed incremental coarsening, here I developed a mixed resolution strategy. For each replica, we define an all-atom 'pose' and a centroid 'dual' such that all moves performed on the all-atom pose are replicated on the centroid dual. Essentially, the dual serves as a centroid deep-copy of the pose. Then, I attribute the replicas with a tuning parameter λ and define the energy potential for a replica to be derived from the AA pose and the CG dual:

$$U_{\text{mix}} = \lambda\ U_{\mathsf{CG}} + (1 - \lambda)\ U_{\mathsf{AA}} \tag{3.3}$$

By performing this operation, we switch the resolution exchange to a Hamiltonian exchange task, where instead of the energy potentials focusing on specific subsets of the proteins, as in prior work, it derives energy from each AA and CG configuration. The tuning parameter is adjusted from 0 to 1 allowing transition from a completely AA stage to a completely CG stage. The Metropolis criterion now changes dependent on

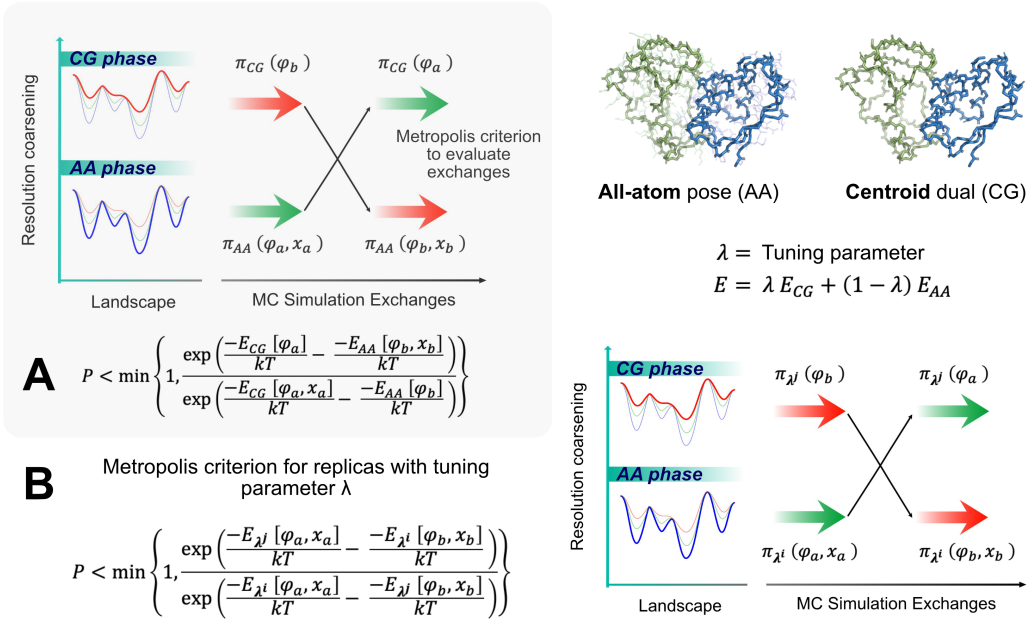


Figure 3.3: Resolution exchange method. (A) Schematic illustration of conventional resolution exchange method with two replicas, an all-atom (AA) replica and a coarse-grained (CG) replica. The distributions are represented by π and the Metropolis Criterion for exchanges with the two replicas are determined as shown by the equation. (B) Schematic illustration of our updated mixed resolution strategy that uses a mixed energy potential (E). Here, multiple replicas can be initiated and the exchange between two replicas i and j is dependent based on the energy contributions of the CG and AA phases to respective replicas. The Metropolis Criterion for our ResEx strategy is stated in the equation.

the tuning parameter λ between two replicas i and j so that the acceptance probability of an exchange is (Figure 3.3):

$$P < min \left\{ 1, \frac{\exp\left[\frac{-E_{\lambda j}(\phi_a,\, x_a)}{k_B T} - \frac{-E_{\lambda i}(\phi_b,\, x_b)}{k_B T}\right]}{\exp\left[\frac{-E_{\lambda i}(\phi_a,\, x_a)}{k_B T} - \frac{-E_{\lambda j}(\phi_b,\, x_b)}{k_B T}\right]} \right\} \tag{3.4}$$

This is illustrated in Figure 3.3. Further, in our MC simulation, to ensure that all moves implemented within the replica are equipped with the mixed potential, I created a `MixedMonteCarlo` mover that utilizes the pose, the dual, and the tuning parameter of the replica to evaluate MC moves. This approach of mixed resolutions allows us to implement multiple replicas enabling a smooth transition between CG and AA modes.

3.4.3 Mixed resolution energy distributions overlap allowing successful MC exchanges

To evaluate the efficiency of this mixed resolution approach, I next tested the energy landscapes of the replicas to check whether there is an overlap to exchange conformations with its neighbours. Higher exchange rates would be possible if the distribution of the sampled decoys overlap with each other. To determine the energy distributions, I simulated docking on protein target 1LFD[28] with 20 independent replicas. Each of the 20 replicas used a fixed tuning parameter for the entire simulation. Here, the aim was to assess if the energy landscape sampled independently by these replicas overlap with the adjacent replicas.

Figure 3.4 compares the probability densities of the replicas. While the pure CG and AA landscapes are relatively distinct, the intermediate replicas connect the landscapes surprisingly well. An interesting aspect of the distributions is highlighted at

the tail end of the AA replicas (*darker shades*). Most of the near-AA replicas overlap exceedingly well, with roughly 70-80% overlap (area under the distribution). However, as we start to move towards the CG replicas, the overlapped area in the distributions decrease. The incremental coarse-graining is demonstrated by broader distributions at the CG tail ends (*lighter shades*). Thus, although the mixed resolution replicas have overlapping landscapes, the overlap varies. Instead of setting equal increments to tuning factors (*i.e.* λ from 0 to 1 with 0.05 increments), one might create more diversity by modulating the increments such that there are larger increments in λ on the AA end of the replicas and smaller increments near the CG end. This, however, is out of scope of this study and will be discussed in our future work.

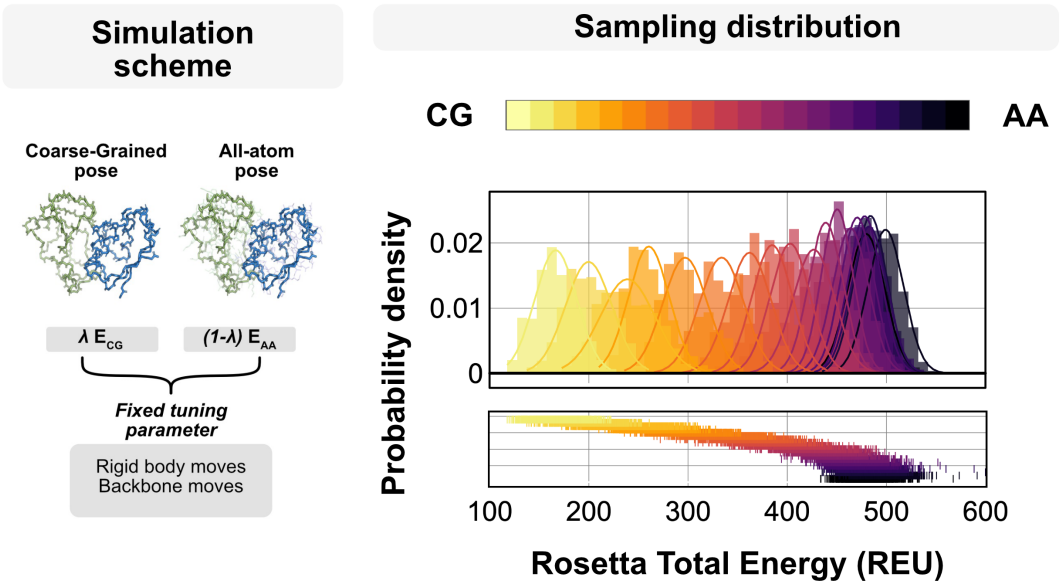


Figure 3.4: Energy landscape of mixed resolution replicas (*left*) Simulation scheme to generate parallel trajectories with the fixed tuning parameters. (*right*) Sampling distribution of the energy landscape obtained from each replica with fixed tuning parameter. Probability density (y-axis) against the Rosetta Total Energy (REU) for each replica with colors highlighting CG replicas (*lighter shade*) to AA replicas (*darker shade*).

3.4.4 Resolution exchange swaps configurations for enhanced sampling

Building on prior knowledge, I developed ResEx for protein-protein docking (Figure 3.5). The ResEx protocol starts from an initial, randomly oriented docking structure. This input pose is then split into a AA pose and a deep-copy is stored in centroid representation as a dual. Both the pose and dual are passed to 20 parallel replicas with tuning parameter λ extending from 0 (completely AA) to 1 (completely CG). The tuning parameter determines resolutions of the intermediate replicas to calculate the net energy potential for exchanges. Each replica undergoes 1,000 MC trials of rigid-body, backbone (Rosetta Backrub[29] and BalancedKIC[30]), and side-chain moves are performed with 2:1:1 probability of sampling. After 1,000 MC trials, a swap is attempted between neighbouring replicas, *i.e.* for replica i, a swap is attempted either with the preceding $(i-1)$ or following $(i+1)$ replicas. I interpret that higher replicas with low resolution (CG) will accept backbone moves that would otherwise be rejected at lower replicas with high resolution (AA). Further, lower replicas will penalize sticky sites at non-native interfaces. For successful simulations, $10^6 - 10^7$ MC steps are necessary and could generate up to 20,000 structures per protein targets. These structures can be clustered and refined to obtain the top-performing decoys with compact interfaces.

3.4.5 Centroid replicas capture large-scale conformational changes while all-atom replicas prevent entrapment in false positive sticky sites

Each replica performs both rigid-body and backbone moves. I wanted to investigate if the replicas show a propensity towards specific type of move based on the score-function. To test this, I obtained the statistics of the trials and extracted the acceptance rates for the rigid-body moves and the backbone moves. Figure 3.6 shows the

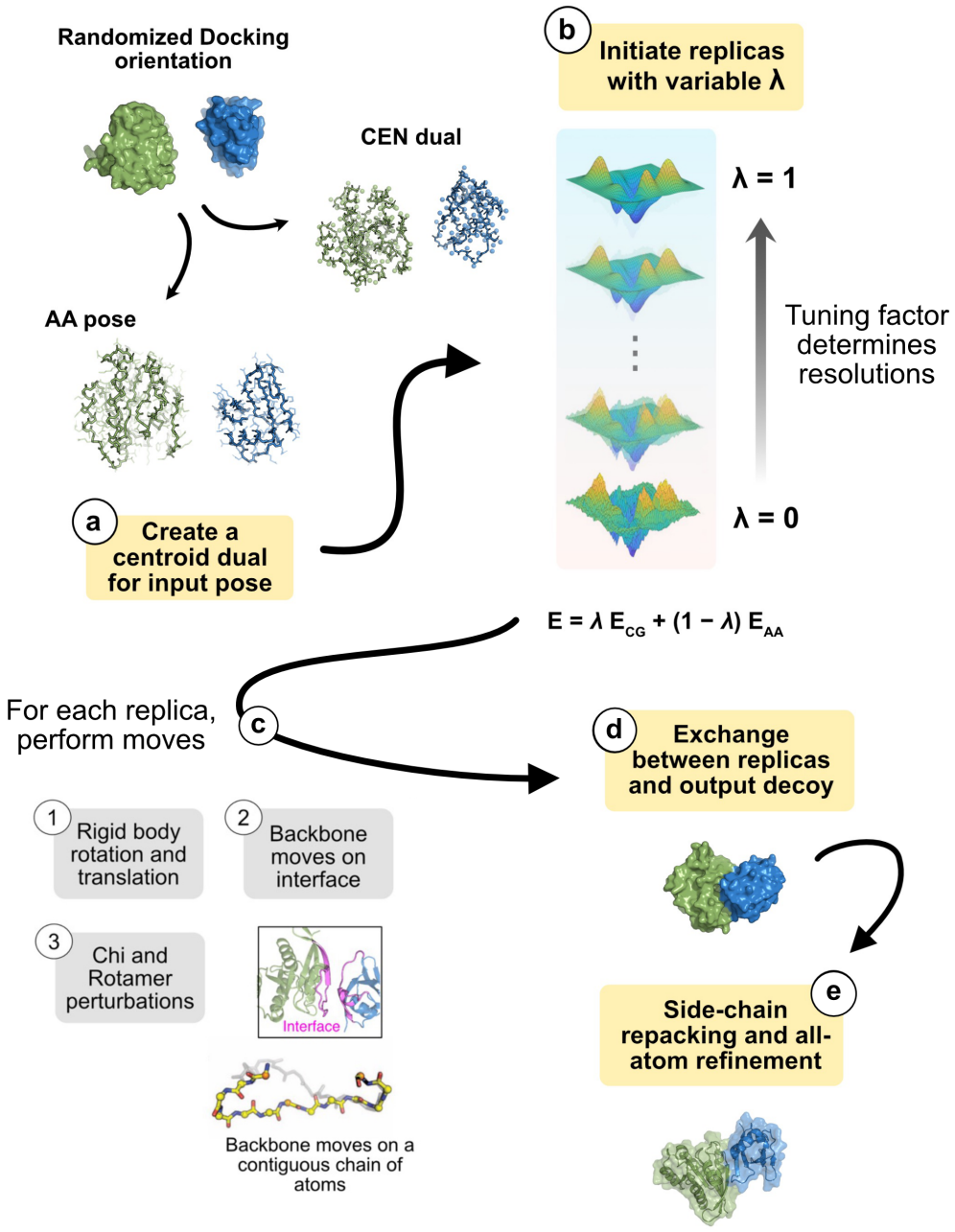


Figure 3.5: Overview of the ResEx docking protocol *Caption follows on next page*

Figure 3.5: Starting from the initial docking pose, the incoming pose is split to an all-atom pose and a centroid dual (a). The pose and dual is passed along to 20 replicas with tuning parameter, λ, ranging from 0 to 1. The mixed energy potential is detemrined and used for exchanges (b). Each replica performs rigid body moves (rotation-translation), backbone moves (Rosetta backrub and balancedKIC), and side-chain moves for each MC trial, followed by exchange between replicas after every 1000^{th} MC trials. For every exchange the Metropolis Criterion was used determining acceptance probabilities. A single trajectory with 20 replicas for 10^6-10^7 MC trial steps and produces ≈20,000 candidate structures. Lastly, all produced structures undergo an all-atom refinement comprising of side-chain packing, smaller rigid-body moves and energy minimization to output final docked structures.

acceptance rates of the moves across the replicas. For rigid-body moves, the near-AA replicas have higher acceptance rates, but the rates drop to less than 0.1 for near-CG replicas (*top-left*). Intuitively this would imply that CG replicas are stuck in a minima of the protein energy landscape and the rigid-body moves performed are not large enough to escape the minima. To assess this hypothesis, I plotted the evolution of Lig-RMSD of the two replicas (marked *a* and *b* denoting near-AA and near-CG replicas respectively) across time while comparing it to the ligand RMSD. The near-AA replica (*blue*) observes a noisy pattern with lig-RMSD varying in a range of roughly 10 Å, consistent with the statistics of rigid-body acceptance rate (being higher for the near-AA replica). On the other hand, the near-CG replica (*yellow*) shows a step-like evolution over time *i.e.* there are broad regions with hardly any perturbations in lig-RMSD, implying that the protein partner is stuck at a binding site with lower acceptance rate of rigid moves. Thus, in these resolution exchange simulations, near-AA replicas can escape from sticky sites on protein surfaces, but CG replicas do not.

I next evaluated the acceptance rates for the backbone moves. As evident from the backbone moves acceptance rate increase dramatically from AA to CG replicas, consistent with the efficient low-resolution sampling of backbones (as demonstrated

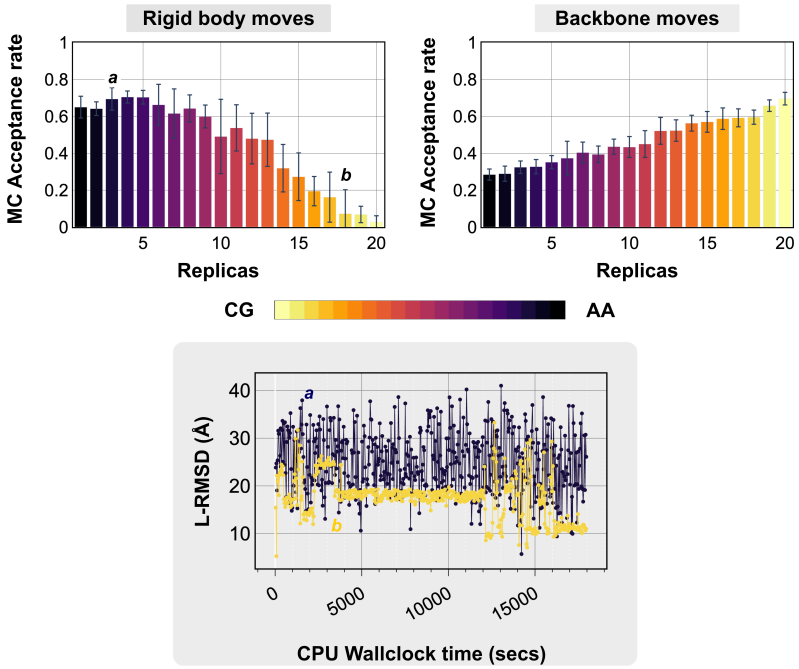


Figure 3.6: Trial statistics for rigid and backbone moves (*top*) MC acceptance rates across replicas for rigid-body moves and backbone moves respectively. Two replicas from rigid-body plot (a: near-AA replica, and b: near-CG replica) were used to determine the evolution across lig-RMSD. (*bottom*) Lig-RMSD (Å) against CPU wallclock time (secs) demonstrates the two replicas.

in Figure 3.6). Further, it suggests that the near-CG replicas in non-native sites allow induced-fit backbone moves to improve the energetics of relatively poor interfaces and thereby biasing the search in a potentially false local minima. Thus in ResEx, near-AA replicas penalize non-native interfaces while near-CG replicas aggressively sample backbones for conformational diversity.

3.4.6 ResEx demonstrates improved performance over prior docking techniques

Next, I tested the ResEx strategy for protein docking was tested on a small benchmark set of 9 protein targets spanning rigid(1), medium(4), and difficult(4) categories of flexibility. For each of the targets, I initiated global docking simulations with ResEx as illustrated in Figure 3.5. For all the protein targets, the evolution of the sampled decoys in terms of interface score (I_sc) and interface-RMSD (Irms) is showcased in Figure 3.7. Each sampled decoy is refined, with lower (AA) replicas highlighted in *darker shades* and higher (CG) replicas highlighted in *lighter shades*. For most of the targets, by sampling in the global protein energy landscape, the protocol identifies an energy funnel to the near-native interface. Further, most of the local minima created in the plots are originating from the near-CG replicas, highlight the impact of induced-fit owing to better backbone sampling in low resolutions.

To obtain a quantitative trend, I next compared the interface RMSD (Irms) and fraction of the native-like contacts (f_{nat}) across the protein targets with ReplicaDock2.0[27] and RosettaDock4.0[31], our prior methods for protein-protein docking. Figure 3.7 shows the results for the targets arranged in an increasing order of flexibility. In 8 out of 9 cases, ResEx docking predictions have lower interface RMS than the other methods. More importantly, the fraction of native-like contacts captured demonstrate

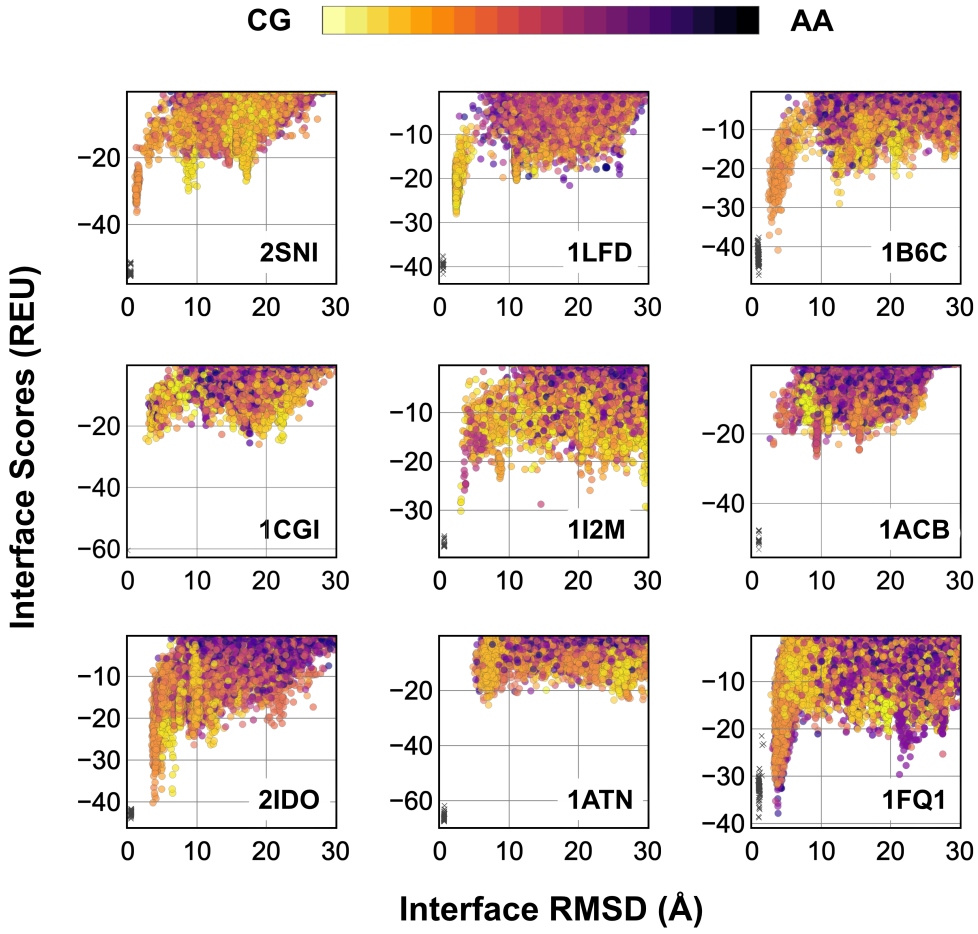


Figure 3.7: Benchmarking protein targets Scatter plot of interface score(REU) and interface-RMSD(Å) for the nine benchmark targets. Each sampled decoy is colored based on the replica it is derived from. Flexibility increases from *top* to *bottom*.

that we successfully recapitulate the native-like interface. These are promising results and highlight the efficacy of ResEx as a new sampling strategy for protein-protein docking.

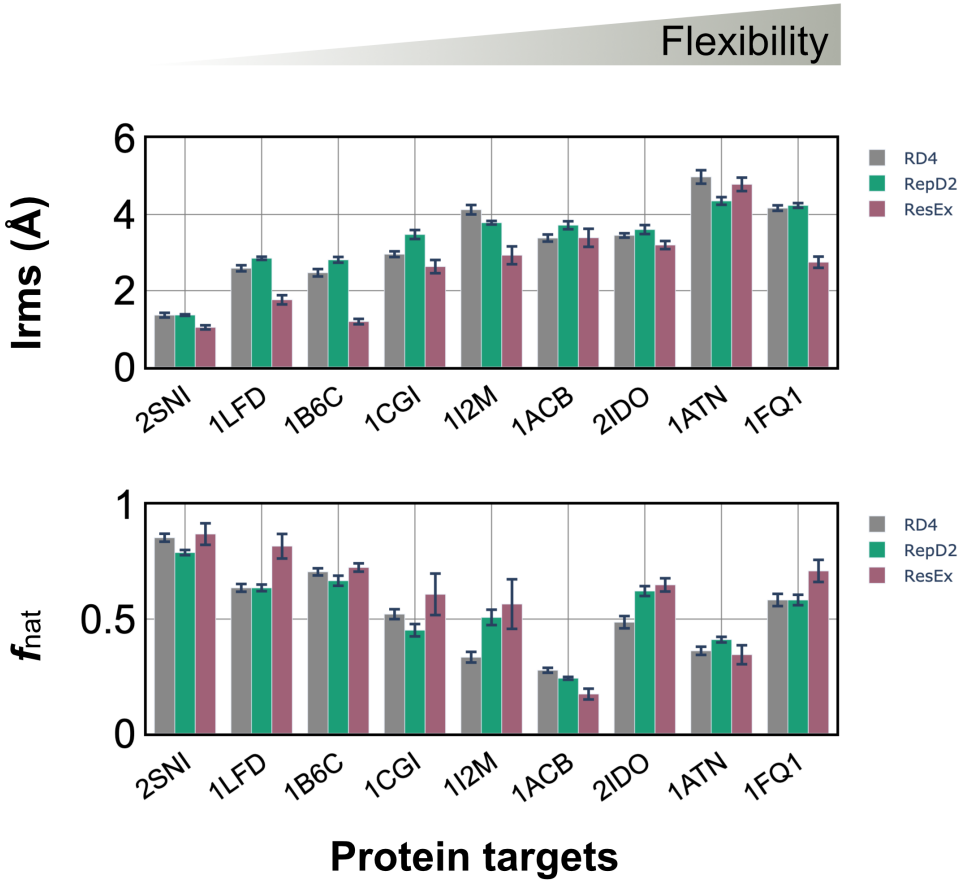


Figure 3.8: Evaluation metrics for the benchmark targets. Interface RMSD (Irms) and fraction of native-like contacts (f_{nat}) for 9 benchmark protein targets. Comparisons are made against our prior work involving ReplicaDock2 and RosettaDock4.

3.5 Discussion and conclusions

In this work, I presented a novel replica exchange approach that combines coarse-grained and all-atom resolutions for improved sampling and scoring in protein-protein docking. The new strategy, resolution exchange (ResEx), is motivated by the efficiency of CG modes to capture better backbones and AA modes to robustly evaluate protein interfaces. Our mixed resolution strategy employs an energy function combined from CG and AA energy functions, modulated via a tuning parameter, to generate intermediate resolutions. Unlike prior methods equipped with only atomistic and CG resolutions, this strategy successfully utilizes mixed resolutions for larger protein systems. The results demonstrate that a ResEx strategy with induced-fit backbone sampling results in better exploration of protein energy landscape and successfully docks protein partners.

One of the key factors in ResEx is the implementation of the tuning factor, λ, that transforms resolution replica exchange into hamiltonian replica exchange by manipulating the energy potentials rather than utilizing ideal intermediate resolutions (partially CG and partially AA). By implementing multiple replicas (20 replicas in this work), parallel simulation trajectories provided a massive boost to sampling in the conformational landscape (20,000 structures generated by ResEx as opposed to 6000 by T-REMC replicas). The diffusion of CG and AA models across trajectories allows exploration of regions that either mode is unable to capture independently in reasonably short simulation timescales. Similar to our prior temperature-REMC approach, ReplicaDock2.0, ResEx requires 5-7 hrs on a computational cluster with 24 cores (roughly 120-170 CPU total wallclock hrs) but samples relatively three-fold more candidate docked structures from each trajectory. For our nine test targets,

I have found that simulating one trajectory with 20 replicas is optimum to create sampling diversity while maintaining computational efficiency of the protocol.

Finally, with multiscale methodologies, the goal is to enable simulations of higher-order assemblies and large, complex biological mechanisms within physically relevant time scales. ResEx serves as an approach to integrate all-atom and coarse-grained modes to explore the uncharted regions of the protein energy landscape. This study provides a proof-of-principle of the application of ResEx to protein docking while demonstrating that mixed resolutions provide a feasible alternative to completely AA models. Further, by incorporating temperature as an adjacent ladder across the replicas, a fused temperature-resolution exchange strategy could be tested. In sum, the ResEx strategy demonstrates promise in capturing the dynamic behaviour of protein complexes. Integrating metadynamics and machine learning approaches with ResEx can be a potential future direction to incorporate dynamics in structure prediction methods.

3.6 Methods

3.6.1 Resolution exchange (ResEx) algorithm for protein-protein docking

The ResEx algorithm for protein docking employs the mixed resolution strategy described prior with the `MixedMonteCarlo` mover in Rosetta. I built ResEx over the temperature replica exchange MC protocol discussed in the last chapter such that majority of the movers for exchange are inherited from the `ThermodynamicMover` parent class in Rosetta for canonical sampling with detailed balance. Starting from an initial configuration, 20 parallel replicas are simulated with replica exchange swaps attempted after every 1000 MC steps. The trial steps for the entire simulation are

pre-determined. For global docking, I found the optimum number of trial steps to be $10^6 - 10^7$ depending on the size of the protein system. The tuning parameter λ is set to be 0 for the lowest replica (high resolution/AA mode) and 1 for the highest replica (low resolution/CG mode) with incremental increases of 0.05 to transition from an AA phase to a CG phase. Exchange attempts are evaluated with the Metropolis criterion as defined in Equation 3.4. To sample the protein conformational space within each replica, rigid body rotational (4°) and translation (2 Å) moves, backbone moves (Rosetta Backrub and BalancedKIC), and side-chain moves are performed while observing detailed balance. After a successful swap, between replicas i and j, the poses exchange the tuning parameter. After the ResEx simulation, all generated structures are refined with side-chain packing and minimization to output docked decoys.

3.6.2 Benchmark evaluation and metrics

To assess our ResEx strategy, I created a small benchmark comprising of rigid, moderate and flexible targets from the Dockground 5.5[15] benchmark set. Each docking simulation was initiated from the unbound monomers with the bound structure as the reference. As stated in CAPRI[32], the performance was evaluated with interface RMSD (I-rms) and fraction of native-like contacts (f_{nat}). Further, acceptance rates of MC moves for every individual replicas were estimated and compare to elucidate the trends in rigid and backbone moves. Going ahead, I intend to benchmark our ResEx algorithm on the entire Dockground benchmark set with 245 protein targets for robust analysis and validation of the protocol.

References

1. Roel-Touris, J. & Bonvin, A. M. Coarse-grained (hybrid) integrative modeling of biomolecular interactions. *Computational and Structural Biotechnology Journal* **18,** 1182–1190. ISSN: 20010370. https://doi.org/10.1016/j.csbj.2020.05.002 (2020).
2. Liu, P., Kim, B., Friesner, R. A. & Berne, B. J. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proceedings of the National Academy of Sciences* **102,** 13749–13754. ISSN: 0027-8424. https://www.pnas.org/content/102/39/13749 (39 2005).
3. Kästner, J. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1,** 932–942. ISSN: 17590884 (6 2011).
4. Limongelli, V., Bonomi, M. & Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 6358–6363. ISSN: 10916490 (16 2013).
5. Camacho, C. J. & Vajda, S. Protein docking along smooth association pathways. *Proceedings of the National Academy of Sciences* **98,** 10636–10641. https://doi.org/10.1073/pnas.181147798 (19 2001).
6. Zacharias, M. ATTRACT: Protein-protein docking in CAPRI using a reduced protein model. *Proteins: Structure, Function, and Bioinformatics* **60,** 252–256. ISSN: 08873585. http://doi.wiley.com/10.1002/prot.20566 (2 2005).
7. Lyskov, S. & Gray, J. J. The RosettaDock server for local protein-protein docking. *Nucleic acids research* **36,** 233–238. ISSN: 13624962 (Web Server issue 2008).
8. Kuroda, D. & Gray, J. Pushing the Backbone in Protein-Protein Docking. *Structure* **24.** ISSN: 18784186 (10 2016).
9. Christen, M. & van Gunsteren, W. F. Multigraining: An algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. *The Journal of Chemical Physics* **124,** 154106. ISSN: 0021-9606. https://doi.org/10.1063/1.2187488 (15 2006).

10. Kar, P. & Feig, M. Hybrid All-Atom/Coarse-Grained Simulations of Proteins by Direct Coupling of CHARMM and PRIMO Force Fields. *Journal of Chemical Theory and Computation* **13,** 5753–5765. ISSN: 15499626 (11 2017).

11. Lyman, E., Ytreberg, F. M. & Zuckerman, D. M. Resolution exchange simulation. *Physical Review Letters* **96,** 1–4. ISSN: 10797114 (2 2006).

12. Liu, P. & Voth, G. A. Smart resolution replica exchange: An efficient algorithm for exploring complex energy landscapes. *Journal of Chemical Physics* **126.** ISSN: 00219606 (4 2007).

13. Lyman, E. & Zuckerman, D. M. Resolution exchange simulation with incremental coarsening. *Journal of Chemical Theory and Computation* **2,** 656–666. ISSN: 15499618 (3 2006).

14. Liu, P., Shi, Q., Lyman, E. & Voth, G. A. Reconstructing atomistic detail for coarse-grained models with resolution exchange. *Journal of Chemical Physics* **129.** ISSN: 00219606 (11 2008).

15. Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., Chaleil, R., Jiménez-García, B., Bates, P. A., Fernandez-Recio, J., Bonvin, A. M. & Weng, Z. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology* **427,** 3031–3041. ISSN: 10898638. http://dx.doi.org/10.1016/j.jmb.2015.07.016 (19 2015).

16. Heath, A. P., Kavraki, L. E. & Clementi, C. From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. *Proteins: Structure, Function, and Bioinformatics* **68,** 646–661. ISSN: 0887-3585. https://doi.org/10.1002/prot.21371 (3 2007).

17. Van der Spoel, D. & Seibert, M. M. Protein Folding Kinetics and Thermodynamics from Atomistic Simulations. *Physical Review Letters* **96,** 238102. https://link.aps.org/doi/10.1103/PhysRevLett.96.238102 (23 2006).

18. Liwo, A., Baranowski, M., Czaplewski, C., Gołaś, E., He, Y., Jagieła, D., Krupa, P., Maciejczyk, M., Makowski, M., Mozolewska, M. A., Niadzvedtski, A., Ołdziej, S., Scheraga, H. A., Sieradzan, A. K., Slusarz, R., Wirecki, T., Yin, Y. & Zaborowski, B. A unified coarse-grained model of biological macromolecules based on mean-field multipole-multipole interactions. eng. *Journal of Molecular Modeling* **20,** 2306. ISSN: 0948-5023 (Electronic) (2014).

19. Koliński, A. Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica* **51,** 34971. ISSN: 0001-527X. http://www.ncbi.nlm.nih.gov/pubmed/15218533 (2 2004).

20. Wang, C., Bradley, P. & Baker, D. Protein–Protein Docking with Backbone Flexibility. *Journal of Molecular Biology* **373,** 503–519. ISSN: 00222836 (2007).

21. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology* **487,** 545–574. ISSN: 00766879 (C 2011).

22. Harmalkar, A. & Gray, J. J. Advances to tackle backbone flexibility in protein docking. *Current Opinion in Structural Biology* **67,** 178–186. ISSN: 1879033X. http://arxiv.org/abs/2010.07455 (2020).

23. Brooks, B. R., III, C. L. B., Jr., A. D. M., Nilsson, L, Petrella, R. J., Roux, B, Won, Y, Archontis, G, Bartels, C, Boresch, S, Caflisch, A, Caves, L, Cui, Q, Dinner, A. R., Feig, M, Fischer, S, Gao, J, Hodoscek, M, Im, W, Kuczera, K, Lazaridis, T, Ma, J, Ovchinnikov, V, Paci, E, Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M, Tidor, B, Venable, R. M., Woodcock, H. L., Wu, X, Yang, W, York, D. M. & Karplus, M. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **30,** 1545–1614. ISSN: 0192-8651. https://doi.org/10.1002/jcc.21287 (10 2009).

24. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *The Journal of Physical Chemistry B* **111,** 7812–7824. ISSN: 1520-6106. https://doi.org/10.1021/jp071097f (27 2007).

25. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T. & Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13,** 3031–3048. ISSN: 15499626 (6 2017).

26. Fallas, J. A., Ueda, G., Sheffler, W., Nguyen, V., McNamara, D. E., Sankaran, B., Pereira, J. H., Parmeggiani, F., Brunette, T. J., Cascio, D., Yeates, T. R., Zwart, P. & Baker, D. Computational design of self-assembling cyclic protein homo-oligomers. *Nature chemistry* **9,** 353–360. ISSN: 1755-4349 (Electronic) (4 2017).

27. Harmalkar, A., Mahajan, S. P. & Gray, J. J. Induced fit with replica exchange improves protein complex structure prediction. *PLOS Computational Biology* **18,** 1–21. https://doi.org/10.1371/journal.pcbi.1010124 (6 2022).

28. Lan, H., Franz, H., Steven, M. & Sung-Hou, K. Structural basis for the interaction of Ras with RalGDS. *Nature Structural Biology* **5,** 422–426. http://link.springer.com/10.1007/978-1-62703-429-6 (6 1998).

29. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of molecular biology* **380,** 742–756. ISSN: 1089-8638 (Electronic) (4 2008).

30. Stein, A. & Kortemme, T. Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLOS ONE* **8,** 1–13. https://doi.org/10.1371/journal.pone.0063090 (5 2013).

31. Marze, N. A., Burman, S. S. R., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics* **34,** 3461–3469. ISSN: 14602059 (20 2018).

32. Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I. & Wodak, S. J. CAPRI: A critical assessment of PRedicted interactions. *Proteins: Structure, Function and Genetics* **52,** 2–9. ISSN: 08873585 (1 2003).

Chapter 4

Critical Assessment of Prediction of Interactions : a global community-wide intiative for protein docking

> Given the three-dimensional structures of any two proteins, is it possible to predict whether they will associate, and if so, in what way?
>
> Michael Connolly, 1986

4.1 Overview

The Critical Assessment of PRediction of Interactions (CAPRI) is a community-wide, global assessment of computational tools to predict biomolecular complexes and interactions. It serves as a blind challenge to evaluate the performance of state-of-the-art docking methods, develop innovative strategies, and further our understanding in improving the computational modeling of protein-protein interactions. Often linked with CASP (Critical Assessment of Structure Prediction), the sequence-to-structure prediction challenge, CAPRI primarily differs in its focus towards protein

assemblies, evaluating protein-protein interfaces and characterizing performance based on docking metrics. Each CAPRI round brings an esoteric challenge with protein targets categorized from easy to difficult. Easy targets consist of cases with available *a priori* information, either via homologous templates or known binding sites over protein surfaces. On other hand, difficult targets have no *a priori* information available and exhibit higher degree of binding-induced conformational changes. Throughout my PhD, I have participated in seven CAPRI rounds from 2019-2023 comprising 45 targets (Table 4.1). In this chapter, I highlight several interesting CAPRI targets and address the major challenges in blind prediction of protein-protein interactions. I will also discuss the impact of AlphaFold - a deep learning approach that revolutionized CASP - and its application for the protein docking problem, thereby setting the premise for the development of a protein docking pipeline in the chapter ahead.

4.2 Introduction

The advent of high-throughput sequencing has exploded the availability of genomic data and higher-order definitions of the protein interaction landscapes.[1,2] Development in structural biology techniques, such as cryo-EM, has further led to the characterization of three-dimensional (3D) structures of these interacting proteins. Despite these advances, the 3D structures of protein complexes deposited in the protein data bank (PDB) are relatively scarce.[3] Computational approaches aim to model these structures and explore the uncharted landscape of protein association. As computational approaches are often assessed on a set of known (published) benchmark targets, *i.e.* proteins where both the unbound monomers and the bound complex

structure are known, CAPRI provides the developers with an opportunity to assess their tools on blind targets prior to their release to the PDB.[4,5] The essentially 'blind' assessment in CAPRI has therefore become a gold-standard evaluation of docking performance, and it has pushed the field forward by highlighting the limitations in docking algorithms.

Each CAPRI round constitutes of multiple targets with an underlying theme. Prior rounds were often limited to protein assemblies, but lately challenges have expanded to include oligosaccharides, nucleic acids, and peptides in association with proteins.[5] Every target presents a prediction challenge and a scoring challenge. The predictors aim to model the protein complex structure by depositing five top models along with five alternative decoy models. The scorers are then provided with a curated set of all predicted models (100 from each group, including the top five models) with the task of discriminating native-like models from the set. On one hand, where predictors are evaluated on their ability to sample the near-native structure, the scorers are evaluated on their scoring efficacy. Assessors characterize each predicted model as high, medium, acceptable or incorrect based on the DockQ score of the model.[6] The DockQ score is comprised of three metrics: the root-mean-square-deviation of the backbone and side-chain atoms of the interface with reference to the bound interface (Irms), the root-mean-square-deviation of the backbone atoms of the ligand with reference to the native ligand when the model is superimposed over the receptor (Lrms), and the fraction of native-like contacts recovered at the interface (f_{nat}).

The Gray lab has been a longstanding participant in CAPRI evaluating our Monte-Carlo minimization (MCM) based docking methods, primarily RosettaDock and SnugDock.[7–13] During my tenure as a Gray lab CAPRI team participant, I fused our

conventional docking approaches with an exhaustive global docking search and aggressive backbone sampling (ReplicaDock 2.0).[14,15] Rounds 46 through 50 involved multimeric assemblies with limited structural information about respective domains. The SARS-Cov2 pandemic in 2020 resulted in the CAPRI initiative to model the complexes of ACE2-receptor with SARS-Cov2 domains in round 51 and 52. In the midst of the global pandemic, these rounds presented a real-life scenario benefiting from computational prediction. Owing to time constraints and our limited expertise in handling protein-nucleic acid complexes, we did not participate in round 53. Round 54, the latest edition of CAPRI in conjunction with CASP15, presented the docking challenge in an era post-AlphaFold, with easier availability of monomeric structures[16] Here, I examine the nuances over multiple targets, summarize our methodology and highlight our performance for available assessments.

4.3 AlphaFold2: the disruptive breakthrough in structural biology

Critical Assessment of protein Structure Prediction (CASP), the *ab initio* structure prediction challenge, enables the prediction of protein structures from sequences alone. Over the past few CASPs, many deep learning algorithms exploited structure and sequence information to improve prediction. CASP14 unveiled one of the disruptive breakthroughs in structural biology that dramatically changed the field. Google's artificial intelligence (AI) subsidiary, Deepmind, presented AlphaFold2 (AF2), a deep learning architecture trained on all available sequences and structures from the protein data bank (PDB) to predict protein folding.[16] AF2 leveraged physical information about protein structure along with multiple-sequence alignments to obtain a score of 90 out of 100 on the Global Distance Test (GDT) assessment metric (Figure 4.1A).

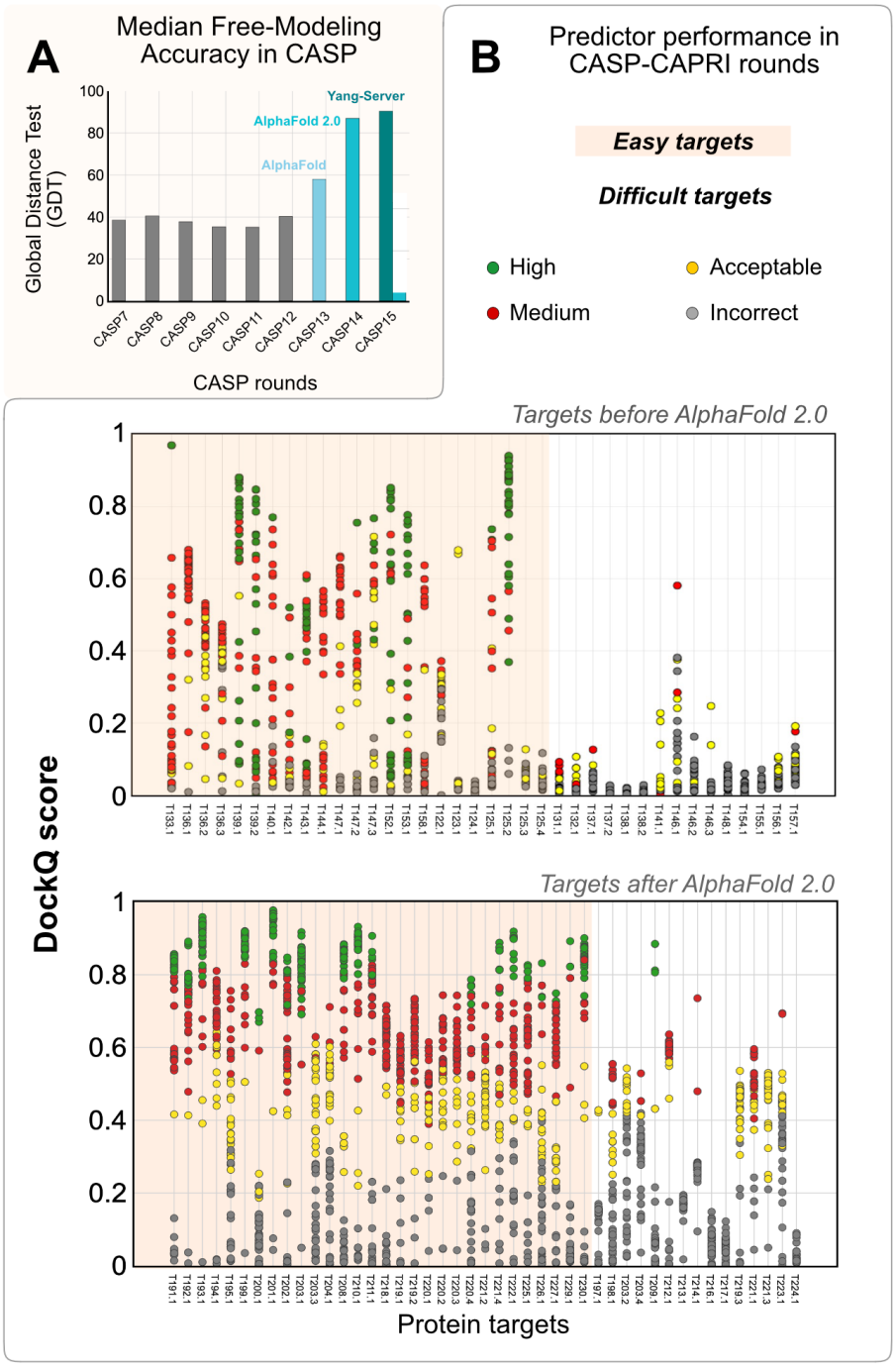


Figure 4.1: Performance of predictors in protein structure prediction challenges. *caption follows on next page*

Figure 4.1: (A) Median accuracy of predictions in the free modeling category (sequence-to-structure) for the best team in CASP. AlphaFold made a disruptive breakthrough in CASP14. CASP15 had Yang server as the best performer. Yang server incorporated AlphaFold with customized MSAs, resulting in better performance. (B) Distribution of the DockQ score for the best model submitted by each predictor group for each individual target interface in rounds 50 (*top*) and 54 (*bottom*) (x-axis). DockQ measures acombination of intermolecular residue-residue contacts, interface RMSD, and ligand RMSD on a scale of 0 (incorrect) to 1(matching the experimental structure). DockQ scores are color-coded by CAPRI model quality ranking: green, high; red, medium; yellow, acceptable; gray, incorrect. (*top*) targets prior to AlphaFold commercial release in 2021.[4,5] (*bottom*) targets in recent CASP15-CAPRI. AlphaFold was openly accessible to all predictor groups for these targets. Data adapted from Harmalkar *et al.*[17] and graciously provided by Mark Lensink.

Essentially, AF2 took a protein sequence as an input and generated a protein structure as an output, highlighting the confidence of each residue via pLDDT (predicted Local Distance Difference Test) and the overall confidence in prediction via PAE (predicted alignment error). With an open-access code available for academic and commercial groups alike, AF2 changed the protein sequence-to-structure space considerably.

Following this release, many groups later employed AF2 for the protein complex prediction task. This re-utilization of AF2 ranged from adding an gap between chains in the amino acid sequence being fed to AF2 (*i.e.* AlphaFold-Gap[18,19]), or incorporating a linker by adding glycines between chains[20,21], all the way up to feeding paired multiple-sequence alignments for predicting protein complexes and multimeric assemblies.[18,22] Owing to these academic developments, Deepmind rushed to release AlphaFold-multimer[19], an updated version of AF2 trained on more than one protein chains that aimed to improve performance over multimeric inputs. However, the lack of blind assessments challenged the accuracy of AF2-multimer for predicting protein assemblies.

The 15th edition of CASP in Summer 2022 thus served as the primary assessment

of AF2-multimer. Many groups built over AlphaFold's existing architecture to diversify model predictions at inference by tuning dropout[23], building better multiple sequence alignment (MSAs) or defining structural templates[24]. AF2 models were no longer the top performers, however, all top performers employed AF2. The joint CASP15-CAPRI initiative embarked upon the assessment for protein assemblies and interactions, and although better than earlier CASPs, many challenges of structural biology remained unsolved (Figure 4.1B). Predictors lacked accuracy in highly flexible complexes, predicting ensembles of conformations, antibody-antigen complexes, and large multimeric assemblies. We participated in CASP15 employing AF2 predicted models as baselines, often refining these structures prior to submission, resulting in a mediocre performance. However, over the course of the competition and assessment, I identified that AF2 models, although inaccurate, can reveal underlying information about flexibility and docking accuracy. With this premise, I will next discuss the assemblies in CAPRI over the years and lay the foundation for building a docking pipeline over the progress of AlphaFold.

4.4 Flexibility still hampers docking accuracy

Binding-induced conformational changes have long confounded state-of-the-art docking algorithms[5,17] Prior to my initiation into the Gray lab CAPRI team, predictors had tackled multiple protein targets with varying degrees of flexibility. Out of the 38 protein-protein targets curated in Figure 4.1B (*top panel*), predictors achieved high-quality structures (DockQ ≥ 0.8) for all 23 easy targets. Here, assessors defined 'easy' as those with little-to-no backbone motion (unbound-to-bound Cα root mean square deviation ($RMSD_{UB}$) of less than 1.5 Å. The remaining 15 targets were categorized

Table 4.1: Summary of CAPRI targets, Rounds 46-54 The table lists the round, target number, name of the complex, the nature of the challenge, stoichiometry of target and the methods used to model the complex. The results for all targets except the ones in Round 50(CASP14-CAPRI experiment) and 54(CASP15-CAPRI experiment) are yet to be announced. Targets are classified based on four categories: multimers or heteromers, homodimers or heterodimers, and antibody-antigen targets.

Round	Target	Complex	Classification	Stoichiometry	Method
47	160	S-layer protein from Bacillus anthracis	Multimer	A1B1C1D1E1F1	RosettaDock + FloppyTail + Loop remodeling
48	161	RnlA homodimer	Multimer	A2	FloppyTail + SymDock
48	162	RnlA-RnlB complex	Multimer	A1B1	RosettaDock + FloppyTail
49	163	SYCE2/TEX12 helical components of the human synaptonemal complex	Heteromer	A2B2	RosettaDock + SymDock
50	164	Structural maintenance of chromosomes flexible hinge domain-containing protein	Homodimer	A2	SymDock
50	165	Monoclonal antibody bound to varicella-zoster virus glycoprotein gB	Heteromer	A3H3L3	ReplicaDock
51	182	SARS-CoV-2 Nsp15 bound to Nuclear Transport Factor (NFT2)	Heteromer	A1B2	SymDock + ReplicaDock
51	183	SARS-CoV-2 Nsp8 bound to human exosome complex component RRP43	Heterodimer	A1B1	ReplicaDock + RosettaDock
51	184	SARS-CoV-2 Nsp7 bound to human transforming protein RhoA	Heterodimer	A1B1	ReplicaDock + RosettaDock
54	191	YscX-YscY complex from Yersinia enterocolitica	Heterodimer	A1B1	ReplicaDock
54	192	D180A mutant of isocyanide hydratase	Homodimer	A2	SymDock
54	193	Wildtype isocyanide hydratase	Homodimer	A2	SymDock

Round	Target	Complex	Classification	Stoichiometry	Method
54	194	GP2 bacteriophage protein with role in phage replication	Homodimer	A2	SymDock
54	195	Human peripheral membrane protein that forms a higher order oligomer	Multimer	A16	SymDock
54	197	Bacterial condensin-like complex to prevent plasmid transformation	Homodimer	A2	SymDock
54	198	Likely viral receptor binding domain	Homodimer	A2	SymDock
54	199	Dimeric N-acetyltransfarase protein	Homodimer	A2	SymDock
54	200	Phage protein with bacterial membrane receptor	Heterodimer	A1B1	ReplicaDock
54	201	Antibiotic biosynthesis monooxygenase (ABM) domain protein	Multimer	A6	SymDock
54	202	Type 6 secretion system (T6SS) lipase effector	Heterodimer	A1B1	ReplicaDock
54	203	LINC complex that connects cytoskeleton and nuclear components across the nuclear membrane	Multimer	A9B3	ReplicaDock + SymDock
54	203	LINC complex that connects cytoskeleton and nuclear components across the nuclear membrane	Multimer	A9B3	ReplicaDock + SymDock
54	203	LINC complex that connects cytoskeleton and nuclear components across the nuclear membrane	Multimer	A9B3	ReplicaDock + SymDock

Round	Target	Complex	Classification	Stoichiometry	Method
54	203	LINC complex that connects cytoskeleton and nuclear components across the nuclear membrane	Multimer	A9B3	ReplicaDock + SymDock
54	205	Complex of mammalian CNPase phosphodiesterase domain with nanobody	Antibody	A1B1	IgFold + ReplicaDock
54	206	Complex of mammalian CNPase phosphodiesterase domain with nanobody	Antibody	A1B1	IgFold + ReplicaDock
54	207	Complex of mammalian CNPase phosphodiesterase domain with nanobody	Antibody	A1B1	IgFold + ReplicaDock
54	208	Complex of mammalian CNPase phosphodiesterase domain with nanobody	Antibody	A1B1	ReplicaDock
54	210	probable transcriptional regulator WhiB6 (P9WF37) from Mycobacterium tuberculosis	Heterodimer	A1B1	ReplicaDock
54	211	Endonuclease domain (245-543) of human EEPD1 involved in homologous recombination repair.	Homodimer	A2	SymDock
54	212	Endoplasmic Reticulum Associated Degradation complex	Heterodimer	A1B1	ReplicaDock
54	213	Ancient protein reconstruction-type A	Heterodimer	A1B1	ReplicaDock
54	214	Ancient protein reconstruction-type B	Heterodimer	A1B1	ReplicaDock

Round	Target	Complex	Classification	Stoichiometry	Method
54	216	Ab against Nucleocapsid phosphoprotein (SARS-CoV-2 N-terminal domain)	Antibody	A1B1C1	IgFold + ReplicaDock
54	217	Ab against Nucleocapsid phosphoprotein (SARS-CoV-2 N-terminal domain)	Antibody	A1B1C1	IgFold + ReplicaDock
54	218	Ab against Nucleocapsid phosphoprotein (SARS-CoV-2 N-terminal domain)	Antibody	A1B1C1	IgFold + ReplicaDock
54	219	ATP-hydrolysing RuvAB complex representing a part of the holiday junction	Multimer	A6	SymDock
54	220	ATP-hydrolysing RuvAB complex representing a part of the holiday junction	Multimer	A6B1	ReplicaDock + SymDock
54	221	ATP-hydrolysing RuvAB complex representing a part of the holiday junction	Multimer	A6B2	ReplicaDock + SymDock
54	222	C-terminal domain of putative adhesin (Q6MNC5)	Multimer	A3	SymDock
54	224	Stoichiometry A8	Multimer	A3	SymDock
54	225	Viral capsid spike protein recombinant construct	Homodimer	A2	SymDock
54	226	Capsid protein (Murine astrovirus)	Homodimer	A2	SymDock
54	229	Tobacco Nictaba plant lectin	Homodimer	A2	SymDock
54	230	RAD52 Human DNA repair protein	Multimer	A10	SymDock

as 'difficult' ($RMSD_{UB}$ over 2.2 Å and/or poor monomer template availability). For these targets, predictors only achieved acceptable quality in 8 of 15 targets (53%) and high quality in only 2 (13%). In the post-AlphaFold era, the performance of predictor groups increased slightly as evident in Figure 4.1B-*bottom panel* (higher average dockQ scores across all targets). However, the difficult targets with higher range of conformational changes had relatively lower improvement in DockQ (53% acceptable, 26% medium and 6% high-quality predictions for 15 difficult targets), demonstrating the limitation in capturing the intrinsic flexibility of proteins.

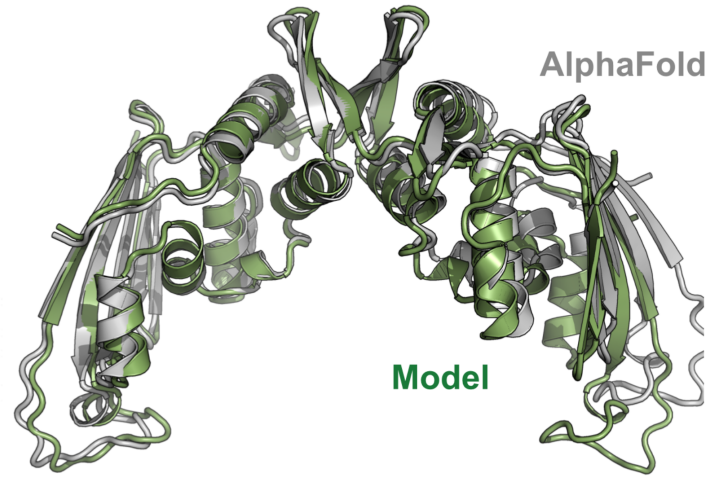


Figure 4.2: Prediction for target T194, a GP2 bacteriophage protein with role in phage infection. Target T194 represents a homodimer with A2 symmetry. Our model (*olive*) superimposed over AF2 prediction (*gray*). Starting from the AF2 template, our backbone sampling and docking routines identified a tighter interface and improved the DockQ score.

The recent CASP15-CAPRI round presented multiple heterodimeric and homodimeric complexes with considerable flexibility. For heterodimeric complexes we used both RosettaDock 4.0[25], a conformer-selection method, and ReplicaDock 2.0[15], an induced-fit method. Since these targets were in the post-AlphaFold era of modeling,

our initial structural poses were often generated by AlphaFold.[16] For conformer-selection with RosettaDock 4.0, we first produced an ensemble of diverse backbone conformations with Rosetta Relax[26], Rosetta Backrub[27], and anisotropic normal modes. Then, the structures were docked together with backbone swapping as described in Marze *et al*.[25] For induced-fit, we initiated docking from the predicted binding pose (from AlphaFold[16,19]) and allowed on-the-fly backbone motions on the putative interface while docking. Each ReplicaDock local docking simulation spans multiple trajectories (default is 8) with three temperature replicas per trajectory. The replicas are set with inverse temperatures, β, of 1.5^{-1} kcal$^{-1} \cdot$ mol, 3^{-1} kcal$^{-1} \cdot$ mol, and 5^{-1} kcal$^{-1} \cdot$ mol respectively. Replica exchange swaps are attempted every 1,000 MC (Monte Carlo) steps generating 6,000 decoys at a local binding site. Modeling of symmetric complexes and homodimers was performed with the symmetry framework in our SymDock 2.0 docking protocol.[28] SymDock 2.0 incorporates backbone ensembles and relaxation in high-resolution stage to sample tighter, complementary interfaces with better packing for symmetric complexes. Figure 4.2 demonstrates a symmetric homodimeric target of the GP2 bacteriophage protein with a role in phage replication. Although the native crystal structure is unavailable, I demonstrate our performance with reference to the AF2 model. Our model results in a medium CAPRI-quality prediction (DockQ score of 0.77 v/s 0.62 for AF2-multimer). The shift in orientation of one of the protein partners shows how our docking routines can capture binding-induced conformational changes and pack a tighter interface.

Conformational changes were of particular interest for targets 182-184 of the CAPRI COVID-19 open science initiative (Round 51): complexes of SARS-COV-2 viral proteins with human host proteins. These interactions were identified from the

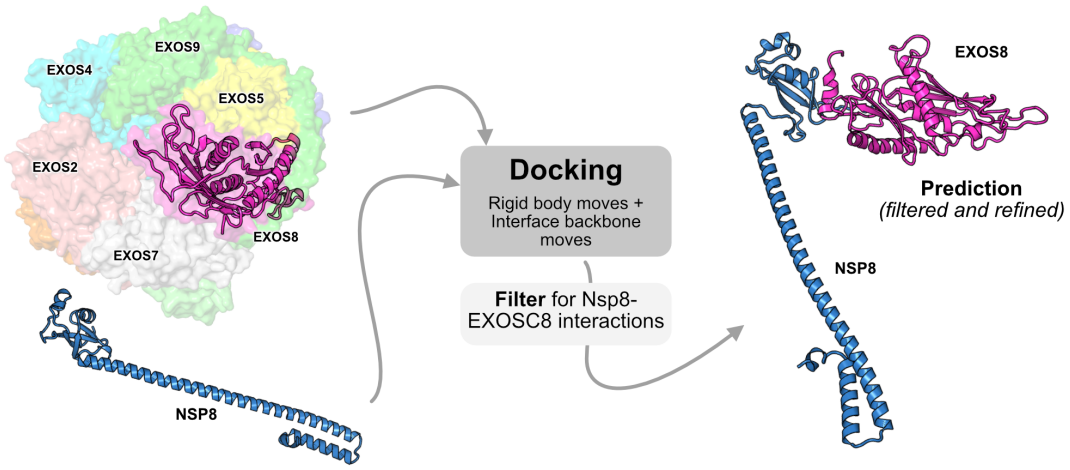


Figure 4.3: Prediction for the SARS-CoV-2 NSP8-EXOS8 complex. Available homologous templates of exosome complex highlight EXOS8 (*magenta*) in complex with other exosome subunits (*top left, in surface representation)*. Non-structural protein (NSP)-8 had been crystalized with NSP7. Our strategy docked these complexes with each other (rigid body roto-translation with induced-fit backbone moves, with focus on the NSP8-EXOS8 interactions. Highlighted on the right is the predicted complex for the NSP8-EXOS8 interaction after removing occluded contacts or potential clashes to other domains/subunits in the exosome. NSP8 has a golf-club fold and multiple sequence alignments with corresponding coronavirus proteins has revealed high conservation across the head domain suggesting that it might play an important role in interactions.[29] In our docked decoy, the head domain of the NSP interacts with the EXOS8 domain, which is in agreement with reported studies. The results of this rounds are not announced yet.

proteomics study of Gordon *et al.* and presented potential druggable host proteins.[30] Unbound crystal structures of a few viral proteins from SARS (PDB ID: 2GTH[31], 2OZK[32]) and potential templates crystalized for SARS-COV-2 (PDB ID: 6WLC[33]) highlighted conformational changes elevating the complexity of the docking challenge. Prior studies demonstrated that these targets were integral in the infection pathway of SARS-COV-2 with the human host proteins, with protein association hypothesized to occur in a co-dependent fashion.[30,34] For example, Target 183 viral protein NSP8 co-existed in complex with NSP7[29], and the human host protein Exos8 was a subunit in the eukaryotic exosome, a multisubunit complex responsible for cellular RNA degradation and processing.[35] Due to the potential complexes co-binding together, we[i] performed docking simulations for the entire complexes with ReplicaDock while focusing at the Exos8 protein surface. This aided in eliminating interfaces bound to other proteins and better elucidating the interaction between the viral protein and host protein of interest (Figure 4.3). For these targets, we benefited from the induced-fit backbone moves of ReplicaDock to model the flexible interfacial regions in the viral protein.

4.5 Multimeric protein targets are difficult to model

Modeling large protein assemblies and multimeric complexes is another intriguing challenge in structural biology. Over the years, the performance of Rosetta-based docking tools has improved for symmetric multimers demonstrated with SymDock 2.0. However, heteromers, *i.e.* distinct protein subunits associating to form larger complex assemblies, have been a limitation owing to the exponentially large sample space. This was highlighted in target 160, the assembly of a surface-layer SAP protein

[i]Dr. Rahel Frick and Dr. Rituparna Samanta aided in docking and analysis for this target

derived from *Bacillus anthracis*. Surface layer proteins (SLPs) are synthesized in a highly regulated fashion with the protein first folding into discrete subunits followed by a self-assembly into a two-dimensional array.[36] The challenge entailed modeling the 3D assembly, *i.e.* the intermolecular contacts between the six discrete subunits (domains D1-D6) and the conformations of the connecting loops. To replicate this biological procedure of hierarchical assembly, I developed a fold-and-dock approach for predicting this assembly.

The organizers facilitated us with crystal structures of domains (D1-D6) oriented randomly in cartesian space along with sequence information of the linkers connecting the domains. The prediction task now entailed modeling the three-dimensional structure of the SAP assembly domain, *i.e.*, the intra-molecular contacts between the SAP domains, as well as the conformations of the connecting loops. Figure 4.4 illustrates the docking strategy that I employed for this assembly prediction. First, I identified the root linker and the base domains. The root linker denotes the linker loop that I first aimed to build with the base domains signifying the domains it connects. The smallest residue sequence linker was chosen as the root, as that would allow less mobility of the domain and allow us to narrow down the conformational search space. Domains 3 and 4 with the three-residue linker 'KEP' were chosen as the base domains. To build the loop, the C-termini of domain 3 and N-termini of domain 4 were translated to be roughly 10 Å apart (note that the Cα-Cα distance in proteins is approx. 3.8 Å). I then incorporated the residues on the termini and used the Rosetta Kinematic Closure (KIC) protocol[37] to close the loop. Once the loop was built, I used the FloppyTail protocol[38] to sample diverse orientations of the domains relative to each other and evaluated the generated decoys on interface score

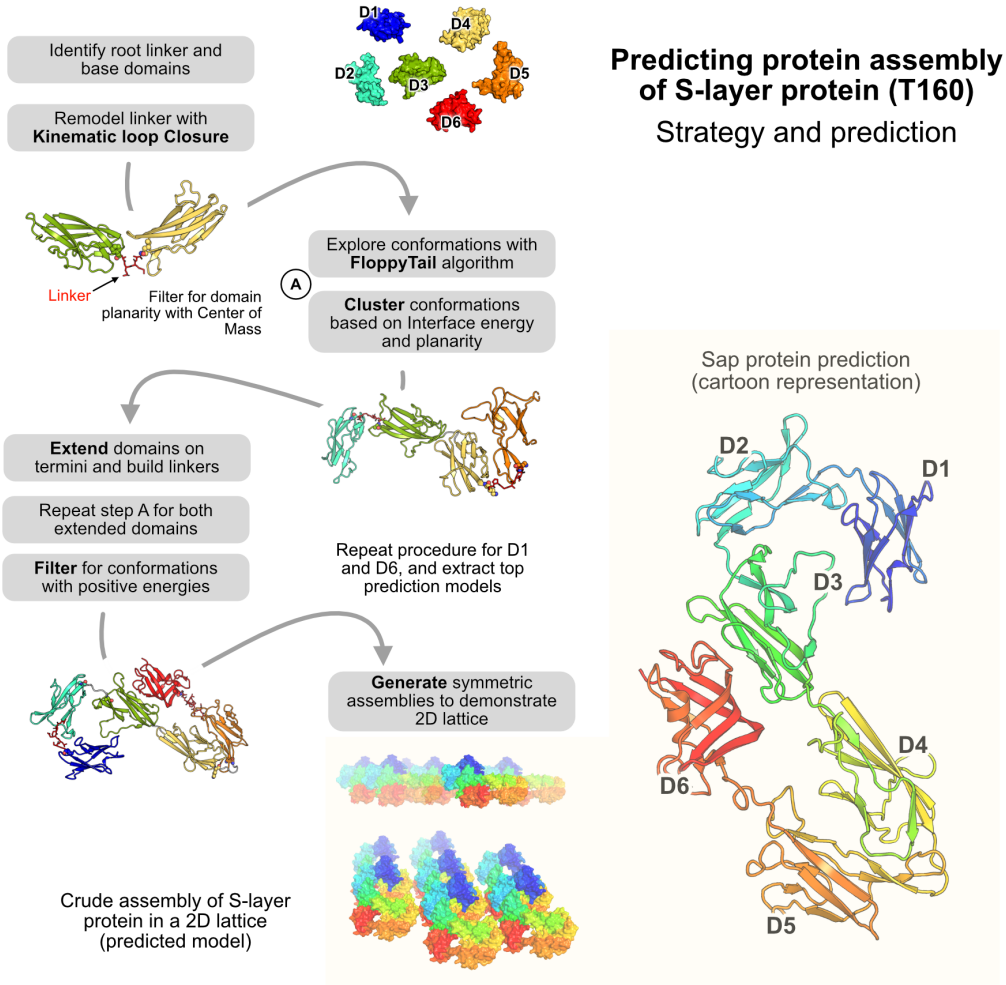


Figure 4.4: Prediction of surface-layer protein assembly. (*left panel*) The fold-and-dock strategy for self-assembly of S-layer proteins. Starting from two base domains (D3 and D4), we first built the linker and perform FloppyTail motions to sample conformations. The generated decoys were evaluated with energies and filtered with planarity score (*i.e.* a metric to measure if the Center of Mass of the domains and the linker are in the same plane to ensure 2D geometry). Top decoys were clustered (5 clusters are selected) and domains were extended at the termini (D2 and D5). I repeated the evaluation and filtering cycle, but also took into account docking metrics such as interfacial contacts within domains to identify top models. Next, I added the next set of domains (D1 and D6) till I obtained the final assembly. Decoys were clustered based on root-mean-square-deviation, and the top clusters were relaxed and submitted for assessment. Highlighted is a surface representation (side view and top view) of the 2D lattice assembly via symmetry-mates in PyMol to visualize a potential 2D lattice. (*right panel*) Predicted S-layer protein in cartoon representation highlights inter-domain interactions between D1-D3, D2-D3, D3-D6 and D4-D5 respectively. Note that the results of this round are to be announced.

and planarity filter. Since the domains were supposed to form a two-dimensional lattice, the center of mass of all the domains would be restricted to a 2D space. This filter aimed at removing any deviants and helping cluster structures relevant for the overall assembly. This procedure was repeated by extending on the edges of the two domains and building the next sets of loops connecting first to domains 2 and 5, and then to domains 1 and 6 respectively (Figure 4.4-*left panel*). To distinguish good predictions over others, the decoys sampled were also evaluated on interfacial contacts (with other domains) and potential clashes. With this strategy, our final pool of decoys was sorted by total Rosetta energy, and the top 10 decoys were submitted for assessment. Figure 4.4 (*right panel*) shows our modeled prediction. Further, I generated symmetry mates to visualize if there is a feasible 2D array being assembled (*side view and top view*). This is illustrated with the surface representation of the SAP protein assembly. Even though the results of this round are not published yet, I speculate that this fold-and-dock strategy could capture good native-like contacts owing to its aggressive sampling and filtering loops. Further, it also demonstrates a hierarchical approach for multi-body docking and protein self-assembly.

On other hand, larger symmetric assemblies with known stoichiometries were easier to adapt with our symmetric docking routines described earlier. One such assembly is target 230, a human DNA repair protein complex with A10 stoichiometry. For large multimeric assemblies, AlphaFold predictions are often incorrect due to inadequacy of handling larger sequence lengths owing to the limits in the multiple sequence alignments (MSAs). As Figure 4.5A demonstrates, the monomer prediction is of poor accuracy with disordered regions, limiting its utility for docking. Here, we relied on Robetta ab-initio modeling to obtain a monomer sub-unit, and then

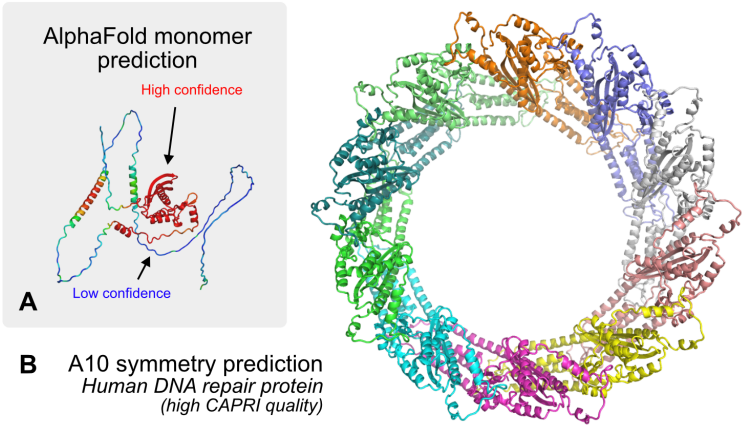


Figure 4.5: Prediction for target T230, Human DNA repair protein (multimeric complex) (A) Prediction of monomeric subunit with AlphaFold. Target T230 presents a A10 stoichiometry multimer difficult to model with AlphaFold-multimer. The monomeric subunit was first modeled with AlphaFold to obtain a starting point, however, the predicted model is inaccurate (high confidence core and low confidence termini) with many disordered regions. A refined, *ab-initio* model from Robetta was employed for our docking calculations with SymDock 2.0. (B) Symmetric assembly of the multimer for target T230. We obtained a high CAPRI-quality ranking for our prediction highlighting the accuracy of our symmetric docking routines.

passed this sub-unit to SymDock 2.0. As SymDock 2.0 incorporates virtual atoms to build a symmetric assembly with the known A10 stoichiometry, it predicted a high CAPRI-quality model for target T230 (Figure 4.5B).

4.6 The challenges in modelling antibody-antigen interactions

Antibody-antigen interactions are the core of immune recognition and signaling, leading to the development of therapeutics for the treatment of increasingly complex diseases. The fundamental units of antibody-antigen interaction constitute the sequence regions on antibody and antigen that comprise the interacting residues (paratope and epitope respectively). The binding interface is predominantly the structural arrangement of a set of loops that constitute the complementarity determining

region (CDR). With the advent of deep learning, tools such as DeepAb[39], ABlooper[40] and IgFold[41] have aimed at better modeling of antibody structures. Despite these advancements, predicting the binding interface or the CDRs structures (particularly CDR H3) has been challenging. The CASP15-CAPRI experiment(Round 54) unveiled three antibody and five nanobody interactions with antigens. Owing to the known inefficiency of the available tools to accurately model these structures, all of these interactions were deemed as 'challenging'. For these 8 targets, we[ii] employed our docking tools, specifically ReplicaDock 2.0, with the structural templates from the deep-learning tools (IgFold and AF2) to model these structures.

First, to model the antibody and nanobody structures, we utilized IgFold, a multi-track deep learning prediction tool by Ruffolo *et al*.[41] The complex structures were obtained from AF2-multimer[18,19], and the antibody/nanobody in the AF2 generated model were replaced with the structures generated from IgFold[41] thereby preserving the binding region identified by AlphaFold. IgFold structures have demonstrated better performance on modeling CDR H3 loops and can be better starting templates to model docked complexes.[41] To focus on sampling paratopic regions of the antibody, I utilized the directed induced-fit strategy illustrated in Chapter 2 to narrow the conformational search towards relevant residues.[15] Figure 4.6 shows our performance in modeling nanobody-antigen complexes (nanobody in complex with mammalian CNPase phosphodiesterase) for two targets T205 (A) and T206 (B), with respect to the top predictions. Unfortunately in all cases, AlphaFold predicted an incorrect binding interface (epitope) for the nanobodies, skewing the search in a false-positive conformational space. In both cases, ReplicaDock 2.0 modeled decoys were able to identify alternate binding sites. However, the predicted structure is off considerably

[ii] Lee-Shin Chu assisted in the modeling of nanobodies and antibodies

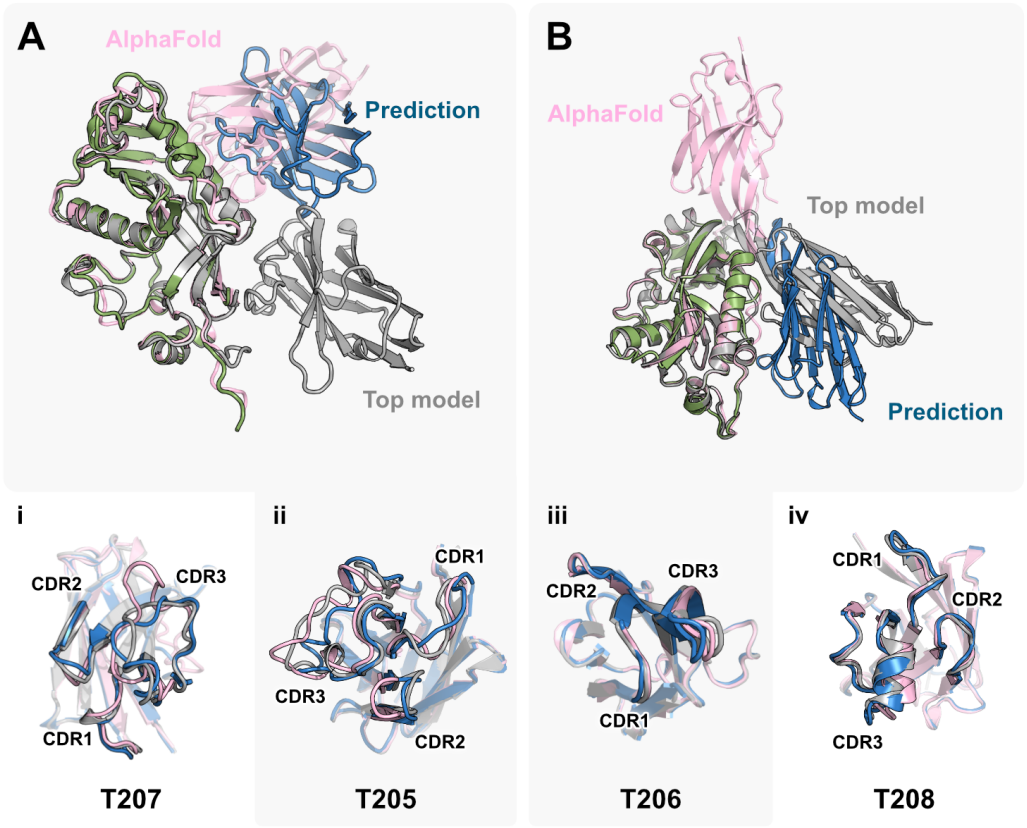


Figure 4.6: Prediction of antibody/nanobody-antigen complexes. Targets 205-208 were nanobodies in complex with mammalian CNPase phosphodiesterase domain. Comparison is provided for our prediction (*green-blue*) with AlphaFold prediction (*pink*) and the top model for each target (*gray*). All structures are aligned over the CNPase phosphodiesterase domain for ease in visualization of nanobody orientation. (A) Target T205 was incorrectly predicted by our model. Because the AF starting template was incorrect, binding interface was potentially skewed by using it as a starting point. (B) Target T206 was an acceptable prediction where the docked structure captured the correct binding interface but had an incorrect orientation and less native-like contacts. (i-iv) Superimposition of nanobody backbones for each target over the nanobody backbone of the top-model. (i) and (ii) demonstrate cases where our backbone sampling is aggressive (owing to longer CDR H3 loops) and attains a structural similarity to the top-model. (iii) and (iv) demonstrate cases where the nanobody prediction is fairly accurate.

for target T205 (*panel*A) and in an incorrect orientation for target T206 (*panel*B). Figure 4.6.i-iv shows the nanobodies in the our top models (*blue*) aligned over respective AlphaFold structures (*pink*) and the best prediction (*gray*). In all the cases, our conformational sampling predicts CDR backbones in agreement with the top model. This results suggest a need for aggressive global sampling to identify better binding poses and capture high quality structures.

4.7 Discussion and conclusions

In the last few years, the field of structural biology, particularly protein structure prediction has changed dramatically. The immense influx of deep-learning tools for protein sequence-to-structure prediction has brought a paradigm shift in the field. Prior rounds of CASP have presented predictors with unique challenges contributing to the development of modeling tools. But with AlphaFold's high accuracy prediction, the field has transitioned to not just predicting assemblies - the focus of CAPRI - but also dynamic structures, *i.e.* protein ensembles, multi-state proteins, and lately, protein-nucleic acid complexes. The goal of this community-driven, blind prediction challenge is to stimulate engagement and move the field ahead, and Deepmind has contributed to this step change.

In this chapter, I discussed a few highlights of our predictions over the last four years, both *pre-* and *post*-AlphaFold, pointing out the shift from aggressive global search to focused backbone sampling and refinement respectively. First, I highlighted that despite the improvement in overall *static* structure prediction accuracy, capturing large-scale conformational changes remains a challenge, even for AlphaFold. Our sampling and scoring strategy within Rosetta is robust but is insufficient for larger

conformational changes. Developing aggressive and relevant backbone movers would be paramount to obtain high CAPRI-quality predictions.

Second, the performance of Rosetta docking routines for multimers has demonstrated better results with more accurate monomeric subunits models. Alternatively, multi-body docking, *i.e.* docking more than two chains or protein partners, remains a major limitation of our protocols. Adapting our docking protocols, RosettaDock 4.0 or ReplicaDock 2.0, to handle multiple chains could be a promising future direction. Unlike symmetric multimers, where a subunit could be virtually extended based on stoichiometry, multi-body docking for heteromers would be computationally expensive.

Third, for antibody-antigen complexes, relying on AlphaFold templates hampers prediction accuracy. Performing a global rigid docking search with ReplicaDock 2.0[15] or ClusPro[42] could have enabled us to obtain a picture of potential binding interfaces to refine with local backbone sampling. Alignment of nanobodies (Figure 4.6) highlights that by focusing on flexible regions of proteins, we can capture diverse, relevant backbones with our current backbone movers. Further, I observed that AlphaFold confidence scores at a residue-level correlate well with flexibility (for nanobodies and antibodies, these would be the paratopic CDR regions). This observation suggests that we can identify mobile regions in blind protein targets improving our backbone sampling. I will investigate this premise in more detail in the next chapter. To conclude, CAPRI targets have led to the development and upgrade of docking protocols in the Gray lab. While demonstrating the benefits of our MCM approach, these blind targets have demarcated the deficiencies in our score-functions and the limitations of our backbone sampling. For antibody-antigen targets particularly, it has highlighted

the benefit of exhaustive global sampling. Clearly, the docking success rates have improved over the decade, but for docking to sustain as a reliable, stand-alone tool, we need higher success rates in all cases with the ability to capture binding-induced conformational changes.

References

1. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins* **87,** 1011–1020. ISSN: 1097-0134 (Electronic) (12 2019).
2. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics* **89,** 1607–1617.
3. Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Costanzo, L. D., Duarte, J. M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D. S., Ghosh, S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Lawson, C., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Persikova, I., Randle, C., Rose, A., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Tao, Y.-P., Voigt, M., Westbrook, J., Young, J. Y., Zardecki, C. & Zhuravleva, M. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research* **49,** D437–D451. ISSN: 0305-1048.
4. Lensink, M. F., Velankar, S., Baek, M., Heo, L., Seok, C. & Wodak, S. J. The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins: Structure, Function and Bioinformatics* **86,** 257–273. ISSN: 10970134 (July 2017 2018).
5. Lensink, M. F. *et al.* Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins: Structure, Function and Bioinformatics,* 1200–1221. ISSN: 10970134 (May 2019).
6. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE* **11,** 1–9.
7. Gray, J. J., Moughon, S. E., Kortemme, T., Schueler-Furman, O., Misura, K., Morozov, A. V. & Baker, D. Protein-protein docking predictions for the CAPRI experiment. *Proteins: Structure, Function and Genetics* **52,** 118–122. ISSN: 08873585. http://doi.wiley.com/10.1002/prot.10384 (1 2003).

8. Daily, M. D., Masica, D., Sivasubramanian, A., Somarouthu, S. & Gray, J. J. CAPRI rounds 3-5 reveal promising successes and future challenges for RosettaDock. *Proteins* **60,** 181–186. ISSN: 1097-0134 (Electronic) (2 2005).

9. Chaudhury, S., Sircar, A., Sivasubramanian, A., Berrondo, M. & Gray, J. J. Incorporating biochemical information and backbone flexibility in RosettaDock for CAPRI rounds 6-12. *Proteins* **69,** 793–800. ISSN: 1097-0134 (Electronic) (4 2007).

10. Sircar, A., Chaudhury, S., Kilambi, K. P., Berrondo, M. & Gray, J. J. A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13-19. *Proteins: Structure, Function and Bioinformatics* **78,** 3115–3123. ISSN: 08873585 (15 2010).

11. Kilambi, K. P., Pacella, M. S., Xu, J., Labonte, J. W., Porter, J. R., Muthu, P., Drew, K., Kuroda, D., Schueler-Furman, O., Bonneau, R. & Gray, J. J. Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20-27. *Proteins: Structure, Function and Bioinformatics* **81,** 2201–2209. ISSN: 08873585 (12 2013).

12. Marze, N. A., Jeliazkov, J. R., Burman, S. S. R., Boyken, S. E., DiMaio, F. & Gray, J. J. Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28-35. *Proteins* **85,** 479–486. ISSN: 1097-0134 (Electronic) (3 2017).

13. Burman, S. S. R., Nance, M. L., Jeliazkov, J. R., Labonte, J. W., Lubin, J. H., Biswas, N. & Gray, J. J. Novel sampling strategies and a coarse-grained score function for docking homomers, flexible heteromers, and oligosaccharides using Rosetta in CAPRI rounds 37–45. *Proteins: Structure, Function, and Bioinformatics* **88,** 973–985.

14. Lensink, M. F. *et al.* Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics* **89,** 1800–1823. ISSN: 0887-3585.

15. Harmalkar, A., Mahajan, S. P. & Gray, J. J. Induced fit with replica exchange improves protein complex structure prediction. *PLOS Computational Biology* **18,** 1–21.

16. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589. ISSN: 14764687.

17. Harmalkar, A. & Gray, J. J. Advances to tackle backbone flexibility in protein docking. *Current Opinion in Structural Biology* **67,** 178–186. ISSN: 1879033X.

18. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. ColabFold - Making protein folding accessible to all. *bioRxiv.*

19. Evans, R., Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Ž´, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J. & Hassabis, D. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021).

20. Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A. & Schueler-Furman, O. Harnessing protein folding neural networks for peptide–protein docking. *Nature Communications* **13,** 176. ISSN: 2041-1723.

21. Ko, J. & Lee, J. Can AlphaFold2 predict protein-peptide complex structures accurately? *bioRxiv.*

22. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876. ISSN: 0036-8075 (6557 2021).

23. Wallner, B. AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling (2022).

24. Elofsson, A. *Protein Structure Prediction until CASP15* 2022. arXiv: 2212.07702 [q-bio.BM].

25. Marze, N. A., Burman, S. S. R., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics* **34,** 3461–3469. ISSN: 14602059 (20 2018).

26. Gregorio, R. A. N. D. B. D. N. L. & Moretti. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLOS ONE* **8,** 1–5.

27. Smith, C. A. & Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *Journal of Molecular Biology* **380,** 742–756. ISSN: 0022-2836.

28. Burman, S. S. R., Yovanno, R. A. & Gray, J. J. Flexible Backbone Assembly and Refinement of Symmetrical Homomeric Complexes. *Structure* **27,** 1041–1051.e8. ISSN: 18784186.

29. Zhai, Y., Sun, F., Li, X., Pang, H., Xu, X., Bartlam, M. & Rao, Z. Insights into SARS-CoV transcription and replication from the structure of the nsp7–nsp8 hexadecamer. *Nature Structural Molecular Biology* **12,** 980–986. ISSN: 1545-9985.

30. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583,** 459–468. ISSN: 1476-4687.

31. Xu, X., Zhai, Y., Sun, F., Lou, Z., Su, D., Xu, Y., Zhang, R., Joachimiak, A., Zhang, X. C., Bartlam, M. & Rao, Z. New Antiviral Target Revealed by the Hexameric Structure of Mouse Hepatitis Virus Nonstructural Protein nsp15. *Journal of Virology* **80,** 7909–7917.

32. Joseph, J. S., Saikatendu, K. S., Subramanian, V., Neuman, B. W., Buchmeier, M. J., Stevens, R. C. & Kuhn, P. Crystal Structure of a Monomeric Form of Severe Acute Respiratory Syndrome Coronavirus Endonuclease nsp15 Suggests a Role for Hexamerization as an Allosteric Switch. *Journal of Virology* **81,** 6700–6708.

33. Kim, Y., Jedrzejczak, R., Maltseva, N. I., Wilamowski, M., Endres, M., Godzik, A., Michalska, K. & Joachimiak, A. Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Protein Science* **29,** 1596–1605.

34. Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., Cinatl, J. & Münch, C. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* **583,** 469–472. ISSN: 1476-4687.

35. Liu, Q., Greimann, J. C. & Lima, C. D. Reconstitution, Activities, and Structure of the Eukaryotic RNA Exosome. *Cell* **127,** 1223–1237. ISSN: 0092-8674.

36. Kern, J., Wilton, R., Zhang, R., Binkowski, T. A., Joachimiak, A. & Schneewind, O. Structure of Surface Layer Homology (SLH) Domains from Bacillus anthracis Surface Array Protein *. *Journal of Biological Chemistry* **286,** 26042–26049. ISSN: 0021-9258. https://doi.org/10.1074/jbc.M111.248070 (29 2011).

37. Bhardwaj, G., Mulligan, V. K., Bahl, C. D., Gilmore, J. M., Harvey, P. J., Cheneval, O., Buchko, G. W., Pulavarti, S. V. S. R. K., Kaas, Q., Eletsky, A., Huang, P.-S., Johnsen, W. A., Greisen, P. J., Rocklin, G. J., Song, Y., Linsky, T. W., Watkins, A., Rettie, S. A., Xu, X., Carter, L. P., Bonneau, R., Olson, J. M., Coutsias, E., Correnti, C. E., Szyperski, T., Craik, D. J. & Baker, D. Accurate de novo design of hyperstable constrained peptides. *Nature* **538,** 329–335. ISSN: 1476-4687.

38. Kleiger, G., Saha, A., Lewis, S., Kuhlman, B. & Deshaies, R. J. Rapid E2-E3 Assembly and Disassembly Enable Processive Ubiquitylation of Cullin-RING Ubiquitin Ligase Substrates. *Cell* **139,** 957–968. ISSN: 00928674. http://dx.doi.org/10.1016/j.cell.2009.10.030 (5 2009).

39. Ruffolo, J. A., Sulam, J. & Gray, J. J. Antibody structure prediction using interpretable deep learning. *Patterns* **3,** 100406. ISSN: 2666-3899.

40. Abanades, B., Georges, G., Bujotzek, A. & Deane, C. M. ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. *Bioinformatics* **38,** 1877–1880. ISSN: 1367-4803.

41. Ruffolo, J. A., Chu, L.-S., Mahajan, S. P. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *bioRxiv.*

42. Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D. & Vajda, S. The ClusPro web server for protein-protein docking. *Nature Protocols* **12,** 255–278. ISSN: 17502799 (2 2017).

Chapter 5

From sequence to structure to complexes : an in-silico pipeline for protein-protein docking

5.1 Overview

Despite recent breakthrough of AlphaFold (AF) in the field of protein sequence-to-structure prediction, producing accurate structures of most of the monomeric (single-chain) proteins is achievable. Yet, modeling protein interfaces and predicting protein complex structures remains challenging, especially when there is a significant conformational change in one or both binding partners. Prior studies have demonstrated that AF-multimer can predict accurate protein complexes in only up to 43% of cases.[1]. However, these studies have not reflected upon approaches to improve failures. In this work, I combine AlphaFold as a structural template generator with a physics-based replica exchange docking algorithm, ReplicaDock 2.0. Using a curated collection of 254 available protein targets with both unbound and bound structures, I first demonstrate that AlphaFold confidence measures can be repurposed for estimating protein flexibility and docking accuracy for multimers. I incorporate these metrics

within our ReplicaDock 2.0 protocol[2] to complete a robust in-silico pipeline for accurate protein complex structure prediction. AlphaRED (AlphaFold-inspired Replica Exchange Docking) successfully docks failed AF predictions including 97 failure cases in Docking Benchmark Set 5.5. AlphaRED generates CAPRI medium-quality predictions for 60% of benchmark targets. This new strategy integrating deep-learning based architectures trained on evolutionary information with physics-based enhanced sampling approaches predict protein complex structures.

5.2 Introduction

In-silico protein structure prediction *i.e.* sequence to structure, tackles one of the core questions in structural biology. The recent release of AlphaFold[3] has brought a paradigm shift in the field by intertwining deep-learning tools with evolutionary data to predict single-chain structures with higher accuracy. Further, AlphaFold-multimer[4] (AFm) and related work[5] have demonstrated the utility of AlphaFold to predict protein complexes. The association of proteins to form transient or stable protein complexes often involves binding-induced conformational changes. Capturing conformational dynamics of protein-protein interactions has been one of the grand challenges in structural biology. Given that AlphaFold generates a static three-dimensional structure, it has been unclear whether conformational diversity could be captured by AlphaFold. In other terms, given a protein sequence, could AlphaFold generate ensembles of structures that include both unbound and bound conformations? Additionally, can AlphaFold reveal intrinsic conformational heterogeneity?

To diversify model complexes generated with AlphaFold-multimer in the recent round of CASP15, predictors employed tuning parameters such as dropout[6], higher

recycles on inference[7] or modulating the MSA inputs[8,9] with the amino acid sequence. While these approaches demonstrated the ability to generate broader conformational ensembles, AFm performance still worsens with a higher degree of conformational flexibility between unbound and bound targets[1]. Prediction accuracies are especially deteriorated in bound complex regions involving loop motions, concerted motions between domains, rearrangement of secondary structures, or hinge-like domain motions, *i.e.*, large-scale conformational changes, which are also challenging for conventional docking methods.[10] However, unlike state-of-the-art docking algorithms, AlphaFold's output models incorporate a residue-specific estimate of prediction accuracy. This suggests a few interesting questions:

- Do the residue-specific estimates from AF/AFm relate to potential metrics demonstrating conformational flexibility?
- Can AF/AFm metrics deduce information about docking accuracy?
- Can one create a docking pipeline for in-silico complex structure prediction incorporating AFm to convert sequence to structure to docked complexes?

Unlike deep-learning approaches that mined evolutionary information for structure prediction, recent work in physics-based docking approaches equipped induced-fit docking[2], larger ensembles[11], or fast-fourier transforms[12] with improved energy functions to capture conformational changes and better dock protein structures. Coupling temperature replica exchange with induced-fit docking, ReplicaDock 2.0[2] achieved successful local docking predictions on 80% of rigid ($RMSD_{UB} < 1.1$Å) and 61% medium ($1.1 \leq RMSD_{UB} < 2.2$ Å) targets in the Docking Benchmark 5.0 set[13]. However, like most state-of-the-art physics-based docking methods, ReplicaDock

2.0 performance was limited for highly flexible targets: 33% success rate on targets with $RMSD_{UB} \geq 2.2$ Å. Promisingly, focusing backbone moves on known mobile residues (i.e. residues that exhibit conformational changes upon binding), ReplicaDock 2.0 sampling could substantially improve the docking accuracy. But a major caveat was that the flexible residues must somehow be identified. With the advent of deep-learning approaches like AlphaFold-multimer[4] or docking-specific tools such as EquiDock[14], DockGPT[15], and GeoDock, faster high-throughput prediction of protein structures is feasible (0.1-10 mins on single NVIDIA GPU), albeit with lower accuracy. In such a scenario, it is computationally expensive to utilize physics-based tools for long time-scale global docking simulations.

In this work, I aim to combine the features of top deep learning approaches (*i.e.* AlphaFold) with physics-based docking schemes (ReplicaDock 2.0) to systematically dock protein interfaces. The overarching goal of this work is to create a robust pipeline for computationalists and biologists for easier, reproducible, and accurate modeling of protein complexes. Here I investigate the aforementioned questions and create a protocol to use AFm fused with our Rosetta-based replica exchange docking approach (ReplicaDock 2.0[2]) to improve on AFm failures and capture binding-induced conformational changes. First, on a curated benchmark set of unbound and bound protein structures, I assess the utility of AFm confidence metrics to detect conformational flexibility and binding site confidence. Next, I equip these metrics while treating AFm as a structural template generator to our docking routine, thereby detailing the development of AlphaRED (AlphaFold-inspired Replica Exchange Docking). AlphaRED builds over the leading protein structure prediction tool (AlphaFold-multimer) by incorporating biophysical information (via energy functions) to better model protein

complexes.

5.3 Results

5.3.1 Dataset curation

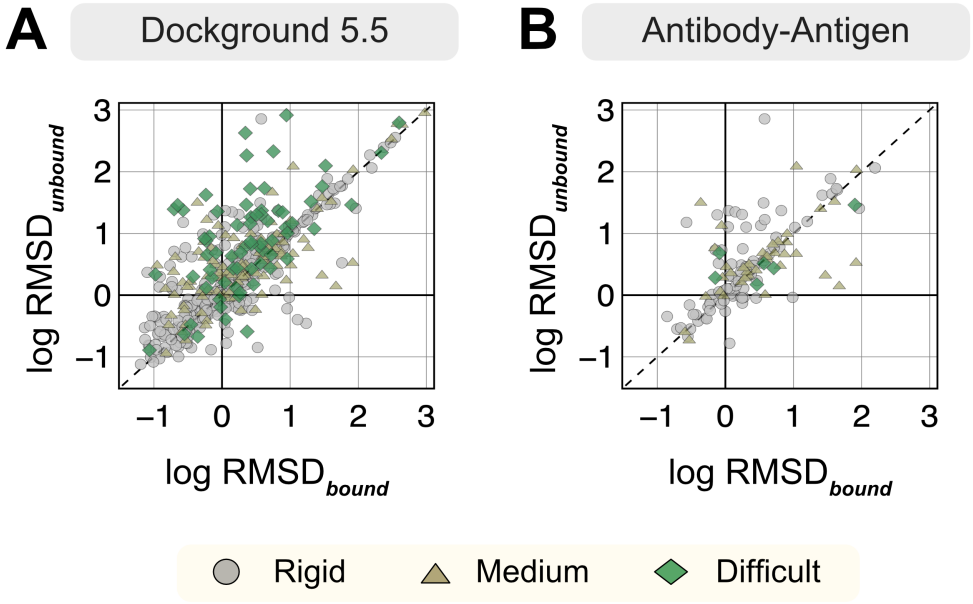


Figure 5.1: RMSDs of AlphaFold-multimer structures from experimental unbound and bound structures. Distribution of the RMSD between the AlphaFold-multimer prediction top-ranked model and the experimental unbound and bound structures. For each target, the protein partners are split into receptor and ligand respectively for comparison. Each symbol represents a category of flexibility (rigid, medium, and flexible). (A) Dockground Benchmark set 5.5; (B) Antibody/nanobody-antigen targets from the benchmark.

I curated a dataset for conformational flexibility from the Docking Benchmark Set 5.5 (DB5.5)[13], which comprises experimentally-characterized (X-ray or cryo-EM) structures of bound protein complexes and their corresponding unbound protein subunits. Each protein target (with unbound and bound structures) is classified based on their unbound-to-bound root-mean-square-deviation ($RMSD_{UB}$) as rigid ($RMSD_{UB} \leq 1.2$ Å), medium (1.2 Å $< RMSD_{UB} \leq 2.2$ Å) or difficult ($RMSD_{UB} \geq 2.2$ Å). Further,

owing to the poor performance of AlphaFold and other predictor groups in predicting antibody-antigen targets in the recent CASP15-CAPRI round[16], I identified a subset comprising only antibody-antigen complexes (including single domain antibodies, or nanobodies) by extracting all the antibody-antigen structures from the DB5.5[13] set. The comprehensive dataset includes 254 protein targets exhibiting binding-induced conformational changes.

For each protein target, I extracted the amino acid sequence from the bound structure and predicted corresponding three-dimensional complex structure with the ColabFold implementation (github.com/YoshitakaMo/localcolabfold) of the AlphaFold-multimer v2.3.0 (released March 2023) for all 254 benchmark targets. Being trained on experimentally-characterized structures deposited in the PDB, AlphaFold is expected to produce models analogous to the PDB structures. However, since both unbound and bound structures exist for the benchmark targets in the PDB, I first investigated whether AFm exhibits any bias towards either unbound or bound forms for the same protein sequence. Figure 5.1 compares the Cα-RMSD of all protein partners of the AFm predicted complex structures from the bound (B) and unbound (UB) crystal structures on a log-log scale (a few AFm predicted models were 20 Å apart from both bound and unbound structures). As evident from Figure 5.1A, the protein partners from the AFm top-ranked model skew more often towards the bound state with structural deviation from both unbound and bound forms. Antibody-antigen targets further demonstrate a similar trend, however with fewer targets predicted within sub-Angstrom accuracy to the bound form (29.7% for Ab-Ag targets as opposed to 41% for DB5.5).

5.3.2 AlphaFold pLDDT provides a predictive confidence measure for backbone flexibility

AlphaFold employs multiple sequence alignments with a multi-track attention-based architecture to predict three-dimensional structures of proteins and complexes. Further, for each structural prediction, it provides a residue-level confidence measure: the predicted local-distance difference test (pLDDT), estimating the agreement between predicted model to an experimental structure based on the Cα LDDT test (*refer Methods*). Tunyasuvunakool *et al.* analyzed pLDDT confidence measures for the human proteome demonstrating the correlation between lower pLDDT scores with higher disordered regions in protein structures.[17] Building on this observation, I evaluated whether there is a correlation between AlphaFold pLDDT confidence metric and the experimental metrics of conformational change between unbound and bound structures. In this regard, I compared the computational (AF-pLDDT) and experimental (per-residue RMSD and LDDT) metrics against each other.

As a reference, I first superimposed the unbound partners over the bound structures and calculated residue-wise Cα deviations, to determine the per-residue RMSD$_{BU}$ values. LDDT$_{BU}$ was measured by calculating the local distance differences in the unbound structure relative to the bound form. These metrics capture the extent of motion in the unbound-bound transitions for each of the protein targets. Next, I compared the per-residue pLDDT score from AFm predicted monomer models with the experimental metrics. Figure 5.2A,B shows the results for two protein targets: kinase-associated phosphastase in complex with phospho-CDK2 (1FQ1[18]) and TGF-β receptor with FKBP12 domain (1B6C[19]). In both cases, pLDDT confidence scores correlate with the experimental measurements of binding: pLDDT decreases as

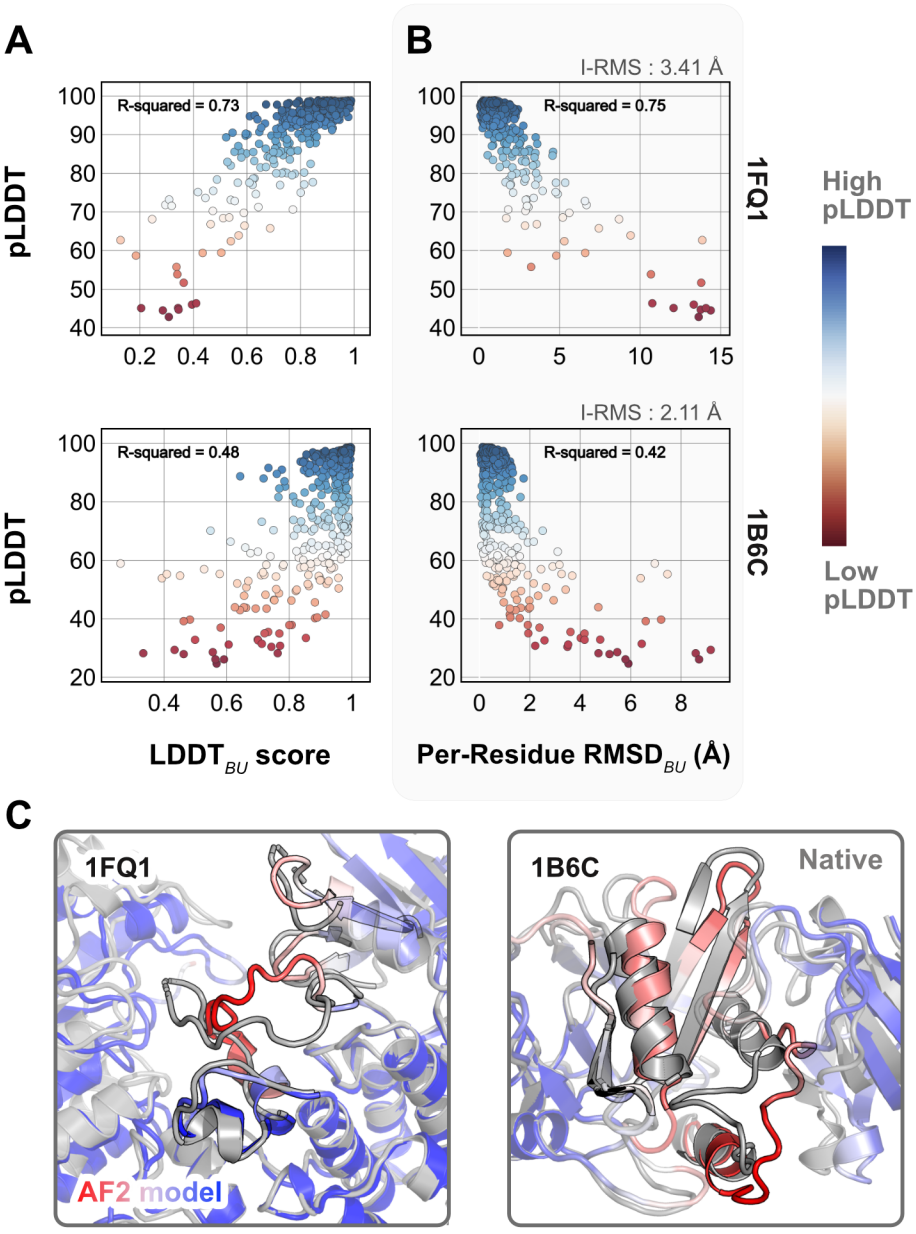


Figure 5.2: Comparison of AFm pLDDT with structural metrics. (A) AlphaFold pLDDT plotted against LDDT$_{BU}$ (local distance difference test). LDDT$_{BU}$ is calculated by comparing the unbound and bound environment for each residue. High scores correlate with high pLDDT (*red*). (B) Per-residue root-mean-square-deviation between unbound-bound structures(Per-Residue RMSD$_{BU}$) against AlphaFold pLDDT. Higher RMSDs correlate with lower pLDDT. (C) Structures for two targets (PDB ID: 1B6C and 1FQ1) with the experimental bound form (*in gray*) and the AlphaFold-multimer predicted model (*colored by spectrum, red-white-blue*). In both cases, the residues with low pLDDT scores (*red*) are the residues with incorrect conformation and more conformational change.

$LDDT_{BU}$ decreases and $RMSD_{BU}$ increases. This is further illustrated with the PyMol representation of the two targets over the bound structures (Figure 5.2C). In regions of low confidence/pLDDT (highlighted in *red*), the prediction is inaccurate, but higher confidence/pLDDT regions (highlighted in *blue*) have high accuracy of prediction with the bound form. The results for the entire benchmark set show similar trends for most targets. The pLDDT, thus, can suggest mobile residues in a protein structure.

5.3.3 Interface-pLDDT correlates with DockQ and discriminates poorly docked structures

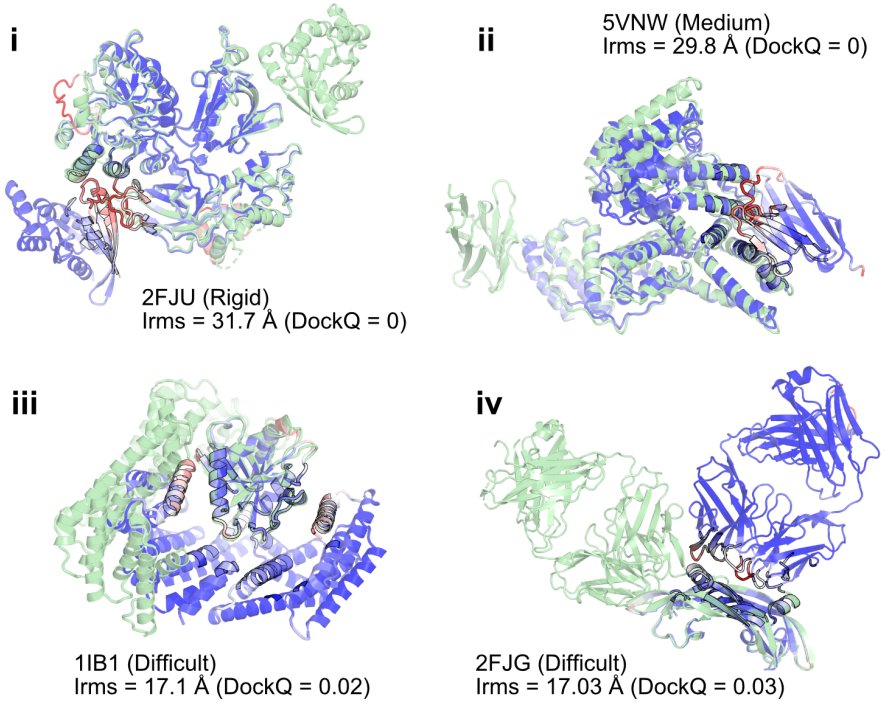


Figure 5.3: AlphaFold predictions with reference to bound experimentally-characterized structures. Here I demonstrate four targets with poor DockQ scores and high interface RMSDs. (i) Activated Rac1 bound to phospholipase Cβ2 (2FJU) - rigid target ($RMSD_{UB}$= 1.04). (ii) Nanobody bound to serum albumin (5VNW) - medium target($RMSD_{UB}$= 1.49) (iii) 14-3-3 Zeta Isoform:Serotonin N-acetyltransferase complex (1IB1) - difficult target ($RMSD_{UB}$= 2.09) (iv) G6 antibody in complex with the VEGF antigen - difficult target ($RMSD_{UB}$= 2.51). Bound structure is highlighted in *gray* and the AlphaFold prediction is highlighted in *green-blue*.

When the prediction accuracy is lower, it is often evident from lower confidence metrics (such as average pLDDT or PAE). However, for AlphaFold-multimer complex predictions, the confidence metrics of the overall prediction do not correlate with the accuracy of the docked prediction, i.e. even if the complex exhibits higher confidence, the docking interfaces could be non-native. Figure 5.3 shows a few examples of failed AFm predictions including rigid (2FJU[20]), medium (5VNW[21]) and flexible targets (1IB1[22], 2FJG[23]). In all the examples, the AFm model (highlighted in *red* to *blue* based on residue-wise pLDDT) is superimposed over an individual binding partner, and the bound structure is highlighted in *pale-green*. AFm models predict the individual subunits (protein partners) accurately in almost all scenarios, however the docking orientation is incorrect.

I investigated whether any of the AlphaFold predictive metrics could be repurposed for distinguishing native-like binding sites from non-native ones. That is, can one could utilize pLDDT or PAE from AFm models to determine whether the predicted docked complex has the accurate binding orientation? To evaluate whether a predicted model lies in the near-native binding region or not, I utilized the DockQ score, the standard metric for docking model quality.[24] DockQ ($\in [0, 1]$) combines interface RMSD (Irms), fraction of native-like contacts (f_{nat}), and Ligand-RMSD (Lrms). DockQ scores above 0.23 correspond to models with CAPRI quality acceptable or higher. As an acceptable quality target implies docked decoys are in the near-native binding region, I chose a binary classification of success with a threshold of DockQ $= 0.23$. I then tested how well DockQ correlated with several AFm-derived metrics:

(a) Interface residues: the number of interface residues (atoms of residues on one partner within 8 Å from an atom on another partner);

(b) Interface contacts: the number of interface contacts between the residues on the interface (Cβ within 5 Å);

(c) Average pLDDT, determined by averaging over the per-residue LDDT score of the entire protein complex;

(d) Interface-pLDDT, determined by averaging the per-residue LDDT score only over the predicted interfacial residues (as identified in case *a*).

Figure 5.4A highlights the classification accuracy of each of these metrics with a receiver-operating characteristics curve. The interface-pLDDT metric stands out with a higher true positive rate (TPR) with an area under curve (AUC) of 0.86. With interface-pLDDT as a discriminating metric, I set an interface-pLDDT of 85 as the cut-off to estimate its accuracy and precision at distinguishing near-native structures (defined as an interface-RMSD < 4 Å). Figure 5.4B summarizes the performance with a confusion matrix. 80% of the targets are classified accurately with a precision of 78%, thereby validating the utility of interface-pLDDT as a discriminating metric to rank the docking quality of the AFm complex structure predictions. This discrimination is also evident in the highlighted interface residues in Figure 5.3, where the AFm predicted models have lower confidence at predicted interfaces (highlighted by *red*). Finally, I show the trend between DockQ scores and interface-pLDDT for each target in Figure 5.4C. The interface-pLDDT threshold of 85 (*dashed line*) thus can serve as the AlphaFold-derived metric to distinguish acceptable quality docked predictions from incorrect models.

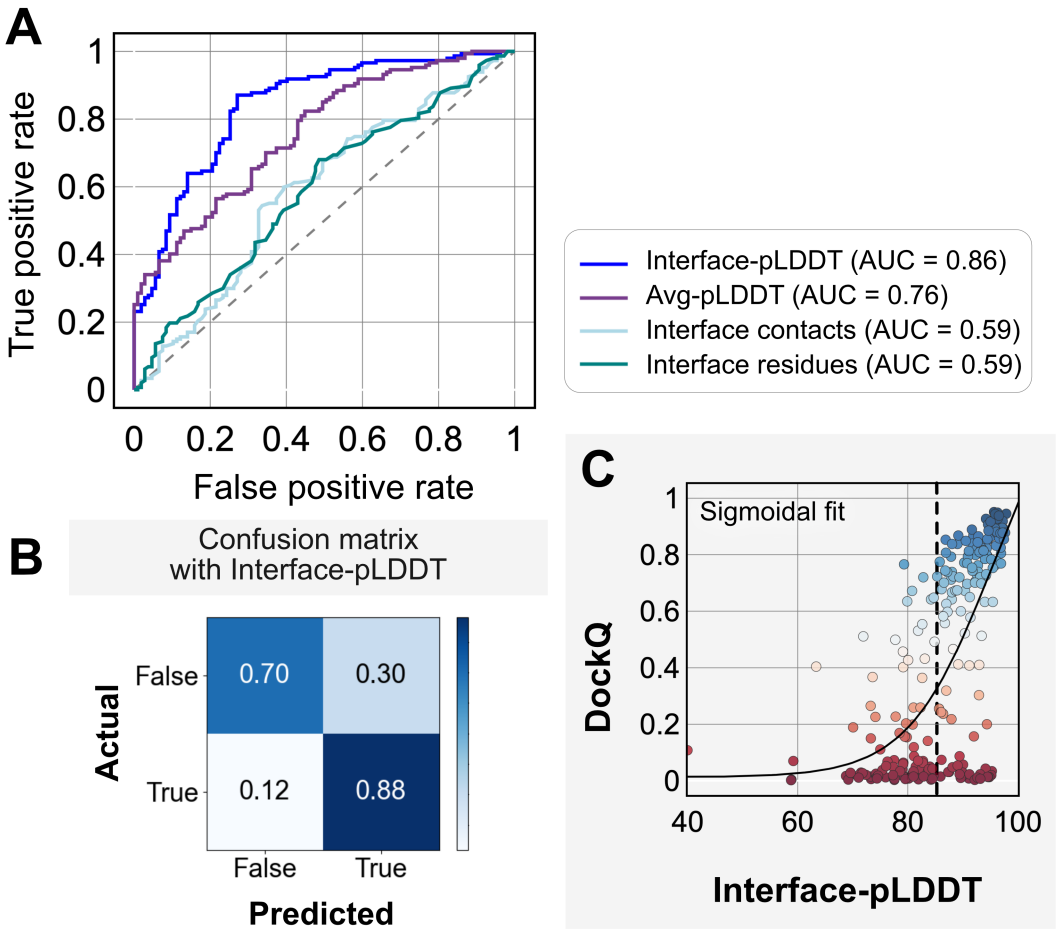


Figure 5.4: Interface-pLDDT is the best indicator of model docking quality. (A) Receiver-operator characteristics (ROC) curve as a function of different metrics for the docking dataset (n=254). Interface residues are defined based on whether atoms of residues on one partner are within 8 Å from atom/s on another partner. Interface-pLDDT is the average pLDDT of interface residues. Avg-pLDDT corresponds to the average pLDDT across all the residues in the predicted model. Interface contacts and interface residues are the counts of the interface contacts and interface residues respectively. Interface-pLDDT has the highest AUC score of 0.87. (B) Confusion matrix with an interface-pLDDT threshold between labels predicted false (<85) and true (≥85) and an interface-RMSD threshold between labels actually true (≤4 Å) and false(>4 Å) actual labels. (C) Interface-pLDDT versus DockQ for all protein targets in the benchmark set. DockQ is calculated from the predicted AlphaFold structure and the experimental bound structure in the PDB. I fit a sigmoidal curve to this available data.

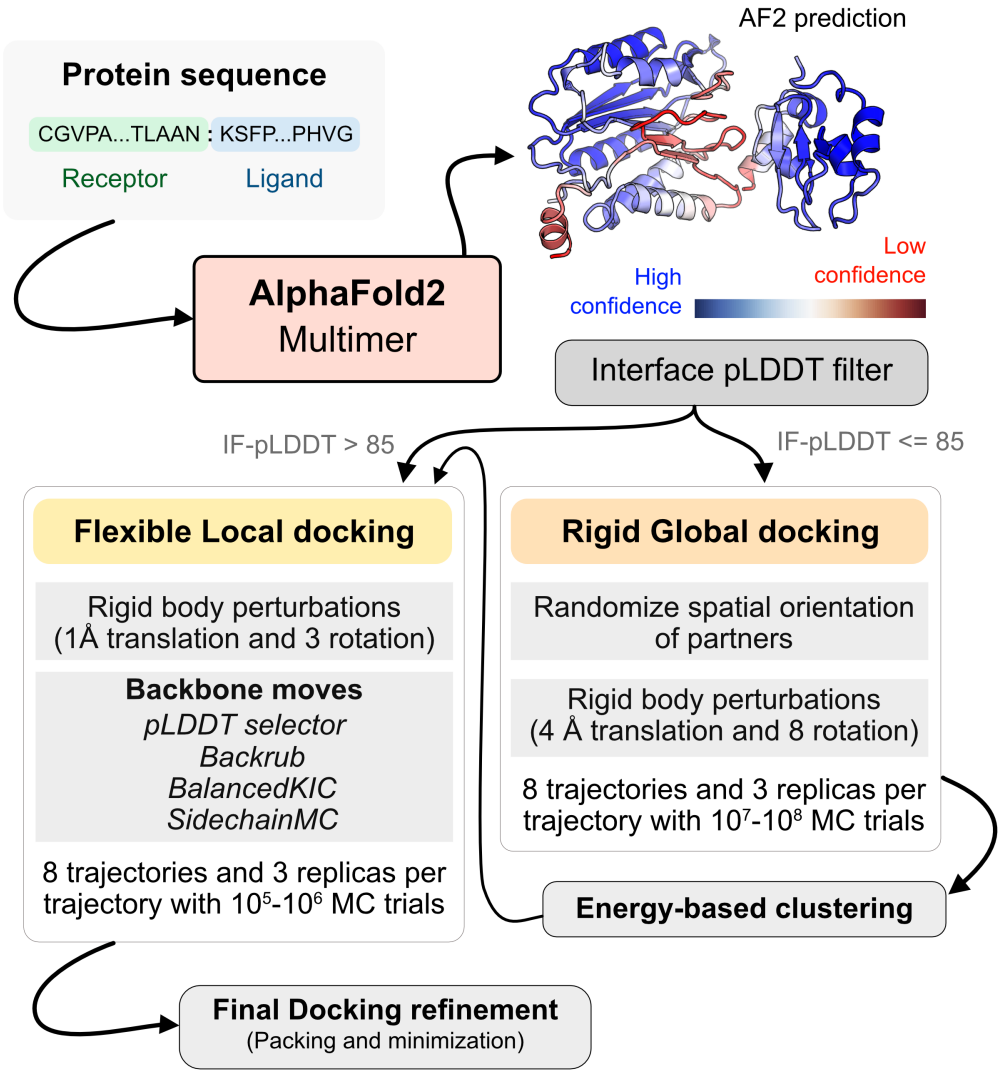


Figure 5.5: AlphaRED protein docking pipeline. Starting with protein sequences of putative complexes, I obtain predicted models from AlphaFold. Each model is accompanied with pLDDT scores, and based on the interface pLDDT I either initiate global rigid-body docking (interface pLDDT < 85), or flexible local docking(interface pLDDT ≥ 85). For global rigid-body docking, the protein partners are first randomized in Cartesian coordinates and then docked with rigid-backbones using temperature replica exchange docking within ReplicaDock2.[2] These decoys are then clustered based on energy clustering before flexible local docking refinement. In flexible local docking, I initiate a directed induced-fit strategy elaborated in ReplicaDock2. The residues are selected as identified by the AlphaFold residue-wise pLDDT scores (threshold of 80). The protocol moves the backbones with Rosetta's Backrub or Balanced Kinematic Closure movers. Output structures are then refined and top scoring structures based on interface energy are selected.

5.3.4 Docking over AlphaFold models improves performance over benchmark targets

With metrics to identify the flexible regions in the protein and the docking accuracy of generated docked models, I next fused AlphaFold-multimer (AFm) with our docking protocol, ReplicaDock 2.0[2], to build a protocol for: (1) improving on incorrect AF docking predictions and producing alternate, near-native binding models and (2) capturing backbone conformational changes with our induced-fit protocol ReplicaDock2.0[2]. With AFm as the structural module translating protein sequences to structure, I create a protocol namely AlphaRED (AlphaFold-inspired Replica Exchange Docking). AlphaRED uses AFm predicted structures as the primary template, estimates docking accuracy metrics, and initiates global docking or refinement protocols as required.

Figure 5.5 illustrates this docking pipeline. Starting from AFm predicted model, I first calculate the interface-pLDDT to determine the docking scheme to follow. If the AFm model is not accurate (interface pLDDT < 85), I initiate a global docking simulation to explore the protein conformational landscape and identify putative binding sites. On the other hand, if the interface-pLDDT > 85 for the AFm predicted model, the docked complex is likely in the correct binding orientation. This implies the global docking stage of the protocol can be skipped and local docking simulations can be directly initiated from the complex coordinates. Global docking follows an exhaustive, rigid-body (no backbone moves) search between the protein partners to sample putative landscapes in the energy landscape. An unbiased global docking simulation is initiated by randomizing the spatial orientation of protein partners from the input structure. The replica exchange MC routine ReplicaDock 2.0 performs rigid-body rotations (8°) and translations (4 Å). Sampled decoys are clustered from

all replicas (based on energies and structural similarity) and the five top clusters are passed along for flexible local docking.

For flexible local docking, I perform aggressive backbone moves (backrub + kinematic closure, *refer Methods*) on candidate encounter complexes (clustered decoys), with fine rigid-body rotations and translations. To narrow conformational sampling, backbone moves are explicitly performed over residues identified as 'mobile' based on the per-residue pLDDT metric (residue pLDDT < 80). Unlike ReplicaDock 2.0 that performs induced-fit over putative interfaces, this approach targets backbone motions over these predicted mobile residues, reducing the sampling space. Local docking decoys are further refined for side-chain packing and minimization to obtain the docked structures (details in *Methods*). The methodological advancements and Rosetta movers in AlphaRED are further detailed in the *Methods* section.

Applying this strategy, I investigated AlphaRED's performance on all 254 benchmark targets (Figure 5.6). 97 targets under the threshold of interface-pLDDT (≤ 85) were passed to the global docking branch. Targets with interface-pLDDT over 85 were input for local docking and refinement. For all benchmark targets, I compared AlphaRED performance of the top-scoring decoys against initial AFm predicted complex structures. Figure 5.6A shows the interface-RMSD (Irms) of the AFm and AlphaRED predictions from the bound structure, respectively. The lower Irms values indicate that AlphaRED improves on existing predictions for almost all targets. For targets where AFm prediction is determined to be a failure (interface-pLDDT ≤ 85, *red*), AlphaRED demonstrates a vast improvement in Irms for 93 out of 97 targets. Additionally, for targets where AFm prediction is considered acceptable (interface-pLDDT > 85), local docking slightly improves performance. AlphaRED captures

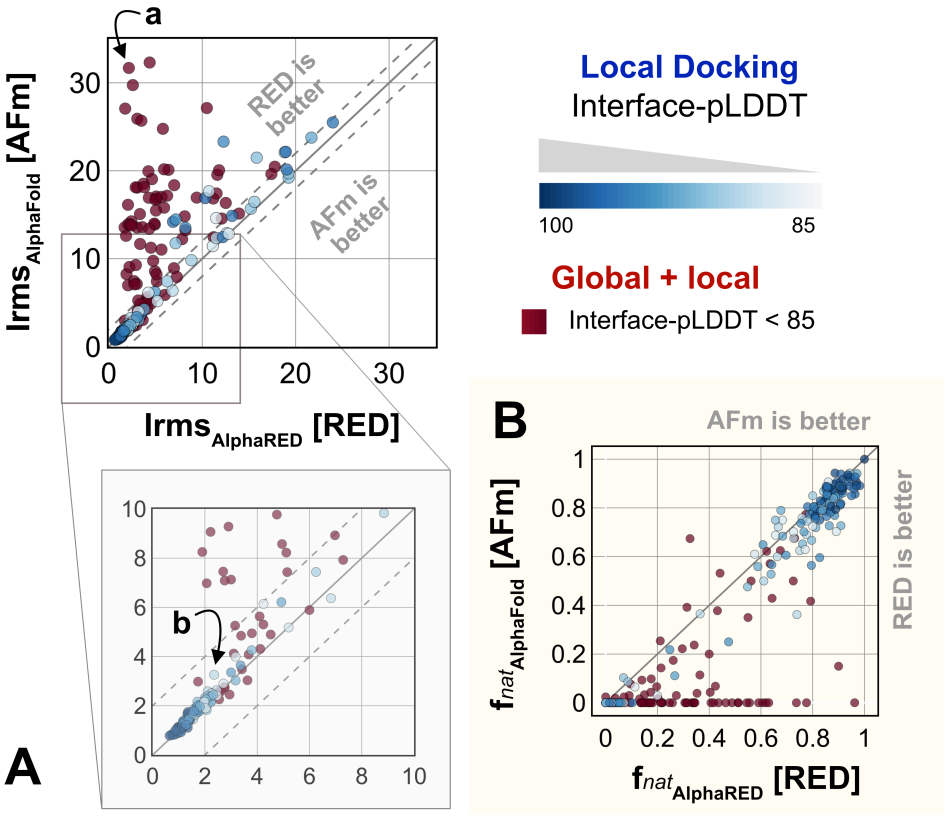


Figure 5.6: Docking performance Targets with Interface-pLDDT $\leq$ 85 were selected and passed in our docking pipeline (*in red*). Targets with interface-pLDDT$>$ 85 were passed for local refinement and are colored based on their interface-pLDDT scores (*in shades of blue*) (A) Interface-RMSD from AlphaFold-multimer predicted models (*y-axis*) in comparison with AlphaRED models(*x-axis*), with under 10 Å measurements in box. (B) Fraction of native-like contacts for models from AFm and AlphaRED respectively. (a) and (b) indicate two targets, 2FJU (global docking) and 5C7X (local docking) respectively highlighted in Figure Figure 5.7.

lower interface-RMSDs (under 10 Å) for targets where AFm models dock at binding sites ∼40 Å away. Figure 5.6B demonstrates the improvement in recapitulating native-like contacts (f_{nat}) with AlphaRED.

Figure 5.7 highlights a global docking (*a*) and local docking (*b*) example for targets 2FJU and 5C7X respectively. Starting from the AFm prediction (*orange*), AlphaRED samples over the conformational landscape to identify a top-scoring decoy (*blue*) with 2.6 Å Irms from the native (*gray*). Figure 5.7b shows the extent of backbone sampling with ReplicaDock 2.0 local docking. The top-scoring decoy (*blue*) samples backbone closer to the bound form improving model quality and docking accuracy for protein target 5C7X.

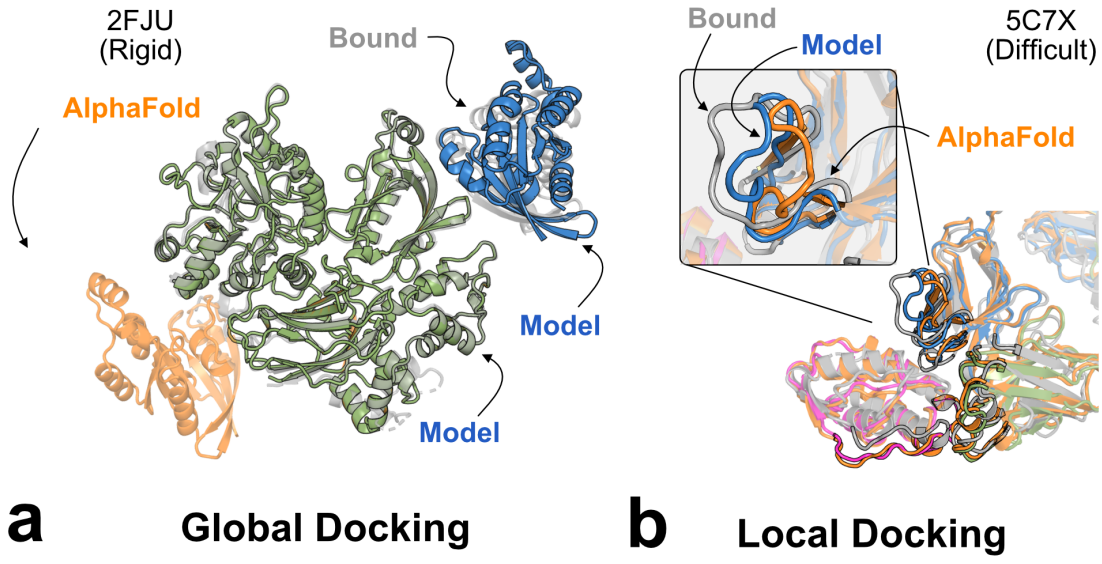


Figure 5.7: Global and local docking performance Docking performance for targets (a) Activated Rac1 bound to phospholipase Cβ2 (2FJU), and (b) Neutralizing anti-human antibody Fab fragment in complex with human GM-CSF (5C7X). Starting from the AFm model (*orange*), global docking performance on 2FJU highlights the improvement in sampling the native-like binding site (*gray*) by sampled decoy (*blue*). For local docking, backbone sampling on mobile residues predicted by residue pLDDT (*highlighted cartoon representation*) shows AlphaRED decoy (*blue*) moves backbone towards bound form(*gray*).

5.3.5 Evaluation on blind CASP15 targets

All results presented thus far may be biased by the fact that these benchmark target structures were used in the AFm training, potentially biasing the outcome of our benchmarking. The ultimate challenge for protein structure prediction protocols is to perform successfully over blind targets such as those in CASP (Critical Assessment of protein Structure Prediction) or CAPRI (Critical Assessment of PRotein Interactions) competitions.[25,26] CASP15 (Summer 2022) provided multiple protein docking targets[16], that were not included in AFm training, aloowing an unbiased evaluation of our AlphaRED pipeline. Thus, I tested the protocol on the five heterodimeric nanobody-antigen complexes where most of the groups performed poorly (Figure 5.8).

For each target, I employed the AlphaRED strategy as described in Figure 5.5. All targets had low interface-pLDDT thereby demanding global docking. This is unsurprising since the targets were nanobody-antigen targets and their CDRs, particularly CDR H3, are not conserved with a scarcity of co-evolution data with the antigen.[27] For target T205, our docking strategy improves the performance drastically (interface RMSD 11.4 Å for AFm model to 2.84 Å for AlphaRED) and binds in the ideal binding site with lower energies. The interface scores versus interface-RMSD plot shows a distinct funnel with low-energy medium-quality structures (Figure 5.8-*top*). Since the crystal structures are not yet released, the reference structure here is the top-model predicted for each category in CASP15. For all the targets, I can visualize how the docking strategy samples the appropriate binding orientation. These cases validate our strategy for blind targets, and demonstrate the ability of AlphaRED to serve as a robust pipeline, integrating AlphaFold with biophysical attributes to better predict

protein complex structures.

5.4 Discussion and conclusions

AlphaFold has dramatically transformed the field of structural biology and is currently the state-of-the-art method to predict protein structures from sequences, not just for monomers but also for complexes and higher assemblies.[28] One of the key elements of its success was the ability to mine evolutionary links between amino acids across protein families and determine structural templates. This approach dramatically improves prediction accuracy for monomers as reflected from prior CASP rounds. However, for larger assemblies and complexes, the evolutionary constraints can be weak and often skew predictions to inaccurate binding sites. Here I demonstrated how augmenting the predictions of AlphaFold with an energy-function dependent sampling approach reveals better backbone conformational diversity and provides accurate prediction of protein complex structures. By utilizing the AlphaRED strategy, I show that failure cases in AFm predicted models are improved for all targets (lower Irms for 97 failed targets) with upto CAPRI medium-quality models (DockQ $\in [0.49, 0.80]$) generated for 59% targets.

First, I showed that AlphaFold confidence measures can be repurposed for estimating flexibility and docking accuracy. This is useful for constructing an efficient docking pipeline. Interface-pLDDT, an average of the per-residue pLDDT only for the interfacial residues, is a robust metric to determine whether AFm predicted binding interfaces are correct. Additionally, thresholds of per-residue pLDDT can ascertain regions of backbone flexibility upon binding. Thus, AFm predicted models can be used as input structures for ReplicaDock 2.0 to dramatically improve sampling and

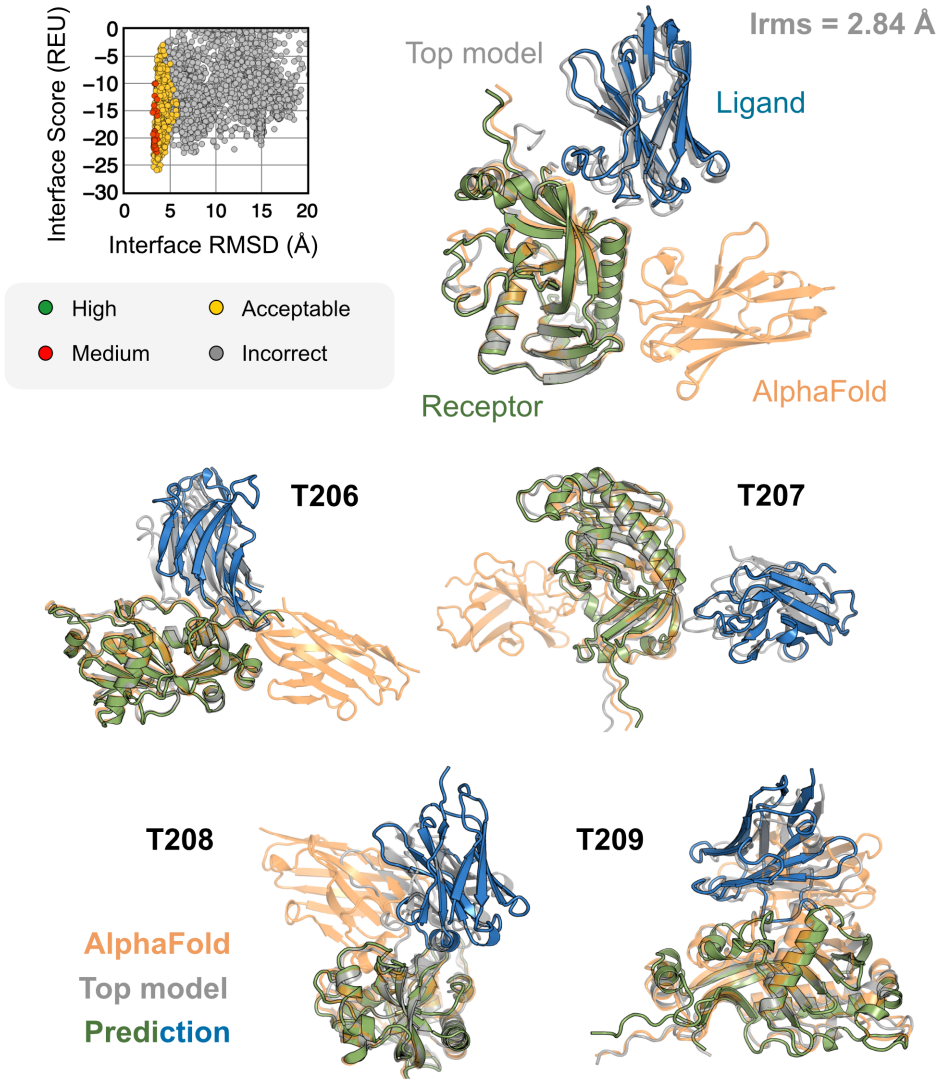


Figure 5.8: AFm and AlphaRED performance on CASP15 targets Docking performance for CASP targets T205-T209. (*top*) T205. Interface score (REU) vs Interface RMSD (Å) for candidate docking structures generated by the AlphaRED docking pipeline. (*top-right*) The top-scoring AlphaRED model (*green-blue*) recapitulates the native interface (*gray*) and has an interface RMSD of 2.84 Å. The distinction between the predicted model with respect to the AFm model(*orange*) is evident (*bottom*) Top-scoring AlphaRED predictions for targets T206, T207, T208 and T209 respectively.

performance over benchmark docking targets. I generated the AlphaFold template structures with the default settings (3 recycles and no dropout) to obtain a structure per target. With DL-methods for structure prediction and downstream sampling with a physics-based energy function, one can efficiently explore the protein energy landscape as demonstrated with AlphaRED performance over DB5.5. Finally, I evaluated recent CASP15 targets to investigate the extrapolation of this strategy over blind protein targets. CASP15 targets were absent from the training routine of AlphaFold and served as blind challenges to determine the efficacy of the protocol. With AlphaRED, I obtained DockQ$>$ 0.23 for all five targets, with medium-quality models (DockQ$>$ 0.49) for targets T205, T207, and T208 respectively. AFSample, a top-performing group in CASP15, employed stochastic perturbation with dropout and increased sampling to obtain medium and high-quality models for these targets. However, AFSample equips couple of GPU simulations to get $\sim$240x models with compute time exceeding to be $\sim$1000x costly than the baseline version. On other hand, I utilized ColabFold[7] to generate 1-5 structures for our docking routine with the baseline version. As opposed to a couple of days on GPU (each GPU node contains upto 48 cores) utilized by AFSample, our docking routine fused with ColabFold uses 5-7 hours on our CPU cluster (runs on 1 node, with 24 cores, approximating to $\sim$100 hours of CPU-hours per target). The AlphaRED docking strategy demonstrates a new and better way to predict protein complex structures within feasible compute times.

With this work, I have built upon the recent advances in structural biology to develop a robust tool for protein docking. Here, I fused deep-learning tools with conventional physics-based sampling tools to develop a pipeline that extracts the

best outcomes of each methodology; where deep-learning methods generate accurate, static structures, and physics-based sampling provides diversity and better discrimination. The protein conformational landscape is vast and deep-learning tools such as AlphaFold provide a snapshot of relevant local minima that can aid in narrowing down the degrees of freedom in sampling.[29] With the paradigm shift in computational structural biology towards deep-learning approaches, integrating physics within these models has tremendous potential towards understanding protein dynamics, modulating protein-protein interactions, and downstream applications to protein design.

5.5 Methods

5.5.1 Prediction of structures

For each target in the DB5.5 dataset, I obtained AlphaFold predicted models with the ColabFold v1.5.2[30] implementation of AlphaFold[3] and AlphaFold-multimer[4]. Each prediction run was performed without templates, with automatic alignments and the default number of recycles to generate five relaxed predictions. Each AlphaFold prediction includes a per-residue pLDDT (predicted LDDT) measurement[31], a confidence measure in prediction accuracy, and predicted template alignment (pTM) score.[32] The models were structurally compared with the unbound and bound structures (deposited in the PDB) for measuring flexibility, similarity and accuracy of docking prediction.

5.5.2 Metrics for backbone flexibility: RMSD and LDDT

Structures of proteins deposited in the PDB[33] provide a static representation of the native-state of the protein. However, structural diversity has been captured by experimental techniques to identify different states of a protein in diverse physiological or chemical states, for e.g. catalysis[34], transport[35], and ligand binding[36]. For protein docking challenges in particular, conformational changes are binding-induced, leading to structural differences between unbound and bound structures of protein targets.

To measure the conformational change in protein structures, I calculated two metrics: Cα root-mean-square-deviation (RMSD) and local distance difference test (LDDT)[31]. In order to get a detailed representation of the intrinsic motion of a protein, I calculated RMSDs at a residue-level, *i.e.*, per-residue Cα RMSD for each residue of a protein target. The sequences of unbound and bound proteins were aligned for ensuring robust measurements, and the RMSDs were calculated for the aligned residues. The total sequence lengths were also matched perfectly and lingering end-termini residues were chopped off to ensure structural and sequential similarity.

Local Distance Difference Test (LDDT) is a superimposition-free score that estimates local distance differences in a model relative to a reference structure.[31] Unlike the Global Distance Test (GDT)[37] score based on rigid-body superimposition, the LDDT score measures the conserved local interactions in the protein model to the reference. For every residue, it computes the distance between all pair of atoms $D(i, j)$ in both the model and the reference structure (bound) within a threshold (defined as the inclusion radius, generally set to 10 Å). For each pairwise distance in both distance vectors, if the distance is within the threshold, the distance is considered

conserved and the fraction of conserved distances is calculated. The final LDDT score is the average of this fraction for the tolerances of 0.5, 1, 2, and 4 Å.

For a protein structure with N number of residues, the overall LDDT score can be given as follows:

$$\text{Overall score} = \text{norm} \cdot \sum_{i,j}^{N} \text{dists\_to\_score}(i,j) \cdot \text{score}(i,j) \tag{5.1}$$

where norm is the normalization factor

$$\text{norm} = \frac{1}{\sum_{i,j} \text{dists\_to\_score}(i,j)} \tag{5.2}$$

and $\text{score}(i,j)$ is the LDDT score for the residue i with respect to every other residue j

$$\begin{aligned} \text{score}(i,j) = 0.25 \cdot \Big\{ & \; bool[\Delta D(i,j) < 0.5] + \\ & \; bool[\Delta D(i,j) < 1.0] + \\ & \; bool[\Delta D(i,j) < 2.0] + \\ & \; bool[\Delta D(i,j) < 4.0] \Big\} \end{aligned}$$

Here, $\Delta D(i,j)$ denotes the absolute difference between $D_{\text{true}}(i,j)$ and $D_{\text{predicted}}(i,j)$ calculated as follows:

$$\Delta D(i,j) = |D_{\text{true}}(i,j) - D_{\text{predicted}}(i,j)| \tag{5.3}$$

$D_{\text{true}}(i,j)$ and $D_{\text{predicted}}(i,j)$ denote the distances between the Cα coordinates of the i^{th} residue and the j^{th} residue for the true (reference) and predicted (model) structures respectively. Let x_i^k and y_i^k represent the k^{th} coordinate of the Cα atom in the i^{th} residue in the reference (true) structure and predicted structure respectively, such that:

$$D_{\text{true}}(i,j) = \sqrt{\sum_{k=1}^{3}(x_i^k - x_j^k)^2} \text{ and } D_{\text{predicted}}(i,j) = \sqrt{\sum_{k=1}^{3}(y_i^k - y_j^k)^2} \tag{5.4}$$

Finally, the distances to score (dists_to_score(i,j)) are computed as those pairwise distances within an inclusion radius (cutoff = 10 Å). m_i^j is the mask value (1 or 0) indicating if the j^{th} coordinate of the Cα atom in the i^{th} residue exists in the true structure.

$$\text{dists\_to\_score}(i,j) = \begin{cases} 1 & \text{if } D_{\text{true}}(i,j) < \text{cutoff} \cdot m_i^j \cdot m_j^i \cdot (1 - \delta_{jN}) \text{ where } \delta = \text{KroneckerDelta} \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

The advantage of the LDDT measurement lies in the estimation of relative domain orientations in multi-domain proteins or concerted motions (for e.g.: hinge-like moves in closed and apo proteins). In these cases, the RMSDs would be relatively high for all residues in the mobile domain, however, since the inter-residue distances within the domains are conserved, they would provide an inaccurate depiction of flexibility for the protein. Estimating both RMSDs and LDDT scores allows us to obtain a nuanced perspective of flexibility during protein association based on experimental structures.

5.5.3 Development of new ResidueSelectors in Rosetta

Protein structures deposited in the PDB often have column for temperature factors/B-factors (B_i) highlighting the temperature dependent flexibility of residues. AlphaFold

predicted structures output their per-residue pLDDT measurement in this column. As highlighted in prior chapters (Chapter 2 and 3), Rosetta utilizes `ResidueSelectors` to perform selections at the protein `pose` level. The residue selections aid in narrowing down specific regions for `movers` to operate their moves within the Rosetta pipeline. I use this strategy for mimicking induced-fit in ReplicaDock2.0[2] and directing induced-fit towards flexible residues as highlighted in Chapter 2, section on *directed induced fit*. As illustrated earlier, one of the challenges in employing directed induced-fit for protein targets is the inadequacy in determining 'flexible' regions in proteins. With AlphaFold metrics however, there is a potential in utilizing pLDDT as a determinant of backbone flexibility and employing it in conjunction with our docking routines. To automate this, I created the `BFactorResidueSelector` within Rosetta that can allow the selection of residues with lower pLDDT values. These selected residues can then be passed in our docking routine as 'flexible' residue units on which I perform aggressive backbone sampling. BFactorResidueSelector is an integral part of our docking pipeline mentioned hereafter and has utility for automating our docking routine on ROSIE server. Details about the implementation are available on *github.com/RosettaCommons*.

5.5.4 Developing a pipeline for protein docking

Using AlphaFold2 as a structural module, I built a pipeline for protein-protein docking to better predict protein complex structures with relatively higher accuracy. As illustrated in Figure 5.5, given a sequence of a protein complex, I use the ColabFold implementation of AF2-multimer to obtain a predictive template. An interface-pLDDT filter determines the accuracy of the docking prediction of the top-ranked model from AFm. If the interface-pLDDT ≤ 85, the prediction has lower confidence in the docking

orientation, and the protocol initiates a rigid, global docking search with ReplicaDock 2.0. Implementation of ReplicaDock 2.0 (global docking) is similar to the version reported in prior work[2]. Each simulation initiates 8 trajectories across 3 temperature replicas with inverse temperatures set to 1.5^{-1} kcal^{-1}.mol, 3^{-1} kcal^{-1}.mol and 5^{-1} kcal^{-1}.mol, respectively. Across each replica within each trajectory, rigid body perturbations (4 Å translations and 8° rotations) are performed for an exhaustive global search. Next, I perform an energy-based clustering of the models to obtain diverse and energetically favourable clusters. Five cluster centers (decoys) are selected and passed to the flexible local docking stage to sample conformational changes.

On other hand, if the interface-pLDDT > 85, the binding orientation has higher confidence and the protocol directly performs a flexible local docking simulation skipping the rigid, global docking. In this stage, I perform smaller rigid-body perturbations (1 Å translations and 3° rotations) and aggressive backbone moves using a set of backbone and side-chain movers: Rosetta `Backrub`[38], `Balanced Kinematic Closure` (BalancedKIC) and `Sidechain`. The sampling weights are biased such that backbone and side-chain movers are weighted higher than rigid body moves (3:1 weightage for backbone:rigid-body moves). I perform directed backbone sampling by focusing on predicted mobile residues (per residue pLDDT < 80). This is automated with the `BFactorResidueSelector` that selects contiguous sets of residues below the specified pLDDT threshold.

However, unlike the induced-fit strategy in ReplicaDock[2], I perform directed backbone sampling directed at the mobile residues (with per residue pLDDT < 80) identified from the AlphaFold model. I automate this using the `BFactorResidueSelector` to select contiguous sets of residues below the specified pLDDT threshold in the

prior section. This residue subset is passed along to the backbone movers to sample backbone moves along with small rigid-body moves. Sampled decoyed are then refined, *i.e.* undergo side-chain packing and minimization, to output docked decoys. Best ranked decoys based on interface scores are then identified as the top-scoring structures.

References

1. Yin, R., Feng, B. Y., Varshney, A. & Pierce, B. G. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Science* **31,** e4379. ISSN: 0961-8368. https://doi.org/10.1002/pro.4379 (8 2022).
2. Harmalkar, A., Mahajan, S. P. & Gray, J. J. Induced fit with replica exchange improves protein complex structure prediction. *PLOS Computational Biology* **18,** 1–21. https://doi.org/10.1371/journal.pcbi.1010124 (6 2022).
3. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589. ISSN: 14764687. http://dx.doi.org/10.1038/s41586-021-03819-2 (7873 2021).
4. Evans, R., Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Ž´, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J. & Hassabis, D. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021).
5. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876. ISSN: 0036-8075 (6557 2021).
6. Wallner, B. AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling. *bioRxiv.* https://www.biorxiv.org/content/early/2022/12/21/2022.12.20.521205 (2022).

7. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. ColabFold: making protein folding accessible to all. *Nature Methods* **19,** 679–682. ISSN: 1548-7105. https://doi.org/10.1038/s41592-022-01488-1 (6 2022).

8. Alamo, D. D., Sala, D., McHaourab, H. S. & Meiler, J. TITLE: Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11.** ISSN: 2050084X (2022).

9. Wayment-Steele, H. K., Ovchinnikov, S., Colwell, L. & Kern, D. Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. *bioRxiv*, 2022.10.17.512570. http://biorxiv.org/content/early/2022/10/17/2022.10.17.512570.abstract (2022).

10. Saldanõ, T., Escobedo, N., Marchetti, J., Zea, D. J., Donagh, J. M., Rueda, A. J. V., Gonik, E., Melani, A. G., Nechcoff, J. N., Salas, M. N., Peters, T., Demitroff, N., Alberti, S. F., Palopoli, N., Fornasari, M. S. & Parisi, G. Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* **38,** 2742–2748. ISSN: 14602059 (10 2022).

11. Marze, N. A., Burman, S. S. R., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics* **34,** 3461–3469. ISSN: 14602059 (20 2018).

12. Yan, Y., Tao, H., He, J. & Huang, S.-Y. The HDOCK server for integrated protein–protein docking. *Nature Protocols* **15,** 1829–1852. ISSN: 1750-2799. https://doi.org/10.1038/s41596-020-0312-x (5 2020).

13. Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., Chaleil, R., Jiménez-García, B., Bates, P. A., Fernandez-Recio, J., Bonvin, A. M. & Weng, Z. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology* **427,** 3031–3041. ISSN: 10898638. http://dx.doi.org/10.1016/j.jmb.2015.07.016 (19 2015).

14. Ganea, O., Huang, X., Bunne, C., Bian, Y., Barzilay, R., Jaakkola, T. S. & Krause, A. Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking. *CoRR* **abs/2111.07786.** arXiv: 2111.07786. https://arxiv.org/abs/2111.07786 (2021).

15. McPartlon, M. & Xu, J. Deep Learning for Flexible and Site-Specific Protein Docking and Design. *bioRxiv*. https://www.biorxiv.org/content/early/2023/04/02/2023.04.01.535079 (2023).

16. CASP15. *15th Community wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction* https://predictioncenter.org/casp15/index.cgi (2022).

17. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J. & Hassabis, D. Highly accurate protein structure prediction for the human proteome. *Nature* **596,** 590–596. ISSN: 14764687 (7873 2021).

18. Song, H., Hanlon, N., Brown, N. R., Noble, M. E. M., Johnson, L. N. & Barford, D. Phosphoproteinx2013;Protein Interactions Revealed by the Crystal Structure of Kinase-Associated Phosphatase in Complex with PhosphoCDK2. *Molecular Cell* **7,** 615–626. ISSN: 1097-2765. https://doi.org/10.1016/S1097-2765(01)00208-8 (3 2001).

19. Huse, M., Chen, Y.-G., Massagué, J. & Kuriyan, J. Crystal Structure of the Cytoplasmic Domain of the Type I TGF x3b2; Receptor in Complex with FKBP12. *Cell* **96,** 425–436. ISSN: 0092-8674. https://doi.org/10.1016/S0092-8674(00)80555-3 (3 1999).

20. Jezyk, M. R., Snyder, J. T., Gershberg, S., Worthylake, D. K., Harden, T. K. & Sondek, J. Crystal structure of Rac1 bound to its effector phospholipase C-2. *Nature Structural Molecular Biology* **13,** 1135–1140. ISSN: 1545-9985. https://doi.org/10.1038/nsmb1175 (12 2006).

21. McMahon, C., Baier, A. S., Pascolutti, R., Wegrecki, M., Zheng, S., Ong, J. X., Erlandson, S. C., Hilger, D., Rasmussen, S. G. F., Ring, A. M., Manglik, A. & Kruse, A. C. Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nature Structural Molecular Biology* **25,** 289–296. ISSN: 1545-9985. https://doi.org/10.1038/s41594-018-0028-6 (3 2018).

22. Vetter, I. R., Arndt, A., Kutay, U., Görlich, D. & Wittinghofer, A. Structural View of the Ranx2013;Importin β; Interaction at 2.3 Å Resolution. *Cell* **97,** 635–646. ISSN: 0092-8674. https://doi.org/10.1016/S0092-8674(00)80774-6 (5 1999).

23. Fuh, G., Wu, P., Liang, W.-C., Ultsch, M., Lee, C. V., Moffat, B. & Wiesmann, C. Structure-function studies of two synthetic anti-vascular endothelial growth factor Fabs and comparison with the Avastin Fab. *The Journal of biological chemistry* **281,** 6625–6631. ISSN: 0021-9258 (Print) (10 2006).

24. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE* **11,** 1–9.

25. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics* **89,** 1607–1617. https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26237 (12 2021).

26. Lensink, M. F. *et al.* Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics*, 1–24. ISSN: 0887-3585 (2021).

27. Adolf-Bryfogle, J., Xu, Q., North, B., Lehmann, A. & Jr, R. L. D. PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Research* **43,** D432–D438. ISSN: 0305-1048. https://doi.org/10.1093/nar/gku1106 (D1 2015).

28. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications* **13,** 1265. ISSN: 2041-1723. https://doi.org/10.1038/s41467-022-28865-w (1 2022).

29. Roney, J. P. & Ovchinnikov, S. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Physical Review Letters* **129,** 238101. https://link.aps.org/doi/10.1103/PhysRevLett.129.238101 (2022).

30. Ovchinnikov, S. *ColabFold online* https://github.com/sokrypton/ColabFold (2021).

31. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29,** 2722–2728. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btt473 (21 2013).

32. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57,** 702–710. ISSN: 0887-3585. https://doi.org/10.1002/prot.20264 (4 2004).

33. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242. ISSN: 0305-1048. https://doi.org/10.1093/nar/28.1.235 (1 2000).

34. Kingsley, L. J. & Lill, M. A. Substrate tunnels in enzymes: Structure–function relationships and computational methodology. *Proteins: Structure, Function, and Bioinformatics* **83,** 599–611.

35. Gora, A., Brezovsky, J. & Damborsky, J. Gates of Enzymes. *Chemical Reviews* **113,** 5871–5923.

36. Gunasekaran, K. & Nussinov, R. How Different are Structurally Flexible and Rigid Binding Sites? Sequence and Structural Features Discriminating Proteins that Do and Do not Undergo Conformational Change upon Ligand Binding. *Journal of Molecular Biology* **365,** 257–273. ISSN: 0022-2836.

37. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research* **31,** 3370–3374. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkg571 (13 2003).

38. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of molecular biology* **380,** 742–756. ISSN: 1089-8638 (Electronic) (4 2008).

Chapter 6

Modeling translocation of bacteriocins though cellular nutrient transporters

6.1 Overview

Decades of excessive use of readily available antibiotics has generated a global problem of antibiotic resistance and, hence, an urgent need for novel antibiotic solutions. Bacteriocins are protein-based antibiotics produced by bacteria to eliminate closely related competing bacterial strains. Colicins are a type of these bacteriocins deployed by *Escherichia coli* that exploit outer membrane (OM) nutrient transporters to penetrate the selectively permeable bacterial cell envelope. In this work, *de novo* Rosetta modeling and live-cell fluorescence imaging uncovers the entry of the pore-forming

toxin colicin B (ColB) into *E.coli* and localizes it within the periplasm. Further, by coupling single-stranded DNA to ColB, the colicin B-translocation pathway has utility as an import route to deliver conjugated DNA cargo into bacterial cells. By applying a combination of photoactivated cross-linking, mass spectrometry, and structural modeling, this work characterizes the molecular mechanism of ColB associated with its OM receptor FepA. The association of ColB with FepA is coincident with large-scale conformational changes in the colicin. Thereafter, active transport of ColB through FepA involves the colicin taking the place of the N-terminal half of the plug domain that normally occludes this iron transporter.

6.2 Introduction

Bacteria are the most common and diverse form of life on earth. The remarkable abundance of different bacterial strains and species capable of surviving in almost any environment frequently leads to competition for space and resources.[1] Competition for scarce nutrients has led to the evolution of nutrient uptake systems, such as the secretion of siderophores to chelate bio-available iron, with the iron-siderophore complex captured by high-affinity receptors and actively transported across the cell envelope.[2] Competing bacteria also deploy weapons in the form of enzymes targeting either components of the cell wall or nucleic acids[3] or depolarizing pores that disrupt the electrochemical potential across the inner membrane.[4] Elimination of competing bacteria while kin bacteria are unharmed is achieved through the coexpression of toxin-specific immunity proteins that render the toxin inactive within producing strains.[5] Cytotoxic proteins can be delivered either in a contact-dependent manner, targeting neighboring cells relying on the assembly of supra-molecular machineries[6],

or through secretion into the milieu as exemplified by bacteriocins.[7]

Colicins, the bacteriocins of *E. coli*, have been extensively studied, with over 20 different examples described.[8] Once released, colicins breach the envelope of their target cell to elicit their cytotoxic activity.[9] The cell envelope of gram-negative bacteria is comprised of an asymmetric outer membrane (OM) with an outer leaflet comprised of lipopolysaccharide and a phospholipid inner leaflet, providing a robust layer of defense surrounding the energized inner membrane (IM) and the intervening periplasm.[10] Colicins are large (29 to 75 kDa) proteins that cannot diffuse through the cell envelope of their target cell[11] and must find a route across the OM.[12] Unlike the proton-motive force (PMF) of the IM, the OM is not directly energized, and energy-dependent processes at the OM such as protein import are coupled to the IM through transperiplasmic complexes. The Tol-Pal system, composed of the TolQ-TolR-TolA complex in the IM, TolB in the periplasm, and Pal anchored to the inner leaflet of the OM, stabilizes the OM during cell division.[13] The structurally related Ton system, composed of the TonB-ExbB-ExbD complex in the IM, powers active transport of nutrients such as siderophores through specialized TonB-dependent receptors in the OM.[14] Both the Tol-Pal and the Ton systems are exploited by colicins to energize their translocation across the cell envelope.

Colicins typically contain three structural domains, a central receptor (R)-binding domain, which anchors the toxin to the cell surface, an N-terminal translocation (T) domain implicated in OM translocation via the Tol-Pal, or a Ton system of a C-terminal cytotoxic domain. Colicin B (ColB) is a pore-forming toxin that was one of the earliest colicins to be described.[15] However, little is known about the cellular translocation process of ColB beyond its dependence on the OM ferric enterobactin

transporter FepA and the Ton system.[16] No additional OM proteins have been identified for ColB toxicity, which may explain why, unlike most other colicins, ColB is composed of only two functional domains: an N-terminal domain that serves as both a receptor-binding domain and a translocation domain (ColB-RT) and a pore-forming, C-terminal cytotoxic domain.[17] The ColB receptor FepA is a 22-stranded b-barrel TonB-dependent transporter (TBDT) with an N-terminal plug domain blocking its lumen.[18–20]

Here, I elucidate the mechanism by which ColB interacts with its receptor FepA and its active transport across the OM. The translocation of the ColB-RT domain to the periplasm of *E. coli* was visualized experimentally by applying live-cell fluorescence microscopy, demonstrating that translocation requires FepA at the OM and depends on colicin's TonB box. This work applies combined approach of *in vitro* and *in vivo* photoactivated cross-linking, mass spectrometry, and structural modeling with Rosetta to monitor the key stages in the ColB-FepA association process. Further, the import route of ColB can be used to import large macromolecules, in this instance single-stranded DNA (ssDNA), into bacteria.

6.3 Results

Previous work has demonstrated that *Pseudomonas aeruginosa*-specific pyocins are capable of transporting fluorophores into target cells.[21,22] Whether this is also possible for colicins and E. coli has yet to be determined. In this chapter, I discuss our computational approach to model the translocation encompassing large-scale conformational change in the ColB and the plug-motion of the FepA. Further, experimental validation demonstrates how this import pathway could be employed for transporting

conjugated single-stranded DNA (ssDNA) into bacteria via FepA.

6.3.1 Computational strategy to model ColB-FepA interactions and translocation

Simulations of protein-protein association and dissociation provide an atomistic view of biological events. However, modeling large-scale conformational changes in biologically relevant association events is infeasible with conventional molecular dynamics (MD) or Monte Carlo (MC) trajectories, owing to longer lifetimes of such events and limitations to simulation timescales. Here, I present a computational strategy to model the ColB-FepA interaction leading to the eventual translocation of ColB through FepA. It is hypothesized that the primary interaction between colicins (ColB) and OM proteins (FepA) induces changes in the structural motifs of both interacting partners eventually leading to translocation. Based on this hypothesis, I developed a computational pipeline as demonstrated in Figure 6.1.

First, I determined putative binding conformations by performing global docking with limited interfacial backbone moves with ReplicaDock 2.0.[25] Upon identifying potential clusters, such that each cluster can be attributed to a putative encounter complex, I next incorporated higher scale of flexibility with Rosetta FloppyTail[24] protocol. Prior studies demonstrated that the N-terminal domain of ColB serves as the receptor-binding domain and induces translocation. To mimic that characteristic of the N-terminal domain, I next model large-scale conformational changes over the docked decoys by extensive backbone sampling over the flexible domain (highlighted in Figure 6.1, *left panel, in blue*). The sampled backbones are further relaxed, docked with small rigid body perturbations (1 Å translation and 2° rotations), and refined to output the top decoys. Finally, to model the translocation pathway, I emulated the FepA

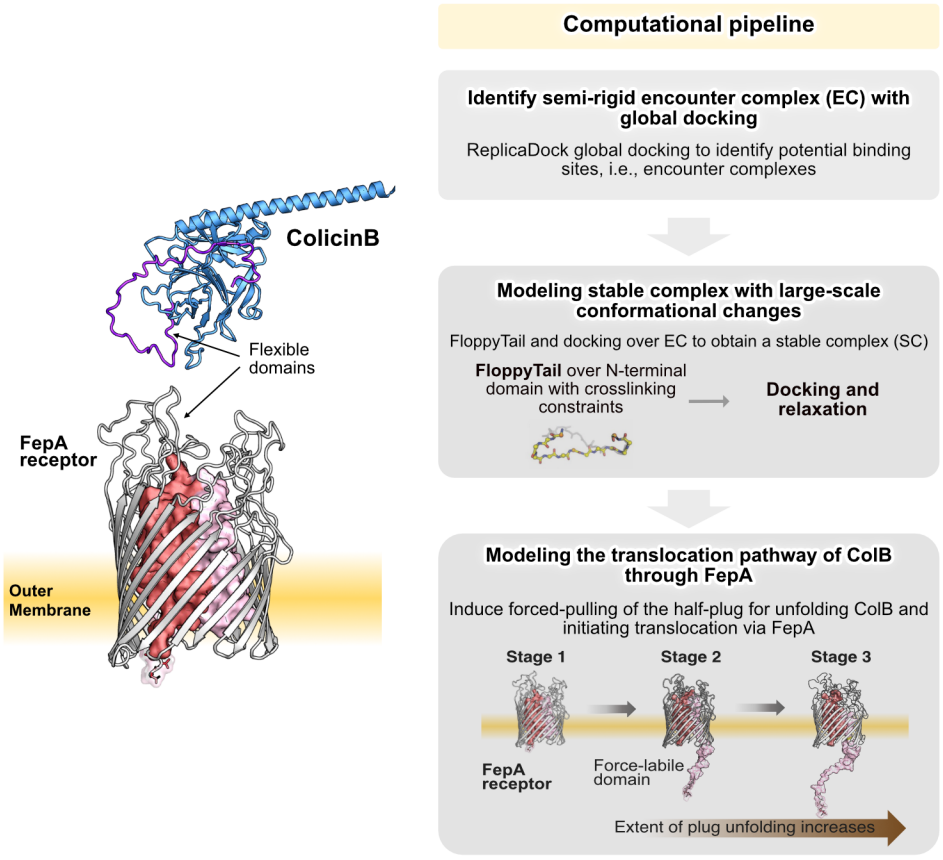


Figure 6.1: Overview of the computational strategy. (*left*) Schematic illustration of ColB (PDB: 1RH1[23]) with outer-membrane receptor FepA (PDB: 1FEP[18]). The flexible domains over ColB are highlighted in *blue*. (*right*) Computational pipeline initiated on the ColB and FepA structures illustrated on *left*. First, a global docking simulation is initiated to determine putative encounter complexes. Then, the flexible N-termini region is modelled with Rosetta FloppyTail[24] followed by docking and relaxation to predict the structure of the stable complex. Finally, translocation stages are modelled by pulling the half-plug domain of the FepA receptor with flexible modeling of ColB and crosslinking constraints.

half-plug unfolding during docking with ColB. Inspired from the TonB-BtuB system[26] that follows a similar TonB-like translocation pathway, the half-plug domain of FepA (residues 1-75) and the N-terminal domain of ColB were modeled with backbone flexibility. As ergodic sampling with MC for this large-scale conformational change would be computationally demanding, I captured the dynamic-unfolding process by instead capturing three putative stages of the unfolding pathway. Briefly, the half-plug domain demonstrated to be labile was pulled into the membrane to generate three models, one with the plug slightly unfolded (N-termini displaced by 4 Å), second with the half-plug unfolded partially (N-termini displaced by 8 Å), and third with the complete unfolding of the half-plug (N-termini displaced by 12 Å). Experimental evidence provided us with cross-linking residues on ColB-FepA interface that retained contact throughout the translocation. Using these residues as constraints (denoted as *crosslinking constraints*), backbone sampling of the N-termini domain of ColB was initiated. Crosslinking constraints were defined based on Cα distance between predicted interacting residues with a flat harmonic penalty. After aggressive backbone sampling (generating over 5000 models for each model), decoys were filtered and evaluated based on interface scores of the structures. The generated structures were assessed for their thermodynamic favorability relative to conformations of other models presented prior and were utilized to hypothesize the unfolding-translocation pathway.

6.3.2 Receptor FepA binding induces large-scale conformational changes in ColB

Many colicins bind multiple OM proteins or even multiple copies of the same OM protein.[9] To address the question of whether additional proteins or copies of FepA

were required for ColB transport, ColB complexes were assembled on the surface of ColB-sensitive *E. coli*, detergent extracted, and purified by nickel affinity, followed by size exclusion chromatography. Native mass spectrometry of this isolated complex revealed the FepA-ColB complex to have a 1:1 stoichiometry, consistent with ColB binding and translocating through a single copy of FepA. Since there are no available structures for the ColB-FepA complex, modeling the ColB-FepA interaction with Rosetta was the only feasible alternative. To understand how colicin associates with TBDT, I equipped docking approaches exploiting available PDB structures of unbound ColB (PDB: 1RH1[23]) and FepA (PDB: 1FEP[18]). The structures were initially positioned with the FepA extracellular loops facing the predicted ColB receptor-binding loops.[27] This calculation revealed a clear energy funnel for an encounter complex (EC) structure (Figure 6.2.A). As an independent test of the Rosetta model predictions, experimental validation were performed with pBPA cross-linking. *para*-benzoyl-L-phenylalanine (pBPA) mutations were introduced into ColB-RT surface loops previously highlighted as potential FepA binding sites.[27] Exposure to UV (365 nm) results in pBPA non-specific cross-linking into C-H bonds within 4 Å.[28] Photo-activated cross-linking experiments were performed both *in vitro*, using an OM protein fraction as a FepA source, and *in vivo*, using live *E. coli* cells. Cross-links were identified by SDS-PAGE and further analyzed these by liquid chromatography-tandem mass spectrometry (LC-MS/MS), as previously described by White *et al.*[21]

Three cross-links were identified *in vitro*, two of which (ColB residues D202X and R205X with FepA residues P642 and K639 respectively) validated the EC computed by Rosetta (Figure 6.2.A). The success of the computational prediction is a likely consequence of recent progress in the Rosetta scoring and sampling strategies.[29,30]

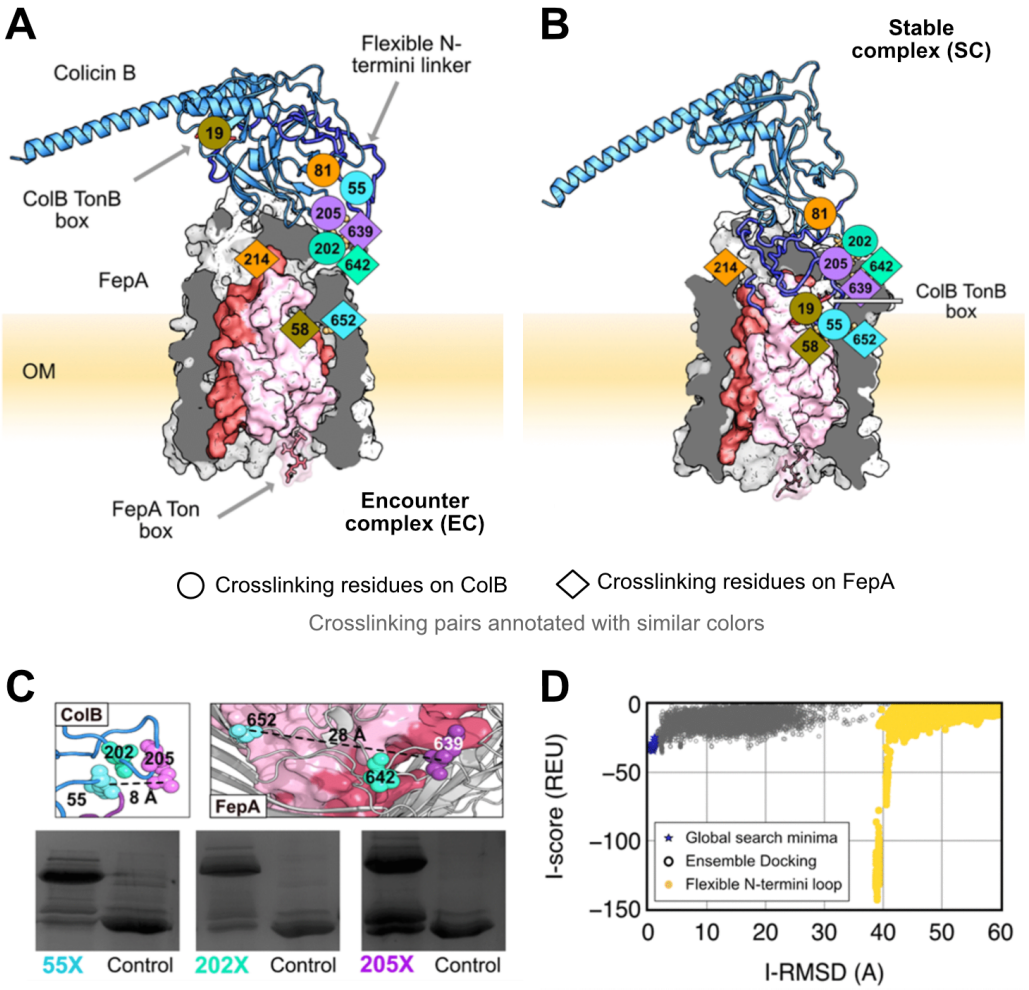


Figure 6.2: Structural insights on the ColB-FepA complex by pBPA cross-linking and Rosetta-based structural modeling. (A) Initial encounter complex (EC) modeled with moderate to little backbone flexibility (under 5 Å root mean square deviations [RMSD]). ColB (*blue*) and FepA (*gray*) form this encounter complex with *in vitro* cross-links, FepA-K639 and ColB-D202 (*teal*), and FepA-P642 and ColB-R205 (*purple*), which lie in proximity in the model. The last *in vitro* cross-link pair, FepA-S652 and ColB-Q55 (*cyan*), and the two in vivo cross-links, FepA-T58 and ColB-M19 (*olive*) and FepA A214 and ColB-G81 (*orange*), are not satisfied in this structure. (B) Fully assembled spontaneously formed stable complex (SC) modeled with the Rosetta FloppyTail algorithm simulating the partially unstructured ColB 1–55 as a floppy tail. (C) Mapped *in vitro* cross-linking sites on the ColB and FepA PDB structures (1RH1 and 1FEP, respectively). Cropped relevant cross-link gels. Self-cross-linking control to the right of each lane. (D) Rosetta interface score (*y axis*) versus interface RMSD (*x axis*) for output structures identified by local docking (ReplicaDock2[25] of ColB to FepA. RMSD is measured relative to the lowest-scoring global docking structure. There is a deep minimum resulting from the arrangement of the flexible N-linker for the FloppyTail models. Measurements corresponding to panel A are in *blue*, measurements corresponding to panel B are in *yellow*.

However, a third cross-link, ColB residue Q55X with FepA residue S652, could not be explained by the computed EC. ColB Q55 is in close proximity (∼8 Å) to ColB D202 and R205 in the ColB PDB structure (PDB: 1RH1), yet its mapped FepA cross-link appears 28 Å apart from the mapped cross-link of ColB 205 (Figure 6.2.B). This disagreement was suggestive of a conformational change accompanying formation of the complex. Hence, to improve the structural model of the ColB-FepA complex, I simulated the N-terminal portion of ColB (residues 1 to 55) as a floppy tail (disordered region with high flexibility), allowing it to sample its environment freely.[29] The resulting model of the stable complex (SC) now explains all three *in vitro*-observed cross-links and is more energetically favorable than the initially calculated EC (Figure 6.2.C,D). The calculated SC also brings the ColB TonB box (residues 17 to 21) closer to the FepA lumen (Figure 6.2.C). In conclusion, using a combination of photo-activated cross-linking and Rosetta-based docking simulations, this work uncovers that ColB associates with its receptor/translocator FepA through an initial encounter complex that then rearranges to the final stable complex, which prepares the toxin for import.

6.3.3 ColB exploits FepA for its active translocation into the cell

The route taken by ColB during FepA-dependent translocation is unknown. Here, I demonstrate how the partially unstructured flexible N-terminal tail of ColB (residues 1 to 55) occupies the channel generated by the TonB-dependent unfolding of the N-terminal half of the FepA plug domain (Figure 6.3). While complex formation is a highly specific step, the translocation mechanism through 22 stranded beta-barrel TBDTs is likely to be applicable to many other systems sharing similar protein folds. The three *in vitro* cross-links obtained prior were also observed *in vivo* along with two additional cross-links (ColB M19X and G81X with FepA T58 and A214), which were

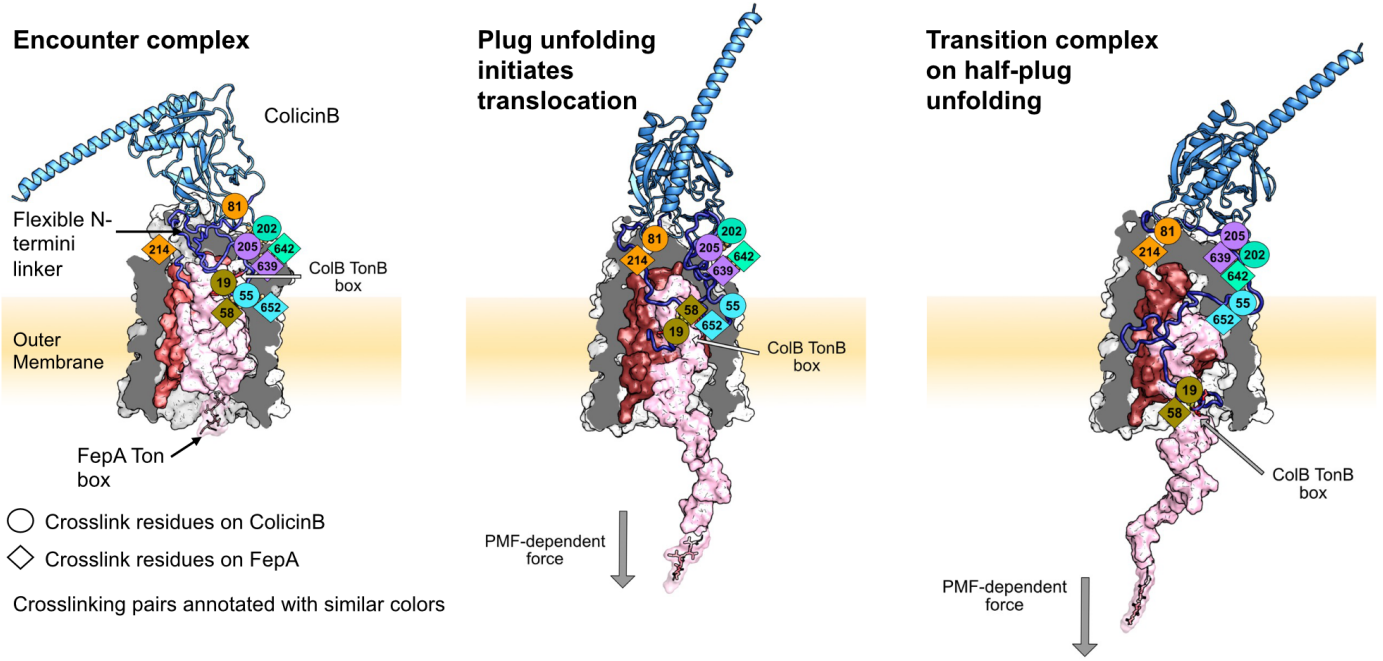


Figure 6.3: Structural insights on the ColB-FepA complex by pBPA cross-linking and Rosetta-based structural modeling. The predicted models in the computational pipeline are illustrated with the crosslinking residues highlighted. Crosslinking residues on ColB are denoted in *circles* and those on FepA are denoted in *diamond*. Starting from the encounter complex (EC), the plug unfolding initiates translocation and the transition complex on the unfolding is predicted. In the predicted model, the ColB TonB box is shown to replace the FepA Ton box.

further mapped by LC-MS/MS. The additional two cross-links did not form in the absence of the energy-transferring protein TonB.

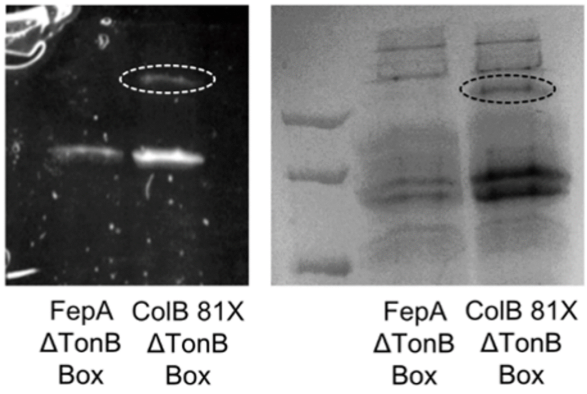


Figure 6.4: Crosslinking data for ColB-FepA interaction. The ability of ColB-81X GFP to cross-link *in vivo* as a function of both ColB and FepA TonB boxes. GFP fluorescence (*right*) and Coomassie blue stain (*left*) are shown. Cross-linked band is circled.

The TonB box of TBDTs and bacteriocins is a conserved pentapeptide sequence essential for interaction with TonB.[31] Two TonB boxes participate in the ColB translocation process: one on colicin itself and the other on its OM receptor, FepA (16, 29).[16,32] The ability of the *in vivo* observed ColB 81-FepA 214 cross-link to form was examined as a function of both the FepA and ColB TonB boxes. The ColB 81-FepA 214 cross-link did not form in the absence of the FepA TonB box, but it still formed in the absence of the ColB TonB box (Figure 6.4). Hence, as both TonB boxes are essential for full colicin translocation, the ColB 81-FepA 214 cross-link appears to capture a stable intermediate translocation step. These experiments were not performed on the second *in vivo*-identified cross-link ColB 19-FepA 58, as ColB 19 is already part of the ColB TonB box.

To investigate the structures during the dynamic translocation process, I equipped conformational modeling in Rosetta to simulate the unfolding of the N-terminal half

(residues 1 to 74) of the globular FepA plug domain, as previously demonstrated for BtuB.[26] I simulated the ColB-FepA translocation process starting with the computed SC structure (Figure 6.2.B) and using the *in vivo*-identified cross-links as guides to generate three intermediate structures in 4 Å increments as discussed earlier (Figure 6.3.B,C). The simulated structures suggest that the translocating N-terminal ColB tail (residues 1 to 55) occupies the cavity generated by the FepA half plug removal with the ColB TonB box now positioned in place of the former FepA TonB box (Figure 6.5).

6.4 Discussion and conclusions

The OM of Gram-negative bacteria excludes several classes of antibiotics.[33] As a means of subverting this impermeability, *trojan horse* antibiotics rely on conjugating antibiotic moieties to siderophores that are actively imported into cells via TBDTs.[34,35] In this work, I elucidate the mechanism of ColB-FepA association and the use of TBDT to translocate across the OM. Further, experimental results demonstrate that the FepA-specific bacteriocin ColB can similarly transport large cargo molecules into *E. coli* under the force of the proton motive force (PMF).

ColB was one of the earliest colicins to be identified[36], yet how this bacteriocin, and its close homologue ColD, recognize FepA has been unclear until now. Using photoactivated cross-linking combined with Rosetta-based simulations, I show that association involves at least two steps in which an initial encounter complex is formed first and then rearranges. The conformational change involves the flexible N-terminal end of the colicin (residues 1 to 55) moving by up to 62 Å to form the final stable complex. An important consequence of these conformational changes is that they

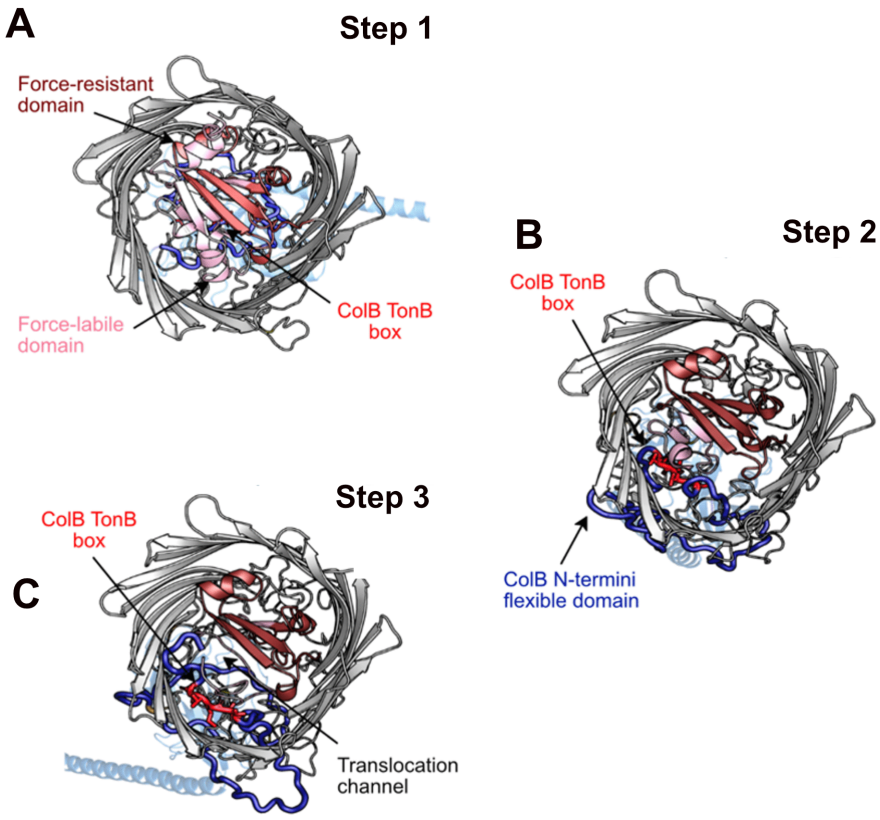


Figure 6.5: Partially unstructured ColB-RT 55-residue N-terminal end occupies the gap generated by the active unfolding of the FepA N-terminal half plug domain. (A) A bottom-to-top view of the hypothesized translocation pathway created with Rosetta by pulling the FepA N terminus into the cell. (A) Step 1, SC complex is formed and the force-labile half-plug domain (*light pink*) begins to unfold; (B) Step 2, the force-labile half-plug is partially unfolded, which allows the ColB N-terminal loop (*blue*) to occupy the void created by the absence of the plug domain; (C) Step 3, the unfolding of the FepA half-plug domain creates a channel for the ColB N-terminal loop to enter.

poise the ColB Ton box close to the channel that subsequently appears during PMF-mediated activation of the TBDT by TonB in the inner membrane. While previous studies have demonstrated that the Ton boxes of both ColB and FepA are important for import[16,32], they do not report on the sequence of events where they are deployed. *In vivo* cross-linking data reveal that the cross-link between ColB-RT G81X and FepA A214 requires the FepA Ton box but not that of ColB, consistent with this cross-link reporting on activation of the TBDT by the PMF. The involvement of the ColB Ton box must be subsequent to this, as has been shown for the import of pyocin S2 through its TonB-dependent transporter FpvAI in *Pseudomonas aeruginosa*.[21]

Past chemical modification data have presented a contradictory picture as to whether ColB translocates across the *E. coli* OM by direct transfer through FepA.[16,37,38] Transport of ColB through FepA would require at least partial unplugging of its central pore. Unplugging of a TBDT to enable uptake of a ligand has been demonstrated by atomic force microscopy for the vitamin B12 transporter BtuB. The N-terminal globular plug domain of BtuB is composed of two mechanically independent half-plug domains. The N-terminal half, which lies proximal to the Ton box, is more amenable to forced unfolding than the C-terminal half.[26] I therefore simulated the unfolding of the N-terminal half-plug of FepA by analogy with that of BtuB[26], within feasible simulation length and timescales. The computed model (Figure 6.2B) emphasizes the importance of the two independent encounters with the energy-transferring protein TonB. The first receptor-mediated encounter allows the translocation of the ColB TonB box to the periplasm (Figure 6.3), while the second activates colicin translocation into the cell. The computed model also suggests that the 55-residue N-terminal end of the translocating colicin mimics the unfolded receptor half-plug and, indeed, replaces

the receptor's TonB box with that from colicin (Figure 6.5).

In summary, the OM translocation of ColB is a highly dynamic process involving two association steps followed by two TonB-dependent events. These simulations also suggest that colicin mimics the part of the FepA half-plug that is removed during import, thereby presenting its own Ton box to the periplasm. The translocation mechanism likely also applies to ColD, which binds FepA through a similar receptor-binding domain and is Ton dependent.[39] The ability of bacteriocin-DNA conjugates to piggy-back the colicin into the cell opens a range of possibilities to utilize bacteriocins for bypassing the Gram-negative bacterial OM. This includes development of novel antibiotic delivery strategies and even genomic manipulations.

6.5 Methods

6.5.1 Structure preparation

The crystal structures of ColB (PDB: 1RH1) and FepA (PDB: 1FEP) were used as starting templates for the computational modeling. Because the crystal structures were missing key loops needed to effectively propagate backbone motions, I built these loops (residues 31 to 44 on ColB and 323 to 335 and 384 to 40 on FepA) using SWISS MODELLER (42). To eliminate energetically unfavorable side chain or backbone clashes, the structures were then relaxed using constraints to the native crystal coordinates using Rosetta Relax.[40]

6.5.2 Modeling the transporter-bacteriocin encounter complex

I determined putative local binding conformations by first performing rigid-body global docking using the ReplicaDock2 protocol (built upon prior work on temperature and Hamiltonian replica exchange Monte Carlo approaches[41,42]) and clustering the lowest-energy docked structures. Starting from each low-energy structure, I refine the structures in a local binding region by using the RosettaDock4.0[43] protocol that adaptively swaps receptor and ligand conformations from a pre-generated ensemble of structures. To diversify the backbone conformations in the ensemble, I used (i) ReplicaDock 2.0[25], (ii) Rosetta Relax[40], and (iii) Rosetta Backrub[44]. Local docking generated ∼6,000 decoys, which are scored based on their interface energies, defined as the energy difference between the total energy of the complex and the total energy of the monomers in isolation.

6.5.3 Predicting stable complex with backbone flexibility

To explore the possibility of the ColB flexible N-terminal domain (residues 1 to 55) interacting explicitly with FepA, I used the Rosetta FloppyTail[24] algorithm, which allows modest sampling of backbone degrees of freedom following a two-stage approach. First, in the low-resolution stage, side chains are represented by a centroid atom and the backbone conformational space is extensively sampled. In the high-resolution stage, all side chain atoms then are returned to refine the structures. I generated ∼5,000 hypothetical decoys starting from the encounter complex obtained in stage 1 (EC). The 5,000 perturbation cycles and 1,000 refinement cycles were used for each decoy. To direct the MC sampling of the FloppyTail algorithm toward possible interacting regions, atom-pair constraints based on the experimental (*in vitro*)

cross-linking residues guided the search. These constraints were calculated based on a harmonic potential with a mean of 6 Å and a standard deviation set to 0.25 Å between the Cα atoms of the candidate residues. Each output decoy was further relaxed to remove unfavorable clashes, and the 100 top-scoring models were then docked using RosettaDock4.0[43] using a fixed backbone. Translational and rotational moves were performed on the top models to generate ~5,000 docking decoys. To confirm the feasibility of these decoys, I evaluated the interface energies and compared the energy landscape of decoys in stage 2 with the prior decoys obtained in stage 1 (Figure 6.2D).

6.5.4 Modeling the translocation pathway by incorporating *in vivo* cross-linking data

Following the partial unfolding of the plug domain in the related TonB-BtuB system[26], I allowed backbone movement in the FepA 75-residue half-plug domain (residues 1 to 75) and the ColB flexible N-terminal domain (residues 1 to 43). Since simulating the dynamic unfolding of FepA half-plug with simultaneous translocation of the ColB via the barrel protein would be computationally demanding, I instead created models to represent three steps along the dynamic pathway of the unfolding translocation process. Briefly, to create each structure along the pathway, I (i) displace the FepA half-plug (residues 1 to 75) using Rosetta FloppyTail to pull the terminus out by 4, 8, and 12 Å, respectively, to begin making each of the three structural steps in the pathway; (ii) translocate the ColB N-terminal domain (residues 1 to 43) using both *in vitro* and *in vivo* cross-linking constraints with Rosetta FloppyTail; and (iii) refine both FepA and ColB conformation and rigid-body displacement using RosettaDock with a flexible FepA half-plug and ColB N-terminal domain. During stages 1 and 2, backbone motions in FloppyTail are propagated toward the closest terminus, but in

stage 3, ColB backbone perturbations during docking are propagated back toward the bulk of ColB to facilitate it finding the optimal rigid-body displacement while the N-terminal domain is translocating.

6.A Appendix

6.A.1 Supplemental Figures

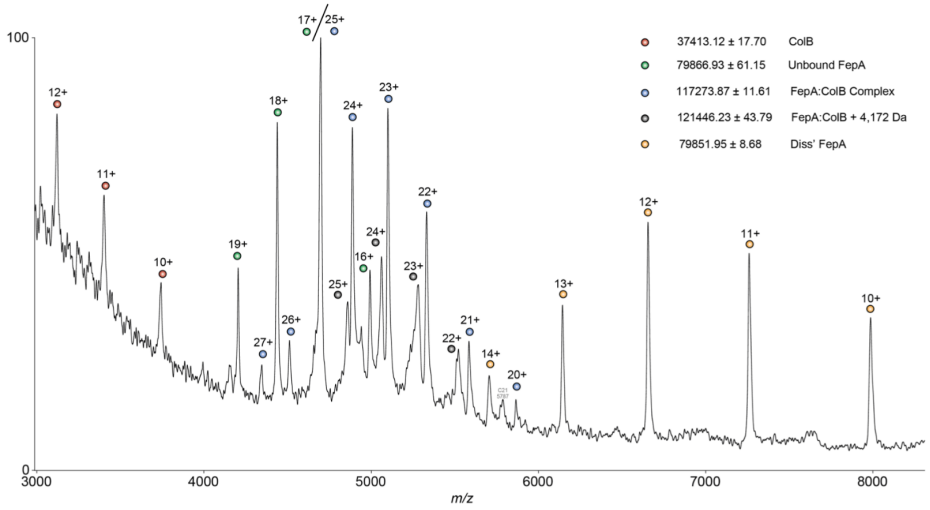


Figure 6.A.1: ColB-FepA stoichiometric complex ratio as observed by native-state mass spectrometry. A clear charge state distribution corresponding to unbound FepA and ColB is observed as well as a 1:1 noncovalent complex composed of one copy of each protein. Charge-reduced species of FepA is also present at higher m/z and indicative of a gas phase-induced dissociation. Also observed is a low-abundance charge state distribution that corresponds to the 1:1 FepA-ColB complex with a discrete mass increase of approximately 4,172 Da. This may correspond to the binding of a single lipopolysaccharide molecule often observed with membrane proteins from the OM, but no further experiments were conducted to further identify the adduct of this low-abundance species. Diss‘ FepA, FepA molecules that have dissociated from a FepA-ColB complex during the run.

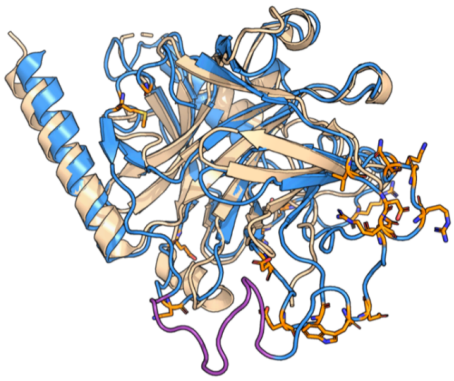


Figure 6.A.2: Structural alignment of ColB and ColE7. Structural alignment of ColB-RT (PDB: 1RH1) (*blue*) and ColE7 T domain (PDB: 2AXC) (*tan*) and positions of pBPA incorporation (*orange* sticks). The average deviation between the corresponding atoms of ColB-RT and ColE7 (RMSD) is 2.38 Å as calculated by PyMol. Both ColB-RT and ColE7 T share a similar pyosin-S fold, yet ColB-RT is the only one forming a complex with FepA. pBPA has been incorporated mainly in ColB exclusive surface loops to examine their role in FepA binding.

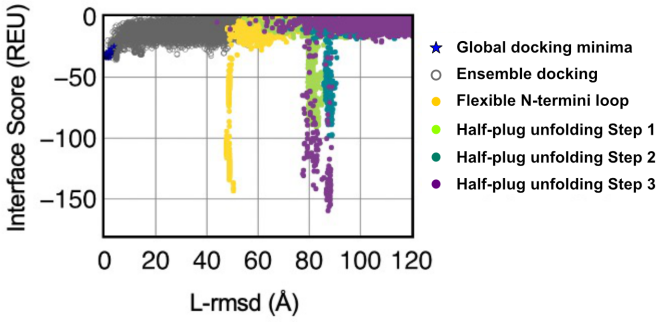
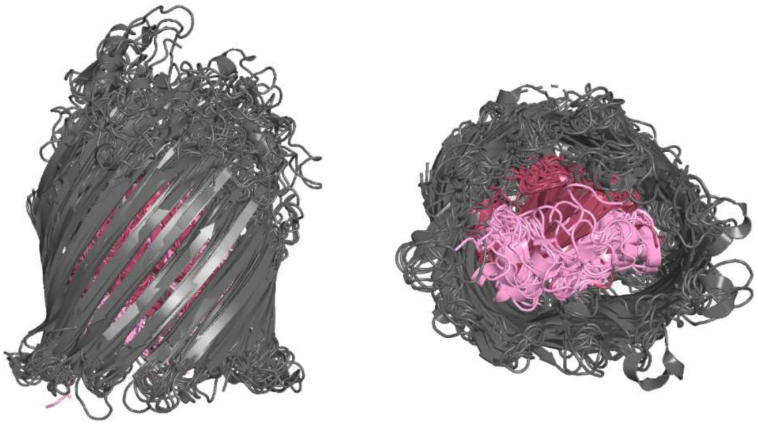


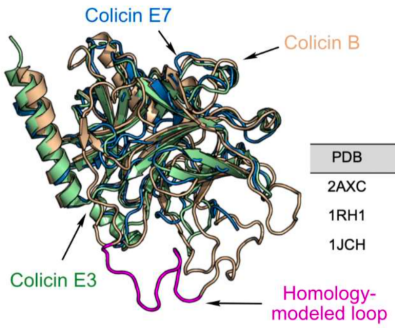
Figure 6.A.3: Computational metrics for predicted structures Interface score v/s l-RMSD (Å) for all the decoys generated for the docking simulations. The global docking minima (blue) are obtained from global docking runs for reference. *Stage 1* ensemble docking models (to create EC) are represented in *gray*, and the *Stage 2* models (for SC) obtained with FloppyTail are represented in *yellow*. *Stage 3* models involving three steps of half-plug unfolding are represented in *green* (step 1), *teal* (step 2), and *purple* (step 3), respectively. As the half-plug is completely unfolded, the interface energies of colicin B in a partial translocation stage with FepA has a deeper energy well than the stage 2 encounter complex.

6.A.2 Supplemental Tables



PDB	RMSD	Protein	Description	Origin
1FEP	0	FepA	Siderophore transporter	Escherichia coli
5MZS	1.23	PfepA	Siderophore transporter	Pseudomonas aeruginosa
5FR8	1.96	PirA	Siderophore transporter	Acinetobacter menangities
1QJQ	2.57	FhuA	Ferric hydroxamate receptor	Escherichia coli
4EPA	3.14	FyuA	Ferric yersiniabactin uptake receptor	Yersinia pestis
1NQH	3.19	BtuB	B12 Transporter	Escherichia coli
1PO0	3.26	FecA	Ferric citrate transporter in complex with iron-free citrate	Escherichia coli
2HDI	3.41	Cir	Colicin I receptor Cir in complex with RBD of Colicin Ia	Escherichia coli
3FHH	3.41	ShuA	Heme/Hemoglobin outer membrane transporter	Shigella dysenteriae
4AIP	3.48	FrpB	Iron transporter	Neisseria meningitdis
4RDT	3.62	ZnuD	Zn-transporter	Neisseria meningitdis
1XKW	3.85	FptA	Pyochelin outer membrane receptor	Pseudomonas aeruginosa
2IAH	3.93	FpvA	Ferripyoverdine receptor bound to substrate	Pseudomonas aeruginosa

Table 6.A.1: Common protein folds for OM receptor FepA Structural alignment of FepA and 12 additional 22-stranded β-barrel OM bacterial proteins identified by the MADOKA server summarized in the presented table. Side view of the alignment (*top-left*), bottom view (*top-right*) of N-terminal half plug domain in pink, C-terminal half plug domain in hot pink.



PDB	RMSD	Protein	Description	Origin
2AXC	0	Colicin E7	N'-terminal translocation domain	Escherichia coli
1RH1	2.38	Colicin B	N'-terminal translocation domain	Escherichia coli
1JCH	9.42	Colicin E3	N'-terminal translocation domain	Escherichia coli

Table 6.A.2: Common protein folds for bacteriocins Structural alignment of ColB (blue) with ColE7 (gray) and ColE3 (green) N-terminal translocation domains identified by Pfam as the pyocin-S domain superfamily. Table summarizing alignment details is on the right.

References

1. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nat Rev Microbiol* **8,** 15–25. https://doi.org/10.1038/nrmicro2259 (2009).
2. Khan, A., Singh, P. & Srivastava, A. Synthesis, nature and utility of universal iron chelator – Siderophore: A review. *Microbiological Research* **212-213,** 103–111. https://doi.org/10.1016/j.micres.2017.10.012 (2018).
3. Aoki, S. K., Diner, E. J., t'Kint de Roodenbeke, C., Burgess, B. R., Poole, S. J., Braaten, B. A., Jones, A. M., Webb, J. S., Hayes, C. S., Cotter, P. A. & Low, D. A. A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria. *Nature* **468,** 439–442. https://doi.org/10.1038/nature09490 (2010).
4. Montville, T. J. & Bruno, M. E. C. Evidence that dissipation of proton motive force is a common mechanism of action for bacteriocins and other antimicrobial proteins. *International Journal of Food Microbiology* **24,** 53–74. https://doi.org/10.1016/0168-1605%2894%2990106-6 (1994).
5. Papadakos, G., Wojdyla, J. A. & Kleanthous, C. Nuclease colicins and their immunity proteins. *Quart. Rev. Biophys.* **45,** 57–103 (2011).
6. Green, E. R. & Mecsas, J. Bacterial Secretion Systems: An Overview. *Microbiol Spectr* **4** (ed Kudva, I. T.) https://doi.org/10.1128/microbiolspec.vmbf-0012-2015 (2016).
7. Simons, A., Alhanout, K. & Duval, R. E. Bacteriocins, Antimicrobial Peptides from Bacterial Origin: Overview of Their Biology and Their Impact against Multidrug-Resistant Bacteria. *Microorganisms* **8,** 639. https://doi.org/10.3390/microorganisms8050639 (2020).
8. Cascales, E., Buchanan, S. K., Duche, D., Kleanthous, C., Lloubes, R., Postle, K., Riley, M., Slatin, S. & Cavard, D. Colicin Biology. *Microbiol Mol Biol Rev* **71,** 158–229. https://doi.org/10.1128/mmbr.00036-06 (2007).
9. Kleanthous, C. Swimming against the tide: progress and challenges in our understanding of colicin translocation. *Nat Rev Microbiol* **8,** 843–848. https://doi.org/10.1038/nrmicro2454 (2010).

10. Masi, M., Réfregiers, M., Pos, K. M. & Pagès, J.-M. Mechanisms of envelope permeability and antibiotic influx and efflux in Gram-negative bacteria. *Nat Microbiol* **2.** https://doi.org/10.1038/nmicrobiol.2017.1 (2017).

11. Braun, V., Pilsl, H. & Gro, P. Colicins: structures, modes of action, transfer through membranes, and evolution. *Arch. Microbiol.* **161,** 199–206. https://doi.org/10.1007/bf00248693 (1994).

12. Egan, A. J. F. Bacterial outer membrane constriction. *Molecular Microbiology* **107,** 676–687. https://doi.org/10.1111/mmi.13908 (2018).

13. Szczepaniak, J., Press, C. & Kleanthous, C. The multifarious roles of Tol-Pal in Gram-negative bacteria. *FEMS Microbiology Reviews* **44,** 490–506. https://doi.org/10.1093/femsre/fuaa018 (2020).

14. Ratliff, A. C., Buchanan, S. K. & Celia, H. Ton motor complexes. *Current Opinion in Structural Biology* **67,** 95–100. https://doi.org/10.1016/j.sbi.2020.09.014 (2021).

15. Arima, K., Katoh, Y. & Beppu, T. Studies on Colicin B Mode of Action and New Extraction Method from Cells. *Agricultural and Biological Chemistry* **32,** 170–177. https://doi.org/10.1271/bbb1961.32.170 (1968).

16. Devanathan, S. & Postle, K. Studies on colicin B translocation: FepA is gated by TonB. *Mol Microbiol* **65,** 441–453. https://doi.org/10.1111/j.1365-2958.2007.05808.x (2007).

17. Pressler, U, Braun, V, Wittmann-Liebold, B & Benz, R. Structural and functional properties of colicin B. *Journal of Biological Chemistry* **261,** 2654–2659. https://doi.org/10.1016/s0021-9258%2817%2935837-4 (1986).

18. Buchanan, S. K., Smith, B. S., Venkatramani, L., Xia, D., Esser, L., Palnitkar, M., Chakraborty, R., Van Der Helm, D. & Deisenhofer, J. Crystal structure of the outer membrane active transporter FepA from Escherichia coli. *Nature structural biology* **6,** 56–63 (1999).

19. Dhar, R. & Slusky, J. S. Outer membrane protein evolution. *Current Opinion in Structural Biology* **68,** 122–128. https://doi.org/10.1016/j.sbi.2021.01.002 (2021).

20. Franklin, M. W., Nepomnyachyi, S., Feehan, R., Ben-Tal, N., Kolodny, R. & Slusky, J. S. Evolutionary pathways of repeat protein topology in bacterial outer membrane proteins. *eLife* **7.** https://doi.org/10.7554/elife.40308 (2018).

21. White, P., Joshi, A., Rassam, P., Housden, N. G., Kaminska, R., Goult, J. D., Redfield, C., McCaughey, L. C., Walker, D., Mohammed, S. & Kleanthous, C. Exploitation of an iron transporter for bacterial protein antibiotic import. *Proc. Natl. Acad. Sci. U.S.A.* **114,** 12051–12056. https://doi.org/10.1073/pnas.1713741114 (2017).

22. Behrens, H. M., Lowe, E. D., Gault, J., Housden, N. G., Kaminska, R., Weber, T. M., Thompson, C. M. A., Mislin, G. L. A., Schalk, I. J., Walker, D., Robinson, C. V. & Kleanthous, C. Pyocin S5 Import into Pseudomonas aeruginosa Reveals a Generic Mode of Bacteriocin Transport. *mBio* **11** (ed Cossart, P. F.) https://doi.org/10.1128/mbio.03230-19 (2020).

23. Hilsenbeck, J. L., Park, H., Chen, G., Youn, B., Postle, K. & Kang, C. Crystal structure of the cytotoxic bacterial protein colicin B at 2.5 Å resolution. *Molecular Microbiology*.

24. Kleiger, G., Saha, A., Lewis, S., Kuhlman, B. & Deshaies, R. J. Rapid E2-E3 Assembly and Disassembly Enable Processive Ubiquitylation of Cullin-RING Ubiquitin Ligase Substrates. *Cell* **139,** 957–968. https://doi.org/10.1016/j.cell.2009.10.030 (2009).

25. Harmalkar, A., Mahajan, S. P. & Gray, J. J. Induced fit with replica exchange improves protein complex structure prediction. *PLOS Computational Biology* **18,** 1–21. https://doi.org/10.1371/journal.pcbi.1010124 (6 2022).

26. Hickman, S. J., Cooper, R. E. M., Bellucci, L., Paci, E. & Brockwell, D. J. Gating of TonB-dependent transporters by substrate-specific forced remodelling. *Nat Commun* **8** (2017).

27. Cheng, Y.-S., Shi, Z., Doudeva, L. G., Yang, W.-Z., Chak, K.-F. & Yuan, H. S. High-resolution Crystal Structure of a Truncated ColE7 Translocation Domain: Implications for Colicin Transport Across Membranes. *Journal of Molecular Biology* **356,** 22–31. https://doi.org/10.1016/j.jmb.2005.11.056 (2006).

28. Dormán, G., Nakamura, H., Pulsipher, A. & Prestwich, G. D. The Life of Pi Star: Exploring the Exciting and Forbidden Worlds of the Benzophenone Photophore. *Chem. Rev.* **116,** 15284–15398. https://doi.org/10.1021/acs.chemrev.6b00342 (2016).

29. Harmalkar, A. & Gray, J. J. Advances to tackle backbone flexibility in protein docking. *Current Opinion in Structural Biology* **67,** 178–186. https://doi.org/10.1016/j.sbi.2020.11.011 (2021).

30. Leman, J. K. *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* **17,** 665–680. https://doi.org/10.1038/s41592-020-0848-2 (2020).

31. Schramm, E, Mende, J, Braun, V & Kamp, R. M. Nucleotide sequence of the colicin B activity gene cba: consensus pentapeptide among TonB-dependent colicins and receptors. *J Bacteriol* **169,** 3350–3357. https://doi.org/10.1128/jb.169.7.3350-3357.1987 (1987).

32. Mende, J. & Braun, V. Import-defective colicin B derivatives mutated in the TonB box. *Molecular Microbiology* **4,** 1523–1533. https://doi.org/10.1111/j.1365-2958.1990.tb02063.x (1990).

33. Ghai, I. & Ghai, S. Understanding antibiotic resistance via outer membrane permeability. *IDR* **Volume 11,** 523–530. https://doi.org/10.2147/idr.s156995 (2018).

34. Górska, A., Sloderbach, A. & Marszałł, M. P. Siderophore–drug complexes: potential medicinal applications of the 'Trojan horse' strategy. *Trends in Pharmacological Sciences* **35,** 442–449. https://doi.org/10.1016/j.tips.2014.06.007 (2014).

35. Kong, H., Cheng, W., Wei, H., Yuan, Y., Yang, Z. & Zhang, X. An overview of recent progress in siderophore-antibiotic conjugates. *European Journal of Medicinal Chemistry* **182,** 111615. https://doi.org/10.1016/j.ejmech.2019.111615 (2019).

36. Fredericq, P. Research on the characteristics and distribution of strains producing colicine B. *Comptes rendus des seances de la Societe de biologie et de ses filiales* **144,** 1287–1289 (1950).

37. Ma, L., Kaserer, W., Annamalai, R., Scott, D. C., Jin, B., Jiang, X., Xiao, Q., Maymani, H., Massis, L. M., Ferreira, L. C., Newton, S. M. & Klebba, P. E. Evidence of Ball-and-chain Transport of Ferric Enterobactin through FepA. *Journal of Biological Chemistry* **282,** 397–406. https://doi.org/10.1074/jbc.m605333200 (2007).

38. Smallwood, C. R., Marco, A. G., Xiao, Q., Trinh, V., Newton, S. M. C. & Klebba, P. E. Fluoresceination of FepA during colicin B killing: effects of temperature, toxin and TonB. *Molecular Microbiology* **72,** 1171–1180. https://doi.org/10.1111/j.1365-2958.2009.06715.x (2009).

39. Rabsch, W., Ma, L., Wiley, G., Najar, F. Z., Kaserer, W., Schuerch, D. W., Klebba, J. E., Roe, B. A., Gomez, J. A. L., Schallmey, M., Newton, S. M. C. & Klebba, P. E. FepA- and TonB-Dependent Bacteriophage H8: Receptor Binding and Genomic Sequence. *J Bacteriol* **189,** 5658–5674. https://doi.org/10.1128/jb.00437-07 (2007).

40. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Science* **23,** 47–55. https://doi.org/10.1002/pro.2389 (2013).

41. Zhang, Z. & Lange, O. F. Replica Exchange Improves Sampling in Low-Resolution Docking Stage of RosettaDock. *PLoS ONE* **8** (ed Zhang, Y.) e72096. https://doi.org/10.1371/journal.pone.0072096 (2013).

42. Zhang, Z., Schindler, C. E. M., Lange, O. F. & Zacharias, M. Application of Enhanced Sampling Monte Carlo Methods for High-Resolution Protein-Protein Docking in Rosetta. *PLoS ONE* **10** (ed Colombo, G.) e0125941. https://doi.org/10.1371/journal.pone.0125941 (2015).

43. Marze, N. A., Burman, S. S. R., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics* **34** (ed Valencia, A.) 3461–3469. https://doi.org/10.1093/bioinformatics/bty355 (2018).

44. Smith, C. A. & Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *Journal of Molecular Biology* **380,** 742–756. https://doi.org/10.1016/j.jmb.2008.05.023 (2008).

Chapter 7

Structure-driven design of orthogonal protein-protein interfaces

Work in this chapter was performed in collaboration with the labs of Dr. Jamie Spangler and Dr. Warren Grayson.

7.1 Overview

Protein-protein interactions (PPIs) are involved in almost all biological processes and understanding the systematic mapping of PPI networks in the cell is instrumental for re-engineering biological functions. Programming protein interfaces to induce novel functionality is a promising new protein engineering strategy with broad applications, from therapeutics to biocatalysis. However, natural proteins are promiscuous and pleiotropy leads to unwanted cross-talk and off-target activity. To address this challenge, I propose a reliable method to generate orthogonal systems (*i.e.* ligand/receptor pairs that interact exclusively with one another and not with any endogenous proteins) from wildtype protein systems through rational

Chapter 8

Rosetta developments and miscellaneous projects

8.1 Overview

Throughout my tenure as a graduate student in the Gray lab, I engaged in several fruitful collaborations, ranging from methods development to design. To unite these disparate research topics while maintaining the flow of my thesis, I briefly describe a few of the interesting developments and protein design projects in this penultimate chapter. The diversity of projects and research topics speaks volumes about the current state of computational protein modeling. Here, I discuss three projects in computational modeling while preserving the underlying theme of protein interactions. First, I discuss our community-wide scientific benchmarking initiative to create robust, automated tests for Rosetta protocols. This initiative demonstrates our dedication, as a software community, to distribute reproducible code with highest scientific rigor. Next, I demonstrate the extension of our docking protocols to membrane environments. Unlike soluble environments, modeling proteins in heterogeneous cell membrane environment is challenging. In this project, I elaborate the development of a flexible docking protocol to predict protein complex structures. Finally, I pivot

to protein design and show a direct application of our orthoPD design strategy on re-engineering histone interfaces. Histones assemble in a nucleosome core to regulate gene expression. By re-engineering histone interfaces to create an asymetric nucleosomes, there is a potential to study important functions such as gene regulation and cell differentiation.

8.2 Automated scientific benchmark for protein docking

Over the past twenty years, the Rosetta modeling suite has grown from a niche software for protein modeling to a software suite encompassing tools to model almost all biomolecules, ranging from nucleic acids and sugars, to peptides and proteins.[1,2] With an ever-increasing codebase with over 3 million lines of code and interlinked functionalities, maintenance and integrity of the code is paramount. Further, for computational models, scientific integrity is intrinsic as protocols should produce similar results irrespective of the computer chips with minor variance. To ensure the integrity and scientific reproducibility of the code-base, I contributed to a community-wide goal of building scientific benchmark tests for core Rosetta protocols.[3] The premise of this was to develop automated tests that were set-up on our test server, with specific goals and requirements. The results of the tests are then displayed on a Dashboard (available at https://benchmark.graylab.jhu.edu/, Figure 8.1), with successful tests displayed in *green* and failure cases highlighted in *red*. Further, each test has adequate documentation highlighting the purpose, benchmark dataset, protocol (.i.e Rosetta

executable used), key results and evaluation metrics for determining success. The contributor to the respective scientific test generally serves an observer (*i.e.* maintainer) and debugs the code that results in the test failures.

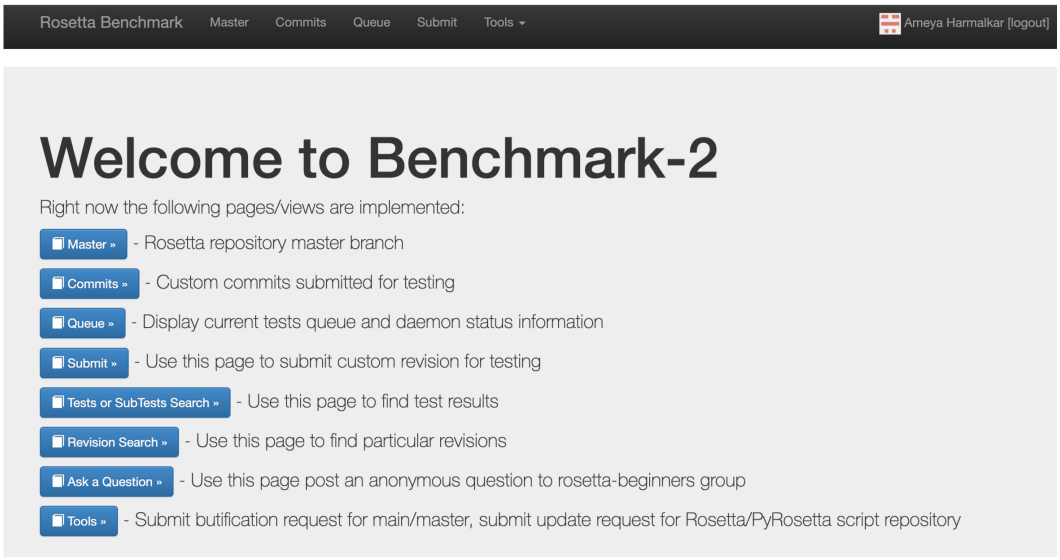


Figure 8.1: Web page for the Testing server dashboard. with options to queue scientific or integration tests for specific versions of the code.

With the framework of scientific testing elaborated above, the community implemented up to 40 scientific benchmark tests. Here, I contributed to the docking scientific tests by creating benchmarks for rigid and flexible protein docking. This results are illustrated in greater detail in Figure 1. This also serves as a small benchmark gauging the development of field over the years, for e.g., as the benchmark results are annually published on our website (https://graylab.jhu.edu/download/rosetta-scientific-tests/), the evolution of the docking protocol (say RosettaDock[4]) and its comparison with newer docking protocols (such as ReplicaDock2) could be performed in an automated fashion.

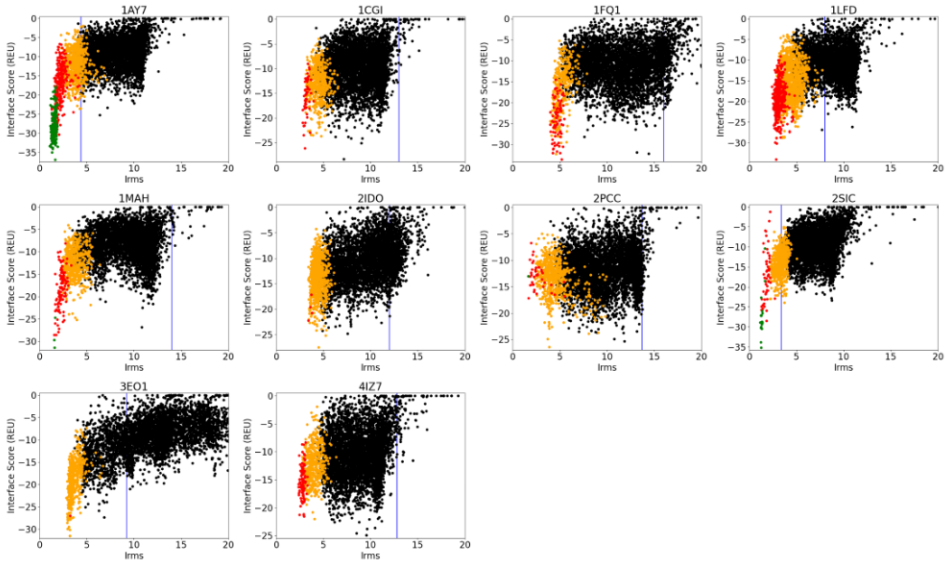
In this work, I made a small contribution to the Rosetta community's goal of ensuring scientific reproducibility of the software. Most of the scientific research labour is short-termed and building automated testing systems guarantees maintenance of software in the long term.

Scientific test: docking_ensemble

FAILURES

None

RESULTS



AUTHOR AND DATE

Adapted for the current benchmarking framework by Ameya Harmalkar (harmalkar.ameya24@gmail.com; Gray Lab), March 2021

PURPOSE OF THE TEST

This benchmark is meant to test how well we discriminate native protein-protein binding orientations from decoys based on the interface score term by performing flexible protein-protein docking experiments with conformer selection across a diverse set of protein-protein complexes.

BENCHMARK DATASET

The dataset consists of 3 protein-protein complexes extracted from the Docking Benchmark 5.0 (Vreven, T. et al. J. Mol. Biol., 2015). The set contains 1 rigid (conformational change < 1.5 Ang), 1 medium-flexible (conformational change between 1.5 and 2.2 Ang), and 1 difficult (conformational change > 2.2 Ang).

Structure preparation:

Figure 8.2: Documentation for the docking scientific test The results page shows the results of the run (10 benchmark targets in this case), the documentation, and the description of whether the test passes or fails (no failures for this case). Results pages are automatically generated at the end of the run for each test as shown here.

8.3 Developing a toolkit for membrane-associate protein docking

A cell is like a fortified city, with lipid membranes as its defensive walls, and membrane proteins as the gatekeepers controlling the transit of molecules and information across these walls.

Membrane proteins (MPs) are of significant importance: they constitute about a third of all proteins and are targets for over 50% of pharmaceuticals.**Alford2017** Despite their importance, MPs represent only 2% of all protein structures in the protein data bank (PDB), and MP complexes are even scarce. Computational approaches have shown promise in capturing membrane protein structures.[5,6] However, unlike soluble proteins (*i.e.* proteins in the cytoplasm or extra-cellular space) that I discussed in most of my thesis, membrane protein modeling involves the challenge of capturing the heterogeneous lipid environment that influences both the structure and function of these entities.[7] In this section, I demonstrate our[i] work on extending protein docking tools to capture the flexible interactions in MP complexes. To incorporate backbone flexibility, here this works adapts existing RosettaDock conformer-selection framework for membrane ecosystems.

Unlike soluble proteins that are modeled with an implicit solvent model in Rosetta, membrane modeling is incorporated with the RosettaMP environment.[8] The central framework of RosettaMP (as illustrated in Figure 2) comprises of three major elements: (1) a membrane residue to define the geometry of the membrane bilayer (*i.e.* membrane bilayer thickness, membrane center and normal), (2) a membrane topology to define the transmembrane region of the protein in the membrane, and (3) FoldTree

[i]This work was initiated by my mentee, Priyamvada Prathima. Dr. Rituparna Samanta and I are handling the completion of this project

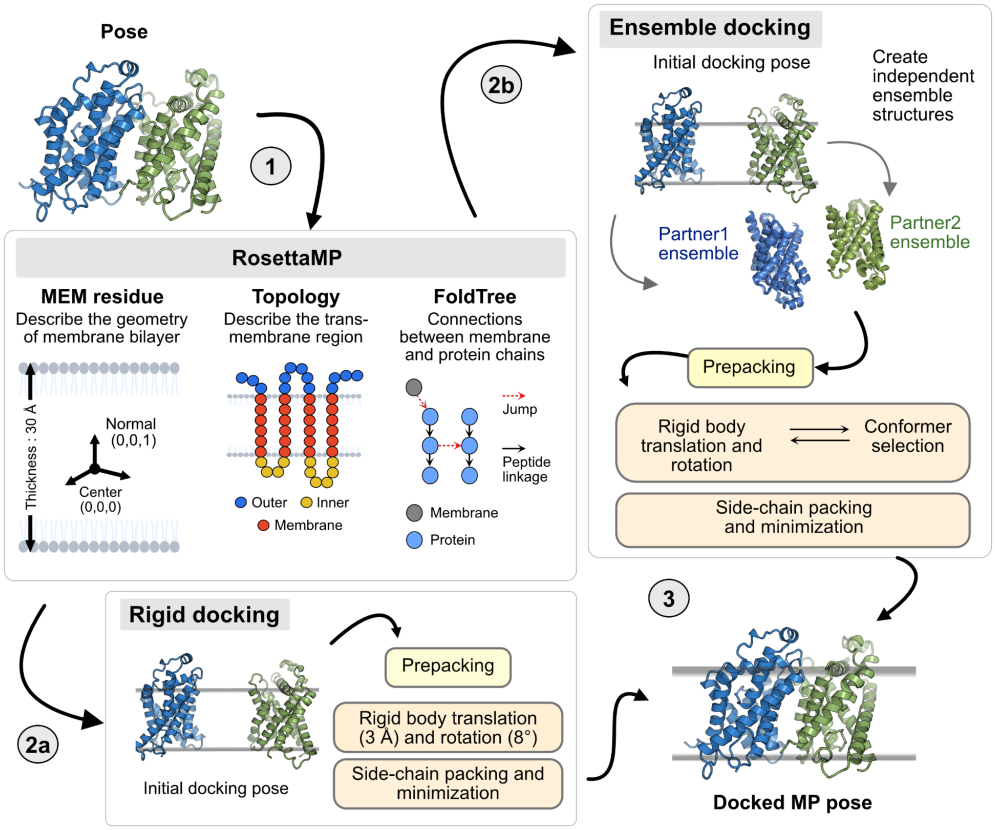


Figure 8.3: Overview of membrane protein docking protocol Starting from a docking pose, RosettaMP generates the membrane environment by adding the MEM atom (describing the geometry of the membrane bilayer by coordinates storing the center, normal and thickness of the bilayer), a spanning topology describing the transmembrane regions of the pose, and a FoldTree to establish connections between the membrane residue and the protein partners. This membrane protein pose is now either passed to rigid docking or ensemble docking. Rigid-docking involves rigid body translations and rotations with side-chain packing and minimization. Ensemble docking creates structural ensembles of individual chains (for backbone diversity) and performs swaps while docking. Generated decoys are packed and minimized to obtain the docked MP pose.

that connects the membrane residue with the protein pose for docking. RosettaMP-Dock utilizes this membrane environment for protein docking and complex structure prediction within a membrane. RosettaMPDock creates a membrane environment for every pose and then initiates a rigid- or ensemble-docking protocol as demonstrated in Figure 8.3.

RosettaMPDock[8] has a rigid-body routine that docks proteins in 6D space within the membrane environment. However, since proteins are intrinsically flexible and conformational dynamics play a crucial role in association, as demonstrated throughout this thesis, we[ii] created a flexible backbone docking protocol, namely the ensemble docking stage. The ensemble-docking stage in RosettaMPDock (Figure 8.3, *right*) draws on the existing functionality of RosettaDock4[4], prior conformer-selection work on heterodimers, and adapts it for membrane proteins. Conformer-selection[9] models for protein interactions obey a statistical mechanical view of protein binding; with unbound states of protein partners existing in an ensemble of low-energy conformations, among which the bound conformations are selected during protein association. RosettaDock 4.0 implements this strategy by pre-generating an ensemble of conformations of the individual protein partners and then employing them as inputs for protein docking. While docking, the ligand (smaller protein partner) and the receptor (larger protein partner) perform rigid body moves coupled with backbone swaps from the pre-generated ensembles. We extended this strategy for RosettaMPDock by implementing a membrane environment for both the pre-generated ensembles and the protein pose that performs backbone swaps and rigid body moves.

Here, I demonstrate the extension of docking approaches to membrane environments for modeling protein interactions. Accurate modeling of membrane complex

[ii]Code development was done in collaboration with Dr. Rituparna Samanta

structures and prediction of the conformational changes in membrane proteins with varying pH, salt concentration, and lipid compositions can pave the way ahead for rational design of MPs. This has tremendous therapeutic potential to treat a vast majority of neurological and cardiovascular diseases.

8.4 Re-engineering the nucleosome core to study the asymmetric histone code

This work was in collaboration with Dr. Evan Worden and Prof. Cynthia Wolberger (JHMI) and was supported by the Johns Hopkins Discovery Award (*awarded in December 2020*).

Regulation of gene expression depends on specific, highly complex combinations of covalent histone modifications that form the basis of the signaling cascade known as the histone code. Over a hundred different post-translational modifications of the four core histone proteins have been identified to date and their functions studied in cell-based and solution studies.[10] The octameric nucleosome core contains two copies each of histones H2A, H2B, H3, and H4, which form a symmetric dimer of tetramers. However, there is mounting evidence that the two histone copies in each nucleosome may not contain the same set of covalent modifications and that this asymmetry serves important functions in gene regulation and cell differentiation. In this work, our goal was to re-engineer the nucleosome so that each protein in the histone octamer can be uniquely distinguished and manipulated. In order to do so, I extended our orthogonal approaches to design interfaces for one pair of histone interactions. This would transform the two identical copies of histones H2A, H2B, H3, and H4 into eight independent polypeptides (Figure 8.4).

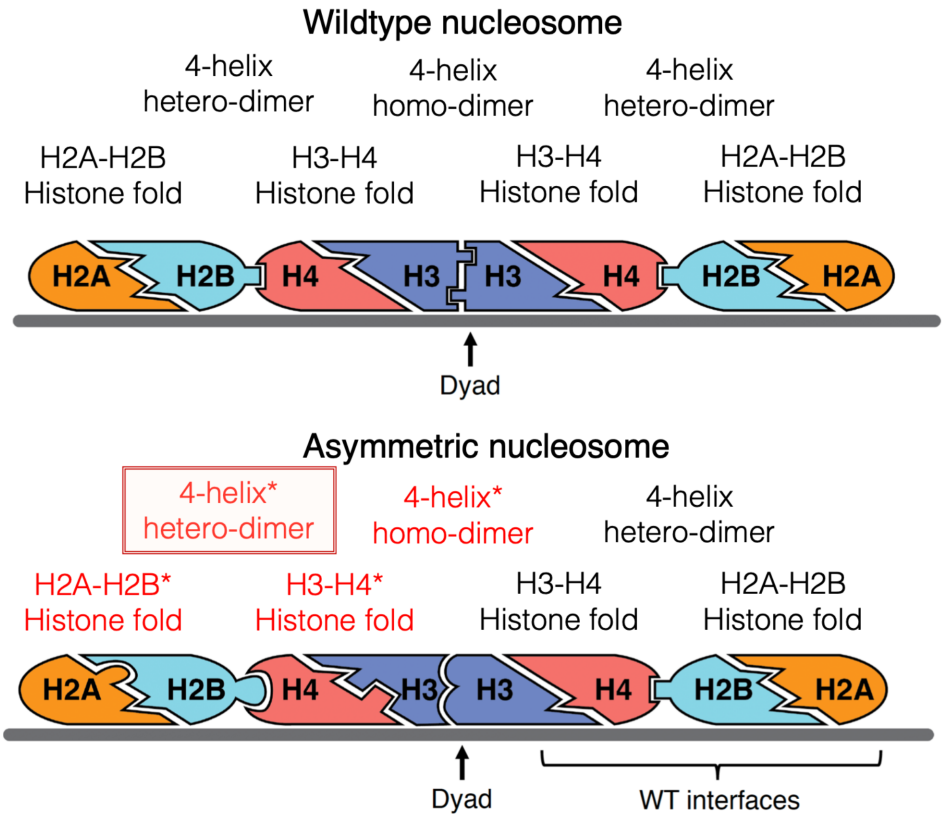


Figure 8.4: Creating asymmetric histones by re-engineering histone interfaces (*top*) Structure of the wildtype nucleosome highlighting the 4 histones (with 2 identical copies each) and the interacting interfaces. (*bottom*) Structure of the asymmetric nucleosome generated by engineering four of the seven histone-histone interfaces.

Here, I equipped Rosetta with our orthogonalization strategy for the histone redesign objective. Prior studies demonstrated the design of histone interface H3-H3 to introduce asymmetry. This served as a proof-of-concept that inducing asymmetry is possible.[11] With OrthoPD strategy (discussed in Chapter 7), co-dependent mutations were introduced on the interfaces. Template structures for each interface were extracted from the nucleosome core crystal structure (PDB: 1KX3) and re-engineered for orthogonality. Re-engineered histones were experimentally validated and will be further reconstituted to full asymmetric nucleosomes. This histone re-engineering work would allow researchers to experimentally probe nucleosome asymmetry in a controlled way for the very first time, making huge contributions to our understanding of gene regulation[12], cell differentiation[13], and cancer[14].

References

1. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology* **487,** 545–574. ISSN: 00766879 (C 2011).
2. Leman, J. K., Weitzner, B. D., Renfrew, P. D., Lewis, S. M., Moretti, R., Watkins, A. M., Mulligan, V. K., Lyskov, S., Adolf-Bryfogle, J., Labonte, J. W., Krys, J., Bystroff, C., Schief, W., Gront, D., Schueler-Furman, O., Baker, D., Bradley, P., Dunbrack, R., Kortemme, T., Leaver-Fay, A., Strauss, C. E., Meiler, J., Kuhlman, B., Gray, J. J. & Bonneau, R. Better together: Elements of successful scientific software development in a distributed collaborative community. *PLoS computational biology* **16,** e1007507. ISSN: 15537358 (5 2020).
3. Leman, J. K., Lyskov, S., Lewis, S. M., Adolf-Bryfogle, J., Alford, R. F., Barlow, K., Ben-Aharon, Z., Farrell, D., Fell, J., Hansen, W. A., Harmalkar, A., Jeliazkov, J., Kuenze, G., Krys, J. D., Ljubetič, A., Loshbaugh, A. L., Maguire, J., Moretti, R., Mulligan, V. K., Nance, M. L., Nguyen, P. T., Conchúir, S., Burman, S. S. R., Samanta, R., Smith, S. T., Teets, F., Tiemann, J. K., Watkins, A., Woods, H., Yachnin, B. J., Bahl, C. D., Bailey-Kellogg, C., Baker, D., Das, R., DiMaio, F., Khare, S. D., Kortemme, T., Labonte, J. W., Lindorff-Larsen, K., Meiler, J., Schief, W., Schueler-Furman, O., Siegel, J. B., Stein, A., Yarov-Yarovoy, V., Kuhlman, B., Leaver-Fay, A., Gront, D., Gray, J. J. & Bonneau, R. Ensuring scientific reproducibility in bio-macromolecular modeling via extensive, automated benchmarks. *Nature Communications* **12.** ISSN: 20411723 (1 2021).
4. Marze, N. A., Roy Burman, S. S., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics* **34,** 3461–3469. ISSN: 14602059 (2018).

5. Roel-Touris, J., Jiménez-García, B. & Bonvin, A. M. Integrative modeling of membrane-associated protein assemblies. *Nature Communications* **11,** 1–11. ISSN: 20411723. http://dx.doi.org/10.1038/s41467-020-20076-5 (2020).

6. Rudden, L. S. & Degiacomi, M. T. Transmembrane Protein Docking with JabberDock. *Journal of Chemical Information and Modeling* **61,** 1493–1499. ISSN: 15205142 (2021).

7. Leman, J. K., Mueller, B. K. & Gray, J. J. Expanding the toolkit for membrane protein modeling in Rosetta. *Bioinformatics* **33,** 754–756. ISSN: 14602059 (5 2017).

8. Alford, R. F., Koehler Leman, J., Weitzner, B. D., Duran, A. M., Tilley, D. C., Elazar, A. & Gray, J. J. An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Computational Biology* **11** (ed Livesay, D. R.) e1004398. ISSN: 1553-7358 (2015).

9. Chaudhury, S. & Gray, J. J. Conformer Selection and Induced Fit in Flexible Backbone Protein–Protein Docking Using Computational and NMR Ensembles. *Journal of Molecular Biology* **381,** 1068–1087. ISSN: 0022-2836. http://www.sciencedirect.com/science/article/pii/S0022283608006086 (2008).

10. Zhao, Y. & Garcia, B. A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harbor Perspectives in Biology* **7.** http://cshperspectives.cshlp.org/content/7/9/a025064.abstract (9 2015).

11. Ichikawa, Y., Connelly, C. F., Appleboim, A., Miller, T. C., Jacobi, H., Abshiru, N. A., Chou, H. J., Chen, Y., Sharma, U., Zheng, Y., Thomas, P. M., Chen, H. V., Bajaj, V., Müller, C. W., Kelleher, N. L., Friedman, N., Bolon, D. N., Rando, O. J. & Kaufman, P. D. A synthetic biology approach to probing nucleosome symmetry. *eLife* **6,** 1–22. ISSN: 2050084X (2017).

12. Shilatifard, A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annual review of biochemistry* **81,** 65–95. ISSN: 1545-4509. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4010150&tool=pmcentrez&rendertype=abstract (2012).

13. Voigt, P., LeRoy, G., Drury, W. J., Zee, B. M., Son, J., Beck, D. B., Young, N. L., Garcia, B. A. & Reinberg, D. Asymmetrically modified nucleosomes. *Cell* **151,** 181–193. ISSN: 00928674. http://dx.doi.org/10.1016/j.cell.2012.09.002 (1 2012).

14. Meeks, J. J. & Shilatifard, A. Multiple Roles for the MLL/COMPASS Family in the Epigenetic Regulation of Gene Expression and in Cancer. *Annual Review of Cancer Biology* **1,** 425–446. ISSN: 2472-3428. http://www.annualreviews.org/doi/10.1146/annurev-cancerbio-050216-034333 (1 2017).

Chapter 9

Conclusions

Proteins are biological polymers that encode the machinery of life. The three-dimensional structure of proteins and protein complexes provides a static snapshot of their physical interactions and can aid the understanding of molecular mechanisms to understand biological processes and suggest disease intervention strategies. With the recent advent of machine learning approaches such as AlphaFold[1] and RoseTTAFold[2], the protein sequence-to-structure prediction challenge has demonstrated unprecedented performance.[3] Yet, protein docking, *i.e.* prediction of protein complex structures and higher-order assemblies, persists as a fundamental challenge owing to large-scale binding-induced conformational changes.[4] Computational methods that model protein interactions and elucidate complex structures can reveal molecular mechanisms and allow us to engineer interactions based on molecular structures, from immunology and cancer to infectious diseases and tissue engineering. The impact of faster, inexpensive, and accurate predictions of PPIs and protein complexes will accelerate the field of structural bio-informatics and protein design. In this dissertation, I have advanced computational modeling approaches for protein-protein docking and created new tools for protein design. Specifically, I have (1) developed novel

enhanced sampling approaches for predicting conformational changes in protein-protein docking mimicking induced-fit approaches of protein binding; (2) presented a sequence-to-structure pipeline for accurate prediction of protein complexes by coupling AlphaFold with our sampling routines; (3) applied modeling to decipher protein dynamics in a biological system which would be otherwise infeasible to model or validate with experimental techniques; and (4) created a computational pipeline for designing orthogonal protein interfaces for a protein signaling system for downstream tissue engineering applications.

9.1 My contributions

My tenure in the Gray lab began with a goal to tackle the sampling challenges in flexible backbone protein docking.[5] Throughout the years, my work has spanned across protein docking, to modeling dynamics and eventually to protein design. In 2019, I participated in Critical Assessment of PRediction of Interactions (CAPRI), a blind community-wide docking challenge, that highlighted the limitations of the state-of-the-art docking tools. To address these limitations, I developed an enhanced sampling approach for protein docking based on the hypothesis that protein association follows a induced-fit mechanism of binding (Chapter 2). This docking approach, namely *ReplicaDock 2.0*[6], employed temperature-replica exchange Monte Carlo (T-REMC) for on-the-fly sampling of backbones in conjunction with rigid body perturbations. On a benchmark of 88 protein complexes with varying degrees of flexibility, ReplicaDock 2.0 is the first method to successfully dock 62% of complexes with conformational changes ($RMSD_{UB}$) up to 2.2 Å.

The development of ReplicaDock2.0 highlighted the efficiency of enhanced sampling approaches to capture native-like backbone moves, however, the low-resolution scoring often skewed the sampling to non-native funnels. To address this shortcoming, I built a novel resolution exchange protocol to swap configurations (all-atom and centroid) across replicas, thereby utilizing the efficiency of the two scoring schemes for better sampling of the conformational landscape. Chapter 3 discusses the development and benchmarking of a resolution exchange protocol and extends its application for the protein docking task. With resolution exchange, the conformational sampling of an high-resolution model is supplemented with low-resolution models while avoiding potential entrapment in non-native sticky sites. I benchmarked this new approach on a small set of 9 flexible protein targets ($RMSD_{UB}$ > 1.2 Å) and demonstrated better performance than prior approaches (ReplicaDock2.0 and RosettaDock 4.0[7]) for 8 targets, with acceptable decoys for all 9 targets.

Chapter 4 describes the performance of the Gray lab prediction team in CAPRI. Over the course of my PhD, I participated in CAPRI rounds 47-54, comprising 45 targets.[4] These rounds highlighted our major limitation in docking antibody-antigen complexes and heteromeric higher-order assemblies. Further, in round 50, DeepMind's model AlphaFold[1], a deep-learning model trained on evolutionary information and protein structural data, demonstrated unprecedented performance. This approach was extended for protein complexes[8,9], peptides[10] and PPIs, however, prior limitations prevailed with poor performance in antibody-antigen complexes and targets with large binding-induced conformational changes. Since AlphaFold metrics for protein structure prediction, such as the predicted local distance difference test (pLDDT)[11] and the predicted alignment error (PAE), correlated well with flexibility,

this made me ask whether AlphaFold could be coupled with our sampling approaches for a complete sequence-to-structure-to-complex pipelines. Using AlphaFold as a structure-generator, I evaluated metrics and developed a pipeline to model complex structures. Out of 245 benchmark targets, AlphaFold identified incorrect binding sites (DockQ<0.2[12]) for 105 targets. With the AlphaFold-RepDock approach, docking all of the 105 targets in the appropriate binding sites was feasible. Moreover, for over 63 targets, the protocol obtained medium CAPRI-quality predictions (DockQ>0.5[12]). My results demonstrate that deep-learning approaches in conjunction with physics-based sampling tools can leverage both evolutionary and biophysical information for improved structure prediction.

My work in computational methods development for docking highlighted the potential avenues to push the boundaries of current modeling routines. In Chapter 6, I applied this principle to examine the interaction between a bacteriocin (ColB) with an outer-membrane receptor (FepA), and devised a potential translocation pathway of ColB through FepA.[13] Here, I learned the utilization of MC-based approaches to reveal stages of a biological pathway. My computational models supported experimental evidence, demonstrating the feasibility of extending MC-approaches to predict transient interactions.

In Chapter 7, I describe the development of OrthoPD, a computational tool to design orthogonal protein interfaces, and demonstrate its utility over the platelet-derived growth factor (PDGF) signaling system.[14,15] Orthogonality induces exclusive selective and prevents pleiotropic effects in endogenous, wildtype proteins.[16] With this computational approach that utilizes structural ensembles and iterative designs mutations, to first disrupt wildtype interaction and then enrich for orthogonality;

one can systematically and conservatively re-engineer existing protein interfaces for customized applications in therapeutics and regenerative medicine. Unlike *de novo* design, re-engineering protein interfaces has applications to modulate existing interaction pathways for specific functionality. OrthoPD results for the PDGF system demonstrate successfully ablation with just two mutations on the receptor with ligand mutant experiments still underway.

In sum, I hope that the methodological advancements and design algorithms presented in this work will contribute a small part to the larger efforts of biomolecular modeling of protein-protein interactions. With machine learning approaches gaining a lot of momentum in the field, I strongly believe that an integration of ML with physics has a huge potential to understand the nuances in structural biology and bring a paradigm shift in molecular discovery and design.

9.2 Future Directions

Reflecting upon the start of my journey as a PhD student in 2018, the protein modeling space has changed dramatically. Accurate modeling at an atomistic level, that seemed like a distant dream, was rationalized with the development of deep-learning tools such as AlphaFold[1] and RoseTTAFold[2]. Despite their drawbacks and limitations, it would be disingenuous to acknowledge the huge role that these structure prediction tools have played in pushing the field ahead. That being said, my efforts have focused on one of its limitations, *i.e.* modeling flexible protein-protein interactions. This influx of deep-learning methods, having tackled the structure prediction challenges, has further enagged in protein design. From structure-agnostic, sequence-only models (ESM like generative models, cite Progen) to learned potentials and diffusion models

(Namrata's paper, Chroma, etc), *de novo* protein design is no longer a niche field but rather reigns as the 'coolest kid' in the disciplines of structural biology and protein engineering. Looking ahead, I am excited at the potential of computational modeling in docking and design. Here, I list future directions stemming from my experience as a computational biologist.

9.2.1 Encoding physics in protein language models for interpretability

With the release of ChatGPT in late 2022[17], scientists and laymen alike were captivated with the abilities of generative large-scale language models (LLMs); whether it was acing the MCATs or writing fragments of programming code. Treating protein sequences as a simple alphabet of 20 amino acids, extension of language models have demonstrated the capability to learn representations across protein families, evolutionary traits, and long-range dependencies in amino acids. Despite the phenomenal growth of LLMs in protein space (for e.g., ESM[18] or auto-regressive models such as UniRep[19], ProGen[20], ProtGPT[21]), the absence of physics and lack of interpretability in these models hamper their generalizable application to protein engineering tasks (for e.g., thermostability prediction, binding, viscosity). Developing models encoded with protein biophysical information would provide utility that would extend beyond the primary tasks of predictive, generative, or representation learning, providing insights that could assist protein engineering efforts.

The scientific rationale for this direction stems from the basis of my thesis: the sequence-structure-function relationship. Language models treat amino acids as the ultimate authority on function. However, as I have demonstrated throughout this thesis, structure is paramount, rather is the primary determinant of protein function. Developing models that integrate the protein language with physical, chemical and

biological properties of residues, could learn about the underlying traits of protein interactions such as higher propensity of hydrophobic patches on binding sites or effects of point mutations on the global structure. From our recent work for predicting antibody thermostability[22], I demonstrated the utility of energetics in predictive models by creating residue-wise contact maps for energies (analogous to distograms for protein residue-wise distances). LLMs supplemented with an energy-track seem to have great potential in decipher the nuances of the protein landscape.

9.2.2 Accelerating enhanced sampling with machine learning approaches

Extending the machine learning tools from the protein structure prediction task to predict protein dynamics is an upcoming area of research. Conventionally, state-of-the-art physics-based approaches model protein interactions and dynamics with an energy function, thereby aiming to map the energy landscape. Here, ML approaches have two avenues to contribute and capture protein dynamics: (1) simplified protein force-field/energy-function to map the conformational landscape, and (2) identify collective variables for enhanced sampling.

All-atom energy landscapes are rugged and development of accurate low-resolution energy functions to mimic these landscapes has been challenging. As deep learning models are universal function approximators, they have potential to deduce energies from atomic coordinates of protein structures rather than estimating a function specified *a priori*.[23] By training on molecular dynamics (MD) or Monte Carlo (MC) trajectories of protein simulations (say folding or docking), a neural network can, in principle, learn both long-range and short-range energies and their relationship with protein coordinates. A shortcoming of this approach would be its inability to extrapolate in poorly sampled conformational space *i.e.* rare events not sampled by

MD or MC methods would be underrepresented. Nevertheless, the development of a model to estimate energies based on protein configurational coordinates would boost computing speeds, and can in turn promote exploratory sampling in low-sampled spaces.

Manipulating collective variables (CVs) can direct sampling and jump across energy barriers, however determining the 'ideal' collective variables is challenging.[24] For sampling, I envision the use of ML-approaches to identify potential CVs for a biomolecular system. By learning over small simulations, ML models could identify putative starting states for following simulations based on learned CVs. Alternatively, sampling could be improved by tuning bias potentials, parameters of the Hamiltonians, temperatures, for approaches such a replica exchange or metadynamics. Invernizzi *et al.* recently demonstrated the use of normalizing flows with replica exchange to sample molecular systems.[25] Extending this approach for some of the enhanced sampling methods described in this thesis, temperatures (T-REMC) or tuning parameters (resolution exchange) could be altered while docking to better explore the conformational landscape

9.2.3 In-silico design of fit-for-purpose antibodies

CoVID-19, Cancer, Celiac disease: the therapeutic potential of these ever-evolving, increasingly complex diseases is encoded in the human adaptive immune system. The human adaptive immune system is capable of mounting a robust response to nearly any foreign pathogen (*i.e.*, viruses, bacteria, fungi, and parasites) by producing proteins (antibodies) that recognize specific regions of the pathogens. The groundbreaking goal for immunology is to generate custom antibody sequences with high

affinity and specificity.[26] Antibody-antigen complexes are one of the most challenging protein docking targets owing to the high flexibility of the interface. Recently, generative and diffusion-based models[27,28] have been applied for generating protein sequences for designs and for building missing motifs on potential binders. To extrapolate these models for antibody-specific tasks, one could transfer learned amino-acid vector representations (such as those from pre-trained language models trained on antibody sequences[29]) and condition models on available antibody-antigen structures. This would allow extension of the generative methods for antibodies and develop a method that could retro-synthesize antibody paratopes for antigen epitopes, *i.e.*, develop antibodies based on custom therapeutic needs for potential antigens.

9.3 Parting thoughts

Life began on earth $\sim$4 billion years ago, written in a chemical language of biopolymers. Today, this chemical language is the basis of our being, with everything from microbes to human-beings evolving as a result of the interactions of the polymers, we now know as proteins. The cusp of 21st century demonstrated major advances in not just understanding this chemical language, but also manipulating it, with humans altering this language as editors. Much of these advances could be contributed to our ability to understand the protein language better, and simulating interactions for navigating the protein sequence-structure space. The versatility of computational approaches, the breakthrough of AI, and the advances in high-throughput experimentation excites me about the possibilities ahead. I hope that the algorithms presented in this work will contribute a small fraction to our overarching goals of discovery and design of protein-protein interactions.

References

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589. ISSN: 14764687. http://dx.doi.org/10.1038/s41586-021-03819-2 (7873 2021).

2. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876. ISSN: 0036-8075 (6557 2021).

3. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics* **89,** 1607–1617. https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.26237 (12 2021).

4. Lensink, M. F. *et al.* Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics* **89,** 1800–1823. ISSN: 0887-3585. https://doi.org/10.1002/prot.26222 (12 2021).

5. Harmalkar, A. & Gray, J. J. Advances to tackle backbone flexibility in protein docking. *Current Opinion in Structural Biology* **67,** 178–186. ISSN: 1879033X. http://arxiv.org/abs/2010.07455 (2020).

6. Harmalkar, A., Mahajan, S. P. & Gray, J. J. Induced fit with replica exchange improves protein complex structure prediction. *PLOS Computational Biology* **18,** 1–21. https://doi.org/10.1371/journal.pcbi.1010124 (6 2022).

7. Marze, N. A., Burman, S. S. R., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics* **34,** 3461–3469. ISSN: 14602059 (20 2018).

8. Evans, R., Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Ž´, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J. & Hassabis, D. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021).

9. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. ColabFold - Making protein folding accessible to all. *bioRxiv*. https://www.biorxiv.org/content/early/2021/10/29/2021.08.15.456425 (2021).

10. Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A. & Schueler-Furman, O. Harnessing protein folding neural networks for peptide–protein docking. *Nature Communications* **13,** 176. ISSN: 2041-1723. https://doi.org/10.1038/s41467-021-27838-9 (1 2022).

11. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29,** 2722–2728. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btt473 (21 2013).

12. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE* **11,** 1–9. https://doi.org/10.1371/journal.pone.0161879 (8 2016).

13. Cohen-Khait, R., Harmalkar, A., Pham, P., Webby, M. N., Housden, N. G., Elliston, E., Hopper, J. T., Mohammed, S., Robinson, C. V., Gray, J. J. & Kleanthous, C. Colicin-Mediated Transport of DNA through the Iron Transporter FepA. *mBio* **12.** ISSN: 21507511 (5 2021).

14. Nevins, M., Giannobile, W. V., McGuire, M. K., Kao, R. T., Mellonig, J. T., Hinrichs, J. E., McAllister, B. S., Murphy, K. S., McClain, P. K., Nevins, M. L., Paquette, D. W., Han, T. J., Reddy, M. S., Lavin, P. T., Genco, R. J. & Lynch, S. E. Platelet-derived growth factor stimulates bone fill and rate of attachment level gain: results of a large multicenter randomized controlled trial. *Journal of periodontology* **76,** 2205–2215. ISSN: 0022-3492 (Print) (12 2005).

15. Shim, A. H. R., Liu, H., Focia, P. J., Chen, X., Lin, P. C. & He, X. Structures of a platelet-derived growth factor/propeptide complex and a platelet-derived growth factor/receptor complex. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 11307–11312. ISSN: 00278424 (25 2010).

16. Sockolosky, J. T., Trotta, E., Parisi, G., Picton, L., Su, L. L., Le, A. C., Chhabra, A., Silveria, S. L., George, B. M., King, I. C., Tiffany, M. R., Jude, K., Sibener, L. V., Baker, D., Shizuru, J. A., Ribas, A., Bluestone, J. A. & Garcia, K. C. Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes. *Science (New York, N.Y.)* **359,** 1037–1042. ISSN: 1095-9203 (Electronic) (6379 2018).

17. Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T. & Ge, B. *Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models* 2023. arXiv: 2304.01852 [cs.CL].

18. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J. & Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America* **118.** ISSN: 10916490 (15 2021).

19. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **16,** 1315–1322. ISSN: 15487105. http://dx.doi.org/10.1038/s41592-019-0598-1 (12 2019).

20. Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: Exploring the Boundaries of Protein Language Models. http://arxiv.org/abs/2206.13517 (2022).

21. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications* **13,** 4348. ISSN: 2041-1723. https://doi.org/10.1038/s41467-022-32007-7 (1 2022).

22. Harmalkar, A., Rao, R., Xie, Y. R., Honer, J., Deisting, W., Anlahr, J., Hoenig, A., Czwikla, J., Sienz-Widmann, E., Rau, D., Rice, A. J., Riley, T. P., Li, D., Catterall, H. B., Tinberg, C. E., Gray, J. J. & Wei, K. Y. Toward generalizable prediction of antibody thermostability using machine learning on sequence and structure features. *mAbs* **15,** 2163584. https://doi.org/10.1080/19420862.2022.2163584 (1 2023).

23. Loose, T. D., Sahrmann, P. G. & Voth, G. A. *Centroid Molecular Dynamics Can Be Greatly Accelerated Through Neural Network Learned Centroid Forces Derived from Path Integral Molecular Dynamics* 2022. arXiv: 2208.07973 [physics.chem-ph].

24. Sidky, H., Chen, W. & Ferguson, A. L. Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Molecular Physics* **118,** e1737742. ISSN: 0026-8976. https://doi.org/10.1080/00268976.2020.1737742 (5 2020).

25. Invernizzi, M., Krämer, A., Clementi, C. & Noé, F. Skipping the Replica Exchange Ladder with Normalizing Flows. *The Journal of Physical Chemistry Letters* **13,** 11643–11649. https://doi.org/10.1021/acs.jpclett.2c03327 (50 2022).

26. Mahajan, S. P., Ruffolo, J. A., Frick, R. & Gray, J. J. Hallucinating structure-conditioned antibody libraries for target-specific binders. *Front Immunol*. https://www.biorxiv.org/content/early/2022/06/06/2022.06.06.494991 (2022).

27. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., Bortoli, V. D., Mathieu, E., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M. & Baker, D. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, 2022.12.09.519842. http://biorxiv.org/content/early/2022/12/14/2022.12.09.519842.abstract (2022).

28. Ingraham, J., Baranov, M., Costello, Z., Frappier, V., Ismail, A., Tie, S., Wang, W., Xue, V., Obermeyer, F., Beam, A. & Grigoryan, G. Illuminating protein space with a programmable generative model (2022).

29. Ruffolo, J. A., Chu, L.-S., Mahajan, S. P. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *bioRxiv*. https://www.biorxiv.org/content/early/2022/04/21/2022.04.20.488972.1 (2022).

Ameya Harmalkar

PhD Candidate – Department of Chemical and Biomolecular Engineering
Johns Hopkins University
✉ aharmal1@jhu.edu • ameyaharmalkar.github.io/

Education

Johns Hopkins University **Baltimore, MD**
Doctor of Philosophy *anticipated June 2023*
Major Chemical and Biomolecular Engineering

Institute of Chemical Technology **Mumbai, IN**
Bachelor of Engineering, 9.11/10 *2014-2018*
Major Chemical Engineering

Research Experience

Graduate Research Assistant, Johns Hopkins University. **Baltimore, MD**
Advisor: Prof. Dr. Jeffrey J. Gray *2018-present*

- Created ReplicaDock2.0 - a Temperature-replica exchange Monte Carlo(T-REMC) approach to improve protein-protein docking by mimicking induced-fit pathways (*62% success rate* on flexible docking targets) .
- Predicted translocation states of bacteriocin via outermembrane receptor for DNA delivery in bacteria.
- Computationally designed *orthogonal* protein interfaces to promote osteogenesis in adipose stem cells.
- Computationally designed histone interfaces to make asymmetric nucleosomes.

DAAD Research Fellow, Technical University of Munich. **Munich, DE**
Advisor: Prof. Dr. Martin Zacharias *Spring 2022*

- Developed enhanced sampling approaches to capture large-scale protein conformational changes.
- Developed a novel "Resolution exchange" methodology for biomolecular simulations.

Graduate Research Intern, Protein Discovery, Amgen Research **Thousand Oaks, CA**
Advisor: Dr. Kathy Wei *2021*

- Created a deep learning based toolkit for identifying antibody thermostability to improve their developability in the drug discovery pipeline.

TEQIP Undergraduate Research Fellow, ICT **Mumbai, IN**
Advisor: Prof. Dr. Vilas Gaikar *2015-2017*

- Performed molecular dynamics and QM-MM calculations to analyze the extraction of Dibenzothiophene and its substitutes from n-dodecane using Ionic liquids as greener extractive solvents.

Publications

1: **Harmalkar A**, Rao R, Gray JJ, Wei K. Towards generalizable prediction of antibody thermo-stability using machine learning on sequence and structure features. *bioRxiv* (accepted in mAbs), 2022. doi: 10.1101/2022.06.03.494724

2: **Harmalkar A**, Mahajan S, Gray JJ. Induced fit with replica exchange improves protein complex structure prediction. *accepted in PLOS Comp. Bio*, 2022. doi: 10.1371/journal.pcbi.1010124

3: Cohen-Khait R*, **Harmalkar A***, Pham P, Webby MN, Housden NG, Elliston E, Hopper JTS, Mohammed S, Robinson CV, Gray JJ, Kleanthous C: Colicin-mediated transport of DNA through the iron transporter FepA. *mBio*, Volume 12, Issue 5, e01787-21, 2021.*(*equal author contribution)*. doi: 10.1128/mBio.01787-21

4: Lensink MF, Brysbaert G,..., **Harmalkar A**,..., Wodak SJ. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins: Struct. Func. Bioinform.*, 89, 1800-1823, 2021. doi: 10.1002/prot.26222

5: Koehler Leman J,..., **Harmalkar A**,..., Gray JJ, Bonneau R. Ensuring scientific reproducibility in bio-macromolecular modeling via extensive, automated benchmarks. *Nature Commun.*, 12, 6947, 2021. doi: 10.1038/s41467-021-27222-7

6: **Harmalkar A**, Gray JJ. Advances to tackle backbone flexibility in protein docking. *Curr.Opin. Struct. Biol.*, 67:178–186, 2021. doi: 10.1016/j.sbi.2020.11.011

7: Singh MB, **Harmalkar AU**, Prabhu SS, Pai NR, Bhangde SK, Gaikar VG. Molecular dynamics simulation for desulphurization of hydrocarbon fuel using ionic liquids. *Journal of Mol. Liq.*, 264, 2018. doi: 10.1016/j.molliq.2018.05.088

In Preparation

1: **Harmalkar A**, Gray JJ. From sequence-to-structure-to-complexes: an in-silico pipeline for protein complex structure prediction. 2023.

2: **Harmalkar A**, Gray JJ, Zacharias M. Resolution exchange for protein-protein docking. 2023.

3: **Harmalkar A***, Samanta R*, Prathima P, Gray JJ. Benchmarking and analysis of flexible backbone membrane protein docking. 2023.

4: **Harmalkar A**, Pietz K, Leonard E, Horenberg AL, Grayson WL, Spangler JB, Gray JJ. Structure-driven design of orthogonal protein-protein interfaces. 2023.

Presentations

Invited Talks

Flatiron Institute, *Center for Computational Biology*, February 2023.
Harmalkar A "Computational modeling and design of protein-protein interactions"

Machine Learning to Accelerate Biology, *Harvard Systems Biology*, September 2022
Harmalkar A, Rao R, Gray JJ, Wei KY, "Towards generalizable prediction of antibody thermostability"

Oral Presentations

American Institute of Chemical Engineers Annual Meeting, Nov 2022, Pheonix, AZ, USA.
Harmalkar A, Zacharias M, Gray JJ, "Resolution exchange Monte Carlo for protein-protein docking."

American Institute of Chemical Engineers Annual Meeting, Nov 2022, Pheonix, AZ, USA.
Harmalkar A, Zacharias M, Gray JJ, "Capturing large-scale conformational changes on protein interfaces."

European Rosetta Conference on Protein Modeling and Design, May 2022, Warsaw, Poland.
Harmalkar A, Mahajan SP, Gray JJ, "Induced fit with replica exchange improves flexible backbone protein docking"

American Institute of Chemical Engineers Annual Meeting, Nov 2021, Boston, MA, USA.
Harmalkar A, Gray JJ, "Replica exchange and backbone sampling methods improve protein-protein docking by mimicking induced-fit pathways."

American Institute of Chemical Engineers Annual Meeting, Nov 2020, Virtual.
Harmalkar A, Gray JJ, "Coupling enhanced sampling with Monte Carlo techniques improves flexible backbone docking."

Annual Summer Rosetta Conference, August 2020, Virtual.
Harmalkar A, Gray JJ, "Coupling enhanced sampling with Monte Carlo techniques improves flexible backbone docking."

Poster Presentations

Keystone symposia on computational modeling of biomolecules (March2023), Banff, Canada.
Harmalkar A, Rao R, Tinberg CE, Gray JJ and Wei KY, "Spying on the sequences: towards generalizable prediction of thermostability using machine learning"

Biophysical Society (Feb 2023), San Diego, CA.
Harmalkar A, Rao R, Tinberg CE, Gray JJ and Wei KY, "Spying on the sequences: towards generalizable

prediction of thermostability using machine learning"

European Rosetta Conference on Protein Modeling and Design (May 2022), Warsaw, Poland.
Harmalkar A, Rao R, Tinberg CE, Gray JJ and Wei KY, "Spying on the sequences: towards generalizable prediction of thermostability using machine learning"

American Institute of Chemical Engineers Annual Meeting (Nov 2021), Boston, MA, USA.
Kizerwetter M, Harmalkar A, Leonard EK, Horenberg AL, Grayson WL, Gray JJ, Spangler JB. "Directed Evolution of PDGFR-β and PDGF-BB for Promotion of Bone Regeneration."

Winter Rosetta Developers Conference (Feb 2020), New York City, NY, USA.
Harmalkar A, Gray JJ, "Coupling enhanced sampling with Monte Carlo techniques improves flexible backbone docking."

7th Annual CAPRI Evaluation Meet (Apr 2019), EMBL-EBI, Hinxton, UK.
Harmalkar A, Gray JJ, "Tackling the conformational search space challenges in flexible protein docking."

Software

TherML
A machine-learning model for antibody thermostability prediction and design
- https://github.com/AmeyaHarmalkar/therML (*public release post publication*)

ReplicaDock 2.0
Induced-fit flexible backbone protein docking method in Rosetta
- https://github.com/RosettaCommons/main
- Demos available in : https://github.com/RosettaCommons/demos/tree/master/public/replicadock2

Scientific Benchmark Tests
Docking benchmark tests in Rosetta
- Code availability: graylab.jhu.edu/download/rosetta-scientific-tests/
- Docking Scientific Test: https://graylab.jhu.edu/download/rosetta-scientific-tests/main/docking_ensemble

Honors and Awards

Deutscher Akademischer Austaschdienst (DAAD) Research Fellowship 2021

Outstanding Graduate Teaching Assistant Award, Johns Hopkins University. 2021

Discovery Award, Johns Hopkins University Provost Research 2020
$100,000 grant for novel collaborative projects under Dr. Jeffrey J. Gray and Dr. Cynthia Wolberger.

Best Student from the penultimate year (Junior Year), ICT 2017

UGC Summer Research Fellowship, ICT 2016

Sir Ratan Tata Trust Scholarship for meritorious students in Chemical Engineering, ICT 2015

Patents and Consulting

US Provisional Patent 2022
Machine learning Techniques for predicting single-chain variable fragment (scFv) thermostability.

Technical Consultant, Baxalta, Inc 2021
Generated models and evaluated binding modes for a patent case.

Provisional Patent, India (*lapsed*) 2018
Process for nutrient recovery from human urine and utilization of treated urine.

Teaching Experience

Co-instructor PyRosetta Bootcamp May 2021

- Tutored and assisted bootcamp participants in the intense week-long crash course to developing in PyRosetta.
- Taught Python and Rosetta software programming to students.

Graduate Teaching Assistant (EN 540.630) Fall 2020
- 4 credit core ChemBE PhD class on Thermodynamics, Statistical mechanics and Kinetics.
- Develop curriculum and taught Thermodynamics and Statistical Mechanics to incoming PhD students.
- Created and graded homework and programming assignments, final exams, and recorded online lectures for students.

Graduate Teaching Assistant (EN.540.409) Fall 2019
- 4 credit core ChemBE undergraduate class on modeling, dynamics and control of biological systems
- Developed curriculum with Dr. Jeffrey J. Gray and taught 3 lectures on instrumentation and process control of biomolecular systems.
- Conducted tutorial sessions every week, developed and graded midterm and final exams.

Mentoring Experience

Priyamvada Prathima: JHU ChemBE undergraduate 2019-2022
- Advised and mentored on a membrane protein modeling and docking project.
- Helped to apply and receive Eleanor Muly Award and currently assisting in finishing a first author publication.
- Current: Research Associate at Harvard Medical School.

Ranjani Ramasubramaniam: JHU BME undergraduate 2020-2022
- Mentored on an computational and experimental collaborative project for design novel bi-specific binders.
- Helped to apply and receive the Provost's Undergraduate Research Award (PURA) for independent undergraduate research.
- Current: Ph.D student, University of Pennsylvania, USA.

Brandon Ameglio: JHU Biophysics undergraduate 2021
- Mentored and advised on developing a new approach for modeling antigen-antibody complexes.

Graylab CAPRI Team 2019-present
- Led the graylab Criticial Assessment of PRedicted Interactions (CAPRI) team in Rounds 46-54
- Helped to develop and document new protocols and methods for structure prediction of complexes.

Service and Outreach

Outreach Lead, National Diveristy in STEM, SACNAS 2022
- Organized and led the RosettaCommons outreach booth in SACNAS.
- Promoted our Post-bac, REU and graduate school programs to recruit students in STEM fields.

Instructor and Organizer, Rosetta Pre-College Intensive Workshop 2022
- Instructor and organizer of BioComp 2022, a computational biology bootcamp for Baltimore high-school students.
- Promoted and secured funding for the program to provide stipend and educational resources to the students.
- Created teaching materials and programming worksheets for the program.

Rosetta Commons Outreach Fellow 2021
- Coordinated and managed RosettaCommons booths in SACNAS National Diversity in STEM, 2021, and ABRCMS 2021 conferences
- Promoted and organized interactive sessions to attract talented young scientists and provide them with resources for graduate school exploration.

Academic Service

Manuscript Review.

Machine Learning for Structural Biology, Nature Methods, PLoS Computational Biology, J.Chem. Theory and Computation

XSEDE allocation grants
Wrote successful research proposals to XSEDE computing resources allocation for Gray lab (>100,000 SUs).

Maximizing Investigators' Research Award (MIRA; R35) for Dr. Jeffrey Gray
Contributed to writing in the MIRA proposal with Dr. Jeffrey Gray

Johns Hopkins Provost's Discovery Award
Contributed to writing the discovery award proposal for a collaboration with Dr. Cynthia Wolberger, JHMI on designing asymmetric histone interfaces.

Skills and Interests

Computational tools:
Programming languages: C++, Python, MATLAB, LaTeX, Bash (Unix shell)
Softwares: Git, Linux, PyTorch, Rosetta, MDAnalysis, Rosetta Biomolecular modeling suite software, MODELLER Homology Modeling, BLAST Search Tool, ClusPro, Robetta.

Hobbies: Rock-climbing and Bouldering, Hiking, Sketching, Painting, Running, Reading.

Languages: English, Marathi, Hindi, German.