THE UNIVERSITY OF CHICAGO

THE PROS AND CONS OF BACKBONE FLEXIBILITY IN MOLECULAR DYNAMICS DOCKING OF PROTEIN COMPLEXES

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

AND

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY

NABIL F. FARUK

CHICAGO, ILLINOIS

DECEMBER 2021

To my family, those present and departed

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I begin by thanking my PhD advisors, Tobin Sosnick and Benoit Roux, for their guidance and example. Tobin is a great example of the power of “vertical integration” in science, his deep knowledge of biophysical fundamentals and phenomena informed through experiments has allowed him to ask the right kinds of questions when developing, using and evaluating models and simulations. It was also motivating to see that Tobin is still hands-on with science, particularly with running his own tests and investigations with our *Upside* model. Benoit also benefits similarly through his tight collaboration with experimental groups, and his theoretical aptitude contributed to the framework of my major project, including how to conclude it on a positive note about what we learned. I also thank Benoit for supporting my application to the Biophysical Sciences graduate program at UChicago after meeting him at a conference, and I thank my former Master’s advisor Pierre-Nicholas Roy for introducing me to him and encouraging me to apply.

I thank my committee members Karl Freed and Tony Kossiakoff. Discussions with them over the years helped expand my work to a more complete, publication and presentation worthy state.

I thank my colleagues and major collaborators, John Jumper and Xiangda Peng. John was my former student mentor for my early years, and I learned much about machine learning and advanced programming from him. He was brilliant and had an independent and innovative mindset but was always patient about explaining concepts and code. He was also a family man and a good inspiration on balancing family life and work. After John left, I thought I would face a lonely battle with the complex *Upside* codebase. But Xiangda soon arrived as a postdoc and quickly learned about much of *Upside*’s inner workings. He contributed to the implementation of my new potential energy terms. I also thank my other lab members for many beneficial discussions and giving their time to help polish my work.

I thank the Graduate Program in Biophysical Sciences. The administrators, my cohort, and other students made an inclusive, engaging, and fun environment to conduct science and share ideas, while enjoying life during grad school. Special thanks to Michele Wittels for keeping me and all other students on track, and Adam Hammond for encouraging broader community service through the Artifice outreach program. It was a privilege to have the opportunity to serve with the admin and my peers on graduate recruitment and BSAB.

Finally, I thank my family and friends. As immigrants, my parents sacrificed much to give more opportunities to my brothers and I and imprinted the importance of education upon us. My grandparents and extended family were also important influences. A special mention goes to my paternal grandmother, who struggled but persevered as single mother with many young children when my grandfather died. She always hoped someone from her family would obtain a PhD, and I am hopefully slated to be the first. I'm also grateful for my brothers, who were my closest companions growing up. And I'm appreciative of the latest additions to my family, my wife Naila and my in-laws. Naila has been a light and motivator for me during the final year of my PhD, and her family members are wonderful and welcoming people.

I have a variety of friends and family friends from different stages of my life who've supported me and made for a wholesome time growing up. Special thanks to my friends in the Muslim community at UChicago and Chicagoland at large for making it feel like a home away from home and helping keep me spiritually grounded.

ABSTRACT

We previously developed *Upside*, a near-atomic, fast molecular dynamics algorithm for protein folding. A key feature of the model's efficiency is the representation of sidechains as single coarse grained beads and a rapid calculation of their rotamer free energies for each time-step, giving a smoother energy surface for the backbone to evolve on. We used the contrastive divergence technique from machine learning to train from simulations of 450 proteins for which our model's efficiency allows for better representations of the Boltzmann ensembles for precise tuning and greater accuracy. The model is afterward able to de novo fold proteins up to 100 residues on a single core in days.

Here we were inspired by *Upside*'s folding performance to adapt the model to predict protein-protein binding. Predicting protein binding is a core problem of computational biophysics. That this objective can be partly achieved with some amount of success using docking algorithms based on rigid protein models is remarkable, although going further requires considering the effect of protein flexibility. However, accurately capturing the conformational changes of the proteins upon binding remains an enduring challenge for docking algorithms. We use *Upside* to investigate when backbone flexibility helps docking predictions, what types of interactions are important, and what is the impact of coarse-graining on accuracy. These efforts also shed new light on the relative challenges posed by folding and docking. After training the *Upside* potential for docking, the model is competitive with established methods, but with some loss of accuracy due to the absence of atomistic side chains. Allowing for backbone flexibility during docking appears to be generally detrimental, as the presence of comparatively minor (3-5 Å) deviations relative to the native folded structure has a negative effect on performance. While this issue appears to be inherent to current forcefield-guided flexible docking methods, antibody-antigen complexes represent a major exception. These systems involve the co-folding of flexible loops that benefit from *Upside*'s backbone flexibility.

CHAPTER 1

INTRODUCTION

Proteins are the major end effectors of biology. They carry out functions of the immune system, signaling, metabolism [1] and even regulate the earlier stages of the central dogma through interactions with nucleic acids [2,3]. The study of proteins is therefore essential to understanding the processes of life and can lead to biomedical applications when investigated in the context of their malfunction. It is also rewarding to study proteins for their own merit as fundamental units of complexity: somehow a balance of energies and entropy expressed through a sequence of simple chemical constructs from a finite library of possible types (the amino acid residues) leads to a well-organized structure that can embody a function. This sequence-structure relationship is broadly described as “the protein folding problem”. Human curiosity cannot but help to distill the principles that govern the emergence of this complexity.

Experimental techniques such as x-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryoElectron microscopy (cryoEM) have revealed the details of many individual protein structures and a fewer but growing number of protein complex structures. However, these techniques suffer from limitations. For example, in x-ray crystallography, these include solubility and aggregation issues and non-directional nucleation stemming from flexible regions of proteins for x-ray crystallography [4]. NMR can face issues of time-consuming sample preparation, difficulty in analysis, and limitations on size and resolution [5,6]. Proteomic experiments suggest that there are many potential protein complexes, often with weak and transient interactions, which would be precluded from study with these structural characterization techniques [7].

Proteins are also not static structures; they fluctuate in their native state and may undergo conformational changes. Furthermore, stabilities and kinetics of proteins and their complexes impact their biological functions. Thus, it is important to study proteins beyond their native structure. Hydrogen-deuterium exchange (HDX) is a powerful experimental technique to provide residue-level information about thermodynamics and dynamics [8,9]. While advances have been made to ease the conduction of HDX experiments [10], they are still labor intensive.

The development of computational approaches for determining protein and protein complex structure, dynamics, and thermodynamics, are therefore promising for the study of cases that are difficult for experimental methods. Computational approaches may also enable higher-throughput study than experiment through parallel computing. This is supported by the remarkable accuracy and efficiency of recent neural network approaches for protein (complex) structure prediction [11–14]. However, these approaches are not able to account for thermodynamics and kinetics, unlike molecular dynamics (MD) simulation methods.

MD methods thus are comprehensive tools for the study of proteins. They also allow for a more physical interpretation of the determinants of protein folding and binding compared to neural network methods because of their use of explicitly defined forcefields that describe interactions and these forcefields are often physically based. However, successful prediction with MD of full protein behavior is predicated on the dual challenges of an appropriate representation and balance of energies in the forcefield to produce an accurate Boltzmann ensemble of states, and efficiently sampling that ensemble.

This thesis presents two major instances of my contributions to *Upside*, a fast, near-atomic MD algorithm [15,16]. *Upside* addresses both issues of energies and sampling. The model contains a single coarse grained (CG) bead in place of explicit sidechains and a quick, iterative procedure is used to determine side chain rotamer probabilities per MD step to give a smoother free energy surface for the backbone to move on. This and other approximations such as a lack of explicit solvent allows for rapid sampling on single CPU cores. Energy balance is achieved through machine learning (ML) based training of physically inspired potential energy terms to best reproduce different experimental observations, which are described below for each chapter. Interestingly, the different energy terms and training for each scenario indicate that *Upside*, as with other MD methods, does not (yet) present a unified model of protein behavior. An examination of *Upside*'s limitations is still instructive in gaining fundamental insight into the nature of proteins, such as what interactions are important for binding compared to folding.

The first chapter describes my major project of extending *Upside* to the prediction of protein-protein interactions, specifically for protein docking. Capturing the conformational changes that occur at protein interfaces for binding remains a considerable challenge according to the community-wide protein docking prediction competition, Critical Assessment of PRedicted Interactions (CAPRI) [17]. This

motivated us to attempt to utilize *Upside*'s protein folding capability for docking. This also involved the addition of new binding specific energy terms and training these terms using a maximum likelihood approach on a protein complex benchmark set [1].

The second chapter highlights my contributions to one of the seminal *Upside* papers where we predicted protein folding [16]. In this paper, sidechain and backbone potential energy terms are trained in concert using contrastive divergence to stabilize the native wells centered on the structures of proteins in our training set taken from the PDB. We achieved de novo, reversible folding for several proteins at an accuracy level compared to much more computationally expensive all-atom methods with better, though not perfect, denatured states.

I also contributed to the other original *Upside* paper that validated our sidechain model [15], but do not include it in this thesis. In this paper, a maximum-likelihood approach was used to train our sidechain interaction parameters to maximize the probability of the native χ1 rotamer given the native backbone configuration observed in our training set of proteins from the Protein Data Bank (PDB) [18]. We obtained similar accuracy as other state-of-the-art methods but using only a fraction of the computational time.

This thesis mainly focuses on structural predictions with *Upside*, with some thermodynamics in the case of the contrastive divergence paper. I also contributed to a study with a fuller account of *Upside*'s thermodynamics through the prediction of HDX patterns [9]. That study demonstrated that, although significant challenges remain for accurate free energy surface generation, MD methods remain relevant for the study of proteins considering the advent of neural network methods. *Upside* in particular is suited for this with its careful consideration of the denatured state ensemble. Additionally, the ideas for training against docking decoys in my protein binding work were adapted for the training of an intermediate improved folding force field against misfolded states [19].

1.1 References

[1] Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, et al. Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. J Mol Biol. 2015 Sep 25;427(19):3031–41.

[2] Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacology. 2013 Jan;38(1):23–38.

[3] Buschbeck M, Hake SB. Variants of core histones and their roles in cell fate decisions, development and cancer. Nat Rev Mol Cell Biol. 2017 May;18(5):299–314.

[4] Holcomb J, Spellmon N, Zhang Y, Doughan M, Li C, Yang Z. Protein crystallization: Eluding the bottleneck of X-ray crystallography. AIMS Biophys. 2017;4(4):557–75.

[5] Snyder DA, Chen Y, Denissova NG, Acton T, Aramini JM, Ciano M, et al. Comparisons of NMR Spectral Quality and Success in Crystallization Demonstrate that NMR and X-ray Crystallography Are Complementary Methods for Small Protein Structure Determination. J Am Chem Soc. 2005 Nov 30;127(47):16505–11.

[6] Puthenveetil R, Vinogradova O. Solution NMR: A powerful tool for structural and functional studies of membrane proteins in reconstituted environments. J Biol Chem. 2019 Nov 1;294(44):15914–31.

[7] Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, et al. CAPRI: A Critical Assessment of PRedicted Interactions. Proteins Struct Funct Bioinforma. 2003 Jul 1;52(1):2–9.

[8] Hu W, Walters BT, Kan Z-Y, Mayne L, Rosen LE, Marqusee S, et al. Stepwise protein folding at near amino acid resolution by hydrogen exchange and mass spectrometry. Proc Natl Acad Sci. 2013 May 7;110(19):7684–9.

[9] Peng X, Baxa M, Faruk N, Sachleben J, Pintscher S, Gagnon I, et al. Prediction and validation of a protein's free energy surface using hydrogen exchange and (importantly) its denaturant dependence. [Submitted Sep. 2021]

[10] Iacob RE, Engen JR. Hydrogen exchange mass spectrometry: Are we out of the quicksand? J Am Soc Mass Spectrom. 2012 Jun;23(6):1003–10.

[11] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Jul 15;1–11.

[12] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science [Internet]. 2021 Jul 15 [cited 2021 Jul 28]; Available from: http://science-sciencemag-org/content/early/2021/07/19/science.abj8754

[13] Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2 and extended multiple-sequence alignments. BioRxiv [Preprint]. 2021 bioRxiv 460468 [posted 2021 Sep 15; cited 2021 Sep 20]. Available from: https://www.biorxiv.org/content/10.1101/2021.09.15.460468v1.

[14] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer [Internet]. BioRxiv [Preprint]. 2021 bioRxiv 463034 [posted 2021 Oct 04; cited 2021 Oct 10]. Available from: https://www.biorxiv.org/content/10.1101/2021.10.04.463034v1.

[15] Jumper JM, Faruk NF, Freed KF, Sosnick TR. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. PLOS Comput Biol. 2018 Dec 27;14(12):e1006342.

[16] Jumper JM, Faruk NF, Freed KF, Sosnick TR. Trajectory-based training enables protein simulations with accurate folding and Boltzmann ensembles in cpu-hours. PLOS Comput Biol. 2018 Dec 27;14(12):e1006578.

[17] Lensink MF, Nadzirin N, Velankar S, Wodak SJ. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. Proteins. 2020 Aug;88(8):916–38.

[18] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

[19] Gaffney KA, Guo R, Bridges MD, Chen D, Muhammednazaar S, Kim M, et al. Lipid Bilayer Induces Contraction of the Denatured State Ensemble of a Helical-Bundle Membrane Protein. BioRxiv [Preprint]. 2021 bioRxiv 444377. [posted 2021 May 17; cited 2021 Oct 10]. Available from: https://www.biorxiv.org/content/10.1101/2021.05.17.444377v1.

CHAPTER 2

PROTEIN DOCKING

2.1 Introduction

Because of the central role protein-protein interactions play in many biological processes ranging from cell signaling pathways to antigen recognition, characterizing and predicting these interactions remains an important challenge of computational biophysics. The ability to accurately predict the conformation and binding affinity of protein complexes using computational approaches would be transformative. In this paper, we focus on the conformational aspect of protein-protein interactions: tools and limitations for flexibly docking proteins and the scoring of docked poses.

Two general approaches to protein docking are template based, which tends to be the most successful if homologous complexes can be identified, and free docking, often using Fast Fourier Transform (FFT) grid representations or basis function expansions to accelerate the generation of docking poses [1, 2]. The top “human” performer in the Round 46 joint Critical Assessment of Predicted Interactions (CAPRI) and Critical Assessment of protein Structure Prediction (CASP) experiment for protein docking used a hybrid pipeline based on the quality of templates but noted that manual intervention using prior knowledge of interface residues during modeling and scoring plays an important part in their success [3, 4]. The incorporation of evolutionary information also contributed to recent improved performance in CAPRI [5, 6].

In a comparison of free docking algorithms, the ones that incorporated protein flexibility were the best performers on Vreven et al.’s benchmark set of complexes [7]. Two methods for including protein flexibility are normal mode deformations (e.g., SwarmDock), and Molecular Dynamics (MD) refinement (e.g., HADDOCK) [7–9]. The HADDOCK approach also made use of bioinformatics predictions of interface residues and antibody loops to guide docking [7].

Two recent forcefield guided (pseudo-)dynamics protein docking methods also warrant mention. At one end of the scale of molecular details and computational resources is the all-atom explicit solvent replica-exchange MD approach of Pan et al. run on the Anton supercomputer [10]. At the other end of the scale is the CABS CG model, consisting of C_α, C_β, united sidechain atom placed at the sidechain center

of mass, and the peptide bond center, with a knowledge-based statistical potential that drives replica-exchange Monte Carlo pseudo-dynamics [19, 20].

AlphaFold 2 [13] and RoseTTAFold [14] are recent neural network approaches that combine evolutionary and structural features for protein structure prediction, with abilities to predict complexes. AlphaFold 2 achieved leading performance by a wide margin in the CASP14 experiment, while RoseTTAFold was developed afterward and comes in second place in a post evaluation of CASP14 targets but is smaller than AlphaFold 2 in terms of model size. Both have been trained only on single protein chains, but the authors of RoseTTAFold report some surprising success in predicting the structure of complexes, with backbone flexibility in the docking intrinsically built into the method due to its construction for protein folding.

The scientific community also started exploring protein complex structure prediction with modified AlphaFold 2 protocols upon its open source release with great success. One study using extended multiple sequence alignments obtained an accuracy of up to ~60% according to certain metrics in a test where traditional docking methods achieved only 22% accuracy [15]. Another study by the AlphaFold team at DeepMind with an AlphaFold model trained on multimers achieved 67% accuracy in predicting at least acceptable-quality heteromeric interfaces [16]. Traditional docking approaches, including template-based and free docking with statistical or physical potentials, may soon become obsolete for the sole purpose of structure prediction. However, it is still extremely important to improve molecular dynamics approaches for docking for studying the thermodynamics and kinetics of protein association, and the associated conformational changes.

The CAPRI experiment [17] has encouraged the development of docking algorithms through a blind prediction challenge open to the scientific community, with 47 rounds held since its inception. Reflections on recent rounds by CAPRI organizers and participants have highlighted the continued challenge of accounting for conformational changes during docking [3, 5, 17]. Round 46 was run jointly between CAPRI and the CASP experiment and continued to emphasize that considerable challenges remain. In this CASP-CAPRI round, only sequence information was provided, whereas in regular CAPRI rounds the unbound structures of the subunits were available. The hard targets that were poorly predicted did not have high quality templates available for homology modeling and so predictors often had only

subunits with large RMSD to their native bound states in their docking pipeline. It is thus attractive to consider leveraging a method with protein folding capability to capture the flexibility required to reach the bound conformations in such situations.

We also examine the recent information-driven antibody-antigen docking study by the HADDOCK developers [19]. They compared four different docking algorithms on antibody-antigen complexes from the Vreven et al. benchmark with different levels of information about the antibody hypervariable loops, also known as complementarity-determining regions (CDRs), and epitope to bias the search. Their own algorithm, HADDOCK, performs the best in part due to the information being incorporated as a restraining potential during the search as opposed to a simple filtering mechanism as used in the other algorithms, and due to their flexible refinement procedure. However, they have mixed performance for predicting the conformation of loop H3, the most variable antibody loop.

We recently developed the coarse-grained (CG) *Upside* model for protein folding simulations and now consider its suitability for docking prediction. *Upside* is a physics-based MD algorithm that is able to fold proteins 10^3-10^4 fold faster than all-atom methods with comparable accuracy. *Upside*'s speed arises from explicitly accounting only for the backbone N, C_α, and C atoms during the dynamics portion, while during force calculation it infers the position of amide hydrogens, carbonyl oxygens, and C_β atoms, and places the multi-position beads that represent the sidechains. Free energies of the side chain rotameric states are solved for between each dynamics step and pushed back onto the backbone atoms, which results in a smoother energy surface for dynamics as opposed to all-atom representations that face side chain friction and kinetic locking [20].

Upside can be slotted into a multitude of docking pipelines to make use of the extra information mentioned earlier in the context of other docking approaches, and below we discuss the case of information driven antibody-antigen docking. However, the primary goal of this paper is to assess the *Upside* model's suitability for predicting protein-protein interactions taking advantage of its backbone folding capability and rapid side chain sampling. To transition from folding, we extend *Upside* for protein docking by training new binding-specific energy terms with a maximum likelihood approach using Vreven et al.'s benchmark set of non-redundant complexes [7].

Our updated model is compared to other docking algorithms using a subset of the benchmark set according to the widely used CAPRI experiment criteria [17]. We identify the penalty for coarse graining sidechains and examine the impact of flexibility, including for other forcefield guided dynamics methods. We then assess flexibility and the performance enhancements of extra information in the case of antibody docking, following the reasonable assumption that *Upside* may perform relatively well for this class of complexes due to the inclusion of backbone dynamics enabling co-folding of the flexible CDR loops during binding. Finally, we conclude on a broader discussion of the different challenges of protein docking versus folding, and the merit of MD approaches in light of the recent performance of neural network methods.

The source code and examples for the docking version of *Upside* can be found at https://github.com/nffaruk/upside-docking. The release tagged v2.0.0 corresponds to this thesis.

2.2 Methods

2.2.1 Data sets for training and testing

The Docking Benchmark v5 provided 230 nonredundant binary complexes, of which 175 of the complexes from the previous version were used for training and the 55 new complexes were used for testing [7]. The benchmark set spans a diverse set of enzyme containing, antibody-antigen, and other types of complexes. The set also contains both bound and unbound forms of the subunits at high sequence identity; the unbound forms are required to compare against other docking algorithms according to the CAPRI methodology. FRODOCK v3 [2], a rigid body docking algorithm based on spherical harmonics, was used to generate 1000 decoys per complex based on the bound conformation of the subunits. FRODOCK has various energy terms that can be used for decoy generation and ranking. Here, we used the defaults for the van der Waals, electrostatics, and SOAP all-atom statistical potential [21], but omitted the desolvation term.

In addition, antibody-antigen (Ab-Ag) complexes were considered to evaluate the impact of backbone flexibility, considering the flexible nature of the antibody hypervariable loops that impart specificity. Ambrosetti et al. recently tested four docking algorithms for Ab-Ag docking using various levels of external information about the Ab loops and epitope to bias the results. Their evaluation set was 16

new Ab-Ag complexes added to the Docking Benchmark v5, and so to enable comparison we also used these complexes. Although these complexes are already included in the general “diverse” set, we docked them anew for two cases of extra information to take advantage of how the CDR loops are known *a priori* and how a rough estimate of the epitope can be sourced from experiment and other predictive tools. Whereas Ambrosetti et al. used the exact Ab loop residues and defined their coarse epitope by those residues within 9 Å of the loops in the native bound structures, for simplicity we defined the loops as residues involved in native interface contacts (Cα distances < 10 Å) plus a zone of up to 3 residues on either side and same for the coarse epitope. While our definitions and use of the loop and epitope information are different from Ambrosetti et al., we think they are sufficiently similar to allow for a meaningful comparison.

For the modeling using only Ab loop information, decoys were again generated with FRODOCK, except beginning with up to 20000 decoys and filtering down to 1000 around the interface by requiring a minimal number of loop contacts. For the Ab loop information plus coarse epitope information case, 200,000 decoys were first filtered down with the *frodockonstraints* program in the FRODOCK suite to keep the two furthest residues of the loops within 55 Å of the two furthest residues on the coarse epitope. These decoys were further filtered down to 1000 or less by selecting those that had contacts between at least one third of the residues of the loops and epitope.

2.2.2 New energy terms for docking

Side chain-side chain (SC-SC) interactions and burial (desolvation) provide important contributions to protein-protein interactions, and, for this reason deserve special attention. *Upside*’s basic potential represents sidechains by a single, directional bead that may be in up to six different states (positions and orientations) to mimic the diversity in the side chain rotamers [20]. The interaction between beads is given by a pairwise potential composed of radial and angular terms using cubic splines that offer flexibility in the form of the potential. This 2-body SC-SC interaction potential is used to determine the SC state probabilities, and in conjunction with intrinsic 1-body rotamer probabilities of the preference to be in a state in the absence of other sidechains, give rise to side chain free energies. These free energies are solved for in a self-consistent iterative procedure during each MD step using belief propagation (method of inference on graphical models) [14, 15]. Sidechain-backbone hydrogen bonding and sidechain-

backbone main atom interactions are incorporated into the 1-body rotamer probabilities during the free energy solution. The forces from these free energies are then back propagated onto the backbone atoms.

In the present treatment, additional corrective terms to the basic *Upside* potential was introduced. An interprotein SC-SC term, $V_{\text{inter\_rot}}$, copying the functional form of the original rotameric SC-SC term was added but it acts only between SC beads on different proteins. This was done such that the additional term would not affect the internal folding behavior of either protein. The SC-BB 1-body terms are excluded in this new term for simplicity. The cutoff was extended from the 7 Å of the base rotameric term to 10.5 Å to better account for possible long-range interactions of electrostatic residues that are more prevalent at protein interfaces [23].

Upside also has a many-body environment term to capture the effects of burial and desolvation. With this term, the number of sidechain beads are counted within a hemisphere above a virtual C_β of a residue, weighted by their rotamer state probabilities and residue types. This count is then coupled to a residue-specific energy composed of cubic splines. A new interprotein environment term, $V_{\text{inter\_env}}$, is added, again copying most of the functional form of the original implementation. However, this new interfacial term requires at least one bead from the opposite protein within the hemisphere for its activation.

The new potential is then given by $V = V_{\text{orig}} + V_{\text{inter\_rot}}(r, \theta_1, \theta_2) + V_{\text{inter\_env}}(N; w)$, where r is the distance between beads, θ_1, θ_2 are the angles between the bead orientation vectors and the displacement vector between the beads, N is the bead count weighted by rotamer probabilities and residue type weights w. We use the *Upside* folding forcefield v1.5 for our V_{orig}, which was developed with more diverse training ensembles and longer training cycles for better results than the first publication [24].

2.2.3 Training of the potential

To train and optimize the potential for protein docking, we initially used our original Contrastive Divergence machine learning methodology that we previously developed for protein folding studies [22]. This approach was based on populations (free energies) and minimizing the difference between the approximate distribution of states generated by *Upside* and the "true" distribution of crystal structures

found through experiment. However, it proved challenging to achieve the thermodynamic sampling required for this method to correct for a myriad possible misbound poses.

Therefore, we used simpler objective of minimizing the native poses' potential energies compared to the decoy poses and therefore maximize their Boltzmann probability. In this new strategy, we consider the average potential energy after short simulations for a set of *i, …, N* proteins complexes starting in *k, …, m* poses

$$\langle E\rangle_k^i, \quad \begin{matrix} k=0 \text{ is native} \\ k\neq 0 \text{ is decoy} \end{matrix} \tag{2.1}$$

We desire to maximize the Boltzman weight (population fraction) of the correct docking poses for the training set by minimizing the negative logarithm of the fraction:

$$\max_{\alpha} \frac{1}{N}\sum_{i}^{N}\left\{\frac{e^{-\beta\langle E\rangle_0^i}}{\sum_{k=0}^{m} e^{-\beta\langle E\rangle_k^i}}\right\} \Rightarrow \min_{\alpha} F = \frac{1}{N}\sum_{i}^{N} -\ln\left(\frac{e^{-\beta\langle E\rangle_0^i}}{\sum_{k=0}^{m} e^{-\beta\langle E\rangle_k^i}}\right) \tag{2.2}$$

where α denotes the energy term parameters. Rearranging and adding a term for regularization gives the objective function

$$F = \frac{1}{N}\sum_{i}^{N}\left(\langle E\rangle_0^i + 1/\beta \ln\sum_{k=0}^{m} e^{-\beta\langle E\rangle_k^i}\right) + \frac{\lambda}{2}\|\alpha\|_2^2 \tag{2.3}$$

Taking the gradient of this expression with respect to the parameters α yields

$$\frac{\partial F}{\partial \alpha_l} \cong \frac{1}{N_{c>5}}\sum_{i,c>5}^{N_{c>5}}\left(\left\langle\frac{\partial E_{0,j}^i}{\partial \alpha_l}\right\rangle_j - \sum_{k=0}^{m}\left\langle\frac{\partial E_{k,j}^i}{\partial \alpha_l}\right\rangle_j \frac{e^{-\frac{\beta E_{k,j}^i}{s}}}{\sum_{k=0}^{m} e^{-\frac{\beta E_{k,j}^i}{s}}}\right) + \lambda\alpha_l, \tag{2.4}$$

where the derivatives are averaged over *j* frames, *s* is a temperature scale factor for numerical stability ($s = 100$ in practice), and $c > 5$ is a condition to exclude well performing complexes. Such complexes were excluded from contributing to the parameter update if their native pose was in the top 5 of all poses because the information content of this pair largely had been extracted. The parameters are then updated for the subsequent training cycle according to

$$\alpha_{t+1} = \alpha_t - r\frac{\partial F}{\partial \alpha_t} \tag{2.5}$$

where *t* denotes the current cycle. In practice, the training set is divided into five minibatches that are cycled, and each training cycle involves 500 *Upside* Time Units ≈ 5 - 50 ns of simulation to relax the

poses of each complex up to 1000 residues. The frame output interval is 2 *Upside* Time Units. For larger complexes, the simulation time is scaled down by the number of residues according to $t'_{\text{sim}} = t_{\text{sim}}/t_{\text{int}} \times \frac{1}{25}\left\lfloor t_{\text{int}}\left(\frac{1000}{n_{\text{res}}}\right)^{3/2}\right\rfloor$ for performance reasons.

Cubic spline parameters for the new protein-protein terms are initialized to low values, $\alpha < 1$. We trained on the top 100 decoys of the bound subunit forms ranked according to *Upside* energy from an initial relaxation run. We did not find much benefit in conducting multiple training rounds, where decoys were reordered according to their new energies and a new top 100 selected for training.

2.2.4 Testing and evaluation using the optimized potential

We ran with two cases of restraints applied to the backbones, a semirigid case and a fully flexible case. In the "semirigid" case, Cα atoms were kept within ~3 Å of their initial positions with spherical flat-bottom quadratic potentials. The flexible case involves no restraints. Run duration was relatively short, 1000 *Upside* Time Units for $n_{\text{res}} \leq 1000$ and scaled down for larger complexes. The first half of each trajectory was discarded as equilibration, and centroid structures and their respective potential energies were selected as representatives for CAPRI Criteria evaluation and ranking.

We also check whether there is a performance benefit with full sidechains. The sidechains of the *Upside* structures from the semirigid restraint case are rebuilt using the SCWRL4 algorithm, which minimizes the energy from an atomic interaction model in conjunction with observed backbone-dependent rotamer frequencies to find the most likely rotamer states [25]. With the full sidechains rebuilt, the poses are rescored with SOAP-PP [21], an atomistic statistical potential used in the consensus scoring method of a former top CAPRI experiment group's docking server [6] and as a component in the scoring model of FRODOCK v3 [2].

For the Ab-Ag antibody set, we examined docking with both fully flexible and semi-rigid loops. For the flexible case, antibody residues involved in the native interface plus along with up to 5 residues on either side were kept flexible, in essence allowing the CDR loop residues to remain flexible, while the rest of the Ab fold was restrained with flat-bottom potentials. The semi-rigid case used the same restraints as the semi-rigid case in the full Diverse Set. Antigens were held within ~2Å C_α-RMSD with harmonic restraints, and able to move as a rigid body up to 10 Å of their starting positions.

To mimic Ambrosetti et al.'s "Scheme 2", whereby a restraining potential between the CDR loops and a coarsely defined epitope is used to bias the results, we apply C_β sigmoid contact potentials of -0.5 k_BT *Upside* energy units between all combinations of antibody interface residues and epitope residues. A note about *Upside* energy units is warranted: the correspondence to physical temperature is not well established for the new training, so we simply provide thermal energies in units of "k_BT". Unlike the Ambiguous Information Restraints used by the HADDOCK algorithm in Ambrosetti et al. that offers a level of smoothness and uniformity, our approach is simply pairwise additive.

Trajectories for each decoy pose are were clustered according to IRMSD for up to 3 clusters of frames within 1 Å IRMSD of the minimum energy structure of each cluster. The minimum energy structures of all clusters for all poses were sorted according to their energies and the top 100 are selected for further CAPRI Criteria evaluation.

2.3 Results

2.3.1 Training and testing of the potential

Upside's forcefield parameters were originally trained on a set of single domain proteins for folding [22]. A preliminary analysis of protein-protein docking performance with the original energy terms and parameters yielded many misbound poses that were energetically favored over the native bound state. In this work we sought to correct for these deficiencies by training new energy terms for protein docking using Vreven et al.'s benchmark set of non-redundant complexes [7]. These new terms are similar in form to existing ones, but only apply at the interface. We explore the question of whether more drastic changes to the single sidechain bead model are required, i.e. if we are limited by our degree of coarse graining, and postulate on the differences between protein folding and docking that gives rise to our differing performance on capturing the two tasks.

We followed the force field training protocol of maximizing the probability of the native pose as explained in Methods. The objective function decreased after training the new energy terms, leveling off after 15 cycles (Fig S2.1). This training procedure improved our ability to distinguish the native from the decoy poses, as compared to the ability of the original forcefield (Fig 2.1). The cumulative counts of finding the native poses at a higher rank (1 being the highest) increases compared to the decoys for all complexes

after relaxing the poses with *Upside* simulations with spherical flat-bottom restraints to keep each semi-rigid. For the training set (Fig 2.1a), the original forcefield predicts the native pose ranking in the top 10 for only about 8 of the 175 complexes (4.6%), whereas the force field trained for docking does so for about 42 complexes (24%), a substantial improvement (dashed lines in figure). The improvement is smaller for the test set, 1.8% → 12.7% (Fig 2.1b). Future training could benefit from a k-fold cross-validation scheme to prevent overfitting.

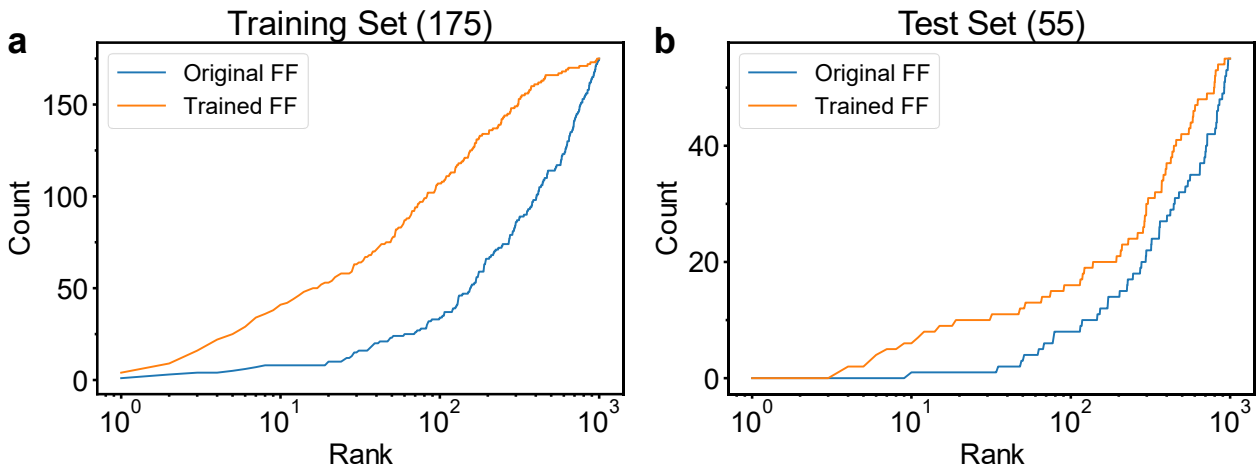

Figure 2.1: **Cumulative counts of native pose rank for trained protein-protein forcefield compared to original forcefield.** a) Training set performance, 175 complexes total. b) Test set performance, 55 complexes total.

The contributing factors to the improved scoring are hinted at by representative plots of the new residue specific potentials (Fig 2.2). Pairwise charge-charge interactions became more pronounced, particularly the repulsive ones (eg. Lys-Lys), with a signal extending beyond the original 7 Å cutoff (Fig 2.2a). Surprisingly, some hydrophobic interactions (eg. Leu-Leu) are among the most strengthened. Although the original folding training contains ample information for hydrophobic sidechain interactions, here at the interface they play a different role in the balance of energies because backbone hydrogen bonding does not play as large a role in binding as it does in folding. Fig S2.2a summaries the magnitudes of change for different residue pairs.

The environmental term potentials (Fig 2.2b) are more difficult to parse, due in in part to the burial number, a measure of desolvation, being determined by a summation of the burial weights of whatever residue types are within the hemisphere of the buried residue. The burial weights of the new potentials do

not appear to have a strong correlation with the size of the neighboring residues (not shown). However, the new potential allows some charged residues to have some favorable combination of burial, as shown for Asp at high burial numbers in Fig 2.2b, and as expected for charged residues to form an interface. Also, a compensation occurs with the pairwise sidechain potential that complicates interpretation. For example, the new Leu potential has unfavorable energies at most low burial numbers that likely helps refine the pairwise attraction term.

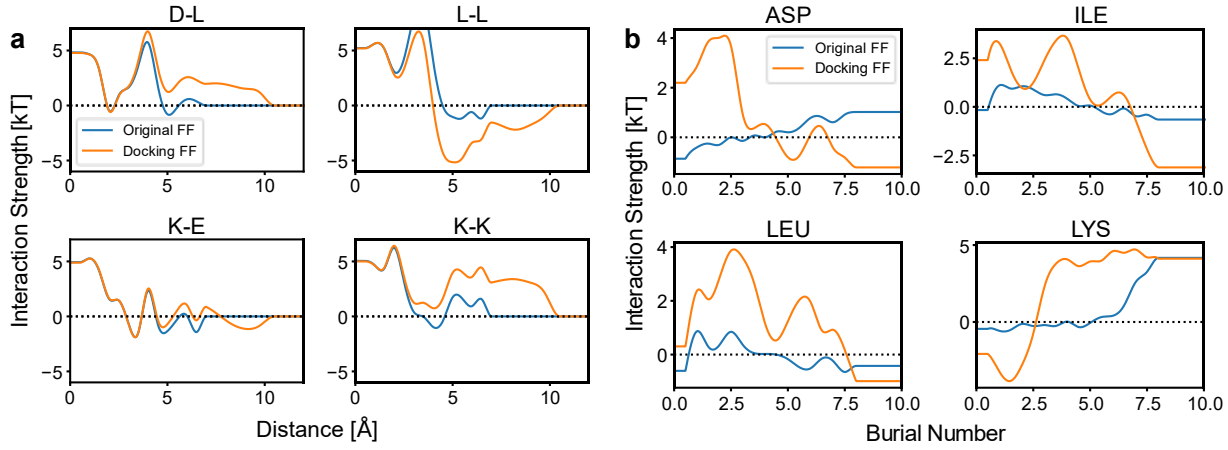

Figure 2.2: **Changes upon training for complexes for representative potentials.** a) Radial part of pairwise SC-SC potentials. The “after" plots are of the original potential along with the corrective contribution of the trained protein-protein potential b) Many body environmental potential and residue weights for the burial number (weights apply only to the new potentials). The “old” and “new” plots are of the separate contributions of the original and new interprotein terms.

2.3.2 Evaluation according to CAPRI criteria

We investigated combinations of subunit starting structures (from bound and unbound conformations) and restraints (semirigid and free) which were assessed with the full CAPRI criteria. In this criteria, the quality level of a prediction is assigned according to three metrics: Interfacial Root Mean Squared Deviation (IRMSD), Ligand Root Mean Squared Deviation (LRMSD), Fraction of Native Contacts (f_{nat}) [17,26]. Given imperfections in our scoring, there may be other poses acceptably close to native-like that fare better than the native pose. An evaluation with CAPRI criteria is more generous than looking at cumulative native ranks because decoys that are native-like according to CARPI and score well are included as successes.

Fig 2.3 shows CAPRI criteria performance of the trained protein-protein *Upside* forcefield in comparison with other docking algorithms featured in Vreven et al. 2015. Fig 2.3a Is adapted from Vreven

et al. and duplicated for comparison to the two subunit starting structure conditions (bound or unbound) for the *Upside* results. The bars indicate how many complexes have predicted native-like structures of a particular quality at different threshold levels of ranking/scoring, with T1 being the top 1 complex and T100 meaning that they are found in the top 100 complexes.

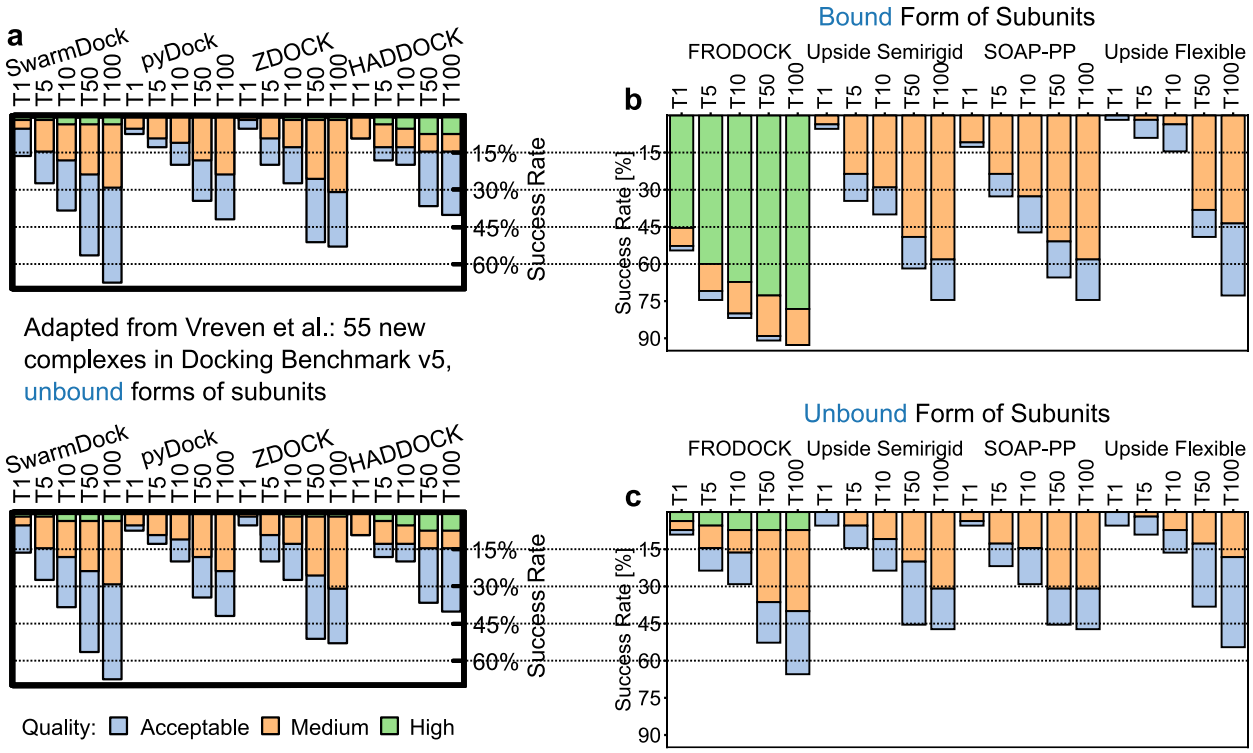

Figure 2.3: **CAPRI criteria evaluation.** a) Performance of four docking algorithms, adapted from Vreven et al., 2015; note that the plots are duplicated for comparison purposes with *Upside* results in b) and c). Their assessment was done on the same set of complexes as our test set. Notation: "T*X*", native pose found within the top *X* predictions. b) *Upside* results using bound forms of the subunits. From left to right are FRODOCK: ranking of FRODOCK poses according to the FRODOCK scores; *Upside* Semirigid: results using *Upside* relaxed poses using spherical flat-bottom restraints for the backbone (with added sidechains) ranked according to the *Upside* energy function; SOAP-PP: results for representative semirigid *Upside* structures scored according to the SOAP-PP atomic statistical potential; *Upside* Flexible: Results for complexes allowing for full backbone motions. c) *Upside* results using unbound forms of the subunits, with same types of subplots as in panel b).

Fig 2.3b highlights the results from the *Upside* pipeline starting from the bound forms of the subunits from the Dockground testing set. From left to right, there are the native-like poses from the initial FRODOCK rigid body docking ranked using their FRODOCK scores compared to the other decoys. Next are the rankings of representative structures from *Upside* after a minor (<3 Å) relaxation of the FRODOCK starting poses using semirigid backbone restraints. These structures are first scored using the *Upside* energy function. Sidechains are then added with SCWRL4 [25] (required for f_{nat} calculation) and

rescored with the all-atom statistical potential SOAP-PP [21]. This rescoring with complete atomistic sidechains examines *Upside*'s loss of accuracy due to the use of a single (albeit multi-position) bead for each side chain. Finally, there are the results for representative structures from *Upside* simulations without restraints.

When starting with the bound forms of the two partners including the native side chain rotamers, the FRODOCK results have a high success rate with a majority of the targets having high quality native-like predictions (red bar). This success is due in large part to the backbones and interfacial sidechains being fixed in their native conformations and rotamers, and FRODOCK can find a well-matched, interdigitated interface between the native positions of the binding partners. Once processed into *Upside*, the exact atomistic positions of the sidechains are lost. This situation could occur with low resolution structures, for which the side chains positions are not as well determined as the backbone (e.g., in nuclear magnetic resonance spectroscopy or cryo-electron microscopy-based structure determination). Nevertheless, *Upside* still performs well on ranking native-like poses compared to the other docking algorithms of Vreven et al. (Fig. 2.3a) when the *Upside* backbones are kept semirigid (note however, that the predictions of the other algorithms use the unbound forms of the subunits, while Upside is using the bound forms. So this comparison is not completely valid).

The SOAP-PP results are calculated using *Upside*'s optimized structures, followed by full sidechain addition by SCRWL4 [25] and then scored using the SOAP-PP energy function. The results with full side chains are notably better only for predicting the native-like pose as the lowest energy structure (T1 performance: 5.4% → 12.7%). For being in the top 5+ lowest energy predictions, however, there is minimal difference apart from the cumulative effect from the T1 performance increase. This may suggest that for most situations, there is only a very mild decrease in performance when using *Upside*'s single sidechain bead at the scoring stage once the backbone is determined. However, the limitations of SCWRL4 and SOAP-PP must be considered. The χ_{1+2} accuracy of SCWRL4 was 80% on a test set of proteins when compared to the crystal positions of sidechains with high electron density, and less for higher χ angles (e.g. 47% χ_3 accuracy for Arg) [25]. Furthermore, in SOAP-PP's original paper, it only had 40% success in placing native-like predictions in the top 10 on a prior version of the Vreven et al. docking benchmark when the subunit backbones and sidechains were in their exact native positions (i.e., it had

the best possible starting structures to score) [21]. Thus, there is still room to benefit from a more accurate model for the sidechains, which is supported by the high success rate of the FRODOCK results with the native sidechain rotamers.

Another major finding is the drop in quality and ranking of native-like structures that occurs when the proteins are allowed to be fully flexible during the *Upside* simulations. The subunits drift away from their native bound forms, indicating that there is insufficient accuracy in the underlying folding component of the *Upside* forcefield. Such deviations are not compensated for by that the optimized interprotein energy terms. We characterize this issue further in subsequent sections and investigate to what extent it may be general for all forcefield-guided dynamics methods. And, although there is movement away from the native state when starting from the bound forms, at this point we were hoping there would be movement towards the native state when starting with the unbound forms, which is the more relevant scenario.

Fig 2.3c shows the results when starting from the subunit in their unbound conformations. In this case, the FRODOCK results are much worse compared to the bound form results since the tight fit at the native interface is lost due to backbone and sidechain conformational differences. FRODOCK still performs better than the other docking algorithms, possibly since it incorporates SOAP-PP in its scoring and may have other advancements since it is a newer algorithm.

Upside's decrease in performance due to coarse-graining of the side chains would be acceptable if it improved on the FRODOCK results of the unbound forms by compensating with its ability to sample different backbone conformations and sidechain rotamers to find a better fit at the interface. However, *Upside* and its conformational sampling exhibit poorer performance, even under semirigid restraints, with both fewer native-like poses ranked highly and more lower quality structures. SOAP-PP is able to improve performance for some of the T1-T5 predictions for the semirigid *Upside* structures rebuilt with full sidechains. Although there is still some backbone deviation, we expect the semirigid case to emphasize the role of the sidechains and new interprotein energy terms. The results indicate that *Upside*'s ability to repack the coarse-grained sidechain beads did not yield an overall scoring advantage over FRODOCK. However, *Upside* with semirigid restraints achieves comparable performance to many of the docking

methods in Vreven et al., which we view as a partial success considering *Upside*'s disadvantage due to coarse-graining of the side chains.

Most importantly, the predicted structures have larger deviations from the native bound state when *Upside* is allowed full backbone flexibility, i.e., even the starting poses with the subunits in their unbound conformations are overall more native-like as compared to those generated when *Upside* is allowed to move the backbone. In the section "**Energetic costs of retaining native-like subunits**", we will return to this issue and characterize the energetic compensation required to shift the subunits from their unbound to bound backbone conformations in order to establish the magnitude of forcefield improvements needed.

The best performing methods in Vreven et al. are SwarmDock and HADDOCK. SwarmDock performs the best in terms of the percentage of acceptable quality native-like structures or better ranked in the Top 10, whereas HADDOCK has the most high-quality native-like structures. SwarmDock's success likely is partly due to its approach to backbone flexibility via normal mode deformation, allowing it to better address the more difficult targets that have a greater change between bound and unbound forms of the subunits. SwarmDock was the only successful method for the sole "easy" target, having a ΔIRMSD < 1 Å between bound and unbound forms. The authors hypothesize that this is due to SwarmDock being able to widen the narrow opening of the receptor binding site. In SwarmDock, the normal mode coefficients are updated in the search procedure in the direction of minimum energy, but the energy evaluations include only the interaction energy between the two binding partners and not their internal normal mode energy [8]. In effect, the range of allowed protein flexibility is somewhat artificial, as it is highly constrained by the selection of the normal modes with lowest frequency. With this strategy, SwarmDock is not penalized by backbone strain during search and scoring, whereas the backbone energy is an integral part of the physical forcefield that guides the dynamical motions in fully flexible models such as *Upside*. With such physical models, structural deformations are governed by the intramolecular potential energy and occurs spontaneously during the sampling, and their influence is implicitly included in scoring poses.

In HADDOCK, Vreven et al. utilized bioinformatics predictions of the interfaces and knowledge of antibody CDR loops to bias the docking results to make use of HADDOCK's "ambiguous information

restraints". HADDOCK's high-quality native-like structures may be a result of this extra information, in combination with HADDOCK's all-atom explicit solvent flexible refinement and an all-atom energy function. In the antibody section of this paper, we also test *Upside*'s performance with additional information to enable a more valid comparison to HADDOCK.

The latest studies for SwarmDock and HADDOCK reassessed their performance for a subset of the docking benchmark set with enhancements to their procedures (cross-docking diversification of the starting conformations for SwarmDock, use of higher levels of informational restraints for HADDOCK). These new studies show that there is still room for improvements in flexible docking for both normal mode and MD methods [10, 11].

2.3.3 Dissecting the impact of backbone flexibility

To quantify the impact of backbone flexibility on docking predictions and separate the latter from any deficiencies in our search process, we next ran a series of "best case" simulations starting from the native pose for all complexes (Fig 2.4). All complexes were run for the same base duration as the CAPRI evaluation runs (the run duration of larger complexes was not reduced). Each point in the plots is the average of three runs of a specific complex, where a centroid representative structure was taken from the second half of trajectories that began from the native bound state of the complex. The left and right columns present two different measures of backbone flexibility using either total subunit RMSD to the native bound state (left) or RMSD of the individual subunit interface (right). Note that the plotted subunit RMSD is taken from the Euclidean norm of both subunit's RMSDs, which was important for observing a pattern ($\|x\|_2 = \sqrt{\mathrm{RMSD}_1^2 + \mathrm{RMSD}_2^2}$). There is a stronger correlation between subunit conformations and the IRMSD of the complex when considering the RMSDs at each individual subunit interface compared to whole subunit RMSDs, as expected.

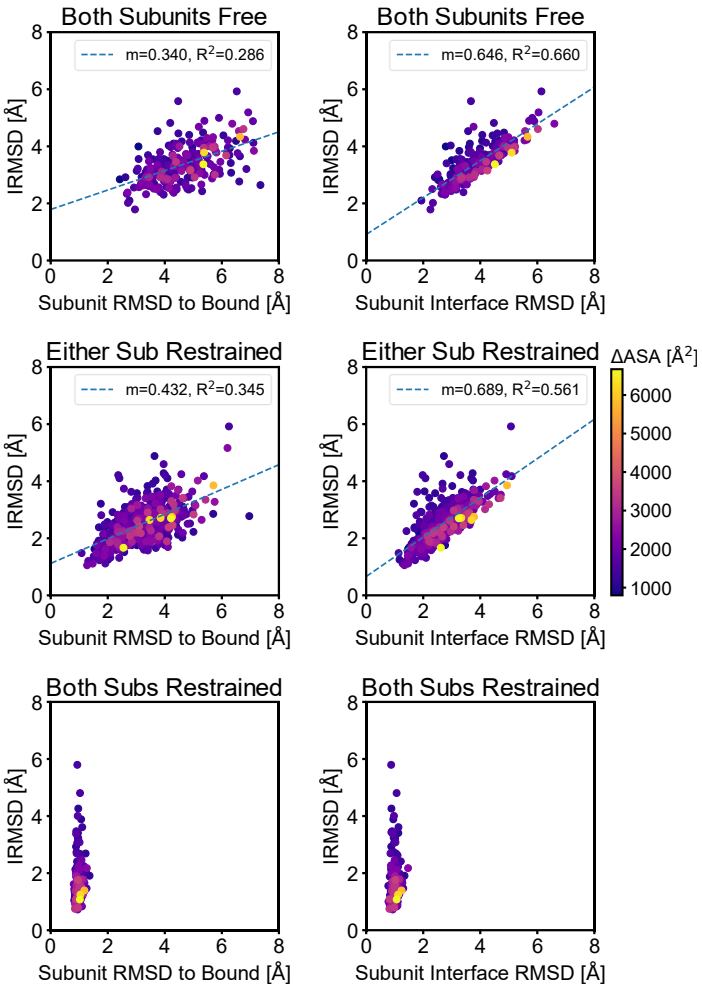

Figure 2.4: **Impact of *Upside* backbone flexibility's on IRMSD.** Left: Whole subunit RMSDs to their native bound forms. Right: RMSDs of the interfaces for each individual subunit. These subunit RMSDs are the Euclidean norm of both subunits. The y-axis is the IRMSD of the complex. The top row: no backbone restraints applied to the subunits; the middle row: harmonic restraints applied to either of the subunits to keep the backbones within ~1 Å of their native bound states (data from both cases combined); bottom row: restraints separately applied to both subunits. The points are colored according to change in accessible surface area, ΔASA.

When no backbone restraints are applied to the subunits (top row), we see that the subunit RMSDs mostly lie between 2 – 5 Å as are the IRMSD values. Conformations that are 2 – 5 Å RMSD from native conformation are generally considered a success for protein folding prediction, so *Upside*'s ability to maintain that RMSD for the subunit conformations could be considered laudable for a *de novo* coarse grained model. However, a substantial number of subunit RMSDs are above 5 Å. For those complexes, medium quality docking results are not achievable, as the requirements are IRMSD < 2 Å or LRMSD < 5 Å (if f_{nat} < 0.5). The net effect of inaccuracies for both subunits, which increases the challenge of predicting the interface, largely explains why allowing for backbone flexibility can be detrimental.

The simulations are improved to 1.5 – 4 Å subunit RMSDs and IRMSD after applying harmonic restraints to either subunit to keep its RMSD to ~1 Å of the original bound conformation (middle row). When both subunits are restrained, the IRMSD is largely below 3 Å. This doubly restrained case emphasizes the role of interprotein interactions and the quality of our training of these interactions by reducing the contributions of the backbone strain caused by imperfections in the folding portion of the forcefield. It is reassuring that for most cases the interprotein interactions are at a high enough fidelity to maintain a low IRMSD. For the complexes where both subunits are restrained yet have a large IRMSD, there are significant rigid body-like translocations of the subunits from the native bound pose. For these examples, our interprotein interaction terms are inadequate. The high IRMSD points correspond to low ΔASA complexes and we further examine the impact of interfacial area in a later section.

Another perspective on the effect of subunit backbone restraints is provided by 2D densities of native state simulations of 16 complexes in Fig Figure S2.3. Here, the densities are shifted to more native-like values of IRMSD and f_{nat} for most cases with subunit backbone restraints.

2.3.4 Energetic costs of retaining native-like subunits

The previous section examined the increase in subunit RMSDs when the proteins are in the complexes. We now focus on the role of backbone flexibility on individual, isolated subunits to examine the magnitude of free energy that would be required to shift the backbones into their native bound conformations, a necessity for obtaining good docking predictions. We again ran the subunits with *Upside* for the same base duration as the CAPRI evaluation runs, starting from their native bound conformations. Fig 2.5a shows representative plots of the Potentials of Mean Force (PMFs) generated from Gaussian kernel density estimates of the RMSDs to the bound native state taken from the second half of the trajectories. The two lines in each plot are for each of the two partners run separately and plotted from the lowest to the highest observed RMSDs. For subunits with RMSD $\geq$ 2+ Å, we observe free energies of up to 4.6 k_BT at their lower RMSD bounds (e.g., 3V6Z in Fig 2.5a). Conformations with lower RMSD would correspond to an even higher free energy cost, indicating that a substantial improvement of the folding portion of *Upside*'s energy function would be needed to consistently obtain structures with RMSD smaller than 2 Å.

We next examined the relationship between the individual RMSDs of the separated subunits and the IRMSDs of the unrestrained simulations of the complexes from Fig 2.4. The results in Fig 2.5b show a

classification of those simulated complexes, with yellow dots corresponding to complexes with one or both subunits are nonnative-like while red dots corresponding to complexes where both partners remain native-like. Complexes are considered native-like when both the partners have 50% of their RMSD distribution below 3 Å when individually simulated. Native-like subunits tend to have native-like IRMSDs, but there is large overlap between the classes. The existence of low IRMSD, but nonnative-like subunits indicate that regions of the proteins away from the binding interface experience the bulk of the conformational difference. On the other hand, a high IRMSD with native-like subunits is a situation where the binding partners experience rigid-body displacement, indicating a deficiency in the inter-protein terms of our energy function for those complexes.

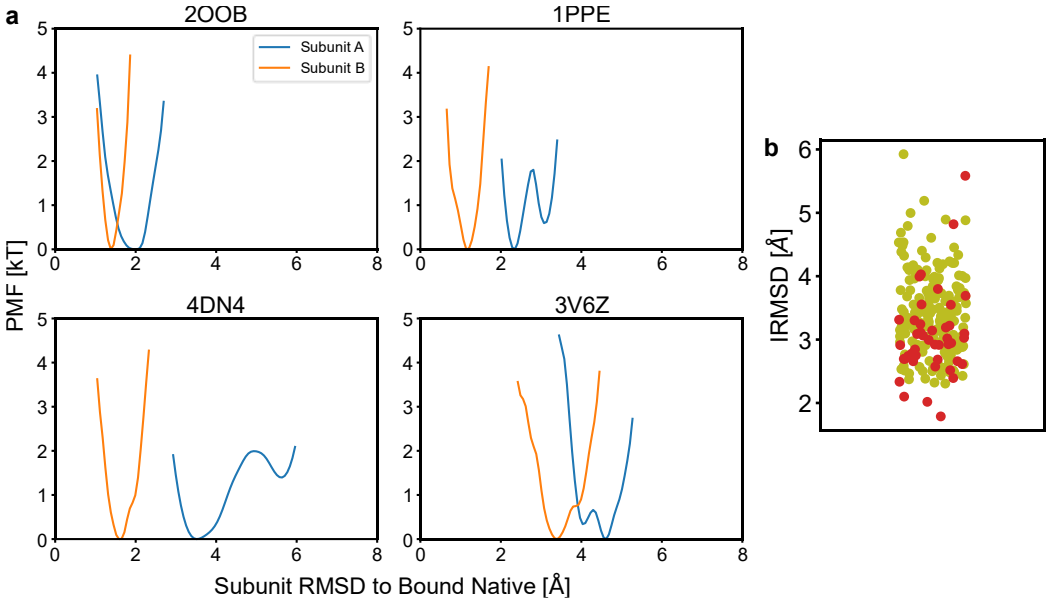

Figure 2.5: **RMSD difference between the native bound structure and the simulated monomers.** a) PMFs from RMSD distributions referenced to the native bound state to visualize the energy required to adopt bound native-like conformations. b) Classification plot of IRMSDs from unrestrained simulations starting from the native complex using the following threshold criterion on the RMSD distributions of the PMFs: complex with native-like subunits if 50% of both of their individual RMSD distributions are below 3 Å (red dots), otherwise they are considered nonnative subunits (yellow dots). The x-axis spread is artificial jitter to aid in the distinguishing of the points.

2.3.5 Backbone flexibility affects other molecular dynamics methods

We now compare *Upside* with other recent forcefield-guided methods to determine the generality of *Upside*'s decrease in performance when backbone movement is allowed. The all-atom explicit solvent MD approach of Pan et al. uses enhanced sampling akin to simulated tempering, which they call tempered binding [10]. The CABS CG model uses Replica Exchange with Monte Carlo moves that capture

transitions on physical timescales, which they call pseudo-dynamics [19, 20]. Both approaches let the dynamics guide the association of the protein partners, unlike the search process used in the *Upside* docking pipeline to start with up to 1000 pre-docked poses. The docking problem can be broken into three parts, the generation of many possibly bound poses followed by their refinement and scoring. In this work we focus on the last two steps for *Upside* as they are sufficient to address whether our model is capable of identifying the true native pose and to what degree does backbone flexibility help in this identification.

The CABS CG model is similar to *Upside*, but has slightly lower level detail, e.g., it includes angular dependance of SC-SC interactions, but single sidechain states [15, 19]. Hence, it offers an independent insight into the potential benefits of backbone flexibility for docking. In the CABS docking study, 12 complexes are free docked, which as mentioned earlier means that the binding partners begin separated at different initial positions for each replica and associate over the course of the simulation. The partners begin in their unbound native conformations and they applied restraints individually to each partner, with strengths such that the receptor fluctuates only around 1 Å and the ligand between 2-12 Å [28]. This setup is overall more flexible than the *Upside* unbound subunit semi-rigid case for the CAPRI evaluations, and the *Upside* case involves FRODOCK pre-docked starting states.

Fig 2.6 is a comparison of the *Upside* unbound subunit semi-rigid results and the CABS results taken from Table 1 of Kurcinski et al. for the 7 complexes common to both test sets. The *Upside* semirigid situation generally performs better than CABS for both lowest IRMSD observed from all poses (Fig 2.6a) and much better for lowest IRMSD in the top 10 ranked poses (Fig 2.6b), indicated by points below the diagonal. Thus, backbone flexibility is a detriment to the CABS CG model as well. “Simple” rigid-body docking, for example using FRODOCK, would have been better, considering that FRODOCK was used for the starting states of *Upside* for these targets and *Upside* tends to do worse than FRODOCK as shown previously. This test should have been ideally conducted as a “CABS semirigid” versus “CABS flexible” setup to remove influence from differences in force fields, but we think that *Upside* as a CG model is a suitable stand-in for the semirigid scenario. We also recognize that the CABS study had a focus on whether low IRMSD states could be sampled at all even if they were not among the top 10 predictions, and the authors acknowledge that scoring improvements for their model are required.

However, our comparison highlights how much of a detriment backbone flexibility can be when sampling and scoring are coupled.

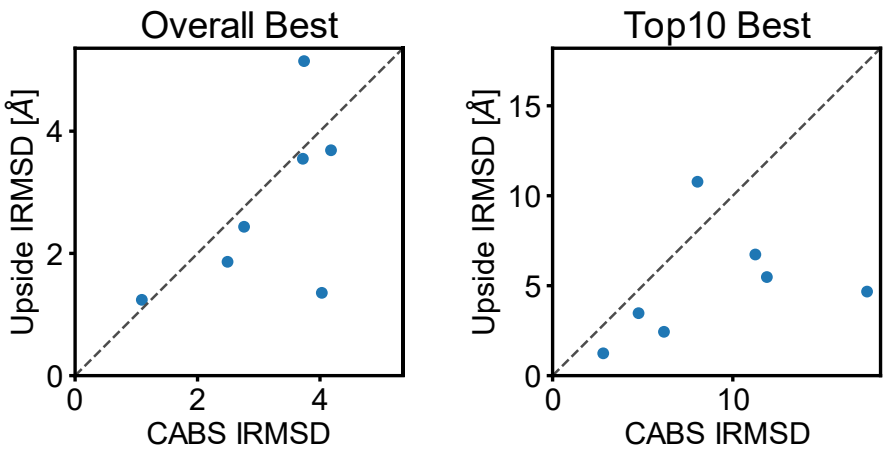

Figure 2.6: **Comparison of *Upside* with CABS docking for 7 complexes common to both studies.** a) Overall best IRMSD out of all poses. b) Best IRMSD among the top 10 ranked poses.

To see whether the issue of backbone flexibility is limited to CG models and their inherent inaccuracies, we next examine the binding MD approach of Pan et al. These all-atom explicit-solvent simulations combined with combined with an Hamiltonian tempering enhanced sampling procedure exploited the computing power of Anton 2 to allow the observation of reversible binding to the native state for five out of six of the complexes (the sixth irreversibly bound into a native-like pose). The native state was the most populated for each, with the IRMSDs of the most stable poses for all six complexes being below 1.3 Å. *Upside* does not have the same complexes as their set, but none of the top 10 poses for any complex were at that level of accuracy.

However, a few issues are worth noting. The chosen complexes in this study bind in a very rigid manner, with an IRMSD smaller than 2 Å between the unbound and bound forms of the subunits. Secondly, the simulations from each individual subunit are started in their bound conformation. And most notably, they apply backbone torsional restraints centered at the bound native structures for both subunits for four complexes, while the remaining two complexes have such restraints applied to one subunit. As we have seen from the *Upside* results, even slight backbone deviations can be very detrimental to the accuracy of the simulation and adding some restraints can substantially improve performance. Pan et al. note that the restraints help prevent conformational degradation at the microsecond timescales they need

to simulate in order to observe reversible association. This is observation is very consistent with the results presented here. While an in-depth assessment of the effects of torsional restraints was not provided (presumably limited by computational feasibility), the association rates of barnase–barstar with and without restraints was also examined with conventional MD. Interestingly, the predicted association rate was about five times slower without restraints than with restraints compared to experiment. In the end, it was concluded that detailed forcefields will have difficulty modeling systems for which the unbound subunit conformations differ significantly from the bound states when torsional corrections cannot be relied upon [10].

2.3.6 Determining features that contribute to performance

To identify areas of weakness in the present model or with the training of the potential function, we conducted an analysis of which factors contribute to the overall performance. We begin with a feature related to backbone flexibility, the amount of conformational change at the interface between bound and unbound forms of the subunits (Fig 2.7). The difficulty categories are as follows, Easy: IRMSD < 1.5 Å and $f_{non\text{-}nat}$ < 0.40, Difficult: I-RMSD > 2.2 Å, Medium: all others [7]. The performance labels for each complex (e.g., Native-like in Top 10, Poor Performers) are taken from the previous CAPRI criteria evaluation from the respective cases of starting conformation and whether backbone flexibility was allowed during the *Upside* simulations.

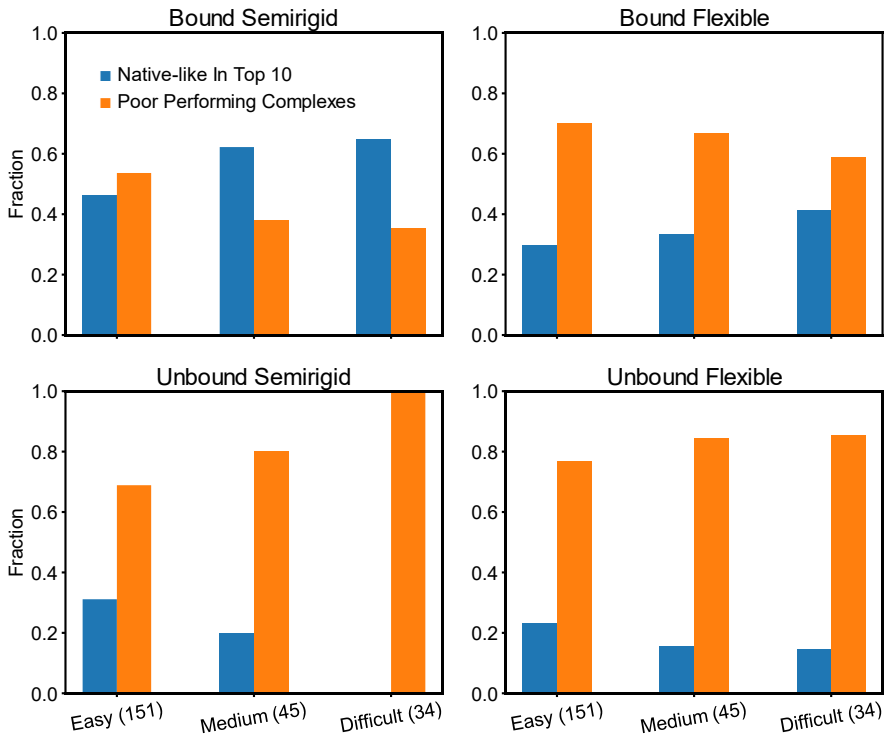

Figure 2.7: ***Upside*'s performance with difficulty class of complexes.** Each subplot corresponds to the subunit starting conformation and backbone flexibility scenarios of the CAPRI evaluation in Fig 3. Easy: IRMSD < 1.5 Å and $f_{non\text{-}nat}$ < 0.40, Difficult: I-RMSD > 2.2 Å, Medium: all others [Vreven 2015].

The bound semirigid case of Fig 2.7 again represents a best-case scenario, since backbone conformational search is removed and backbone drift minimized, thereby focusing on the scoring performance of *Upside*'s energy function. In this situation, we find that the performance is not very different between the difficulty classes. This finding is partly expected since we trained on the bound forms of the complexes, but also it also indicates that *Upside* does not have a harder time learning the properties of the interfaces for the different difficulty classes. In the bound flexible situation, performance decreases across the board due to issues with our and most other MD forcefields, as discussed in the previous section.

In the unbound and semirigid case, we have fewer native-like (top) performers in large part because the subunit backbones are being held in their unbound conformations. A contributing factor to this lack of native-like poses comes from the rigid-body docking stage of our pipeline where FRODOCK is unable to find the general binding interface due to a loss of lock and key fit of the surfaces. And even if the general native binding interface is found, the unbound conformation subunit backbone RMSD can be

a detriment to the IRMSD and CAPRI evaluation. The performance correlates with the difficulty class of the complexes as expected. When we allow for backbone flexibility in the unbound flexible scenario, we lose some performance on the easy complexes, but gain for the difficult complexes, implying that there exist some cases where our energy function can drive the backbones in the correct direction.

We next examine the effects of interface composition and size (Fig 2.8). The performance is judged according to the CAPRI bound semirigid scenario in order to focus on the binding energy instead of backbone conformation search. In Fig 2.8a, we compare performance based on the amount of pairwise interactions between Apolar, Polar, and Charged residues. The values for each type of pair interaction are normalized according to the maximum of the fraction at the interface between the good and poor performers to make it easier to compare the performance (i.e., the greatest value in each subplot is 1.0).

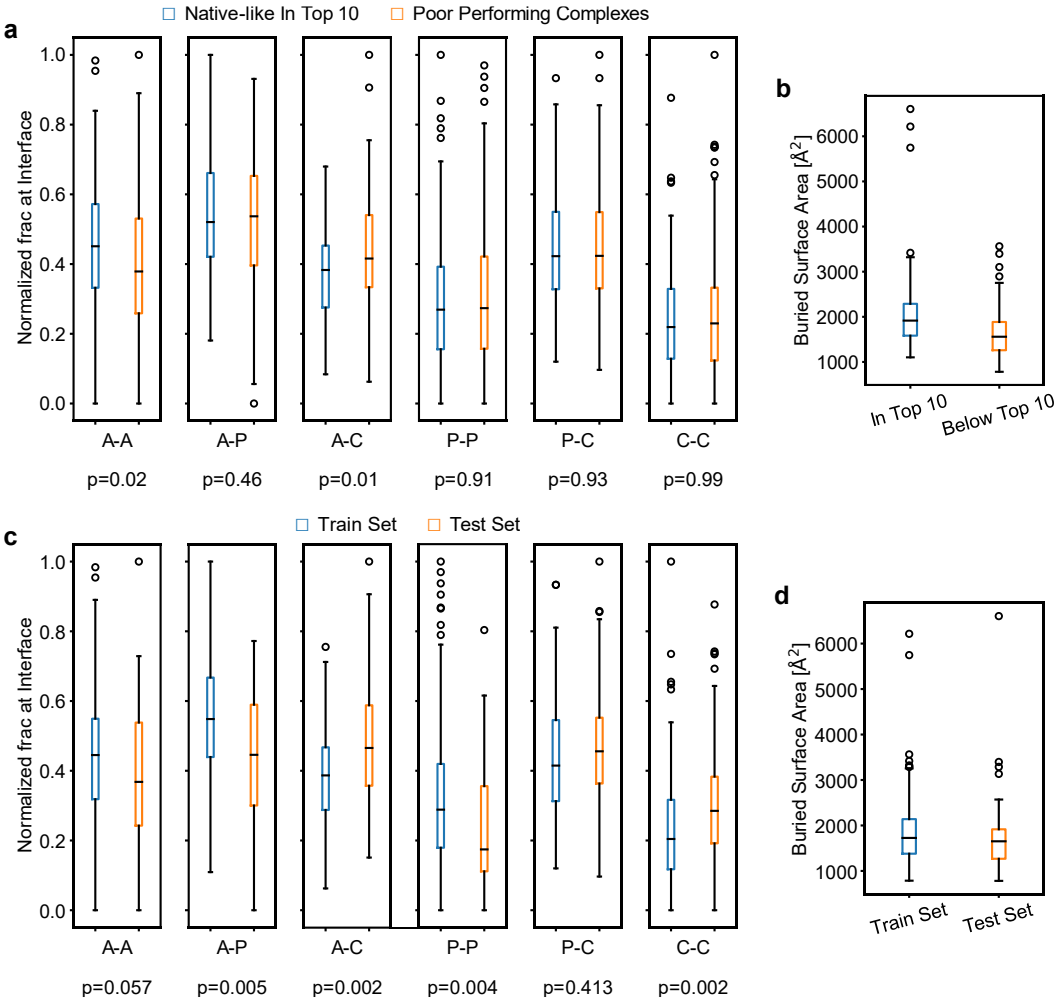

Figure 2.8: ***Upside*'s performance with types of interactions and interface size.** The performance labels correspond to the bound semirigid scenario in the CARPI evaluation. a) Comparison of the different

Figure 2.8: **Upside's performance with types** continued.
types of pairwise interactions between good and poorly performing complexes. A: Apolar, P: Polar, C: Charged. The values for each type of pair interaction are normalized according to the maximum amount between the good and poor performers. b) Comparison of the interface size between the good and poor performers. c) Comparison of the different types of pairwise interactions between the training and test sets. d) Comparison of the interface size between the training and test sets.

There are significant differences between the distributions of Apolar-Apolar and Apolar-Charged interactions for good and poor performers at 95% confidence level according to the Kolmogorov-Smirnov test. In Fig 2.8c, we see several significant differences in the interface compositions between the training set and the test set. Notably, Apolar-Charged interactions tend to be more numerous for the test set and hence, *Upside* had fewer examples of high Apolar-Charged interfaces to learn from during training. This may explain why Apolar-Charged interactions tend to be greater for the poor performing complexes. This suggests that *Upside*'s docking forcefield may be improved in the future by reducing the size of the test set from the current ~24% of the entire set to ~10%, considering that we have relatively few training examples for our number of parameters (albeit multiplied by the number of decoy poses). The training set should be divided further into k-fold cross-validation to find the best stopping point of training to prevent overfitting. Irrespective of whether these suggestions would produce significant improvements, for comparison purposes, we chose the current size and membership of the test set to correspond to that of Vreven et al.

Complexes with larger interfaces tend to perform better (Fig 2.8b). This finding may be because larger interfaces allow more opportunity for the cancellation of errors in the forcefield across all interactions, whereas smaller interfaces have higher variability. Larger native interfaces may also be more separated in size compared to other decoy poses of the complexes, and so generally are more attractive due to the van der Waals component of interactions, which *Upside* may have effectively learned in its training.

Vreven et al. were able to find a separating line of performance of the docking algorithms that they tested based on interface area versus experimental binding energy of the complexes, with 79% success on the side of greater interface area in combination with greater binding energy, whereas each individual feature was only weakly predictive of success. They used a much looser criteria for success in their assessment compared to us (theirs: native-like in top 100, ours: top 10); nevertheless, it suggests that binding energy (Kd), which is easier to obtain experimentally than the structure, could be used in a

filter to select complexes for which we can be more confident in our predictions. The training and test sets have about the same median interface size and lower bounds (Fig 2.8c), so performance gains from a rearrangement in the train-test split and k-fold cross-validation may not be as influenced by interface size.

Next we tested whether we can find a linear combination of features that separate the performance classes. Accordingly, we combined the interface pair interaction type features and the interface size feature of Fig 2.8 for the entire set of complexes and performed Linear Discriminant Analysis (LDA) (Fig 2.9a). We used the labels of good (Native-like in Top 10) and poor *Upside* performance as the classes to separate in this analysis. LDA returns an output one dimension less than the number of labels/classes. However, this one-dimensional treatment does a poor job of separating our performance classes. So we increased the number of labels with information about whether the complex belonged to the training or test set for a total of four labels ([train, test] × [good, poor]), since earlier we noticed that some differences in the distribution of features correspond with differences in their distribution in the training and test sets (Fig 2.9b). However, even this three-dimensional LDA lacks a strong separating surface between the performance classes. Considering that *Upside*'s energy function contains non-linear terms (e.g., the environment energy) and details such as distance and orientation between sidechain beads, classification may not be feasible with a linear method and with the simple features that we have chosen.

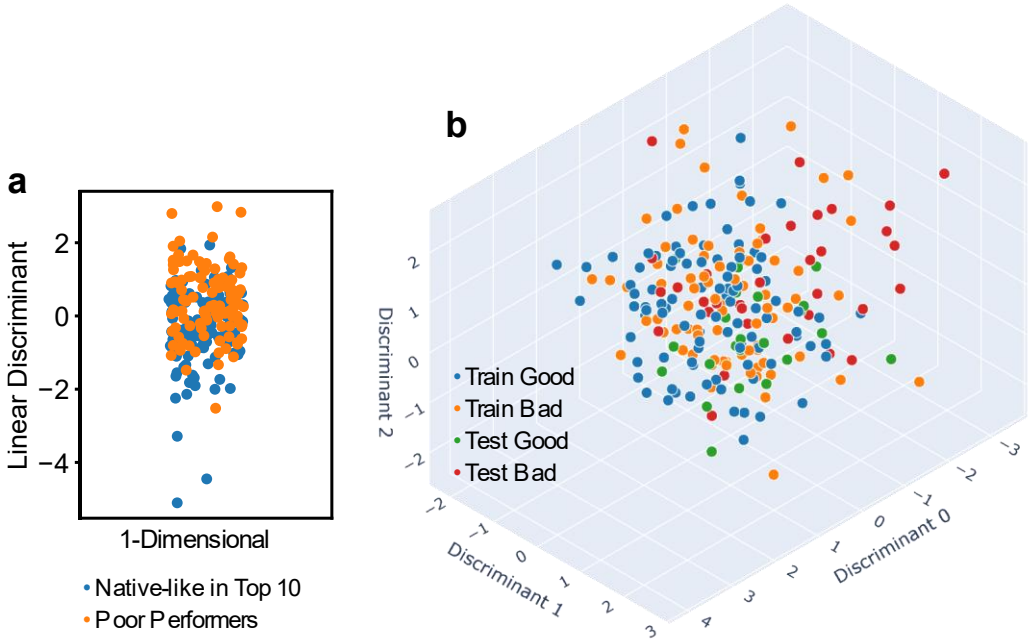

Figure 2.9: **LDA of interface pair interaction features and interface size for the entire set of complexes.** There are five types of pair interactions, which along with the interface size gives six total features. a) One-dimensional LDA using two performance labels (Native-like in Top 10, Poor Performers). The spread in the x-axis is artificial jitter to aid in distinguishing points. b) Three-dimensional LDA using four labels (training set good performers, training set poor performers, testing set good performers, testing set poor performers).

2.3.7 Information driven antibody-antigen docking

In some cases, additional sources of information about a complex exist beyond the structure and sequence of the unbound forms of the binding partners. For example, the use of experimental information for challenging complexes is recognized by CAPRI organizers, and they have provided small angle X-ray scattering (SAXS) and cross linking/mass spectrometry (XL/MS) data in the Round 46 CAPRI-CASP experiment for one such complex [3]. For antibody-antigen docking, the CDR loops can be identified based on the sequence of the conserved protein framework around them, and one could use hydrogen exchange (HX) or mutational scanning experiments to glean information about the location of the epitope [29]. It is illustrative to examine the extent that limitations of docking algorithms can be overcome by the incorporation of this extra information.

Results for the information-driven antibody-antigen docking are presented in Fig 2.10 according to CAPRI criteria with the same ranking of native-like structures as used in Fig 2.3. We follow Ambrosetti et al.'s comparative study [19] that presents the performance of individual complexes as opposed to the

aggregate of all complexes, to obtain finer grained insights on the impact of both flexibility and auxiliary information on prediction accuracy.

Ambrosetti et al. found that their HADDOCK algorithm, which involves torsional and explicit solvent flexible refinement stages, does not perform well with biasing information solely from the antibody HV loops (HV-Surf), with several complexes lacking any native-like pose in the top 100 ranks (Fig 2.10a). However, the biasing of interactions between HV loops along with a coarse definition of epitope residues produces substantial gains (Fig. 2.10a, column heading HV – Epi 9). Likewise, *Upside* lacks native-like predictions for many complexes when just filtering for poses in contact with the Ab loops (Fig. 2.10b, HV Filtered). However, as with HADDOCK, *Upside* shows improvement when augmented with coarse epitope information (HV – Coarse Epi) for filtering to include poses where both loops and epitope are in contact and biasing the interactions between them during the *Upside* runs.

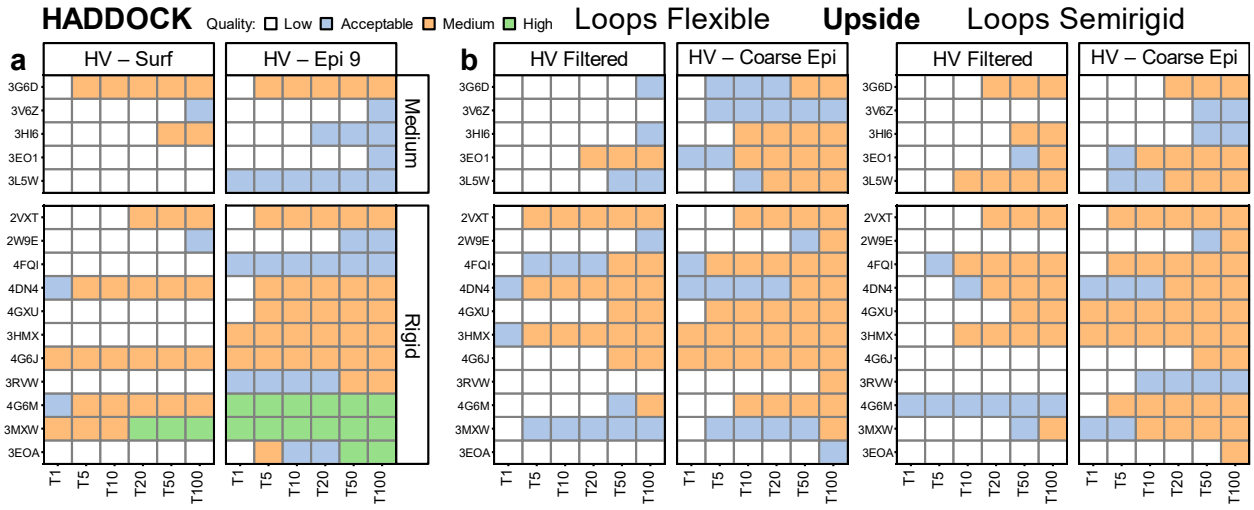

Figure 2.10: **Antibody-antigen information-driven docking predictions.** a) CAPRI criteria docking performance of HADDOCK using different levels of informational restraints (adapted from Ambrosetti et al., 2019). b) *Upside* results with either flexible Ab loops or loops held semi-rigid with spherical flat-bottom restraints. HV Filtered uses loop contact information to filter poses, but with no biasing potential during *Upside* runs to approximately correspond to HADDOCK's HV – Surf protocol. *Upside*'s HV – Coarse Epi protocol uses both loop and coarsely defined epitope information to filter poses, and pairwise sigmoidal contact potentials to bias interactions during *Upside* runs. *Upside*'s HV – Coarse Epi roughly corresponds to HADDOCK's HV – Epi 9. Both HADDOCK and *Upside* results use the unbound forms of the subunits as inputs for docking.

However, the inclusion of loop flexibility has mixed outcomes for *Upside*. First, a comparison of the HV Filtered runs between the flexible and semi-rigid loop cases finds that some complexes either worsen their ranking or quality of their poses when backbone flexibility is allowed (eg. 3G6D, 3HI6, 3LW5,

4FQI, 4G6M) although others gain in ranking or quality (e.g., 2VXT, 4DN4, 3HMX, 3MXW). To obtain more consistent results in the future when only loop information is known, a scheme that combines flexible and semi-rigid poses and ranking them together might help (untested). Only the inter-protein terms of the energy function can be considered for ranking in such a setup because the flexible poses will likely benefit from less backbone strain.

When epitope information is used (e.g., HV – Coarse Epi), flexibility generally produces better predictions (e.g., 3G6D, 3V6Z, 3HI6, 4G6J). The contact biasing potentials in this case compensate for inaccuracies in the folding and protein-protein energy terms. Ambrosetti et al. similarly observed that biasing with higher levels of information was required to help with packing the antibody H3 loops during their flexible refinement stage. Improvement is required in the underlying models and forcefields of both *Upside* and HADDOCK for unaided flexible docking.

This analysis indicates that *Upside* and HADDOCK perform similar but have different strengths, appreciating that their filtering and biasing schemes are not exactly the same. When epitope information is used and loops are flexible, *Upside* does better for some medium difficulty complexes (3V6Z, 3HI6, 3EO1). This improvement demonstrates the benefit of *Upside*'s greater flexibility. However, HADDOCK produces high-quality predictions for some of the more rigid complexes (4G6M, 3MXW, 3EOA), whereas *Upside* is unable to produce any high-quality structures as noted before with the general data set in Fig 2.3. This finding may again reflect the limitations of *Upside*'s coarse-graining of side chains.

As each CDR loop is known to favor certain clusters of canonical structures [30], we also compare our predicted cluster assignments of the loops to those of the native antigen bound structures. This analysis is based on the *Upside* structures from the Loops Flexible CAPRI evaluation scenario of the preceding discussion. The PyIgClassify server [30] was used to assign CDR loop structures to known clusters according to the loop's backbone dihedral angles.

Cluster assignments were done for all 16 antibody complexes, and with the four different scenarios for each complex using either: 1) the known bound forms of the loops; 2) the unbound forms of the loops, which were the starting structures of the *Upside* relaxation simulations; 3) best scoring native-like structure from *Upside* using only CDR loop information to assist the docking; 4) best scoring native-like structure from *Upside* using a bias between the loops and the coarse epitope to assist the docking.

Fig 2.11 summarizes CDR loop prediction accuracy, where the bars count the number of loops that have been assigned to a wrong, non-native cluster. The 3G6D complex, where no bars are visible, indicates that unbound and *Upside* structures for that complex have the same cluster assignments as the native conformations for all six CDRs. At the other end, the 3EO1 complex has three or more CDRs that are not predicted to be in the native cluster.

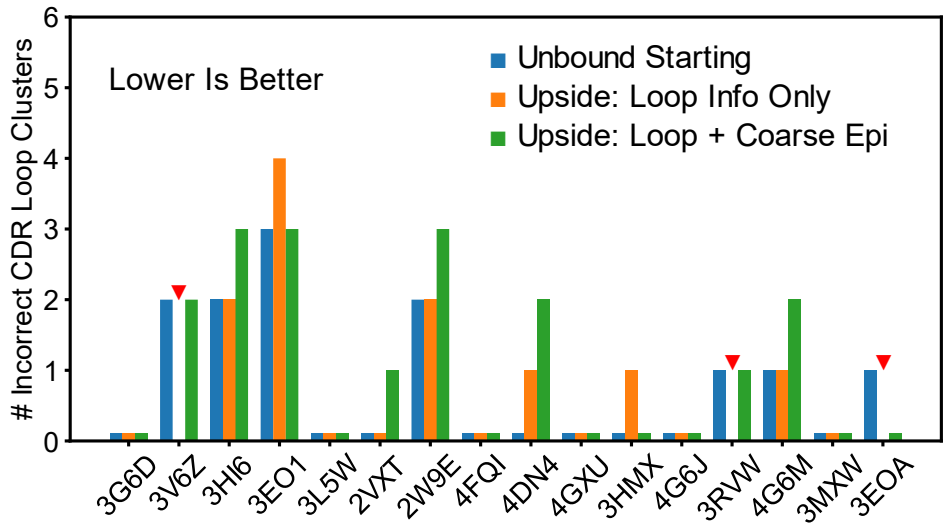

Figure 2.11: **Number of incorrect CDR loop cluster assignments compared to the native assignment by PyIgClassify.** The blue bars represent CDR loops in the unbound conformation, which are the starting points for the *Upside* relaxation simulations. The orange and green bars are from the best scoring native-like *Upside* prediction under the two different information conditions. The orange bars are for the loop information only case, where the antigen was simply filtered to be in contact with the CDR loops. The green bars are for the case with bias between the loops and the coarsely defined epitope. ▼ designate cases where structures are missing for the no bias case because no native pose scored within the top 100 poses.

More importantly, the loops in the *Upside* simulations largely remain in their original clusters irrespective of the information level used to assist the docking. Buried portions likely encounter steric hinderance and interactions that trap them in their original conformation. Literature suggests that the CDR loops (particularly H3) span a spectrum of flexibility, and even among our test set, examples exist of complexes with loops that undergo conformational change between unbound and bound states, so there is motivation to improve the conformational sampling of the more flexible regions [25, 26].

Fig S2.4 further illustrates this point with visualizations of the loop structures for the two complexes (3EO1, 3HI6) that have the largest discrepancy between our predicted loop structure cluster assignments and the native assignments. During the *Upside* simulations, the loops generally do not

change their backbone RMSD by more than ~3 Å from the starting structures. Although the internal loop conformations are retained, the loops still undergo center of mass shifts and tilts when referenced to the rest of the entire antibody (Fig S2.5) and so the binding surface changes and explains we can obtain better CAPRI predictions in the flexible loop with epitope bias scenario.

For comparison, Ambrosetti et al. investigated the RMSD of H3 loop of their predicted models after flexible refinement with alignment of the framework residues to the antigen bound structure (i.e., overall shifts in position and orientation were included in their RMSDs). The accuracy tended to get worse by up to ~1.5 Å for the easy complexes that already start with low H3 RMSD in the unbound forms. They saw an improvement by up to 1.25 Å in some of their predictions for medium difficulty complexes when coarse epitope information was used, but they also saw a degradation for some others. Overall, HADDOCK has mixed performance with CDR loop prediction.

Our temperature replica exchange molecular dynamics (TREMD) simulations of 3EO1 starting from its unbound conformation in the absence of antigen found that the H1 loop can sample lower RMSD states compared to those previously found during the docking (Fig 2.12). The previous docking simulations were done at a relatively low and constant temperature. This suggests that to sample the native conformation, we likely require enhanced sampling procedures.

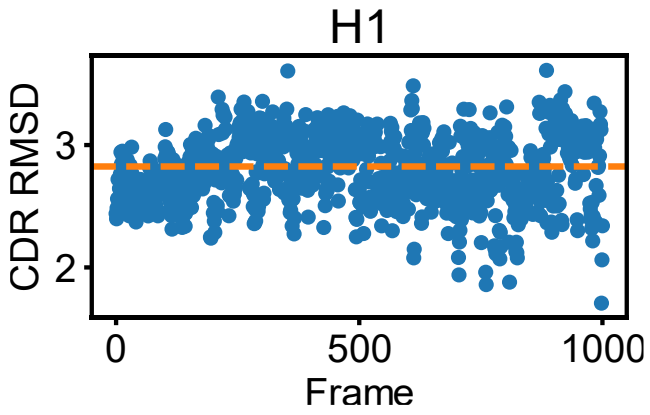

Figure 2.12: **H1 CDR loop RMSD (in Å) from TREMD of 3EO1 antibody separated from antigen.** -- indicates the RMSD from the best scoring model docking model used in the previous figures.

In conclusion, the docking algorithms, including *Upside*, do a relatively poor job of predicting antibody binding with CDR loop information only. Biasing interactions using a coarse definition of the epitope greatly improves the results, with *Upside*'s greater efficiency in flexible backbone sampling over HADDOCK enabling better predictions of medium difficulty complexes. While center of mass position and

orientation of the loops shift during this flexible sampling in the reference frame of the framework residues, and thus the binding surface changes, the internal structures of the loops do not change much over the course of *Upside* simulations in contact with antigen. Enhanced sampling or a conformational selection scheme (e.g., cross docking different pre-sampled unbound conformations) may be required to better predict the internal CDR loop structures.

2.4 Discussion

Ultimately, models representing protein molecules should capture all aspects of their behavior. However, depending on our immediate objectives and available computational resources, we must make approximations. Importantly, identifying which kinds of approximations perform well for a given problem often provides new insight on the fundamental nature of proteins. We explored these issues using *Upside*, a very fast MD algorithm designed to model protein dynamics in which the side chains are represented by multi-position beads. By any standards, *Upside* performs fairly well regarding the protein folding problem [22]. By adapting *Upside* to protein-protein docking, our hope was that its backbone sampling capabilities would spontaneously result in better prediction complexes that undergo conformational change upon binding and overcome the lack of explicit side chains. The reality turned out to be more complicated.

2.4.1 Limitations of model

We found that the addition of specific inter-protein energy terms considerably improved our ability to score native-like structures compared to the original folding forcefield. Combinations of pairwise polar and charge sidechain interactions and the burial of hydrophilic sidechains were underrepresented in the protein folding training set. Therefore, it is not surprising that the original forcefield designed for folding is not optimized for protein-protein docking. One important limitation appeared to originate from the coarse-grained representation of the side chains.

Even with the single bead representation of the side chains, *Upside*'s performance was comparable to traditional full sidechain methods when the subunit structures were restrained in their unbound conformations. Only in a few cases where the subunits did not change conformation by more than a few Å in the *Upside* simulations did the addition of sidechains with SCWRL4 [25] and rescoring

with SOAP-PP [21] produce a significant benefit. The CABS CG method was also able to find some low IRMSD predictions in the top 10 scored predictions [28], which when considered with our results indicates that moderate success in docking can be achieved without explicit side chains.

However, we must note that SCWRL4 and SOAP-PP do not give a perfect reconstruction and scoring of sidechains and it is still likely that explicit sidechains are required for high accuracy predictions. This is supported by the much higher success rate of FRODOCK [2] with bound conformations of the subunits and native full sidechain rotamers compared to *Upside* using the same structures as initial states and restraining the subunits to their bound conformations. The recent neural network approaches to docking also utilize explicit side chains and are successful in high accuracy predictions [13–16].

Surprisingly, inclusion of backbone flexibility exhibited both advantages and disadvantages regarding the protein docking problem. The accuracy of the approach generally decreased when full backbone flexibility was allowed even starting from the native pose. Generally, this outcome is due to the lowest energy structure, which is the product of both the folding and binding energy terms, was not within 3 Å of the bound subunit structures ("Both Subunits Free" panels in Fig. 2.4). The PMF-determined energy needed to shift the structures of the subunits to their bound conformations often was too large to overcome using the energy of the binding terms (e.g., >4 k_BT). This steep penalty for what would seem to be a relatively minor error prevented *Upside* from being successful in flexible docking. To improve our procedure would require better training and/or more sophisticated forcefield terms that contribute to folding as well as binding.

The decrease in performance with backbone flexibility is likely endemic to most current forcefield guided dynamics docking methods, as we found that rigid docking performs better than the large conformational search of the CABS CG method [28]. Even the extensively sampled, MD simulations of Pan et al. benefitted from imposing backbone constraints based on the bound conformation of the subunits to circumvent the structural degradation occurring over time during a simulation based on a classical all-atom MD force field [10]. In the specific case of antibody docking, biased simulations using information on antibody CDR loops and epitopes can overcome forcefield inaccuracies, and flexible docking with *Upside* produced better results than with backbone restraints and was comparable to HADDOCK [19].

For predicting antibody CDR loop conformations during docking, we noticed that conformational selection may play a role, as opposed to induced fit only, because the internal loop conformations did not change very much while in contact with the antigen due to high energy barriers. This is supported by the presence of more native-like loops in the antibody-only simulations. In this scenario, the natural fluctuations of the unbound antibody access the native bound conformation of the CDR loops, and this state binds the antigen. But attempting to improve the loop conformation after initial contact of the unbound forms of the antibody and antigen (as done in the *Upside* docking pipeline) may not succeed because of the aforementioned issue of steric hinderance experienced by the loops. The literature is divided on the relative weight of conformational selection and induced fit in antibody binding [27, 28].

2.4.2 Folding versus binding

Since *Upside* was designed for protein dynamics and folding rather than docking, it is worth discussing how the challenges faced by these two classes of problems may be different. Proteins seemingly have a huge number of possible conformations and yet many manage to fold within seconds [35]. As Rose notes, there are strong organizational constraints imposed by backbone sterics and hydrogen bonding [30, 31]. A 100-residue domain thus may have only ~10 helices and strands, which could be arranged in about 10^3 fundamental folds. Larger proteins consist of such domains so that conformational diversity may grow manageably with length.

To find the correct fold using simulations, the energy terms must be balanced and considerable searching is required. In practice, imperfections in force fields can readily result in kinetic trapping. For example, *Upside* is able to fold some proteins less than 100-residues, in part due to careful consideration of backbone potential terms and training against misfolded structures. But the method is not perfect since some misfolded states are stable and some native states are not the global energy minimum. This issue becomes much more problematic with longer sequences.

In contrast, the search problem is substantially simpler for rigid-body protein docking. According to Janin's model of barnase-barstar binding, about 70,000 poses are needed to find the native well if states are discretized every 14 degrees with respect to each of the angles about their center to center vector [38]. For larger complexes, the number of poses required likely is proportional to the square of the surface area, and the area grows as the mass with a fractional power of 0.7 [39], Hence, the number of

decoys grows manageably and it is possible to generate and score the 10^5-10^6 poses needed to sample docking space.

However, when docking the unbound structures of the subunits and presented with the docked set of 10^5-10^6 poses, the traditional approaches only have a success rate of finding a native-like pose in the top 10 predictions (i.e., ten predictions are needed for one to be native-like) for less than ~30% of the complexes. Improvement likely necessitates generating docking poses with subunits that have structures closer to their true bound conformations. Relaxation of the complexes with *Upside* results in 3-4 Å Cα RMSD for the subunits at best; this accuracy can be considered a good folding prediction, but it still produces only medium quality docking poses and the situation is not much improved. Generating high accuracy binding poses requires having subunit structures that are close to their bound structures, and effectively, the challenge of binding becomes a challenge in high accuracy structure refinement that incorporates elements of folding.

2.4.3 Recent machine learning methods

Our results that pointed to a possible conformational selection mechanism for antibody loops illustrates the utility of molecular dynamics tools to learn about pathways of protein binding and folding, as opposed to the recent neural network approaches for protein structure prediction. In AlphaFold 2 and RoseTTAFold, backbone sterics, especially of the peptide bond, are not explicitly represented in the primary stages of the models such that when coupled to distance information of specific residue pairs, the models are able to unnaturally search for the optimal positions of residues over long distances while the backbone “clips” or “ghosts” through itself [10, 11]. So, these neural network methods would not be able to differentiate between induced fit and conformational selection binding mechanisms. With molecular dynamics simulations, backbone sterics and non-specific interactions hinder the search, but provide an avenue to predict thermodynamics, kinetics, and pathways.

Better antibody loop predictions will require either fully flexible free docking or rigid-body crossdocking of antibodies with different pre-sampled loop conformations followed by refinement. Flexible docking thus encompasses the difficulties of protein folding and adds to it. Indeed, even AlphaFold 2, that although earned the title of solving the protein folding problem during the CASP 14 experiment, still had trouble with a multidomain protein whose domain associations were like that of protein complexes [40]. Further

advancements are required for de novo flexible docking, in the case of *Upside* this may entail joint training of folding and inter-protein energy terms.

2.5 Acknowledgements

This study was supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada (N.F.), and by the National Science Foundation (NSF) through grant MCB-1517221 (B.R.)

2.6 Supporting Information

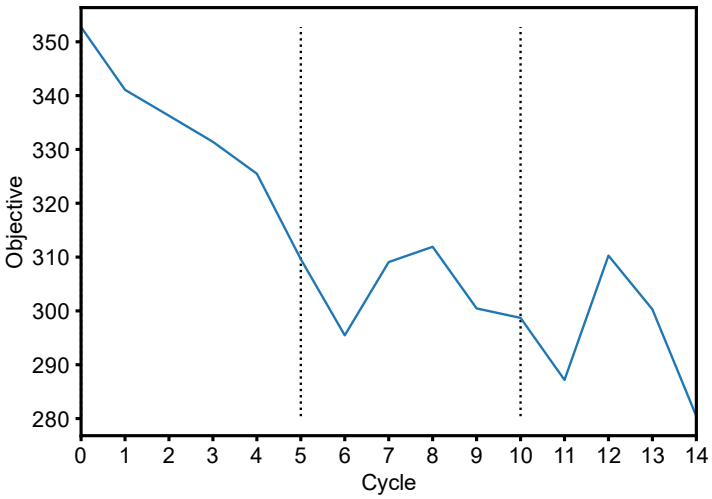

Figure S2.1: **Objective function during training**

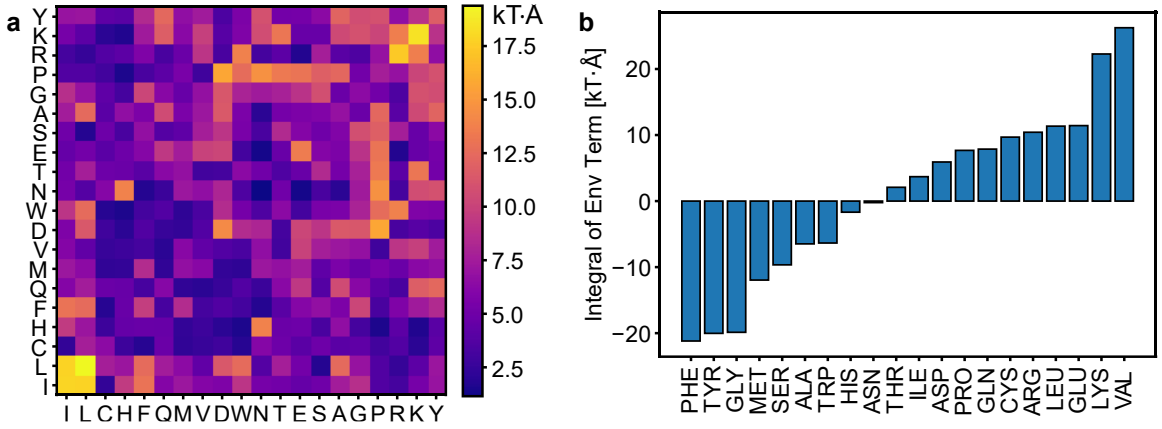

Figure S2.2: **Magnitude of changes to potentials with the new docking FF.** a) Integral of the absolute value of the radial part of the new interprotein pairwise sidechain potential. b) Integral of the new interprotein environment potentials up to 10 Å.

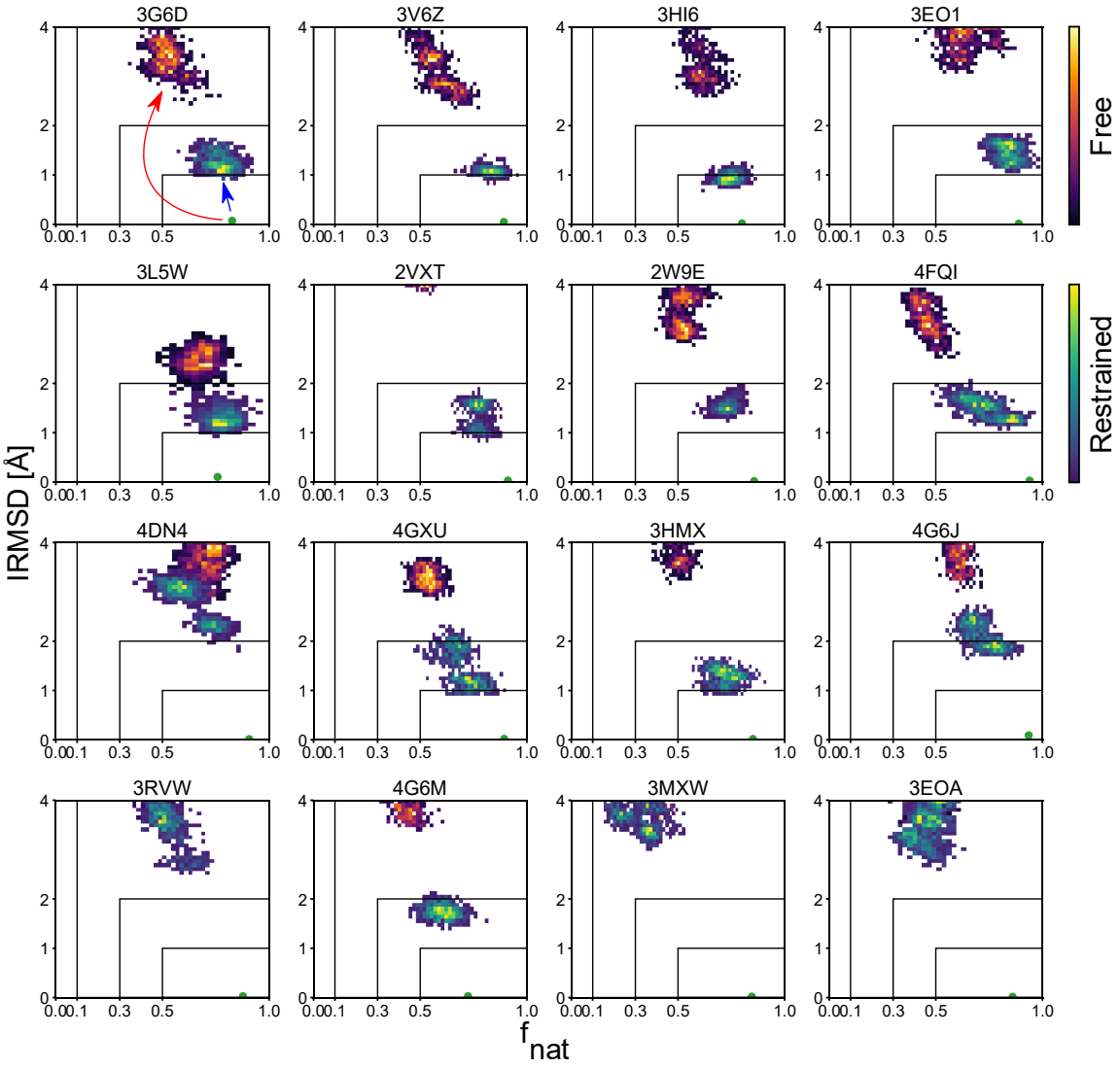

Figure S2.3: **2D probability distributions from native state simulations.** The green dots are the starting native states in *Upside*, which are away from the exact native position in bottom right corners (IRMSD = 0 Å, f_{nat} = 1.0) due to the coarse-graining. The red distributions are simulations run without backbone restraints, while the green ones are with backbone restraints. Each condition is run in triplicate and the second halves of the trajectories concatenated to produce the distributions.

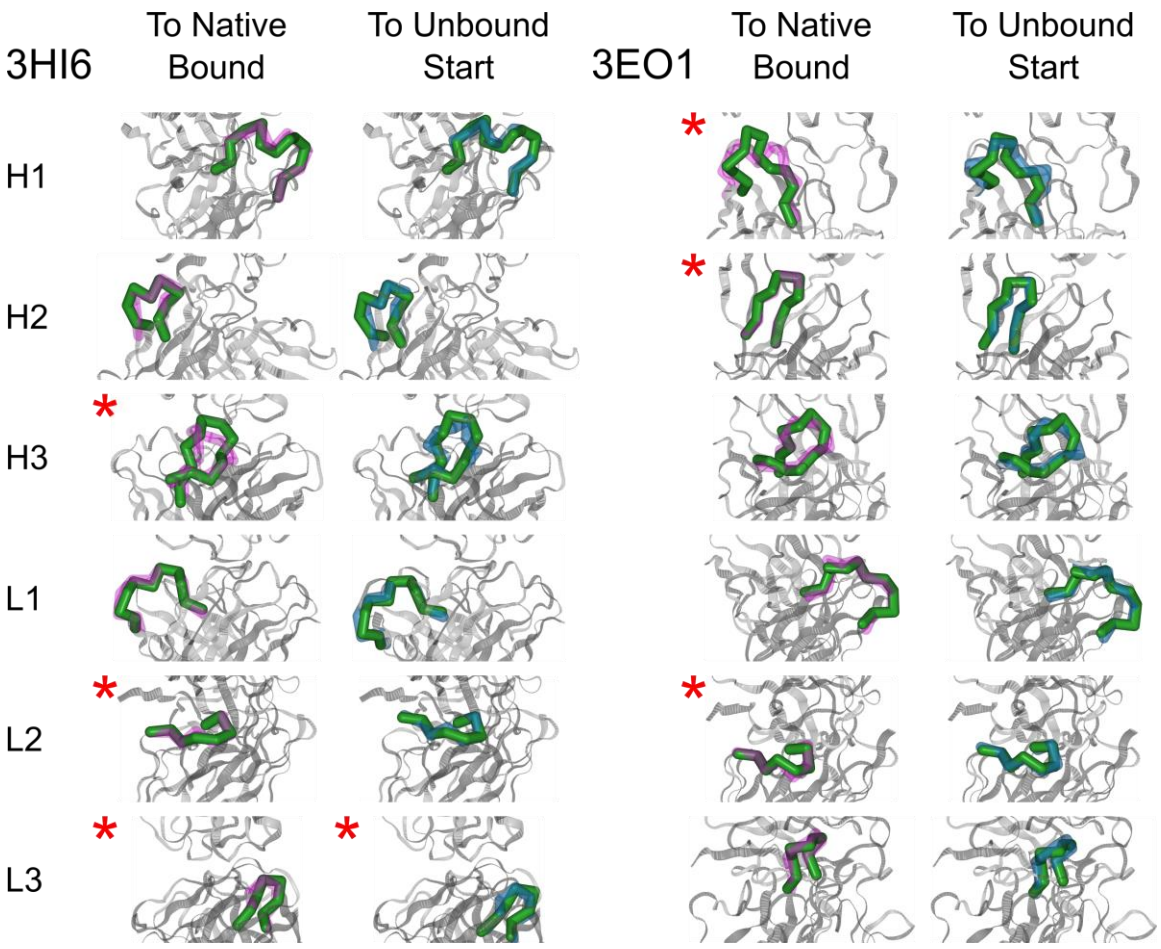

Figure S2.4: **Visualization of CDR loops.** The green segment is from the best ranked native-like *Upside* prediction using epitope biasing superimposed onto the native antigen bound loop structure (magenta) or unbound structure (blue). The rest of the complex is shown in grey in the native bound form. The red stars denote mismatch between the PyIgClassify cluster classifications between the *Upside* prediction and the reference loop structure.

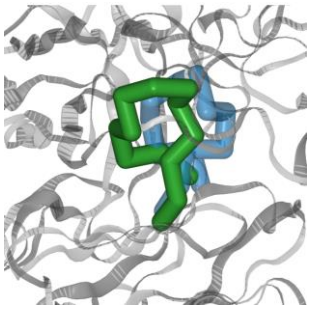

Figure S2.5: **Structural alignment of the entire antibody.** Reveals center of mass shift and tilt of the 3EO1 H3 loop after *Upside* relaxation (green) from the unbound starting structure (blue), although there is not much difference in the loop conformations in terms of dihedrals as shown in Fig Figure S2.4.

2.7 References

[1] Porter KA, Desta I, Kozakov D, Vajda S. What Method to Use for Protein-Protein Docking? Curr Opin Struct Biol. 2019 Apr;55:1–7.

[2] Ramírez-Aportela E, López-Blanco JR, Chacón P. FRODOCK 2.0: fast protein–protein docking server. Bioinformatics. 2016 Aug 1;32(15):2386–8.

[3] Lensink MF, Brysbaert G, Nadzirin N, Velankar S, Chaleil RAG, Gerguri T, et al. Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. Proteins Struct Funct Bioinforma. 2019;87(12):1200–21.

[4] Dapkūnas J, Olechnovič K, Venclovas Č. Structural modeling of protein complexes: Current capabilities and challenges. Proteins Struct Funct Bioinforma. 2019;87(12):1222–32.

[5] Lensink MF, Velankar S, Wodak SJ. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. Proteins Struct Funct Bioinforma. 2017 Mar 1;85(3):359–77.

[6] Quignot C, Rey J, Yu J, Tufféry P, Guerois R, Andreani J. InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. Nucleic Acids Res. 2018 Jul 2;46(Web Server issue):W408–16.

[7] Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, et al. Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. J Mol Biol. 2015 Sep 25;427(19):3031–41.

[8] Moal IH, Bates PA. SwarmDock and the Use of Normal Modes in Protein-Protein Docking. Int J Mol Sci. 2010 Sep 28;11(10):3623–48.

[9] van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol. 2016 Feb 22;428(4):720–5.

[10] Pan AC, Jacobson D, Yatsenko K, Sritharan D, Weinreich TM, Shaw DE. Atomic-level characterization of protein–protein association. Proc Natl Acad Sci. 2019 Mar 5;116(10):4244–9.

[11] Kolinski A. Protein modeling and structure prediction with a reduced representation. Acta Biochim Pol. 2004;51(2):349–71.

[12] Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid AE, Kolinski A. Coarse-Grained Protein Models and Their Applications. Chem Rev. 2016 Jul 27;116(14):7898–936.

[13] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Jul 15;1–11.

[14] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science [Internet]. 2021 Jul 15 [cited 2021 Jul 28]; Available from: http://science-sciencemag-org/content/early/2021/07/19/science.abj8754

[15] Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2 and extended multiple-sequence alignments. BioRxiv [Preprint]. 2021 bioRxiv 460468 [posted 2021 Sep 15; cited 2021 Sep 20]. Available from: https://www.biorxiv.org/content/10.1101/2021.09.15.460468v1.

[16] Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer [Internet]. BioRxiv [Preprint]. 2021 bioRxiv 463034 [posted 2021 Oct 04; cited 2021 Oct 10]. Available from: https://www.biorxiv.org/content/10.1101/2021.10.04.463034v1.

[17] Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, et al. CAPRI: A Critical Assessment of PRedicted Interactions. Proteins Struct Funct Bioinforma. 2003 Jul 1;52(1):2–9.

[18] Lensink MF, Nadzirin N, Velankar S, Wodak SJ. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. Proteins. 2020 Aug;88(8):916–38.

[19] Ambrosetti F, Jiménez-García B, Roel-Touris J, Bonvin AMJJ. Modeling Antibody-Antigen Complexes by Information-Driven Docking. Structure. 2020 Jan 7;28(1):119-129.e2.

[20] Jumper JM, Faruk NF, Freed KF, Sosnick TR. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. PLOS Comput Biol. 2018 Dec 27;14(12):e1006342.

[21] Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, Sali A. Optimized atomic statistical potentials: assessment of protein interfaces and loops. Bioinforma Oxf Engl. 2013 Dec 15;29(24):3158–66.

[22] Jumper JM, Faruk NF, Freed KF, Sosnick TR. Trajectory-based training enables protein simulations with accurate folding and Boltzmann ensembles in cpu-hours. PLOS Comput Biol. 2018 Dec 27;14(12):e1006578.

[23] Skinner JJ, Wang S, Lee J, Ong C, Sommese R, Sivaramakrishnan S, et al. Conserved salt-bridge competition triggered by phosphorylation regulates the protein interactome. Proc Natl Acad Sci. 2017 Dec 19;114(51):13453–8.

[24] Gaffney KA, Guo R, Bridges MD, Chen D, Muhammednazaar S, Kim M, et al. Lipid Bilayer Induces Contraction of the Denatured State Ensemble of a Helical-Bundle Membrane Protein. BioRxiv [Preprint]. 2021 bioRxiv 444377. [posted 2021 May 17; cited 2021 Oct 10]. Available from: https://www.biorxiv.org/content/10.1101/2021.05.17.444377v1.

[25] Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009 Dec;77(4):778–95.

[26] Lensink MF, Méndez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins Struct Funct Bioinforma. 2007 Dec 1;69(4):704–18.

[27] Torchala M, Gerguri T, Chaleil RAG, Gordon P, Russell F, Keshani M, et al. Enhanced sampling of protein conformational states for dynamic cross-docking within the protein-protein docking server SwarmDock. Proteins Struct Funct Bioinforma. 2020;88(8):962–72.

[28] Kurcinski M, Kmiecik S, Zalewski M, Kolinski A. Protein-Protein Docking with Large-Scale Backbone Flexibility Using Coarse-Grained Monte-Carlo Simulations. Int J Mol Sci. 2021 Jul 8;22(14):7341.

[29] Paterson Y, Englander SW, Roder H. An Antibody Binding Site on Cytochrome c Defined by Hydrogen Exchange and Two-Dimensional NMR. Science. 1990 Aug 17;249(4970):755–9.

[30] Adolf-Bryfogle J, Xu Q, North B, Lehmann A, Dunbrack RL Jr. PyIgClassify: a database of antibody CDR structural classifications. Nucleic Acids Res. 2015 Jan 28;43(D1):D432–8.

[31] Krishnan L, Sahni G, Kaur KJ, Salunke DM. Role of Antibody Paratope Conformational Flexibility in the Manifestation of Molecular Mimicry. Biophys J. 2008 Feb 15;94(4):1367–76.

[32] Jeliazkov JR, Sljoka A, Kuroda D, Tsuchimura N, Katoh N, Tsumoto K, et al. Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification. Front Immunol. 2018 Mar 2;9:413.

[33] Keskin O. Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: a case study of antibodies. BMC Struct Biol. 2007 May 17;7(1):31.

[34] Wang W, Ye W, Yu Q, Jiang C, Zhang J, Luo R, et al. Conformational selection and induced fit in specific antibody and antigen recognition: SPE7 as a case study. J Phys Chem B. 2013 May 2;117(17):4912–23.

[35] Dill KA, Ozkan SB, Shell MS, Weikl TR. The Protein Folding Problem. Annu Rev Biophys. 2008 Jun 9;37:289–316.

[36] Rose GD, Fleming PJ, Banavar JR, Maritan A. A backbone-based theory of protein folding. Proc Natl Acad Sci. 2006 Nov 7;103(45):16623–33.

[37] Rose GD. Protein folding - seeing is deceiving. Protein Science. 2021;30(8):1606–16.

[38] Janin J. The kinetics of protein-protein recognition. Proteins Struct Funct Bioinforma. 1997 Jun 1;28(2):153–61.

[39] Glyakina AV, Bogatyreva NS, Galzitskaya OV. Accessible Surfaces of Beta Proteins Increase with Increasing Protein Molecular Mass More Rapidly than Those of Other Proteins. PLOS ONE. 2011 Dec 1;6(12):e28464.

[40] Service RF. ‘The game has changed.’ AI triumphs at solving protein structures [Internet]. Science. 2020 [cited 2021 Sep 14]. Available from: https://www.science.org/content/article/game-has-changed-ai-triumphs-solving-protein-structures

CHAPTER 3

PROTEIN FOLDING

3.1 Attribution and contributions

This chapter has been reproduced almost in full from

Jumper JM, Faruk NF, Freed KF, Sosnick TR. Trajectory-based training enables protein simulations with accurate folding and Boltzmann ensembles in cpu-hours. PLOS Comput Biol. 2018 Dec 27;14(12):e1006578.

in accordance with its Creative Commons Attribution License (CC BY 4.0). I was credited with “formal analysis, Investigation” for the publication, with major contributions of running many of the protein folding simulations, including the CASP targets, and conducting much of the analysis in the sections “3.5.2 Accuracy of structure prediction” and “3.5.3 Comparison with other physics-based approaches”.

3.2 Chapter abstract

An ongoing challenge in protein chemistry is to identify the underlying interaction energies that capture protein dynamics. The traditional trade-off in biomolecular simulation between accuracy and computational efficiency is predicated on the assumption that detailed force fields are typically well-parameterized, obtaining a significant fraction of possible accuracy. We re-examine this trade-off in the more realistic regime in which parameterization is a greater source of error than the level of detail in the force field. To address parameterization of coarse-grained force fields, we use the contrastive divergence technique from machine learning to train from simulations of 450 proteins. In our procedure, the computational efficiency of the model enables high accuracy through the precise tuning of the Boltzmann ensemble. This method is applied to our recently developed *Upside* model, where the free energy for side chains is rapidly calculated at every time-step, allowing for a smooth energy landscape without steric rattling of the side chains. After this contrastive divergence training, the model is able to de novo fold proteins up to 100 residues on a single core in days. This improved *Upside* model provides a starting point both for investigation of folding dynamics and as an inexpensive Bayesian prior for protein physics that can be integrated with additional experimental or bioinformatic data.

3.3 Introduction

Since Anfinsen's original demonstration that a protein's sequence determines its structure, multiple computational strategies have been developed to predict a protein's structure from its sequence. An additional facet of this challenge is to replicate the energy landscape that defines both the folding process and other dynamical properties. In the absence of other information, coarse-grained models with one or a few beads per residue are too simplistic for de novo structure prediction. C_β level models having authentic protein backbones with ϕ/ψ dihedral angles, but lacking side chain rotamers, have achieved some success [1–3]. Within the last decade, all-atom, explicit solvent methods have become successful for the folding of some small proteins, although the ability to replicate the properties outside the native basin requires substantial improvement [4]. For the folding process, it is unclear which representation provides the optimal combination of detail and computational expense to replicate protein folding and dynamics. Integral to the choice of representation is which interactions to include, such as hydrogen bonding, van der Waals interactions and hydrophobic burial.

Another factor is the parameterization of the energy function with the training algorithm needing to balance the influences of all interactions. Protein thermodynamics reflects a delicate balance between the free energy of the folded and unfolded states. If one interaction is slightly too large, the entire landscape can be severely distorted. For example, if backbone hydrogen bonding energies are too large compared to backbone-solvent interactions (which includes hydrogen bonds between the backbone and water), an excess of hydrogen bonding ensues and pathways become dominated by unrealistically stable native- and non-native secondary structures. In an extreme situation, the lowest energy structure may have long helices involving nearly all residues.

The balancing of these various energies has been a major effort, and the balance is continually being adjusted as new force field biases are identified [5]. However, the adjustment of some parameters to correct one deficiency can inadvertently degrade performance of other quantities. In order to achieve the correct balance, all terms in the model should be trained together, rather than adjusted with an ad hoc procedure to correct each identified deficit.

To achieve this balance with a detailed interaction model, we use our recently developed, extremely rapid *Upside* implicit solvent molecular dynamics program [6]. Each residue In *Upside* is

represented with a polypeptide backbone and a side chain interaction site or bead which can adopt up to 6 positions representing up to six different side chain χ_1/χ_2 states. The key advance of the model is the smoothing of the energy surface by approximate analytic integration of free energies for the side chains' discrete states. When trained to predict side chain conformations from the Protein Data Bank (PDB), the method can fold a few small proteins with moderate accuracy in a cpu core-day. The majority of speedup of the procedure is a result of a unique side chain algorithm which directly calculates the side chain probability distribution and the free energy. This free energy calculation, performed at every time step, avoids the steric rattling of the side chains which can occur in the condensed phase in all-atom simulations, and so allows the backbone to move on a smoother energy landscape.

Here, we demonstrate that we can achieve *de novo* folding for a diverse collection of proteins by combining our fast-equilibrating *Upside* model with a contrastive divergence procedure that optimizes the stability of the native well. We demonstrate that gradient descent on energy terms using only data from sampled trajectories is sufficient to parameterize a protein model with tens of thousands of parameters. The resulting parameters are sufficiently balanced and accurate to achieve reversible folding for many proteins in our validation set. In addition, the resulting model is an excellent starting point for large scale protein simulations using more detailed models as well as the integration of large quantities of external information (such as predictions of residue contacts).

3.4 Methods

3.4.1 Coarse-grained model

In our recently-developed *Upside* model, only the N, C_α, and C atoms for each residue undergo dynamics. This simple representation of the protein allows for molecular dynamics on a smooth landscape but also makes it challenging to include the entirety of the protein physics. To address this challenge, we build additional layers of derived coordinates during the energy computation, much like virtual sites in a traditional force field. These layers include amide hydrogens, carbonyl oxygens, hydrogen bonding and residue burial scores, and the possible locations of protein side chains. All of the derivative information required is backpropagated through these layers of representation during the computation of forces for molecular dynamics. The side chain positions are the most challenging to represent because we must

solve a side chain packing problem in order to determine the distribution of side chain positions for a given backbone geometry. To pack the side chains probabilistically and obtain a side chain free energy, we use a rapid self-consistent iteration as described in our recent work [6] (Fig 3.1). The major computational steps are:

Step 1. The loop begins (upper left corner) with each residue in the protein being represented with 3 backbone atoms, the N, Cα and C. Based on the position of these atoms, the carbonyl oxygen, O, and amide proton, H, are deterministically placed.

Step 2. Each side chain, represented by a single oriented bead, is assigned an initial probability for being in 1–6 states, depending on the residue type and the average frequency observed in the PDB. The state of the bead is defined by its position and an orientation, (x,y,z,v), where v is a unit vector, relative to the peptide plane.

Step 3. The pair-wise state probabilities of all side chains are simultaneously and rapidly calculated using belief propagation to produce the lowest system free energy.

Step 4. Forces on the 3 backbone atoms, as well as on the O, H and side chain beads are calculated from the derivative of the free energy.

Step 5. Forces on the O, H and bead are “pulled back” and added to the forces on the 3 backbone atoms by reversing the placement process.

Step 6. Langevin dynamics (implicit solvent with friction) are run on the 3 backbone atoms using the forces calculated in Steps 4 and 5.

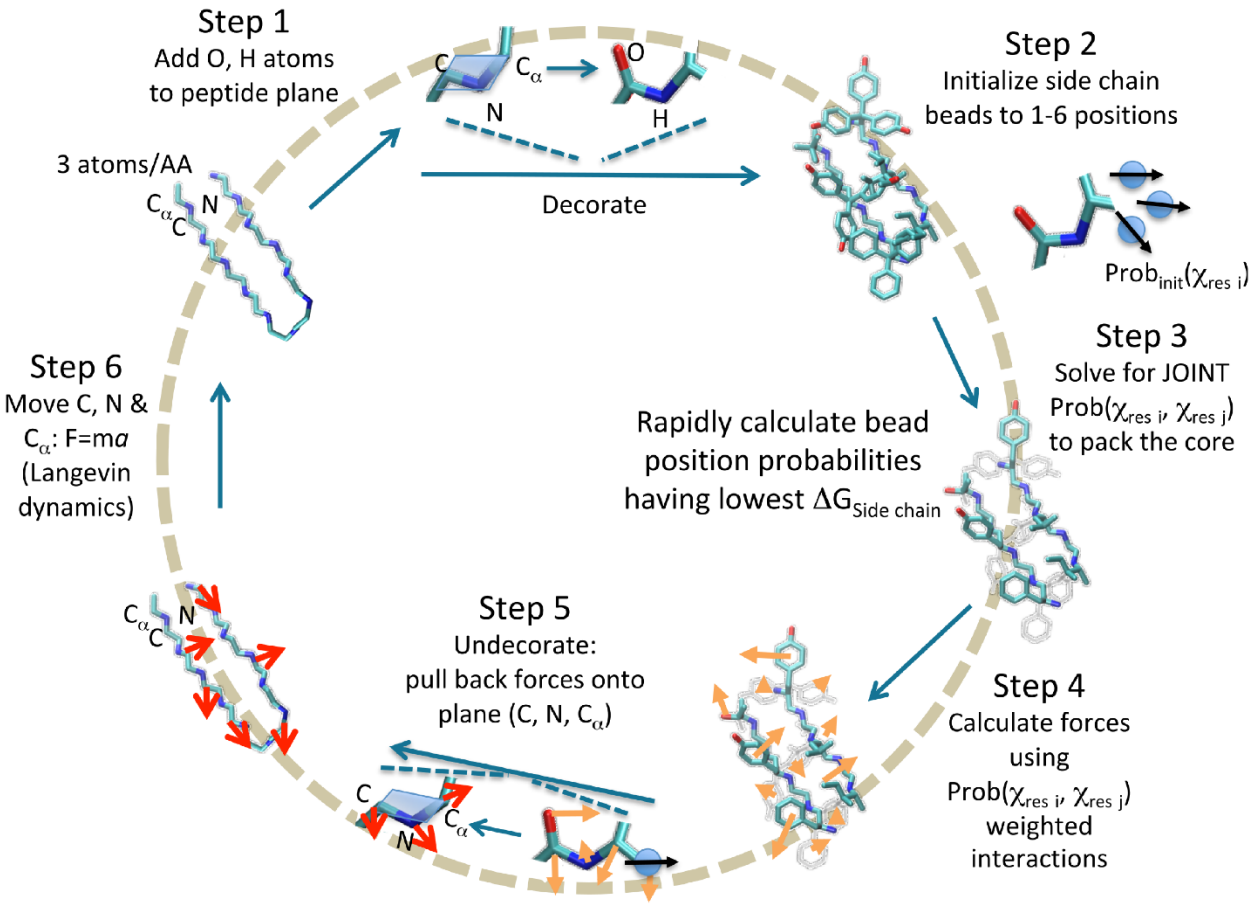

Figure 3.1: **Computational inner loop for *Upside*.** The positions of the protein side chains are added during each energy or force computation, then an approximate Boltzmann distribution is estimated for the side chains, and the free energy of the side chains is computed using the approximate Boltzmann ensemble. The resulting energy derivatives are pulled back to the backbone coordinates to update the backbone momenta.

The majority of parameters in *Upside* define the pairwise interactions between side chains, where each side chain is represented by a single directional bead. Concretely, each interaction pair is described by bead positions y_1 and y_2 and their orientations n_1 and n_2. From the distance $r_{12} = |y_1 - y_2|$ and displacement unit vector $n_{12} = (y_1 - y_2)/r_{12}$ are calculated. All of the pairwise interactions have the functional form

$$V = \kappa\left(V_{\text{radial}}(r_{12}) + \text{ang}_1(-n_1 \cdot n_{12})\text{ang}_2(n_2 \cdot n_{12})V_{\text{angular}}(r_{12})\right), \qquad 3.1$$

where V_{radial}, ang_1, ang_2, and V_{angular} are smooth curves represented by cubic splines for increased flexibility, rather than fixed functional forms such as a van der Waals 6-12 potential. The potential for each of the $\binom{20}{2} + 20 = 210$ types of amino acid pairs are described with 62 spline coefficients per pair, giving

13020 parameters. There are also five interaction sites on the backbone, roughly representing the H, O, N, C_α, and C atoms, with 54 parameters per interaction due to a smaller cutoff distance (10 versus 8 Å). The total number of side chain-backbone interaction parameters is 5400.

We add an additional term to capture desolvation effects by computing the approximate number of side chains N_i within a hemisphere above the C_β (see S1 Text in Supporting Information of Jumper et al.). High values of Ni correspond to buried residues. The total energy is

$$V_{\text{env}} = \sum_i V_{a_i}^{\text{env}}(N_i), \tag{3.2}$$

which is the sum of the values from individual $V_{a_i}^{\text{env}}$ potential curves for each residue i. Although more sophisticated solvation potentials exist, our implementation is very fast and easily optimized by the contrastive divergence procedure, while remaining flexible enough to represent many of the solvation effects omitted by the pairwise side chain potential.

The backbone dihedral angle Ramachandran potential is $\sum_i V_i^{\text{Rama}}(\phi_i, \psi_i)$, where V_i^{Rama} depends on the chemical identity of the $i-1$, i, and $i+1$ residues. The Ramachandran potentials are based on the turn, coil, or bridge (TCB) Ramachandran probability models in the NDRD backbone library [7]. We introduce a single parameter controlling extra stabilization of angles consistent with β-sheet geometries to allow training to counteract an observed tendency for our model to overstabilize helices. The backbone non-bonded interactions are governed by a distance- and angle-dependent hydrogen bonding potential whose magnitude (but not geometry) is chosen by contrastive divergence. The backbone N, C_α, C_β and C feel a steric repulsive interaction when their internuclear distance is approximately 3.0 Å.

Source code for *Upside* can be obtained from https://github.com/sosnicklab/*Upside*-md, and the results of this paper can be reproduced using the version tagged `trajectory_training_paper`.

3.4.2 Contrastive divergence

Our implementation of contrastive divergence considers two ensembles, one closely restrained to the native (crystal) structure and another that is free to diffuse away during simulations (Fig 3.2). In a perfect model, an unrestrained ensemble would remain close to the native structure. For an inexact model, differences arise, such as an excess of backbone-backbone hydrogen bonding in the free ensemble. Reducing the hydrogen bond energy would shift the free ensemble closer to the native ensemble. The

parameter modification must be small, however, because shifting the hydrogen bond energy may adversely affect other features of the ensemble, e.g., by reducing the burial of hydrophobic residues. Accordingly, after simulations are run on the first set or “minibatch” of 12 proteins in our 456 protein training set, we modify all the parameters with small updates to shift the simulation ensemble to better match the native-restrained ensemble. Simulations are repeated on the next of the 38 subsets of 12 proteins, and the parameters are updated again. The algorithm is converged when no parameter can be altered to shift the free ensemble closer to the native-restrained ensemble.

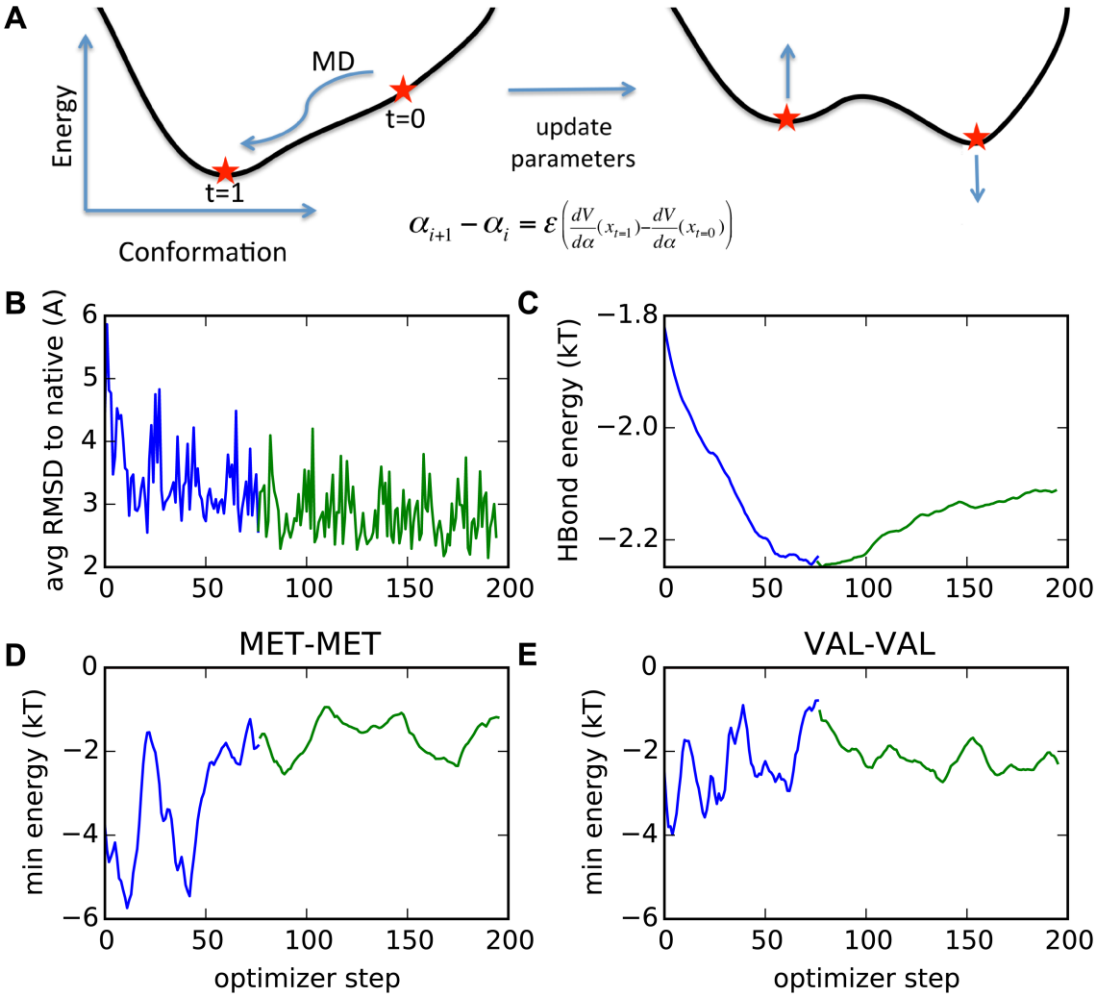

Figure 3.2: **Contrastive divergence training.** (A) Schematic of the training procedure depicting how the native state is stabilized relative to other states upon parameter updates. (B)-(E) In all plots, the blue curves indicate larger initial step-size training and the green plots indicate smaller step-size (fine-tuning). (B) The upper left plot shows the decline in minibatch-averaged RMSD over the course of the optimization. The remaining plots show (C) the convergence of the hydrogen bonding and side chain-side chain interaction parameters over the optimization for (D) Met-Met and (E) Val-Val potential. The larger step-size optimization of the side chain parameters exhibits large oscillations that inhibit convergence.

The free ensemble is generated using 5000 time units of dynamics (approximately 10 wall-clock minutes), with the first half being discarded as equilibration. Unless the native state is particularly unstable, this time is insufficient for exploration of the conformational landscape much beyond the native basin (RMSD within 6 Å) and so produces only a locally-equilibrated ensemble.

The native ensemble is traditionally defined as a single conformation. This δ-function distribution is problematic for proteins because they are dynamical molecules. Additionally, the solution ensemble may differ from the crystal structure for multiple reasons, including crystallographic packing. To reduce the impact of these issues, we replace the exact ensemble structures with the ensemble restrained to be near the crystal structure, within approximately 1 Å C_α-RMSD. This procedure is analogous to the restrained equilibration of crystal structures required to prepare systems for all-atom molecular dynamics. To account for changing parameters, we apply the restrained relaxation at every optimizer step.

After generation of the free and native-restrained ensembles, we change the energy parameters α_i, where i is the optimizer step, in proportion to the amount that the change can differentiate the two ensembles. This procedure is a form of gradient descent to reduce the "distance" between the free and native-restrained ensembles,

$$\alpha_{i+1} = \alpha_i + \frac{\epsilon}{M}\sum_{a=1}^{M}\left(\left\langle\frac{dV}{dx_i}\right\rangle_{\text{restrained}} - \left\langle\frac{dV}{dx_i}\right\rangle_{\text{free}}\right), \tag{3.3}$$

where ϵ is the step size, M is the number of proteins, and a indexes the simulated proteins. The quantity $\left\langle\frac{dV}{dx_i}\right\rangle_{\text{restrained}} - \left\langle\frac{dV}{dx_i}\right\rangle_{\text{free}}$ represents a pseudo-derivative of the free energy of restraining the simulation to be near the crystal structure (see SI for details). In the limit that the simulation duration is infinite, this difference is the exact derivative of the free energy. In practice, this difference chooses a suitable direction to improve the parameters.

The simulations use temperature replica exchange with eight replicas to enhance barrier crossing [8], while the temperature intervals of the replicas scale with to encourage efficient replica exchange for proteins of various sizes. The progress of the replica exchange is monitored by the average RMSD-to-crystal structure over the simulation for each "minibatch", the 12 protein subset used for a single gradient-descent step.

3.5 Results

3.5.1 Training

The parameters are initially set to those used to optimize side chain (χ1) accuracy [6]. The contrastive divergence training rapidly improves the model's average RMSD over a minibatch from 6 Å to 3 Å. This decline is accompanied by rapid change in the parameters. To reduce parameter fluctuations and fine-tune the results, we reduce the optimizer step size by a factor of four after two full passes through the 38 minibatches.

Although the slope has greatly decreased of RMSD change with respect to the number of steps over the iterations, there are indications that the parameters have not yet converged. Earlier tests, however, showed that continuing the contrastive divergence until convergence does not necessarily produce better results, as has been previously observed [9]. When large barriers surround the native states, minimal relaxation of the conformation occurs, which in turn provides little new information, and further fine-tuning may even *reduce* the accuracy of the model. Potentially the decreased exploration in the native well in the later stages overtrains the model to distinguish between native and near-native structure at the expense distinguishing against a more diverse ensemble. Early termination of optimization has been observed to favor simpler models [10].

The hydrogen bond strength unexpectedly appears to converge to a significantly smaller value during the late, fine-tuning stage than during the early phase with larger optimizer steps. We speculate that the extra noise in the side chain interactions during the larger optimizer steps may in aggregate cause stronger side chain interactions for the protein. This effect would necessitate a large hydrogen bond energy to balance the increase in side chain interactions. The final pair-wise energy functions between the side chain beads and either the backbone carbonyl oxygen or the amide proton, and the bead-bead interactions are shown in Fig 3.3.

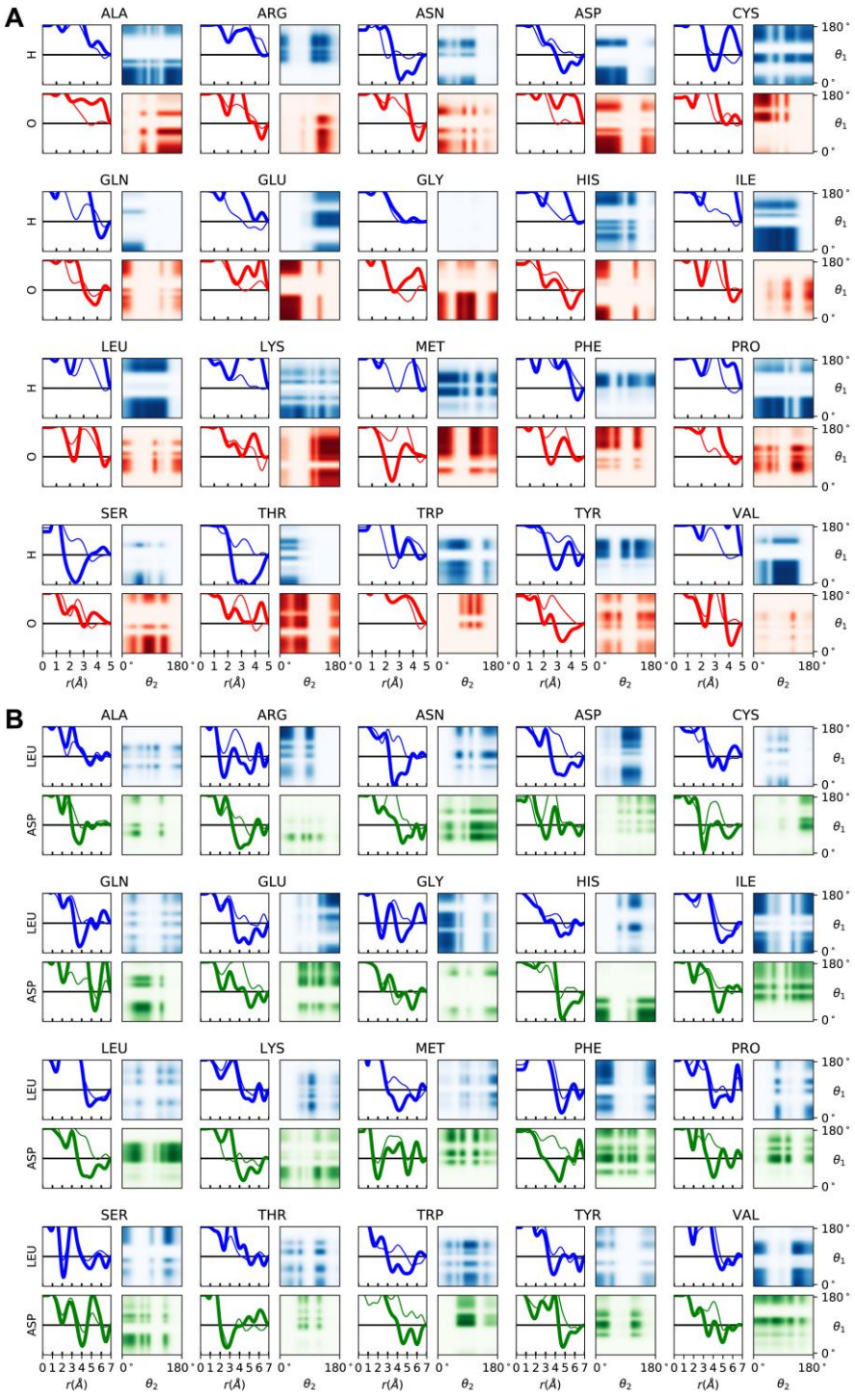

Figure 3.3: **Representative pair interaction potentials from the contrastive divergence training.** (A) Side chain bead-to-carbonyl oxygen and bead-to-amide proton (blue/red) and (B) side chain bead-to-bead (blue/green). Thin lines indicate $V_{\text{radial}}(r)$ while thick lines indicate $V_{\text{radial}}(r) + V_{\text{angular}}(r)$ with a plot range of (−6kT, 6kT). The heat maps show the angular product $\text{ang}_1(\theta_1)\text{ang}_2(\theta_2)$.

3.5.2 Accuracy of structure prediction

Contrastive divergence training has been shown to be effective for many machine learning problems [11], even without having simulations that converge to the Boltzmann ensemble. To test the accuracy of contrastive divergence on our protein model, we attempt de novo folding of a benchmark set of small, fast-folding proteins similar to those used in references [12–14] as well as various CASP11 targets investigated by other physics-based approaches (Figs 3.4 and 3.5) [15, 16]. Before training, we remove homologous proteins from the training set to help ensure that this would be a true *de novo* prediction.

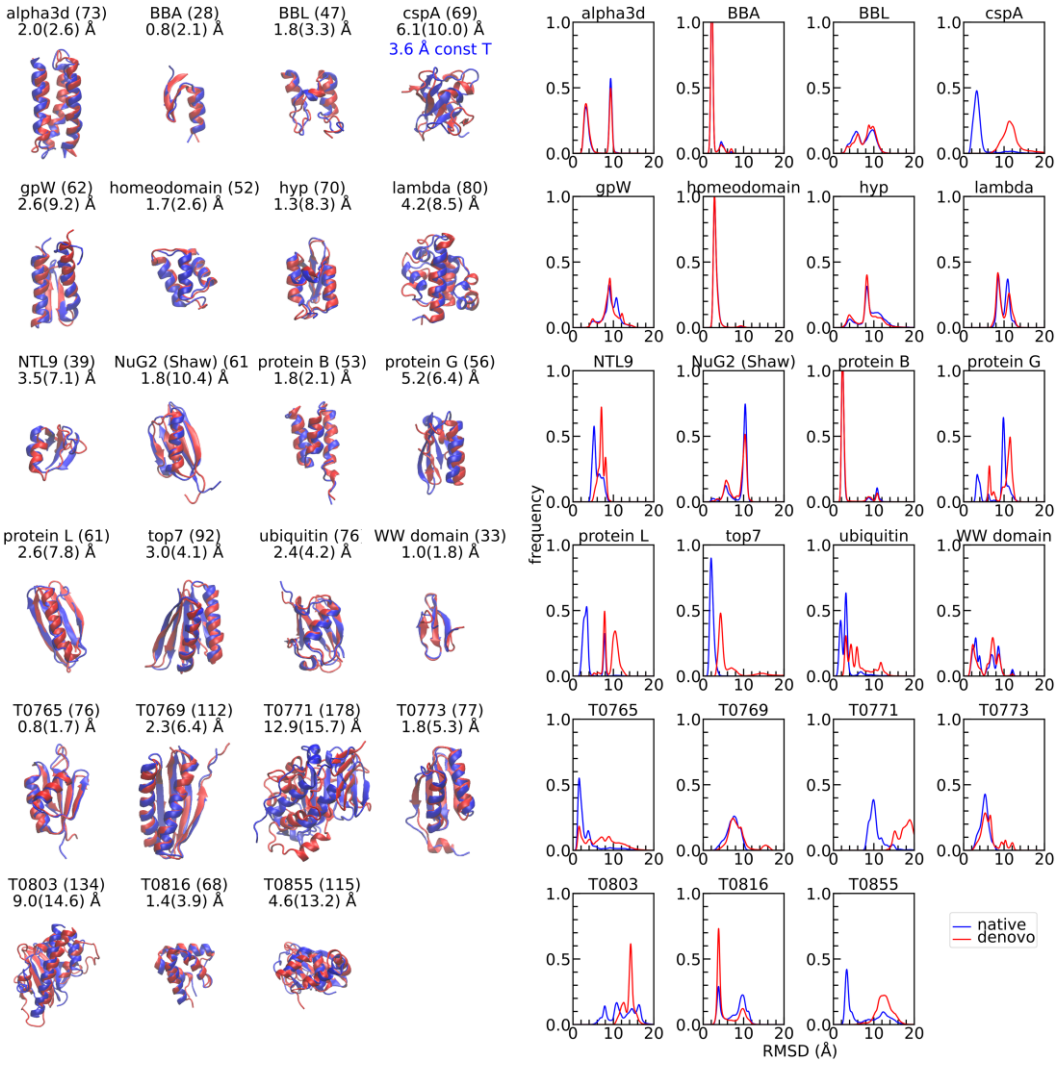

Figure 3.4: **Predicted structures and C_α-RMSD distributions.** After equilibration phase for the lowest temperature of replica exchange simulations (see S1 Text of Jumper et al.). The simulations start from either the native (blue) or a random unfolded state (red). For the refolding simulations, the lowest C_α-RMSD to native structures is provided along with the value for the centroid of largest cluster (in parentheses). RMSD calculations exclude three residues at the amino- and carboxy-termini to account for possible disorder at the ends. Each replica is run for about three days with one CPU-core.

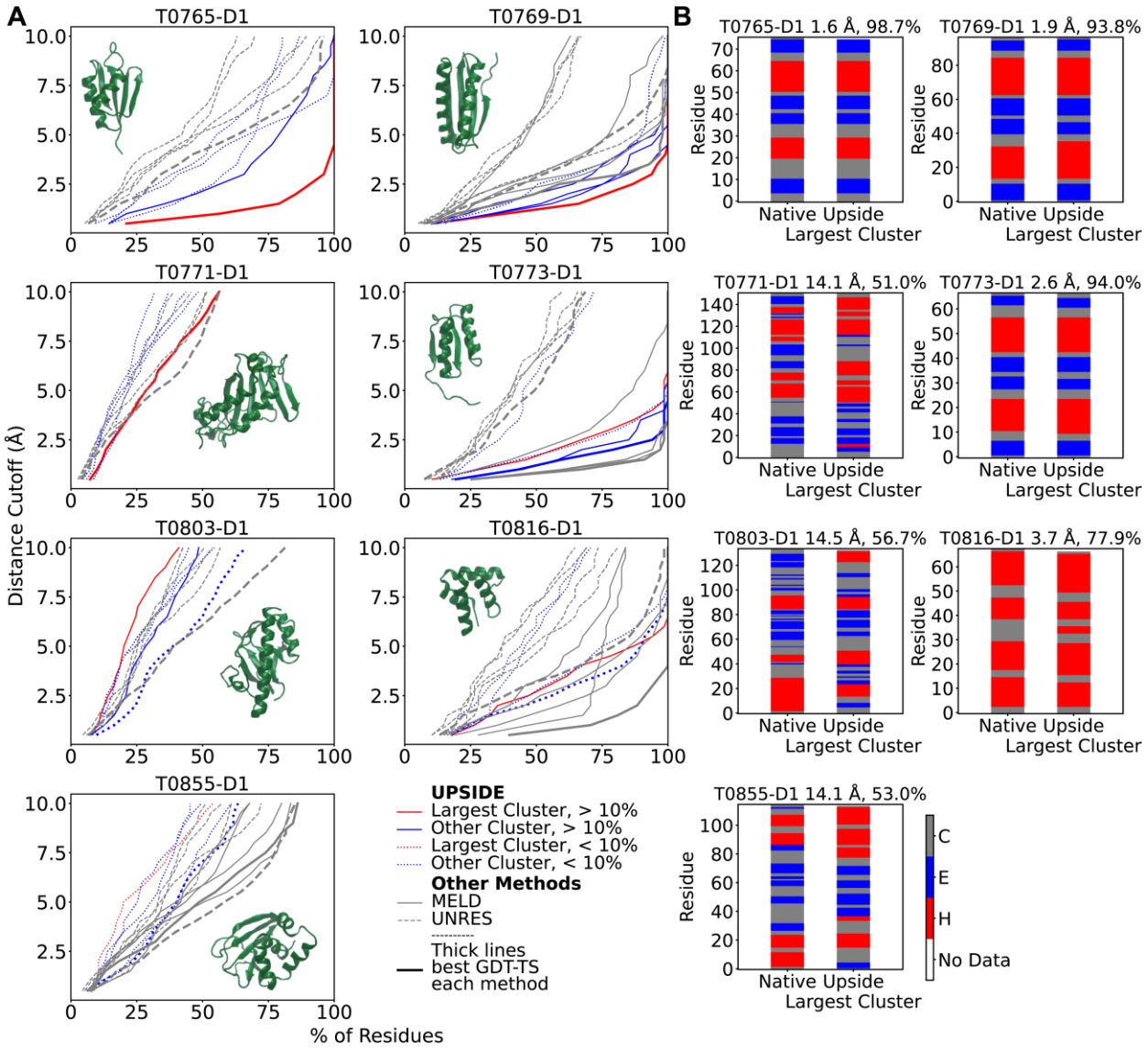

Figure 3.5: ***Upside*, UNRES and MELD's performance on seven CASP11 Targets.** (A) Hubbard plots for the centroid of *Upside*'s top five clusters are compared to the UNRES's and MELD's five submitted structures. The length and the relevant residue range used in CASP11 analysis for each protein is shown along with the structure. (B) *Upside*'s secondary structure predictions for the centroid of the top cluster (Cα-RMSD and secondary structure accuracy provided at top). The sequences provided by CASP11 organizers can be longer than the sequences used for evaluation due to disorder (e.g., for T0769-D1, simulations are conducted on 112 residues, but only the 97 folded residues are evaluated). The RMSD values provided are based on the CASP11-defined folded regions, and hence may differ slightly than those provided in Fig 3.4.

Two temperature replica exchange simulations are run for each of the 23 proteins (14 replicas each). The first set is initialized from the native configuration to assess the stability of the experimental structure for the potential obtained from contrastive divergence training. The second set is initialized from an unfolded state (random Ramachandran ϕ and ψ angles) to test *Upside*'s capability to find the native

structure which is reflection of both the accuracy of the energy function and the method's ability to search conformational space. Each range of temperatures is chosen to be large enough to cover the unfolding transition for each given protein. We judge the accuracy and equilibration from the histograms of the C_α-RMSD from the native structure after discarding the initial third of the simulation as equilibration (Fig 3.4).

The majority of the proteins show a small number of well-defined basins that represent the dominant conformations with the current potential. While the simulations often produce several conformations quickly, equilibration of their populations takes longer, on the order of CPU-days for some proteins, though still extremely short in comparison to typical molecular dynamics simulations.

For all 20 proteins below 100 residues, the lowest Cα-RMSD structure obtained starting from an unfolded state is within 5 Å of the native state (54% within 3 Å). In some cases, the lowest Cα-RMSD structure is in the largest cluster, while for other proteins, the best structure is in a minor cluster even when it is within 3 Å (e.g., gpW, NTL9). The designed 3-helix bundle, α3d [17], has a mirror image as a second heavily populated cluster.

When the native-initialized and unfolded-initialized structures have similar C_α-RMSD distributions, the simulations are likely converged. Half of the proteins are approximately converged by this criterion (e.g., BBA, protein B, homeo domain, α3d and WW), but others are not, (e.g., protein L and ubiquitin). Convergence is achieved for a variety of proteins with the native or near-native structure being the dominant conformation (e.g., BBA, homeo domain, protein B). These proteins represent the ideal scenario in terms of both accuracy and convergence. But, convergence can be achieved even when the native conformation is not the dominant conformation (e.g., BBL, λ-repressor, NuG2). This result indicates that for these proteins, our energy function is inadequate in regards to identifying the native structure even though there is adequate sampling. For cspA, a relatively small protein having a complex all β fold, additional simulations run at constant temperature can find a stable structure having significantly lower Cα-RMSD (3.6 rather than 6.1 Å); this finding points to the search process being the limiting factor rather than *Upside*'s energy function.

The *Upside* simulations tend to achieve the correct secondary structure with a small number of distinct tertiary arrangements. This diversity in tertiary structures occurs as mirrored three helix bundles

for α3d and protein B, as well as the subtle re-arrangements of NuG2. For the three largest CASP11 targets we investigated (115–178 residues), the secondary structure performance is noticeable poorer, implying a strong coupling between secondary and tertiary structure formation for these larger systems (Fig 3.5).

3.5.3 Comparison with other physics-based approaches

Simmerling and coworkers folded 17 sub93 residue proteins using GPUs to obtain a microsecond of simulation time per day with their pairwise Generalize Born (GB) model trained to reproduce Poisson–Boltzmann solvation along with their ff99SB force field [14]. Impressively, their replex protocol folded 16 of the 17 proteins to within 3 Å Cα-RMSD although the top cluster was greater than 10Å for five of the six largest proteins. Over-all, the performance is very similar to *Upside*'s in that 1-3 Å Cα-RMSD structures are achievable on most proteins but the structures are not always in the largest cluster.

For seven CASP11 targets between 65-178 residues, we compared *Upside* with two physics-based approaches that participated in CASP11 (Fig 3.5): the Cornell-Gdansk group's coarse-grained united residue model "UNRES" [16] and MacCallum, Perez and Dill's highly accelerated molecular simulation method "MELD" (Modeling Employing Limited Data), a Bayesian approach that utilizes physically-based heuristics combined with atomistic implicit solvent simulations [15]. It should be noted that both methods employ PsiPred, a secondary structure predictor employing evolutionary information [18]. In contrast, *Upside*'s secondary structures emerge during folding solely are a result of our energy function.

For T0765-D1, a 76 residue α/β protein, *Upside*'s major cluster contains the native fold (Fig 3.5). The performance is reflected in a low flat trace for the cluster centroid in the Hubbard plot of the Global Distance Test (GDT) versus sequence percentage. This performance is superior to all five of UNRES's submissions (there were no MELD submissions). For T0769-D1, a 112 residue α/β protein, both *Upside* and MELD perform very well, with UNRES's best submission being only slightly worse. For T0771-D1 and T0803-D1, 178 and 134 residue α/β proteins, respectively, neither *Upside* nor UNRES's performance is very good (no MELD submissions). For T0773-D1, a 77 residue α/β protein, MELD performs extremely well while one of *Upside* structure also has the native fold. UNRES performance is much poorer. For T0816-D1, a 68 residue helical bundle, MELD performs astonishingly well while *Upside*'s and UNRES's

performances also are commendable. For T0855-D1, a 115 residue α/β protein, both MELD and UNRES perform similarly and better than *Upside*, but none succeed in finding the native fold. Generally, the three approaches are capable of folding proteins of up to 94 residues, but are challenged with larger proteins.

3.5.4 Characterization of folding behavior

In constant temperature simulations, we observe reversible folding to the native state for a number of proteins in our test set in core-days (Figs 3.6 and 3.7). The time scales of folding indicated by these trajectories imply that the time scales we employed in the contrastive divergence simulations are far less (often a factor of 100 or more) than required to equilibrate these proteins, implying that contrastive divergence is optimizing only over fluctuations in or near the native well.

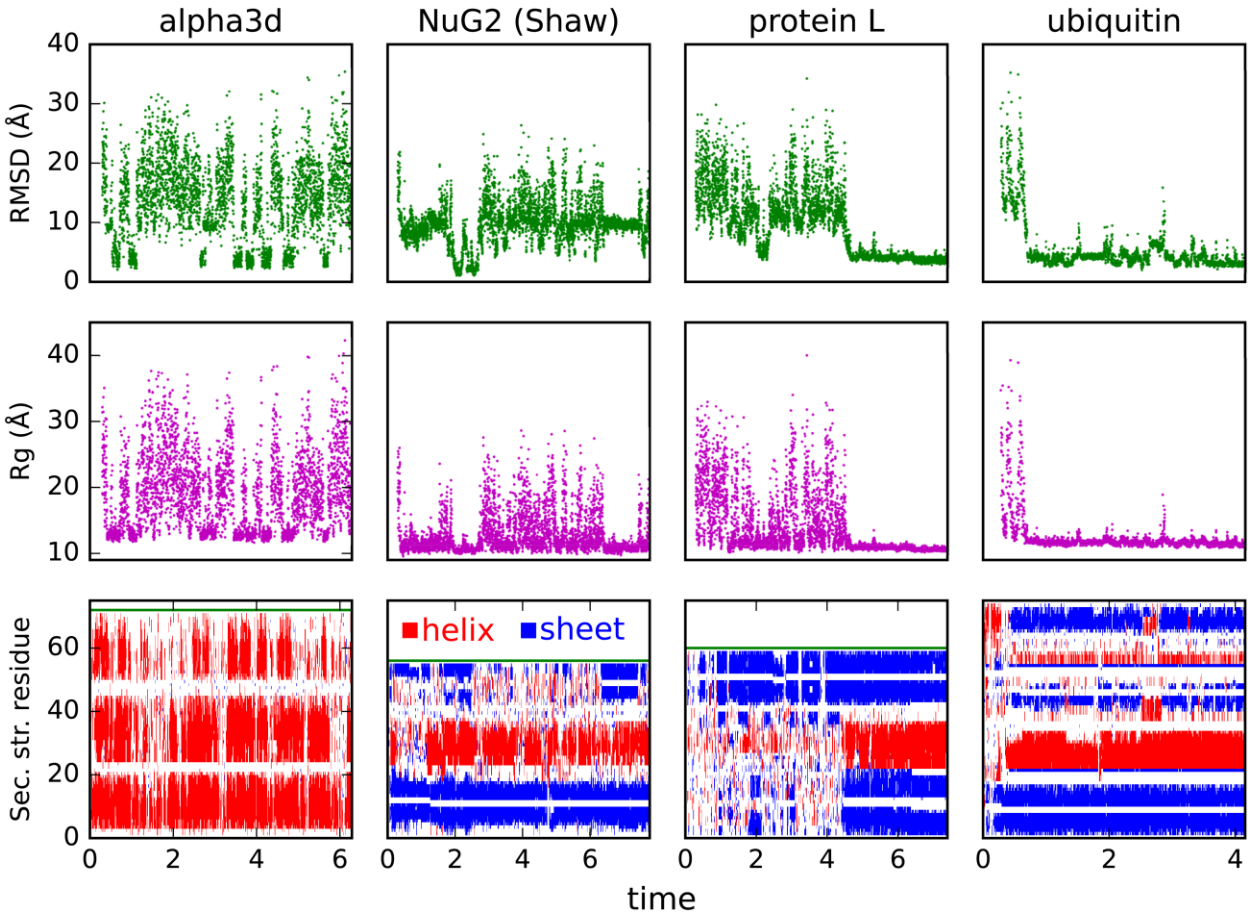

Figure 3.6: **Constant temperature simulations.** Trajectories are selected by the highest temperatures that still produce a significant population for the native state. Note that pivot Monte Carlo moves are attempted periodically which has little effect on folded dynamics but greatly decreases correlation time in the unfolded state.

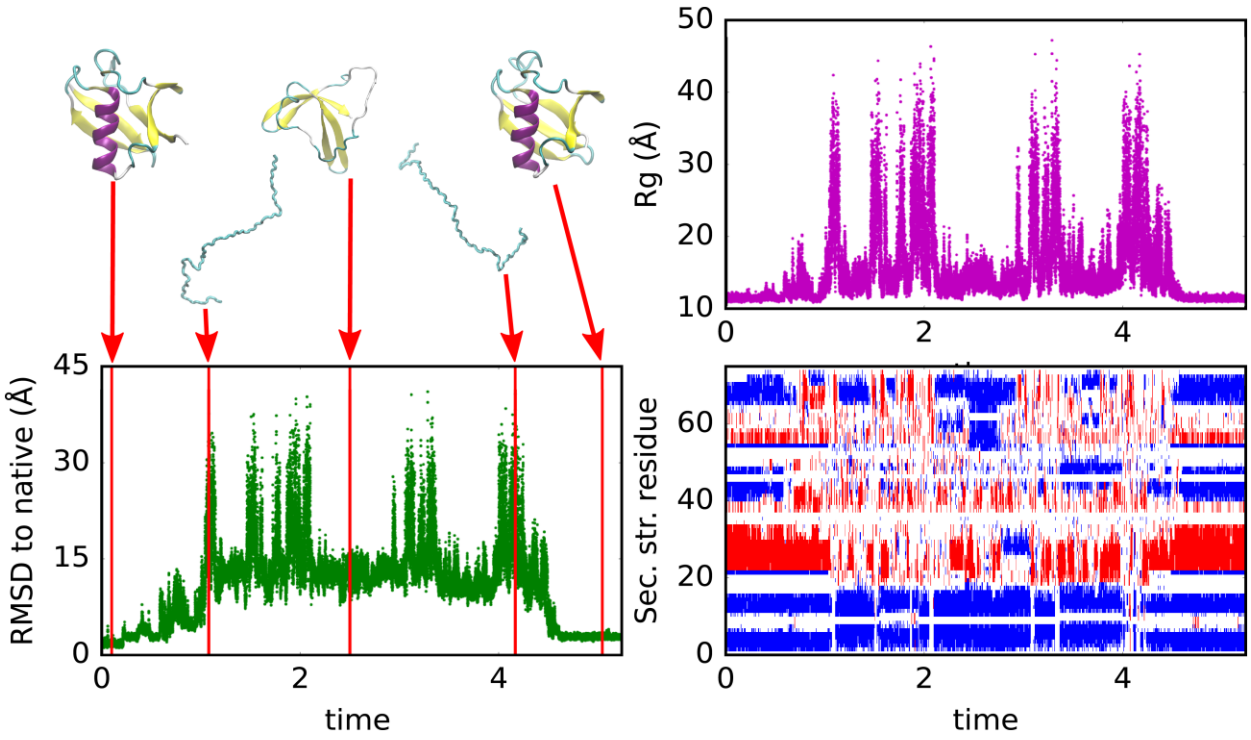

Figure 3.7: **Constant temperature trajectory of ubiquitin.** Simulation conducted at T = 1.00, initialized from the native structure, with representative structures along the trajectory highlighted. The 2nd and 4th structures are chosen for having a high R_g while the last structure is chosen based on minimum RMSD (2.3 Å.) after achieving full unfolding. Red and blue colors in lower right panel refer to helical and sheet secondary structures.

Note that conditional on low hydrogen bonding, the radius of gyration (R_g) at high temperature and at the peak of the heat capacity are quite similar. This suggests the increase in R_g for the unfolded state as temperature increases is driven by a reduction in backbone-backbone hydrogen bonds rather than side chain effects.

Based on these results, two observations should be reconciled. The first observation is the presence of a sharp phase transition with a single peak for the heat capacity. The shape of the phase transition, but not its amplitude, is consistent with a cooperative folding transition. The second observation is the unrealistically large level of residual hydrogen bonding in the denatured state at temperature of the maximum in the heat capacity. Although the hydrogen bonding is less than that in the native state, the residual hydrogen bonding indicates that the transition is not fully cooperative. These observations may be explained by the essential feature of the contrastive divergence process, that it must balance the competing energy terms of the model so that no one energy dominates. More extensive training, for

example using a more diverse ensembles that contain conformations outside the native well, may remove the excess hydrogen bonding.

The *Upside* model exhibits concerted melting behavior over a small range of temperatures (Fig 3.8). While the temperature of the model in *Upside* is not exactly comparable to a physical temperature, it is reasonable to assume T = 1 corresponds roughly to a temperature of 300-310 K. The ubiquitin transition occurs over a temperature range of approximately 0.07 temperature units, or approximately a 20 K range, similar to that observed experimentally [20].

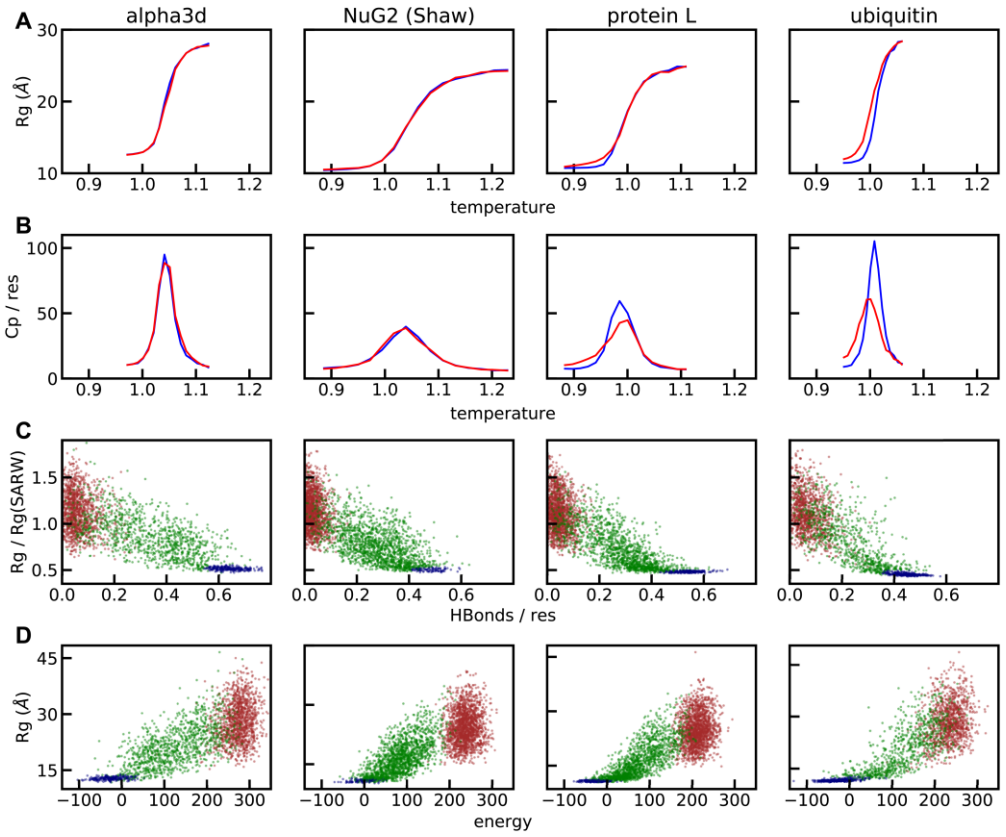

Figure 3.8: **Thermodynamic behavior.** The heat capacity is computed using the fluctuation relation $C_{\mathrm{p}} = (\mathrm{var}\, E)/T^2$. The self-avoiding random walk R_g is computing using $R_g = (1.9\text{Å})(N_{\mathrm{res}})^{0.6}$ chemically-denatured proteins [19]. In the upper two panels (A) and (B), the C_{p} and R_g values are obtained from simulations started from either the native (blue line) or a random unfolded state (red line). In the lower two panels (C) and (D), the brown points are from high temperature simulations, while the green points (unfolded state) and blue points (folded state) are from simulations at the peak of the heat capacity. The simulation units are converted to physical units by assuming that the physical energy unit is 0.6 kcal/mol and that T = 1 corresponds to 300 K.

Furthermore, our temperature-denatured states have high R_{g} near the midpoint of the transition, consistent with experimental results and inconsistent with many all-atom molecular dynamics folding

simulations [4, 21]. At the peak of the heat capacity, the R_g is ~15% smaller than the predicted from experimental data while the R_g at high temperature is ~10% larger than the experimental value. Both R_g values are significantly larger than those in most atomistic molecular dynamics simulations [4].

3.6 Discussion

A major challenge in protein chemistry is to extract from a set of proteins the underlying interaction energies that capture the physiochemistry governing their folded structures and dynamics. We addressed this challenge by showing that a strong connection exists between properties of the native basin and the rest of the protein's conformational landscape, and this connection is strong enough to train a potential for de novo folding simulations. Furthermore, the resulting potential is inexpensive enough to equilibrate simulations of small proteins in CPU core-days on a commodity computer.

Specifically, we have developed a procedure involving extremely short simulations in the native energy well, coupled with optimization using contrastive divergence, to parameterize a sophisticated coarse-grain model. Underlying the model is a re-evaluation of the common assumption that increased detail is the path to greater accuracy. This requirement for detail is mitigated with trajectory-based training because less expensive models allow more extensive exploration leading to higher accuracy. We have also shown that very large numbers of parameters (even ~20000 in our case) are no obstacle to producing accurate proteins models using trajectory-based training. While over-fitting is always a concern, the severity is greatly reduced because contrastive divergence is training against the vast possibilities of alternative protein conformations explored by conformational sampling. Additionally, contrastive divergence automatically obtains balanced parameters such that no particular interaction overwhelms the others. We contend that this balance between parameters is more important than the accuracy of any particular term.

Decoupling representations of protein physics is a key aspect of the *Upside* model. In particular, *Upside* decouples the representation of the protein used for dynamics, an N–C_α–C backbone model, from the representation used for computing energies and forces, a complex representation that includes oriented side chain interactions. This combination allows us to build up the sophisticated coordinates needed to represent solvent exposure of side chains, geometry of hydrophobic packing, and side chain-backbone hydrogen bonding without the cost of running dynamical simulation on a complex model with

slow equilibrium. The largest improvement comes from applying belief propagation to the side chain degrees of freedom so that we represent detailed side chain physics at the χ_1/χ_2-level without incurring the roughening of the energy landscape and slowing of the dynamics normally associated with detailed sterics of side chain interactions. It is an open question to determine how much molecular detail must be retained for accurate protein energetics, but *Upside* provides a flexible framework to explore these issues without compromising the simple backbone representation of dynamics.

3.6.1 Related work

Contrastive divergence optimization has been applied to Gō-like protein potentials sampled with crankshaft Monte Carlo moves [22, 23]. These works optimized only tens of parameters, and the resulting model is used to fold protein G and 16-residue peptides.

Other studies have trained protein energy functions using libraries of decoys [24]. Such efforts are challenging because atomic energy functions have rugged energy landscapes where even small structural differences can produce large energy differences. This ruggedness implies that scoring decoys by energy without first relaxing them is problematic for the sharply-defined force fields necessary to describe protein physics, a problem that contrastive divergence avoids.

A distinction between contrastive divergence and traditional training methods, such as Z-score optimization [25], relates to the goal and the source of the decoys. In contrastive divergence, the critical task is to produce a high population of low RMSD structures with the model. Z-scoring training attempts to make the energy of the native state much lower than the average energy of an pre-constructed decoy library. This is problematic because the decoys may not have structures that exhibit the pathologies of a poorly-trained model. Additionally, we believe optimization should concentrate on the lowest energies that have significant Boltzmann probability, not the average energy which is dominated by highly-unlikely structures. Furthermore, it is difficult to evaluate the reliable energies of decoys without relaxing the decoys. Methods based on simulation ensembles (such as maximum likelihood and contrastive divergence) are well-defined and do not need pre-constructed decoy libraries.

Podtelezhnikov et al. [26] apply contrastive divergence to few-parameter protein models to optimize the parameters of hydrogen bond geometry. Their work is similar to this paper but narrower in scope.

The maximum likelihood method requires the computation of the derivative of the free energy, which involves a summation over an equilibrium ensemble. Such a requirement necessitates a very long simulation to update parameters. Still, this approach can be viable when used with very small proteins on which the simulations converge quickly. A variant of maximum likelihood is given in Ref. [27], where decoys are generated and a maximum likelihood model is fit to adjust the parameters to distinguish between near-native and far-from-native conformations. The potential is trained on a single protein, tryptophan cage, and then the resulting potential is applied to a number of α-helical proteins with some success.

3.6.2 Time and temperature scale

The precise time scale and temperature scale of the *Upside* models is intentionally left arbitrary because the coarse-graining process may leave us without a linear relationship to physical time and temperature. The speed-up of *Upside* simulation due to the smoothing of side chain interactions is likely to have a disproportionate effect on time scales for condensed structures as compared to extended structures. Regardless, the equilibrium population distribution that determines the free energy is expected to be approximately correct, as well as the order of dynamical folding events. The precise relationship of *Upside* time scales to physical time scales is left to future work.

3.6.3 Conclusion

By employing the computationally fast yet detailed *Upside* model, we can use multiple trajectories to train tens of thousands of parameters simultaneously to simulate protein folding and dynamics. The training successfully produces low-energy, native or near-native structures with sharp folding transitions for most of our validation proteins. The strategy's success argues that simpler (in atomic representation) models that can be globally parameterized can rival more detailed but slower models whose parameterization is more challenging. We achieve success for some proteins in terms of accurately folding to low energy native state and achieve thermodynamic equilibration, but still fail on others. We hypothesize that the short-time contrastive divergence we are using does not provide a sufficient library of large changes in the tertiary structure to enable the potential to properly distinguish the various conformations. This issue will be addressed in future studies. Coupling large computational resources with Markov state models [28]

should improve training of the *Upside* model by exploring a larger and more diverse conformational landscape on each contrastive divergence step.

The ready generation of Boltzmann ensembles allows for a wide range of computational studies of protein folding, dynamics, and binding. For example, computational screening of large numbers of proteins for foldability should be tractable as is the study of hydrogen exchange and folding kinetics. Additionally, in studies that incorporate experimental or bioinformatics data, including contact predictions, *Upside* provides an inexpensive Bayesian prior distribution over protein structures that may be updated using experimental information. This provides accurate predictions that make essential use of the totality of protein physics as encoded in the *Upside* model, while being inexpensive enough to allow validation and iteration on large numbers of proteins.

3.7 Supporting information

The supporting information for the original paper, “S1 Text. Derivation, model, and optimization details.” can be obtained from https://doi.org/10.1371/journal.pcbi.1006578.s001 (PDF).

3.8 Acknowledgments

We would like to thank Sheng Wang for helpful discussions during this research. We thank Carolyn Jumper for proofreading and editorial assistance.

3.9 References

[1] Adhikari AN, Freed KF, Sosnick TR. Simplified protein models: Predicting folding pathways and structure using amino acid sequences. Physical review letters. 2013;111(2):028103. pmid:23889448

[2] Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins: Structure, Function, and Bioinformatics. 1999;37(S3):171–176.

[3] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. Proteins: Structure, Function, and Bioinformatics. 2016;84(S1):4–14.

[4] Skinner JJ, Yu W, Gichana EK, Baxa MC, Hinshaw JR, Freed KF, et al. Benchmarking all-atom simulations using hydrogen exchange. Proceedings of the National Academy of Sciences. 2014;111(45):15975–15980.

[5] Best RB, Hummer G. Optimized molecular dynamics force fields applied to the helix- coil transition of polypeptides. The Journal of Physical Chemistry B. 2009;113(26):9004–9015. pmid:19514729

[6] Jumper JM, Faruk NF, Freed KF, Sosnick TR. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. PLOS Comput Biol. 2018 Dec 27;14(12):e1006342.

[7] Ting D, Wang G, Shapovalov M, Mitra R, Jordan MI, Dunbrack RL Jr. Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. PLoS Comput Biol. 2010;6(4):e1000763. pmid:20442867

[8] Salakhutdinov RR. Learning in Markov random fields using tempered transitions. In: Advances in neural information processing systems; 2009. p. 1598–1606.

[9] Desjardins G, Courville A, Bengio Y, Vincent P, Delalleau O. Parallel tempering for training of restricted Boltzmann machines. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. MIT Press Cambridge, MA; 2010. p. 145–152.

[10] Duvenaud D, Maclaurin D, Adams RP. Early Stopping as Nonparametric Variational Inference. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics; 2016. p. 1070–1077.

[11] Carreira-Perpinan MA, Hinton G. On Contrastive Divergence Learning. In: AISTATS. vol. 10. Citeseer; 2005. p. 33–40.

[12] Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science. 2011;334(6055):517–520. pmid:22034434

[13] Adhikari AN, Freed KF, Sosnick TR. De novo prediction of protein folding pathways and structure using the principle of sequential stabilization. Proceedings of the National Academy of Sciences. 2012;109(43):17442–17447.

[14] Nguyen H, Maier J, Huang H, Perrone V, Simmerling C. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. Journal of the American Chemical Society. 2014;136(40):13959–13962. pmid:25255057

[15] Perez A, Morrone JA, Brini E, MacCallum JL, Dill KA. Blind protein structure prediction using accelerated free-energy simulations. Science advances. 2016;2(11):e1601274. pmid:27847872

[16] Krupa P, Mozolewska MA, Wiśniewska M, Yin Y, He Y, Sieradzan AK, et al. Performance of protein-structure predictions with the physics-based UNRES force field in CASP11. Bioinformatics. 2016;32(21):3270–3278. pmid:27378298

[17] Walsh ST, Cheng H, Bryson JW, Roder H, DeGrado WF. Solution structure and dynamics of a de novo designed three-helix bundle protein. Proceedings of the National Academy of Sciences. 1999;96(10):5486–5491.

[18] McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000;16(4):404–405. pmid:10869041

[19] Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, et al. Random-coil behavior and the dimensions of chemically unfolded proteins. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(34):12491–12496. pmid:15314214

[20] Jacob J, Krantz B, Dothager RS, Thiyagarajan P, Sosnick TR. Early collapse is not an obligate step in protein folding. Journal of molecular biology. 2004;338(2):369–382. pmid:15066438

[21] Jensen MR, Blackledge M. Testing the validity of ensemble descriptions of intrinsically disordered proteins. Proceedings of the National Academy of Sciences. 2014;111(16):E1557–E1558.

[22] Podtelezhnikov AA, Wild DL. Inferring Knowledge Based Potentials Using Contrastive Divergence. In: Bayesian Methods in Structural Bioinformatics. Springer; 2012. p. 135–155.

[23] Várnai C, Burkoff NS, Wild DL. Efficient parameter estimation of generalizable coarse-grained protein force fields using contrastive divergence: a maximum likelihood approach. Journal of chemical theory and computation. 2013;9(12):5718–5733. pmid:24683370

[24] Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins: Structure, Function, and Bioinformatics. 2003;53(1):76–87.

[25] Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Ołdziej S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization. Journal of computational chemistry. 1997;18(7):874–887.

[26] Podtelezhnikov AA, Ghahramani Z, Wild DL. Learning about protein hydrogen bonding by minimizing contrastive divergence. Proteins: Structure, Function, and Bioinformatics. 2007;66(3):588–599.

[27] Zaborowski B, Jagieła D, Czaplewski C, Hałabis A, Lewandowska A, Zmudzińska W, et al. A Maximum-Likelihood Approach to Force-Field Calibration. Journal of chemical information and modeling. 2015;55(9):2050–2070. pmid:26263302

[28] Bowman GR, Beauchamp KA, Boxer G, Pande VS. Progress and challenges in the automated construction of Markov state models for full protein systems. The Journal of chemical physics. 2009;131(12):124101. pmid:19791846

CHAPTER 4

CONCLUSION

This thesis has demonstrated *Upside*'s capability of folding many sub-100 residue proteins to near-native structures and how the model is competitive with other physics-based approaches. *Upside* was established on the paradigm that ineffective model parameterization was a more limiting factor for MD simulation models compared to the level of force field detail. This view is upheld by the comparable folding results to other methods with our CG model using the native state simulation contrastive divergence training procedure. However, further improvements to folding performance in subsequent work required better training and expansion of energy terms [1, 2] to give a more accurate account of the full Boltzmann ensemble. This ability to probe thermodynamics and kinetics sets MD methods, and *Upside* in particular, apart from recent neural network native structure prediction methods, and makes *Upside* a worthwhile model to continue developing.

We also found that *Upside* is comparable to traditional atomistic docking methods after addition of inter-protein energy terms and training on a docking benchmark set. We consider this a notable achievement for a CG model that is a new entrant to the world of protein docking. The new energy terms, including a pairwise sidechain-sidechain term and a multibody sidechain burial term to capture desolvation effects, were informed by the greater role of sidechain interactions in protein binding. Reconstruction of explicit sidechains with SCWRL4 [3] and rescoring with the atomistic SOAP-PP potential [4] only provided a minor boost to our performance, which supports again our view that effective parameterization of our CG model outweighs limitations from its level of detail. However, explicit side chains are likely necessary for high accuracy predictions, considering how tight packing at the native interface helps with its selectivity.

We further found that allowing for backbone flexibility tends to be detrimental. Prediction of native-like poses among the top predicted docking poses decreased for the fully flexible setup of Upside compared to a semirigid setup with backbone restraints. Allowing for backbone flexibility appeared to make the subunit conformations move further away from their native bound conformations, which was confirmed with native state simulations. Through these native state simulations, we observed how our

folding performance indicated via subunit RMSDs provided a bound on our docking performance, which comes from a joint measure of the IRMSDs and subunit RMSDs. Although our folding performance of maintaining 2–5 Å subunit RMSDs for many complexes is commendable, we have greater than 5 Å RMSD for many other cases, which precludes the prediction of medium quality complexes. The distribution of CAPRI metrics moves to more native-like values for most complexes when subunit backbone restraints are applied. In this case, it is satisfying that we can maintain a local minimum near the native state. There are still some high IRMSD deviations for complexes that have low interfacial area, which highlights deficiencies in our interprotein terms. In a comparison with the CG CABS method [5], and all-atom explicit solvent tempered binding method [6], we argued how problems with backbone flexibility are general to MD/MC methods where backbone energies are involved in driving the conformational search.

The case of antibody-antigen docking showed that *Upside*'s backbone flexibility and sampling efficiency can be beneficial when inaccuracies in the forcefield are compensated for by sequence-based biasing information. *Upside* with fully flexible CDR loops achieved more native-like top 10 predictions than HADDOCK [7] and Upside with semirigid loops for medium difficulty antibody complexes. However, we found that internal structures of the CDR loops do not change much from their unbound starting states during our simulations. The loops encounter steric hinderance and attractive interactions that lock them into a conformation, although there may be overall center of mass shifts and rotations in the reference frame of the entire antibody. Simulations of the free antibody allow the CDR loops to sample lower RMSD conformations. This points to a conformational selection mechanism for the folding of the CDR loops, which highlights the continued utility of MD methods to investigate mechanisms.

In the future, we can attempt joint training of the folding and binding components of the *Upside* forcefield in a procedure similar to [1]. In this method, instead of a maximum likelihood approach on potential energies, we can optimize on free energies using contrastive divergence by using replica exchange of decoy decoys and the native pose. We may also benefit from adding an additional bead to the sidechain model.

We conclude on a summary of the discussion on the differences between protein folding and protein binding and challenges to model them. In Rose's model of protein folding, backbone hydrogen

bonding and sterics impose the major organizational constraints that reduces a myriad of conformations to about 10,000 fundamental folds [8]. Larger proteins are constructed as modules of these folds, such that complexity grows tractably. The challenge for computer models is representing the perfect balance of energy terms and undertaking considerable sampling to find the native structure; imperfections in the force field lead to kinetic trapping and misfolded states being more stable than the native state.

For protein binding, sidechain interactions with other sidechains and solvent dominate. One challenge for computer models is in tuning these interactions with a lower margin of error than in folding since the backbone interactions cannot help to compensate to the same extent. In terms of sampling, Janin's simple model of precomplex formation indicates about 67,000 poses are required to find the native pre-complex of barnase-barstar [9]. For larger complexes, $N_{\text{decoy}} \propto (SA)^2 \propto \left(M_{prot}^{0.73}\right)^2$ [10], which can be easily covered by rigid-body docking algorithms that can generate 100,000s of poses in minutes with modern parallel computing. However, we have demonstrated how high accuracy binding is limited by accuracy of refining subunits to their bound conformations, and in this way protein binding has its own challenges in addition to those of protein folding. Although other MD studies have looked at the interplay between binding and flexible refinement, such as with HADDOCK and the prediction of H3 loop conformations [7], we believe that *Upside* is ideally situated for a comprehensive analysis because of its ground-up design and efficiency for protein folding.

4.1 References

[1] Gaffney KA, Guo R, Bridges MD, Chen D, Muhammednazaar S, Kim M, et al. Lipid Bilayer Induces Contraction of the Denatured State Ensemble of a Helical-Bundle Membrane Protein. BioRxiv [Preprint]. 2021 bioRxiv 444377. [posted 2021 May 17; cited 2021 Oct 10]. Available from: https://www.biorxiv.org/content/10.1101/2021.05.17.444377v1.

[2] Peng X, Baxa M, Faruk N, Sachleben J, Pintscher S, Gagnon I, et al. Prediction and validation of a protein's free energy surface using hydrogen exchange and (importantly) its denaturant dependence. [Submitted Sep. 2021]

[3] Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009 Dec;77(4):778–95.

[4] Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, Sali A. Optimized atomic statistical potentials: assessment of protein interfaces and loops. Bioinforma Oxf Engl. 2013 Dec 15;29(24):3158–66.

[5] Kurcinski M, Kmiecik S, Zalewski M, Kolinski A. Protein-Protein Docking with Large-Scale Backbone Flexibility Using Coarse-Grained Monte-Carlo Simulations. Int J Mol Sci. 2021 Jul 8;22(14):7341.

[6] Pan AC, Jacobson D, Yatsenko K, Sritharan D, Weinreich TM, Shaw DE. Atomic-level characterization of protein–protein association. Proc Natl Acad Sci. 2019 Mar 5;116(10):4244–9.

[7] Ambrosetti F, Jiménez-García B, Roel-Touris J, Bonvin AMJJ. Modeling Antibody-Antigen Complexes by Information-Driven Docking. Structure. 2020 Jan 7;28(1):119-129.e2.

[8] Rose GD. Protein folding - seeing is deceiving. Protein Science. 2021;30(8):1606–16.

[9] Janin J. The kinetics of protein-protein recognition. Proteins Struct Funct Bioinforma. 1997 Jun 1;28(2):153–61.

[10] Glyakina AV, Bogatyreva NS, Galzitskaya OV. Accessible Surfaces of Beta Proteins Increase with Increasing Protein Molecular Mass More Rapidly than Those of Other Proteins. PLOS ONE. 2011 Dec 1;6(12):e28464.