Ensemble Methods in Computational Protein and Ligand Design: Applications to the Fcγ Immunoglobulin, HIV-1 Protease, and Ketol-acid Reductoisomerase Systems

by

Nathaniel White Silver

A.B. in Chemistry, Bowdoin College (2006)

Submitted to the Department of Chemistry
in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy in Physical Chemistry

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2012

Author .....
Department of Chemistry
December 15, 2011

Certified by .....
Bruce Tidor
Professor of Biological Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Robert W. Field
Chairman, Department Committee on Graduate Students

Thesis Committee

Accepted by .....

Jianshu Cao
Professor of Chemistry
Chairman of Thesis Committee

Accepted by .....

Bruce Tidor
Professor of Biological Engineering and Computer Science
Thesis Supervisor

Accepted by .....

Moungi G. Bawendi
Lester Wolfe Professor in Chemistry
Thesis Committee Member

Ensemble Methods in Computational Protein and Ligand Design: Applications to the Fcγ Immunoglobulin, HIV-1 Protease, and Ketol-acid Reductoisomerase Systems

by

Nathaniel White Silver

Submitted to the Department of Chemistry
on December 15, 2011, in partial fulfillment of the
requirements for the degree of
Doctorate of Philosophy in Physical Chemistry

Abstract

This thesis explores the use of ensemble, free energy models in the study and design of molecular, biochemical systems. We use physics based computational models to analyze the molecular basis of binding affinity in the context of protein–protein and protein–ligand binding as well as reaction rate enhancement in enzyme catalysis. First, we evaluate the solvent screened energetics of immunoglobulin G (IgG):Fcγ receptor binding using molecular mechanics, Poisson–Boltzmann surface area (MM-PBSA) models. We assess the role IgG_1–linked glycans play in binding to human FcγR-III and computationally evaluate experimentally designed Fc mutations that recover binding affinity in the absence of glycosylation. Using the insight gained from this study, we developed novel murine IgG variants with engineered Fcγ receptor binding patterns via the computational design and experimental validation of Fc mutations that are predicted to knock out binding to FcγR-IV. Our design and analysis highlight the importance of solvent screened electrostatic interactions and electrostatic complementarity in protein–protein binding. Second, we develop novel, ensemble methods to measure configurational free energy and entropy changes in protein–ligand binding and use it to predict the relative binding affinity of a series of previously designed HIV-1 protease inhibitors. We find that using configurational free energies to evaluate inhibitor efficacy significantly improves relative ranking of inhibitors over traditional, single–point energy metrics, but that only a relatively small number of low energy configurations are necessary to capture the ensemble effect. Finally, we present a joint study of the redesign and dynamic analysis of ketol-acid isomeroreductase (KARI). We first develop and apply a novel, end–point method to rationally design enzyme variants that reduce the free energy of activation, and present the computational and experimental analysis of a series of designed KARI mutants. Our analysis reveals that this transition–state theory based approach is effective at reducing the enthalpy of activation, but also increases entropic activation penalties that ultimately overpower the enthalpic gains. A dynamic analysis of these

KARI variants is also presented, in which the transition path ensemble is explored using transition path sampling. We find that this ensemble approach is better able to predict relative enzyme activities and suggests a conserved, dynamic mechanism for catalysis. The results and analysis presented herein demonstrate novel, computational approaches to account for ensemble effects in the study and design of effective biomolecules.

Thesis Supervisor: Bruce Tidor
Title: Professor of Biological Engineering and Computer Science

Acknowledgements

First and foremost, I would like to thank my family for all the love and support they have given me. Their sage advice and constant encouragement has been invaluable during my time before and at MIT. I would also like to thank Kristina Fenn for all her support, especially during my hectic, final few months of graduate school.

I would like to thank my advisor, Bruce Tidor, whose guidance and passion for rigorous science has been instrumental in both shaping this work as well as my development as a scientist. In addition, I am greatly indebted to all the members of the Tidor lab that I have had the pleasure of interacting with during my time at MIT. A heartfelt thanks to Yuanyuan Cui, Pradeep Ravindranath, Sudipta Samanta, Devan Raghunathan, Ishan Patel, David Hagen, Nirmala Paudel, Gil Kwak, Tina Toni, Thomas Gurry, David Witmer, Bo Kim, Aurore Zyto, Mala Radhakrishnan, Shuan Lippow, Kathryn Loving, Caitlin Bever, Katharina Wilkins, Jared Toettcher, Josh Apgar, Michael Altman, and Dave Huggins. Through weekly suggestions at group meeting and daily lunch time conversation, I have grown both professionally and personally. I would like to especially thank Yang Shen, Filipe Gracio, Jason Biddle, and Harold Hwang for their friendship and support over the years, as well as excellent advice regardless of whether I asked for it. Also, a special thanks to Bracken King, my long time desk mate and fellow system administrator, for his words of wisdom and support.

I would like to acknowledge and thank all of my collaborators in the Dupont–MIT Alliance, Schiffer, Rana, and Wittrup labs, especially those individuals who took the time to experimentally test my theoretical predictions. My thanks to Tiffany Chen for her work characterizing Fc mutants, Hong Cao for her work measuring HIV protease inhibitor affinities, Akbar Ali and Kiran Reddy for their work synthesizing said inhibitors, and Katharine Gibson and Mike McCluskey for their work characterizing KARI mutants.

Finally, thank you to the Forbes Family Cafe, Anna's Taqueria, the Kendall Square food court, Jose's Mexican food truck, and Gooseberries for lunch time sustenance as well as Tommy Doyle's and the John Stamos Theatrical Extravaganza/Explosion for mind expanding trivia.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction and Motivation

Over the past 50 years, the fields of computational chemistry and structural biology have seen enormous growth in both the development of analytical *in silico* tools as well as the amount of structural data available. Since the first crystal structure of myglobin was solved in 1958 [1], more than 70,000 protein structures have been found and deposited in the Protein Data Bank [2]. Additionally, the proliferation of fast, inexpensive computers has driven the rapid development in computational methods to analyze these biological data. Decades of work by many computational scientists has resulted in the development of accurate force fields to measure the atomic scale interactions of proteins in a given conformation as well as a host of algorithms to explore the dynamics of these and other biological structures [3, 4, 5, 6, 7, 8]. Recent applications of these computational tools have shown a demonstrated ability to predict structural and energetic properties of proteins and their relevant ligands [9, 10, 11]. It is well known, however, that a single, static conformation of any biologically relevant chemical species and the sum of all its inter-atomic interactions fails to accurately represent all the physical parameters that define that chemical species [12, 13]. Crystallographic models primarily reveal ensemble averaged structures of macromolecules, and while they provide a great deal of insight into the structure of the complex interaction network in biochemical systems, they are not the whole

story. They represent a single point on the potential energy landscape or a single point in phase space when assigned momenta, with limited information about the relative thermal fluctuations of any given atom (B-factors). Experimental measurements of binding affinity, stability, or reaction rates correspond to average properties of the system and include contributions from the average enthalpy as well as entropic effects. As such, accurate absolute or relative prediction of these properties often requires not just accurate force fields, but effective sampling of configurational and/or momentum space such that converged, thermodynamic or kinetic averages can be obtained. Furthermore, this sampling must be done in accordance with the underlying Boltzmann distribution that governs the equilibrium probability of observing any given point in phase space. Common sampling methods include molecular dynamics as well as Monte Carlo based search routines. When modeling complex biological systems, however, it is not always possible to exhaustively explore phase space due to computational constraints and the sheer number of degrees of freedom present. As such, approximations are often made such that only selected degrees of freedom are treated in an ensemble versus static manner; however, appropriate partitioning is non-trivial. Successful use of approximate models requires careful assessment of all those physical effects important to the system under study. Failure to account for these effects often yields poor agreement between model and experiment, while accounting for far more than is necessary can result in computational intractability.

In the case of protein–protein docking, there are often a huge number of relevant degrees of freedom in both the bound and unbound states, including side chain motions, backbone motions, and solvent degrees of freedom. Current approaches to modeling bound complexes are often iterative and hierarchical, in which different degrees of freedom are considered at different times during the docking process [14]. Full, explicit solvent molecular dynamics methods can be used to assess binding, but when the goal is to compare multiple variants or binding modes, such approaches come with a large computational cost. Simulations must be run for long enough to sample all relevant, high probability portions of phase space. Rather than treating all solvent molecules explicitly, a common approximation is to use continuum sol-

vent methods such as Poisson–Boltzmann Surface Area (PBSA) [15] or Generalized Born Surface Area (GBSA) [16] models, which account for solvent degrees of freedom through an implicit, mean field approach and are thus much faster to evaluate. For a given protein conformation, they allow for the rapid computation of the solvent free energy. When combined with a conformational sampling scheme, this method allows much faster exploration of protein configurational space. However, the utility of such continuum models is still an open question. In some systems, they compare very well to explicit solvent models and are highly predictive [17, 18, 19], but in others they fail to capture important solvent properties without system-specific parametrization [20, 21], limiting their broad application.

In the case of ligand–protein binding, there are similar issues of scale, with large numbers of ligand and protein degrees of freedom contributing to the thermodynamics of binding. Continuum solvent approximations are used, often with an additional simplification, the rigid binding approximation. This assumes that the bound complex is well defined by a single, predominant configuration, and that both the ligand and receptor adopt identical conformations in the unbound state. Using this to compare many chemically related ligands and predict relative activity, one implicitly assumes negligible differences in entropies of binding between different ligands. Some models correct for this by assuming a constant entropic penalty proportional to the number of ligand rotatable bonds [22, 23], but studies have shown that this can often be inaccurate [24, 25]. Such rigid binding models are commonplace in high throughput virtual screens, and useful when one is interested in simply discriminating binders from nonbinders [26]. When absolute or relative affinity measurements are required, however, these approximations are inappropriate, and can result in poor correlation with experiment [27]. As such, developing accurate, efficient ways of computing ensemble free energies remains an active field of study.

In a similar vein, when modeling enzymatic reactions, one is often interested in evaluating the free energy difference between the transition–state and reactant–state ensembles, as this difference is hypothesized to be proportional to the phenomenological rate of reaction [28]. It is difficult, however, to map out the conformational

ensemble of the transition–state without information about the specific reaction mechanism. Additionally, accurate modeling of bond breaking and bond forming reactions requires a quantum mechanical treatment, which is very computationally expensive. As such, it is common to assume a specific reaction coordinate and find transition state(s) along that path [29]. In reality, however, for enzymatic systems, there are a multitude of paths with different transition states, and single coordinate or single transition state models can miss important regions of phase space, leading to poor assessment of the activation free energy [30, 31]. Recent sampling algorithms such as the string and nudged elastic band methods [32, 33] have attempted to improve sampling rare events and their associated transition states by trying to find minimum free energy paths, but these methods often require either a smooth potential energy surface or a good initial guess of the path and do not actually yield dynamic trajectories. A recently developed transition path sampling algorithm [34] allows for unbiased sampling of reactive paths without *a priori* knowledge of the reaction mechanism, and can generate appropriately weighted ensembles of transitions that carry systems across reaction barriers. Its utility has been validated on very small systems but has only seen limited application to enzymes.

In this thesis, we test the validity many of these approximations via the application and development of physics-based, free energy, ensemble models to biochemical systems, with specific focus on the application of such models in the context of ligand and protein design. One of the long–term goals in the application of computational chemistry and biology has been the *de novo* design of small molecules and proteins for medicinal or scientific purposes. Experimental methods to design effective small molecule or protein therapeutics rely on expensive, large–scale screens to explore the vast chemical or sequence space; initial chemical screens in modern pharmaceutical operations commonly explore chemical libraries on the order of 10^6 compounds [35]. As such, computational design and virtual screening methods are often employed to offset experimental cost and improve the odds of finding successful candidates [26]. Overall, this thesis examines the use of ensemble, free energy models in the evaluation and design of specificity in protein–protein interactions, high–affinity small molecule

inhibitors, and active enzyme catalysts. We explore the role of solvent mediated electrostatic interactions through continuum electrostatics, configurational entropy and free energies in ligand binding via ensemble enumeration techniques, as well as activation free energies and protein dynamics in enzyme catalysis through transition state theory and transition path sampling.

In Chapter 2 we examine the physical driving forces behind protein–protein affinity in the antibody–receptor, IgG:Fcγ system through the application of mean-field, continuum electrostatic models. The motivation for this work was the discovery of an aglycocylated, human IgG antibody that successfully bound to and activated the Fcγ-III receptor. It is known that removal of the glycan attached to the Fc region of the hIgG molecule normally results in reduced affinity or loss of binding to all Fcγ receptors [36, 37, 38, 39, 40, 41, 42]. However, through the rational redesign of the three residue glycosylation tag, mutants were found that recovered binding. Previous computational studies had indicated the importance of electrostatic interactions in protein association [43], specificity [44], and binding [45, 19], with the interplay between desolvation penalties and coulombic interactions being notably important [46]. We explore the ability of an MM-PBSA, implicit solvent model, combined with a hierarchical configurational search scheme, to capture these interactions and develop homology models of the wild–type and aglycocylated, mutant structures. These models are used to examine the role played by the glycan as well as the mutants upon binding to the Fcγ receptor. We find that much of the observed binding pattern can be explained by changes in the solvent mediated electrostatic interactions that improve electrostatic complementarity at the binding interface. Building on this discovery, in Chapter 2 we also present the negative design of mouse IgG antibodies to knock out binding to the Fcγ-IV receptor to identify IgG mutants with limited immune response. We sought to test the use of electrostatic complementarity as a design paradigm and present the design of a series of mutations that are predicted to disrupt binding by reducing complementarity. Overall, we find that coulombic and solvent mediated electrostatic interactions are critical in protein binding and that electrostatic optimization is effective as a protein design strategy. High electrostatic

complementarity is predictive of improved binding affinity, and low electrostatic complementarity is predictive of reduced binding affinity and activation.

In Chapter 3 we incorporate an additional level of complexity into our ensemble models and examine the effect of configurational entropy and free energy in the computational evaluation of a series of small–molecule, HIV-1 protease inhibitors. This work was motivated by a previous computational study that designed and evaluated the binding affinity of the same inhibitor series without accounting for configurational entropy effects [47]. Comparison with experiment showed poor correlation between this approximate screening method, and we sought to assess the differences and sources of the configurational entropy change for ligand degrees of freedom upon binding. Previous studies of the thermodynamics of ligand binding have suggested that the ensemble nature of ligand–receptor interactions play a key role in both their absolute and relative binding affinities [24, 25]. It is well known that ligands can adopt multiple conformations in their respective binding pockets [48] and lose a great deal of configurational freedom (e.g., translational, rotational, internal) upon binding [49, 50]. Similarly, proteins can undergo comparable losses in conformational freedom upon binding as active site residues free in the unbound state become constrained when interacting with the ligand in the bound state [51]. We explored the role of ligand conformational entropy using a novel, end–point approach that effectively enumerates the conformational space of the ligand in the bound and unbound states. This work employs a hierarchical energy evaluation scheme and also accounts for ensemble solvent effects via a PBSA continuum model. Free energies, enthalpies, and entropies of binding are explicitly computed by applying the exact Boltzmann weight to each conformation and taking the appropriate average over configurational space. We find that inclusion of ensemble effects results in relative binding free energies that correlate strongly with experiment and that the free energies are well defined by a small portion of the configurational space in both the bound and unbound states. Furthermore, we identify major sources of configurational entropy loss and explore the effect of entropic coupling between different ligand degrees of freedom upon binding via a novel, information theory based approach. As a whole, this work outlines

a generalized framework through which the role of ensemble binding effects due to both ligand and receptor can be explored.

In Chapter 4 we revisit ensemble based protein design and explore the utility of this design paradigm in the context of enzyme redesign. In so doing, we assess the efficacy of transition–state theory as a protein design framework and explore the dynamics of an enzyme catalyzed reaction using transition path ensembles. The goal of this work was to redesign and improve the specific activity of the enzyme ketol-acid reductoisomerase (KARI) by stabilizing the transition state of the isomerization/alkyl migration reaction that it catalyzes. There have been a number of recent successful studies presenting both the redesign and *de novo* design of enzymes [52, 53, 54, 55, 56] using design schemes based on classical transition–state theory. The fundamental hypothesis is that given a model of the transition state and reactant (i.e., ground) state one can improve activity by energetically stabilizing the relative free energy difference between the two states. We sought to test this design principle using static models of the transition and ground state in conjunction with an end–point, protein design scheme based on the work presented in Chapter 2 and Lippow *et al.* [19]. We computationally explore possible mutations that decrease the free energy difference between transition and reactant state models, select, and characterize a panel of mutants predicted to improve activity. Experimental measurements of the temperature dependence of the rate of reaction indicate that our designs appear successful at reducing the enthalpy of activation relative to wild type, but unfavorably affect the entropy of activation and counteract the enthalpic improvement. In order to fully capture the effect these mutants have on catalysis, we analyze the rates and dynamics of these mutants using transition path sampling and explore the ensemble of reactive trajectories that connect reactants to products. This ensemble based method yields relative rates that are in qualitative agreement with experiment and reveals a unique reaction mechanism for both wild type and mutant variants based on the vibrational activation of the breaking bond. Overall, this work presents a critical analysis of the utility of both single structure transition state based design as well as transition path ensembles in deciphering enzymatic, reaction mechanism.

Collectively, this work examines the use of ensemble, free energy models in the study and design of proteins and small–molecule therapeutics. We present novel, predictive, ensemble methods that are used to address both basic science and engineering questions in ligand and protein binding as well as enzyme catalysis. We assess the effect of solvent mediated interactions in protein–protein binding and their application in antibody design, the import of configurational entropy in ligand binding, the efficacy of transition–state theory in enzyme design, and the role of dynamics in enzymatic reactions. These studies demonstrate the utility of computational analysis in the design and study of biophysical macromolecular systems, and as methods improve, we believe that computational methods become increasingly important in the design of effective protein and small molecule therapeutics.

Chapter 2

Computational Analysis and Rational Design of IgG:Fcγ Receptor Binding [1,2]

Abstract

In this study we examine the the binding interaction of two immunoglobin G (IgG):Fcγ receptor complexes, human IgG_1:FcγRIIA and mouse IgG_{2a}:FcγRIV, using molecular mechanics, Poisson–Boltzmann surface area (MM-PBSA) models to assess the role solvent screened electrostatics play in binding. In the human IgG:FcγRIIA system, we present the experimental, rational design of the Fc (fragment crystallizable) region to remove the conserved, N-linked glycan at residue 297 while retaining affinity to Fcγ receptors. Our results indicate that removal of the N297 glycan reduces or knocks out binding to almost all Fcγ receptors, but that the two mutations, S298G and T299A, can recover binding to FcγRIV. Using homology models of the wild type (WT) as well as aglycosylated, S298G/T299A mutant, we explore the role these mutations and the glycan play in modulating binding affinity. We find that the Fc glycan has limited, direct engagement with the receptor, but can modulate affinity via indirect

[1] All experimental work presented in this chapter was performed by collaborators in the laboratories of K. D. Wittrup at the Massachusetts Institute of Technology and J. V. Ravetch at Rockefeller University. Specifically, S. Sazinsky and R. G. Ott performed the original experimental $hIgG_1$ mutant screen and subsequent validation, and T. Chen constructed and performed affinity measurements on the designed $mIgG_{2a}$ mutants.

[2] Portions of this chapter have been published as: Sazinsky S. L., Ott R. G., Silver N. W., Tidor B., Wittrup K. D., and Ravetch J. V. Aglycosylated immunoglobulin G_1 variants productively engage activating Fc receptors. *Proc. Natl. Acad. Sci. U.S.A.* 105:20167-20172 (2008).

electrostatic shielding of the essential D265/K117 salt bridge. We also find that the S298G/T299A mutant is able to rescue binding in part through a favorable reduction in desolvation penalties paid upon binding and improvement of electrostatic complementarity at the contact surface. In the mouse IgG_{2a}:FcγRIV system, we further explore the electrostatic complementarity of both binding partners via the computational, negative design of the Fc region to knock out binding to and activation of FcγRIV without seriously affecting stability. We find that 5 of the 6 mutants that were designed and experimentally characterized successfully abrogate binding relative to WT. Analysis of the predicted decrease in affinity reveals that by increasing the positive charge at specific locations along the binding interface, either by removing negative charge or adding positive charge, one can further reduce electrostatic complementarity and knock out binding.

2.1 Introduction

The binding of immunoglobin G (IgG) to Fc gamma receptors (FcγR) is a critical step in the activation and regulation of immune response. IgG molecules, once bound to their target epitope, are recognized by a host of FcγRs through specific interactions with the conserved, Fc (fragment crystallizable) region of the antibody. Sequence and structural variability in this region determines the relative binding affinity of antibodies to each of the three human FcγR (hFcγR) subtypes (FcγRI, FcγRII, and FcγRIII), and ultimately, cell-mediated effector function [57, 58]. In both humans and mice, it is known that antibodies with high affinity to FcγRI and FcγRIII (and FcγRIV in mice) activate signaling pathways and result in cell-mediated or complement-dependent cytotoxicity. By contrast, those that interact with FcγRII engender an inhibitory response, which prevents activation [59, 60, 57]. Much of the effort to develop effective antibody therapeutics has focused on selective induction of these responses by modulating the binding interaction between the Fc region and FcγRs through Fc mutagenesis [36, 61] and glycoengineering [62, 63]. In the case of anticancer therapeutics, optimizing for FcγRI/III affinity has been shown to improve cytotoxicity toward targeted cancer cells [64]. Similarly, therapeutic antibodies involved in treatment of autoimmune disorders such as rheumatoid arthritis, psoriasis, and transplant rejection have focused on inhibiting immune response via modulation of FcγRII activation [65] or minimizing FcγR activation all together [66]. For example, in the case of immunosuppresive therapies that target CD3 and modulate T-cell activity, antigen binding alone dictates efficacy, and target activation is both unnecessary and unwanted [67, 68]. Identification of specific Fc residues important to FcγR specificity and affinity is critical to the development of effective, tunable antibody therapeutics.

Critical to FcγR activation is the glycosylation of the CH2 domain of the Fc region via an N-linked glycosylation of asparagine 297 (N297). Antibody–receptor affinities are highly sensitive to the specific glycoform attached at N297, and removal of this glycan through point mutations at position 297 [36, 37], enzymatic Fc deg-

lycosylation [38], or expression in prokaryotes [39, 40] results in reduced affinity to all FcγRs and abrogation of binding for low affinity IgG/FcγR complexes [41, 42]. From a therapeutic antibody production perspective, this glycan dependency makes controlled production of antibodies both expensive and difficult, as production strains are limited to mammalian cell lines, which often generate heterogeneous antibodies with a wide variety of glycoforms, making purification difficult [41].

In this study we explore the binding interactions of two IgG:FcγR complexes, human IgG_1/FcγRIIA^{131R} and mouse IgG_{2a}/FcγRIV through a combination of experimental and computational, structure based design. In the case of the former, we present an experimental site directed mutagenesis study of the Fc region that identified mutations (S298G/T299A) capable of productively engaging FcγRs in absence of glycosylation *in vitro*. We computationally examine the effect that these mutations and the glycan have on the binding interaction using structural, homology models, and evaluate their respective, energetic contributions to affinity using a molecular mechanics, Poisson–Boltzmann surface area (MM-PBSA) model. Our homology models are constructed using the efficient, dead-end-elimination (DEE) [69, 70, 71, 72] and A* algorithms [73, 74] to find low energy binding conformations, the individual interactions of which are evaluated using an MM-PBSA based component analysis. Our models of the Fc:FcγR bound complexes suggest that electrostatic interactions play an essential role in determining relative binding affinity. We find that the N-linked glycan present in the hFc has limited interaction with hFcγRIIA and plays an indirect role in determining binding affinity. Component analysis of the bound Fc/FcγRIIA complex suggests that the two oligosaccharides have limited electrostatic as well as van der Waals (VDW) contact with FcγRIIA and are primarily interacting with their respective Fc chains. However, they shield an important intermolecular salt bridge between D265 (Fc) and K117 (FcγR) that is predicted to become stronger in the absence of the large glycan. Our models also predict that the S298G/T299A hFc mutations augment FcγRIIA binding by reducing the desolvation penalty paid by these residues upon binding, which improves the electrostatic complementarity at the protein–protein interaction surface.

In the case of mouse IgG:FcγR, we present the computational, negative design of the mouse Fc region to find specific, affinity tuning mutations that abrogate binding to FcγRs without affecting stability. Using similar methodology to that used in the development of the human IgG:FcγR homology model, we leverage the DEE/A* search scheme to explore all possible, single residue interface mutations to selectively disrupt binding. By selecting mutants that disrupt the binding interaction without affecting folding stability, we find a number of single mutations that are predicted to knock out binding. We report here on the design of these mutants as well as their experimental validation. In addition, we use an interaction free energy component analysis to provide energetic hypotheses for the mechanism of inhibition for each. Interestingly, we observe the opposite complementarity principle, specifically, by mutating away specific interactions and reducing electrostatic complementarity, one can disrupt binding without affecting stability. In our negative design of the mFc domain, we see that packing the binding interface with positively charged residues successfully abrogates Fc:FcγRIV binding. Taken together, we find that by rationally designing mutations in the Fc region to make the binding interface more or less complimentary in electrostatics, one can both positively and negatively modulate the binding free energy. Furthermore, we find that this MM-PBSA model combined with a DEE/A* configurational search strategy is predictive of relative, *in vitro* protein–protein binding affinity.

2.2 Methods

2.2.1 Computational Design of hIgG$_1$:FcγRIIA and mIgG$_{2a}$:FcγRIV Homology Models

The starting point for both human and murine homology models was the crystal structure of the extracellular portion of the human FcγRIIIB receptor bound to the Fc region of the human IgG$_1$ (PDB ID 1E4K) [75]. Hydrogen atoms were placed using the CHARMM computer program [3, 76], CHARMM27 force field [77], and HBUILD module [78]. The side chains of H116(C) and H131(C) on the receptor were flipped by 180° around χ_2 and treated in their neutral, ϵ-protonated form. In the Fc fragment, all histidine side chains were neutral and protonated as indicated, to maximize hydrogen bonding potential: 268(A)-δ, 268(B)-ϵ, 285(A)-δ, 285(B)-δ, 310(A)-δ, 310(B)-ϵ, 429(A)-δ, 429(B)-δ, 433(A)-δ, 433(B)-δ, 435(A)-δ, and 435(B)-δ. Homology models of the human FcγRIIA as well as murine FcγRIV complexes were constructed on the FcγRIIIB backbone as follows: all non-alanine, non-glycine residues further than 5 Å from an interface residue were replaced by alanine; both glycosylated and aglycosylated forms of the structure were prepared; and in the glycosylated structure, a sliding, restrained harmonic minimization was performed on the side chain of the N-glycosylated N297(B). Note that the non-interfacial alanine mutations were performed to allow an unbiased placement of the new sequence. Partial atomic charges for the N-glycosylated N297(C) residues were derived by fitting to the electrostatic potential using the restrained fitting methods of Bayly *et al.* [79] for each monosaccharide. The charges associated with hydrogens missing in the polysaccharide were added to their parent atoms to ensure charge conservation.

To generate either the FcγRIIA^{R131} or FcγRIV receptor structure, all FcγRIIIB interfacial residues were mutated to their associated, structural counterparts using the dead-end elimination and A* protocol described by Lippow *et al.* [19]. Similarly, mutant hFc or mFc was generated using DEE/A* in the presence of their respective binding partners. For each mutant sequence, the global minimum energy

conformation, as well as a collection of progressively higher energy conformations, was identified in the context of discrete rotameric conformational freedom of all placed side chains, except for the glycosylated form of N297. All of the Fc mutants examined were generated in the presence of the receptor during this conformational search. Note that one interfacial residue in the linker region of the FcγR structure (E86 in the FcγRIIA sequence) was left as a glycine, as all glutamate rotamers searched had a van der Waals clash with the receptor backbone. In the unbound FcγRIIA crystal structure [80], the two domains of the receptor separate slightly to accommodate this larger residue. The lowest energy structures of each sequence found via the DEE/A* aided search of configurational space were re-evaluated using a molecular mechanics, Poisson–Boltzmann Surface Area (MM-PBSA) based model, as described in Section 2.2.3.

2.2.2 Rational Design Protocol of mIgG$_{2a}$:FcγRIV Mutants

To identify Fc mutations that inhibit the binding and subsequent activation of murine FcγRIV, we performed a computational, enumerative, single mutant screen of all residues within 5 Å on both Fc chains (A/B). All possible amino acids except for proline, including all three possible protonation states of histidine (δ,ϵ, and the positively charged tautomer), were evaluated as possible mutations at each position, yielding a total of 456 possible single mutants. Due to the asymmetric contact of FcγR with the Fc chains, 16 of the 24 total positions examined were on the B chain. Similar to the protocol used in creating both homology models, possible Fc variants were evaluated using a DEE/A* methodology to explore sequence and conformational space of each mutation. Minimum energy conformations of each variant were also ranked using the same MM-PBSA model. Each single mutant was evaluated based on its contribution to the binding interaction with FcγRIV as well as its contribution to the stability of the folded Fc region. In order to knock out binding, the mutant list was screened for those mutations that incurred large binding penalties relative to WT (i.e. $\Delta\Delta G_{\text{bind}} \geq 1$ kcal/mol) while minimizing unfavorable folding interactions (i.e. $\Delta\Delta G_{\text{fold}} \leq 0$ kcal/mol). Additionally, mutants were chosen in order to max-

imize possible cooperative effects by selecting for multiple different mechanisms of inhibition.

2.2.3 MM-PBSA Electrostatic Model

All solvent-screened electrostatic interactions, desolvation penalties, as well as the residual electrostatic potential upon binding for these structural models were computed by solving the linearized Poisson–Boltzmann equation as described by Lee and Tidor [81] using a locally modified version of the finite difference solver, DELPHI [82, 83, 84, 85], with PARSE radii and charges [86]. A dielectric of 4 was used for all internal, receptor, ligand, and explicit water regions, while a dielectric constant of 80 was used for all implicit solvent regions. A salt concentration of 0.145 M was used along with a Stern layer of 2.0 Å and a 1.4–Å probe to determine the dielectric boundary. When solving for the potential, an average value was obtained using the potential collected from ten translations of a 129 x 129 x 129 cubic grid. The grid error associated with this average was on the order of 0.01 kcal/mol, which is much smaller than the interaction and desolvation energies reported here. To obtain more accurate potentials, a focusing procedure was used: initial calculations used a 23% fill with Debye-H"*u*kel boundary conditions followed by higher resolution calculations using a 92% fill. Component interactions or desolvation penalties were computed based on the work of Hendsch and Tidor [87]. We make use of the linearity of the model and the superposition principle to decompose the potential and free energy into components for each charged side chain and backbone chemical group. For these specific interactions, the focusing procedure included an additional 184% fill centered on the specific component being examined.

The free energy of desolvation of hydrophobic residues was modeled using a linear expression (Eq. 2.1) proportional to the solvent accessible surface area (SASA).

$$\Delta G_{\text{desolv}}^{\text{hydrophobic}} = \alpha * SASA + \beta \tag{2.1}$$

A proportionality constant, α, of 5 cal/mol/Å was used [86]. Given that all en-

ergies were computed relative to wild–type, the constant term, β, always canceled and was ignored. All van der Waals and covalent energies were computed using the CHARMM27 force field with no nonbond cutoffs.

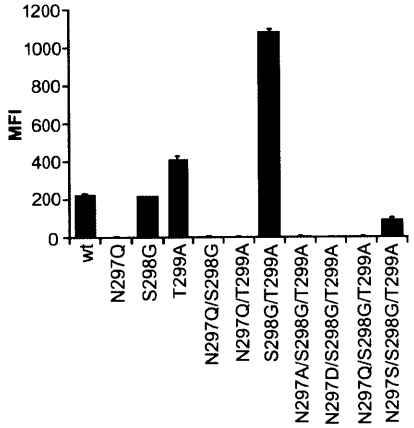
2.3 Results and Discussion

2.3.1 Experimental Screen for Aglycosylated Fc Variants that Bind Fcγ Receptors

The importance of glycosylation for IgG_1/FcγRIIA binding was assessed by experimentally constructing saturation mutagenesis libraries at positions 296 through 300 in the CH2 domain of $hIgG_1$ and screening via a yeast cell surface display system [88, 89]. $hIgG_1$ variants were displayed on yeast cell surfaces and their binding to fluorophore-labeled tetramers of FcγRIIA using fluorescence-activated cell sorting (FACS). After three rounds of FACS screening, three variants were identified that lacked the canonical, Asn-X-Ser/Thr, N-linked glycosylation motif: S298G/T299A, S298G/T299G, and T299A. Affinity measurements of each single/double mutant show that either S298G or T299A alone is able to engage and modestly improve binding to FcγRIIA, but that in combination they greatly improve affinity (Fig. 2-1a). Interestingly, this double mutant is unable to rescue binding in a N297A/D/Q/S mutant background. Affinity measurements (assessed by measuring the K_d of the bound complex) of WT, and mutants N297Q and S298G/T299A show differential binding to different FcγRs (Fig. 2-1b). The single mutant experiences a 10 fold drop in affinity to the R1a variant relative to WT and abrogated binding to all other measured receptors. The aglycosylated double mutant also sees a 10 fold drop in affinity to R1a, but is able to bind to the RII receptor variants with near WT affinities. Specifically, the double mutant experiences a 3- and 2-fold improvment in affinity relative to WT when binding with $RIIA^{131R}$ and RIIB, respectively, and a small drop in affinity with $RIIA^{131H}$ (80% of WT affinity).

2.3.2 Models of Human IgG_1/Fc$\gamma$$RIIA^{131R}$ Interaction

To explore the structural basis for FcγR binding of aglycosylated Fc domain variants, we constructed homology models of Fc:FcγRIIA complexes based on the previously solved structures of the IgG_1 Fc, the FcγRIIA structure [80], and the Fc:FcγRIII



(a)

Affinity: K_D (x10^{-6} M)

Variant	RIa	RIIA^{131H}	RIIA^{131R}	RIIB	RIIIA^{176F}	RIIIA^{176V}	C1q
wt	0.04	5.5	5.0	9.8	13	4.6	0.3
N297Q	0.4	n.b.	n.b.	n.b.	n.b.	n.b.	n.b.
S298G/ T299A	0.3	7.0	1.7	5.7	n.b.	n.b.	n.b.

(b)

Figure 2-1: **Experimental hIgG$_1$:FcγR Binding Affinity Measurements** (a) Binding of hIgG1 Fc variants to fluorophore-labeled FcγRIIA tetramers, measured by median fluorescence intensity (MFI) of the labled receptor during FACS. Intensities shown here are the average of two trails. (b) Dissociation constant (K_d) measurements of aglycosylated Fc mutant:FcγR/complement component C1q complexes. n.b., no binding detected.

complex [75] (Fig. 4). Four features emerge from this modeling. First, in the model of the WT interaction, there is only limited interaction between the two N-linked glycans and FcγRIIA (Fig. 2-2). The asymmetric nature of the IgG_1 Fc:FcγRIIA interaction predicts that the glycan attached to the B chain of the Fc dimer may interact with residues K117, T119, F121, S126, and F129 of the receptor, whereas the glycan attached to the A chain does not make contact with FcγRIIA. These glycan–FcγR contacts provide negligible screened electrostatic intermolecular interactions in our calculations, compared to the much larger intramolecular ones between glycan and Fc, roughly –1.3 kcal/mol, with a dominant contribution from N297/glycan(B)-D265(B). Analysis of van der Waals interactions yields similar results. Both oligosaccharides are predicted to make favorable interactions worth approximately 6 kcal/mol with FcγRIIA, the majority of which is due to the B chain glycan, while the intramolecular glycan/Fc-fragment interactions are worth over 40 kcal/mol. These data suggest that both oligosaccharides are primarily interacting with their respective Fc chains and make limited contact with the Fcγ receptor.

The second, emergent feature of this model is the importance of N297 to the Fc:FcγRIIA interaction. Aglycosylated N297 has the potential to make hydrogen bond interactions across the interface with S126 of the receptor (Fig. 2-3). These interactions may be mediated by a bridging water molecule that can be observed nearby in an unbound FcγRIIA crystal structure [80]. Replacement of N297 with glutamine or alanine disrupts this interaction (and fails to make similar, stabilizing ones) and is consistent with the observed absence of binding for such mutants (see Fig. 2-1a). Interestingly, replacement with aspartic acid may be able to make a similar interaction; however, the greater desolvation penalty of the charged side chain upon FcγR binding likely results in the reduced binding of this variant.

The third feature that results from our Fc/FcγR model is the favorable, indirect effect the loss of the glycan has on the electrostatic binding interactions in the S298G/T299A double mutant. Both our glycosylated WT and aglycosylated mutant models predict an intermolecular salt bridge between D265 on the B chain of the Fc dimer and K117 on the FcγR. In the WT structure, this interaction is shielded from

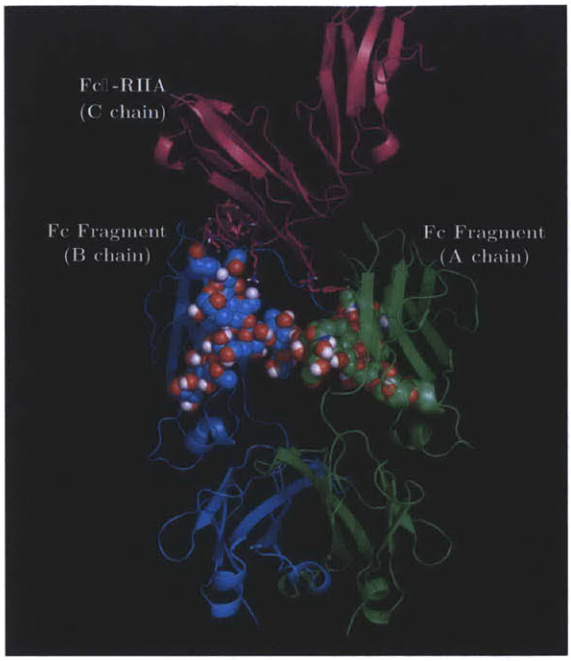


Figure 2-2: **Structure of the WT Fc fragment bound to the constructed Fc-γ-RIIA^{131R} homology model.** The portion of Fc-γ-RIIA (chain C) highlighted as purple sticks shows those side chains within 5 Å of the two glycans.

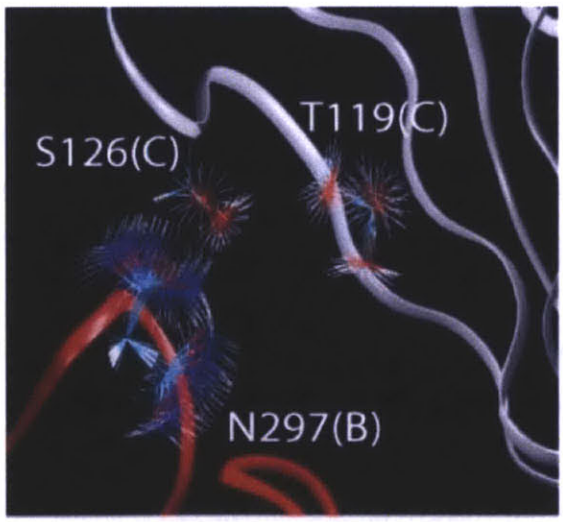


Figure 2-3: **Structure of intermolecular aglycosylated N297(B) interactions** Shown here are possible rotameric states of N287(B) interacting with possible rotameric states of T119(C) and S126(C) in the aglycosylated, S298G/T299A model.

solvent by the oligosaccharide chain. In the aglycosylated S298G/T299A mutant, this salt bridge is exposed to the solvent (Fig. 2-4), which nearly halves the screened electrostatic interaction energy compared to WT (~5 kcal/mol versus ~10 kcal/mol). However, this effect is more than compensated in the S298G/T299A mutant by a reduction in the desolvation penalty [81], a measure of the loss of electrostatic interactions with solvent upon binding, paid by both salt bridge partners. Specifically, the desolvation penalty of the mutant drops by about 7 kcal/mol, including 5 kcal/mol from D265 and nearly 2 kcal/mol from K117. Therefore, the net effect of this deshielding provides stabilization worth about 2 kcal/mol. This effect is illustrated (Figs. 2-5a and 2-5b) by a reduction in the residual electrostatic potential present on D265(B) in the mutant compared to the WT, where the residual potential is defined as the desolvation potential of the IgG Fc region plus the interaction potential due to the FcγR.

Finally, our model predicts that the net increase in affinity of the aglycosylated S298G/T299A mutant to Fcγ-RIIA^{131R} relative to WT is electrostatic in nature and due to both the favorable deshielding of the D265/K177 salt bridge (*vide supra*) and the favorable change in relative desolvation of residue 298 upon mutation from serine to glycine, which contributes approximately 2 kcal/mol to the stability of the mutant complex. Figs. 2-5a and 2-5b illustrate this change via a decrease in the residual electrostatic potential present on residue 298 in the mutant compared to WT. Here, a decrease in the residual potential indicates an increase in the electrostatic complementarity of the binding partners, as the serine at position 298 was not making strong electrostatic interactions with the receptor but paid a desolvation penalty upon binding. Note that while T299A is also a polar to non-polar mutation, the threonine side chain is highly solvent exposed in the WT complex, and pays effectively no desolvation penalty upon binding. In total, the predictions made by this homology model highlight the importance of electrostatic interactions in Fc/FcγR binding and provide a hypothetical mechanism for the stability of the aglycosylated Fc:FcγR complex, resulting from hydrogen bonding and electrostatic interactions altered in the aglycosylated mutant.

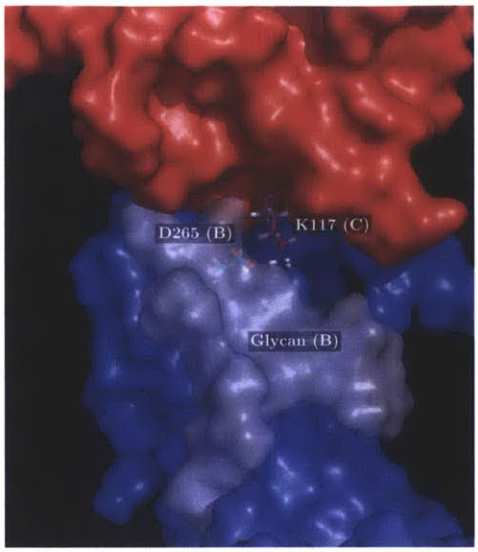
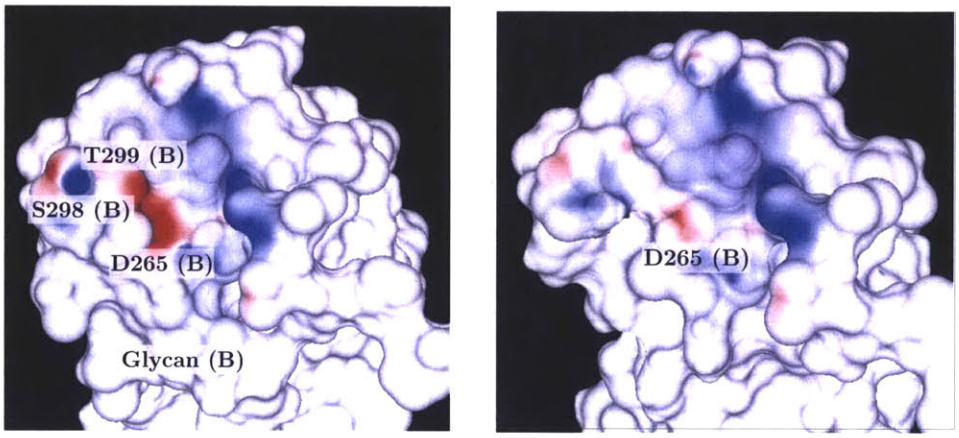


Figure 2-4: **Structure of the WT K117(C)-D265(B) salt bridge**. Both interacting partners are enclosed within the binding cavity partially covered by the glycan (shown in transparent grey). The S298G/T299A mutant lacks the glycan, exposing the the K117(C)-D265(B) interaction to solvent.

2.3.3 Computational Design of Mouse IgG_{2a}/FcγRIV Knockout Mutations

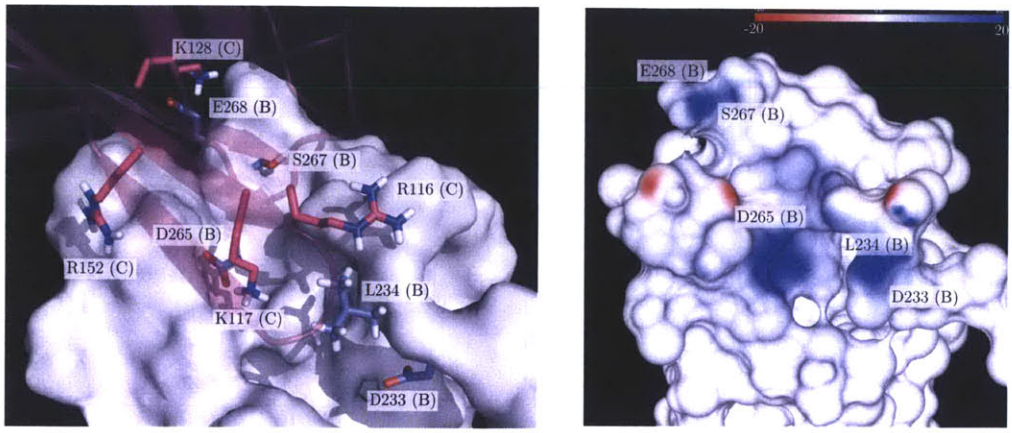
Of the 456 mutants in our exhaustive virtual screen, six were selected based on their predicted ability to reduce binding affinity without affecting stability. All six mutants can be organized into two categories: those mutations that affect binding via removal of favorable electrostatic contacts (D265A and E268I/A) and those that affect binding via introduction of unfavorable electrostatic contacts (L234K, S267R, and D233R) (Fig. 2-6a). In almost all cases the mutations are predicted to operate by increasing the total positive charge at the interface near regions of large positive interaction potential (i.e. the potential felt on the Fc surface due to charges on FcγRIV), either by removing negative charge and/or by adding positive charge (Fig. 2-6b). This addition or subtraction has an unfavorable effect on the free energy of binding, but has either a neutral or favorable effect on the free energy of folding for the complex (Table 2.1). The first group of mutations, D265A and E268I/A, are all predicted to lose approximately 1 kcal/mol of binding free energy relative to WT with negligible



(a) (b)

Figure 2-5: **Mutant and WT Residual Potentials** (a) Residual potential after binding mapped onto the interaction face of the B chain of the Fc fragment in the WT structure. White regions indicate areas of ideal complementarity between the Fc fragment and Fcγ-RIIa^{131R}, while deep red or blue regions indicate areas of poor complementarity because of ligand desolvation costs uncompensated by interactions made upon binding. Red corresponds to negative residual potential potentials and blue to positive residual potentials. The coloring scale, identical in both panels, covers the range of +20/-20 kT/e. (b) Residual potential after binding mapped onto the interaction face of the B chain of the Fc fragment in the S298G/T299A mutant structure.

or slightly favorable effects on folding stability. In the case of D265A, the bulk of this affinity change can be traced to the loss of favorable electrostatic interactions of the B chain residue, D265(B), with its salt bridge partner on FcγRIV, K117(C), as well as neighboring R152(C). Nearly 10 kcal/mol of interaction free energy is lost, with a concomitant, favorable 9 kcal/mol reduction in the desolvation penalties paid by both partners upon binding, a net loss of 1 kcal/mol. The E268I mutation is also predicted to lose affinity from the removal of a favorable intermolecular salt bridge, in this case between E268(B) and K128(C), which results in an unfavorable 2 kcal/mol decrease in interaction energy and favorable 1.4 kcal/mol reduction in desolvation penalty, a net loss of 0.6 kcal/mol. Finally, in the case of E268A, which also disrupts the E268(B)-K128(C) salt bridge, the computed affinity decrease stems primarily from a loss of van der Waals contact between the two residues. In contrast to E286I, the electrostatic change is predicted to stabilize binding by nearly 1 kcal/mol due to the increased solvent exposure of K128(C) in the bound complex. This reduction of the desolvation penalties paid by residues 268(B) and 128(C) more than compensates for the loss in interaction energy, resulting in a net favorable affect on the free energy of binding. This component analysis suggests that success of these designed knockout mutations hinges on their ability to disrupt specific, hot-spot interactions with positively charged residues on the receptor interface by removing negatively charged salt-bridge partners, which in all mutants aside from E268A also reduces electrostatic complementarity. Examining the desolvation and interaction potentials projected onto the Fc binding surface of WT, D265A, and E268I, we can visualize this reduction in complementarity through the loss of negative potential that the Fc face presents to the FcγR (Fig. 2-7). It is interesting to note that even in WT, both of these salt-bridge interactions (D265(B):K117(C) and E268(B):K128(C)) have a net unfavorable affect on binding due to the relative difference of their desolvation costs and electrostatic interaction gains. Both salt bridges pay more in desolvation penalties than they get back in interaction energy by +2.6 kcal/mol and +1.7 kcal/mol, respectively, consistent with the hypothesis that salt bridges are more important for engineering specificity [90]. The variants found here further penalize the interactions by removing one of the



(a) (b)

Figure 2-6: **mIgG$_{2a}$/FcγRIV homology model**. (a) WT mouse IgG$_{2a}$/FcγRIV binding interface. The Fc B chain is shown as a white surface with mutation sites highlighted in blue sticks, and FcγRIV is shown in transparent purple ribbons with positively charged interface residues shown in purple sticks. (b) The interaction potential of FcγRIV projected on the Fc binding surface. Red corresponds to negative interaction potentials and blue to positive interaction potentials. The coloring scale covers the range of –20/+20 kT/e.

charge pair partners, which zeros the favorable interaction energy and leaves only unfavorable desolvation penalties.

The second group of mutations, L234K, S267R, and D233R are predicted to lose roughly 5, 1.4, and 1 kcal/mol of binding free energy, respectively, with favorable relative folding stabilities compared to WT. Our model predicts that each mutant reduces binding affinity by introducing repulsive electrostatic interactions with the receptor and/or increasing desolvation penalties upon binding. L234K incurs a 4 kcal/mol penalty relative to WT via unfavorable interactions with R116(C) and K117(C) ($\sim$2 kcal/mol) as well as unfavorable increases in the desolvation penalty paid by itself and residue K117(C) ($\sim$2 kcal/mol). Similarly, while the net affinity change is not as drastic (due to van der Waals gains), S267R pays large electrostatic penalties due to repulsive interactions with K128(C) (1.2 kcal/mol) and increased desolvation of both itself and neighboring Fc residue D269(B) (approx. 2.6 and 0.8 kcal/mol, respectively). Finally, in the case of D233R the loss of binding affinity is predicted to

		Total	Elec	vdW	Geo	SASA
D265A	Bind	1.3	0.8	0.4	0.0	0.1
	Fold	0.0	−1.0	2.9	−2.1	0.2
E268I	Bind	1.1	0.7	0.4	0.0	0.0
	Fold	−0.8	−0.6	−0.3	0.2	−0.1
E268A	Bind	1.2	−0.9	1.8	0.0	0.3
	Fold	−0.1	0.1	0.4	−0.7	0.1
L234K	Bind	5.0	4.6	0.5	0.0	−0.1
	Fold	−2.0	1.4	−2.4	−0.5	−0.5
S267R	Bind	1.4	4.3	−2.6	0.0	−0.3
	Fold	−3.3	−0.4	−2.8	0.3	−0.4
D233R	Bind	1.0	1.2	−0.2	0.0	0.0
	Fold	−5.4	−0.8	−3.4	−0.8	−0.4

Table 2.1: **Computed Folding and Binding Energies of mIgG$_{2a}$ Fc mutants to FcγRIV Relative to Wild Type**. Elec values are solvent screened electrostatic energies, vdW values are pairwise van der Waal energies, Geo values are covalent strain energies, and SASA values are solvent accessible surface area energies. All values are reported in kcal/mol relative to wild type.

result primarily from long range, action–at–a–distance, electrostatic repulsion with R116(C) and K117(C) (2 kcal/mol). Interestingly, this position faces away from the receptor and is charged in the WT structure; as such, it desolvation changes are negligible. Thus, in contrast to the charged to neutral mutations (D265A, E268I/A), we find that these mutations modulate affinity primarily by adding positive charge at the interface. The effect on complimentarity, however, is similar. We see this effect illustrated for the L323K and S267R mutants in Fig. 2-8. Positive potentials appear in the desolvation projection at each mutation site, indicating that the Fc face pays larger desolvation penalties upon binding and presents a positive potential to the FcγR where positive charges already exists.

These six mutants were experimentally characterized by measuring their affinity to mouse FcγRIV via yeast cell surface display [91] and FACS. Figure 2-9 shows the measured knock-down in affinity relative to WT. We see that five of the six mutations successfully reduce binding affinity by 10 fold to FcγRIV, all except E268I. This suggests that negative modulation of electrostatic complementarity at the binding

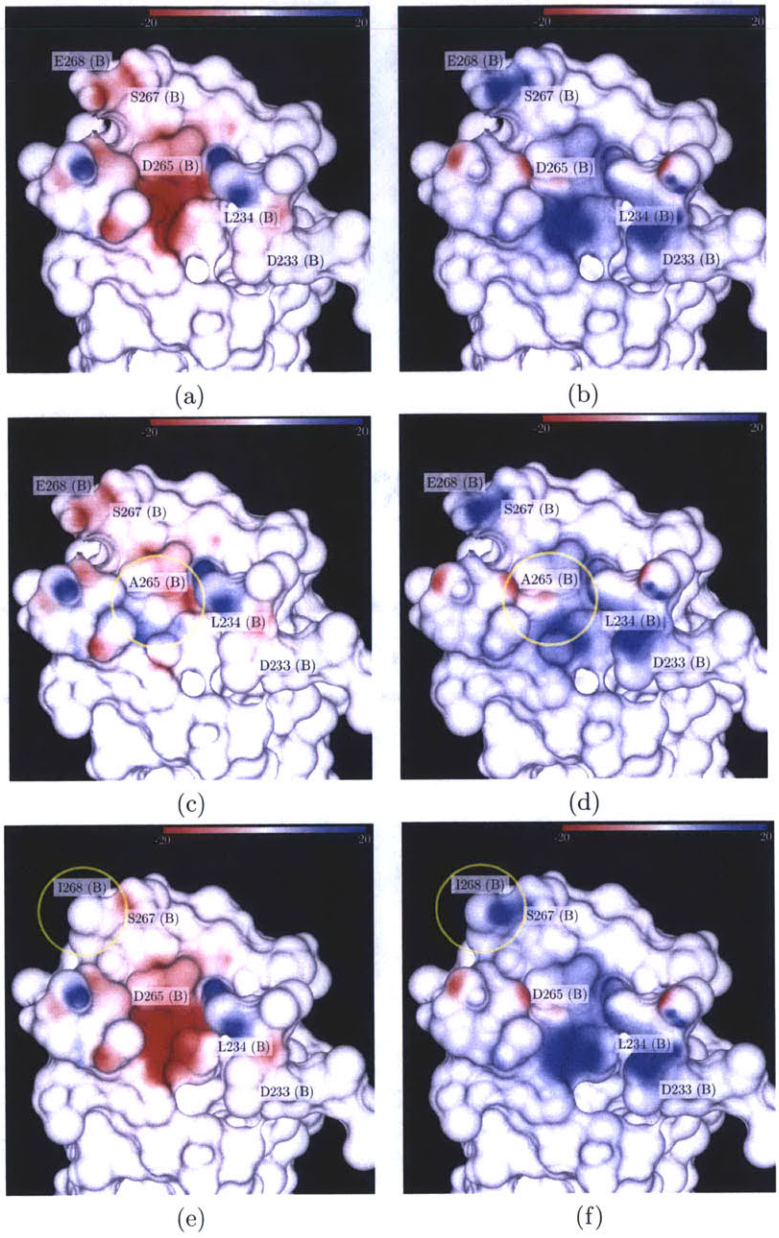


Figure 2-7: **mIgG$_{2a}$/FcγRIV D265A and E268I Mutant and WT Desolvation and Interaction Potentials**. (a) WT desolvation and (b) interaction potentials projected onto the Fc binding interface. The coloring scale, identical in all panels, covers the range of –20/+20 kT/e. (c) D265A desolvation and (d) interaction potentials (e) E268I desolvation and (f) interaction potentials.

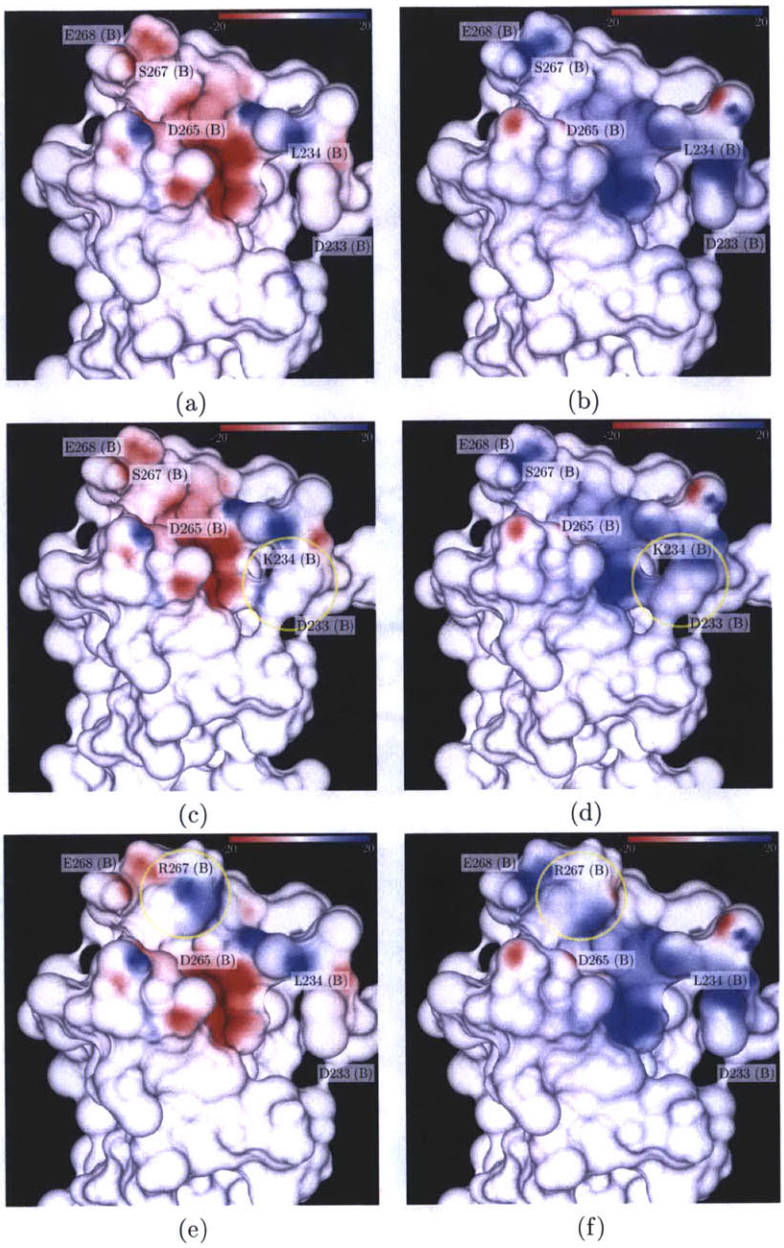


Figure 2-8: **mIgG$_{2a}$/FcγRIV L234K and S267R Mutant and WT Desolvation and Interaction Potentials**. (a) WT desolvation and (b) interaction potentials projected onto the Fc binding interface. The coloring scale, identical in all panels, covers the range of –20/+20 kT/e. (c) L234K desolvation and (d) interaction potentials (e) S267R desolvation and (f) interaction potentials.

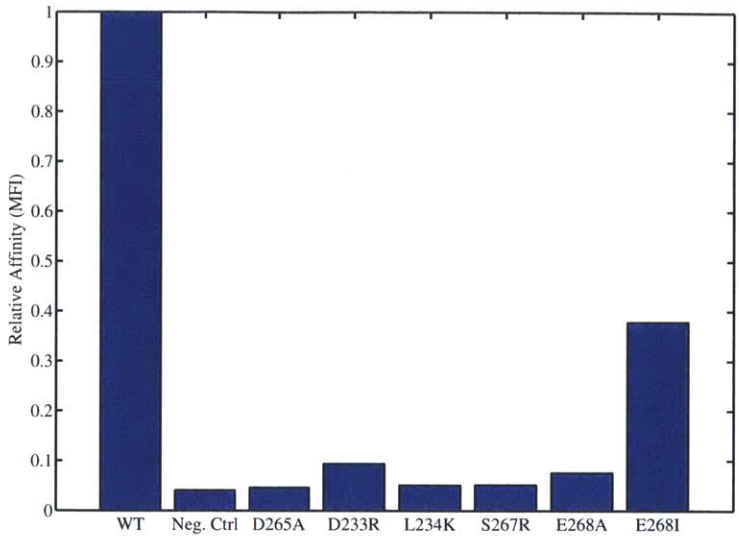


Figure 2-9: **Experimental Mutant mIgG$_{2a}$:FcγRIV Binding Affinities**. Relative binding affinity of mIgG$_{2a}$ Fc variants to fluorophore-labeled mFcγRIV, measured by median fluorescence intensity (MFI), relative to WT.

surface is an effective means of disrupting IgG:FcγR binding. We note that similar, positively charged mutations (G236R and L328R) were found by Lazar *et al.* [92] to inhibit binding to FcγRs. In our homology model, these residues are also at the surface, and neighbor K117 of FcγR. It is likely that these mutants also operate by increasing the desolvation penalty of the Fc region and/or introduce unfavorable positive–positive charge interactions.

2.4 Conclusion

In this study we have examined the binding of the Fc domain of IgG variants to both human FcγRIIA as well as mouse FcγRIV. In total, our results suggest that this ensemble, MM-PBSA implicit solvent model accounts for the important, solvent mediated effects essential to protein–protein binding. Additionally, they indicate that such models can be used both as an exploratory tool to probe the specific interactions important to binding as well as a predictive tool for design. In our examination of the human IgG:FcγRIIA complex, we presented structural models of the binding interaction for WT and rationally designed Fc mutants. Using energetic component analysis tools, we also examined the electrostatic contributions to binding from the N-linked glycan and aglycosylated S298G/T299A mutant. Our analysis suggests that the glycan is not directly involved in the Fc:FcγR interaction, as it makes limited VDW contact and has negligible solvent screened electrostatic interaction with the receptor. We do find, however, that it can indirectly affect binding by modulating the strength of charge-pair interactions inside the binding cleft. Energetic analysis of aglycosylated mutant suggests that the D265(B)-K117(C) salt bridge is in fact strengthened in the absence of the glycan due to a favorable decrease in the desolvation penalty paid by both residues upon binding. We also observe that the S298G/T299A mutation is able to rescue binding to FcγRIIA (relative to aglycosylated WT) due, in part, to the reduced desolvation penalty paid by residue 298 relative to WT. Both this as well as the desolvation change associated with glycan removal result in improved electrostatic complementarity at the binding site. We note that this strategy of optimizing for complementarity has also been used in the design of HIV-1 cell entry inhibitors [93], and observed in the analysis of lead progression of effective neuraminidase [94]. Our results indicate that this design principle is not limited to ligand design, and that it should be possible to use this approach to develop highly complimentary, tight-binding protein–protein complexes like those observed in nature [81].

In the case of the mouse Fc:FcγRIV complex, we have presented the successful,

computational design of Fc domain mutants that disrupt binding. We find that binding can be knocked out by increasing the net positive charge present at the binding interface. The mutations explored in this study are predicted to disrupt binding either by removing negatively charged residues at salt bridge sites (D265A, E268I/A) or by adding positively charged residues near regions where the FcγR already presented a large, positive potential (L234K, S268R, D233R). In the former case, loss of favorable interactions between charge pairs relative to favorable changes in desolvation results in loss of binding affinity. In the latter case, increases in desolvation penalties as well as electrostatic repulsion between positively charged residues across the interface reduce the free energy of binding. We find that for all knockout mutations that are predicted to affect binding via electrostatic changes (all but E268A), there is reduced electrostatic complementarity between binding partners relative to WT. In total, our analysis suggest that the binding affinity of the Fc region to Fcγ receptors is tunable based on the electrostatic complementarity of the binding interface.

Chapter 3

Efficient Computation of Small Molecule Configurational Binding Entropy and Free Energy Changes by Ensemble Enumeration [1]

3.1 Abstract

Here we present a novel, end-point method using the dead-end-elimination and A* algorithms to efficiently and accurately calculate the free energy, enthalpy, and configurational entropy of binding for ligand-receptor systems. We apply it to a series of human immunodeficiency virus (HIV-1) protease inhibitors to examine the effect ensemble re-ranking has on correlation with experiment as well as estimate the absolute and relative configurational entropy losses upon binding for structurally related inhibitors. Our results suggest that most thermodynamic parameters can be estimated using only a small fraction of the full configurational space, and we see significant improvement in correlation with experiment when using an ensemble versus static approach to ligand ranking. We also find that using approximate metrics based on the single conformation enthalpy differences between the global minimum energy configuration in the bound as well as unbound states also correlates well with experiment. Using a novel, additive entropy expansion based on conditional mutual information,

[1] All experimental affinity measurements presented in this chapter were performed by H. Cao, and the synthesis of all inhibitors was performed by A. Ali and K. K. Reddy, collaborators from the University of Massachusetts Medical School.

we also analyze the source of configurational entropy loss upon binding in terms of both uncoupled per degree of freedom losses as well as changes in coupling between inhibitor degrees of freedom. We estimate entropic free energy losses of approximately +24 kcal/mol, 12 kcal/mol of which stems from loss of translational and rotational entropy. Coupling effects contribute only a small fraction to the overall entropy change (1–2 kcal/mol), but suggest differences in how inhibitor dihedral angles couple to each other in the bound versus unbound states. The importance of accounting for flexibility in drug optimization and design is also discussed.

3.2 Introduction

One of the goals of rational drug design is to understand the thermodynamics of small-molecule–receptor binding in order to design effective, high-affinity therapeutics. Lead compound development is expensive and requires a great deal of experimental effort to explore the large combinatorial space of chemical functionality. To expedite the process, computational methods are often used to optimize the search and examine the binding thermodynamics of lead compounds. It is difficult, however, to compute accurately both the enthalpy (ΔH_{bind}) and entropy changes (ΔS_{bind}) upon binding and to rank compounds based on a true free energy of binding ($\Delta G = \Delta H - T\Delta S$). Most approaches based on physical force-fields include enthalpic binding contributions and perhaps solvent entropy contributions, but both are estimated from a single conformation. The neglect of configurational entropy changes for binding partners is a significant omission in common applications. Nonetheless, such calculations are valuable as they can provide a more accurate and detailed breakdown of the thermodynamic changes. Experimental methods such as isothermal titration calorimetry (ITC) [95] can only report on ensemble averaged binding enthalpies and entropies, and they cannot determine the source of the change (e.g. ligand, receptor, or solvent).

Computational approaches, in principle, account for the contributions to the free energy of binding from ligand, receptor, and solvent degrees of freedom. Standard molecular mechanical treatments of ligand binding separate the enthalpic changes into separate terms for internal, van der Waals, coulombic, and solvation interactions [3, 5]. Similarly, binding entropies are often decomposed into conformational entropy terms for the ligand, receptor, and solvent, but compared to the enthalpic terms, are much more time consuming to calculate. To accurately compute an ensemble free energy or entropy change upon binding, one must fully explore and integrate over the conformational space of the solvent, the ligand, its receptor, and the complex [96], which is a formidable task even for small systems. Given this difficulty, the configurational entropy change is often assumed to be the same for different ligands in a series or is approximated with an empirical term that assumes a constant change in

entropy per frozen rotatable bond [22]. However, these approaches have been criticized recently as they lack significant experimental evidence and theoretical support. Chang *et al.* calculated the change in configurational entropy of the clinically approved inhibitor amprenavir binding to HIV-1 protease (as $-T\Delta S$) and found it to oppose binding by ~ 25 kcal/mol, which differs significantly from empirical estimates [49]. It is still unknown, however, whether the configurational binding entropy change is similar for different, related protease inhibitors.

Modern computational methods that are used to compute free energies of binding and their component enthalpies and entropies generally fall into one of two categories, perturbation and end-point methods. The former includes free energy perturbation (FEP) [97, 98] and thermodynamic integration (TI) [99, 100, 101], which often rely on molecular dynamics (MD) or Monte Carlo (MC) simulations to perturb a system from one state to another (e.g. wild type to mutant, one ligand to another, unbound to bound). The total free energy change can then be computed as a function of the perturbation coordinate. While widely used, these methods are often very slow to converge, as there are significant challenges [102]. End-point methods determine free energy changes by calculating absolute free energies of the final and initial states of the system and taking the difference [96]. These absolute free energies can also be found via MD or MC simulations and have been successfully used to study ligand binding in a variety of molecular systems [103, 104]. Recent, alternative formulations make use of the single or predominant state approximation, in which a single or multiple low-energy structures are identified, and the local configurational space about each initial structure is sampled [105, 24]. Implicit in these methods is the assumption that high-energy conformations contribute negligibly to the ensemble entropy and enthalpy averages, and that the potential energy surface is well described using a single or set of local minima. Further approximations are often made to analytically integrate over local minima using the harmonic or quasiharmonic approximation [103, 106, 107]. The former assumes the potential energy surface about the initial structure can be modeled using a multidimensional harmonic potential, while the later also assumes that conformational fluctuations are governed by a multivariate Gaussian probability

distribution. While these methods are efficient, they are not guaranteed to search all of phase space and often predict the free energy change to be more favorable than it actually is [108].

This study seeks to evaluate a number of these assumptions using an ensemble, configurational free energy of binding to accurately rank computationally designed human immunodeficiency virus (HIV-1) protease inhibitors. Previous studies of the examined inhibitors have shown that using a single, low-energy configuration to evaluate each inhibitor can successfully predict binding geometries, but often fails to correctly rank inhibitors with binding free energies within 2–3 kcal/mol of each other [47]. We sought to improve upon this static, predominantly enthalpic treatment by accounting for ensemble effects both in the bound and the unbound state. To this end, we developed a novel, deterministic, end-point method for computing the free energy of binding of ligand–receptor complexes that uniformly searches conformational space and explicitly accounts for both enthalpic and configurational entropic effects. This approach fundamentally differs from the aforementioned methods in that it does not sample from a Boltzmann distribution of configurations to collect an average, but instead uses uniform, rotameric enumeration [109] of ligand torsional degrees of freedom to map out and explicitly integrate over the potential energy landscape. While normally an intractable problem, searching through this high-dimensional space is enabled through the use of the dead-end elimination (DEE) [69, 70, 71, 72] and the A* branch-and-bound algorithms [73, 74]. DEE is used to prune high-energy rotamers, which excludes low probability configurations from the search space, while A* is used to rapidly enumerate the accessible configurational states of the structure. Both of these algorithms are global optimizers, and when used in conjunction are guaranteed to both find the global minimum energy configuration (GMEC) and eliminate all those configurations with energies greater than a user supplied energy cutoff above the GME. Using this method, we were able to generate an energy-ranked, gapless list of low-energy ligand configurations in a computationally tractable amount of time, and evaluate the bound and unbound state partition functions to compute the free energy, enthalpy, and configurational entropy of binding in the context of a rigid

receptor.

The configurational entropy changes of all the protease inhibitors explored in this study were further analyzed using a novel, additive entropy expansion. By decomposing the entropy into a series of marginal entropy and mutual information (coupling) terms, we were able to extract the entropic contribution of each degree of freedom as well as the contributions from entropic coupling between pairs, triplets, and higher-order combinations. Similar entropy expansions have been described in the literature to examine the configurational entropies of liquids [110, 111], spin frustrated systems [112], as well as biological systems [113, 114]. However, given the aforementioned difficulty associated with effectively sampling the potential energy landscape of complex biological systems, previous applications to such systems have been limited to approximating the full entropy of the system. These methods assume that only a low-order subset of the entropy terms contribute significantly, as they are unable to accurately evaluate the remaining high-dimensional terms. Additionally, while these estimation attempts have been reasonably successful at describing the larger distribution, the individual terms are often difficult to physically interpret as they contain overlapping entropic contributions that are successively added and removed as the level of approximation improves [113]. The expansion used in this study is similar to that presented by Killian *et al.* and Matsuda [112], as it is based on the generalized Kirkwood superposition approximation [113, 115, 116], which approximates a high-order probability distribution using a series of successively lower-order distributions. It differs, however, in that each entropy term in the expansion is conditioned on the remaining degrees of freedom of the system, which aids in the physical interpretation of these terms by separating their contributions into non-overlapping pieces. Each term describes either the conditional marginal entropy of each degree of freedom or the conditional mutual information (coupling) between sets of degrees of freedom. By appropriately conditioning each term, these conditional couplings are measures of the coupling between degrees of freedom that are not mediated by another degree of freedom of the system, which avoids the layered, compensating additions and subtractions of the same physical effect present in other methods.

Applying this novel, conditional mutual information expansion (CMIE) and DEE/A* enumeration method on a series of protease inhibitors, we have been able to interpret configurational variation both within a given ensemble as well as between equilibrium ensembles in terms of specific thermodynamic changes. We can accurately evaluate the contribution of each marginal and coupling term to the full entropy as well as provide some insight into how physical coupling of degrees of freedom affects configurational entropy of binding. Our results analyze the efficacy of our approach by exploring thermodynamic convergence, comparisons with experimental measurements of binding affinity, and the role configurational entropy plays in binding. We find that our computed free energies correlate strongly with experiment, and that most thermodynamic averages are well defined by only a small portion of configurational space. Compared to previous computational studies of the inhibitors examined here [47], the enhanced sampling methods employed in this study provide better single conformation in the bound and unbound states as well as average enthalpy and entropy estimates that correlate with experimental free energy measurements. We also observe that each inhibitor loses a significant amount of configurational entropy upon binding, and that relative to each other, the entropic losses are significant (1-3 kcal/mol). Our entropy expansions show that the majority of both the absolute and relative entropic losses can be traced to changes in marginal, conditional entropy, and that changes in entropic coupling play a more subtle role in the thermodynamics of inhibitor binding.

3.3 Methods

3.3.1 Binding Theory

The theoretical framework for binding thermodynamics has been presented in recent literature [96, 50]; here we summarize the relevant portions to place our work in context. The standard free energy of binding for a ligand (L) and receptor (R) in solution can be evaluated using the standard chemical potential for each species:

$$\Delta G^{\circ}_{\text{sol,bind}} = \mu^{\circ}_{\text{sol},LR} - \mu^{\circ}_{\text{sol},L} - \mu^{\circ}_{\text{sol},R}. \tag{3.1}$$

The standard chemical potential for a dilute solution of ligand is defined as [117]

$$\mu^{\circ}_{\text{sol},L} = -RT \ln \left(\frac{1}{V_{N,L}\, C^{\circ}} \frac{Q_{N,L}}{Q_N} \right) + P^{\circ} V_L. \tag{3.2}$$

Here, $V_{N,L}$ is the volume of the system containing N solvent molecules and one ligand molecule. C° is the standard state concentration, assumed to be 1 M, which is equivalent to $1000\, N_A\, \text{m}^{-3}$, where N_A is Avogadro's constant. $Q_{N,L}$ and Q_N are the partition functions for systems containing N solvent molecules and one ligand molecule, and only N solvent molecules, respectively. The last term, $P^{\circ}V_L$, corresponds to the work associated with moving the ligand from the gas phase to a solvated state at constant pressure, where V_L is the volume of a single ligand and P° is the standard state pressure. This last term will be very small except at very high pressure, and in the present analysis it is assumed that binding occurs at 1 atm where this pressure–volume term will be negligible. The ratio of partition functions is expanded as follows,

$$\frac{Q_{N,L}}{Q_N} = \frac{\int\int e^{-\beta \left(\sum_i^{M_S+M_L} \mathbf{p}_i^2/2m_i + U(\mathbf{q}_S, \mathbf{q}_L) \right)} d\mathbf{p}_S d\mathbf{p}_L d\mathbf{q}_S d\mathbf{q}_L}{h^{3M_L} \sigma_L \int\int e^{-\beta \left(\sum_j^{M_S} \mathbf{p}_j^2/2m_j + U(\mathbf{q}_S) \right)} d\mathbf{p}_S d\mathbf{q}_S} \tag{3.3}$$

where $\beta = \frac{1}{k_B T}$, σ_L is the symmetry number of the ligand, $\mathbf{q}_{S/L}$ and $\mathbf{p}_{S/L}$ refer to the set of all position and momentum degrees of freedom of the solvent and ligand, respectively, and M_S and M_L define the total number of solvent and ligand atoms, respectively. U is the internal energy, T is the absolute temperature, h is Planck's constant, and k_B is Boltzmann's constant. This expression can be simplified by analytically integrating over the momentum portion of phase space (from $-\infty$ to $+\infty$) for each atom i of both the solvent and ligand ($\mathbf{p}_S$, $\mathbf{p}_L$) and cancelling the resulting expressions for the solvent momentum.

$$\frac{Q_{N,L}}{Q_N} = \prod_i^{M_L} \left(\frac{2\pi m_i k_B T}{h^2} \right)^{\frac{3}{2}} \frac{\int \int e^{-\beta U(\mathbf{q}_S, \mathbf{q}_L)} d\mathbf{q}_S d\mathbf{q}_L}{\sigma_L \int e^{-\beta U(\mathbf{q}_S)} d\mathbf{q}_S} \tag{3.4}$$

Further simplification is possible by defining a potential of mean force $W(\mathbf{q}_L)$ to make use of an implicit solvent treatment and avoid explicit integration over solvent degrees of freedom. This is done by defining the interaction potential between the ligand and the solvent for a fixed configuration of the system and averaging the Boltzmann factor of this potential over all solvent degrees of freedom.

$$U_{\text{int}}(\mathbf{q}_S, \mathbf{q}_L) = U(\mathbf{q}_S, \mathbf{q}_L) - U(\mathbf{q}_L) - U(\mathbf{q}_S) \tag{3.5}$$

$$W(\mathbf{q}_L) = -k_B T \ln \left(\frac{\int e^{-\beta U_{\text{int}}(\mathbf{q}_L, \mathbf{q}_S)} e^{-\beta U(\mathbf{q}_S)} d\mathbf{q}_S}{\sigma_L \int e^{-\beta U(\mathbf{q}_S)} d\mathbf{q}_S} \right) \tag{3.6}$$

Substituting Eq. 3.6 into Eq. 3.4 yields a reduced expression in which the position of the ligand no longer depends upon the exact configuration of the solvent,

$$\frac{Q_{N,L}}{Q_N} = \prod_i^{M_L} \left(\frac{2\pi m_i k_B T}{h^2} \right)^{\frac{3}{2}} \int e^{-\beta [U(\mathbf{q}_L) + W(\mathbf{q}_L)]} d\mathbf{q}_L \tag{3.7}$$

The integral over the position of the ligand ($\mathbf{q}_L$) can also be simplified by defining an internal reference frame that does not depend on absolute external coordinates (i.e., translational and rotational coordinates) of the ligand. This coordinate frame is

defined using a set of three bonded atoms in the ligand to specify the 6 external degrees of freedom and a set of $3N-6$ bond length (r_L), bond angle (θ_L), and torsional angle (ϕ_L) (BAT) coordinates to recursively specify the position of each subsequent atom relative to the position of the first three atoms. This coordinate change allows the integral over ligand configurational space to be separated into external and internal pieces, where the potential of the solvated ligand ($U(\mathbf{q}_L)$) is now independent of the external degrees of freedom. Analytically integrating over these external degrees of freedom of the ligand yields a constant factor of $8\pi^2 V_{N,L}$ [96]. The remaining integral over internal degrees of freedom can be computed numerically, and doing so in a BAT coordinate system often results in improved accuracy, as BAT sampling corresponds to natural motions of the molecule and a smoother exploration of the potential energy surface compared to a Cartesian coordinate system [118]. After simplification, the resulting expression for $\mu^{\circ}_{\mathrm{sol},L}$ is,

$$\mu^{\circ}_{\mathrm{sol},L} = -k_B T \ln\left(\frac{8\pi^2}{C^{\circ}\sigma_L}\prod_i^{M_L}\left(\frac{2\pi m_i k_B T}{h^2}\right)^{\frac{3}{2}} Z_L\right) \tag{3.8}$$

$$Z_L = \int J_L e^{-\beta[U(\mathbf{r}_L)+W(\mathbf{r}_L)]} d\mathbf{q}_L \tag{3.9}$$

where Z_L is a configurational integral over the solvated ligand, internal degrees of freedom and $J_L = \prod_L r_L^2 \sin\theta_L$ is the Jacobian weight for sampling in a BAT space [119, 120]. Note that limits of integration for r_L are defined by the volume of the system ($V_{N,L}$) which is ultimately normalized by the standard state concentration of the ligand in solution. Limits of integration for θ_L are defined from 0 to π, and limits for ϕ_L are defined from 0 to 2π. In this study only torsional degrees of freedom of the ligand were explored. Bond lengths and bond angles were held fixed at their equilibrium values, as it has been suggested that these degrees of freedom experience only small changes in configurational freedom upon binding and contribute negligibly to the free energy change [49]. Receptor degrees of freedom were held fixed due to issues of computational tractability and the large number of receptor degrees of freedom.

The derivations for the standard chemical potential of the receptor and complex are similar and will not be repeated here. It should be noted, however, that in the complex the six external degrees of freedom of the bound ligand become internal degrees of freedom of the complex, and integration over these new internal degrees of freedom is limited to only those conformations in which the ligand is actually bound and contained entirely within the receptor's active site cavity. Combining Eq. 3.1 with Eq. 3.8 for the ligand, receptor, and complex, the following expression for the standard free energy change is obtained

$$\Delta G^{\circ}_{\text{sol,bind}} = -k_B T \ln \left(\frac{C^{\circ}}{8\pi^2} \frac{\sigma_L \sigma_R}{\sigma_{LR}} \frac{Z_{LR}}{Z_L Z_R} \right). \tag{3.10}$$

Note that this expression is only dependent upon the configurational degrees of freedom of the complex, unbound receptor, and unbound ligand; all factors resulting from integration over the momentum portion of phase space exactly cancel when taking the difference between the bound and unbound states.

Once the partition functions in the bound and unbound states have been found, the enthalpy change (excluding negligible pressure–volume terms) can be found by calculating the appropriate averages over solute configurational space

$$\Delta H^{\circ}_{\text{sol,bind}} = \langle U(\mathbf{q}_R, \mathbf{q}_L) + W(\mathbf{q}_R, \mathbf{q}_L) \rangle_{\mathbf{q}_R, \mathbf{q}_L} - \langle U(\mathbf{q}_R) + W(\mathbf{q}_R) \rangle_{\mathbf{q}_R} - \langle U(\mathbf{q}_L) + W(\mathbf{q}_L) \rangle_{\mathbf{q}_L}, \tag{3.11}$$

where $\langle \rangle_{q_{R/L}}$ defines the configurational ensemble average over ligand and receptor degrees of freedom, respectively. The configurational entropy change upon binding can be found through the canonical equation [12]

$$S = k_B \ln Z + k_B T \left(\frac{\partial \ln Z}{\partial T} \right), \tag{3.12}$$

which results in the follow expression

$$\Delta S^{\circ}_{\text{sol,bind}} = \frac{1}{T}\left(\Delta H^{\circ}_{\text{sol,bind}} - \Delta G^{\circ}_{\text{sol,bind}}\right) - \left(\left\langle \frac{\partial W(\mathbf{q}_R, \mathbf{q}_L)}{\partial T} \right\rangle_{\mathbf{q}_R,\mathbf{q}_L} - \left\langle \frac{\partial W(\mathbf{q}_R)}{\partial T} \right\rangle_{\mathbf{q}_R} - \left\langle \frac{\partial W(\mathbf{q}_L)}{\partial T} \right\rangle_{\mathbf{q}_L}\right) \tag{3.13}$$

The final three terms that appear in the above expression for the entropy change result from the introduction of a potential of mean force (Eq. 3.6) to implicitly deal with solvent degrees of freedom. This formulation partitions the entropy change into additive solute and conditional solvent components in a mathematically and thermodynamically rigorous fashion [50, 121]. The first two terms in Eq. 3.13 correspond to the configurational entropy change of the solute, while the remaining terms correspond to the change in solvent entropy conditioned on the configurational state of the solute, averaged over all solute configurational degrees of freedom.

$$\Delta S^{\circ}_{\text{sol,bind}} = \Delta S^{\text{solute}}_{\text{config}}(\mathbf{q}_{LR}) + \Delta S^{\text{solute}}_{\text{config}}(\mathbf{q}_S|\mathbf{q}_L, \mathbf{q}_R) - \Delta S^{\text{solute}}_{\text{config}}(\mathbf{q}_S|\mathbf{q}_R) - \Delta S^{\text{solute}}_{\text{config}}(\mathbf{q}_S|\mathbf{q}_L) \tag{3.14}$$

In the present study, all reported free energy differences include enthalpic and entropic contributions from both solute and solvent degrees of freedom.

3.3.2 Conditional Mutual Information Expansion

The configurational entropy of each ligand was decomposed into individual, per degree of freedom entropy and higher-order coupling terms using a conditional mutual information expansion (CMIE). Similar to the mutual information expansion presented by Matsuda [112] and Killian [113], this expansion divides the full entropy into a sum of sequentially higher-order mutual information terms. However, rather than partition the total entropy into a set of overlapping entropic contributions that are added and subtracted with successive terms, we partition the space into a set of mutually exclusive terms, each of which captures the entropy content of either a single degree of freedom or the coupling between a group of degrees of freedom. This is done by adding up the mutual information of all possible combinations of degrees

of freedom, given that the distributions of the remaining variables are known. This can be expressed as

$$\begin{aligned} S(x_1, x_2, ..., x_N) = &\sum_{i=1}^{N} I(x_i|\{x_i\}^c) \quad + \quad \sum_{\substack{i,j=1 \\ i<j}}^{N} I(x_i; x_j|\{x_i, x_j\}^c) \quad + \\ &\sum_{\substack{i,j,k=1 \\ i<j<k}}^{N} I(x_i; x_j; x_k|\{x_i, x_j, x_k\}^c) \quad + \quad ... \quad + \\ &I(x_1; x_2; x_3; ...; x_n), \end{aligned} \tag{3.15}$$

where N is the total number of degrees of freedom of the system, $\{x\}^c$ is the complement of $\{x\}$, and $I(\{x\}|\{x\}^c)$ is the mutual information of a set of variables $\{x\}$ conditioned on the complementary set $\{x\}^c$ or simply conditional entropy when $|\{x\}| = 1$. This decomposition follows from the set measure-theoretic definition of multivariate mutual information [122, 123], where each conditional information term corresponds to a non-overlapping subset of an information diagram.

As an example, consider a system with three degrees of freedom $\{x, y, z\}$. The CMIE for this system is

$$\begin{aligned} S(x, y, z) = &I(x|y, z) + I(y|x, z) + I(z|x, y) + \\ &I(x; y|z) + I(x; z|y) + I(y; z|x) + \\ &I(x; y; z), \end{aligned} \tag{3.16}$$

where the first three terms are of first order, the second three terms are of second order, and the last term is of third order. As illustrated in Fig. 3-1, the first-order terms define the conditional entropy due solely to each individual degree of freedom; this corresponds to the average entropy due to in a degree of freedom, given that the remaining degrees of freedom are known. That is, first-order measures define the entropy due to each degree of freedom that is not mediated by any other degrees of freedom through coupling. Similarly, the second-order terms define the conditional mutual information between each pair of degrees of freedom, which correspond to

measures of the coupling present between pairs of variables that is not mediated by higher-order coupling. The third-order term defines the higher-order coupling present among all variables. It is important to note that while this expansion partitions the entropy into non-overlapping pieces, only first- and second-order terms are guaranteed to be positive [121]. As such, higher order mutual information terms can either increase or decrease the total entropy of the system. Additionally, as with any entropy expansion, all of these terms are fundamentally dependent upon the choice of the reference frame and thus represent a potentially non-unique but still useful interpretation [96, 50].

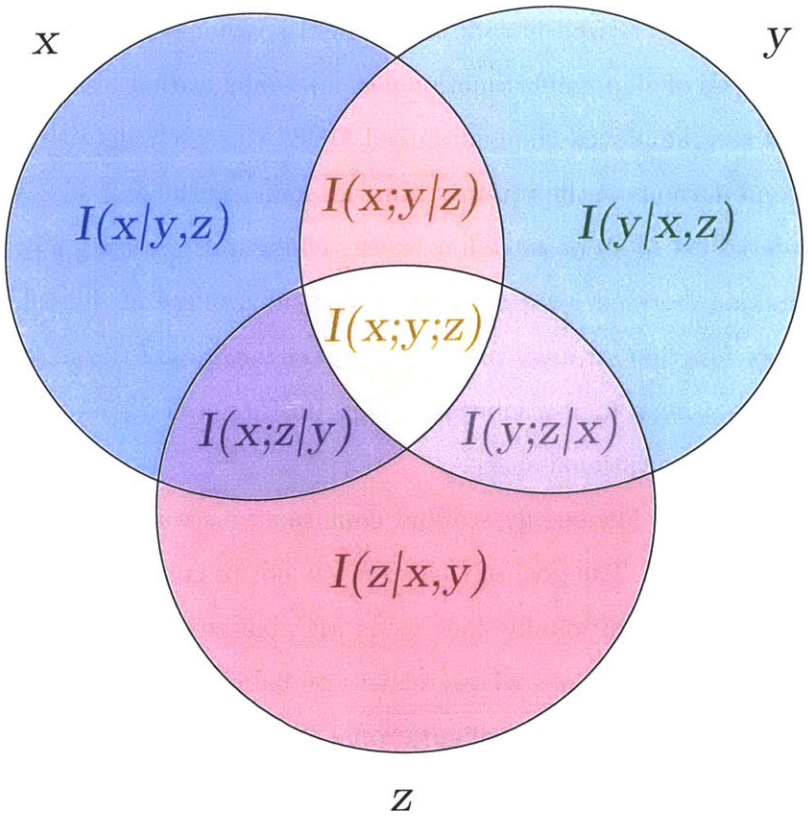


Figure 3-1: **Three body conditional mutual information expansion.** The entropy of a three body system with degrees of freedom x, y, and z corresponds to the union of all three circles. This total entropy is broken up according to Eq. 3.16 into marginal entropies (blue, green, and red areas), pairwise coupling entropies (purple, orange, and brown areas), and a single three body or third order entropy (yellow area).

3.3.3 Ensemble Enumeration and Partition Function Determination

The bound and unbound state configurational integrals of five HIV-1 protease inhibitors (Fig. 3-2) were evaluated via a three-step, rotamer based, enumerative configurational search. All internal torsions as well as external (i.e., orientational relative to the receptor) degrees of freedom were rotamerized using uniform step sizes to exhaustively explore configurational space at different levels of discretization. All examined ligands were comprised of a common chemical scaffold with potentially variable functional groups at five positions (R1–R5). The first step of the search involved generating separate discretized libraries of scaffold positions and orientations as well as rotamer libraries of all possible functional group configurations relative to the scaffold. The second step employed the guaranteed DEE/A* search algorithms to explore all possible combinations of the rotamer libraries found in the first step and generate an energy-ordered list of all possible low-energy configurations using a pairwise additive energy function (termed low-resolution). The third phase of the calculation used a tiered energy function strategy to re-evaluate the energies of the collected low-energy configurations using a high-resolution energy function and numerically integrate over the explored configurational space.

The ensemble of low-energy scaffold conformations was generated using an enumerative MC search. The goal of this step was not to collect a Boltzmann ensemble via sampling, as is traditionally done using MC, but to mine for an ensemble of low-energy scaffold configurations whose relative probabilities will be explicitly computed after exploring the remaining configurational space. For all simulations the move set included all torsional rotations, excluding methyl and amide bond rotations, as well as overall translations and rotations in the bound state. The upper bounds on step sizes for overall translations and rotations were set to 0.5 Å and 30°, and individual torsional moves were capped at 15° and 180° in the bound and unbound state, respectively, with an equal weight applied to all moves. Ten, independent simulations of 50,000 steps each were performed for each ligand in both the bound and unbound

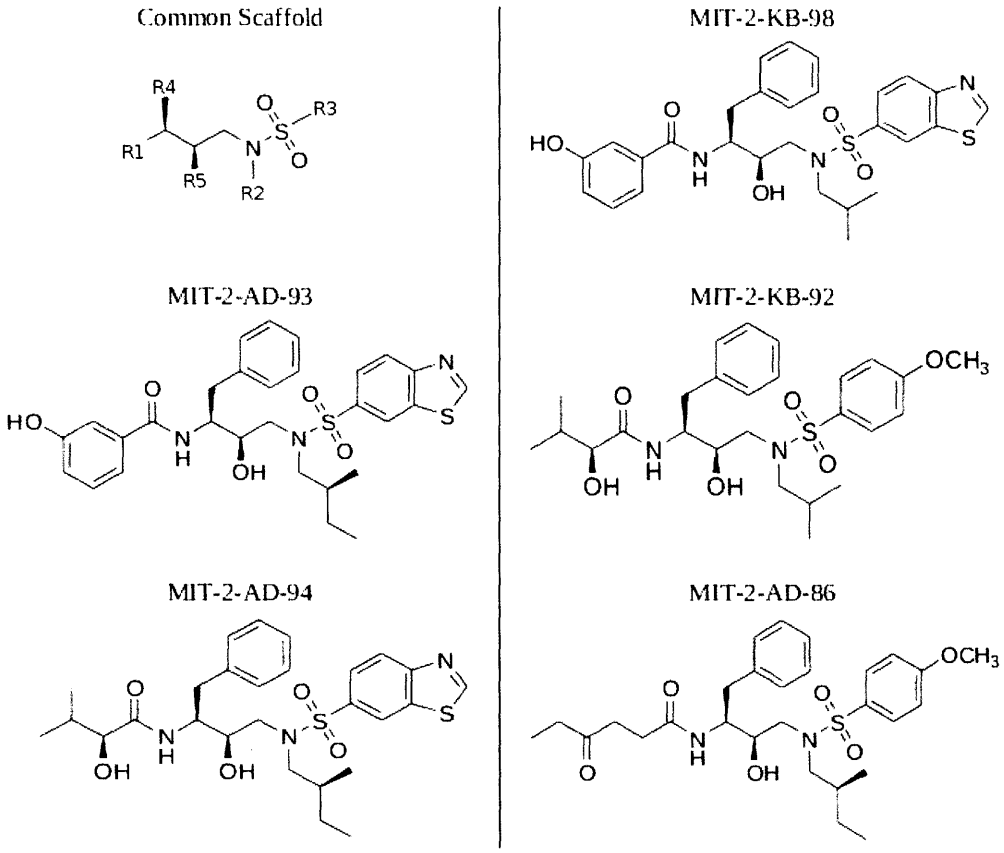


Figure 3-2: **Selected HIV-1 protease inhibitor structures.** These five inhibitors were originally designed by Altman *et al.* [47] to test the substrate envelope hypothesis. They are derived from the Darunavir/Amprenavir scaffold and all exhibit nanomolar binding affinity.

states, and the external and scaffold degrees of freedom of all collected configurations were snapped to a uniform rotamer grid with a resolution of 0.1 Å and 10° (20°) for the bound (unbound) state. All search parameters and grid resolutions were selected to maximize spatial coverage over the course of a simulation, while ensuring that the rate of configurational exploration approached zero with a computationally feasible number of scaffolds. Simulations were performed using the CHARMM computer program [3, 76] with the CHARMm22 force field [124] and a distance-dependent dielectric constant of 4. The functional group rotamer libraries were generated by brute force enumeration of torsional space.

All possible rotamer combinations from the constructed libraries were energetically evaluated to map out ligand configurational space using an in-house implementation of the DEE and A* algorithms [47]. Given a set of uniformly sampled, complete scaffold and functional group rotamer libraries, these algorithms are guaranteed to find the GMEC as well as all other near-optimal structures up to a user supplied energy cutoff. Using the list of collected energies produced by this guaranteed search, Boltzmann factors ($e^{-\beta U_i}$) were calculated for each configuration i and used to compute a low-resolution estimate of the configurational integral (Eq. 3.9) by numerical quadrature. For these high throughput energy evaluations, a pair-wise decomposable energy function was used that included all pairwise van der Waals and Coulombic, intra- and intermolecular interactions, computed with the CHARMm22 parameter set [124]. The energy cutoff used in this low-resolution estimate was always $\leq$ 10 kcal/mol, as this provided enough coverage of the potential energy landscape to guarantee partition function convergence (*vide infra*).

The final step of the search included the energetic re-evaluation of the collected ensemble using a higher resolution energy function to account for solvation effects and to obtain a more accurate measure of the free energy change upon binding. The improved energy function included all pairwise van der Waals interactions, continuum electrostatic solvation energies collected from a converged linearized Poisson–Boltzmann calculation found using the DelPhi computer program [84, 125], as well as solvent accessible surface area energies to model the hydrophobic effect [86]. Solvation energies were calculated using an internal dielectric of 4 and a solvent dielectric of 80. A grid resolution of 129 x 129 x 129 with focusing boundary conditions [85] was used, along with a Stern layer of 2.0 Å and an ionic strength of 0.145 M.

The high-resolution (HRes) configurational integral was computed using a bootstrapping method that breaks up ligand low-energy configurational space into two regions: one described in terms of explicit configurational states evaluated using the HRes energy function, and the other in terms of distributions of HRes energy levels

(E_{HR}) inferred from the low resolution (LRes) energy level (E_{LR}) distribution:

$$Z_{HR} = \int_A J(\mathbf{q}_L)\, e^{-\beta E_{HR}(\mathbf{q}_L)} d\mathbf{q}_L + \int_B g(E_{HR}) e^{-\beta E_{HR}} dE_{HR} \tag{3.17}$$

Here, $g(\mathbf{E}_{HR})$ is the degeneracy of the HRes energy levels, and A and B define complementary regions of configurational space that together cover the entire space. Note that this approximation of the HRes energy space was made for computational efficiency, as it is currently computationally intractable to explicitly re-evaluate all of the millions of configurations collected from the LRes DEE/A* search. The first term of Eq. 3.17 corresponds to the HRes partition function defined by some fraction of the total number of low-energy configuration, and it is calculated by explicitly re-evaluating thc energies of the top 50,000 configurations and integrating over these states. The second term estimates the contributions made by the remaining, higher-energy members of the full ensemble to the HRes partition function and can be viewed as a correction to the first term. It was computed by finding an approximate distribution of HRes energy levels as a function of the known LRes energy level distribution to estimate $g(E_{HR})$, using a probabilistic formalism similar to those hierarchical evaluation methods used in molecular design [126]. The LRes energy space (minus the top 50,0000 structures) was divided into 0.1 kcal/mol bins, and a randomly selected set of 1000 configurations were re-evaluated in each bin. Each 1000-configuration sample was used to approximate the distribution of HRes energy levels observed in each LRes bin i. The resulting set of conditional high-resolution energy level probability distributions, $P(E_{HR}|E_{LR})_i$, were empirically fit to either single or double skewed Gaussian distributions to determine the approximate shape of the distribution. Each was then weighted by the number density of configurations in that particular bin $\rho(E_{LR})_i$, which yielded the HRes energy level degeneracy in each bin, $g(E_{HR})_i$.

$$g(E_{HR})_i = \rho(E_{LR})_i P(E_{HR}|E_{LR})_i \tag{3.18}$$

The total contribution of all bins to the high-resolution partition function was then calculated by integrating over the HRes energy levels in each bin via numerical quadra-

ture and then summing over each LRes bin

$$Z_{HR} = \int J(\mathbf{q}_L)\, e^{-\beta E_{HR}(\mathbf{q}_L)} d\mathbf{q}_L + \sum_i \int \rho(E_{LR})_i P(E_{HR}|E_{LR})_i\, e^{-\beta E_{HR}} dE_{HR}. \quad (3.19)$$

3.3.4 Structure Preparation

The receptor structure used in this study was a darunavir bound x-ray crystal structure obtained from the Protein Data Bank (PDB) [127] (Accession code 1T3R) [128], prepared using methods and structural modifications from Altman *et al.* [47]. Partial atomic charges for each inhibitor were determined by fitting to the electrostatic potential of an optimized ground state structure using the restrained fitting methods of Bayly *et al.* [79]. Geometry optimizations as well as electrostatic potential calculations were performed with the GAUSSIAN03 computer program [129] using the Restricted Hartree-Fock method with the 3-21G and 6-31G* basis sets, respectively.

3.4 Results and Discussion

3.4.1 Rotamer Grid Resolution and Thermodynamic Convergence

Conformational space of all inhibitor degrees of freedom was explored at multiple resolutions. Ultimately, the grid resolution used to compute all thermodynamic parameters was selected based on the rate of exploration of scaffold degrees of freedom and the numerical convergence observed in the computed free energy. As the configurational search was performed in multiple steps, we examined the convergence of each step separately. In step one the external and internal scaffold degrees of freedom were explored via an enumerative MC simulation in which collected configurations were snapped onto a uniform rotamer grid of user defined resolution. The coverage of low-energy space was measured at multiple grid resolutions to find the highest resolution grid possible while maximizing coverage (Fig. 3-3). Simulation convergence was quantified via the number of unique, grid-snapped configurations found at each MC step, as this growth rate should approach zero as the simulation length increases and more configurational space is explored. A 0.1 Å/10° grid was used in the bound state and a 20° grid in the unbound, as the final growth rates at these resolutions were converged to within 0.03 unique configurations per step (3% of the maximum growth rate) for all bound/unbound inhibitors. Increasing the grid resolution to 5° and 10° increased the number of unique configurations found per step, which required collection of a computationally intractable number of scaffold positions in order to ensure the configurational space was adequately sampled (i.e., a growth rate of approximately zero). Note that in all the computed inhibitor ensembles, the unbound state required a coarser grid in order to obtain a comparable rate of convergence using a similar number of overall configurations. This indicates that there is a more densely populated low energy scaffold space in the unbound compared to the bound state. Without the receptor present to constrain the torsional motions of the inhibitor, the scaffold is able to adopt a much wider variety of conformations without paying large

energetic penalties. It is important to note, however, that while the unbound state has a more densely populated low energy space compared to the bound state, the structural differences between individual unbound conformations are also smaller, and capturing these features requires less resolution.

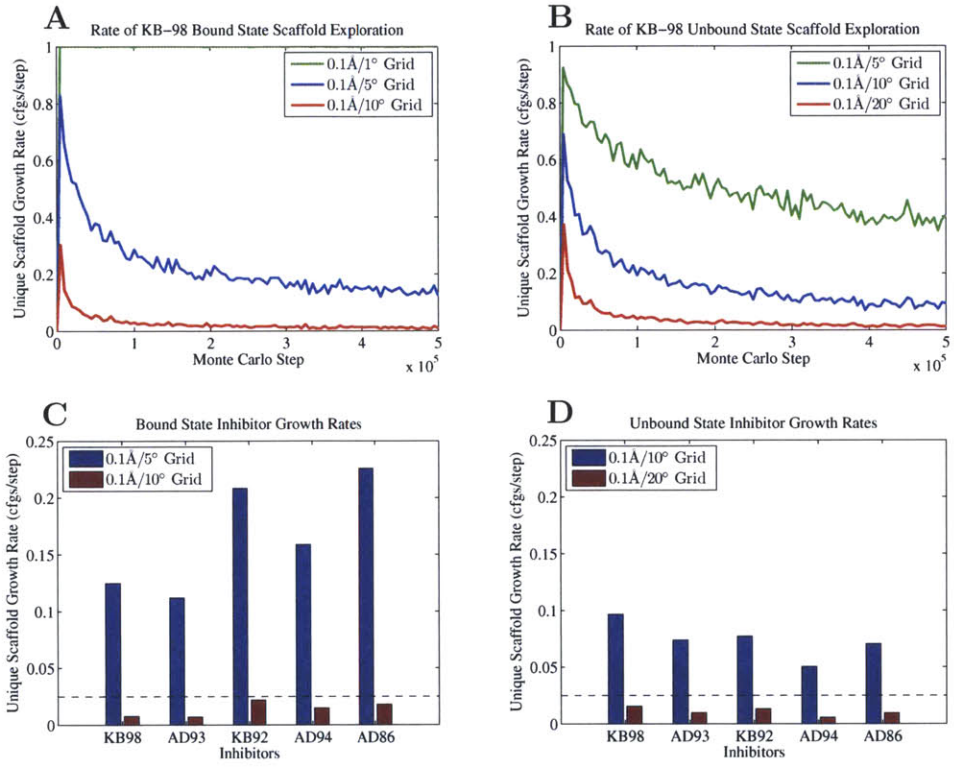


Figure 3-3: **KB-98 enumerative Monte Carlo scaffold grid resolution convergence.** (A,B) The unique scaffold/external configuration growth rates measured as a function of grid resolution and Monte Carlo simulation length in the bound and unbound states. (C,D) Final unique configuration growth rates for bound and unbound state. The dotted line indicates the 0.03 configurations per step acceptance cutoff used.

In the second configurational search step the remaining functional group degrees of freedom were enumerated using a user specified specified rotamer step size, yielding a low-resolution measure of free energy. Convergence of this absolute free energy was measured using the derivative of the free energy as a function of the grid resolution and offset calculations (Fig 3-4). Using maximum possible sampling resolutions of

15° in the bound state and 30° in the unbound state, we observe free energy changes of $\sim$ 0.2 kcal/(mol·degree) and $\sim$ 0.015 kcal/(mol·degree), respectively. Offset calculations at these resolutions also yield small errors of less than $\pm$1 kcal/mol. Both of these measures of free energy error suggest that accurate measures of the free energy of binding can be found with moderate functional group grid resolutions. These data also imply that the unbound state has a more degenerate, low-energy configurational space than the bound state with wider, less rugged potential energy wells, i.e. vast regions of unbound configurational space are well described by coarser sampling with limited grid error.

3.4.2 Ensemble Size and Thermodynamic Convergence

The Boltzmann distributions computed for each inhibitor ensemble was truncated at a range of energy cutoffs to explore the effect of collected ensemble size on free energy convergence. These cutoffs define the ensemble of all configurations with energies within a particular energy range above global minimum energy (Fig. 3-5). Our results suggest that only a very small portion of configurational space is necessary to achieve a very high level of convergence for the free energy. The top 1 kcal/mol of the ensemble brings the computed free energy within 2% ($\sim$1.5 kcal/mol) of the converged free energy in the bound state and within 5% ($\sim$2.5 kcal/mol) in the unbound state. At the highest degree of rotamerization, this corresponds to less than 200 and 700 configurations in the bound and unbound states, respectively. This behaviour was observed in all configurational ensembles, and all thermodynamic averages were well converged to within less than 1 kcal/mol when the full set of configurations was included. This rapid convergence suggests that the most relevant portions of configurational space are low-energy wells and that the average thermodynamic properties of these systems are well described by low-energy configurational ensembles, supportive of the predominant state hypothesis.

One should note, however, that given a fixed ensemble size, not all averages reach the same level of convergence. We observe that ensemble enthalpies and entropies show slower rates of convergence compared to free energies. Using a 15° grid in the

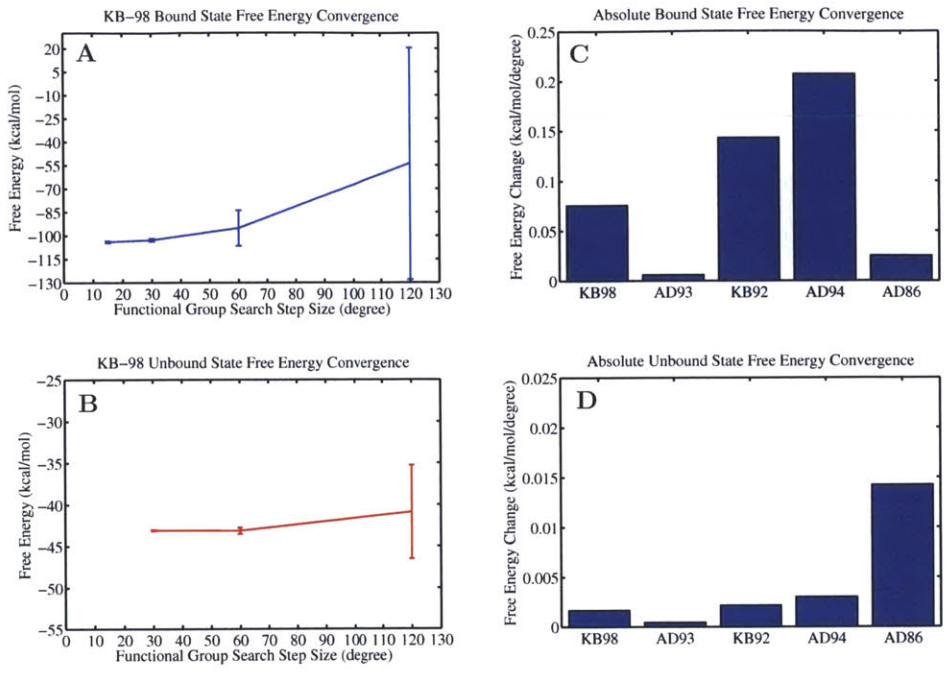


Figure 3-4: **KB-98 functional group grid resolution convergence.** (A,B) Low-resolution free energy convergence of KB-98 in the bound and unbound state as a function of the rotamer step size used when searching the functional group degrees of freedom. Error bars indicate the standard deviation computed from the two offset calculations performed at each grid resolution. All calculations use a starting geometry collected from the lowest energy configuration found during the bound state scaffold search. The offset calculations begin from this configuration with each degree of freedom offset by half the step size so as to escape from the initial, low energy well. (C,D) Final rate of free energy change for each examined inhibitor in the bound and unbound state.

bound state and a 60° grid in the unbound state, the ensemble free energies exhibited an average final rate of convergence of 0.001 and 0.1 kcal/mol per kcal/mol of the ensemble, respectively. By comparison, both the enthalpy and entropy averages showed average final rates of 0.02/0.2 and 0.02/0.3 (bound/unbound). These data suggest that larger fractions of configurational space are required to accurately gauge enthalpic or entropic contributions to binding compared to the full free energy and that the sampling error associated with these contributions partially cancels when computing the free energy. In particular, accurate estimation of entropic changes

requires elaboration of low density regions of the distribution, which becomes increasingly difficult to explore as the degeneracy of configurational space increases. In all cases, a significantly smaller portion of the less degenerate bound state space was required to obtain comparable levels of convergence compared to the unbound state.

3.4.3 Experimental versus Calculated Inhibitor Affinities

One of the primary motivations in this study was the lack of strong correlation between the originally computed design energies for these inhibitors and experiment (Fig. 3-6A). We sought to improve upon our original relative affinity prediction and examine which methodological enhancements brought about the most significant correlative improvement of calculated affinity with experiment. Examining the r^2 correlation coefficients for calculated vs. experimental affinity, the old static enthalpy metric shows almost no correlation (0.05). Note that the single conformations used to evaluate binding in this study were found using a coarser rotamer library, a constrained scaffold set, and the rigid binding approximation. Using our more exhaustive search method, we first examined the effect of searching both the bound and unbound states in tandem for low energy structures with a higher resolution rotamer set (Fig. 3-6B). By selecting the lowest energy structure in both the bound and unbound ensembles, we computed a static enthalpy of binding that shows marked correlation with experiment ($r^2 \approx 0.7$) and significant reduction in the variance weighted sum of squared error ($\chi^2 \approx 47$ vs. 230). While the relative affinity ordering is still not correct, we have effectively separated the cluster of drugs that were previously all predicted to bind with the same affinity. This large improvement in correlation likely stems from two factors: searching conformational space more finely and appropriately accounting for the significant conformational change that each inhibitor undergoes upon binding with independent bound and unbound state searches. Heavy atom, least squares alignments of the best unbound to bound conformations show root mean square deviations of greater than 3.4 Å for each of the five inhibitors. Additionally, when bound inside the protease, each inhibitor takes on an extended shape to fit within the active site, and when unbound, each inhibitor undergoes a structural collapse in order to

maximize solvation as well as intramolecular interactions. We find that the solvent accessible surface area of the best inhibitor configurations decreases on average by $80 \pm 20\text{Å}^2$ upon binding.

We next examined the correlative effect of ensemble averaging by computing the average enthalpy change upon binding for each of the five inhibitors (Fig. 3-6C). This further improves correlation with experiment, yielding an r^2 value of 0.85, an additional two-fold reduction in χ^2 error, and the correct relative ranking of each inhibitor. Incorporating the ensemble average into the calculation of the enthalpy change has a significant effect on both the bound and unbound state and drags the average enthalpy up by 2.3 ± 0.2 kcal/mol and 2.8 ± 0.7 kcal/mol in the bound and unbound state, respectively. Examining the net difference, we find that KB-98 and AD-86 experience the largest change relative to their static enthalpy evaluation ($\sim$1 kcal/mol). This is due to the fact that for these two inhibitors, the average unbound state enthalpy is pushed up by 1 kcal/mol more than in the bound state, which suggests that the unbound low-energy state space of these two compounds is more densely populated compared to the remaining inhibitors. These changes result in residual improvement for all points except AD-93, with KB-98, KB-92, and AD-94 showing the largest improvement.

We also examined the correlative effect of ensemble averaging by correcting the new static enthalpy estimates with our computed entropy penalties (Fig. 3-6D). Interestingly, this also introduces a clear separation between the high affinity inhibitors (KB-98 and AD-93) and the less effective ones (AD-94, KB-92, and AD-86), and significantly improves correlation with experiment, giving a correlation coefficient of 0.84. Comparing this correlation with that of the new static enthalpy change, the observed improvement is primarily the result of bringing KB-98 and AD-94 closer to the best fit line. KB-98 shifts because its computed entropic free energy loss is much smaller relative to the other computed entropic penalties, while AD-94 shifts because is experiences the largest entropic free energy loss. Note that attempts to similarly correct the new static enthalpy with a constant, entropic penalty per rotatable bond fail to significantly improve correlation. We explored possible constant corrections up

to 2 kca/mol/bond, and the most effective (0.5 kcal/mol/bond) only yielded an r^2 value of 0.74.

The final effect we explored was accounting for both the configurational entropy change of the ligand (ΔS_{bind}) as well as the the average enthalpy change (ΔH_{bind}) upon binding, which together correspond to the full configurational free energy of binding, ΔG_{bind} (Fig. 3-6E). Surprisingly, while separately including either ensemble measure significantly improves correlation, together there is only a slight improvement over previously examined metrics ($r^2 \sim 0.87$). Each ensemble measure captures similar information such that together they have only a small, coupled effect. In total, our results show that finding a better estimate of the global minimum energy conformation (GMEC) in both the bound and unbound states can substantially improve relative inhibitor rankings, but that further improvement necessitates an ensemble treatment. We find that including information about the shape of the minimum energy potential well and surrounding wells, in addition to its relative position, is required to resolve more subtle differences between inhibitors.

A similar quantitative picture emerges when examining the calculated enthalpy–entropy breakdown of these inhibitors (Table 3.1). All of the inhibitors are computed to be enthalpically driven binders, as ΔH is favorable and nearly twice the size of the unfavorable, configurational entropy loss. Given this large relative difference and the functional form of the Boltzmann distribution, the importance of the GMEC in ranking can be rationalized, as this distribution is strongly biased towards and peaked about low energy configurations. It is interesting to note that the entropy changes ($-T\Delta S$) are quite large (+22–26 kcal/mol) relative to ΔG, and they are very similar to previous estimates of the configurational entropy loss of chemically similar HIV-1 protease inhibitors using different methodology [49]. In contrast to cheaper, empirical measures of configurational entropy loss, these entropies show only marginal correlation with the number of rotatable bonds explored ($r^2 \approx 0.5$). In particular, the entropy losses of AD-94 and AD-86 deviate significantly from the trend exhibited by KB-98, AD-93, and KB-92, which show a consistent loss of $\sim$1 kcal/mol per rotatable bond. Both KB-92 and AD-94 have 15 rotatable bonds, yet AD-94 loses nearly 1.5

kcal/mol more in entropic free energy upon binding. Structurally, these two inhibitors are also very similar, and differ only in the identity and flexibility of their R2 and R3 functional groups. In AD-94, R2 is more flexible than R3, while the reverse is true for KB-92. In the case of AD-86, we find that it loses nearly 2 kcal/mol less entropic free energy than would be expected assuming a constant entropy penalty. It is more flexible than KB-92 at the R1 and R2 positions and AD-94 at the R1 and R3 positions, yet it loses much less entropy than expected given its flexibility. These deviations highlight the fact that both the number of rotatable bonds as well as their location influence configurational entropy losses.

Table 3.1: Calculated thermodynamic changes upon binding for the five tested HIV-1 protease inhibitors. All values reported in units of kcal/mol.

	KB-98	**AD-93**	**KB-92**	**AD-94**	**AD-86**
ΔG	-24.3	-26.0	-19.8	-19.5	-20.1
ΔH	-46.9	-49.7	-44.5	-45.6	-45.1
$-T\Delta S$	22.6	23.7	24.8	26.1	24.96
Num. Rot. Bonds	13	14	15	15	17

3.4.4 Analysis of Marginal Configurational Entropy Changes

In order to understand the subtle entropic differences between these inhibitors and discern the major contributions to the absolute entropy loss, we decomposed inhibitor entropy changes using a conditional mutual information expansion. This decomposition separates out the entropy change into additive components that quantify the marginal conditional entropy losses and all higher order changes in coupling entropy. We find that the majority (95%) of the total entropy change can be traced back to the first-order conditional entropy terms, with the remaining entropy difference bound up in coupling entropy. These first-order terms are measures of the average change in configurational entropy present in each degree of freedom, given fixed conformations for the rest of the inhibitor torsions, i.e. they measure the per degree of freedom change in entropy independent of any coupling. As can be see in Fig. 3-7,

all inhibitors experience large losses in external entropy (~12.3 kcal/mol), as binding to the receptor creates a highly constrained environment, severely limiting inhibitor rotation and translation. For comparison, assuming a standard state concentration of 1 M, which corresponds to a volume of 1660 $\AA^3$ per molecule, and a free rotational volume of $8\pi^2$, the unbound state has an external, standard state entropic free energy ($-TS^\circ$) of approximately -7 kcal/mol, with -4.4 kcal/mol stemming from translational entropy and -2.6 kcal/mol stemming from rotational entropy.[2] This estimate of the external entropy loss compares very well with a variety of alternative formulations. In particular, Chang *et. al* estimate an external entropy loss of 12.3 kcal/mol for a chemically very similar inhibitor, Amprenavir, binding to HIV-1 protease using the second generation mining minima algorithm, and a loss of 11.6 kcal/mol using the quasiharmonic approximation [49]. Using molecular dynamics in conjunction with the quasiharmonic approximation as well as Schlitter's entropy formula, Carlsson and Aqvist estimate the combined translational/rotational entropy loss of benzene binding to fixed T4-Lysozyme to be ~11 kcal/mol [132].

The remaining internal, inhibitor degrees of freedom exhibit net losses of 9-10 kcal/mol, with a coupling independent, average loss of 0.7 ± 0.3 kcal/(mol·rot. bond) of configurational entropy. This average loss is in agreement with previously reported values of 0.4 to 0.9 kcal/mol/bond, which were estimated from the experimentally measured thermodynamics of fusion of small hydrocarbons [133]. Page *et. al* report losses of 1–1.4 kcal/kcal/mol/bond, which were estimated from the entropy loss measured upon hydrocarbon cyclization [134]. Note that the former estimate is derived from the entropy loss as a molecule is captured inside a crystal lattice, while the latter is a measure of the entropic cost of completely freezing out a degree of freedom into a constrained ring structure. The slight difference between our estimate and that of Page *et. al* likely stems from the fact that individual torsional angles are not

[2]Note that while this implies that the bound state entropy of external ligand degrees of freedom, which are now internal to the complex, is negative, it comprises only a fraction of the total entropy, which need not be positive. Additionally, these numbers do not account for the entropic contribution of the momentum portion of phase space, which exactly cancels when taking the difference upon binding (see Eq. 3.10).

completely frozen upon binding. As we can see from examining the marginal probability distributions of individual degrees of freedom, many torsional angles retain a considerable amount of conformational freedom upon binding (Fig. 3-8). Further exploration of these distributions shows unique differences in how each distribution changes upon binding. We observe two major trends among all the inhibitors, which we will illustrate using KB-98 as an example. First, moving down either column in Fig. 3-8, we see the distributions becoming increasingly spread out, indicating that as one moves farther away from the scaffold of the molecule, each degree of freedom is increasingly more mobile. Small-angle rotations about scaffold torsions (degrees of freedom one and two) swing large lever arms, which results in large displacements. Comparatively, rotations around terminal dihedral angles (e.g., degree of freedom five) swing small lever arms, and as a result tolerate much larger changes. Second, loss of configurational freedom upon binding for these inhibitors is due both to the disappearance of populated wells and to well contraction. Comparing the unbound- and bound-state distributions for degrees of freedom one, two, and four, we observe a collapse from a multimodal distribution to a unimodal one. Additionally, making the same comparison for degrees of freedom one, three, and four, we see the corresponding wells contract from a width of 80°, 120°, and 90° in the unbound state to 10°, 60°, and 60° in the bound. The marginal distribution changes for scaffold torsions are particularly stark, as one sees unimodal collapse and contraction upon binding for almost all of these core degrees of freedom across all the examined inhibitors. The window of occupied configurational states for these motions is always less than 20° in the bound state, which implies that accurate sampling of these highly constrained degrees of freedom requires very small step sizes. Comparatively, the most free motions correspond to hydroxyl rotations, as they have the shortest associated lever arm. All such groups exhibit broad, nearly uniform distributions in the unbound state, which become more peaked (widths of $\sim 180°$) and centered around an ideal hydrogen bond position upon binding.

Examining these marginal distributions for the variable functional groups R1, R2, and R3 across all of the studied inhibitors and their marginal entropy losses, we noted

a spatial dependence of the marginal entropy loss, with differential losses per degree of freedom depending on where the structural group docks within the active site. Table 3.2 shows the averages entropy loss upon binding per degree of freedom for the external, scaffold, and functional groups. Averaging over all five inhibitors, we find that for this scaffold the R3 group, which binds in the P2' pocket, experiences the largest entropic loss per degree of freedom (1.2 ± 0.5 kcal/mol/bond). By comparison, the R1, R2, and R4 groups, which sit in the P2, P1', and P1 pockets, lose 0.6, 0.8, and 0.7 kcal/mol/bond), respectively. This suggests that rigid functional groups are preferred at this site as they experience the greatest loss in entropy upon binding. Interestingly, when examining the experimentally measured affinities of the larger MIT-2 inhibitor library, we see a similar trend where the binding free energy becomes more unfavorable as the functional groups become more flexible [47].

Table 3.2: Entropy loss per rotatable bond. All values reported in units of kcal/mol.

	$-T\Delta S$
External	2.05 ± 0.03
Scaffold	0.65 ± 0.05
R1	0.6 ± 0.1
R2	0.8 ± 0.3
R3	1.2 ± 0.5
R4	0.7 ± 0.1
R5	0.49 ± 0.06
Avg. Internal	0.7 ± 0.3

3.4.5 Analysis of Configurational Coupling Entropy Changes

The remaining contributions to the configurational entropy loss are bound up in higher-order coupling terms. Individually, these terms are often much smaller than the first-order losses, but their net effect is significant, accounting for the loss of 1–2 kcal/mol of entropy as well as being informative of gross intramolecular coupling trends. Examining the relative error in the entropic free energy as a function of coupling order, we see that the source of this coupling entropy differs between the bound

and unbound states (Fig. 3-9A–E). In the bound state, we see that the largest source of coupling appears in second-order terms (as the error drops precipitously upon the addition of second order coupling terms), but that the higher-order effects only become negligible after the addition of ninth or tenth order terms when the cumulative error reaches 0. In the unbound state, we again see that the largest source of coupling arises from second order terms, but note that higher-order effects become negligible by the addition of fourth or fifth order terms. The size of the relative drop in the error between the bound and unbound states upon addition of second order terms suggests that there is more second order coupling in the unbound versus bound state, which translates to a loss of entropic free energy upon binding. The relative importance of higher order coupling terms in the bound state suggests that there is more significant higher-order coupling in the bound versus unbound state, which translates to a net gain in entropic free energy upon binding. Averaging over all inhibitors, we find that the net change in entropic free energy upon binding is unfavorable for all coupling interactions involving five or less degrees of freedom ($-T\Delta S \approx +1$ kcal/mol) and generally favorable for all higher order coupling interactions ($-T\Delta S \approx -0.3$ kcal/mol) (Fig. 3-9F). This suggests that upon binding, the receptor restricts not only the independent motions of individual, inhibitor rotatable bonds, but many of the pairwise, three-, four-, and five-body coupling interactions present in the unbound state as well. This also intimates that in the bound state, higher-order coupling between inhibitor torsional degrees of freedom arise as inhibitors adopt specific conformations to adapt to the constrained, receptor binding site.

We examined the large number of individual coupling interactions in both the bound and unbound ensembles and found that the majority of specific coupling terms each contribute less than 0.05 kcal/mol, and that the largest individual coupling terms never contribute more than ~0.3 kcal/mol. Consistent with our analysis of the gross changes in coupling, we see that most of the large magnitude couplings terms appear in the unbound state and stem from second order coupling between scaffold degrees of freedom. In particular, we observe strong coupling between adjacent scaffold torsions or between torsions one bond apart. Interestingly, these unique pairs of torsions can

modulate the van der Waals packing of the functional groups with each other, and Fig. 3-10A shows the two-dimensional probability distribution for two such coupled torsions in KB-98. We see that these two dihedral angles can manipulate the position of R3 relative to the rest of the inhibitor, and cooperatively interact to maximize intramolecular van der Waal interactions between the R3 ring and either the scaffold backbone or the R4 phenyl ring (Fig. 3-10C moving from left to right, top to bottom). These data suggest that unbound state coupling arises as a result of cooperative motions that maximize intramolecular hydrophobic and ring stacking interactions. Note, however, that none of these couplings individually contribute more than 0.15 kcal/mol to the overall entropy change. By comparison, we observe far fewer, large coupling terms in the bound state, and the most significant ($\sim$0.3 kcal/mol) arise between the two dihedral angles surrounding the amide moiety in the R1 functional group. These two torsions are coupled as it appears they can modulate the position of the distal, R1 hydroxyl group, which forms a hydrogen bond with the sidechain of D29 (KB-98/AD-93) or the backbone carbonyl of G48 (AD-94/KB-92). As in the unbound case, these two torsions compensate for each other, although here they affect intermolecular interactions with the receptor. The higher order, bound state coupling terms that we see are predicted to couple external degrees of freedom to core, scaffold torsions, but individually rarely contribute more than 0.1 kcal/mol.

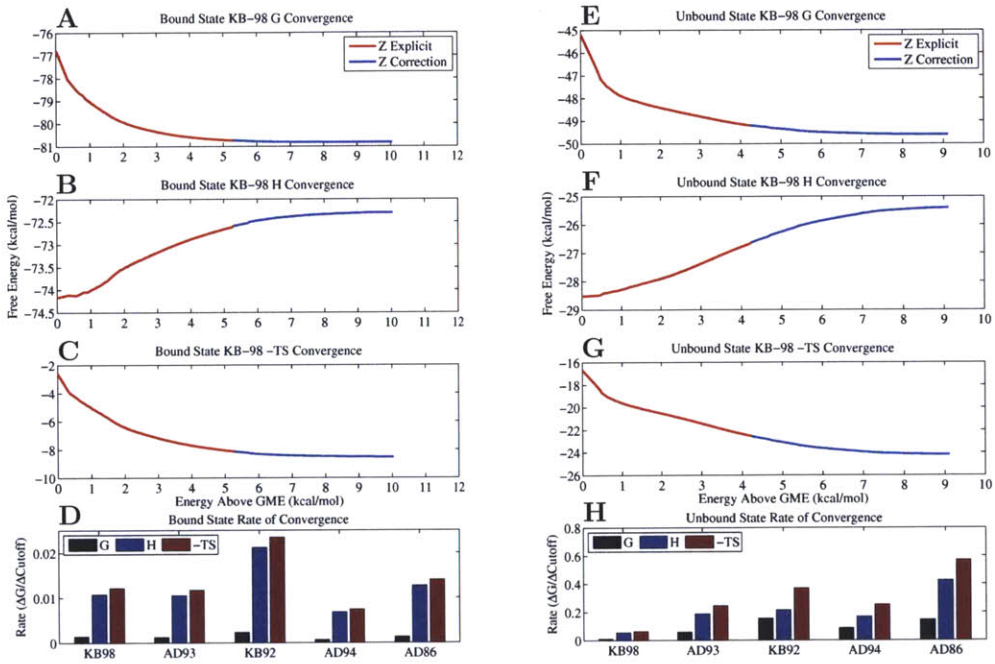


Figure 3-5: **Configurational free energy convergence.** (A-C) The convergence of the free energy, enthalpy, and entropy of KB-98 in the bound state, using a 10° scaffold and 15° functional group grid, as a function of ensemble size, measured in kcal above the global minimum energy. The red portion of curve shows the contribution to the average from the top 50,000 configurations (computed by explicit re-evaluation), and the blue portion shows the contribution from the remaining millions of configurations (computed via high resolution energy level inference). (D) The final rate of convergence of each thermodynamic parameter for each inhibitor in the bound state measured as a function of change in ensemble cutoff (kcal/mol per kcal/mol of the ensemble). (E-H) Measures of convergence for the unbound state using a 20° scaffold and 60° functional group grid.

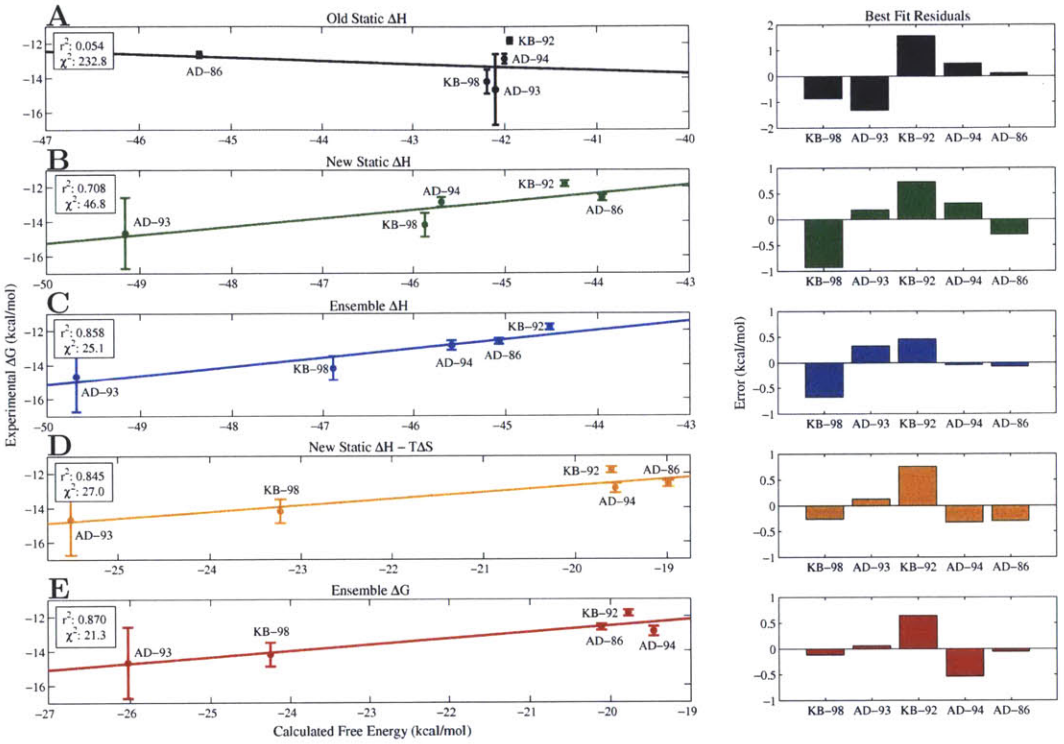


Figure 3-6: **Correlation between Calculated and Experimental Binding Affinity**. Effect of approximation on the correlation of the calculated inhibitor affinities with the experimental binding free energies. All experimental measurements come from Altman *et al.* [47], which were collected using an enzymatic inhibition assay [130, 131]. (A) Correlation between previously calculated static enthalpy metric (single conformation energy assuming rigid binding) and experiment. (B) Correlation between updated, static enthalpy metric (single conformation energy difference between bound and unbound states) and experiment. (C) Correlation between ensemble enthalpy and experiment. (D) Correlation between updated static enthalpy with ensemble entropy change and experiment. (E) Correlation between ensemble free energy change and experiment.

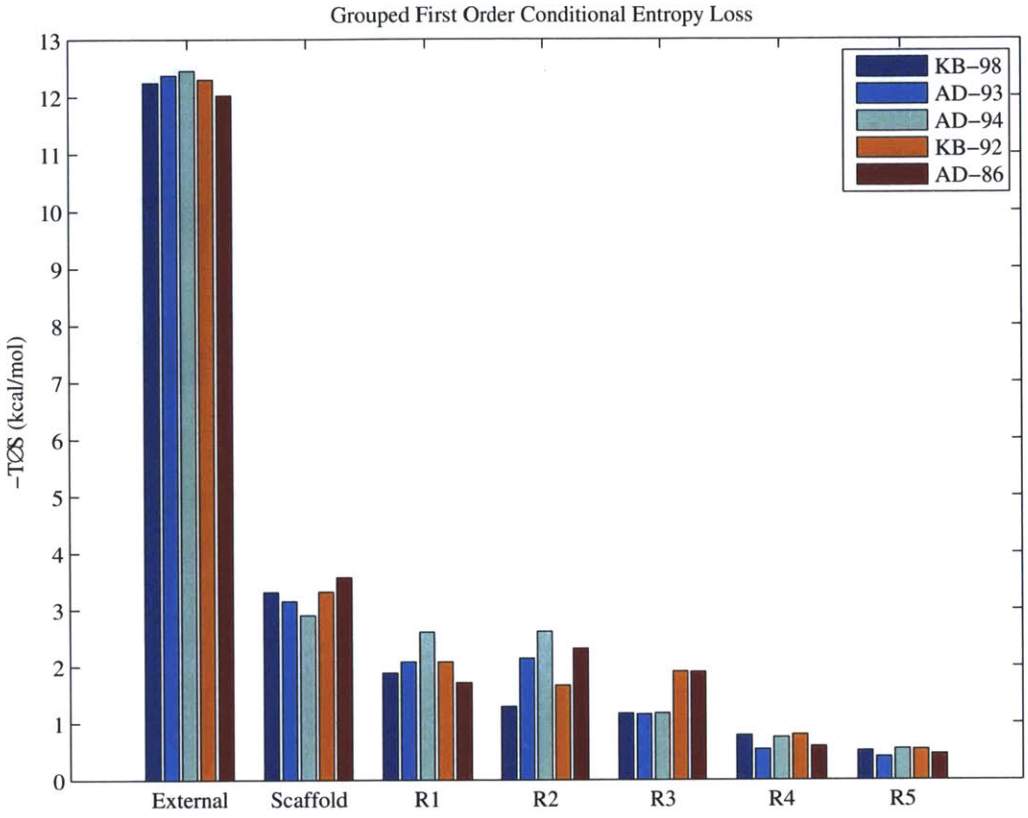


Figure 3-7: **First order conditional entropy losses**. The reported entropy loss for each group is defined as the sum of the first-order conditional entropy losses for each degree of freedom contained within the structure group. The scaffold contains 5 torsions, R1 contains up to 5 (KB-98/AD-93 – 3, AD-94/KB-92 – 4, AD-86 – 5), R2 contains up to 2 (KB-98/KB-92 – 1, AD-93/AD-94/AD-86 – 2), R3 contains up to 2 (KB-98/AD-93/AD-94 – 1, KB-92/AD-86 – 2), and both R4 and R5 contain 1 torsional degree of freedom.

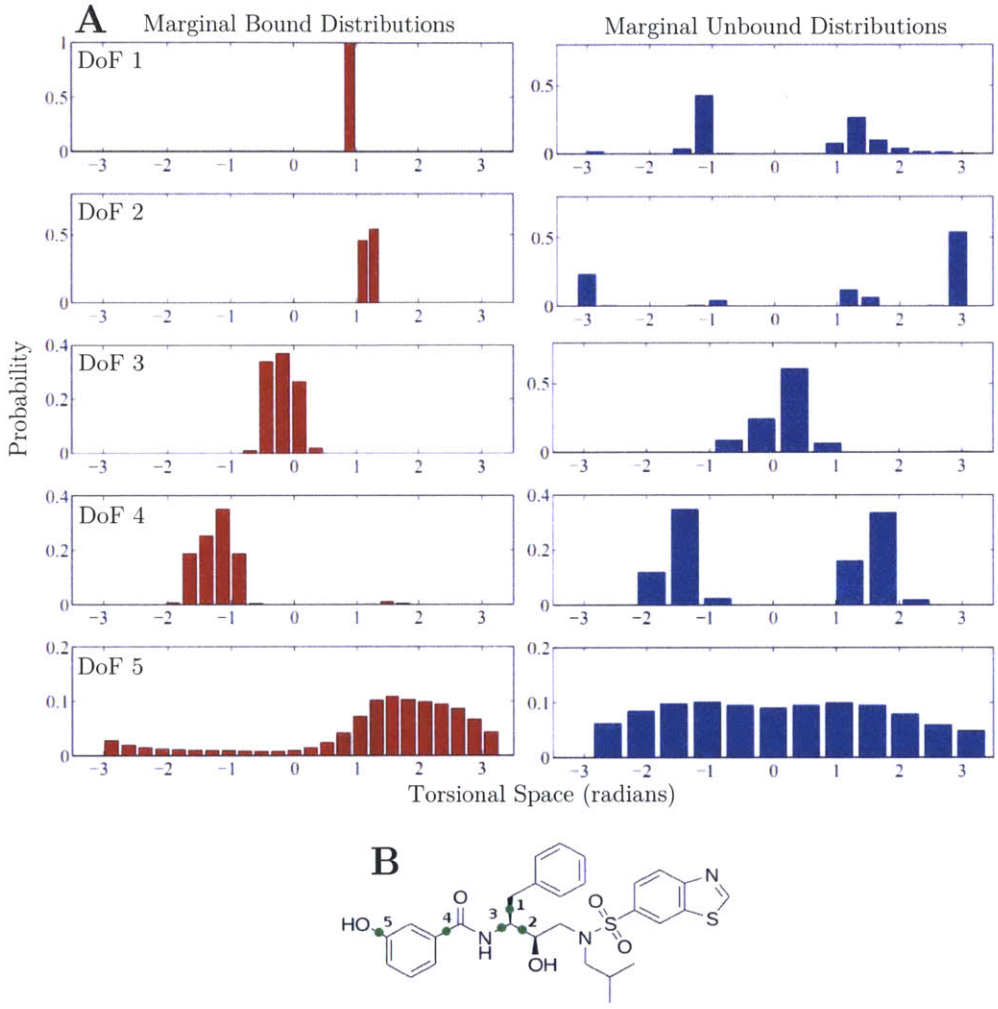


Figure 3-8: **Selected marginal distributions of KB-98** (A) Marginal distributions for selected torsional degrees of freedom in the bound and unbound states of KB-98. (B) Structure of KB–98 marked with degrees of freedom 1–5.

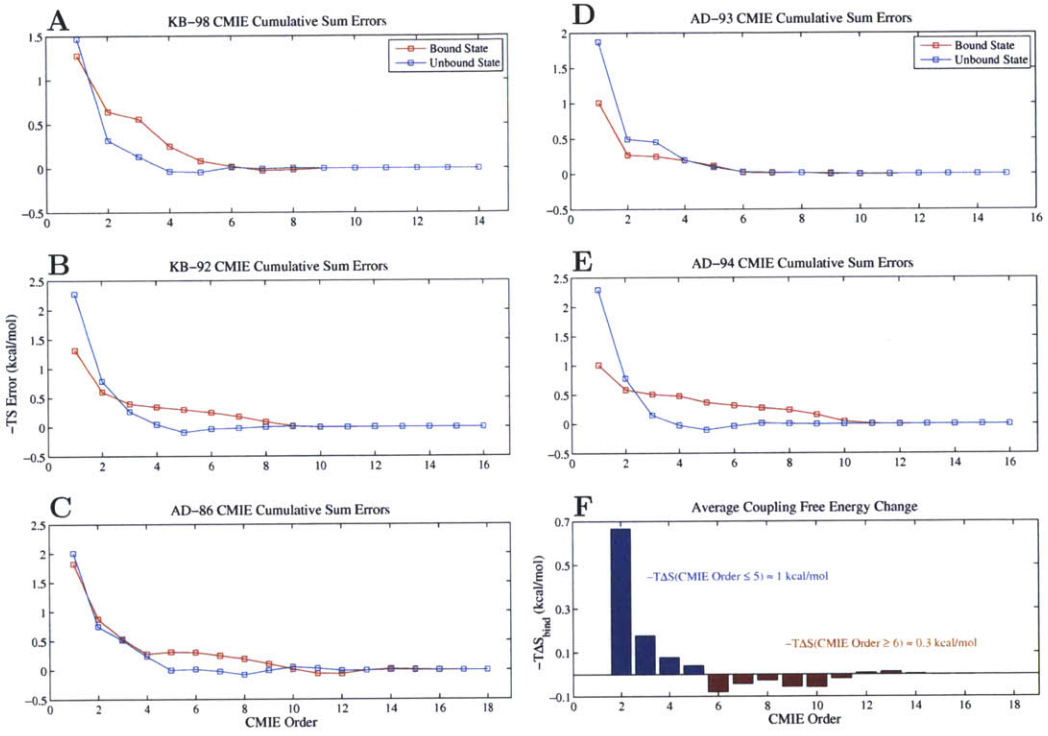


Figure 3-9: **Cumulative CMIE summation errors.** (A-E) Error associated with the cumulative summation of all conditional mutual information terms as a function of term order for each inhibitor in the bound and unbound states. (F) Average, net change in entropic free energy as a function of term order.

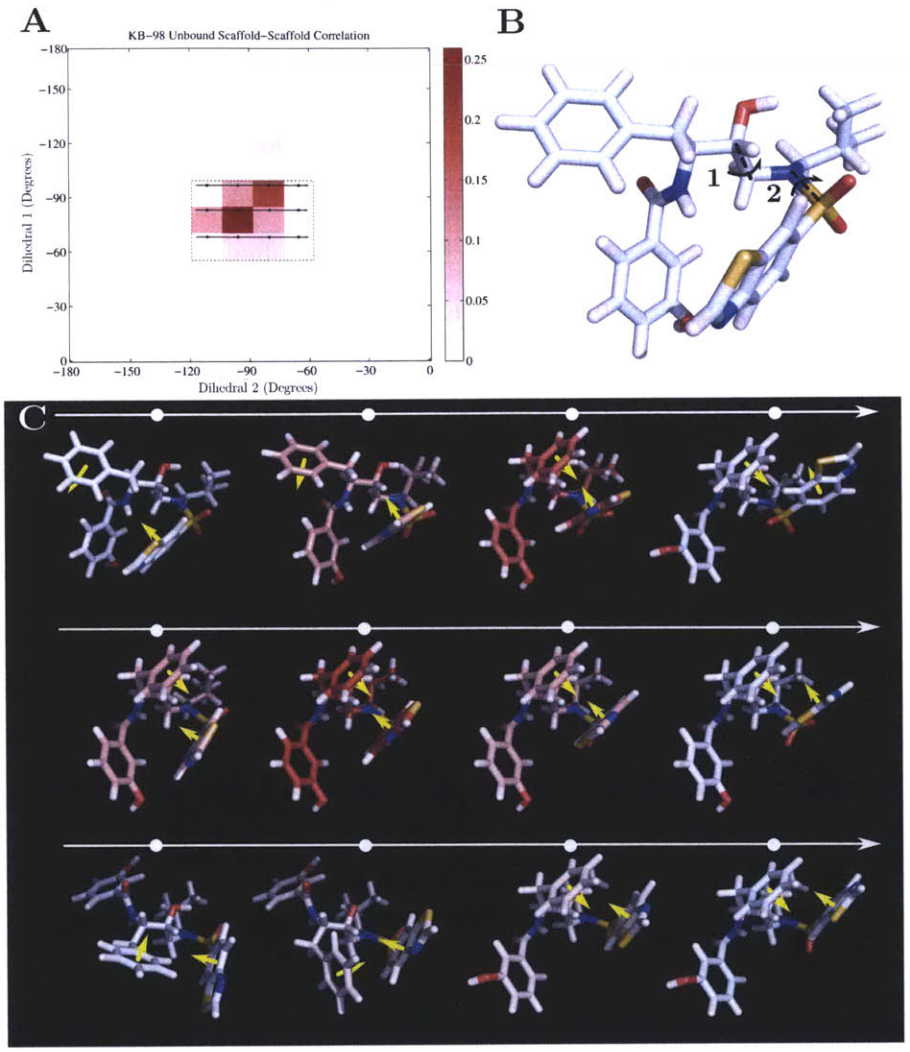


Figure 3-10: **KB-98 Unbound State Coupling.** (A) Pairwise probability distribution of two coupled scaffold dihedral angles in the unbound ensemble of KB-98. The inset rectangle and arrows highlight the sequence of configurations shown in part C. (B) Unbound KB-98 labeled with dihedrals 1 and 2. Arrows show direction of rotation moving from left to right (dihedral 1) and top to bottom (dihedral 1). (C) Sequence of most probable unbound configurations where dihedrals 1 and 2 have the values indicated by the inset rectangle in part A. Moving from right to left corresponds to clockwise rotation of dihedral 2 with a dihedral 1 fixed, and moving from top to bottom corresponds to counter-clockwise rotation of dihedral 1 with dihedral 2 fixed. Configurations are colored based upon their relative probabilities, with red indicating high probability, and the yellow arrows indicate the normal vector of the aromatic rings present in functional groups R3 and R4.

3.5 Conclusion

We have presented here a novel method to enumerate and study changes in the potential energy landscape of inhibitors upon binding. Using this enumerative, rotamer based approach, we obtain converged binding free energies, enthalpies, and entropies for flexible HIV protease inhibitors that accurately rank inhibitor affinities relative to one another. We find that using a fine grain configurational search to find just the global minimum energy conformation in the bound and unbound states to rank inhibitors correlates well with experiment, but that ensemble effects are critical for more accurate resolution of affinity differences. Breaking the free energy change apart, we observe that average enthalpies and entropies of binding are highly sensitive to the shapes of the global minimum energy well and surrounding wells, but that the sampling errors associated with these sensitivities partially cancel when computing the free energy. Additionally, we find that the predominant state assumption is valid for these high affinity inhibitors in both the bound and unbound states. The majority of configurational space contributes only marginally to the ensemble free energy, and converged free energy, enthalpy, and entropy values are obtained when truncating the configurational integral to include only those configurations within 10 kcal/mol of the GMEC. Compared to the free energy, however, computing accurate entropy and enthalpy changes requires larger low-energy ensembles that account for lower probability regions of phase space.

Analysis of the low-energy thermodynamic ensembles collected in this study reveals both how the shape of this landscape changes upon binding and how these differences translate into changes in the thermodynamic properties of the system. By decomposing the entropy change using an additive, conditional mutual information expansion, we see that the large computed differences in configurational entropy upon binding originate primarily from losses in external and uncoupled internal entropy, with average losses consistent well with previously reported experimental and computational estimates. From a potential energy landscape perspective, these changes arise from both well contraction and well disappearance. Changes in coupling entropy

play a more subtle, less pronounced role, and while their net effect is significant and critical to the rank ordering of inhibitor affinity, the entropy present in individual coupled motions is small. We find that most significant coupling interactions are of low (second or third) order, and most often appear between neighboring dihedral angles that can cooperatively modulate intermolecular or intramolecular interactions in the bound and unbound state, respectively. Examining the change in coupling between the bound and unbound state, we observe a net loss of low-order coupling interactions present in the unbound state between core degrees of freedom, and a net gain of high-order coupling interactions that appear only in the bound state. It is interesting to note that this entropy decomposition could be used to inform the optimization of future inhibitors, as it provides a way to estimate the spacial dependence of entropy loss for a given scaffold and determine the ideal position to include either flexible or rigid chemical groups.

Overall, these results suggest that inhibitor flexibility plays an important role in binding, but that the thermodynamic properties of these high affinity inhibitors are fundamentally determined by a small fraction of the full configurational ensemble. Low-energy configurations dominate the ensemble averages and coupling between inhibitor degrees of freedom has only a small effect. It is interesting to note that all of these results and conclusions arise without approximating the geometry of the potential energy landscape or inordinate sampling times. Our method is structured around the use the DEE/A* algorithm, which sorts configurations by their internal energies and explicitly computes their contribution to the Boltzmann distribution. As a result, the free energy is computed from the bottom up without having to approximate or account explicitly for landscape and well geometry. This ensures that the enthalpic and entropic contributions of all spatially distinct, low energy minima are included according to their level of import, and constructs a convergent, minimal configurational ensemble. By comparison, perturbative methods (FEP, TI) are very slow to converge, and alternative end-point approaches focus on computing free energies on a well-by-well basis using Boltzmann sampling to explore configurational space and map out low energy wells. The often used harmonic and quasiharmonic approxima-

tions assume the shape of potential energy well(s) can be accurately modeled as a collection of harmonic oscillators, which provides an analytical expression for the free energy contribution of each well. The former is the basis for normal mode analysis, which has been widely used to estimate entropy changes in biological systems [135], and the latter is used in the mining minima approach to similar effect [24]. The strong agreement found between ours and more approximate methods speaks to the accuracy of the predominant state and harmonic assumptions made for this system, but this may not be true for all systems.

Finally, it is also interesting to note that all the inhibitors examined in this study were originally developed to test the substrate envelope hypothesis and were designed to bind inside the substrate envelope. Four of these inhibitors (KB-98, AD-93, AD-94, and AD-86) were experimentally shown to exhibit relatively flat binding profiles to a variety of HIV protease mutants [47]. Considering just the top 50,000 configurations in each ensemble, we find that the vast majority of configurations in each of the respective ensembles also fit inside the substrate envelope, suggesting that the envelope hypothesis may be applicable in a more dynamic context. For an inhibitor to be insensitive to mutations in its target, the low energy ensemble of ligand configurations must fit inside the substrate envelope or substrate envelope ensemble.

The methods outlined here offer a flexible framework in which to study ensemble binding effects, and while the current study only explored ligand configurational freedom, receptor flexibility can easily be incorporated into this rotamer, DEE/A* based search scheme, given enough computational power. Nonetheless, there are clear limitations to the study presented here, which only considers flexibility in the ligand, binding to a rigid receptor. Proteins have significant numbers of degrees of freedom with local and global motions that will be affected, perhaps differentially, by ligand binding. Moreover, if ligand degrees of freedom couple effectively with receptor ones, then the ligand configurational entropy losses computed here will be overestimates.

Chapter 4

Mechanistic Analysis and Rational Design of Enzyme Catalysis in Ketoacid Reductoisomerase [1]

Abstract

We present here a combined study of the rational redesign and dynamic analysis of wild type and mutant variants of ketol-acid isomeroreductase (KARI). We develop an end-point, enzyme design protocol based on transition state theory (TST) to select for mutants that reduce the free energy of activation, and apply it to the redesign of KARI. Our results suggest that the TST based approach used here is effective at reducing the enthalpy of activation, but that in this naturally optimized system, this reduction is paired with a concomitant increase in the entropic free energy of activation. We further explore the wild type and designed mutants using transition path sampling (TPS) to estimate rate constants and study the reaction pathway. Analysis using this dynamic, ensemble method yields relative rates in qualitative agreement with experiment and indicates that TPS derived rates can discriminate between active and inactive enzyme variants. We also find that the isomerization reaction occurs via a vibrational, resonant energy transfer mechanism whereby the breaking bond is vibrationally pumped by its local environment before being pushed over the reaction barrier. This mechanism is observed in both wild type and mutant enzymes, which differ in the likelihood of reaching this activated state. These results highlight the role of dynamics in enzyme function and redesign.

[1] All experimental work presented in this chapter was performed by collaborators K. J. Gibson and M. P. McCluskey at the DuPont Central Research and Development Experimental Station.

4.1 Introduction

Ketol-acid reductoisomerase (KARI; EC 1.1.1.86) is a critical enzyme in the biosynthesis of the essential branched chain amino acids. Present in plants, bacteria, and some fungi, it catalyzes two chemical transformations, an isomerization and NADPH reduction reaction, along the pathway converting pyruvate into either valine or isoleucine [136]. In its most active form, it is known to bind two cofactors, one molecule of NADPH as well as two divalent magnesium ions [137] along with one of two related substrates, acetolactate (AL; (2S)-2-hydroxy-2-methyl-3-ketobutanoate) or acetohydroxybutyrate (AHB; (2S)-2-hydroxy-2-methyl-3-ketopentanoate). The binding of these cofactors is ordered and required for both isomerization and reduction reactions [138, 139, 140]. In the initial isomerization reaction, KARI catalyzes the intramolecular, alkyl transfer of either a methyl (AL) or ethyl group (AHB) as well as the interconversion of ketone and alcohol moeities, producing either 3-hydroxy-3-methyl-2-ketobutanoate (HMKB) or (3R)-3-hydroxy-3-methyl-2-ketopentanoate (HMKP) (Fig. 4-1). It is thought that this reaction is the rate limiting step for overall turnover [139]. The isomerization intermediate is not released and subsequently undergoes an NADPH dependent reduction, yielding the diols dihydroxyisovalerate (DHIV; (2R)-2,3-dihydroxy-3-methylbutanoate) or dihydroxymethylvalerate (DHMV; (2R,3R)-2,3-dihydroxy-3-methylpentanoate) [141]. Interestingly, KARI is known to also bind and reduce a number of 2-ketoacids without prior isomerization [142]. Recent crystal structures of multiple KARI variants reveal a wide variety of structural isomorphisms including dimeric (spinach) and icosameric forms (*Pseudomonas aeruginosa*) with highly conserved, both in sequence and structure, charged active sites [143, 144, 145, 146, 147]. Multiple studies have validated the necessity of these charged active site residues, with many active site mutations showing reduced substrate binding and/or loss of activity [142]. Even for wild type (WT) enzymes, however, the turnover rate is quite slow (1 s^{-1}), and given its essential role in plant growth, it is an attractive target for enzyme redesign.

Recent experimental and computational studies have proposed similar reaction

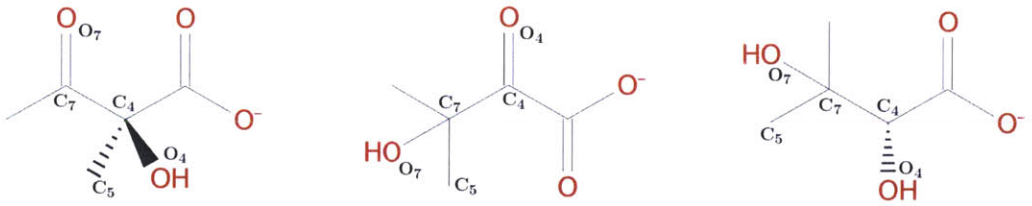


Figure 4-1: **KARI Substrates** Shown above are the methyl substituted reactant (acetolactate), the product of the isomerization reaction, HMKB), and the product of the reduction reaction (DHIV).

mechanisms for KARI [148, 149]. They suggest that during the initial isomerization reaction, a hydroxyl or active site residue abstracts a proton from the substrate, while the C_5 alkyl group migrates from the C_4 to the C_7 carbon. Both parts of this process are thought to be facilitated by the bound magnesium ions, which polarize the reactant O_4 alcohol and O_7 ketone dipoles, respectively. It has also been suggested that the transition state for this reaction is further stabilized via coupling to a proton transfer reaction between the substrate and an active site glutamate proton donor, E496 [148]. Similarly, for the reduction reaction it is thought that the magnesium ions polarize the O_4 ketone dipole, increasing the partial positive charge on the C_4 carbon, and electrostatically facilitate hydride transfer. These descriptions of the reaction mechanism, however, are couched in the language of transition state theory (TST) [150, 151], which assumes the rate of a reaction is proportional to the free energy difference between the ground and transition state (i.e., the free energy of activation). As such, they are limited to describing how the transition state is stabilized relative to the ground state and unable to provide a kinetic mechanism for how the reaction is initiated or climbs the reaction barrier.

A number of recent studies have been successful at (re)designing enzymes to either improve or develop novel activity using transition state theory in conjunction with computational methods to explore reaction pathways. Marti *et al.* used a combined quantum mechanics/molecular mechanics (QM/MM) approach along with free energy perturbation (FEP) and umbrella sampling to explore transition states for

the catalytic antibody 1F7 and side reaction of isochorismate pyruvate lyase, both of which catalyze the conversion of chorismate into prephenate. By optimizing interactions between the enzyme and observed transition states, they found or predicted mutations that lowered the activation free energy and improved activity [52, 53]. Similarly, transition state stabilization methods have been used in the *de novo* design of enzymes that catalyze novel reactions [54, 55, 56] as well as improve upon those initial designs [152, 153]. It has not yet been explored, however, whether the observed improvements in activity are due to enthalpic or entropic effects.

The goal of this study is two–fold: assess the efficacy of transition state based design in the re-design of spinach KARI as well as elucidate the mechanism of the catalyzed isomerization reaction for both WT and designed mutants. Both aims were accomplished using a combined approach using end–point TST as well as dynamic, transition path sampling (TPS) methods to assess transition state structure and reactive pathway dynamics. We developed structural models of both reactant and transition state species and computationally optimized for improved enzyme activity by evaluating the relative free energy differences between the transition and ground states for all possible single mutants within 10 Å of the active site. A panel of six mutant candidates was selected based upon predicted ability to reduce the free energy of activation relative to WT, and subsequently evaluated both experimentally as well as computationally using TPS to assess mutation effect in a dynamic context. We analyzed mutant activity from both TST– and TPS–based approaches and find that our designed mutants were effective at reducing activation enthalpies, but also incur entropic activation penalties that overpower the enthalpic gains. Re-evaluating mutant activities using TPS to compute rates and ensembles of reactive trajectories, we found that this dynamic, ensemble method was able to resolve the relative activity differences reported by experiment and suggests that these mutants are less active because the probability of fluctuating across the barrier is much lower compared to WT. We also find that the most active KARI variants are unique in that they not only facilitate reactive fluctuations, but they also increase the likelihood of non-reactive fluctuations that carry the system up but not across the reactive barrier. Finally, we

find that the isomerization reaction of both WT and mutant enzymes is consistent with a "pump–and–push" mechanism. Energy is pumped into the breaking bond via vibrational, resonant energy transfer and this activated bond is pushed by a conserved active site residue toward product state. We note that this is the first study to explicitly use TPS in the calculation of enzyme rates across a panel of enzymes with varying activities.

4.2 Methods

4.2.1 Computational Design Strategy

Possible activity–improving mutations were designed to minimize the activation free energy barrier of the enzyme catalyzed, isomerization reaction relative to WT. Using a transition state theory based approach, we assumed that the forward reaction rate (k_{cat}) will be related to the free energy barrier of the rate limiting step of the reaction ($\Delta G^{\ddagger}_{\text{rls}}$) according to the Eyring equation [150, 151]

$$k_{\text{cat}} = \left(\tau \frac{k_{\text{B}}T}{h}\right) e^{-\frac{\Delta G^{\ddagger}_{\text{rls}}}{k_{\text{B}}T}}, \tag{4.1}$$

where τ, k_{B}, T, and h are the transmission coefficient, Boltzmann constant, absolute temperature, and Planck constant, respectively. In our approach, we define $\Delta G^{\ddagger}_{\text{rls}}$ as the free energy of the folded, bound transition state (TS) relative to that of the folded, bound reactant state (RS). Note that in this study, we are primarily interested in relative improvement of mutant (Mut) activity to WT activity $\frac{k^{\text{Mut}}_{\text{cat}}}{k^{\text{WT}}_{\text{cat}}}$ via reduction of the relative free energy of activation $\Delta\Delta G^{\ddagger}_{\text{rls}}$, which translates to a favorable relative free energy difference of activation, assuming a negligible effect on the transmission coefficient upon mutation (Eq. 4.2)

$$\frac{k^{\text{Mut}}_{\text{cat}}}{k^{\text{WT}}_{\text{cat}}} = \left(\frac{\tau^{\text{Mut}}}{\tau^{\text{WT}}}\right) e^{-\frac{\Delta\Delta G^{\ddagger}_{\text{rls}}}{k_{\text{B}}T}}. \tag{4.2}$$

We calculated this relative free energy of activation using a two–state approach, in which the binding and folding free energy of the ground state, reactant structure as well as the binding and folding free energy of the transition state structure were calculated separately, for each possible mutant relative to WT. The free energy for a mutant or WT structure is defined as follows:

$$\Delta G^{\text{fold+bind}} = G^{\text{folded}}_{\text{complex}} - G^{\text{unfolded}}_{\text{receptor}} - G_{\text{ligand}}. \tag{4.3}$$

The relative free energy difference of mutant to WT is thus

$$\Delta\Delta G^{\text{fold+bind}} = \left(G^{\text{folded}}_{\text{Mutcomplex}} - G^{\text{folded}}_{\text{WTcomplex}}\right) - \left(G^{\text{unfolded}}_{\text{Mutreceptor}} - G^{\text{unfolded}}_{\text{WTreceptor}}\right), \quad (4.4)$$

where here, the unbound ligand free energy has cancelled. The difference between this expression for the TS and RS yields the relative free energy of activation of the mutant relative to WT

$$\Delta\Delta\Delta G^{\ddagger}_{\text{Mut-WT}} = \Delta\Delta G^{\text{fold+bind}}_{\text{TS}} - \Delta\Delta G^{\text{fold+bind}}_{\text{RS}}. \quad (4.5)$$

4.2.2 Structure Preparation

The crystal structure used in this study was a transition state analogue (N-hydroxy-N-isopropyloxamic acid) bound structure of the spinach KARI variant, obtained from the Protein Data Bank (PDB) [127] (accession code 1YVE) [143]. The structure was prepared with procedures outlined by Lippow *et al.* [19] using CHARMM [3, 76] with the CHARMM27 force field [77]. In all designs and simulations, only the chain A monomer was used to improve computational efficiency. All histidines were neutral and protonated as indicated to maximize hydrogen bonding potential: 103(A)-δ, 215(A)-δ, 226(A)-δ, 232(A)-δ, 280(A)-ϵ, 328(A)-ϵ, 484(A)-δ, 506(A)-ϵ, and 564(A)-ϵ. Non-active site crystalographic waters were removed if they failed to make at least three hydrogen bonds with the protein (using a maximum heavy atom hydrogen bond distance of 3.33 Å). A total of 61 water molecules remained (resids 72, 75, 87, 93, 106, 109, 179, 194, 379, 405, 429, 440, 474, 481, 838–841, 852, 862, 878, 883, 887, 894, 895, 941–949, 965, 967–969, 975, 998, 999, 1023–1025, 1032, 1072, 1089, 1093–1095, 1097, 1105, 1108, 1206, 1250, 1252, 1253, 1257, 1304, 1305, 1779).

4.2.3 Enzyme Redesign

Relative enzyme activity calculations were performed using a combined quantum mechanics/molecular mechanics (QM/MM) configurational search strategy. Configurational space of reactant and transition state ligands and active site residues were

initially explored using *ab initio* quantum theory. The remaining enzyme degrees of freedom were explored using a rotameric, molecular mechanics based approach. The initial QM search incorporated the ligand, magnesium centers, five magnesium coordinating water molecules, as well as the side chains of three surrounding active site residues, E319, D315, and E496 that are involved in ligand and magnesium coordination. In this model the E496 residue was protonated as previous studies had indicated its importance in reducing the computed activation energy barrier [148]. QM calculations were done *in vacuo* using the GAUSSIAN03 computer program [129] at the rhf/3-21g* level of theory. The reactant structure was found via ground state energy minimization (keyword OPT), and a transition state was found using a first–order saddle point search method (keyword QST3). This TS was validated by following the vibrational eigenmode corresponding to the single negative eigenvalue to ensure it connected the isomerization reactants and products. Once optimized and validated, each QM–derived structure was incorporated into the larger MM model of the enzyme and docked into the active site via aligning to the carbon backbone of the bound transition state analogue from the original crystal structure. This aligned structure underwent ten rounds of sliding, constrained minimization to properly fit the active site backbone to the QM optimized residues. All substrate, magnesium, and coordinating aspartate/glutamate oxygen atoms were held fixed during this minimization. The remaining active site residues were harmonically constrained using a force constant of 50 kcal/mol/Å. The sliding minimization procedure consisted of 100 steps of steepest descent minimization followed by 100 steps of adopted basis Newton–Raphson minimization, where the harmonic constraints were reset after each round of minimization. Partial atomic charges were fit to each QM atom using restrained electrostatic potential fitting methods of Bayly *et al.* [79].

The evaluation of each possible mutant was performed using a hierarchical, design strategy similar to the dead-end elimination and A* based approach of Lippow *et al.* [19]. Binding and folding energies for all possible single mutants within 10 Å of the bound ligand were computed in conjunction with a guaranteed, rotameric search of side chain degrees of freedom in both the bound complex and unfolded states. For

each single mutant examined, all side chains within 5 Å of the mutant residue were allowed to relax and included in the rotameric, conformational search. In this initial screen a pairwise decomposable energy function was used, accounting for changes in the pairwise coulombic, van der Waals, and geometric strain energies upon binding and folding. Top hits from each design were re-evaluated using a Poisson–Boltzmann Surface Area (PBSA) implicit solvent model, and the best, individual, minimum energy structures were compared to their homologous mutations in the transition state or reactant state design. Potential mutations were selected based upon their predicted relative ability to stabilize the transition state relative to the reactant state without significantly destabilizing the folded reactant state species. Note that in this design scheme protein and substrate configurational entropy changes upon activation are assumed to be constant across all mutants.

4.2.4 Transition Path Sampling

In order to assess sources of error in our original design, we determined the important, dynamic factors that influence enzyme activity and recomputed reaction rates using transition path sampling (TPS) [154, 155, 156, 34, 157, 158, 159, 160]. This method was used to collect unbiased ensembles of dynamic, reactive transitions from the reactant bound equilibrium state to the product bound equilibrium state as well as compute rates for the forward reaction. The collected transitions were sampled according to their underlying probability distribution, and were not dependent upon any assumed reaction coordinate. The basic theoretical foundations of TPS as it is used for ensemble generation and rate calculations are reviewed below, but for more detailed presentations we direct the reader to descriptions by P. G. Bolhuis and C. Dellago [155, 156, 34].

Ensemble Generation

In the TPS framework, path ensembles connecting the reactant basin to alternate regions of phase space are sampled using a Monte Carlo algorithm that samples

chains of points in phase space (i.e., paths) according to their relative probabilities. Assuming deterministic dynamics, any path, $\chi(t)$, is simply a sequence of points in phase space

$$\chi(t) \equiv \{\chi_0, \chi_1, \chi_2, ..., \chi_t\}, \tag{4.6}$$

and is uniquely determined by its starting point, χ_0. As such, the probability of any path in phase space, $\rho[\chi(t)]$, is equal to the probability of its initial point in phase space $\rho(\chi_0)$. For paths starting in an equilibrium basin, A, this initial probability is simply the equilibrium likelihood of χ_0

$$\rho(\chi_0) = \frac{e^{-E(\chi_0)/k_B T}}{Z_A}, \tag{4.7}$$

where $E(\chi_0)$ is the initial, total energy of the system, k_B is Boltzmann's constant, T is the absolute temperature, and Z_A is the partition function for the equilibrium basin, A.

Collecting an ensemble of paths that connects two regions, A and B, amounts to sampling the distribution $\rho_{AB}[\chi(t)] = h_A(\chi_0)\rho(\chi_0)h_B(\chi_t)$, where h_A and h_B are Heaviside step functions of phase space that equal 1 when the system occupies A and B, respectively. Given an initial trajectory of length τ that connects these regions, TPS provides a simple algorithm for the efficient sampling of this distribution. New trajectories are generated from an initial path through either shifting or shooting moves [155]. In the former, the path is shifted in time by moving backward or forward Γ time steps whilst keeping the length of the path constant. In the latter move, an isoenergetic velocity perturbation (δ) is applied to all atomic coordinates at some time point, χ_n such that $0 < n < \tau$, along the trajectory. This is done for each atom by drawing a new velocity vector from a Maxwell–Boltzmann distribution at the desired temperature, adding it to the current velocity, and scaling the sum such that the kinetic energy is unchanged. From this perturbed point, a new path of the same length is generated by projecting forward $\tau - n$ and backward n steps in time. The probability of this new, shifted or shot trajectory, $\chi'(t)$, is then compared to the

probability of the starting trajectory and accepted with the probability

$$p_{\text{accept}} = \min\left(1, \frac{h_A(\chi_0')\rho(\chi_0')h_B(\chi_t')P(\chi(t),\Gamma,n,\delta)}{h_A(\chi_0)\rho(\chi_0)h_B(\chi_t)P(\chi'(t),\Gamma,n,-\delta)}\right) \tag{4.8}$$

where $P(\chi(t),\Gamma,n,\delta)$ is the probability of attempting the shooting or shifting move, with parameters, Γ, n, and δ, from path $\chi(t)$ and $P(\chi'(t),\Gamma,n,-\delta)$ is the probability of attempting the reverse shooting or shifting move, with parameters Γ, n, and $-\delta$, from path $\chi'(t)$. As long as the probability of selecting Γ for forward and backward shifting moves is identical, and the probability of selecting n and δ is identical to that of n and $-\delta$ for shifting moves, detailed balance is maintained and paths are accepted according to their true relative probabilities [34].

Rate Calculations

We first consider a simple two–state system, A and B, with forward and reverse rate constants, k_{f} and k_{r}

$$A \underset{k_{\text{r}}}{\overset{k_{\text{f}}}{\rightleftharpoons}} B, \tag{4.9}$$

where A denotes a reactant bound enzyme complex and B denotes a product bound enzyme complex. Assuming a single dynamical bottleneck between A and B, the phenomenological forward reaction rate (k_{f}) for this two state system is related to the microscopic time correlation function, $C(t)$ [161, 162]

$$\dot{C}(t) \equiv \frac{\left\langle h_A(\chi_0)\,\dot{h}_B(\chi_t)\right\rangle}{\langle h_A(\chi_0)\rangle} \approx k_{\text{f}} e^{-t/t_{\text{rxn}}}, \tag{4.10}$$

where h_A and h_B are Heaviside step functions of phase space that equal 1 when the system occupies the equilibrium basins of A and B, respectively, and $\chi_{0/t}$ denote the system's initial location in phase space or at some time, t, later. Note that $\dot{C}(t)$ is simply the derivative of the probability that the system ends in the product basin, B, some time, t, after starting in the reactant basin, A. For times, t, greater than the molecular transit time of the reaction (t_{mol}), but much shorter than the time scale of the exponential approach of the system to equilibrium ($t_{rxn} = (k_{\text{f}} + k_{\text{r}})^{-1}$), the

exponential term in equation 4.10 will be approximately 1. Thus, we are left with a time correlation function that should grow linearly on time scales $t_{mol} < t << t_{rxn}$ with a slope equal to the forward rate constant.

In principle, for a given t, this correlation function can be computed by a windowing technique similar to umbrella sampling [163]. Using an order parameter, λ, that uniquely distinguishes the regions of phase space that correspond to basin A and basin B, one can divide phase space into a series of overlapping windows defined by overlapping ranges of λ that connect basin A to basin B. Using the transition path sampling method described above with a fixed path length of t, one can collect ensembles of transitions that start in basin A and end within each one of these λ-windows. From these ensembles, one can compute the normalized probability of observing any value of λ within each window. To obtain a full, normalized probability distribution of λ, $p(\lambda, t)$, one can scale neighboring windows relative to a single window. $C(t)$ can be calculated by integrating $p(\lambda, t)$ along λ over the portion of this distribution that corresponds to basin B

$$C(t) = \frac{\int_B p(\lambda, t) d\lambda}{\int_{-\infty}^{\infty} p(\lambda, t) d\lambda}. \tag{4.11}$$

This processes must then be repeated to accurately compute the derivative, $\dot{C}(t)$, and ensure that $C(t)$ is growing linearly in time. Given the computational requirements for even a single $C(t)$ calculation, however, it is more efficient to further decompose eq. 4.10 into a time dependent piece, $\dot{v}(t)$, and time independent piece, P,

$$\dot{C}(t) = \dot{v}(t)P = \left(\frac{\left\langle h_A(\chi_0)\, \dot{h}_B(\chi_t) \right\rangle}{\langle h_A(\chi_0)\, h_B(\chi_\tau) \rangle} \right) \left(\frac{\langle h_A(\chi_0)\, h_B(\chi_\tau) \rangle}{\langle h_A(\chi_0) \rangle} \right) \tag{4.12}$$

where τ is some time long enough such that $\dot{C}(t)$ has reached a plateau. Note that P is simply $C(t)$, evaluated when $t = \tau$. As such, it is the probability that the system is in the product state at time τ and can be found using the aforementioned windowing technique to compute $p(\lambda, \tau)$. $\dot{v}(t)$ is the probability of being in the product basin at time t, given that the system started in the reactant basin, normalized by the probability of being in the product basin at time τ, and its derivative is effectively

the τ-normalized rate of crossing the reactant barrier. This ratio can be found for all times $t' < t$ from a single transition path simulation, by ensuring that the trajectory ensemble includes those trajectories that visit the product basin at some point *within* t. Note that if they are required to end in the product basin at t, the trajectory ensemble collected by transition path sampling will ignore trajectories that make it to the product basin by t' but leave before t, yielding incorrect estimates of $\langle h_B(t' < t)\rangle$. To accomplish this, we introduce the path functional $H_B(\chi_\tau)$, which equals unity if the trajectory visits the product basin at some point along the trajectory χ_τ and is zero otherwise. Inserting this into Equation 4.12 yields

$$\dot{C}(t) = \dot{v}(t)P = \left(\frac{\left\langle h_A(\chi_0)\,\dot{h}_B(\chi_t)\,H_B(\chi_\tau)\right\rangle}{\langle h_A(\chi_0)\,h_B(\chi_\tau)\,H_B(\chi_\tau)\rangle}\right)\left(\frac{\langle h_A(\chi_0)\,h_B(\chi_\tau)\rangle}{\langle h_A(\chi_0)\rangle}\right), \tag{4.13}$$

which can be computed much more efficiently because $\dot{v}(t)$ can be calculated using a single TPS simulation.

Simulation Methods

All simulations were performed using the TPS module as implemented in CHARMM36 [164] compiled with SQUANTUM. Additional AM1 parameters for magnesium were added based on those used by J. J. P. Stewart [165]. Energies were evaluated using a combined QM/MM approach in which the active site was treated with semi-empirical quantum mechanics (AM1), and the remainder of the protein was treated with molecular mechanics (CHARMM27 force field). The QM treated active site included the substrate, acetolactate, both magnesium centers, five magnesium coordinating active site water molecules, side chains of E319, D315, and E496, and the nicotinamide group of NADPH. The QM/MM boundary atoms were treated using the Generalized Hybrid Orbital method [166] and included the Cα atoms of residues 319, 315, and 496, as well as the C5' atom of the ribose ring in NADPH that contains the nicotinamide moiety. Here, as in previous computational studies, the substrate was deprotonated and the coordinating E496 was protonated [148]. All dynamics were

run using a leapfrog integrator with a time step of 1 fs at 300 K.

In this study, the distance of the breaking C_4–C_5 bond ($r_{C_4-C_5}$) minus that of the forming C_7–C_5 bond ($r_{C_7-C_5}$) was used as an order parameter (λ) to distinguish reactants from products of the isomerization reaction. The reactant and product basins were defined as λ=[–2.0 , –0.15] and [0.15 , 2.0], respectively, with units of Å. The initial path used to seed the TPS simulation was found by computing a potential of mean force (PMF) along the order parameter λ using umbrella sampling and the weighted histogram analysis method (WHAM) [167]. This provided an estimate of the location of the transition region along λ, from which multiple configurations were collected. For each collected snapshot, the system was seeded with random momenta drawn from a Boltzmann distribution at 300 K, and projected forward and backward in time until the system fell into one of the two basins of attraction (reactants or products). This was repeated until a trajectory was found that connected the reactant basin to the product basin. The umbrella sampling was performed in CHARMM using the RXNCOR module with umbrellas 0.05 Å in width and harmonic constraints of 200 kcal/mol/Å^2.

In the calculation of $\dot{v}(t)$ and P, a path length (τ) of 101 fs was used, as this was found to be longer than the molecular transit time of the reaction and long enough to capture the linear growth of $\langle h_B \rangle$ as a function of time (*vide infra*). Path ensembles for $\dot{v}(t)$ were collected in two stages. The initial path was equilibrated by running 200, 1000–step simulations with different initial random seeds, and the final path from each simulation was then used as the seed path for another 200, 1000–step, production simulations. $\dot{v}(t)$ was computed by averaging over all of the independently collected paths ensembles, fitting the linear regime to a first–order polynomial and extracting the slope. Shifting and shooting moves were attempted with equal probability, the maximum shift length was 50 fs, and all time points were possible shooting locations.

For the calculation of P, the order parameter was partitioned into a large number of windows in order to accurately compute the relative probabilities of traversing order–parameter space. Given the high energetic barrier found from a potential of mean force (PMF) calculation along λ, very narrow windows were used to ensure

accurate sampling within each window. As the system surmounted the barrier ($\lambda <$ 0.05 Å), we employed windows 0.02 Å in length with an overlap of 0.01 Å. For all λ–windows beyond 0.05 Å, we used windows 0.1 Å in length with an overlap of 0.05 Å. Note that if windows along the uphill climb are too large, the relative probability difference between ending near the proximal edge of the window versus ending near the distal edge will be too great and poor relative probabilities will result. Ultimately, $p(\lambda)$ was found by binning the probability of ending at each value of the order parameter within each λ–window and scaling the nth window by the ratio of the probability of the overlapping region in window n-1 to that of window n. In this way, we were able to normalize the probability distribution of λ relative to the first window, and compute P using Equation 4.11. For each of these windows, shifting and shooting moves were attempted with equal probability, the maximum shift length was 5 fs, and only the final 10 fs were possible shooting positions to ensure non–zero move acceptance ratios. Finally, as in the calculation of $v(t)$, multiple independent simulations were run for each window. For each, 40–80 differentially seeded simulations were equilibrated for 1000 steps, followed by a 1000–step production run. Note that simulations were run in batches of 20, and the total number of per window simulations performed was dictated by the convergence of P for each enzyme variant (*vide infra*).

The same set of parameters described above were used for the generation of path ensembles and the calculation of rates for WT KARI as well as the four single mutants examined in this study: L323N, E488M, S276G, and E319D.

4.3 Results and Discussion

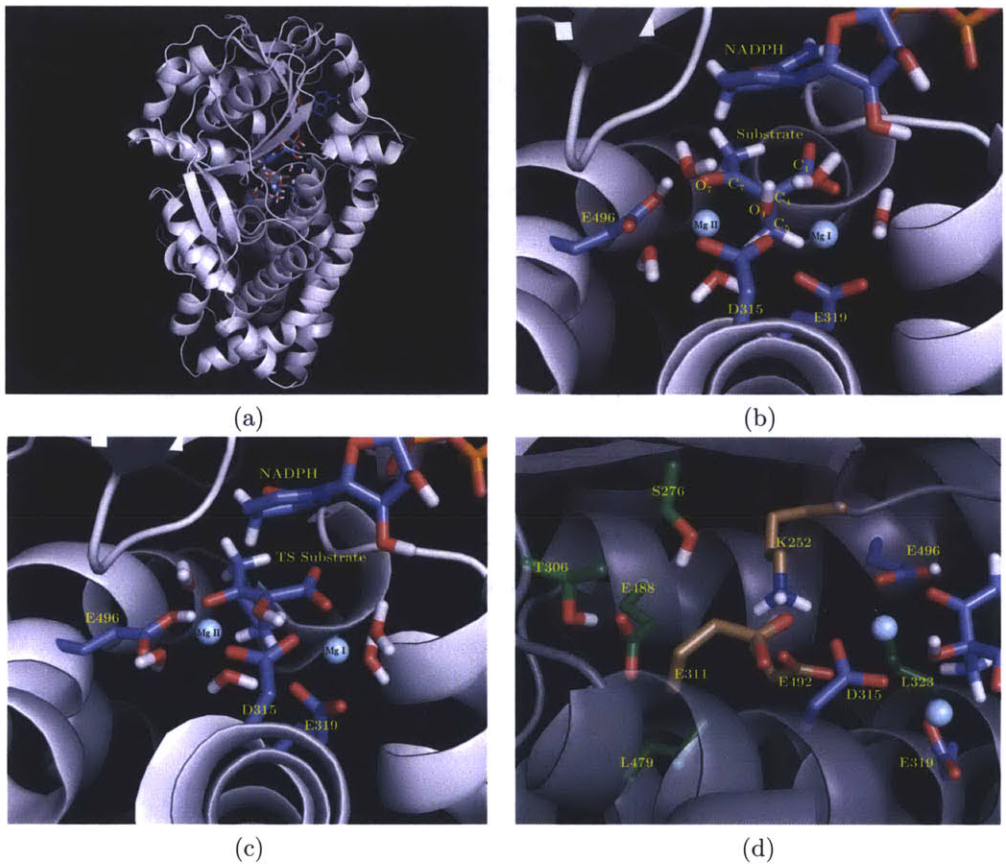
4.3.1 Reactant and Transition State Models

Our computational models of the KARI ground and transition states show subtle differences in how the bound substrate interacts with the active site and indicate a concerted isomerization and proton transfer mechanism. They also highlight the importance of the bound magnesium ions, as well as residues E496 and E315 in stabilizing the transition state. Examining the geometry of the substrate, we note significant differences between the reactant and transition states (Fig. 4-2). In the reactant state, we find that the C_7 carbon atom exhibits a trigonal planar geometry, consistent with its expected sp^2 hybridization, the C_4 carbon atom exhibits a tetrahedral geometry, consistent with its expected sp^3 hybridization, and the C_5 methyl group is bound to C_4. In the transition state, however, both C_4 and C_7 exhibit trigonal planar geometries, and the C_5 methyl group sits equidistant (1.9 Å) from both carbon atoms. Additionally, the C_7–O_7 bond expands from 1.2 to 1.3 Å and the C_4–O_4 bond shrinks from 1.4 to 1.3 Å as they interconvert between an alcohol and a ketone. We also observe differences in the relative coordination distances between the magnesium ions and their respective ligands, as well as between the substrate and the active site. In the ground state structure, we find that both of the magnesium ions are hexacoordinated by the substrate, five water molecules, and active site residues D315, E319, and E496. On average the coordination distance for these ligands is 2.2 Å for both metal ions. The substrate is held in the active site via electrostatic interactions with both of these magnesium ions, as well as through a hydrogen bond with the protonated sidechain of E496, which has a coordination distance of 2.0 Å. By comparison, in the transition state, coordination to the magnesium I ion by the substrate is disrupted by the transfer of a proton to D315, and the remaining ligands bind more tightly to their respective metal centers with average coordination lengths of 2.0 Å for both metal ions (Fig. 4-2c). Similarly, the substrate–E496 distance also shrinks by 0.2 Å, indicative of a stronger interaction. Taken together, these relative differences are suggestive of a highly concerted methyl transfer, ketone/alcohol iso-

merization, and ligand to enzyme proton transfer process. They also suggest that the magnesium ions, E496, and D315 play an integral role in stabilizing this concerted transition state. As the system undergoes both methyl and proton transfer reactions, the magnesium ions polarize and bind more tightly to the substrate and active site residues, D315 accepts the substrate proton, and the protonated E496 residue stabilizes the build up of negative charge on O_7, which in our models decreases from –0.36e to –0.50e. We note that much of this picture is consistent with previous studies of the isomerization reaction mechanism [148, 137, 149]. These investigations posit similar roles for the magnesium ions as well as E496, and also suggest an active site residue as a possible proton sink.

4.3.2 Computational Enzyme Designs

Using our WT reactant and transition state models, we exhaustively explored all possible single mutants at all possible positions within 10 Å of the active site. In our initial single mutant search, we found three mutations (E488M, S276G, and L323N) that were predicted to reduce the activation energy of the isomerization reaction. Each is predicted to improve $\Delta G^{\ddagger}$ relative to WT primarily by disrupting the folding stability of the reactant state more than that of the transition state (Table 4.1). Interestingly, the E488M and S276G mutants are one layer removed from the active site (Fig. 4-2d) but are still predicted to reduce the activation energy by differentially interacting with active site residues. Specifically, E488M is predicted to significantly improve the local van der Waals packing in the TS relative to the RS, at the cost of losing hydrogen bonds with neighboring residues H484 and T306 as well as favorable long–range electrostatic interactions with K252 and both magnesium ions. These losses are mitigated in the TS relative to the RS, with the net effect being a destabilized mutant ground state with a favorable, relative TS–RS free energy difference. By contrast, the S276G mutant experiences comparable losses to local van der Waals packing in both TS and RS, but favorable relative electrostatic differences due to relative changes in the folding desolvation penalties paid by residues G276, K252, and E488. As with E488M, the net effect is a favorable relative free energy of



(a) (b) (c) (d)

Figure 4-2: **Reactant and Transition State Design Models** (a) Buried active site of the spinach KARI monomer in the complexed reactant state (b) Ground–state, active site model bound to acetolactate, magnesium ions, and NADPH. (c) Transition state, active site model bound to acetolactate, magnesium ions, and NADPH. (d) Ground state structure of KARI active site with neighboring residues predicted to improve relative activity. Mutant positions are shown here in green, neighboring residues that mediate second shell mutation effects are shown in tan, and active site residues are shown in blue.

activation via selective, ground state destabilization. Finally, L323N is an active site mutant in contact with the substrate and catalytic residues E319, E496, and D315. It is predicted to improve the the TS–RS gap primarily through differential van der Waals packing of the mutant side chain with residue E496, which directly interacts with the TS and RS.

In order to reduce the electrostatic and van der Waals folding penalties observed for these single mutants, secondary mutations were also explored to stabilize the RS. To this end, the double mutants T306M/E488A, S276G/L479M, and E488M/T306G were selected and assayed for potential improvements to activity (Table 4.1). Similar to the E488M single mutant, the T306M/E488M and T306G/E488A double mutants retained their relative improvement of the local van der Waals packing in both TS and RS. More importantly, however, they significantly reduced the destabilizing, electrostatic penalties paid by the single mutant alone (from approximately 8 to 2 kcal/mol). This was accomplished by also removing one of WT E488's hydrogen bonding partners, T306, which reduced folding desolvation penalties. Interestingly, these double mutants also improved local electrostatic packing such that the neighboring active site residues (K252, D315, E311, E319, and E492) made more favorable interactions compared to the single mutant case. Similar to the initial designs, however, the favorable predicted effect on catalysis stemmed from selective destabilization of the ground state. Finally, the S276G/L479M mutant was predicted to maintain the favorable relative electrostatic stabilization of the S276G mutation, and reduced the 3 kcal/mol van der Waals penalty by introducing a fold stabilizing, but catalytically neutral mutation (L479M). Ultimately, it was predicted to stabilize the TS more than it destabilized the RS.

These six mutants were synthesized and assayed experimentally by measuring both their specific activity as well as the activation energy of the overall reaction using acetolactate as the substrate. We find that the mutants designed using this single structure, end–point method show varying activities between 33 and 66% of WT (Table 4.2), inconsistent with our predictions. Using the Eyring equation (4.1) to further decompose these activities into component enthalpies and entropies of ac-

		Total	Elec	vdW	Geo	SASA
E488M	$\Delta\Delta G_{TS}$	1.56	7.71	–9.6	3.02	0.43
	$\Delta\Delta G_{RS}$	2.75	8.53	–8.32	2.11	0.43
	$\Delta\Delta G^{\ddagger}$	–1.19	–0.82	–1.28	0.91	0.0
S276G	$\Delta\Delta G_{TS}$	0.29	–3.72	3.09	0.68	0.23
	$\Delta\Delta G_{RS}$	1.60	–2.5	3.16	0.69	0.24
	$\Delta\Delta G^{\ddagger}$	–1.31	–1.22	–0.07	–0.01	–0.01
L323N	$\Delta\Delta G_{TS}$	–2.36	1.7	–3.56	–0.66	0.16
	$\Delta\Delta G_{RS}$	5.92	–2.82	7.78	0.58	0.38
	$\Delta\Delta G^{\ddagger}$	–8.28	4.52	–11.34	–1.24	–0.22
T306M/ E488A	$\Delta\Delta G_{TS}$	1.06	1.03	–1.32	1.31	0.04
	$\Delta\Delta G_{RS}$	2.39	3.37	–3.15	2.14	0.03
	$\Delta\Delta G^{\ddagger}$	–1.33	–2.34	1.83	–0.83	0.01
T306G/ E488M	$\Delta\Delta G_{TS}$	1.75	0.74	–1.6	2.43	0.18
	$\Delta\Delta G_{RS}$	2.41	1.98	–2.55	2.81	0.17
	$\Delta\Delta G^{\ddagger}$	–0.66	–1.24	0.95	–0.38	0.01
S276G/ L479M	$\Delta\Delta G_{TS}$	–0.73	–2.88	0.42	1.53	0.18
	$\Delta\Delta G_{RS}$	0.58	–1.59	0.43	1.54	0.58
	$\Delta\Delta G^{\ddagger}$	–1.31	–1.29	–0.01	–0.01	–0.01

Table 4.1: Computed folding and binding energies in transition and reactant states for KARI mutants. Elec values are solvent screened electrostatic energies, vdW values are pairwise van der Waals energies, Geo values are covalent strain energies, and SASA values are solvent accessible surface area energies. All values are reported in kcal/mol relative to wild type.

tivation, however, we observe that while the free energy of activation was higher than WT, most of these mutants actually reduced the enthalpy of activation by 0.5–2 kcal/mol. The net reduction in activity was the result of compensatory increases in the entropy of activation by 1–3 kcal/mol. Thus, our end–point design based on selective destabilization of the ground state was successful in improving the relative, TS/RS minimum energy configurations. However, this destabilization, whether mediated through van der Waals or electrostatic interactions, unfavorably disrupted the configurational distributions in one or both states and increased the entropic barrier to activation.

	Rel. Activity	$\Delta G^{\ddagger}$	$\Delta H^{\ddagger}$	$-T\Delta S^{\ddagger}$
WT	1	0	0	0
E488M	0.66	0.75	–2.3	3.0
S276G	0.33	1.0	–1.5	2.5
L323N	0.60	n.m.	n.m.	n.m.
T306M/E488A	0.52	0.39	–0.81	1.2
T306G/E488M	0.52	0.95	–0.51	1.5
S276G/L479M	0.65	0.49	0.17	0.32

Table 4.2: Specific activities and activation parameters for the combined isomerization and reduction reactions. All activities are shown as a fraction of WT activity, and activation free energies are shown in kcal/mol, relative to WT fit values. n.m., not measured.

4.3.3 Transition Path Sampling Calculated Rates

Given the success of our end–point design method at reducing the enthalpy of activation, but inability to account for entropic effects, we re-evaluated a subset of our mutants using transition path sampling, an ensemble method that captures the dynamics of the reaction, to both improve our predictive capacity and elucidate the source of the reduced activity. Using TPS, ensembles of reactive trajectories were collected and used to compute reaction rates for the isomerization reaction (Table 4.3). We found that TPS derived rates correctly showed reduced activity of the mutants relative to WT, and they were appropriately clustered compared to experientially

measured rates of the complete reaction. The WT enzyme is computed to be the most active variant, and E319D, a known inactivating mutation [149], was computed to be 6 orders of magnitude less active than WT. The remaining mutants all fell in between E319D and WT with E488M and L323N predicted to be more active than S276G, which matched the experimental results, although with the order of E488M and L323N switched. The improved predictive power of this method indicates that TPS effectively captures the ensemble effects ignored in our original evaluation and can correctly discern active from inactive enzymes (WT vs. E319D), as well as active from partially active (WT vs. L323N/E488M/S276G). Interestingly, we find that the absolute rates computed with this method are many orders of magnitude smaller than those measured by experiment, but consistent with the literature overestimates of the barrier height as found using transition state theory with similar force fields [148].

	Rel. Exp. Rate	Rel. Calc. Rate	Exp. Rate	Calc. Rate
WT	1.0	1.0 ± 0.2	1.34	$2.9 \pm 0.4 \times 10^{-12}$
E488M	0.66	$2.2 \pm 0.5 \times 10^{-2}$	0.89	$6.5 \pm 0.1 \times 10^{-14}$
L323N	0.60	$4 \pm 2 \times 10^{-2}$	$0.54^{\dagger}$	$1.2 \pm 0.6 \times 10^{-13}$
S276G	0.33	$1.1 \pm 0.7 \times 10^{-2}$	0.45	$3.1 \pm 0.2 \times 10^{-14}$
E319D	0.0	$1.2 \pm 0.2 \times 10^{-6}$	$0^{\ddagger}$	$3.6 \pm 0.5 \times 10^{-18}$

Table 4.3: Experimental and Calculated Rates for KARI Variants. All absolute values have units of s^{-1}, and calculated errors correspond to one standard error of the mean. †Collected in a separate series where WT activity was measured to be 0.9. ‡This value is drawn from Dumas *et al.*, and was shown to be an inactivating mutant in the presence of magnesium [149]

The high degree of correlation with experiment observed in this study suggests that the TPS framework can be accurately applied to large scale enzymatic systems and used to resolve the relative activities of single enzyme mutant libraries. While previous studies have used TPS to examine reaction dynamics [160, 158, 159], to our knowledge, this is the first time TPS has been used to compute enzyme activities. Additionally, given the significant improvement in predictive power, it suggests that TPS can effectively capture the ensemble effects ignored in our original, static evaluation and discern small differences in relative activity.

4.3.4 Convergence of Transition Path Sampling Calculated Rates

To assess the convergence of the TPS computed rates, we examined the convergence of $\dot{v}(t)$ as well as P as a function of the number of samples included in the ensemble average. For $\dot{v}(t)$ we observed rapid convergence with small fluctuations in the normalized frequency factor after averaging over 2×10^5 trajectories for all five of the KARI variants explored (Fig. 4-3). In all cases the mean of $\dot{v}(t)$ converged to within 3% of its final value within 1×10^5 trajectories. We observed much slower convergence for P (Fig. 4-4), as it depends upon the convergence of each λ window, and P converged to within 15% of its final value only within the last 5,000 samples collected. We note here that the value of P was many orders of magnitude smaller than that of the frequency factor, $\dot{v}(t)$, consistent with previous TPS studies [168, 169].

4.3.5 Comparison of $\dot{v}(t)$ and P for Wild Type and Mutant Enzyme Variants

Comparing the converged values of $\dot{v}(t)$ and P for WT and mutant variants, we find that the primary difference between calculated rates originates from P rather than $\dot{v}(t)$. Examining $\dot{v}(t)$ as a function of time (Fig. 4-5a), we see that for all variants, it starts and remains at zero for some time (approximately 12 fs), which corresponds to the minimum time necessary for the system to cross from the reactant to product basin. It then rises rapidly and plateaus after some molecular transit time ($t_{mol} \approx 60$ fs), validating our use of 101 fs as a path length (τ) long enough to capture reactive transitions. Additionally, we find that the average time spent in the transition region, i.e. after the system has left the reactant basin ($\lambda < -0.15$ Å) but before it has entered the product basin ($\lambda > 0.15$ Å), is approximately 28 fs for both WT and mutants. This value is consistent with the transient times found for other enzyme catalyzed reactions examined by TPS, but it falls on the shorter end of the spectrum. It is slower than the 10 fs observed for the hydride transfer reaction catalyzed by lactase dehydrogenase [157, 158], consistent with the fact that a more massive methyl

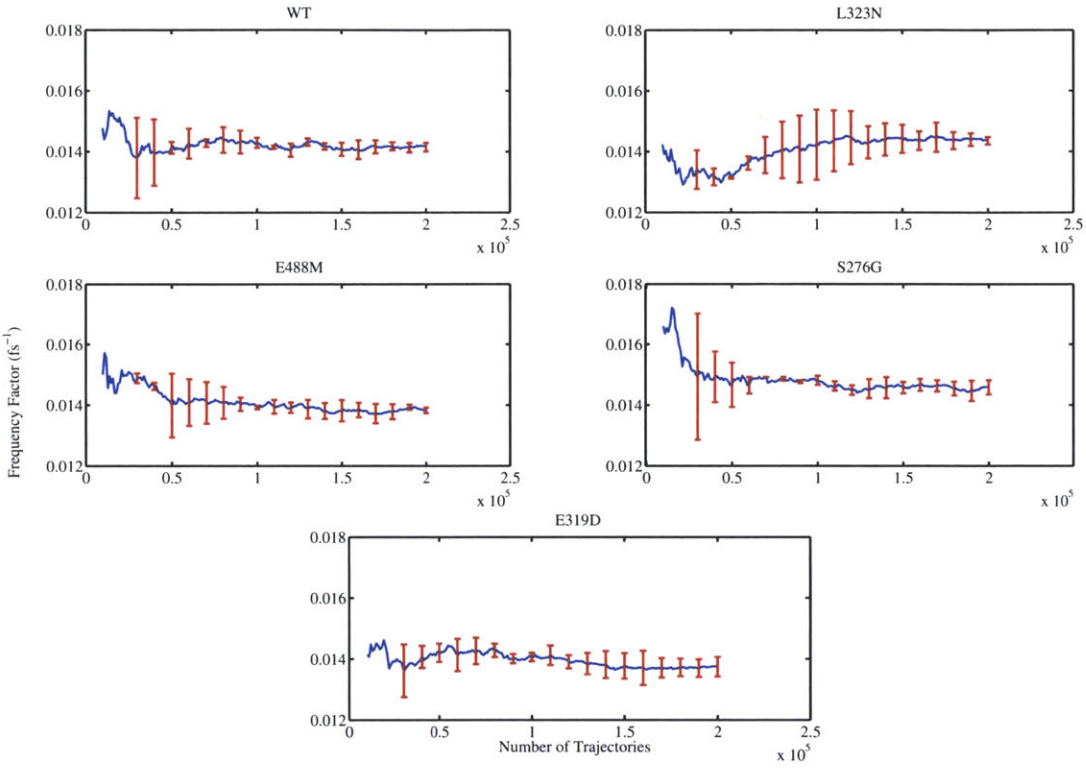


Figure 4-3: **Convergence of rate frequency factor,** $\dot{\nu}(t)$. Convergence is shown as a function of the number of trajectories included in the average. Errors correspond to one standard error and were computed at each point by separating the collected trajectories into initial and final halves and computing a standard error of the mean for these two measurements. All values are reported in units of fs^{-1}.

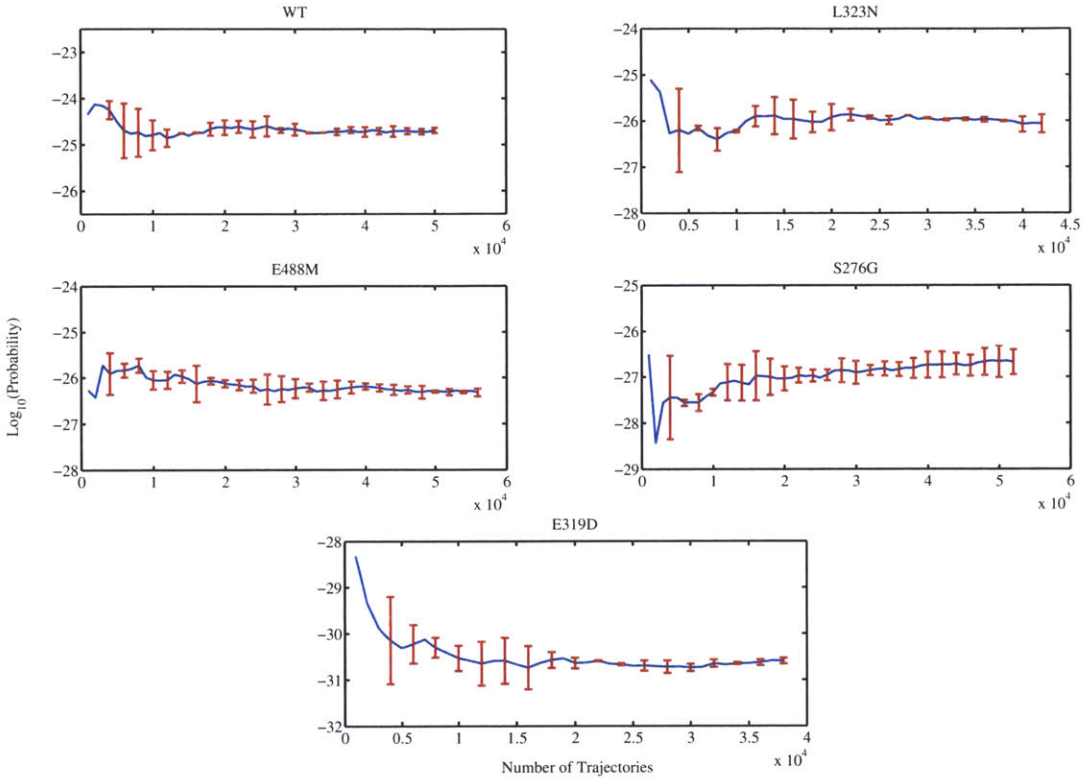


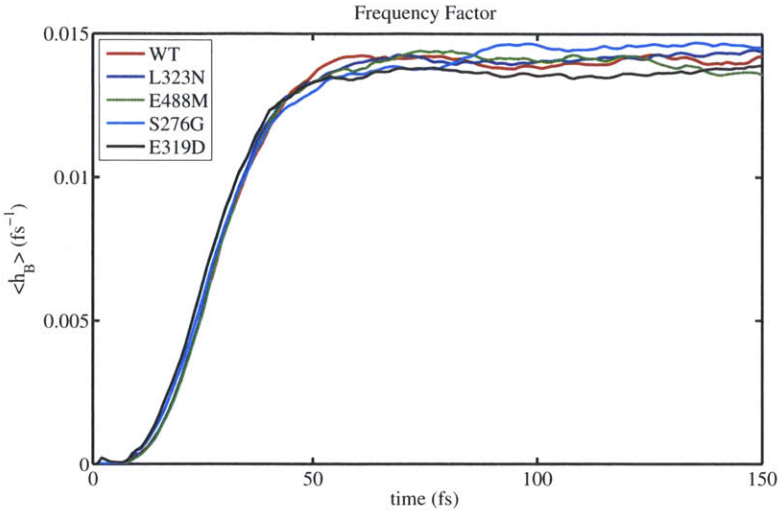
Figure 4-4: **Convergence of the log of the rate probability factor,** P. Convergence is shown as a function of the number of trajectories included in the average per λ-window. Errors correspond to one standard error and were computed at each point by separating the collected trajectories into initial and final halves and computing a standard error of the mean for these two measurements. Note that the probability span shown for each plot is identical (4 orders of magnitude), and the number of trajectories included in the average was dictated by the 15% error convergence criteria.

group is being transferred in this case, but notably faster than the 100 fs and 130 fs observed for the phosphorolysis catalyzed by human purine nucleoside phosphorylase [159] and Claisen rearrangement catalyzed by chorismate mutase [160], respectively. Thus, we see that there are negligible differences in both the transient time of the reaction as well as $\dot{v}(t)$ between enzyme variants, which implies that, on average, the rate at which each system enters the product basin along reactive trajectories is not significantly affected by these mutations.

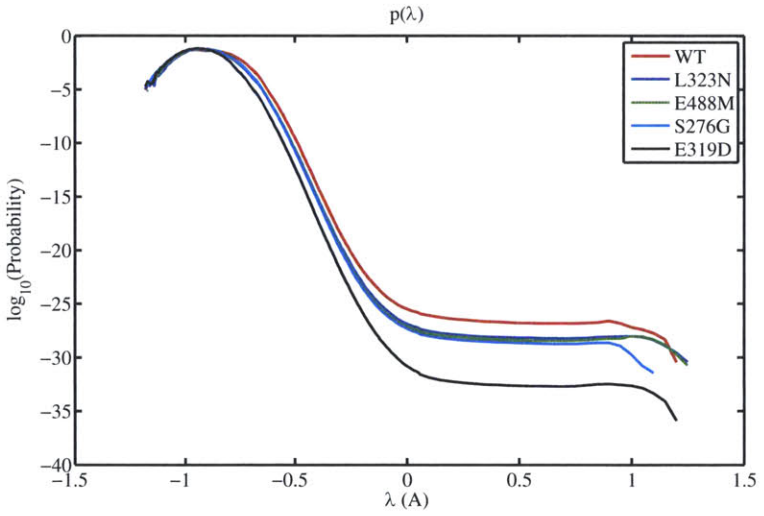
By comparison there are large differences in the probability of reaching the product basin in time τ given that each system starts in the reactant basin. In Figure 4-5b, we see that the WT system has a higher probability of making it across the reaction barrier and into the product state compared to the mutants. The same transition is slightly less likely for E488M and significantly less likely for the remaining mutants. Perhaps most interestingly, however, we also observe that the WT system (the most active variant considered) is more likely to reach almost every value of the order parameter outside of the high probability region of reactant basin ($\lambda > -0.7$ Å) compared to the less active mutants. These data show that WT is more active than mutant variants because the probability of reaching the product basin from the reactant basin in time τ is higher. Put another way, they suggest that reactive points in phase space (i.e., those coupled configuration/momentum states that reach the product basin in time τ) are more probable in highly active enzymes compared to less active or inactive variants. Furthermore, these data indicate that fluctuations that carry the system close to the product basin but not across the barrier are also stabilized in active enzymes — the most active enzymes are those that can facilitate fluctuations both toward as well as across the reactive barrier.

4.3.6 Reactive Trajectory Ensembles

To determine the dynamic, structural factors that led to the observed differences in P, we sought to answer the question of why some trajectories made it farther along the order parameter coordinate and crossed the barrier while others failed in the same amount of time. To this end we examined ensembles of WT and mutant trajectories



(a)



(b)

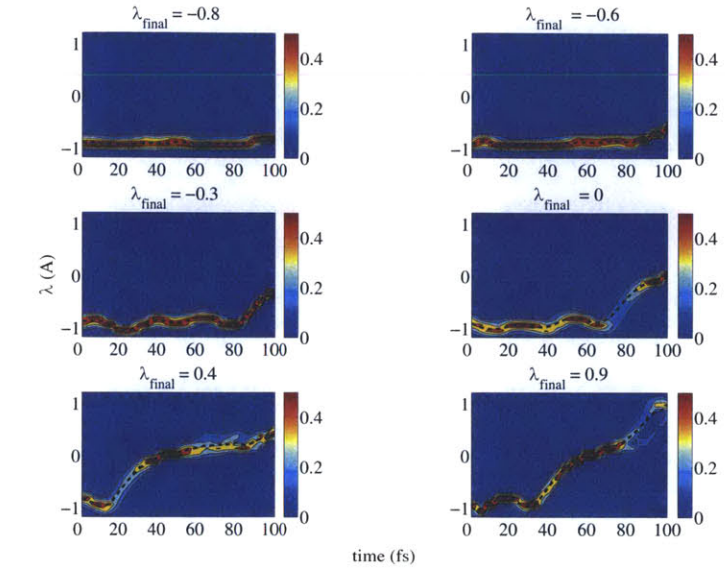
Figure 4-5: **Enzyme $\dot{v}(t)$ and P components** (a) $\dot{v}(t)$ as a function of time for WT and mutant enzymes. A normalization time (τ) of 101fs was used to compute $\dot{v}(t)$ and P. (b) The probability of transitioning from the reactant basin to values of the order parameter, λ, that separate reactants from products for WT and mutant enzymes.

that connected the reactant basin to successive points along the order parameter coordinate in identical amounts of time (101 fs). Comparing the value of the order parameter (λ) and its velocity for trajectories that fluctuated out of the reactant basin but did not cross the reaction barrier, we observed three distinct dynamic stages (Fig. 4-6). First, as the system fluctuated out of the reactant basin, it quickly accelerated along the order–parameter coordinate. Second, within approximately 20 fs of leaving the initial basin, the velocity peaked and began to slow as the system climbed the barrier. Finally, the velocity eventually reached zero by the end of the trajectory. Comparatively, those trajectories that made it across the barrier underwent similar initial dynamics, but differed in the final stage in that they retained enough forward momentum along the order parameter coordinate after approaching the top of the barrier to coast across before rapidly descending into the product basin.

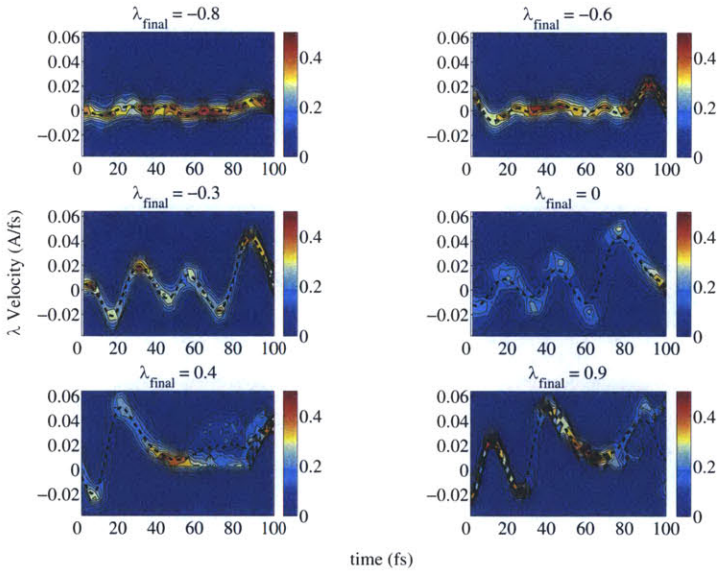
We find that the rapid acceleration and subsequent deceleration of the order parameter corresponded directly to the dynamics of one of its components, the C_4–C_5 bond distance that ultimately breaks during the isomerization reaction (Fig. 4-7). The initial acceleration of the order parameter toward the product basin corresponded to this bond shrinking, and returning to equilibrium from a compressed state. As it stretched past equilibrium ($\langle r_{C_4-C_5} \rangle = 1.54$ Å), the bond restoring force grew, causing the bond and the order parameter to decelerate and their respective velocities to eventually peak. In cases where the system fell short of the barrier, this restoring force was strong enough to prevent the bond from breaking and halt further bond extension. This resulted in an order parameter velocity of approximately 0 by the end of the trajectory. By comparison, in cases where trajectories did cross the barrier, the restoring force was not strong enough to prevent bond breaking and simply slowed the reaction by mediating the conversion of kinetic into potential energy. The net result was a decreased, but positive velocity, and a slow crossing of the barrier region.

4.3.7 C_4–C_5 Bond Breaking Dynamics

Examining the dynamics of the C_4–C_5 bond vibration in reactive and non-reactive trajectories, we find that the C_4–C_5 bond appears to have undergone vibrational

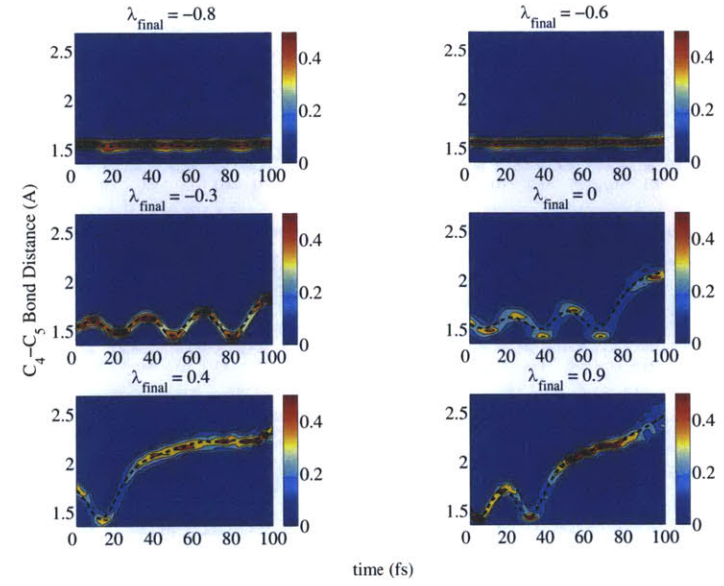


(a)

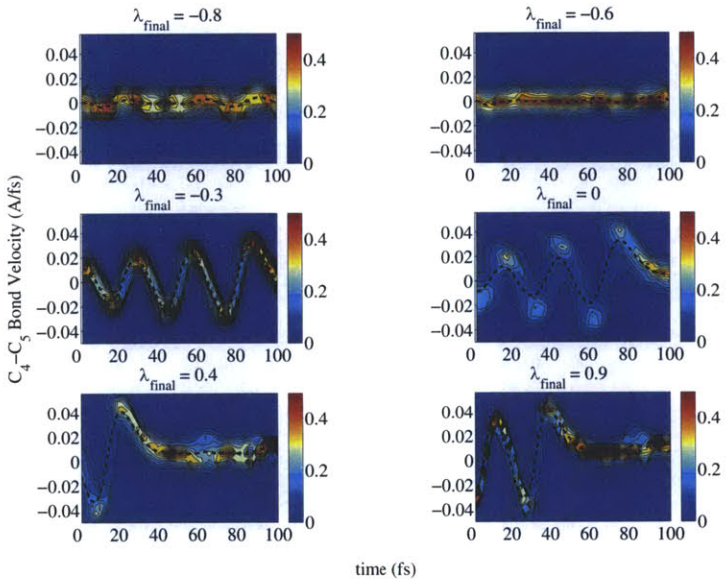


(b)

Figure 4-6: **Order parameter probability distributions for WT trajectory ensembles**. (a) Probability distributions of the order parameter for trajectories that ended at λ = –0.8, –0.6, –0.3, 0, 0.4, and 0.9 Å. The dashed line shows the mean value as a function of time. (b) Corresponding probability distributions of the velocity of the order parameter. Note that for ensembles that crossed the transition barrier (λ = [–0.15 , 0.15] Å), trajectories were aligned such that they cross the λ = 0 Å point at the same time.



(a)



(b)

Figure 4-7: **C_4–C_5 bond distance probability distributions for WT trajectory ensembles**. (a) Probability distributions of the C_4–C_5 bond distance for trajectories that ended at λ = –0.8, –0.6, –0.3, 0, 0.4, and 0.9 Å. The dashed line shows the mean value as a function of time. (b) Corresponding probability distributions of the velocity of the C_4–C_5 bond. Note that for ensembles that crossed the transition barrier (λ = [–0.15 , 0.15] Å), trajectories were aligned such that they cross the λ = 0 Å point at the same time.

resonant energy transfer immediately preceding departure from the reaction basin. In each of multiple, sequential fluctuations up the barrier, the bond vibrated with successively larger amplitude and velocity (Fig. 4-7). Additionally, comparing trajectories that progressed farther along the order–parameter coordinate (e.g., $\lambda = -0.6$ vs. $\lambda = 0$), these vibrations became increasingly energetic before the bond eventually broke upon crossing the barrier. Individual trajectories in each of these ensembles resembled a classical, forced, undamped, harmonic oscillator undergoing resonance with a neighboring vibration. To quantify this relationship, we fit individual time traces of the C_4–C_5 bond to the model

$$x(t) = \frac{F_0}{m(\omega_0^2 - \omega^2)} cos(\omega t + t_0) + Acos(\omega_0 t + t_0) + Bsin(\omega_0 t + t_0) + x_0, \qquad (4.14)$$

which is the general solution to the canonical, periodic forced harmonic oscillator equation [170]

$$\frac{d^2x}{dt^2} = -m\omega_0^2 x + F_0 cos(\omega t). \qquad (4.15)$$

We note here that F_0 (amplitude of the external force), ω_0, ω (natural and forcing frequencies, respectively), A, B (intial condition constants of integration), t_0, and x_0 (time phase and distance offset factors), were all free to vary in the fit.

This model fits well to the data, with an average r^2 value greater than 0.85 for each ensemble examined, suggesting that fluctuations toward the product basin were consistent with being driven by resonant energy transfer into the C_4–C_5 bond from some neighboring vibration. Additionally, we found that as sequential trajectory ensembles carried the system closer to the product basin, the degree of resonance increased. The difference between the forcing and natural frequency shrunk as the forcing frequency approached the natural frequency from above, a redshift of approximately 5 THz, and the amplitude of the mass normalized, external force increased by nearly an order of magnitude from 0.5 to 3.5 $\times 10^{-3}$ Å/fs^{-3} (Fig. 4-8). The vibrational frequencies of the C_4–C_5 bond dynamics at equilibrium had strong contributions at approximately 32 and 37 THz (Fig. 4-9a), consistent with the frequencies predicted by our simulations for trajectories that only move very short distances along the order–parameter

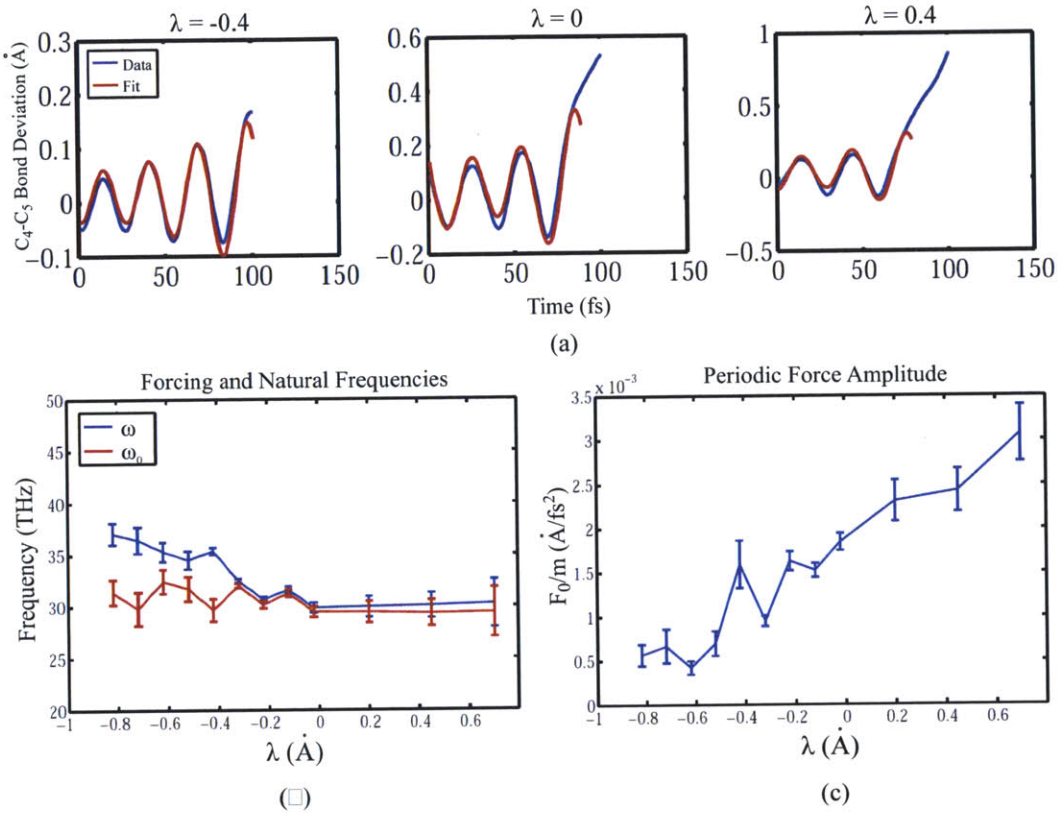


Figure 4-8: **Fitting results of WT trajectory traces to a forced harmonic oscillator model**. (a) Example fits of the model to the data at three locations along the order parameter coordinate ($\lambda = -0.4, 0, 0.4$ Å). Each trace was fit to all time points up to 20 fs after the final bond compression to avoid fitting a classical model to a non-classical bond breaking event. (b) Average parameter fits for the natural and forcing frequencies. (c) Mass normalized, external forcing amplitude as a function of λ. Error bars in panels b and c correspond to one standard error.

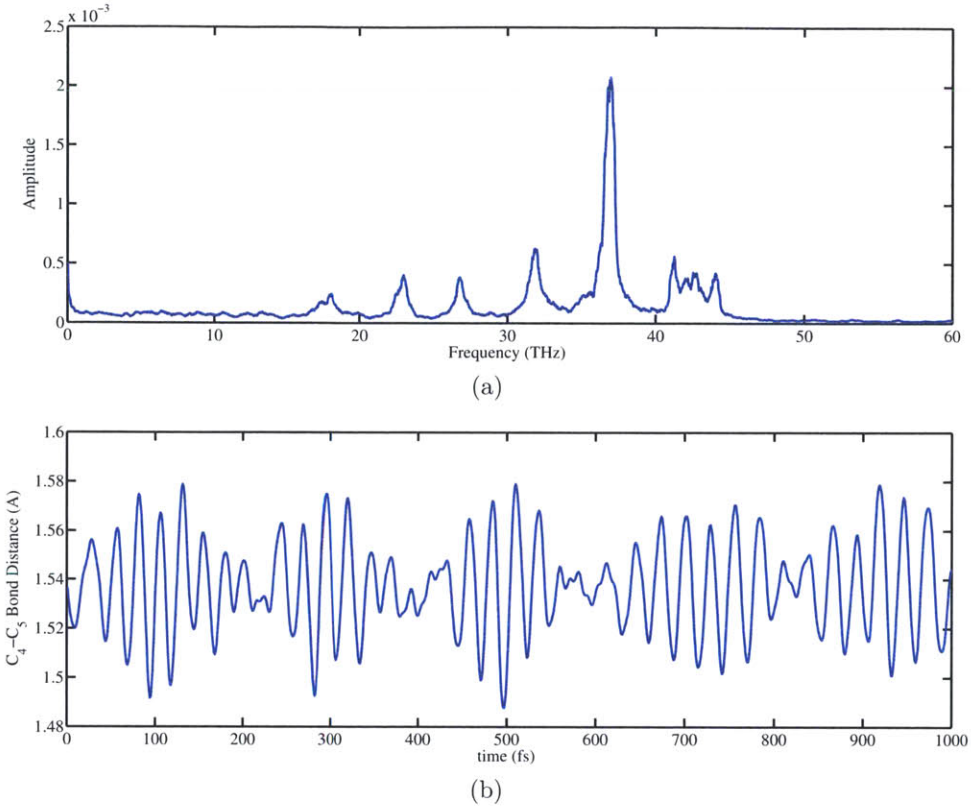


Figure 4-9: **Equilibrium WT C_4–C_5 bond vibrations**. (a) C_4–C_5 bond vibration frequencies found via a Fourier transform of 50 ps of equilibrium C_4–C_5 dynamics. From left to right, resolved peaks are centered at 18, 23, 27, 32, 37, and 41–44 THz. (b) Trace of the C_4–C_5 bond distance oscillating while displaying a beating pattern over a period of 1 ps.

coordinate ($\lambda = -0.8$ Å). We also found that this bond exhibited transient beating patterns at equilibrium (Fig. 4-9b), suggesting similar resonant oscillations to those observed in reactive trajectories, albeit with much smaller amplitude vibrations. The transient nature of the beating pattern observed at equilibrium and the frequency shift seen in reactive trajectories indicates that this is a dynamic phenomenon. As such, it is likely dependent on other degrees of freedom that create an environment conducive to resonant energy flow.

Similar results were obtained for all mutants examined in this study. Each showed C_4–C_5 bond dynamics that appear in resonance with an external force as the system

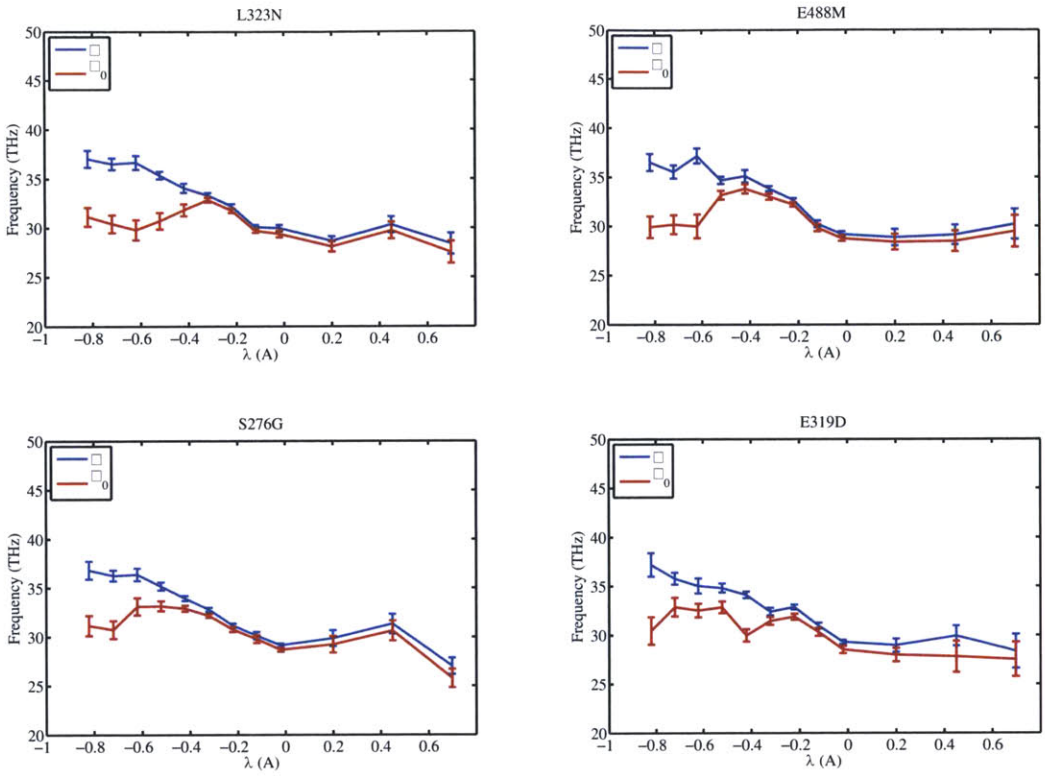


Figure 4-10: **Average natural and forcing frequencies of mutant trajectory traces fit to a forced harmonic oscillator model**. Each trace was fit to all time points up to 20 fs after the final bond compression to avoid fitting a classical model to a non-classical bond breaking event. Error bars correspond to one standard error.

moved farther along the order–parameter coordinate (Fig. 4-10), as well as similar, equilibrium C_4–C_5 bond frequencies and dynamics (data not shown). We note, however, that for the mutants examined in this study, the fitted natural frequencies were far less stable, and shifted toward the blue to meet the redshifting forcing frequency for intermediate, non-reactive trajectories, indicating that in mutant enzyme simulations there were dynamic shifts in the vibrational frequencies of both the breaking bond as well as the external force.

In exploring additional geometric coordinates that influence the C_4–C_5 bond during the reaction, we found that conserved, active–site residue E319 effectively pushed the C_5 methyl group out of the reactant well and directed it toward its position in

the product. As the system traversed the reaction coordinate and the C_4–C_5 bond became vibrationally activated, the distance between C_5 and one of the carboxylate oxygens of the E319 side chain shrunk before reaching a minimum when the C_4–C_5 bond was maximally extended during its penultimate oscillation (Fig. 4-11A). At this point, there was a van der Waals clash between the two atoms and the C_5 methyl group ricocheted off of the side chain. Examining the ensemble averages of the C_5–O_{E319} distance for transitions in the WT and mutants systems, the average minimum distance is approximately 2.7 Å, and a large, unfavorable VDW interaction appeared at this minimum value for all but the E319D mutation (Fig. 4-11b,c). Notably, the E319D mutation removed the E319 wall and replaced it with a smaller, aspartate side chain that had little interaction with substrate. The C_5–O_{E319} distance never dropped below 4 Å, and relative to other mutants there was very little van der Waals interaction (2 kcal/mol, compared to 20 kcal/mol in WT) between the two groups.

Thus, we propose that the dynamics of the KARI isomerization reaction are dominated by a two–stage, "pump–and–push" mechanism in both WT and mutant variants (Fig. 4-12). Energy is pumped into the C_4–C_5 bond via resonant energy transfer from an externally modulated, driving vibration. This yields an excited C_4–C_5 bond that becomes bumped by the E319 active–site side chain residue, pushing the substrate toward the product configuration. If enough energy has been driven into the C_4–C_5 bond and the C_5–O_{E319} collision is sufficiently forceful, the bond stretches and eventually breaks as the system crosses the reaction barrier. Given this conserved mechanism, KARI enzyme activity is proportional to the probability that this mechanism occurs (P); thus, the calculated differences in activity are reflective of the differences in this probability. In particular, it implies that mutant enzymes are less able to redshift the external forcing frequency and/or push the activated C_4–C_5 bond. In the case of mutant E319D, we hypothesize that its inactivity arises because it cannot effectively push the activated bond in the right direction, as the E319D mutation removes the side chain wall present in WT.

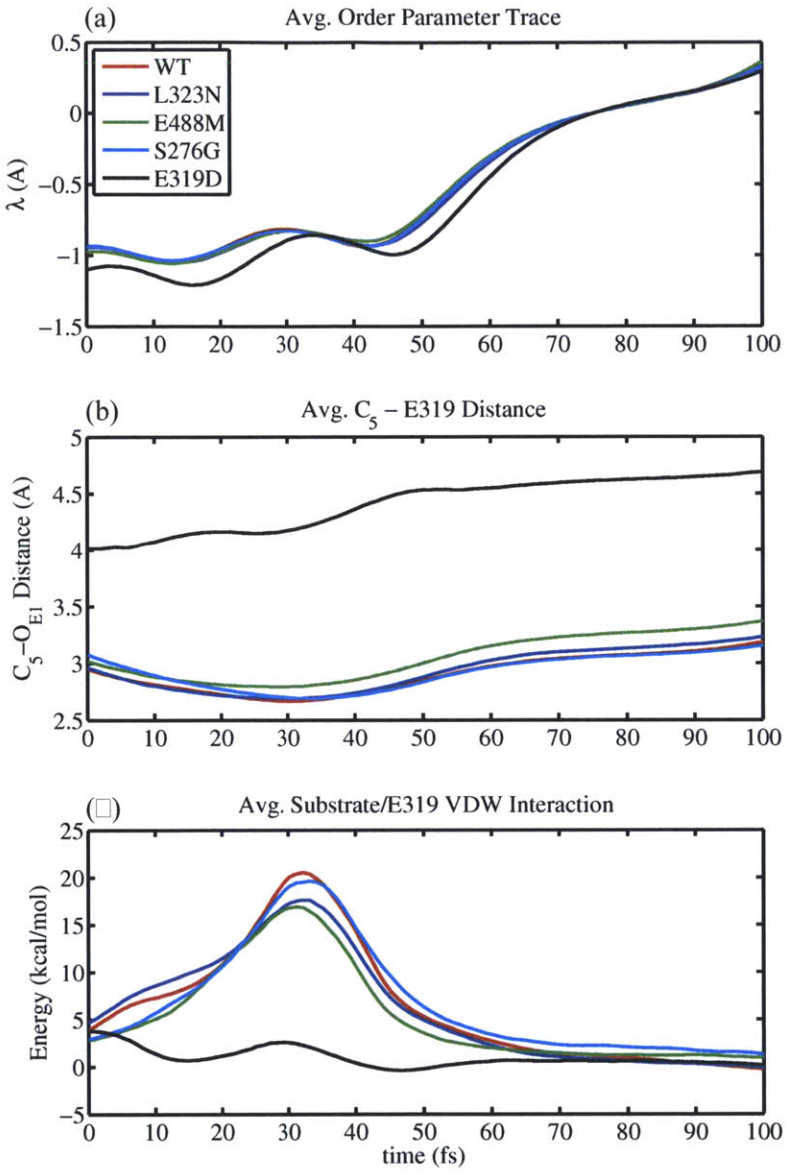


Figure 4-11: **Average traces of the substrate C_5/E319 interaction for WT and mutant variants.** (a) Average traces of the order parameter, λ, for WT and mutants. (b) Average traces of the C_5–O_{E319} distance. (c) Average traces of the van der Waals interaction energy between the substrate and the side chain of residue E319. All trajectories make the successful transition from the reactant to product basin in 101 fs and were aligned such that they cross the $\lambda = 0$ point at the same time (approximately 75 fs) before averaging.

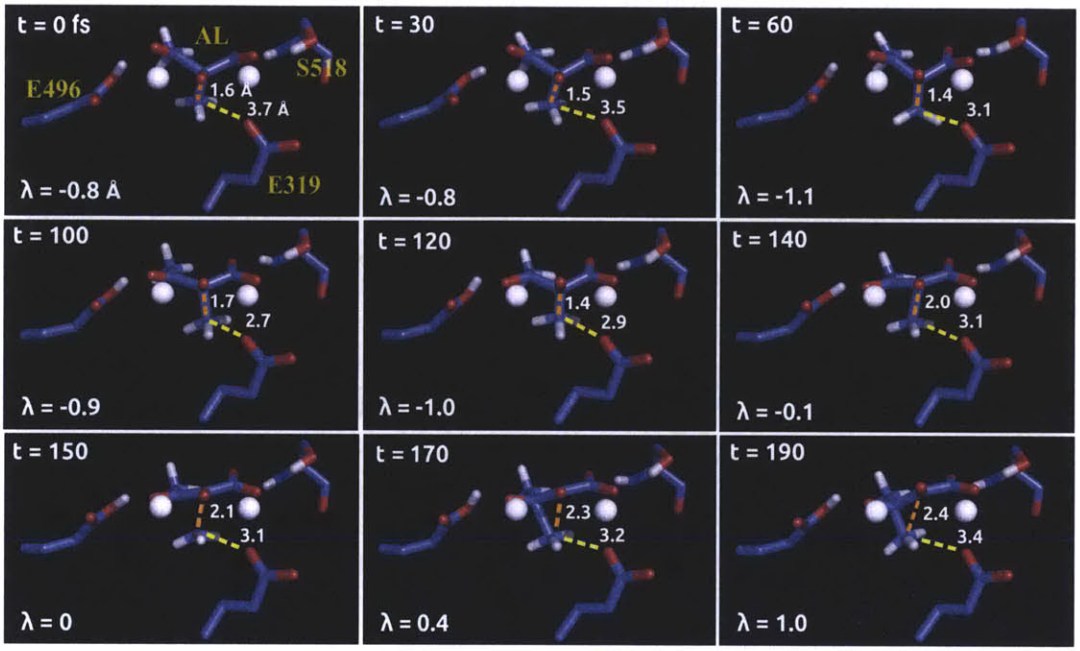


Figure 4-12: **Snapshots along a single, WT reactive trajectory**. In each snapshot the bound substrate, acetolactate (AL), both magnesium ions, and coordinating residues E496, E319, and S518 are shown. Moving from left to right, top to bottom, while the system sat in the reactant basin ($\lambda < -0.15$ Å), the C_4–C_5 bond distance underwent large–scale vibrations between 1.4 and 1.7 Å. Additionally, the C_5–E319 distance compressed from 3.7 to 2.7 Å (t=100 fs), as the C_5 methyl group swung into and collided with the carboxyl group of E319. The methyl group then rebounded, underwent one more vibration (t=100–140 fs), and proceeded across the barrier (t=150 fs) into the product basin.

4.4 Conclusions

We have presented an analysis of transition–state theory based enzyme design as well as a comprehensive study of reaction mechanism of WT and rationally designed mutant KARI variants. Using a TST based framework, we first developed an end–point method to minimize the free energy of activation by focusing on enthalpic contributions and single ground–state and transition–state energetics. Our models highlighted the importance of both the magnesium ions as well as the surrounding active–site residues E496 and D315 in stabilizing the observed transition state. We designed six KARI mutants predicted to reduce the activation energy relative to WT via ground–state destabilization and experimentally validated them with measurements of specific activities and component activation free energies. We find that enthalpic reaction barriers can be reduced by selective destabilization of the ground state, but that increases to entropic barriers often overwhelm these benefits. In four of the six mutants explored in this study, we found that this method was successful at reducing the enthalpic component of the activation free energy by 0.5–2.0 kcal/mol, but did not account for entropic effects, which ultimately increase the free energy barrier by 1–3 kcal/mol. Thus, this method is able to select for active mutants but has difficulty finding those with better activities than WT. These designs were then re-evaluated using transition path sampling to asses the importance of ensemble effects on activity prediction, and the relative activities derived using TPS were found to be in agreement with experiment. The success of the TPS methodology over our initial transition state–theory approach implies that effective catalyst design requires an ensemble treatment of the enzymatic process. Optimizing for reduced activation free energies via minimizing energy differences between single transition and ground states is effective for finding active mutants, but can have the unintended consequence of increasing entropic barriers. For such high dimensional systems, there are likely many different transition states connecting reactants to products [171, 30], and trying to reduce the energetic cost to reach a single one may increase the barrier to reach the rest, increasing the entropic cost of the reaction. A dynamic, ensemble modeling

approach, which does not assume a particular reaction coordinate and/or transition state, is able to account for multiple transition states as well as the various ways the system can reach them. We suspect that for enzymes that have not been optimized by evolution, substantial gains can be made in reducing the enthalpy of activation by improving relative transition–state interactions. There is likely, however, and upper bound on the efficacy of this approach [152], and additional improvement necessitates treatment of dynamic and ensemble effects.

Analysis of the components that make up the kinetic rate constant as found by TPS, $\dot{v}(t)$ and P, reveals that for all KARI variants explored in this study, the actual time required for the system to cross the reaction barrier once it has left the reactant basin is only 28 fs. We note here the difference between the transient time observed for the KARI catalyzed isomerization and the measured ensemble turnover rate of 1 s^{-1}. The disparity found here is similar to that for other enzyme catalyzed reactions, and suggests that there are likely additional, slower protein motions involved in enzyme turnover that act on significantly different time scales from those observed in this study. Our component analysis also leads us to conclude that the real difference between WT and less active mutants lies in P, the probability of being in reactive regions of phase space in the reactant basin (i.e., points that connect to the product basin within τ). Furthermore, the most active variants not only stabilize reactive points and thus reactive trajectories in phase space more than variants that are less active; they also stabilize points that result in non-reactive fluctuations toward the product basin more so than their less active variants. Thus, we find that highly active enzymes facilitate fast fluctuations out of the reactant toward the product basin independent of whether they cross the reaction barrier. Taken together, these findings suggest a novel enzyme redesign strategy based on selective stabilization of reactive points in *phase* space. Increasing the relative probability of coupled configuration and momentum states should increase P and the overall rate of reaction.

Finally, by analyzing trajectory ensembles that carry substrate bound enzymes close to and into the product basin in both WT and mutant systems, we propose that the KARI isomerization reaction proceeds by a pump-and-push mechanism involving

the vibrational activation of the breaking C_4–C_5 bond. In all fluctuations toward the product basin, we observed C_4–C_5 bond vibrations increasing in amplitude over time, the time traces of which fit well to a classical, forced, undamped, harmonic oscillator undergoing resonance. As trajectories approached the barrier, the amplitudes of these vibrations increased, along with the degree of resonance between the bond vibration and the external, oscillating force. For trajectories that crossed the barrier, the forcing frequency had red-shifted by 5 THz from its initial, equilibrium value of 37 THz and overlapped with the natural frequency of the C_4–C_5 bond (32 THz). The system did not fluctuate toward the product basin, however, until it was pushed by the carboxylate group of the active site E319 side chain, which acted as a van der Waals bumper. The C_5 methyl group caromed off the E319 side chain before crossing the reaction barrier. Thus, we propose that the reaction occurs via resonant, vibrational pumping by an externally modulated force and a choreographed push from the E319 side chain. It is interesting to note that the effect of the driving vibration observed here is similar to those proposed for other enzyme catalyzed reactions involving both phosphorolysis [159] as well as hydride transfer [158, 172]. In these cases vibrations facilitate the reaction by destabilizing the reactant via dynamic short-lived correlated compression of the substrate, which pushes the system across, or enhances tunneling through the reaction barrier. We observe similar reactant destabilization, via rapid energy flow into a single bond vibration and a dynamic kick from the active site. In closing, we note that many of our findings point to a unified, dynamic picture of enzyme catalysis, in which enzymes couple motions on many different time scales to accelerate reactions [173, 174, 172, 175]. Enzymes explore phase space through both slow, large–scale conformational changes as well as fast, local fluctuations, and both contribute to the reaction rate. In the case of KARI, we measured a slow, ensemble turnover rate and rapid reaction dynamics. The bond-breaking, bond-forming process occurs on the order of tens of femtoseconds, and it is nearly instantaneous compared to either the initial, slow search through phase space for reactive points that result in barrier crossing or the time necessary for product release.

Chapter 5

General Conclusions

This work has explored the use of ensemble, free energy models in the study and design of protein and small molecule therapeutics. We employed a combination of rigorous, physical theory as well as experiment to develop and test accurate computational models in the areas of protein engineering as well as ligand and enzyme design.

In Chapter 2, we validated the use of implicit solvent electrostatic models in the study and design of antibody–antibody signaling receptors. We found that these efficient, ensemble models are able to capture many of the physically important solvent effects without explicit sampling of solvent degrees of freedom, and can be used to understand the thermodynamics of IgG antibodies binding to Fcγ receptors. In particular, these models provide physical explanations for the observed binding affinity in aglycosylated, mutant IgG antibodies. We found that the glycan contributes negligibly to direct Fc–FcγR interactions, indicating that it has more of an indirect effect on binding, perhaps through stabilization of the folded, unbound immunoglobin [176]. We also found that the affinity recovering mutations act primarily by increasing the electrostatic complementarity of the binding interface. They reduce desolvation penalties paid upon binding by polar, IgG residues that are not directly interacting with polar, FcγR residue partners. These finding highlight the importance of solvent–protein interactions and how they mediate the binding and activation of immune signalling receptors. They also suggest a protein design paradigm based on electrostatic tuning of protein binding interfaces; one that we successfully used in the

negative design of a related, mouse IgG:FcγR system. By adding positive charge at the IgG surface (via addition of lysine or arginine, or removal of aspartate or glutamate), we were able to reduce the electrostatic complementarity at the binding interface, and knock out binding to Fcγ-RIV.

We extended the use of continuum solvent models in Chapter 3, where we explored the role of configurational freedom in the thermodynamics of ligand–protein binding. We developed a powerful, DEE/A* based design framework to enumerate configurational ensembles and efficiently compute ligand free energies, entropies, and enthalpies of binding. This method was applied to a series of previously designed HIV-1 protease inhibitors, and we found that inclusion of ensemble, configurational effects significantly improves correlation between computed and experimentally measured binding affinities. Interestingly, however, the number of important configurations that define the bound and unbound configurational ensembles is quite small given the number of degrees of freedom present in these ligands. Additionally, using the lowest energy configurations in the bound and unbound state to approximate the free energy of binding correlates quite well with experiment. These findings indicate that while the rigid binding approximation is not always accurate, the predominant state(s) hypothesis is likely valid, and the configurational, entropic losses experienced by related ligands are indeed similar. As such, it seems that low energy regions of configurational space dominate the landscape in both the bound and unbound states, and that finding those regions should be paramount for any accurate virtual screening method. Nonetheless, this ensemble methodology is highly accurate and is a valuable lead optimization tool. We hope to see this method applied to a wider variety of compounds to test its applicability in different ligand–receptor systems. Additionally, as computational power increases, we hope to see this design and the CMIE methodology expanded to incorporate receptor degrees of freedom into the calculation in order to explore the relative changes in receptor configurational entropy as well as the relative coupling between ligand and receptor degrees of freedom upon binding. This framework for exploring entropic effects offers a unique way to identify the role of specific degrees of freedom in mediating the thermodynamics of binding as well as

how receptors respond to changes in inhibitor flexibility. We are particularly excited to see it applied to the study of adaptive inhibitors (i.e., those ligands that bind to a broad range of related targets and are effective against drug resistant targets), as recent work highlights the importance of inhibitor flexibility as a mode of adaptation [177, 178, 179].

Finally, in Chapter 4, we combined the lessons learned in previous chapters and assessed the predominant, transition state approximation as it applies to enzyme redesign using MM-PBSA as well as quantum mechanical models. We developed an end-point approach designed to optimize the relative difference in the free energy of the transition versus reactant state, and applied it to the redesign of ketol-acid reductoisomerase. We found that this design approach is successful at reducing the activation enthalpy, but also results in unfavorable increases in the activation entropy. This indicated that configurational activation entropy differences between similar enzymes cannot be assumed to be negligible. We re-explored our panel of mutants in an ensemble context using transition path sampling to map out ensembles of reactive paths, and found that accounting for the multitude of ways the system can cross the barrier results in much more accurate rate predictions. Furthermore, this dynamic method of computing rates revealed that both WT and mutant enzymes have similar reaction mechanisms involving the external, vibrational activation of the breaking bond followed by a dynamic push by an active site residue over the barrier. We also observed the interesting trend that active enzymes not only stabilize reactive trajectories more so than less active variants, but also stabilize non-reactive fluctuations toward the products. These findings highlight the importance of dynamic fluctuations in enzyme catalysis and suggest that future enzyme design methods must take a more complete view of the catalytic process and consider transition state stabilization as well as the effect of mutations on enzyme dynamics when ranking and evaluating potential designs. In particular, we hope to see future design efforts focus on stabilizing not just reactive configurations, but momentum states as well (i.e., points in phase space).

Overall, in this work we presented a critical examination of the utility and ap-

plication of ensemble models in three major areas of computational biochemical engineering: protein–protein binding via antibody design, protein–ligand binding via inhibitor design, and enzyme catalysis via enzyme design. We find that by taking a principled, physics based approach to modeling, we are able to develop significant insight into the inner workings of many biological systems. Furthermore, models developed in this way can be used to intelligently inform experimental design and develop medicinal proteins and small molecules. Collectively, the work presented here highlights the importance of using rigorous statistical mechanical approaches in computational screening and design, even in the absence of perfect force fields. By virtue of the size of these biochemical systems and limits on current computing power, none of the force fields used here are accurate at all length scales. Molecular mechanical and semi-empirical treatments are only approximate, and more expensive, *ab initio* quantum methods will yield more accurate potentials. However, the success observed here suggest that one can compensate for imperfect energy functions by exploring enough of phase space to effectively average out local inaccuracies. Thus, by incorporating more accurate statistical mechanical theory into future computational models (in addition to more accurate force fields), and using experimental work to help guide their application, the field of computational biophysics will undoubtedly speed the scientific exploration of biology and development of high impact therapeutics.

Appendix A

Efficient calculation of molecular configurational entropies using an information theoretic approximation [1]

Abstract

Accurate computation of free energy changes upon molecular binding remains a challenging problem, and changes in configurational entropy are especially difficult due both to the potentially large numbers of local minima, anharmonicity, and high-order coupling among degrees of freedom. Here we propose a new method to compute molecular entropies based on the maximum information spanning tree (MIST) approximation that we have previously developed. Estimates of high-order couplings using only low-order terms provide excellent convergence properties, and the theory is also guaranteed to bound the entropy. The theory is presented together with applications to the calculation of the entropies of a variety of small molecules and the binding entropy change for a series of HIV protease inhibitors. The MIST framework developed here is demonstrated to compare favorably with results computed using the related mutual information expansion (MIE) approach, and an analysis of similarities between the methods is presented.

[1] This work was done in collaboration with Bracken M. King and Bruce Tidor and has been submitted for publication.

A.1 Introduction

A fundamental goal of computational chemistry is the calculation of changes in thermodynamic properties for physical processes, such as chemical potential, enthalpy, and entropy changes. Accurate calculation of such properties can enable computational design and screening at a scale infeasible experimentally, and provides tools for detailed computational analysis of molecules and processes of interest. Design work often focuses on evaluating single configurations in the bound and unbound state, frequently representing the global minimum energy conformation, for instance; however, after filtering out infeasible candidates, additional effort may be warranted to investigate configurational ensemble properties, with significant corrections resulting from entropic or enthalpic sources possible [24, 113, 106]. This work, as well as recent experimental studies using NMR, has so far highlighted the importance of configurational solute entropy in a variety of systems [24, 180]. As such, improving the accuracy and speed of molecular ensemble based calculations, particularly in larger systems, is an area of active research.

One class of approaches for computing configurational averages centers around the use of sampling based simulations, such as molecular dynamics (MD) and Monte Carlo. Such methods may be particularly well suited for larger systems, including proteins, where explicit enumeration and characterization of all relevant minima is infeasible [113]. One of the better known methods in this field is the quasiharmonic approximation, which approximates the system as a multidimensional Gaussian using the covariance matrix computed across aligned simulation frames [106]. While successful in many cases, the quasiharmonic approximation has been shown to significantly overestimate entropies in systems containing multiple unconnected minima, which are poorly modeled by a single Gaussian [108, 181]. Recent phrasings have instead focused on more directly estimating probability densities over the configurational space of a molecule using the frames from MD simulations [182, 113]. As system size grows, however, direct estimation of the density over all molecular degrees of freedom (DOF) becomes infeasible, due to exponential scaling of the sampling

requirements with respect to the effective dimensionality of the system [183].

While estimates of the probability distribution over all DOF of reasonably-sized molecules generally cannot reach convergence given current sampling capabilities, the distributions over each individual DOF, or the joint distribution for groups of small numbers of DOF at a time may converge with the sample sizes accessible from MD simulations [184]. Because of this, recent directions have focused on taking advantage of the entropies computed over these subsets (also called marginal entropies). In general, the motivation for such methods is to combine these well-converged low-order marginal entropies of pairs or triplets of DOF in a molecule to provide an approximation of the ensemble properties of the full molecular system; rather than estimating higher-order contributions directly, they are approximated from low-order terms or neglected through assumptions of the theory. Essentially, these methods effect a trade-off by introducing errors due to theoretical approximations, yet recouping enhanced convergence by avoiding direct estimation of the high-order terms. A net benefit results if the approximation errors introduced are smaller than the estimation (convergence) errors avoided.

One such method is the mutual information expansion (MIE) approximation, recently developed by Gilson and co-workers, which enables approximation of configurational entropies as a function of lower-dimensional marginal entropies [113]. This method expresses the configurational entropy of a system as a series of couplings between all possible subsets of degrees of freedom, placed in order from lowest to highest order. To provide a low-order approximation, and typical for an expansion, MIE assumes the higher-order terms expressed in its expansion can be neglected and truncates all couplings including a large number of DOF (generally omitting all sets of 4 or more DOF). The MIE framework has proved accurate in the analysis of a variety of small-molecule systems [113], and it has been combined with nearest-neighbor methods to improve convergence [185]. It has also been used in the analysis of side-chain configurational entropies to identify residue–residue coupling in allosteric protein systems [186].

In parallel work developed in the context of gene expression and cell signaling

data, we have generated a similar framework, maximum information spanning trees (MIST), that provides an upper bound to Shannon's information entropy as a function of lower-order marginal entropy terms [187]. For multiple synthetic and biological data sets, we found that, in addition to acting as a bound, the MIST approximations generated useful estimates of the joint entropy. Due to the mathematical relationships between information theory and statistical mechanics, application of MIST to the calculation of molecular entropies proved feasible with relatively little adaptation. While similar in spirit to MIE, MIST represents a distinct framework for approximating high-dimensional entropies by combining associated low-order marginal entropies. In particular, whereas MIE explicitly includes all couplings of a particular order in the approximation (e.g., accounting for the couplings between all pairs of DOF), MIST chooses a subset of the same couplings to include so as to maintain a guaranteed lower bound on the entropic contribution to the free energy. In effect, MIST can be thought to infer a model of relevant couplings between molecular degrees of freedom and only include these couplings in the approximation.

Here we examine the behavior of MIST when used to calculate molecular configurational entropies from MD simulation data and also in the context of idealized rotameric systems. The behaviour of MIST is compared directly to MIE in both of the scenarios for a number of systems, and explanations for the differences between the two methods are explored.

A.2 Theory

In this section, we review the Maximum Information Spanning Tree (MIST) approximation in the context of configurational entropies. Further details of MIST have been published previously in the context of analyzing mRNA expression data for cancer classification [187]. In addition, we highlight the theoretical differences between MIST and MIE. In both cases, the goal is to generate an approximation to the confiugrational entropy of a molecule by combining marginal entropy terms calculated over subsets of the degrees of freedom. In so doing, one seeks to introduce a small

approximation error, in favor of faster convergence relative to calculations over all degrees of freedom.

The information theoretic phrasing of the calculation of configurational entropies has been well described previously [113]. The key step of the phrasing comes from representing the partial molar configurational entropy, S°, of a molecule as

$$-TS^\circ = -RT \ln \frac{8\pi^2}{C^\circ} + RT \int \rho(\mathbf{r}) \ln \rho(\mathbf{r}) d\mathbf{r}, \tag{A.1}$$

where R is the gas constant, T is the absolute temperature, C° is the standard state concentration, and ρ is the probability density function (PDF) over the configurational degrees of freedom, $\mathbf{r}$. For the purposes of this report, $\mathbf{r}$ is represented in a bond–angle–torsion (BAT) coordinate system, as opposed to Cartesian coordinates. BAT coordinates tend to be less coupled than Cartesian coordinates for molecular systems and are thus well suited for low-order approximations [188]. The first term on the RHS represents the entropic contribution of the six rigid translational and rotational degrees of freedom and is found via analytical integration, assuming no external field. When negated, the second term can be recognized as the continuous Gibbs entropy [189], which is also identical to RT times the continuous information entropy, S, as described by Shannon [190], providing the equation

$$-TS^\circ = -RT \ln \frac{8\pi^2}{C^\circ} - RTS, \tag{A.2}$$

$$S = -\int \rho(\mathbf{r}) \ln \rho(\mathbf{r}) d\mathbf{r}. \tag{A.3}$$

This relationship allows techniques developed in the context of information theory to be used for the calculation of configurational entropies. Also note that here we generally report the configurational entropy contribution to the free energy or free energy change ($-TS^\circ$ or $-T\Delta S^\circ$), which we refer to as the entropic free energy (change).

The MIST framework provides an upper bound to the information entropy using marginal entropies of arbitrarily low order. The approximation arises from an exact

re-expression of the entropy as a series of conditional entropies, or alternatively, as a series of mutual information terms,

$$S_n(\mathbf{r}) = \sum_{i=1}^{n} S_i(r_i|\mathbf{r}_{1,\ldots,i-1}) = \sum_{i=1}^{n} \left[S_1(r_i) - I_i(r_i;\mathbf{r}_{1,\ldots,i-1})\right], \tag{A.4}$$

$$I(\mathbf{x};\mathbf{y}) = \int \rho(\mathbf{x},\mathbf{y}) \ln \frac{\rho(\mathbf{x},\mathbf{y})}{\rho(\mathbf{x})\rho(\mathbf{y})} d\mathbf{x}d\mathbf{y}, \tag{A.5}$$

where $I_i(r_i;\mathbf{r}_{1,\ldots,i-1})$ is the mutual information (MI) between DOF r_i and all DOF that have already been included in the sum. Throughout this section, subscripts on S or I are included to indicate the order of the term (i.e., the number of dimensions in the PDF needed to compute the term). Notably, while these MI terms are functions of probability distributions of dimension i, they are still pairwise mutual information terms between the single variable r_i and the set of variables represented by $\mathbf{r}_{1,\ldots,i-1}$, in contrast to the multi-information terms employed in the MIE expansion. As a result, these high-dimensional terms maintain key properties of mutual information, including non-negativity [121]. The MI phrasing can be thought of as adding in the entropy of each DOF one at a time (S_1 terms in A.4), then removing a term corresponding to the coupling between that DOF and all previously considered DOF (I_i terms).

The MIST approximation consists of limiting the number of DOF included in these information terms. For example, for the first-order approximation, all coupling is ignored, and the I term is completely omitted from the formulation. By the non-negativity of MI [121], the first-order approximation is thus an upper bound to the exact entropy

$$S_n(\mathbf{r}) = \sum_{i=1}^{n} \left[S_1(r_i) - I_i(r_i;\mathbf{r}_{1,\ldots,i-1})\right] \leq \sum_{i=1}^{n} S_1(r_i) = S_n^{MIST_1}(\mathbf{r}), \tag{A.6}$$

where the superscript $MIST_i$ indicates the MIST approximation of order i.

For the second-order approximation, when each DOF is added, its coupling with a single previously chosen DOF is accounted for, as opposed to considering the coupling

with all previously included terms

$$S_n(\mathbf{r}) \leq S_n^{MIST_2}(\mathbf{r}) = \sum_{i=1}^{n} \left[S_1(r_i) - I_2(r_i; r_j) \right]; j \in \{1, \ldots, i-1\}. \tag{A.7}$$

Comparing A.4 to A.7, one can see that the $I_2(r_i; r_j)$ replaces $I_i(r_i; \mathbf{r}_{1,\ldots,i-1})$ in the summation. This replacement provides the upper-bounding behavior of MIST_2 due to the fact that $I(\mathbf{r_i}; \mathbf{r_j}) \leq I(\mathbf{r_i}; \mathbf{r_j}, \mathbf{r_k})$, for all vectors $\mathbf{r_i}$, $\mathbf{r_j}$, and $\mathbf{r_k}$ [121]. Additional discussion and demonstration of this relationship can be found in the supplementary information.

To generate approximations of arbitrarily high order k, we include an increasing number of DOF in the mutual information term,

$$\begin{aligned} S_n(\mathbf{r}) \leq S_n^{MIST_k}(\mathbf{r}) = & \sum_{i=1}^{n} \left[S_1(r_i) - I_{k'}(r_i; \mathbf{r_j}) \right]; \\ & k' = \min\{i, k\}; j \in \{1, \ldots, i-1\}; \\ & |\mathbf{r_j}| = k' - 1 \end{aligned} \tag{A.8}$$

where $\mathbf{r_j}$ is a vector of length $k'-1$ representing any subset of DOF $\in \{r_1, \ldots, r_{i-1}\}$. As with A.7, the bounding properties of the approximation are guaranteed by the fact that including additional DOF in the MI terms can not decrease the information.

The ordering of terms to consider in the summation over i, and the terms included in the MI terms, j, in A.7 and A.8 may impact the resulting approximation. Because any choice of ordering and information terms will still result in an upper bound to the true entropy, the choices that minimize this expression will provide the best approximation. For small systems, exhaustive enumeration of ordering of indices and information terms may be feasible, but for larger systems, an optimization method is called for. Here, we have chosen to employ a greedy selection scheme that maximizes

the information term selected in each step of the summation,

$$S_n(\mathbf{r}) \leq S_n^{MIST_2}(\mathbf{r}) = \sum_{i=1}^{n} \left[S_1(r_i) - \max_{j \in \{1,\ldots,i-1\}} I_2(r_i; r_j) \right] \tag{A.9}$$

$$\begin{aligned} S_n(\mathbf{r}) \leq S_n^{MIST_k}(\mathbf{r}) = & \sum_{i=1}^{n} \left[S_1(r_i) - \max_{j \in \{1,\ldots,i-1\}} I_{k'}(r_i; \mathbf{r_j}) \right]; \\ & k' = \min\{i, k\}; \\ & |\mathbf{r_j}| = k' - 1. \end{aligned} \tag{A.10}$$

Thus, a third-order approximation is constructed by stepping through each degree of freedom in sequence and adding the $S_1(r_i)$ one-dimensional entropy for that degree of freedom to and subtracting one pairwise third-order mutual information term $I_3(r_i; \mathbf{r_j})$ from the accumulated sum. The term subtracted is the one that gives the largest mutual information between the current DOF r_i and a pair of previously considered DOF $\mathbf{r_j}$; for the first DOF considered there is no mutual information term, for the second degree of freedom the mutual information term is just the second-order mutual information between the second and first DOF $I_2(r_2; r_1)$, and for the third degree of freedom there is no choice in the pairwise third-order mutual information term as there is only one possibility. Higher-order approximations are constructed in the same manner, but the bulk of the pairwise mutual information terms are order k for a k-th order approximation, with lower-order mutual information terms used for the first $k-1$ DOF considered.

In the context of approximations to thermodynamic ensemble properties, MIST bears a strong resemblance to the Bethe free energy (also known as the Bethe approximation) [191]. In fact, the second-order MIST approximation is equivalent to the Bethe approximation, and the full MIST framework may thus be thought of as a high-order generalization of the Bethe free energy. While a full comparison of MIST and the Bethe approximation is outside the scope of the current work, a number of modifications and applications of the Bethe approximation have been explored that may be extensible to MIST [192, 193].

In contrast to MIST, MIE [113] expands the entropy as a series of increasingly

higher-order information terms, as previously formulated by Matsuda [112]:

$$S_n(\mathbf{r}) = \sum_{i=1}^{n} S_1(r_i) - \sum_{i=1}^{n}\sum_{j=i+1}^{n} M_2(r_i;r_j) + \sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{k=j+1}^{n} M_3(r_i;r_j;r_k) - \ldots, \quad \text{(A.11)}$$

where M is the multi-information, defined as

$$M_n(r_1;\ldots;r_n) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{i_1<\ldots<i_k} S_k(r_{i_1},\ldots,r_{i_k}), \quad \text{(A.12)}$$

and the second summation runs over all possible combinations of k DOF from the full set of $\{r_1,\ldots,r_n\}$. Note that for $n = 1$, the multi-information is equivalent to entropy, and for $n = 2$, it is equivalent to the mutual information defined in A.5. MIE generates a k^{th}-order approximation to the full entropy by truncating all terms of order larger than k in A.11. The approximation will converge to the true entropy when no relationships directly involving more than k DOF exist in the system. Notably, MIE does not carry any bounding guarantees, but it does not require the optimization utilized in MIST.

Despite relying on different expansions, MIST and MIE share many similarities. The first-order approximation is identical in both cases (summing all first-order entropies). For the second-order approximation, MIE adds in all first-order entropies and subtracts off all possible pairwise mutual information terms,

$$S_n^{MIE_2}(\mathbf{r}) = \sum_{i=1}^{n} S_1(r_i) - \sum_{i=1}^{n}\sum_{j=i+1}^{n} I_2(r_i;r_j) \quad \text{(A.13)}$$

In contrast, MIST adds in all first-order entropies, and then subtracts off $n-1$ of the information terms (where n is the number of DOF in the system), as is seen in A.9. These terms are chosen to account for as much information as possible, while still guaranteeing an upper bound. The second-order approximations highlight the theoretical differences between MIST and MIE. Whereas MIE removes all pairwise couplings, effectively assuming that that all couplings are independent of each other, MIST removes a subset of couplings, effectively assuming a network of higher-order

dependencies in which each DOF is primarily coupled to the system through a single dominant interaction. In particular, MIST can provide a good approximation if the majority of the degrees of freedom in the system are directly coupled only to a small number of other DOF. Such a system can be well covered by the $n-1$ terms included in MIST. The maximization procedure leading to the tightest upper bound effectively selects these direct couplings when sufficient data exist to accurately estimate their relative magnitudes.

In contrast, MIE may not provide a good approximation in such a system due to indirect couplings that are likely to exist between DOF, and must be removed by higher-order terms. These indirect couplings arise when the configuration of a DOF is coupled to a second DOF only through its interactions with a third intermediate DOF. In such a case, all three DOF will exhibit pairwise coupling with each other, as well as a strong third-order coupling. In MIE, these high-order couplings may be missed, whereas in MIST, one of the pairwise terms may be omitted, providing the possibility for a good approximation of highly coupled triplets of DOF, even when using a second-order approximation. Alternatively, in systems containing a larger number of direct pairwise interactions and relatively few higher-order couplings, MIST may provide a poor approximation relative to MIE. Given these differences in representation, we have performed a series of computational experiments to evaluate the performance of MIST and MIE in a variety of molecular systems, which have helped to reveal how coupled coordinates contribute to configurational entropy.

A.3 Methods

A.3.1 Molecular dynamics simulations of small molecules

All molecular dynamics simulations were run using the program CHARMM [76, 3] with the CHARMm22 all-atom parameter set [194, 195]. Partial atomic charges were fit using the RESP procedure [79, 196] and the program GAUSSIAN 03 [129], with the 6-31G* basis set [196]. All simulations were run at a temperature of 1000 K using a

distance-dependent dielectric of $4r$ with a 1-fs time step, Langevin dynamics, and the leapfrog integrator. A 1-ns equilibration was performed prior to a 50-ns production run from which frames were extracted at a frequency of 1 frame per 10 fs, yielding 5 million frames per simulation.

For each molecule, an internal coordinate representation was chosen in the BAT framework. The internal coordinate representation consisted of the selection of three seed atoms, as well as a single bond, angle, and torsion term for each subsequent atom so as to use improper dihedrals whenever possible, and to place heavy atoms prior to hydrogens. Improper torsions were selected to produce an effect similar to the phase angle approach used by Abagyan and co-workers and Gilson and co-workers [197, 113]. Only bond, angle, and torsion terms between chemically bonded atoms were allowed as coordinates. Other than these restrictions, the specific coordinates were chosen arbitrarily. The values of each bond, angle, and torsion were extracted from the simulations and binned. Marginal PDFs of all single, pairs of, and triplets of coordinates were computed using the frequencies from the simulation. These PDFs were then used to compute the first-, second-, and third-order entropies and information terms. All first- and second-order terms were computed using 120 bins per dimension, and all third-order terms were computed using 60 bins per dimension. In both cases, the ranges of the bins were defined by the minimum and maximum value observed in the simulation. For MIE, third-order information terms containing any bond or angle DOF were set to zero, as was done previously to improve convergence [113]. For MIST, all third-order terms were included, as doing so did not dramatically impact numerical stability. All calculations included a Jacobian term of $\prod_i b_i^2 \sin\theta_i$ where b_i and θ_i are the bond length and bond angle used to place atom i, and the product runs over all DOF included in the marginal term.

A.3.2 Mining minima implementation

In order to enable comparison to the Mining Minima (M2) method using our specific parameters and energy function, we implemented a version of the method within

CHARMM consistent with the original method as described [24, 118]. Briefly, a list of candidate minima was first identified by combinatorially combining all observed minima in each torsional degree of freedom from the MD simulations. Each minimum was then further minimized in CHARMM and duplicates were omitted. For each remaining minimum the Hessian was computed in Cartesian coordinates in CHARMM and converted to the same BAT coordinate system used for analysis with MIST. The energy of each minimum was also extracted. The BAT Hessians were diagonalized, and the product of the eigenvalues computed, with a correction applied to include no more than 3 standard deviations or 60° in any dimension along each eigenvector. Modes with force constants less than 10 kcal/mol were also integrated numerically to check for anharmonicities. For the current work, no modes were found to differ from the harmonic approximation by more than 1 kcal/mol, so the integrated results were not used in the final calculation. Finally, the ensemble average energy was computed and subtracted from the potential to yield the entropic contribution to the free energy, $-TS$.

A.3.3 Discrete rotameric treatment of HIV protease inhibitors

Discrete rotameric systems representing four candidate HIV protease inhibitors, either unbound or in the binding pocket of a rigid HIV-1 protease were generated. Each system consists of the 5×10^4 lowest-energy rotameric configurations, accounting for $> 99\%$ of the contributions to the free energy at 300 K in all cases. For the current work, these 5×10^4 configurations were treated as the only accessible states of the system, enabling exact calculation of all ensemble properties.

The low-energy configurations were determined via a two-step, grid based, enumerative configurational search. All ligands are comprised of a common chemical scaffold with potentially variable functional groups at 5 possible positions (see A-5). We first collected an ensemble of low-energy scaffold conformations using an enumerative Monte Carlo (MC) search. Ten independent simulations of 5×10^4 steps were performed for each ligand in both the bound and unbound states, and the external and scaffold degrees of freedom of all collected configurations were idealized to a uniform

grid with a resolution of 0.1 Å and 10° or 20° (bound or unbound state, respectively). All simulations were performed using CHARMM [76] with the CHARMm22 force field [124] and a distance-dependent dielectric constant of $4r$. The result of the first step was a set of energetically accessible rotameric scaffold configurations.

The second step exhaustively searched the configurational space of the remaining functional group degrees of freedom for each collected scaffold using a combination of the dead-end elimination (DEE) [69, 198, 199] and A* algorithms [74] as described previously [47]. For high throughput energy evaluations, a pairwise decomposable energy function was used that included all pairwise van der Waals and Coulombic, intra- and inter-molecular interactions, computed with the CHARMm22 force field and a distance-dependent dielectric. Uniformly sampled rotamer libraries for each functional group with resolutions of 15° or 60° for the bound or unbound states, respectively, were used. The 5×10^4 lowest-energy configurations across all scaffolds were enumerated and their energies computed.

These lowest-energy configurations from each ensemble were re-evaluated using a higher resolution energy function to account for solvation effects and to obtain a more accurate estimate of the energy. The enhanced energy function included all pairwise van der Waals interactions, continuum electrostatic solvation energies collected from a converged linearized Poisson–Boltzmann calculation using the Delphi computer program [200, 125], and solvent accessible surface area energies to model the hydrophobic effect [86]. Solvation energies were calculated using an internal dielectric of 4 and a solvent dielectric of 80. A grid resolution of $129 \times 129 \times 129$ with focusing boundry conditions [85] was used, along with a Stern layer of 2.0 Å and an ionic strength of 0.145 M.

Given the energies of all configurations in the idealized rotameric systems, entropies of arbitrary order were computed analytically by integrating through the Boltzmann distribution determined from the 5×10^4 molecular configurations included in the ensemble. To evaluate the convergence properties of the metrics in the context of the discrete rotameric systems, we randomly drew from the 5×10^4 structures representing each system with replacement according to the Boltzmann weighted

distribution. The resulting samples were then used to estimate the single, pair, and triplet PDFs as for the MD systems. Because the exact marginal entropies are analytically computable, convergence for these systems was examined with respect to the same approximation computed using the analytically-determined marginal terms. No symmetry adjustments were applied for the discrete systems.

A.4 Results

A.4.1 Molecular dynamics simulations of small molecules

To investigate the behavior of the MIST framework in the context of configurational entropies, we first examined a set of small molecules including hydrogen peroxide, methanol, 1,2-dichloroethane, and linear alkanes ranging in size from butane to octane. Configurational entropies for all of these systems have been previously computed using MIE and were shown to agree well with M2 calculations [113]. As was done in those studies, we collected 5×10^6 frames from a 50-ns molecular dynamics trajectory for each molecule and computed the single, pair, and triplet entropies of all BAT degrees of freedom as described in Methods. We then combined these marginal entropies according to the MIST (A.8) or MIE (A.11) framework, using approximation orders of one, two, or three. The resulting values for the entropic contribution to the free energies, $-TS^\circ$ (computed using A.2), are shown in A-1, where they are compared to the gold standard estimation from the M2 method.

As seen in the previous studies using MIE (red bars), the second-order approximation (MIE_2) shows good agreement with M2 (dashed line) for all molecules, particularly the smaller systems. MIE_3 generally shows similar agreement with M2 for the small molecules and worse agreement (> 10 kcal/mol in some cases) for the alkanes, while MIE_1 shows worse agreement in all cases. The MIST approximations (blue bars) show somewhat different behavior than MIE. As inherent in the theory, the first-order MIST and MIE approximations are identical. $MIST_2$, shows somewhat larger deviations from M2 for the smallest molecules compared to MIE_2 but provides

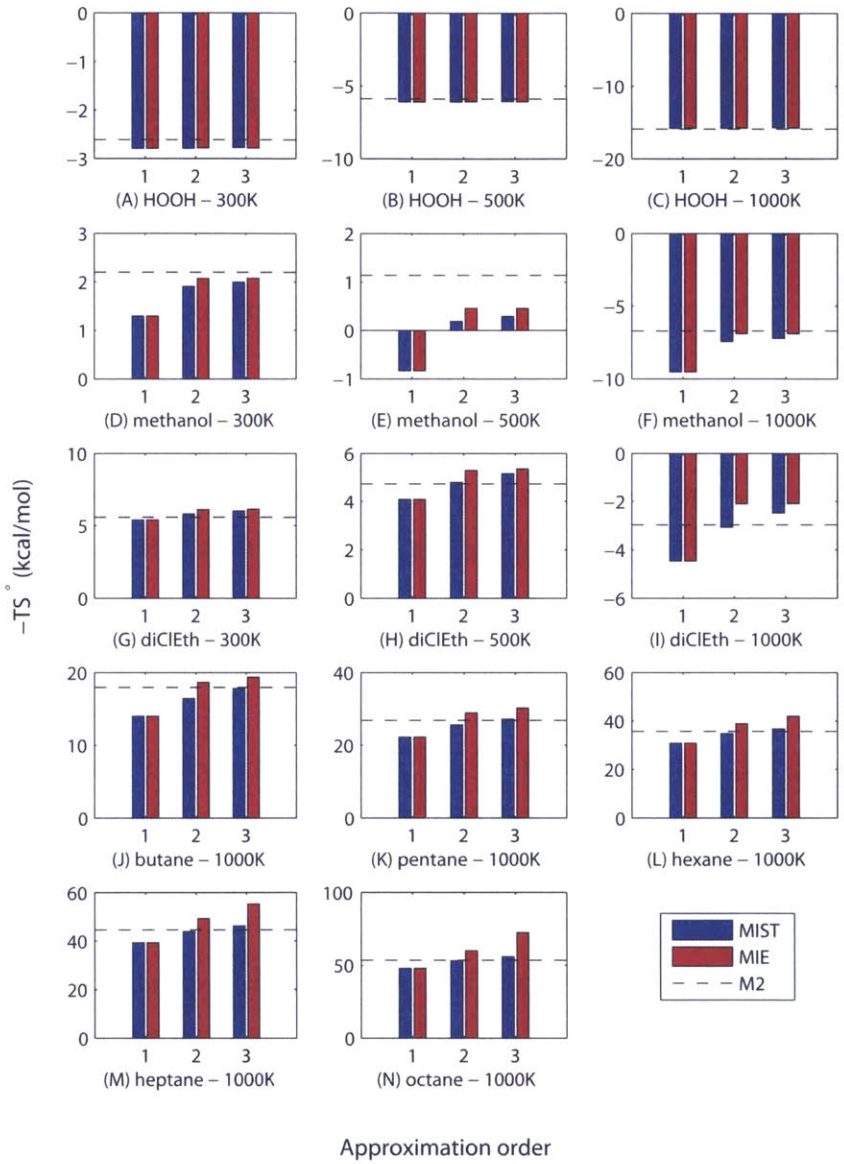


Figure A-1: **MIST and MIE results for small alkanes**: Hydrogen peroxide, methanol, 1,2-dichloroethane and five linear alkanes ranging in size from butane to octane were simulated using MD, and the resulting 5×10^6 frames were used to estimate the marginal entropies. These entropies were then combined according to MIST (blue bars) or MIE (red bars) to generate the first-, second-, or third-order approximation to the configurational entropy of each molecule. Results are compared to calculations using the Mining Minima method (dashed black line).

better agreement for 1,2-dichloroethane and the linear alkanes. Also, whereas MIE_3 generally shows worse agreement with M2 than MIE_2, $MIST_3$ improves upon $MIST_2$ for many systems, showing deviations from M2 less than 1.0 kcal/mol for all systems other than heptane and octane, where the deviations are 1.6 and 2.4 kcal/mol, respectively. While $MIST_3$ is guaranteed to yield at least as accurate a result as $MIST_2$ when both are fully converged, here we see that behavior in the context of finite sample sizes.

A.4.2 Convergence for small molecules

In addition to looking at the MIE and MIST values computed using the full 50-ns simulation, we also examined the behavior of the approximations when using only frames corresponding to shorter simulation times, obtained by truncating the existing simulations. Because each approximation order is converging to a different value and the fully converged values are not known, we track the approach to the value computed with the full 50 ns. The results are shown in A-2 and A.1. For all systems, $MIST_2$ (solid blue lines) exhibits faster convergence than MIE (red lines). While the third-order approximations (dashed lines) converge more slowly than the corresponding second-order ones (solid lines), $MIST_3$ demonstrates comparable convergence to MIE_2 for hydrogen peroxide, methanol, and 1,2-dichloroethane and faster convergence for the alkanes.

Previous work showed that MIE_3 was poorly converged for many of the alkanes, particularly the larger ones, as is observed here [113]. Over the last 10 ns of the hexane, heptane, and octane simulations, the MIE_3 estimate changes by 1.0–3.5 kcal/mol. Notably, the third-order MIE approximation already omits a number of terms to improve numerical stability (all three-way information terms containing a bond or an angle are set to zero). In contrast, the third-order MIST implementation shown here includes all of these terms, and still demonstrates significantly faster convergence. Though we have not explored higher-order MIST approximations for these systems, the good convergence of $MIST_3$ suggests that fourth- or fifth-order approximations may be feasible.

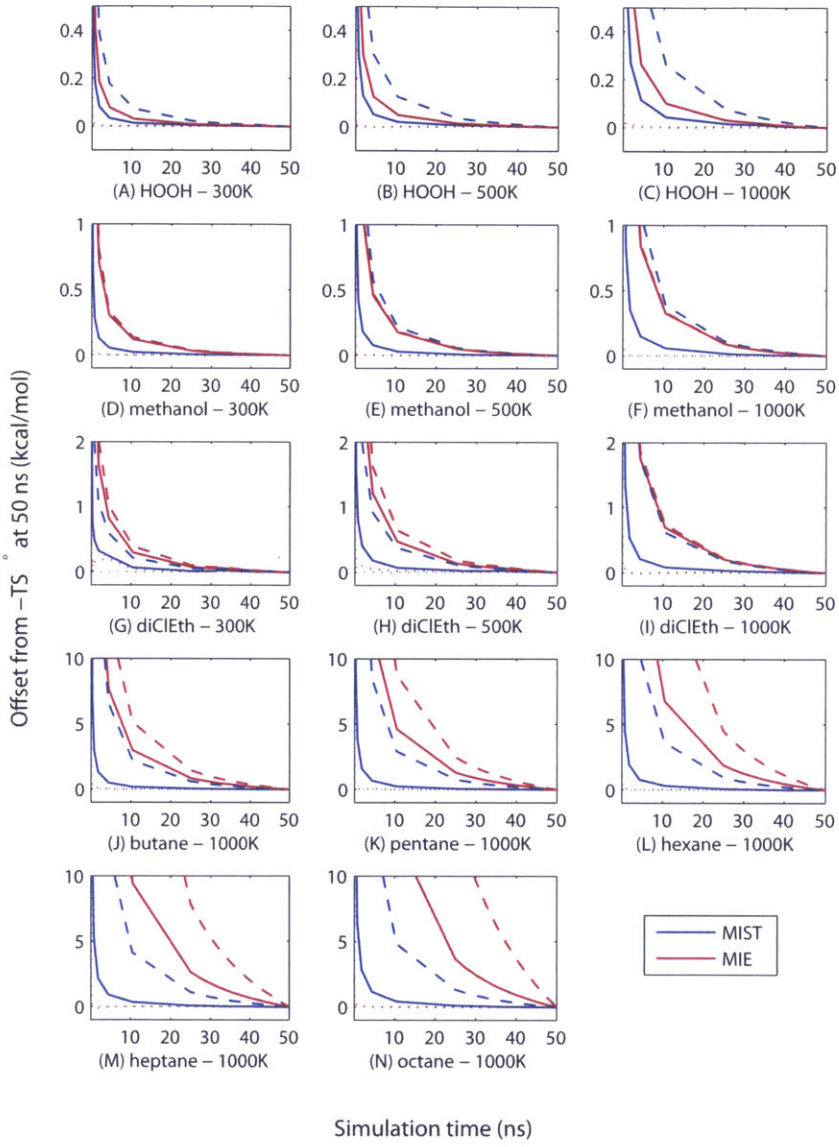


Figure A-2: **Convergence of MIST and MIE for small molecules**: MD simulations of various small molecules were subsampled to include frames corresponding to shorter simulation times, and the resulting sets of frames were used to compute the MIST (blue lines), and MIE (red lines) approximations. The convergence of first- (dotted line), second- (solid lines), and third-order (dashed lines) approximations is shown. Each line shows the deviation from the same value computed using the full 50-ns trajectory. MIE_3 overlaps MIE_2 for HOOH and methanol because each system contains only a single torsional term.

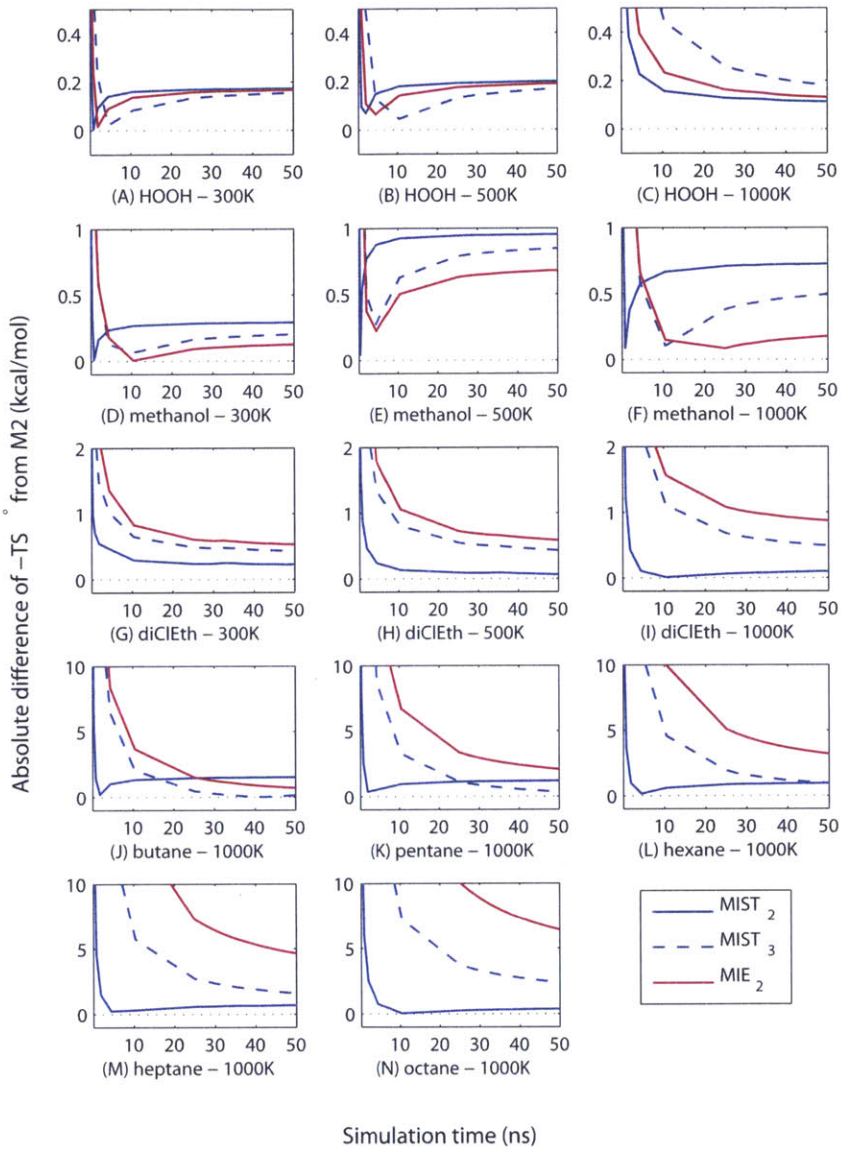


Figure A-3: **Agreement with M2 across sampling regimes**: MIST (blue lines) and MIE (red lines) approximations were computed as a function of simulation times as described in A-2, and the absolute deviation from M2 results were plotted, demonstrating that different approximations provide the best agreement with M2 in different sampling regimes.

molecule	T (K)	$MIST_1$=MIE_1	$MIST_2$	$MIST_3$	MIE_2	MIE_3
HOOH	300	−0.00	−0.00	−0.01	−0.00	−0.00
HOOH	500	−0.00	−0.00	−0.01	−0.00	−0.00
HOOH	1000	−0.00	−0.00	−0.02	−0.01	−0.01
methanol	300	0.00	−0.00	−0.01	−0.01	−0.01
methanol	500	−0.00	−0.00	−0.02	−0.01	−0.01
methanol	1000	0.00	−0.00	−0.03	−0.02	−0.02
diClEth	300	−0.01	−0.01	−0.02	−0.02	−0.03
diClEth	500	−0.01	−0.02	−0.04	−0.04	−0.06
diClEth	1000	−0.00	−0.01	−0.05	−0.05	−0.06
butane	1000	0.00	−0.01	−0.15	−0.21	−0.37
pentane	1000	−0.01	−0.03	−0.20	−0.34	−0.64
hexane	1000	0.00	−0.02	−0.23	−0.48	−1.15
heptane	1000	−0.01	−0.03	−0.29	−0.68	−2.01
octane	1000	0.00	−0.03	−0.34	−0.93	−3.67

Table A.1: Change in estimation of $-TS^\circ$ from 40 ns–50 ns (kcal/mol)

Taken together with the previous section demonstrating the agreement between MIST, MIE, and M2, the results show that sampling regimes may exist in which any of the MIE or MIST approximations give the smallest error. To gain a sense of how the approximations may behave in this regard, we can treat M2 as a comparison point. Although the M2 result may not be equivalent to the full entropy to which MIE and MIST would ultimately converge, treating it as a standard can be instructive about the combined behavior of the methods when weighing accuracy and convergence. To this end, A-3 shows the absolute error of the approximations as a function of simulation time when treating M2 as a gold standard.

For hydrogen peroxide regimes exist for which $MIST_2$, $MIST_3$, or MIE_2 provide the smallest error. In particular, the rapid convergence of $MIST_2$ produces the best agreement with M2 for very short simulation times. With more samples MIE_2 tends to reach adequate convergence to provide the best estimate until $MIST_3$ converges to the point that it provides the closest agreement. For methanol, the faster convergence of $MIST_2$ again provides the best agreement for small sample sizes before MIE_2 converges to give the best agreement. Across 1,2-dichloroethane and the alkanes, MIST provides

better agreement than MIE_2. Either $MIST_2$ or $MIST_3$ provides the best agreement depending on the number of samples. In particular, $MIST_3$ seems to provide better overall agreement with M2 when converged, but the fast convergence of $MIST_2$ again creates regimes for which it demonstrates the best agreement. For heptane and octane (the largest systems examined here), $MIST_2$ provides the best agreement even after 50 ns, possibly due to $MIST_3$ having not fully converged.

A.4.3 Source of differences between MIE_2 and $MIST_2$ for small molecules

To understand the differences in accuracy and convergence between MIE and MIST, we examined the terms of the expansions that differ between the two approximation frameworks. In particular, for the second-order approximations, $MIST_2$ includes a subset of the mutual information terms considered by MIE_2, as can be seen by comparing A.13 and A.7. As such, these omitted terms are entirely responsible for the differences between the two approximations. The values of the terms used for both approximations when applied to butane are shown in A-4.

For each plot, the lower triangle of the matrix shows the pairwise mutual information between each pair of degrees of freedom, all of which are included in the calculation of MIE_2. The upper triangle shows the subset of these terms that are used by $MIST_2$, chosen to minimize A.7 while maintaining an upper bound on the entropy. Focusing on panel D, which shows the results using the full 50-ns simulation, one can see that most of the terms omitted in $MIST_2$ are relatively low in value, whereas the high MI terms are included (to satisfy the maximization in A.7). Panels A–C show the same information computed over the first 4, 10, or 25 ns of the simulation, respectively. In contrast to the 50-ns results, the shorter simulations show dramatic differences between $MIST_2$ and MIE_2. While roughly the same set of terms is omitted by $MIST_2$ in these cases as in the 50-ns case (because the largest MI terms come from the same couplings in the shorter and 50-ns calculations), the omitted terms are much larger, due to their relatively slow convergence. These plots indicate that slow

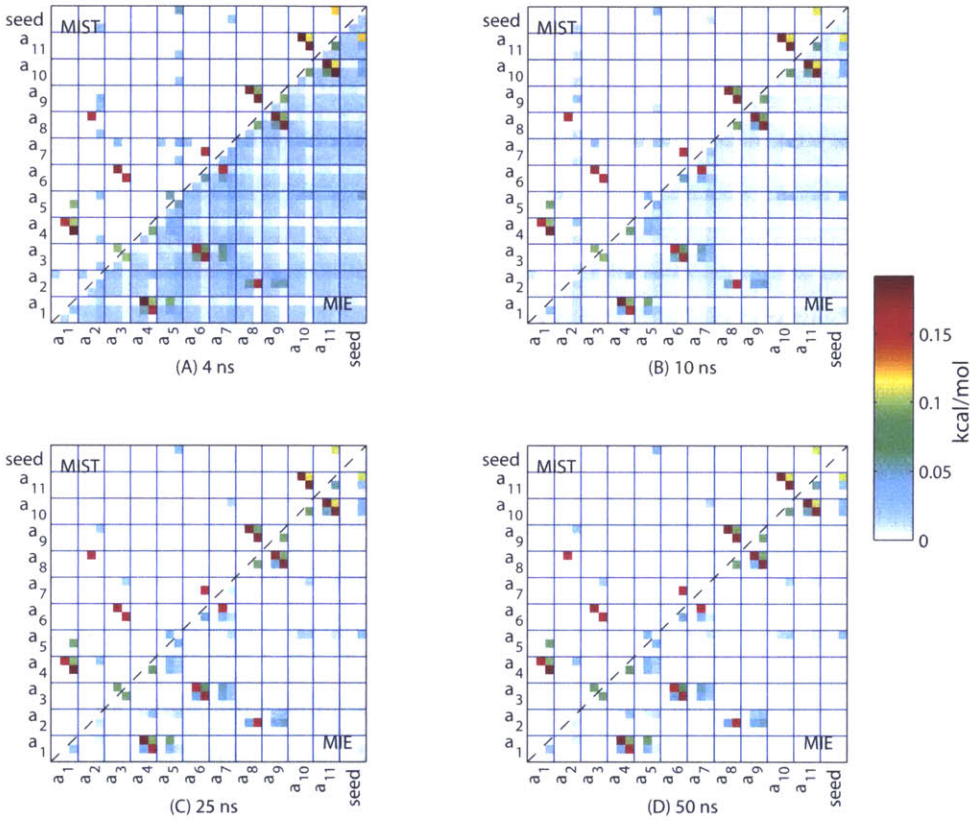


Figure A-4: **Convergence of MI matrix for butane**: The pairwise mutual information terms between all pairs of degrees of freedom in butane computed using the first (A) 4 ns, (B) 10 ns, (C) 25 ns, or (D) 50 ns are show in the lower triangles. The upper triangles indicate the terms that were chosen to be included in the second-order MIST approximation, according to A.7. The dark blue lines separate the atoms from each other, with each atom being represented by three degrees of freedom associated with its placement (bond, angle, torsion from bottom to top and left to right in each box). All values are reported in kcal/mol.

convergence of MIE_2 relative to $MIST_2$ is a result of the many terms in the MI matrix that are slowly converging to very small values. In particular, because MI values tend to be consistently overestimated, the quadratic number of MI terms included in MIE_2 slow convergence more than the linear number of MI terms included in $MIST_2$. Furthermore, the terms that are included by $MIST_2$ are the larger MI terms which tend to converge more quickly than small MI terms. For sufficiently short simulations neglecting these small and slowly converging terms (as done by $MIST_2$) appears to be better than trying to estimate them (as done by MIE_2).

molecule	$x \geq 0.05$	$0.05 > x \geq 0.01$	$0.01 > x \geq 0.00$
butane	29.7	30.4	39.9
pentane	28.4	30.1	41.5
hexane	24.4	26.7	49.0
heptane	19.5	26.3	54.2
octane	17.5	23.8	58.7

Table A.2: Percentage of ($MIE_2 - MIST_2$) accounted for by terms of various magnitudes

To further examine the source of differences between $MIST_2$ and MIE_2, we looked at how much of the difference between the approximations was accounted for by terms of various sizes for the linear alkanes. The results of this analysis using the full 50-ns simulations are shown in A.2. As suggested by A-4, much of the difference between $MIST_2$ and MIE_2 comes from the large number of omitted small terms. For example, for butane 39.9% of the 2.22 kcal/mol difference comes from MI terms with magnitudes less than 0.01 kcal/mol. Furthermore, the importance of these small terms grows as the system size increases, accounting for nearly 60% of the disparity for octane. Taken in conjunction with the slow convergence of these small terms, these results suggest that, while some real representational differences do exist between MIE and MIST, much of the difference may in fact be explained by differences in convergence, even at 50 ns.

A.4.4 Discretized inhibitor molecules as an analytical test case

While the good agreement that both MIST and MIE show with the M2 results is an important validation step in evaluating the overall accuracy of the approximations, some fundamental differences in the methodology can make the results somewhat difficult to evaluate. There are two primary issues that can confound the interpretation. Firstly, M2 calculations and MD simulations represent similar but ultimately different energy landscapes. Whereas the MD landscape represents the exact energy function, M2 approximates the landscape by linearizing the system about a set of relevant minima. Although mode-scanning is employed to account for some anharmonicities in the systems, M2 still operates on an approximation of the energy landscape sampled during MD. As such, even given infinite samples, and without making any truncation approximations (i.e., directly generating $\rho(\mathbf{r})$ for use in A.1), the entropy estimate would not necessarily converge to the M2 result. Secondly, because application of MIST and MIE relies upon estimating the low-order marginal entropies from a finite number of MD frames, it is difficult to separate the error introduced by the approximation framework from the error introduced by estimating the marginal terms.

To address these issues, we examined MIST and MIE in the context of a series of discrete rotameric systems in which the energy of all relevant states was calculated directly. Given this distribution of rotameric states, the full configurational entropy and all marginal entropies can then be computed exactly. As such, for these systems we can separately evaluate the approximation errors due to the MIST or MIE frameworks as well as sampling errors due to estimating the marginal terms; here the marginal terms are known exactly. These discrete ensembles were originally generated to analyze a series of candidate HIV-1 protease inhibitors [47], but their primary importance for the current work is as a test case in which entropies of arbitrary order can be computed exactly. The chemical structures of the four inhibitors are given in A-5. Additional details on the generation of these systems is described in Methods.

We employed eight different discrete ensembles, representing bound and unbound

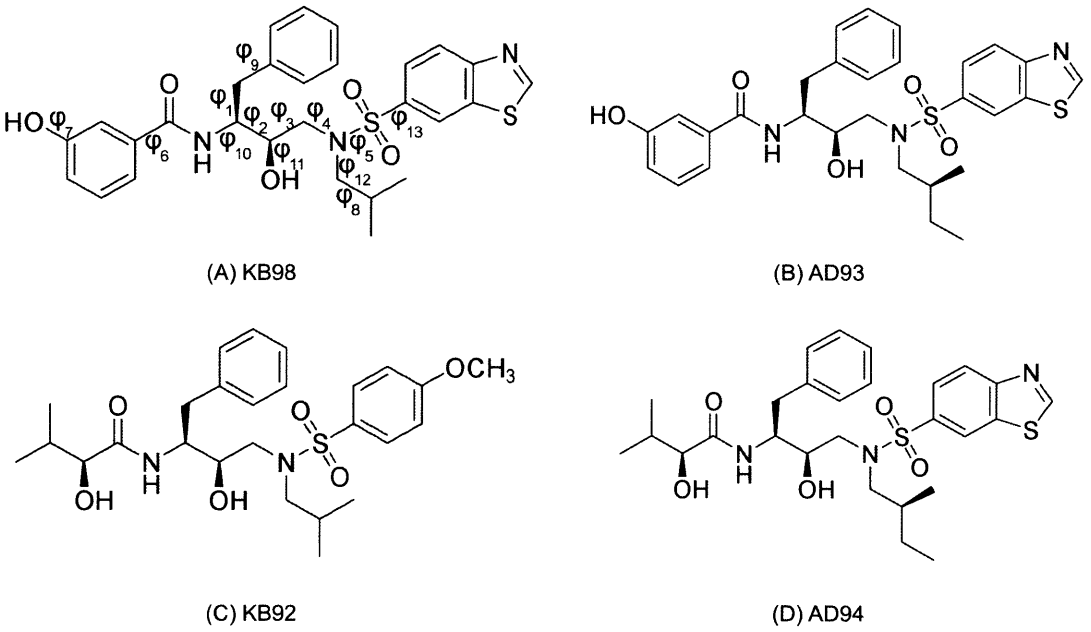


Figure A-5: **Chemical structures of HIV-1 protease inhibitors**: The four molecules shown were previously designed as candidate HIV-1 protease inhibitors [47]. For the current work, idealized rotameric systems in which the exact energies of 50,000 rotameric states were generated in both bound and unbound states, as described in Methods. All torsional degrees of freedom for each inhibitor were rotamerized, and all other DOF (bonds, angles, impropers) were fixed to idealized values. In the bound state overall translations and rotations (external DOF) were also enumerated. Torsions about the bonds labeled in (A) correspond to numbering used in A-8.

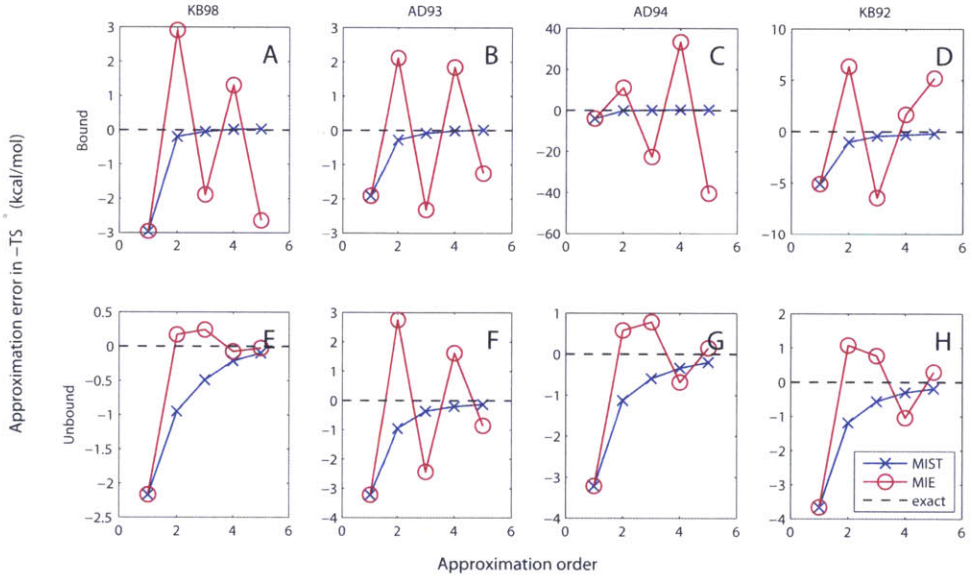


Figure A-6: **Accuracy in rotameric systems**: For each of the four inhibitors, either in the unbound state (bottom row), or in the context of a rigid binding pocket (top row), we computed the exact marginal entropies for all combinations of 1–5 torsions, according to the Boltzmann distribution across the 5×10^4 configurations representing each system. Using these exact marginal entropies we computed the MIST (blue lines) or MIE (red lines) approximations to the entropy ($-TS^\circ$) of each system. The convergence as a function of approximation order is shown in comparison to the analytically determined entropy of the full system (dashed black line).

states of the four inhibitors. All bonds, angles, and non-torsional dihedrals were idealized and fixed, leaving 13–15 torsional degrees of freedom in each inhibitor. We also included an additional variable (referred to as external or ext) representing the six translational and rotational degrees of freedom in the bound cases to model the position of the inhibitor with respect to the rigid binding pocket. For each system we computed exactly all entropy terms containing 1, 2, 3, 4, or 5 degrees of freedom by marginalizing the full Boltzmann distribution of each ensemble, which consisted of the 5×10^4 lowest-energy molecular configurations. We then computed approximations to the total entropy of each system using either MIST or MIE. As such, we were able to examine the approximation error associated with both methods when the low-order terms are known exactly. The results are shown in A-6. For all eight systems the MIST approximations (blue lines, $\times$'s) monotonically approach the full entropy (dashed black line) as the approximation order increases. All MIST approximations also provide a lower bound to the entropic free energy (or an upper bound to the associated Shannon entropy) when the low-order terms are known exactly. Both of these properties are guaranteed for MIST when the marginal terms are known exactly, so seeing them hold in our test system is important validation but not surprising. For all cases the second-order MIST approximation provides an estimate within 1.2 kcal/mol of the full analytic entropic free energy, with particularly good performance in the bound systems (top row of Figure).

For the four unbound systems (bottom row of A-6), MIE (red lines, $\circ$'s) shows similar accuracy to MIST, generating a lower-error estimate once (KB98, panel E), a worse estimate once (AD93, panel F), and comparable error for two cases (AD94 and KB92, panels G and H). Unlike MIST, MIE is not guaranteed to monotonically reduce the approximation error as the order increases, and in some cases, such as unbound KB98 and AD94, the third-order approximation performs worse than the second-order one. In general, however, for the unbound cases the MIE approximations converge towards the true entropy as the approximation order is increased, with exact low-order terms.

In contrast to its performance in the unbound systems, MIE demonstrates erratic

behavior in the bound systems. For all four inhibitors and all approximation orders, MIST results in considerably lower error than the corresponding MIE approximations. Furthermore, increasing the approximation order does not dramatically improve the performance of MIE in the bound systems, and actually results in divergent behavior for orders 1–5 in AD94 (panel C). Notably, the bound systems represent identical molecules to those in the unbound systems; the only differences lie in the level of discretization, and the external field imposed by the rigid protein in the bound state.

A.4.5 Convergence properties in discrete systems

Having investigated the error due to the MIST and MIE approximation frameworks in our analytically exact discrete systems, we next looked to explore the errors associated with computing the approximations from a finite number of samples. To do this we performed a series of computational experiments in which we randomly drew with replacement from the 50,000 structures representing each system according to the Boltzmann distribution determined by their energies and a temperature of 300 K. For each system we drew 10^6 samples, and estimated the PDF over the 50,000 states using subsets of the full 10^6. These PDFs were then used to compute the marginal entropies used in MIST and MIE. For each system this procedure was repeated 50 times to evaluate the distribution of sampling errors for the two methods.

In order to quantify the sampling error separately from the approximation error (which we examined in the previous subsection), we compared the approach of each approximation to the value computed when using the exact low-order terms (i.e., we examined the convergence of each approximation to its fully converged answer, as opposed to the true joint entropy). The results for the bound and unbound KB98 systems are shown in A-7. Results for the other inhibitors were similar and are shown in Figures S1, S2, and S3. As expected, the lower-order approximations converge more quickly, as the low-order PDFs require fewer samples to estimate accurately. For the unbound case (bottom row), both MIE (red) and MIST (blue) exhibit consistent steady convergence for all 50 runs. For the bound case (top row), while MIST exhibits similar convergence behavior as in the unbound system, MIE shows much

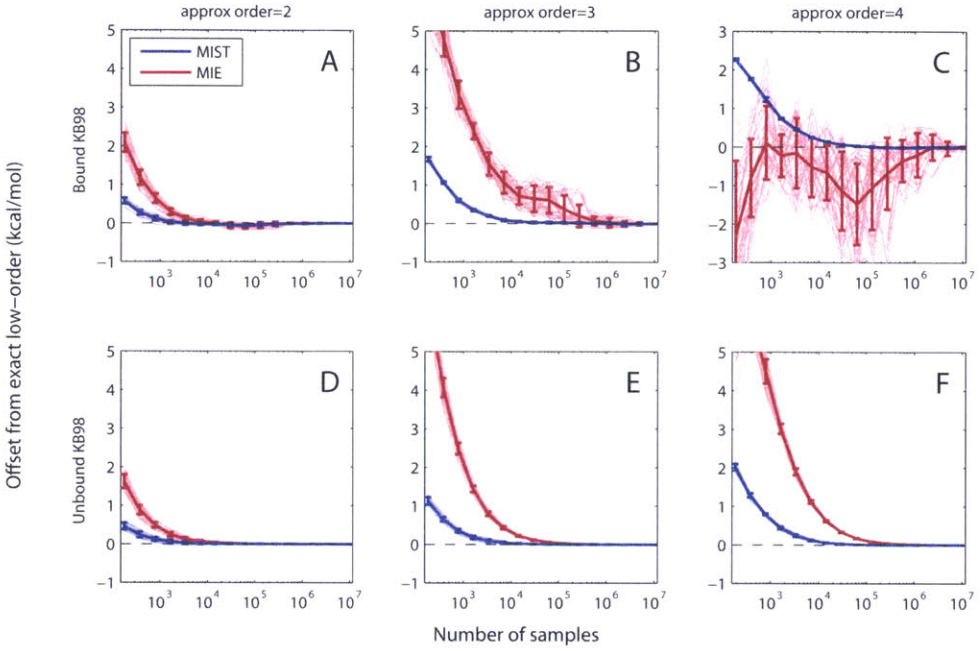


Figure A-7: **Convergence in KB98 rotameric systems**: For each of the eight idealized rotameric systems, we sampled with replacement from the 5×10^4 configurations representing the system, according to the Boltzmann distribution determined by the relative energies of each configuration. These samples were then used to estimate the marginal entropies of all combinations of 1–4 torsions prior to application of MIST (blue lines) or MIE (red lines) to compute $-TS°$. This procedure was repeated 50 times for each system, and the deviation of each run from the exact result to the same order approximation are shown (pale lines), as well as the mean and standard deviation across the 50 runs (thick lines). Results for bound (top row) and unbound (bottom row) KB98 are shown here. Results for other molecules were similar and can be seen in Figures S1, S2, and S3.

larger variations across the 50 runs. As with the MD analysis, MIST demonstrates considerably faster convergence than MIE for all approximation orders examined and all systems.

A.4.6 Source of differences between MIE_2 and $MIST_2$ for discrete systems

We next examined the MI terms accounting for differences between the two approximation frameworks. As with the analysis of the alkanes, the similarities between the second-order approximations enable a direct comparison of the MI terms that are included by MIE but omitted in MIST. Unlike the alkane studies, however, because the low-order terms can be determined directly for these discrete cases, the convergence errors, which played an important role in differences for the alkanes, can be eliminated in the current analysis. Doing so allows direct examination of the differences for the two approximation frameworks, independent of errors introduced due to sampling. The MIs between all pairs of degrees of freedom for bound and unbound KB98, as well as the terms chosen by $MIST_2$ are shown in A-8.

The results for the unbound case (panel B), for which MIE_2 provides lower error, are qualitatively similar to those seen for the alkanes. Most of the differences between MIE_2 and $MIST_2$ in the unbound inhibitor arise from the omission of a number of relatively small terms, less than 0.2 kcal/mol each. The larger MI terms are all included in both approximations. In contrast, the differences between the two methods for the bound case come from a different source: $MIST_2$ omits three of the seven largest MI terms in the bound system, together accounting for nearly 2 kcal/mol of the 2.91 kcal/mol difference between MIE_2 and $MIST_2$. In particular, whereas all six pairwise relationships among the external, ϕ_2, ϕ_3, and ϕ_5 degrees of freedom show strong (and nearly equivalent) couplings, $MIST_2$ only includes three of these terms.

The qualitative differences in the terms accounting for the disparity between $MIST_2$ and MIE_2 in bound KB98 compared the unbound KB98 and the alkanes may be particularly relevant given the relatively poor accuracy of MIE for the bound sys-

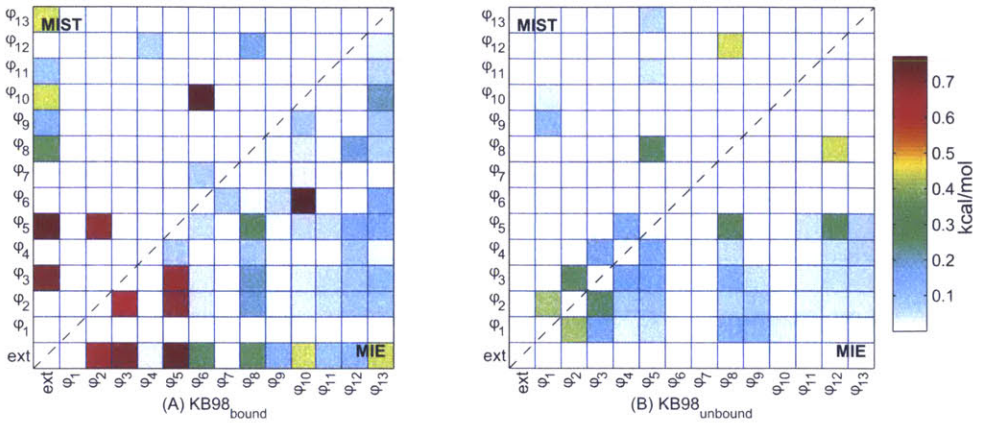


Figure A-8: **MI matrix for discretized KB98**: The pairwise mutual information terms between all pairs of degrees of freedom (DOF) in (A) bound or (B) unbound KB98 are show in the lower triangles. The upper triangles indicate the terms that were chosen to be included in the second-order MIST approximation to $-TS^\circ$, according to A.7. All values are reported in kcal/mol. Numbering of DOF corresponds the labels in Panel A of A-5

tems. The strong couplings between the four degrees of freedom of focus (external, ϕ_2, ϕ_3, and ϕ_5), suggest a high-dimensional transition in which all four DOF are tightly coupled to each other and must change in concert to adopt different energetically relevant states. In particular, the values of the couplings, all of which are near $RT \ln 2$, are consistent with these four degrees of freedom together occupying two dominant states. Furthermore, the coupling between all subsets of three, and the full set of four DOF also are near $RT \ln 2$, further demonstrating the strong high-dimensional coupling between these four. Due to the structure of the MIE approximation, in which all low-order couplings are treated as independent from each other, a highly coupled system may result in errors due to the double-counting of low-order relationships. In contrast, the MIST approximation, which treats each DOF to be predominantly coupled to the system through a single low-order coupling, can appropriately describe such a highly-coupled system with a small number of effective states.

A.5 Discussion

Here we have examined the behavior of our Maximum Information Spanning Trees (MIST) approximation framework in the context of computing molecular configurational entropies. Though we originally developed MIST to pursue high-dimensional information theoretic phrasings in the analysis of experimental biological data, the generality of the method, coupled with the mathematical relationships between information theory and statistical mechanics, enabled application to molecular configurational entropy with relatively little modification. The adaptation of the method was largely inspired by the approach taken previously with the Mutual Information Expansion (MIE) method [113]. We have compared to both MIE and the well established Mining Minima (M2) method in the context of MD simulations of a variety of small molecules. While MIE showed better agreement with M2 for some systems (notably methanol simulated at various temperatures), the MIST approximations tended to provide improved agreement, particularly for larger systems. Furthermore, for all but the smallest molecules, both MIST_2 and MIST_3 demonstrated faster convergence than the MIE approximations. While MIST_3 seemed to provide the best converged answers across all systems, the fast convergence of MIST_2 resulted in it providing better agreement in many sampling regimes, particularly for the larger alkanes. These results suggest that the MIST approximations are likely to be particularly useful in larger systems where simulation times may be limiting.

While the agreement with M2 is an important validation for the overall accuracy of the methods, it does not provide an ideal testing framework, as M2 and the MD simulations represent different energy landscapes. As such, separate examination of the errors due to approximation and sampling was not possible. To address this we also examined MIST and MIE in the context of a series of idealized rotameric systems for HIV protease inhibitors in which the exact entropies could be computed directly. In these systems, we observed that while MIE and MIST both showed good behavior in systems representing unbound molecules, MIE demonstrated poor accuracy in the more restricted bound systems, even for the fifth-order approximation with exactly

determined marginal terms. In contrast MIST exhibited small approximation errors in the bound systems, even for the second-order approximation. Furthermore, when sampling from the known analytical distribution, the fast convergence of MIST relative to MIE seen in the MD systems was also observed for these discretized molecular systems.

In addition to improved convergence, MIST carries useful properties that are not shared by MIE. For fully converged systems, the approximation error of MIST is guaranteed to monotonically decrease with increasing approximation order. This behavior can be easily seen for the discrete systems in A-6, and stands in contrast to the behavior of MIE in the same systems. In application to novel systems where the behavior of the approximations is untested, this property means that the highest approximation order to have reached convergence provides the best estimate of the full entropy. In the absence of such a guarantee, it is unclear how to select the appropriate approximation order.

Furthermore, all converged MIST approximations provide a lower bound on the entropic contribution to the free energy, $-TS^\circ$ (or an upper bound on the Shannon information entropy, S). The bounding behavior may prove particularly useful in identifying optimal coordinate representations. In the previous MIE work the choice of coordinate system has been demonstrated to significantly impact the quality of the approximation [113]. In particular, removing high-order couplings between coordinates, such as those present in Cartesian coordinates, can dramatically improve the accuracy of low-order approximations like MIST and MIE. Because MIST applied to any valid coordinate system will still provide a lower bound on $-TS^\circ$, a variety of coordinate systems may be tested, and the one that yields the largest converged answer is guaranteed to be the most accurate. While additional work is needed to fully enable such a method, even brute-force enumeration is likely to improve performance.

The results of MIE and MIST in the context of the discrete systems also highlight the ability of MIST to provide a good approximation at low orders, even when direct high-order couplings are known to exist. As has been described previously [113, 112], low-order MIE approximations truncate terms in A.11 representing only direct

high-order relationships. The poor accuracy of low-order MIE metrics for the bound idealized systems therefore implies that these systems contain significant high-order terms. Despite the presence of such complex couplings, MIST still provides a good approximation in these same systems. For systems such as proteins that are known to exhibit high-dimensional couplings, the ability to capture high-order relationships in the context of a low-order approximation may prove crucial.

Since the original development of the MIE framework, additional work has been done to extend and apply the method. Nearest-neighbor (NN) entropy estimation has been used to compute the low-order marginal terms utilized by the MIE framework, resulting in significantly improved convergence [185]. Given that MIST relies upon the same low-order marginal terms as MIE, it is likely that NN methods would also be useful in the context of MIST. MIE has also been used to analyze residue side-chain configurational freedom from protein simulations [186]. These studies were able to identify biologically relevant couplings between distal residues in allosteric proteins. Given the relative computational costs of simulating large proteins, and the strong high-dimensional couplings that surely exist in the context of proteins, application of MIST in similar studies may be particularly useful. Preliminary results from ongoing studies have proved promising in the calculation of residue side-chain configurational entropies in the active site of HIV-1 protease.

A.6 Conclusion

In summary, we have adapted our existing information theoretic-based approximation framework to enable calculation of configurational entropies from molecular simulation data. Having characterized its behavior in a variety of molecular systems, we believe MIST can serve as a complement to existing methods, particularly in poorly sampled regimes. A variety of existing extensions and applications for MIE are also likely to be useful in the context of MIST, though further exploration is needed. Finally, in addition to improved convergence, MIST carries monotonicity and bounding guarantees that may prove valuable for future applications.

A.7 Supplementary Material

Demonstration that additional terms cannot decrease MI

We aim to demonstrate that additional terms included to the mutual information cannot decrease its value. That is

$$I(\mathbf{r_i};\mathbf{r_j}) \leq I(\mathbf{r_i};\mathbf{r_j},\mathbf{r_k}). \tag{A.14}$$

We start with the fact that the conditional entropy is less than or equal to the unconditioned entropy, and further that additional conditioning terms can only decrease the entropy further [121]

$$S(\mathbf{r_i}) \geq S(\mathbf{r_i}|\mathbf{r_j}) \geq S(\mathbf{r_i}|\mathbf{r_j},\mathbf{r_k}). \tag{A.15}$$

We next negate both sides of the inequality, and add $S(\mathbf{r_i})$ to both sides,

$$S(\mathbf{r_i}) - S(\mathbf{r_i}|\mathbf{r_j}) \leq S(\mathbf{r_i}) - S(\mathbf{r_i}|\mathbf{r_j},\mathbf{r_k}). \tag{A.16}$$

By the definition of mutual information, we can rewrite this expression as

$$I(\mathbf{r_i};\mathbf{r_j}) \leq I(\mathbf{r_i};\mathbf{r_j},\mathbf{r_k}), \tag{A.17}$$

which is the relationship we aimed to develop.

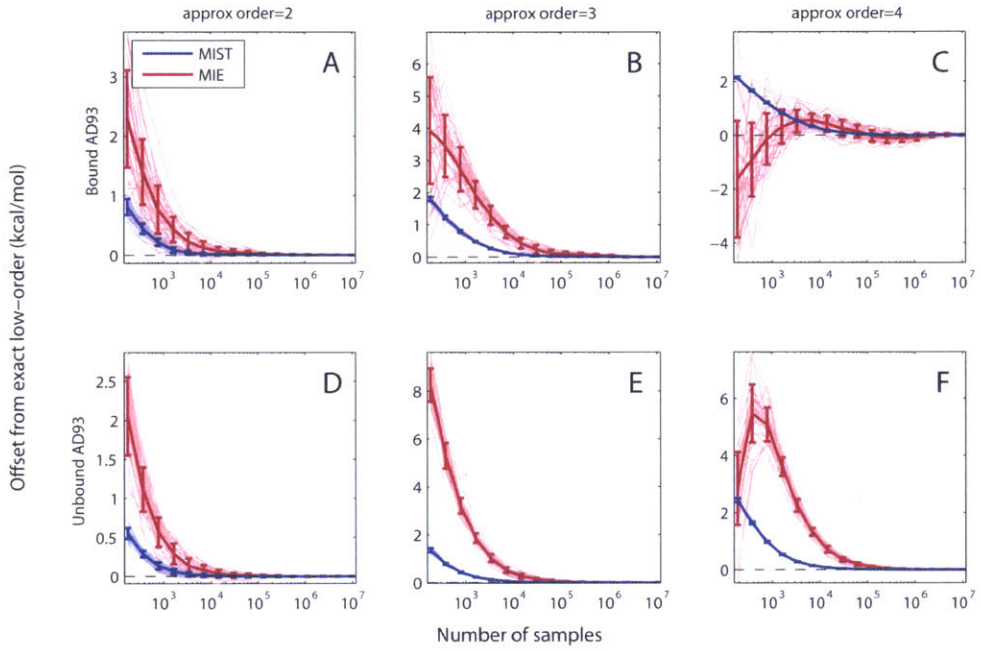


Figure A-9: **Convergence in AD93 rotameric systems**: For each of the eight idealized rotameric systems, we sampled with replacement from the 5×10^4 configurations representing the system, according to the Boltzmann distribution determined by the relative energies of each configuration. These samples were then used to estimate the marginal entropies of all combinations of 1–4 torsions prior to application of MIST (blue lines) or MIE (red lines) to compute $-TS^\circ$. This procedure was repeated 50 times for each system, and the deviation of each run from the exact result to the same order approximation are shown (pale lines), as well as the mean and standard deviation across the 50 runs (thick lines). Results for bound (top row) and unbound (bottom row) AD93 are shown here.

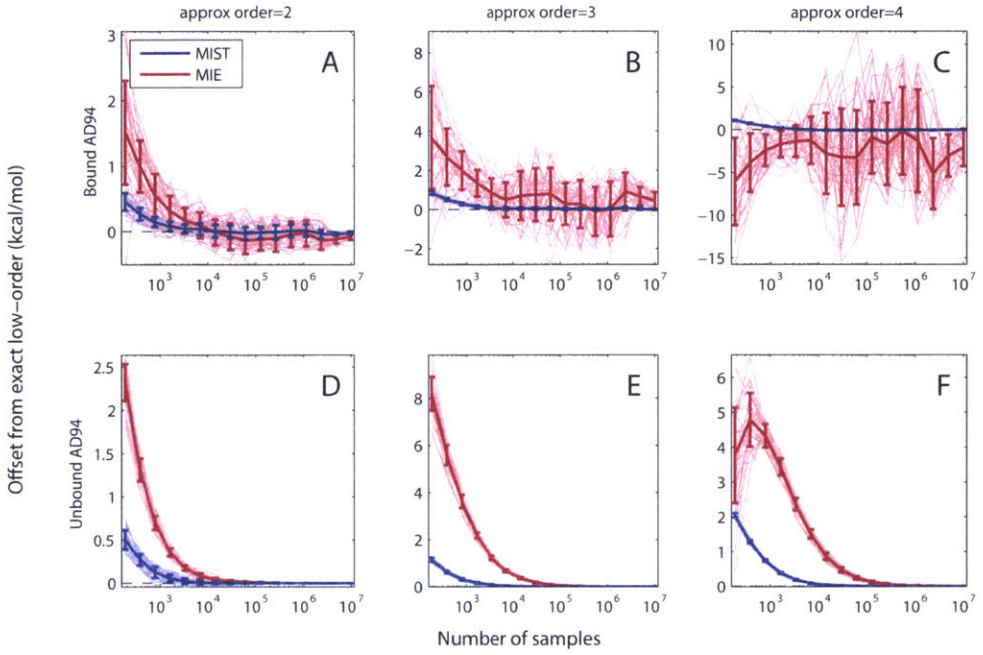


Figure A-10: **Convergence in AD94 rotameric systems**: For each of the eight idealized rotameric systems, we sampled with replacement from the 5×10^4 configurations representing the system, according to the Boltzmann distribution determined by the relative energies of each configuration. These samples were then used to estimate the marginal entropies of all combinations of 1–4 torsions prior to application of MIST (blue lines) or MIE (red lines) to compute $-TS^\circ$. This procedure was repeated 50 times for each system, and the deviation of each run from the exact result to the same order approximation are shown (pale lines), as well as the mean and standard deviation across the 50 runs (thick lines). Results for bound (top row) and unbound (bottom row) AD94 are shown here.

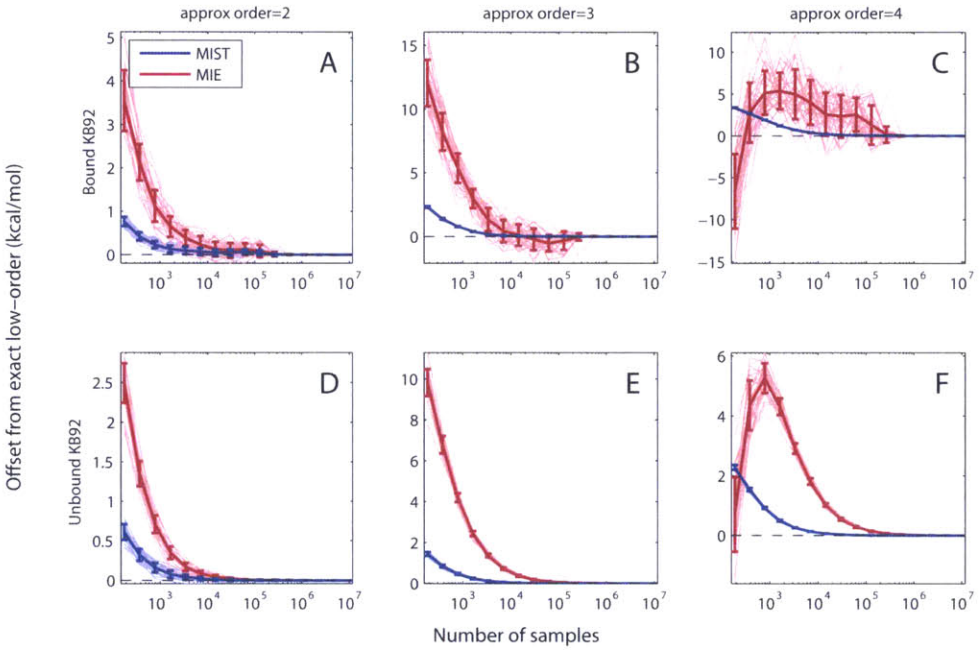


Figure A-11: **Convergence in KB92 rotameric systems**: For each of the eight idealized rotameric systems, we sampled with replacement from the 5×10^4 configurations representing the system, according to the Boltzmann distribution determined by the relative energies of each configuration. These samples were then used to estimate the marginal entropies of all combinations of 1–4 torsions prior to application of MIST (blue lines) or MIE (red lines) to compute $-TS°$. This procedure was repeated 50 times for each system, and the deviation of each run from the exact result to the same order approximation are shown (pale lines), as well as the mean and standard deviation across the 50 runs (thick lines). Results for bound (top row) and unbound (bottom row) KB92 are shown here.

Bibliography

[1] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181, 1958.

[2] H. M. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nat. Struct. Biol.*, 10, 2003.

[3] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: The biomolecular simulation program. *J. Comp. Chem.*, 30:1545–1614, 2009.

[4] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comp. Chem.*, 24, 2003.

[5] D. A. Case, T. Chatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, and R. Woods. The amber biomolecular simulation programs. *J. Comp. Chem.*, 26, 2005.

[6] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4, 2008.

[7] W. L. Jorgensen and J. Tirado-Rives. The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110, 1988.

[8] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard III, and W. M. Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.*, 114, 1992.

[9] J. Moult, K. Fidelis, A. Kryshtafovych, and A. Tramontano. Critical assessment of methods of protein structure prediction (casp)– round ix. *Proteins: Struct., Funct., Bioinf.*, 79, 2011.

[10] M. F. Lensink and S. J. Wodak. Docking and scoring protein interactions: Capri 2009. *Proteins: Struct., Funct., Bioinf.*, 78, 2010.

[11] S. F. Sousa, P. A. Fernandes, and M. J. Ramos. Protein-ligand docking: Current status and future challenges. *Proteins: Struct., Funct., Bioinf.*, 65, 2006.

[12] D. A. McQuarrie. *Statistical Mechanics.* University Science Books, Sausalito, CA, 2000.

[13] D. L. Mobley, , and K. A. Dill. Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure*, 17:489–498, 2009.

[14] M. Zacharias. Accounting for conformational changes during protein-protein docking. *Curr. Opin. Struct. Biol.*, 20:180–186, 2010.

[15] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham III. Calculating structures and free energies of complex

molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33:889–897, 2000.

[16] D. Bashford and D. A. Case. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, 51:129–152, 2000.

[17] J. Wagoner and N. A. Baker. Solvation forces on biomolecular structures: A comparison of explicit solvent and poisson-boltzmann models. *J. Comp. Chem.*, 25:1623–1629, 2004.

[18] H. Gohlke and D. A. Case. Converging free energy estimates: Mm-pb(gb)sa studies on the protein-protein complex ras-raf. *J. Comp. Chem*, 25:238–250, 2003.

[19] S. M. Lippow, K. D. Wittrup, and B. Tidor. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, 25, 2007.

[20] H. Gouda, I. D. Kuntz, D. A. Case, and P. A. Kollman. Free energy calculations for theophylline binding to an rna aptamer: comparison of mm-pbsa and thermodynamic integration. *Biopolymers*, 68:16–34, 2003.

[21] R. D. Gorham, C. A. Kieslich, A. Nichols, N. U. Sausman, M. Foronda, and D. Morikis. An evaluation of poisson-boltzmann electrostatic free energy calculation through comparison with experimental mutagenesis data. *Biopolymers*, 95:746–754, 2011.

[22] T. Lazardis, A. Masunov, and F. Gandolfo. Contributions to the binding free energy of ligands to avadin and streptavidin. *Proteins: Struct., Funct., Genet.*, 47:194–208, 2002.

[23] M. K. Gilson and H. X. Zhou. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36:21–42, 2007.

[24] C. Chang and M. K. Gilson. Free energy, entropy, and induced fit in host-guest recognition: Calculations with the second-generation mining minima algorithm. *J. Am. Chem. Soc.*, 126(40):13156–13164, 2004.

[25] W. Chen, C. E. Chang, and M. K. Gilson. Calculation of cyclodextrin binding affinities: Energy, entropy, and implications for drug design. *Biophys. J.*, 87:3035–3049, 2004.

[26] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discovery*, 3, 2004.

[27] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet, and K. A. Dill. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, 371:1118–1134, 2007.

[28] M. Garcia-Viloca, Jiali Gao, M. Karplus, and D. G. Truhlar. How enzymes work: Analysis by modern rate theory and computer simulations. *Science*, 303:186–195, 2004.

[29] K. E. Ranaghan and A. J. Mulholland. Investigations of enzyme-catalysed reactions with combined quantum mechanics/molecular mechanics (qm/mm) methods. *Int. Rev. Phys. Chem.*, 29, 2010.

[30] P. G. Bolhuis, C. Dellago, and D. Chandler. Reaction coordinates of biomolecular isomerization. *Proc. Natl. Acad. Sci. U.S.A.*, 97, 2000.

[31] J. K. Lassila. Conformational diversity and computational enzyme design. *Curr. Opin. Chem. Biol.*, 14:676–682, 2010.

[32] E. Weinan, W. Ren, and E. Vanden-Eijnden. String methods for the study of rare events. *Phys. Rev. B*, 66:052301, 2002.

[33] D. Sheppard, R. Terrell, and G. Henkelman. Optimization methods for finding minimum energy paths. *J. Chem. Phys.*, 128:134106, 2008.

[34] C. Dellago, P. G. Bolhuis, and P. L. Geissler. Transition path sampling. *Adv. Chem. Phys.*, 68, 2001.

[35] K. H. Bleicher, H. Bohm, K. Muller, and A. I. Alanine. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discovery*, 2, 2003.

[36] R. L. Shields, A. K. Namenuk, K. Hong, Y. G. Meng, J. Rae, J. Briggs, D. Xie, J. Lai, A. Stadien, B. Li, J. A. Fox, and L. G. Presta. High resolution mapping of the binding site on human igg1 for fc gamma ri, fc gamma rii, fc gamma riii, and fcrn and design of igg1 variants with improved binding to the fc gamma r. *J. Biol. Chem.*, 276, 2001.

[37] M. H. Tao and S. L. Morrison. Studies of aglycosylated chimeric mouse-human igg. role of carbohydrate in the structure and effector functions mediated by the human igg constant region. *J. Immunol.*, 143, 1989.

[38] Y. Mimura, P. Sondermann, R. Ghirlando, J. Lund, S. P. Young, M. Goodall, and R. Jefferis. Role of oligosaccharide residues of igg1-fc in fc gamma riib binding. *J. Biol. Chem.*, 276, 2001.

[39] Y. Mazor, T. V. Blarcom, R. Mabry, B. L. Iverson, and G. Georgiou. Isolation of engineered, full-length antibodies from libraries expressed in escherichia coli. *Nat. Biotechnol.*, 25, 2007.

[40] L. C. Simmons, D. Reilly, L. Klimowski, T. S. Raju, G. Meng, P. Sims, K. Hong, R. L. Shields, L. A. Damico, P. Rancatore, and D. G. Yangsura. Expression of full-length immunoglobulins in escherichia coli: rapid and efficient production of aglycosylated antibodies. *J. Immunol. Meth.*, 263, 2002.

[41] R. Jefferis. Antibody therapeutics: isotype and glycoform selection. *Expert Opin. Biol. Ther.*, 7, 2007.

[42] R. Jefferis. Glycosylation as a strategy to improve antibody-based therapeutics. *Nat. Rev. Drug. Discov.*, 8, 2009.

[43] Y. Shaul and G. Schreiber. Exploring the charge space of protein-protein association: A proteomic study. *Proteins: Struct., Funct., Bioinf.*, 60:341–352, 2005.

[44] D. F. Green and B. Tidor. Escherichia coli glutaminyl-trna synthetase is electrostatically optimized for binding of its cognate substrates. *J. Mol. Biol.*, 342:435–452, 2004.

[45] S. M. Lippow and B. Tidor. Progress in computational protein design. *Curr. Opin. Biotech.*, 18:305–311, 2007.

[46] B. A. Joughin, D. F. Green, and B. Tidor. Action-at-a-distance interaction enhance protein binding affinity. *Protein Sci.*, 14:1363–1369, 2005.

[47] M. D. Altman, A. Ali, K. K. Reddy, M. N. L. Nalam, S. G. Anjum, H. Cao, S. Chellappan, V. Kairys, M. X. Fernandes, M. K. Gilson, C. A. Schiffer, T. M. Rana, and B. Tidor. Hiv-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants. *J. Am. Chem. Soc.*, 130(19):6099–6113, 2008.

[48] K. H. M. Murthy, E. L. Winborne, M. D. Minnich, J. S. Culp, and C. Debouck. The crystal-structures at 2.2-angstrom resolution of hydroxyethylene-based inhibitors bound to human-immunodeficiency-virus type-1 protease show that the inhibitors are present in 2 distinct orientations. *J. Biol. Chem.*, 267, 1992.

[49] C. A. Chang, W. Chen, and M. K. Gilson. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U.S.A.*, 104:1534–1539, 2007.

[50] H. X. Zhou and M. K. Gilson. Theory of free energy and entropy in noncovalent binding. *Chem. Rev.*, 109:4092–4107, 2009.

[51] R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman. Side-chain flexibility in proteins upon ligand binding. *Proteins: Struct., Funct., Bioinf.*, 39:261–268, 2000.

[52] S. Marti, J. Andres, V. Moliner, E. Silla, I. Tunon, and J. Bertran. Predicting an improvement of secondary catalytic activity of promiscuous isochorismate pyruvate lyase by computational design. *J. Am. Chem. Soc*, 130, 2008.

[53] S. Marti, J. Andres, V. Moliner, I. Tunon, and J. Bertran. Computer-aided rational design of catalytic antibodies: The 1f7 case. *Angew. Chem. Int. Edit.*, 46, 2007.

[54] L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Rothlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas III, Donald Hilvert, K. Houk, B. Stoddard, and D. Baker. De novo computational design of retro-aldol enzymes. *Science*, 319, 2008.

[55] D. Rothlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, and D. Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 453, 2008.

[56] J. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St.Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael, and D. Baker. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science*, 329, 2010.

[57] F. Nimmerjahn and J. V. Ravetch. Fcgamma receptors as regulators of immune responses. *Nat. Immunol.*, 8, 2008.

[58] S. Siberil, C. A. Dutertre, W. H. Fridman, and J. L. Teillaud. Fcr: The key to optimize therapeutic antibodies? *Crit. Rev. in Oncol./Hematol.*, 62, 2007.

[59] W. L. W. Hazenbos, J. E. Gessner, F. M. A. Hofhuis, H. Kulpers, D. Meyer, I. A. G. M. Heijnen, R. E. Schmidt, M. Sandor, P. J. A. Capel, M. Daeron, J. G. J. van de Winkel, and J. S. Verbeek. Impaired igg-dependent anaphylaxis and arthus reaction in fcriii (cd16) deficient mice. *Immunity*, 5, 1996.

[60] S. Wernersson, M. C. I. Karlsson, J. Dahlstrom, R. Mattsson, J. S. Verbeek, and B. Heyman. Igg-mediated enhancement of antibody responses is low in fc receptor chain-deficient mice and increased in fcrii-deficient mice. *J. Immunol.*, 163, 1999.

[61] G. A. Lazar, W. Dang, S. Karki, O. Vafa, J. S. Peng, L. Hyun, C. Chan, H. S. Chung, A. Eivazi, S. C. Yoder, J. Vielmetter, D. F. Carmichael, R. J. Hayes, and B. I. Dahiyat. Engineered antibody fc variants with enhanced effector function. *Proc. Natl. Acad. Sci. U.S.A.*, 103, 2006.

[62] P. Umana, J. Jean-Mairet, R. Moudry, H. Amstutz, and J. E. Bailey. Engineered glycoforms of an antineuroblastoma igg1 with optimized antibodydependent cellular cytotoxic activity. *Nat. Biotech.*, 17, 1999.

[63] R. L. Shields, J. Lai, R. Keck, L. Y. O'Connell, K. Hong, Y. G. Meng, S. H. A. Weikert, and L. G. Presta. Lack of fucose on human igg1 n-linked oligosaccharide improves binding to human fcriii and antibody-dependent cellular toxicity. *J. Biol. Chem.*, 277, 2002.

[64] G. Cartron, H. Watier, J. Golay, and P. Solal-Celigny. From the bench to the bedside: ways to improve rituximab efficacy. *Blood*, 104, 2004.

[65] X. R. Jiang, A. Song, S. Bergelson, T. Arroll, B. Parekh, K. May, S. Chung, R. Strouse, A. Mire-Sluis, and M. Schenerman. Advances in the assessment and control of the effector functions of therapeutic antibodies. *Nat. Rev. Drug. Discov.*, 10, 2011.

[66] A. C. Chan and P. J. Carter. Therapeutic antibodies for autoimmunity and inflammation. *Nat. Rev. Immunol.*, 10, 2010.

[67] M. L. Alegre, L. J. Peterson, D. Xu, H. A. Sattar, D. R. Jeyarajah, K. Kowalkowski, J. R. Thistlethwaite, R. A. Zivin, L. Jolliffe, and J. A. Bluestone. A non-activating "humanized" anti-cd3 monoclonal antibody retains immunosuppressive properties in vivo. *Transplantation*, 57, 1994.

[68] S. Bolt, E. Routledge, I. Lloyd, L. Chatenoud, H. Pope, S. D. Gorman, M. Clark, and H. Waldmann. The generation of a humanized, non-mitogenic cd3 monoclonal antibody which retains in vitro immunosuppressive properties. *E. J. Immunol.*, 23, 1993.

[69] J. Desmet, M. D. Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

[70] B. I. Dahiyat and S. L. Mayo. Protein design automation. *Protein Sci.*, 5:895–903, 1996.

[71] B. I. Dahiyat and S. L. Mayo. De novo protein design: Fully automated sequence selection. *Science*, 278:82–87, 1997.

[72] N. A. Pierce, J. A. Spriet, J. Desmet, and S. L. Mayo. Conformational splitting: A more powerful criterion for dead-end elimination. *J. Comp. Chem.*, 21:999–1009, 2000.

[73] P. E. Hart, N. J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. on Sys. Sci. and Cybern. SSC4*, 2:100–107, 1968.

[74] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33(2):227–239, 1998.

[75] P. Sondermann, R. Huber, V. Oosthuizen, and U. Jacob. The 3.2-a crystal structure of the human igg1 fc fragment-fc gammariii complex. *Nature*, 406, 2000.

[76] B. R. Brooks. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4(2):187–217, 1983.

[77] A. D. MacKerell Jr., N. Banavali, and N. Foloppe. Development and current status of the charmm force field for nucleic acids. *Biopolymers*, 56, 2001.

[78] A. T. Brunger and M. Karplus. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. *Proteins*, 4, 1988.

[79] C.I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The resp model. *J. Phys. Chem.*, 97:10269–10280, 1993.

[80] K. F. Maxwell, M. S. Powell, M. D. Hulett, P. A. Barton, I. F. McKenzie, T. P. Garrett, and P. M. Hogarth. Crystal structure of the human leukocyte fc receptor, fc gammariia. *Nat. Struct. Biol.*, 6, 1999.

[81] L. P. Lee and B. Tidor. Optimization of binding electrostatics: Charge complementarity in the barnasebarstar protein complex. *Prot. Sci.*, 10, 2001.

[82] K. A. Sharp and B. Honig. Calculating total electrostatic energies with the nonlinear poisson-boltzmann equation. *J. Phys. Chem.*, 94, 1990.

[83] K. A. Sharp and B. Honig. Electrostatic interactions in macromolecules - theory and applications. *Annu. Rev. Biophys. Bio.*, 19, 1990.

[84] M. K. Gilson and B. Honig. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins*, 4:7–18, 1988.

[85] M. K. Gilson, K. A. Sharp, and B. H. Honig. Calculating the electrostatic potential of molecules in solution - method and error assessment. *J. Comp. Chem.*, 9:327–335, 1988.

[86] D. Sitkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free-energies using macroscopic solvent models. *J. Phys. Chem.*, 98:1978–1988, 1994.

[87] Z. S. Hendsch and B. Tidor. Electrostatic interactions in the gcn4 leucine zipper: substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Sci.*, 8:1381–1392, 1999.

[88] E. T. Boder, K. S. Midelfort, and K. D. Wittrup. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc. Natl. Acad. Sci. U.S.A.*, 97, 2000.

[89] J. A. Rakestraw, A. R. Baskaran, and K. D. Wittrup. A flow cytometric assay for screening improved heterologous protein secretion in yeast. *Biotech. Prog.*, 22, 2009.

[90] Z. S. Hendsch and B. Tidor. Do salt bridges stabilize proteins? a continuum electrostatic analysis. *Protein Sci.*, 3:211–226, 1994.

[91] E. T. Boder and K. D. Wittrup. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.*, 15, 1997.

[92] G. A. Lazar, A. J. Chirino, W. Dang, J. R. Desjarlais, S. K. Doberstein, R. J. Hayes, S. B. Karki, and O. Vafa. Optimized fc variants and methods for their generation, 2010.

[93] D. F. Green and B. Tidor. Design of improved protien inhibitors of hiv-1 cell entry: Optimization of electrostatic interactions at the binding interface. *Proteins: Struct., Funct., Bioinf.*, 60:644–657, 2005.

[94] K. A. Armstrong, B. Tidor, and A. C. Cheng. Optimal charges in lead progression: A structure-based neuraminidase case study. *J. Med. Chem.*, 49:2470–2477, 2006.

[95] E. Freire, O. L. Mayorga, and M. Straume. Isothermal titration. *Anal. Chem.*, 62:950–959, 1990.

[96] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.*, 72:1047–1069, 1997.

[97] D. L. Beveridge and F. M. Dicapua. Free-energy via molecular simulation - applications to chemical and biomolecular systems. *Annu. Rev. Biophys.*, 18:431–492, 1989.

[98] J. Carlsson and J. Aqvist. Calculations of solute and solvent entropies from molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, 8:5385–5395, 2006.

[99] T. P. Straatsma and J. A. McCammon. Computational alchemy. *Annu. Rev. Phys. Chem.*, 43:407–435, 1992.

[100] S. B. Dixit and C. Chipot. Can absolute free energies of association be estimated from molecular mechanical simulations? The biotin-streptavidin system revisited. *J. Phys. Chem. A*, 105:9795–9799, 2001.

[101] Y. Cai and C. A. Schiffer. Decomposing the energetic impact of drug resistant mutations in hiv-1 protease on binding drv. *J. Chem. Theory Comput.*, 6:1358–1368, 2010.

[102] L. Zheng, M. Chen, and W. Yang. Random walk in orthogonal space to achieve efficient free-energy simulation of complex simulations. *Proc. Natl. Acad. Sci. U.S.A.*, 105:20227–20232, 2008.

[103] J. M. J. Swanson, R. H. Henchman, and J. A. McCammon. Revisiting free energy calculations: A theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.*, 86:67–74, 2004.

[104] I. Stoica, S. S. Kashif, and P. V. Coveney. Rapid and accurate prediction of binding free energies for saquinavir-bound hiv-1 proteases. *J. Am. Chem. Soc.*, 130:2639–2648, 2008.

[105] M. S. Head, J. A. Given, and M. K. Gilson. "Mining minima": Direct computation of conformational free energy. *J. Phys. Chem. A*, 101:1609–1618, 1997.

[106] M. Karplus and J. N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981.

[107] H. Luo and K. Sharp. On the calculation of absolute macromolecular binding free energies. *Proc. Natl. Acad. Sci. U.S.A.*, 99:10399–10404, 2002.

[108] C. Chang, W. Chen, and M. K. Gilson. Evaluating the accuracy of the quasi-harmonic approximation. *J. Chem. Theory Comput.*, 1(5):1017–1028, 2005.

[109] R. L. Dunbrack. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.*, 12:431–440, 2002.

[110] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3:300–313, 1935.

[111] J. G. Kirkwood. Molecular distribution in liquids. *J. Chem. Phys.*, 7:919–925, 1939.

[112] H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Phys. Rev. E*, 62(3):3096–3102, 1999.

[113] B. J. Killian, J. Y. Kravitz, and M. K. Gilson. Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.*, 127(2):02417, 2007.

[114] B. J. Killian, J. Y. Kravitz, S. Somani, P. Dasgupta, Y. Pang, and M. K. Gilson. Configurational entropy in protein–peptide binding: Computational study of tsg101 ubiquitin e2 variant domain with an hiv-derived ptap nonapeptide. *J. Mol. Biol.*, 389:315–335, 2009.

[115] P. Attard, O. G. Jepps, and S. Marcelja. Information content of signals using correlation function expansions of the entropy. *Phys. Rev. E*, 56:4052–4067, 1997.

[116] A. Singer. Maximum entropy formulation of the kirkwood superposition approximation. *J. Chem. Phys.*, 121:3657–3666, 2004.

[117] T. L. Hill. *Cooperativity Theory in Biochemistry.* Springer-Verlag, New York, 1985.

[118] C. A. Chang, M. J. Potter, and M. K. Gilson. Calculation of molecular configuration integrals. *J. Phys. Chem. B.*, 107(4):1048–1055, 2003.

[119] K. S. Pitzer. Energy levels and thermodynamic functions for molecules with internal rotation. *J. Chem. Phys.*, 14:239–243, 1946.

[120] N. Go and H. A. Sheraga. On the use of classical statistical mechanics in the treatment of polymer chain conformation. *Macromolecules*, 9(4):535–542, 1976.

[121] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, New York, 2006.

[122] F. M. Reza. *An Introduction to Information Theory.* Dover, New York, 1994.

[123] R. M. Fano. *Transmission of information; a statistical theory of communications.* M.I.T. Press, Cambridge, Mass., 1963.

[124] F. A. Momany and R. Rone. Validation of the general-purpose quanta(r)3.2/charmm(r) force-field. *J. Comp. Chem.*, 13(7):888–900, 1992.

[125] A. Nicholls and B. Honig. A rapid finite-difference algorithm, utilizing successive over-relaxation to solve the poisson-boltzmann equation. *J. Comp. Chem.*, 12:435–445, 1991.

[126] D. F. Green. A statistical framework for hierarchical methods in molecular simulation and design. *J. Chem. Theory Comput.*, 5:1682–1697, 2010.

[127] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[128] D. L. Surleraux, A. Tahri, W. G. Verscheuren, G. M. Phille, H. A. de Kock, T. H. Jonckers, A. Peeters, S. de Meyer, H. Azijn, R. Pauwels, M. P. de Bethune, N. M. King, M. Prabu-Jeyabalan, C. A. Schiffer, and P. B. Wigerinck. Hiv protease wild-type in complex with tmc114 inhibitor. *J. Med. Chem.*, 48:1813–1822, 2005.

[129] M.J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara,

K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. *Gaussian 03, Revision B. 05.* Pittsburgh, PA: Gaussian Inc., 2003.

[130] A. Ali, G. S. K. K. Reddy, H. Cao, S. G. Anjum, M. N. L. Nalam, C. A. Schiffer, and T. M. Rana. Discovery of hiv-1 protease inhibitors with picomolar affinities incorporating n-aryl-oxazolidinone-5-carboxamides as novel p2 ligands. *J. Med. Chem.*, 49:7342–7356, 2006.

[131] E. D. Matayoshi, G. T. Wang, G. A. Krafft, and J. Erickson. Novel fluorogenic substrates for assaying retroviral proteases by resonance energy transfer. *Science*, 247:954–958, 1990.

[132] J. Carlsson and J. Aqvist. Absolute and relative entropies from computer simulation with applications to ligand binding. *J. Phys. Chem. B*, 109:6448–6456, 2005.

[133] M. S. Searle and D. H. Williams. The cost of conformational order: Entropy changes in molecular associations. *J. Am. Chem. Soc.*, 114:10690–10697, 1992.

[134] M. I. Page and W. P. Jenks. Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *PNAS*, 68, 1971.

[135] B. Tidor and M. Karplus. The contribution of vibrational entropy to molecular association: The dimerization of insulin. *J. Mol. Bio.*, 238, 1994.

[136] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry.* W. H. Freeman, 2008.

[137] R. Dumas, V. Biou, F. Halgand, R. Douce, and R. G. Duggleby. Enzymology, structure, and dynamics of acetohydroxy acid isomeroreductase. *Acc. Chem. Res.*, 34, 2001.

[138] R. Dumas, D. Job, J. Ortholand, G. Emeric, and A. Greiner. Isolation and kinetic properties of acetohydroxy acid isomeroreductase from spinach (*Spinacia oleracea*) chloroplasts overexpressed in *Escherichia coli. Biochem. J.*, 288, 1992.

[139] S. K. Chunduru, G. T. Mrachko, and K. C. Calvo. Mechanism of ketol acid reductoisomerase–steady-state analysis and metal ion requirement. *Biochemistry*, 28, 1989.

[140] O. Laprevote, L. Serani, B. C. Das, F. Halgand, E. Forest, and R. Dumas. Stepwise building of a 115-kda macromolecular edifice minitored by electrospray mass spectrometry: The case of acetohydroxy acid isomeroreductase. *Eur. J. Biochem.*, 256, 1999.

[141] S. M. Arfin and H. E. Umbarger. Purification and properties of the acetohydroxy acid isomeroreductase of *Salmonella typhimurium. J. Biol. Chem.*, 244, 1969.

[142] R. Tyagi, Y. Lee, L. W. Guddat, and R. G. Duggleby. Probing the mechanism of the bifunctional enzyme ketol-acid reductoisomerase by site-directed mutagenesis of the active site. *FEBS J.*, 272, 2005.

[143] V. Biou, R. Dumas, C. Cohen-Addad, R. Douce, D. Job, and E. Pebay-Peyroula. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with nadph, two magnesium ions and a herbicidal transition state analog determined at 1.65 a resolution. *EMBO J.*, 16, 1997.

[144] K. Thomazeau, R. Dumas, F. Halgand, E. Forest, R. Douce, and V. Biou. Structure of spinach acetohydroxyacid isomeroreductase complexed with its reaction

product dihydroxymethylvalerate, manganese and (phospho)-adp-ribose. *Acta. Crystallogr. D Biol. Crystallogr.*, 56, 2000.

[145] R. Tyagi, S. Duquerroy, J. Navaza, L. W. Guddat, and R. G. Duggleby. The crystal structure of a bacterial class ii ketol-acid reductoisomerase: domain conservation and evolution. *Protein Sci.*, 14, 2005.

[146] H. J. Ahn, S. J. Eom, H. J. Yoon, B. I. Lee, H. Cho, and S. W. Suh. Crystal structure of class i acetohydroxy acid isomeroreductase from pseudomonas aeruginosa. *J. Mol. Biol.*, 328, 2003.

[147] E. W. Leung and L. W. Guddat. Conformational changes in a plant ketol-acid reductoisomerase upon mg(2+) and nadph binding as revealed by two crystal structures. *J. Mol. Biol.*, 389, 2009.

[148] F. P. Martin, R. Dumas, and M. J. Field. A hybrid-potential free-energy study of the isomerization step of the acetohydroxy acid isomeroreductase reaction. *J. Am. Chem. Soc.*, 122, 2000.

[149] R. Dumas, M. Butikofer, D. Job, and R. Douce. Evidence for two catalytically different magnesium-binding sites in acetohydroxy acid isomeroreductase by site-directed mutagenesis. *Biochemistry*, 34, 1995.

[150] M. G. Evans and M. Polanyi. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Trans. Faraday Soc.*, 31, 1935.

[151] H. Eyring. The activated complex in chemical reactions. *J. Chem. Phys.*, 3, 1935.

[152] O. Khersonsky, D. Rothlisberger, O. Dym, S. Albeck, C. J. Jackson, D. Baker, and D. S. Tawfik. Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the ke07 series. *J. Mol. Biol.*, 396, 2010.

[153] O. Khersonsky, D. Rothlisberger, A. M. Wollacott, P. Murphy, O. Dym, S. Albeck, G. Kiss, K. N. Houk, D. Baker, and D. S. Tawfik. Optimization of the *In-Silico*-designed kemp eliminase ke70 by computational design and directed evolution. *J. Mol. Biol.*, 407, 2011.

[154] L. R. Pratt. A statistical method for identifying transition states in high dimensional problems. *J. Chem. Phys.*, 85:5045–5048, 1986.

[155] P. G. Bolhuis, C. Dellago, and D. Chandler. Sampling ensembles of deterministic transition pathways. *Farad. Discuss.*, 110, 1998.

[156] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.*, 108, 1998.

[157] J. E. Basner and S. D. Schwartz. How enzyme dynamics helps catalyze a reaction in atomic detail: A transition path sampling study. *J. Am. Chem. Soc.*, 127, 2005.

[158] S. L. Quaytman and S. D. Schwartz. Reaction coordinate of an enzymatic reaction revealed by transition path sampling. *Proc. Natl. Acad. Sci. USA*, 104, 2007.

[159] S. Saen-oon, S. Quaytman-Machleder, V. L. Schramm, and S. D. Schwartz. Atomic detail of chemical transformation at the transition state of an enzymatic reaction. *Proc. Natl. Acad. Sci. USA*, 105, 2008.

[160] R. Crehuet and M. J. Field. A transition path sampling study of the reaction catalyzed by the enzyme chorismate mutase. *J. Phys. Chem. B*, 111, 2007.

[161] D. Chandler. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem. Phys.*, 68, 1978.

[162] C. Dellago, P. G. Bolhuis, and D. Chandler. On the calculation of reaction rate constants in the transition path ensemble. *J. Chem. Phys.*, 110, 1999.

[163] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comp. Phys.*, 23, 1977.

[164] M. F. Hagan, A. R. Dinner, D. Chandler, and A. K. Chakraborty. Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in dna. *Proc. Natl. Acad. Sci. USA*, 100, 2003.

[165] J. J. P. Stewart. Optimization of parameters for semiempirical methods iv: extension of mndo, am1, and pm3 to more main group elements. *J. Mol. Model*, 10, 2004.

[166] J. Gao, P. Amara, C. Alhambra, and M. J. Field. A generalized hybrid orbital (gho) method for the treatment of boundary atoms in combined qm/mm calculations. *J. Phys. Chem. A*, 102, 1998.

[167] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comp. Chem.*, 13, 1992.

[168] C. Dellago, P. G. Bolhuis, and D. Chandler. Efficient transition path sampling: Application to lennard-jones cluster rearrangements. *J. Chem. Phys.*, 108, 1998.

[169] P. L. Geissler C. Dellago and D. Chandler. Chemical dynamics of the protonated water trimer analyzed by transition path sampling. *Phys. Chem. Chem. Phys.*, 1, 1999.

[170] R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman Lectures on Physics.* Addison Wesley Longman, 1970.

[171] R. J. Dimelow, R. A. Bryce, A. J. Masters, I. H. Hillier, and N. A. Burton. Exploring reaction pathways with transition path and umbrella sampling: Application to methyl maltoside. *J. Chem. Phys.*, 124, 2006.

[172] N. Boekelheide, R. Salomon-Ferrer, and T. F. Miller III. Dyanmics and dissipation in enzyme catalysis. *Proc. Natl. Acad. Sci. USA*, 108, 2011.

[173] S. D. Schwartz and V. L. Schramm. Enzymatic transition state and dynamic motion in barrier crossing. *Nat. Chem. Biol.*, 5, 2009.

[174] S. Hammes-Schiffer and S. J. Benkovic. Relating protein motion to catalysis. *Annu. Rev. Biochem.*, 75, 2006.

[175] S. C. L. Kamerlin and A. Warshel. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins*, 78, 2009.

[176] R. Jefferis and J. Lund. Interaction sites on human igg-fc for fcgammar: current models. *Immunol. Lett.*, 82:57–65, 2002.

[177] H. Ohtaka and E. Freire. Adaptive inhibitors of the HIV-1 protease. *Prog. in Biophys, and Mol. Biol.*, 88:193–208, 2005.

[178] K. Das, A. D. Clark Jr., P. J. Lewi, J. Heeres, M. R. de Jonge, L. M. H. Koymans, H. M. Vinkers, F. Daeyaert, D. W. Ludovici, M. J. Kukla, B. de Corte, R. W. Kavash, C. Y. Ho, H. Ye, M. A. Lichtenstein, K. Andries, R. Pauwels, M. P. de Bethune, P. L. Boyer, P. Clark, S. H. Hughes, P. A. J. Janssen, and E. Arnold. Roles of conformational and positional adaptability in structure-based design of tmc125-r165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant hiv-1 variants. *J. Med. Chem.*, 47:2550–2560, 2004.

[179] Y. Shen, M. Radhakrishnan, and B. Tidor. Unpublished results.

[180] A L Lee, S A Kinnear, and A J Wand. Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nat. Struct. Biol.*, 7(1):72–77, 2000.

[181] Vladimir Hnizdo, Eva Darian, Adam Fedorowicz, Eugene Demchuk, Shengqiao Li, and Harshinder Singh. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *J. Comput. Chem.*, 28(3):655–668, 2007.

[182] Vladimir Hnizdo, Adam Fedorowicz, Harshinder Singh, and Eugene Demchuk. Statistical thermodynamics of internal rotation in a hindering potential of mean force obtained from computer simulations. *J. Comput. Chem.*, 24(10):1172–1183, 2003.

[183] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley and Sons: New York, 1992.

[184] Emily B Walton and Krystyn J Van Vliet. Equilibration of experimentally determined protein structures for molecular dynamics simulation. *Phys. Rev. E*, 74(6 Pt 1):061901, December 2006.

[185] Vladimir Hnizdo, Jun Tan, Benjamin J. Killian, and Michael K. Gilson. Efficient calculation of configurational entropy from molecular simulations by combining the Mutual-Information expansion and Nearest-Neighbor methods. *J. Comput. Chem.*, 29(10):1605–1614, July 2008.

[186] CL McClendon, G Friedland, DL Mobley, H Amirkhani, and MP Jacobson. Quantifying correlations between allosteric sites in thermodynamic ensembles. *J. Chem. Theory Comput.*, 5(9):2486–2502, 2009.

[187] Bracken M. King and Bruce Tidor. MIST: maximum information spanning trees for dimension reduction of biological data sets. *Bioinformatics*, 25(9):1165–1172, May 2009.

[188] Michael J. Potter and Michael K. Gilson. Coordinate systems and the calculation of molecular properties. *J. Phys. Chem. A*, 106(3):563–566, 2002.

[189] J. W. Gibbs. *Elementary Principles in Statistical Mechanics.* C. Scribner's sons, 1902.

[190] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[191] H. A. Bethe. Statistical theory of superlattices. *Proc. R. Soc. Lond. A*, 150:552–575, 1935.

[192] Andrea Montanari and Tommaso Rizzo. How to compute loop corrections to the Bethe approximation. *J. Stat. Mech.: Theory Exp.*, 2005(10):P10011, 2005.

[193] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Bethe free energy, Kikuchi approximations and belief propagation algorithms. Technical Report 16, Mitsubishi Electric Research Lab, 2001.

[194] A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

[195] A. D. MacKerell, Jr., Michael Feig, and C. L. Brooks, III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.*, 25(11):1400–1415, 2004.

[196] David F. Green and Bruce Tidor. Evaluation of ab initio charge determination methods for use in continuum solvation calculations. *J. Phys. Chem. B*, 107(37):10261–10273, 2003.

[197] R. Abagyan, M. Totrov, and D. Kuznetsov. ICM — A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.*, 15:488–506, 1994.

[198] B I Dahiyat and S L Mayo. Protein design automation. *Protein Sci.*, 5(5):895–903, 1996.

[199] B I Dahiyat and S L Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, 1997.

[200] M K Gilson and B Honig. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins*, 4(1):7–18, 1988.