

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Molecular Simulation Guided Protein Engineering and Drug Discovery

Permalink

<https://escholarship.org/uc/item/39k2m5vd>

Author

King, Edward

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Molecular Simulation Guided Protein Engineering and Drug Discovery

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Edward King

Dissertation Committee:
Professor Ray Luo, Chair
Associate Professor Han Li
Professor Greg Weiss
Professor Shiou-Chuan (Sheryl) Tsai
Assistant Professor Shane Gonen

2021

Chapter 1 © 2021 Frontiers Media
Chapters 2 and 5 © 2020-2021 American Chemical Society
Chapter 3 © 2020 Elsevier
All other materials © 2021 Edward King

TABLE OF CONTENTS

	Page
LIST OF FIGURES	v
LIST OF TABLES	vii
ACKNOWLEDGMENTS	viii
VITA	ix
ABSTRACT OF THE DISSERTATION	xi
1 Free energy calculations in drug discovery	1
1.1 Abstract	2
1.2 Introduction	2
1.3 Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA)	4
1.3.1 MM-PBSA developments and benchmarks	10
1.4 Linear Interaction Energy (LIE)	12
1.4.1 LIE developments and benchmarks	16
1.5 Absolute alchemical simulations	17
1.5.1 Absolute alchemical developments and benchmarks	23
2 Estimating the roles of protonation and electronic polarization in absolute binding affinity simulations	50
2.1 Abstract	50
2.2 Introduction	51
2.3 Methods	55
2.3.1 Structure preparation for molecular simulations	55
2.3.2 Alchemical simulation protocol	58
2.3.3 Minimization and equilibration	58
2.3.4 Imposing restraints	59
2.3.5 Alchemical simulation parameters	61
2.3.6 Estimation of electronic polarization with MBAR/PBSA	62
2.4 Results and discussion	63
2.4.1 Structural agreement between simulation and experiment	63
2.4.2 Crystal structure analysis	65
2.4.3 Effect of force field choices on ligand binding modes	65

2.4.4	Benchmarking the effects of simulation conditions on predictive accuracy	67
2.4.5	Default setup leads to no correlation with experiment	68
2.4.6	Use of salt and consistent protonation state improves predicted affinities	68
2.4.7	6DOF versus 1DOF in predicted affinities	70
2.4.8	Possible protonation states at active sites	71
2.4.9	Estimation of electronic polarization by MBAR/PBSA	72
2.5	Conclusion	75
2.6	Acknowledgements	78
2.7	Supplementary information	78
2.7.1	Apparent binding free energies with one titratable group in the active site	78
2.7.2	SI Tables	80
2.7.3	SI Figures	83
3	Engineering natural and noncanonical nicotinamide cofactor-dependent enzymes: design principles and technology development	113
3.1	Abstract	114
3.2	Introduction	114
3.3	Design principles in engineering natural cofactor-dependent enzymes	117
3.4	Design principles in engineering noncanonical cofactor-dependent enzymes	118
3.5	Technology development for engineering natural cofactor-dependent enzymes	134
3.6	Technology development for engineering noncanonical cofactor-dependent enzymes	137
3.7	Conclusion	139
4	Semi-rational design of <i>E. coli</i> gapA to utilize the artificial redox cofactor NMN⁺	148
4.1	Abstract	148
4.2	Introduction	149
4.3	Methods	152
4.3.1	Plasmid and strain construction	152
4.3.2	Protein expression and purification	153
4.3.3	gapA enzymatic assays and kinetics study	153
4.3.4	Rosetta ligand docking and enzyme design	154
4.3.5	High-throughput library screening	155
4.4	Results	156
4.4.1	Rational design of NMN ⁺ binding gapA	156
4.4.2	Advancing orthogonality by disrupting native cofactor binding	159
4.4.3	High-throughput library screening for NMN ⁺ activity	160
4.5	Discussion	165
4.6	Supplementary information	167
4.6.1	SI Tables	167
4.6.2	SI Figures	169

5	Analysis of mutations altering oxygenase conformational dynamics and substrate specificity	176
5.1	Abstract	177
5.2	Introduction	177
5.3	Methods	180
5.3.1	Docking acenaphthene into P450-BM3	180
5.3.2	<i>Ac</i> CHMO homology modeling	181
5.3.3	Molecular dynamics simulations	181
5.3.4	<i>Ac</i> CHMO cofactor binding analysis	183
5.4	Results and discussion	183
5.4.1	P450-BM3 GVQ-AL shows improved binding pocket complementarity to ACN	183
5.4.2	P450-BM3 GVQ-D222N favors the catalytically active, lid-closed state	184
5.4.3	<i>Ac</i> CHMO cofactor binding pose	187
5.4.4	Mutations reshape <i>Ac</i> CHMO conformational landscape and hydride transfer potential	188
5.5	Supplementary information	190
6	Conclusions and future directions	201

LIST OF FIGURES

	Page
1.1 Simulation of protein-ligand binding interactions	1
1.2 Citations per year for free energy methods	4
1.3 MM-PBSA thermodynamic cycle	5
1.4 LIE binding free energy calculation	13
1.5 Absolute alchemical simulation thermodynamic cycle	18
2.1 Chemical structures of the 10 evaluated UPA inhibitors	54
2.2 Example UPA inhibitor binding poses	64
2.3 Relieving steric clash between the ligand phenol and Ser-198	66
2.4 Baseline absolute alchemical binding predictions for UPA inhibitors	69
2.5 MBAR/PBSA binding affinity calculations.	73
2.6 Ligand binding pKa process	79
2.7 Illustration of Boresch 6DOF orientational restraints	83
2.8 Free energy transitions during the decharging phase for the complex trajectories in the baseline simulation	84
2.9 Free energy transition during the decharging phase for the ligand trajectories in the baseline alchemical simulation	85
2.10 Free energy transition during the VDW phase for the complex trajectories in the baseline simulation	86
2.11 Free energy transition during the VDW phase for the ligand trajectories in the baseline simulation	87
2.12 Illustration of the UPA binding pocket with all residues within 6 Å of the ligand highlighted	88
2.13 Analysis of binding pocket flexibility through normalized B-factor Z-scores	89
2.14 Inhibitor equilibration poses from GAFF and GAFF2 compared to starting crystal poses	90
2.15 Backbone CA RMSD development over equilibration with GAFF and GAFF2 force fields	90
2.16 Binding pocket CA RMSD development over equilibration with GAFF and GAFF2 force fields	91
2.17 Ligand heavy atom RMSD development over the equilibration with GAFF and GAFF2 force fields	92
2.18 Comparison of 1DOF and 6DOF restraint schemes	93

2.19	Binding affinity predictions with standard alchemical simulation with different protonation states	94
2.20	Binding affinity predictions with outlier 1GJD removed for standard alchemical simulation with different protonation states	95
2.21	MBAR/PBSA binding affinity calculations including the outlier 1GJD	96
3.1	Cofactor engineering graphical abstract	113
3.2	Chemical structures of natural nicotinamide redox cofactors and mNADs	115
3.3	Representative screening methods used to facilitate the directed evolution of oxidoreductases	135
3.4	Outline of protocol to engineer enzymes for mNAD activity	137
4.1	Redox cofactor generation with native and engineered glycolysis.	150
4.2	Modeled NMN ⁺ binding pose and first round gapA variant activities	157
4.3	Orthogonal gapA specific activities and cofactor binding poses	161
4.4	High-throughput colorimetric screening of gapA for NMN ⁺ activity	162
4.5	gapA specific activities with the native cofactor NAD ⁺	170
4.6	Designed gapA specific activities with NMN ⁺ , NAD ⁺ , and NADP ⁺	170
5.1	Docked models of the P450-BM3 ACN binding poses	184
5.2	Selected mutations alter P450-BM3 conformational dynamics	186
5.3	Homology model of CHMO DTNP	188
5.4	Root mean square fluctuation (RMSF) analysis of CHMO DTNP	189
5.5	Evaluation of hydride transfer efficiency in CHMO variants	191
5.6	Proposed P450-BM3 ACN reaction mechanism	192
5.7	P450-BM3 structural flexibility	193
5.8	Free energy landscapes for DTN, DTNP, and WT CHMO with NADH or NADPH bound	194

LIST OF TABLES

	Page
2.1 Summary of simulation conditions.	57
2.2 Summary of UPA error and correlation statistics	75
2.3 MBAR/PBSA binding affinity accuracy with optimized Radiscale and Protscale parameters	81
2.4 Binding affinity prediction accuracy versus solute interior dielectric (Epsin) with MBAR/PBSA	81
2.5 Full binding predictions at all conditions tested compared to experimental values	82
3.1 Performance of wild type enzymes with noncanonical cofactors	124
3.2 Performance of engineered enzymes with noncanonical cofactors	132
4.1 gapA Michaelis-Menten kinetic parameters	165
4.2 gapA strains table	168
4.3 gapA plasmid table	169
4.4 Counts of gapA variants obtained from high-throughput screen	169
4.5 Michaelis-Menten parameters for HT9-G10R	169

ACKNOWLEDGMENTS

I owe a debt of gratitude to many brilliant colleagues that I've met on this journey.

I am grateful for members of the Li Lab (Linyue Zhang, Will Black, Sarah Maxel, Derek Aspacio, Francis Nicklen, Yulai Zhang, etc.), members of the Luo Lab (Andrew Schaub, Vy Duong, D'ary Greene, Ruxi Qi, Erick Aitchison, Shiji Zhao, Haixin Wei, etc.), and others for their support and friendship. Much of the success reported here is due to their contributions.

Outside of UCI, I would like to thank members of the Siegel Lab at UC Davis (Youtian Cui, Wilson Mak, etc.), coworkers at Hitachi Chemical (Nancy Vi, Tatsu Matsunaga, Anthony Tsai, etc.), and the data science team at Zymergen (Lauren Fitch, Norton Kitagawa, etc.) for their help and everything they've taught me along the way.

I would like to thank my advisors, Han Li and Ray Luo. Han always had high expectations and challenged me to be a better scientist than I thought I could be. Ray gave me the opportunity to work on computational biology when I had zero previous experience and taught me to see biology from a more mathematical perspective. Both advisors provided significant mentorship and encouragement to make this dissertation possible. I would also like to thank Greg Weiss, Sheryl Tsai, and Shane Gonen for serving on my dissertation committee.

Lastly I would like to thank my parents and family. This dissertation is dedicated to them.

VITA

Edward King

EDUCATION

Doctor of Philosophy in Biological Sciences 2021
University of California, Irvine *Irvine, CA*

Bachelor of Science in Biochemistry/Cell Biology 2011
University of California, San Diego *La Jolla, CA*

RESEARCH EXPERIENCE

Graduate Research Assistant Jun 2016–Nov 2021
University of California, Irvine *Irvine, CA*

Data Science Intern Jun 2021–Sep 2021
Zymergen *Emeryville, CA*

Research Intern Nov 2015–Nov 2016
Hitachi Chemical *Irvine, CA*

Research Associate Mar 2012–Aug 2015
Easel Biotechnologies *Los Angeles, CA*

TEACHING EXPERIENCE

Teaching Assistant 2016–2021
University of California, Irvine *Irvine, CA*

Microbiology Lab (M118L), Virology (M124A), Viral Pathogenesis and Immunity (M124B), Synthetic Biology Lab (M130L), Cell Biology (D130) Biochemistry (BIO98), Computational Biology (MB223), Molecular Biology Lab (M116L), Biochemistry Lab (M114L), Molecular Biology (BIO99)

JOURNAL PUBLICATIONS

1. Maxel S*, Saleh S*, **King E***, Aspacio D, Zhang L, Luo R, et al. Growth-Based, High-Throughput Selection for NADH Preference in an Oxygen-Dependent Biocatalyst. *ACS Synth Biol.* 2021.
2. **King E**, Aitchison E, Li H, Luo R. Recent Developments in Free Energy Calculations for Drug Discovery. *Frontiers in Molecular Biosciences.* 2021.
3. **King E**, Qi R, Li H, Luo R, Aitchison E. Estimating the Roles of Protonation and Electronic Polarization in Absolute Binding Affinity Simulations. *J Chem Theory Comput.* 2021.
4. Maxel S*, **King E***, Zhang Y, Luo R, Li H. Leveraging Oxidative Stress to Regulate Redox Balance-Based, In Vivo Growth Selections for Oxygenase Engineering. *ACS Synth Biol.* 2020.
5. Black WB, Aspacio D, Bever D, **King E**, Zhang L, Li H. Metabolic engineering of *Escherichia coli* for optimized biosynthesis of nicotinamide mononucleotide, a non-canonical redox cofactor. *Microb Cell Fact.* 2020.
6. Maxel S, Zhang L, **King E**, Acosta AP, Luo R, Li H. In Vivo, High-Throughput Selection of Thermostable Cyclohexanone Monooxygenase (CHMO). *Catalysts.* 2020.
7. **King E***, Maxel S*, Li H. Engineering natural and noncanonical nicotinamide cofactor-dependent enzymes: design principles and technology development. *Curr Opin Biotechnol.* 2020.
8. Maxel S, Aspacio D, **King E**, Zhang L, Acosta AP, Li H. A Growth-Based, High-Throughput Selection Platform Enables Remodeling of 4-Hydroxybenzoate Hydroxylase Active Site. *ACS Catal.* 2020.
9. Black WB, Zhang L, Mak WS, Maxel S, Cui Y, **King E**, et al. Engineering a nicotinamide mononucleotide redox cofactor system for biocatalysis. *Nat Chem Biol.* 2020.
10. Black WB, **King E**, Wang Y, Jenic A, Rowley AT, Seki K, et al. Engineering a Coenzyme A Detour To Expand the Product Scope and Enhance the Selectivity of the Ehrlich Pathway. *ACS Synth Biol.* 2018.
11. Zhang L, **King E**, Luo R, Li H. Development of a High-Throughput, In Vivo Selection Platform for NADPH-Dependent Reactions Based on Redox Balance Principles. *ACS Synth Biol.* 2018.

CONFERENCE PUBLICATIONS

Electrostatics and Polarization in Protein-Ligand Binding Affinity Predictions

Apr 2021

American Chemical Society Spring 2021

ABSTRACT OF THE DISSERTATION

Molecular Simulation Guided Protein Engineering and Drug Discovery

By

Edward King

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2021

Professor Ray Luo, Chair

Targeted protein-ligand binding interactions drive the metabolic processes essential for life and biochemical manufacturing. Binding interactions between enzymes and small molecules are mediated by the sum of weak, non-covalent interactions including: hydrophobic packing, steric effects, electrostatics, and hydrogen-bonding. Characterization of these interactions is limited by the difficulty in obtaining high resolution structural data of the active binding poses. Furthermore, static models from crystallography are unable to capture the dynamic conformational changes that occur during the transition from the protein unbound to bound states. By resolving how these transitory contacts affect protein function, we accelerate the design of enzymes with target activities and discovery of small molecule inhibitors.

We investigate protein-ligand interactions from two directions: 1) From the perspective of protein engineering in answering the question, what mutations should be made in a protein's amino acid sequence to enhance its binding affinity toward a target ligand. 2) From the field of drug design, how can we accurately predict the absolute binding free energies of small molecules. This work demonstrates how computational methods utilizing physical modeling can be applied in combination with high-throughput, directed-evolution experiments to advance biomolecular design.

Molecular dynamics (MD) simulations account for the effects of atomic flexibility and ex-

plicit solvent that are key to biomolecular interactions. In Chapter 1, we review the basis of free energy calculations based on the Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA), Linear Interaction Energy (LIE), and alchemical simulation approaches in drug development. We perform absolute alchemical simulations in Chapter 2 with inhibitors targeting the Urokinase Plasminogen Activator (UPA) system and analyze how a range of simulation parameters such as counter-ion concentration and alternative binding pocket protonation states impact the binding free energy predictions. We improve predictive accuracy by adapting the protocol to utilize the continuum PBSA solvent model with charge polarization corrections through scaling of the solute dielectric.

In Chapter 3, we describe current approaches to engineering proteins for altered redox cofactor specificity, which has industrial value in specific delivery of electron energy and reduction of feedstock costs in biomanufacturing. We integrate molecular modeling with site-saturated mutagenesis to efficiently navigate protein sequence space with *Escherichia coli* glyceraldehyde 3-phosphate dehydrogenase (*Ec gapA*) to enable utilization of the artificial redox cofactor nicotinamide mononucleotide (NMN⁺) in Chapter 4. Lastly, we investigate how mutations fine-tune oxygenase conformational dynamics to modify substrate specificity and turnover in Chapter 5.

Metabolic pathway engineering with enzymes specific for NMN/H provides direct control over electron flow in living organisms. Application of our developed molecular modeling tools will improve the accuracy and speed of MD simulations, facilitating routine usage to reduce the costs required to construct and screen protein variants, expedite identification of potential pharmaceuticals, and allow study of dynamic biomolecular interactions that are inaccessible through experiment.

Chapter 1

Free energy calculations in drug discovery

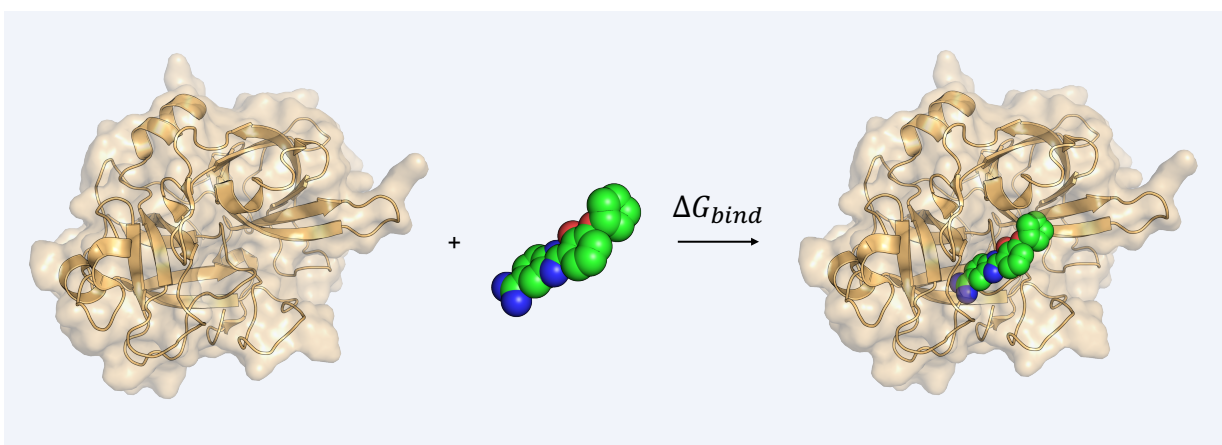


Figure 1.1: Simulation of protein-ligand binding interactions.

Adapted from:

Authors: Edward King, Han Li, Erick Aitchison, Ray Luo

Front Mol Biosci. 2021;8: 712085.

doi: [10.3389/fmolb.2021.712085](https://doi.org/10.3389/fmolb.2021.712085)

Publication Date (Web): August 11, 2021

1.1 Abstract

The grand challenge in structure-based drug design is achieving accurate prediction of binding free energies. Molecular dynamics (MD) simulations enable modeling of conformational changes critical to the binding process, leading to calculation of thermodynamic quantities involved in estimation of binding affinities. With recent advancements in computing capability and predictive accuracy, MD based virtual screening has progressed from the domain of theoretical attempts to real application in drug development. Approaches including the Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA), Linear Interaction Energy (LIE), and alchemical methods have been broadly applied to model molecular recognition for drug discovery and lead optimization.

1.2 Introduction

Modern drug development requires screening over vast regions of chemical space to identify potential binders against a protein target. This approach is costly in time and material resources[1]. Even after identification of potential ligands from initial screening assays, further refinement must be carried out to improve binding properties, ensure that off target effects are minimized, and optimize pharmacokinetic properties. Evaluation of binding free energies through virtual screening has shown promise in efficiently narrowing the chemical search space for candidate compounds and streamlining the process of lead compound optimization. Outside of the pharmaceutical field, binding affinity predictions find additional uses in protein engineering, and guide the rational

design of mutations altering enzyme substrate/product specificity[2–6], structural stability[7–10], and catalytic efficiency[11, 12].

Here we discuss recent developments and applications of molecular dynamics to calculate absolute binding free energies in protein-ligand binding interactions. Through utilization of the Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA)[13–20], Linear Interaction Energy (LIE)[21–24], and absolute methods[25–31], researchers are able to evaluate biomolecular interactions that drive molecular recognition at atomic resolution and derive accurate predictions for binding free energies. These methods rigorously account for conformational dynamics and solvent interactions that are key to protein-ligand interactions and absent in coarser-grained approaches such as ligand docking. The value in these methods for advancing drug discovery is highlighted by their widespread application. Within the last 20 years the number of citations for each method has grown from a small handful to several thousand, notably the MM-PBSA method was found in over 2,000 citations in the last year (Figure: 1.2).

These three methods differ in their treatment of solvent and required simulation data, either involving only the end point states of bound and unbound species, or demanding simulation of a complete binding pathway traversing intermediate states between the end points for determination of binding free energy. These differences result in trade-offs between predictive accuracy and computational cost that must be weighed by the user to select the best approach for their application. In this review, discussion of approaches for the calculation of relative binding free energies is skimmed over as having been recently reviewed elsewhere[32, 33]. We focus on describing the fundamental principles of each method, recent developments enhancing their usability by improving accuracy and computational efficiency, and successful applications in drug discovery projects.

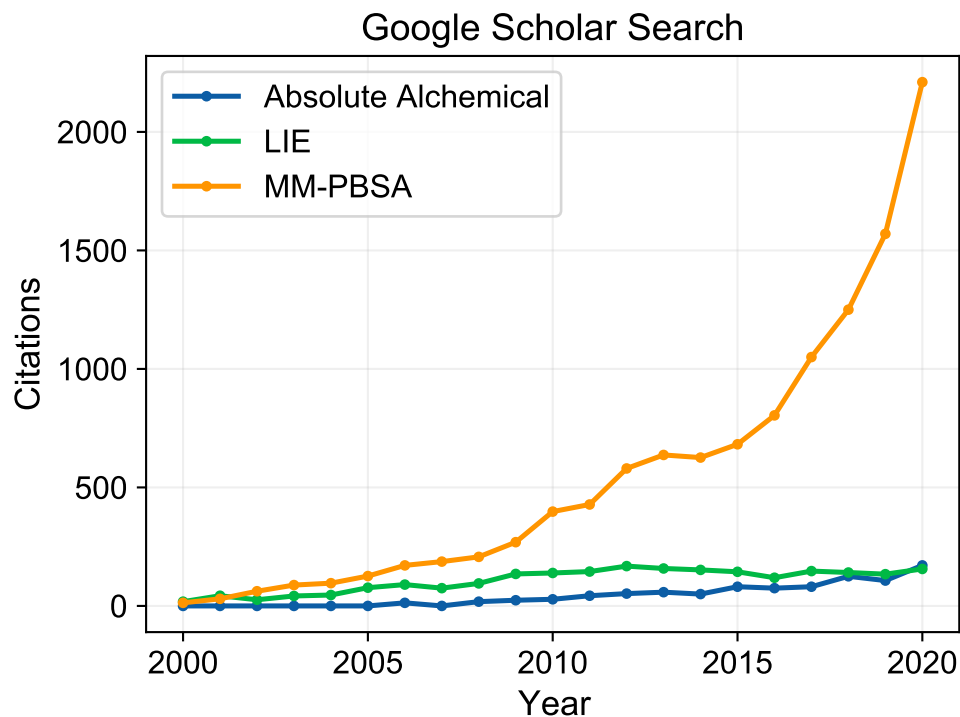
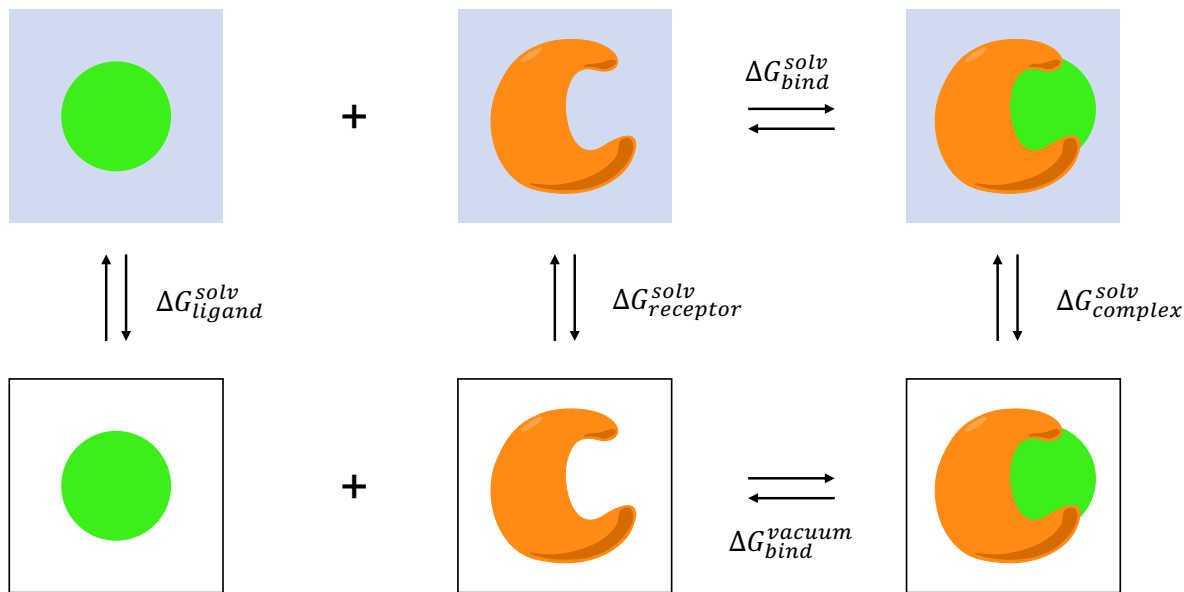


Figure 1.2: Citations per year for free energy methods. The development and utilization of molecular simulation to guide drug discovery has grown dramatically in recent years. The MM-PBSA method, which balances simulation rigor, high speed, and minimal setup complexity to allow high throughput screening, has seen extensive application reaching over 2,000 citations in 2020. Steep computational costs and challenges in generalizing protocols to work on broad sets of protein-ligand systems have limited the usage of absolute alchemical and LIE based approaches.

1.3 Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA)

The MM-PBSA method as applied to small molecule binding is an end-point method estimating the binding free-energy difference between the protein-ligand complex and the separate unbound components, the complex, ligand, and protein alone[13–19, 34–36] (Figure: 1.3). MM-PBSA provides a balanced approach characterized by improved rigor and accuracy over molecular docking, and with reduced computational demands compared to pathway methods such as alchemical transformations that require involved experimental setup to sample intermediate states through the decoupling of ligand interactions[37–40].



$$\Delta G_{bind}^{solv} = \Delta G_{bind}^{vacuum} + \Delta G_{complex}^{solv} - (\Delta G_{ligand}^{solv} + \Delta G_{receptor}^{solv})$$

Figure 1.3: MM-PBSA thermodynamic cycle. The binding free energy in aqueous environment is calculated as the difference between the sum of binding in vacuum and solvating the complex with solvating the receptor and ligand individually. The information necessary to complete this cycle can be obtained by decomposing a single trajectory into the ensemble desolvated receptor, ligand, and complex configurations, and computing the solvation free energies for each state with the Poisson-Boltzmann equation. Normal mode analysis can be performed to determine the contribution of entropy to the binding process.

In addition to only requiring end-point data, a further approximation with MM-PBSA that enables efficient free-energy calculation is the utilization of implicit solvation. By coarse-graining solvent as a continuum with uniform dielectric constant the treatment of solvent interactions is greatly simplified. However, this may lead to difficulties modeling highly charged ligands and recent works have focused on minimizing these errors[36].

Two main approaches are employed to generate the data for MM-PBSA binding energy predictions with both starting from molecular dynamics (MD) simulation in explicit solvent: multiple trajectories with the three components, complex, apo receptor, and ligand separately, or a single trajectory with the bound protein-ligand complex that is divided into the three components afterward[15, 19]. MD is carried out with explicit solvation to maximize accuracy of conformational

sampling, and frames are post-processed by removal of solvent and ion molecules. The converged trajectory is evaluated with each frame as an individual sample point to generate ensemble averages and uncertainty values for the energy quantities. The single-trajectory approach is favored for its straightforward implementation and cancellation of covalent energy errors as conformations for the complex and separated receptor and ligand are based on shared configurations. However, the single-trajectory method may not be optimal due to its reliance on the problematic assumption that ligand binding does not involve large-scale conformational changes[19, 41]. The multi-trajectory approach is better suited for binding events associated with large conformation changes, but is noted to produce noisier estimates and require longer simulation time to reach convergence as the complex and individual components can sample diverged conformations[17, 42].

The binding free energy between the ligand (L) and receptor (R) is defined as:

$$\Delta G_{bind} = G_R - G_L \tag{1.1}$$

The difference in free energy between the complex and individual components can be decomposed into enthalpic (ΔH) and entropic ($-T\Delta S$) terms evaluating changes in bonding interactions and conformational disorder with binding. The enthalpic energy term can be approximated as the gas-phase molecular mechanics energy (ΔE_{MM}) and solvation free energy (ΔG_{solv}). The configurational entropy ($-T\Delta S$) can be estimated with the normal mode or quasi-harmonic analysis[17, 43], but is often omitted due to high computational cost and difficulty obtaining convergence.

$$\Delta G_{bind} = \Delta H - T\Delta S \tag{1.2}$$

$$\approx \Delta E_{MM} + \Delta G_{solv} - T\Delta S \tag{1.3}$$

ΔE_{MM} is computed from the molecular mechanics force field and consists of the covalent energy ($\Delta E_{covalent}$), electrostatic energy (ΔE_{elec}), and van der Waals dispersion and repulsion energy

(ΔE_{vdW}). The covalent term includes changes in bonds (ΔE_{bond}), angles (ΔE_{angle}), and torsion ($\Delta E_{torsion}$) energies.

$$\Delta E_{MM} = \Delta E_{covalent} + \Delta E_{elec} + \Delta E_{vdW} \quad (1.4)$$

$$\Delta E_{covalent} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{torsion} \quad (1.5)$$

ΔG_{solv} describes the contribution of polar and non-polar interactions to the transfer of the ligand from gas phase to solvent. The polar solvation component (ΔG_{polar}) specifies the interaction energy of the solute’s charge distribution in the continuum solvent and is found by evaluation of the Poisson-Boltzmann equation (PBE)[20, 20, 34, 44–65]. The non-polar solvation term ($\Delta G_{non-polar}$) measures the energy from the solute forming a cavity in the solvent and the van der Waals interactions at the cavity interface between solute and solvent[66, 67], so that the total solvation free energy can be expressed as:

$$\Delta G_{solv} = \Delta G_{polar} + \Delta G_{non-polar} \quad (1.6)$$

The basis of the PBE is the Poisson equation with dielectric distribution $\varepsilon(r)$, electrostatic potential distribution $\psi(r)$, and fixed atomic charge density $\rho(r)$, where each function is dependent on the solute atom position vector (r).

$$\nabla\varepsilon(r) \nabla\psi(r) = -4\pi\rho(r) \quad (1.7)$$

To account for electrostatic interactions from ionic salt molecules in the solution, the electrostatic potential ($\psi(r)$) is solved with the PBE with the additional terms $\lambda(r)$ representing the ion-exclusion function set to 0 inside the Stern layer and molecular interior and 1 outside, and salt-related term $f(\psi(r))$ that depends on the electrostatic potential, the valence (z_i), electron charge (e), bulk

concentration (c_i), and temperature (T), with summation over all ion types (i).

$$\nabla\varepsilon(r) \nabla\psi(r) + \lambda(r) f(\psi(r)) = -4\pi\rho(r) \tag{1.8}$$

$$f(\psi(r)) = 4\pi \sum_i^n z_i e c_i \exp\left(-\frac{z_i e \psi(r)}{k_b T}\right) \tag{1.9}$$

The PBE can be linearized for easier numerical computation under conditions where the ionic strength and electric field are both weak. The linear PBE equation includes the modified Debye-Hückel parameter (κ^2), solvent dielectric constant (ε_{solv}), and solution ionic strength (I) where $I = z^2 c$.

$$\nabla\varepsilon(r) \nabla\psi(r) - \varepsilon_{solv} \kappa^2 \psi(r) = -4\pi\rho(r) \tag{1.10}$$

$$\kappa^2 = \frac{8\pi e^2 I}{\varepsilon_{solv} k_B T} \tag{1.11}$$

MM-PBSA is often used in tandem with the closely related Molecular Mechanics Generalized Born Surface Area (MM-GBSA) approach as both utilize the same set of inputs for the prediction of binding free energies with continuum solvation[36, 68]. The difference between the methods lies in the calculation of ΔG_{polar} where the GB model is based on an analytical expression approximating the PBE. This leads to large speed improvements, but predictive performance is generally reduced compared to PBE, though this is system dependent[35, 68]. The GB equation is composed of terms describing solute atoms as spheres with partial charge (q), internal dielectric (ε) and solvent dielectric (ε_0), distance between particles i and j (r_{ij}), and the effective Born radius (α).

$$\Delta G_{GB} = - \left(\frac{1}{\varepsilon} - \frac{1}{\varepsilon_0} \right) \sum_{ij} \frac{q_i q_j}{f_{GB}} \quad (1.12)$$

$$f_{GB} = \sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp \left(- \frac{r_{ij}^2}{4\alpha_i \alpha_j} \right)} \quad (1.13)$$

$\Delta G_{non-polar}$ has classically been determined as proportional to the solute’s solvent accessible surface area (*SASA*)[66, 67] as:

$$\Delta G_{non-polar}^{SA} = \gamma \cdot SASA + b \quad (1.14)$$

The surface tension constant (γ) describing the free energy of forming a cavity in water and the offset (b) are determined empirically and set as constants for all solute molecules. These variables are assigned as $\gamma = 0.00542$ kcal/mol-Å² and $b = 0.92$ kcal/mol in the AMBER package[18, 69]. Alternative methods with atom-specific surface tension constants have also been explored[70, 71].

More updated methods to resolve $\Delta G_{non-polar}$ incorporate the van der Waals dispersion free-energy as a separate term, treating the process as two events where a cavity is created and the non-polar solute is inserted into the cavity[67]. The separation of terms additionally allows individual scaling of the cavity formation and dispersion terms as a function of solute size. ΔG_{cavity} is calculated with similar linear regression as the classical $\Delta G_{non-polar}$ equation with *SASA* replaced with solvent accessible volume (*SAV*) and the attractive dispersion energy is computed through surface-integration. The updated scaling factors are set as $\gamma = 0.0378$ kcal/mol-Å³ and $b = -0.569$ kcal/mol in the AMBER package[18, 69].

$$\Delta G_{non-polar}^{CD} = \Delta G_{dispersion} + \Delta G_{cavity} \quad (1.15)$$

$$\Delta G_{cavity} = \gamma \cdot SAV + b \quad (1.16)$$

1.3.1 MM-PBSA developments and benchmarks

Improvements to the MM-PBSA method include more rigorous treatment of the dielectric constants and electrostatic polarization for better predictive accuracy on highly charged ligands, faster PB solvers, extension to pKa calculation, and novel schemes for determination of entropy. Scaling of the solute dielectric constant to tune the screening of electrostatic interactions in the non-polar protein environment is found to have a critical, receptor-dependent role on predictive accuracy[72]. Heterogenous dielectric values are applied to implicit membrane models where the dielectric is discretely varied with membrane depth[73], and with Gaussian dielectric to smoothly distribute the interface over protein cavities[74]. Integration of a Gaussian based model for molecular volume and surface area determination with the Gaussian dielectric distribution removes sharp surfaces separating the solute and solvent for a surface free approach to MM-PBSA calculation[75].

Electronic polarization effects can be incorporated through the use of polarizable force fields such as AMOEBA, this is implemented in the boundary integral PBE solver PyGBe[76]. Combination of the polarizable Drude oscillator force field with PBSA lowers RMSE from 2.5 kcal/mol with the standard CHARMM36 force field to 0.8 kcal/mol in calculation of solvation free energies for 70 molecules in addition to reducing errors in alanine scanning[77]. Coupling PBE calculation with Monte Carlo sampling of protonation states is applied to estimation of protonation free energies leading to pKa values within 2.05 pKa units RMSE of experiment using the Drude-PB method and within ~ 0.8 pKa units RMSE using PypKa[77, 78].

There are also updates to the PBE solvers through geometric multigrid on CPU allowing massively parallel scaling to 100 CPUs and a grid size of 10^9 [79], and GPU implementation leading to ~ 100 times speed up compared to CPU[80]. Introduction of analytical interface and surface regulation for the immersed interface method is proposed to improve stability and convergence and GPU implementation leads to 20 times speed up[81]. Regularization methods are investigated under the matched interface and boundary framework for proper treatment of charge singularities for higher numerical accuracy[82]. Finally extensions of the harmonic average method are proposed for fully taking advantage of the dense data parallelism to enhance the performance of PBE solvers on GPU

platforms[81].

Ensemble MM-PBSA calculation through use of multiple independent trajectories and maintenance of an explicit ligand hydration shell on the bromodomain-containing protein 4 system, a key regulator of transcription, showed robust reproducibility[83]. Menzer et al.[84] implement a confining potential on ligand external degrees of freedom and higher order cumulant expansion terms for average receptor-ligand interaction energies for more effective treatment of entropy.

A number of recent benchmarks identify best-practices to achieve optimal accuracy and directly compare MM-PBSA with other binding free energy prediction methods to highlight its advantages and disadvantages in drug discovery. When testing of MM-PBSA was performed on over 250,000 ligands for the GPCR superfamily following docking[85, 86], utilization of a single energy minimized structure is found to be the most computationally efficient method for virtual screening. In prediction of binding free energies and correct binding pose from 55 protein-RNA complexes, MM-PBSA (r_p -0.51) shows slightly lower performance than MM-GBSA (r_p -0.557)[87]. Molecular mechanics 3-dimensional reference interaction site model (MM-3D-RISM) is shown to have similar predictive performance as MM-PBSA, but differs in decomposition of polar and non-polar solvation energies[88]. Mishra and Koca[89] investigate the effects of simulation length, VDW radii sets, and combination with QM Hamiltonian on MM-PBSA predictions of protein-carbohydrate complexes. The conditions with optimal agreement to experiment are found to be 10 ns simulation with the mbondi radii set, and PM6 DFT method with QM resulting in the highest correlation of 0.96.

Entropic effects are further studied by Sun et al.[90] through comparison of normal mode analysis (NMA) and interaction entropy on over 1,500 protein-ligand systems with varying force fields. The most accurate results are obtained with the truncated NMA method, but due to high computational costs the authors recommend the interaction entropy approach instead, and force field choice made only minor differences. Enhanced sampling methods including aMD and GaMD are compared to conventional MD with MM-PBSA on protein-protein recognition, although the enhanced sampling methods are beneficial in encouraging exploration of conformational space, they do not improve binding affinity predictions on the timescales tested[91]. The effect of including a small

number of explicit water molecules and performing NMA for entropy calculation is examined for the bromodomain system[92]. Using a limited number of solvent molecules (~ 20) and entropy estimate improved MM-PBSA accuracy, although performance does not surpass absolute alchemical approaches the results came at significantly lower compute requirements.

The ease of performing MM-PBSA analysis and balance of speed and accuracy make it a popular method to use as an initial filter to rank drug candidates. Estimation of binding affinities with MM-PBSA for small-molecule protein-protein interaction inhibitors is automated with the farPPI web server[93] and prediction of changes in protein-DNA binding affinities upon mutation with the Single Amino acid Mutation binding free energy change of Protein-DNA Interaction (SAMPDI) web server[94]. Furthermore, due to its reliability MM-PBSA is often used as a baseline comparison or in combination with alternative methods for higher performance. Machine learning methods based on extracting protein-ligand interaction descriptors as features from MD simulation are compared to MM-PBSA on the tankyrase system[95]. Machine learning also accelerates pose prediction methods based on short MD simulation combined with MM-PBSA through the Best Arm Identification method to obtain the correct binding pose with minimal number of runs[96].

QM approaches allow more accurate consideration of nonbonded electrostatic interactions, but their usage is limited by high computational costs. This problem is addressed through fragment-based methods where localized regions of the protein-ligand system are treated with QM and the more global effects of solvation, entropy, and conformational sampling are evaluated through MM-PBSA analysis[97–100].

1.4 Linear Interaction Energy (LIE)

The Linear Interaction Energy (LIE) approach is another end-point method that predicts absolute binding free energies based on the change in free-energy from transferring the ligand from the solvated receptor-bound state to the aqueous free state[23, 24] (Figure: 1.5).

Bound Ligand



$U_{\text{complex+water}}$

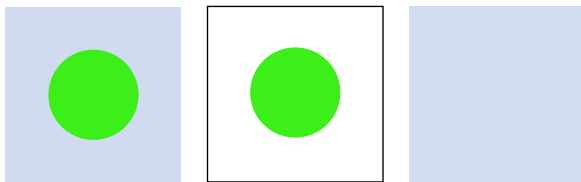
$U_{\text{receptor+water}}$

U_{ligand}

$$U_{\text{bound}}^{\text{elec}} = U_{\text{complex+water}}^{\text{elec}} - U_{\text{receptor+water}}^{\text{elec}} - U_{\text{ligand}}^{\text{elec}}$$

$$U_{\text{bound}}^{\text{vdW}} = U_{\text{complex+water}}^{\text{vdW}} - U_{\text{receptor+water}}^{\text{vdW}} - U_{\text{ligand}}^{\text{vdW}}$$

Free Ligand



$U_{\text{ligand+water}}$

U_{ligand}

U_{water}

$$U_{\text{free}}^{\text{elec}} = U_{\text{ligand+water}}^{\text{elec}} - U_{\text{ligand}}^{\text{elec}} - U_{\text{water}}^{\text{elec}}$$

$$U_{\text{free}}^{\text{vdW}} = U_{\text{ligand+water}}^{\text{vdW}} - U_{\text{receptor+water}}^{\text{vdW}} - U_{\text{ligand}}^{\text{vdW}}$$

$$\Delta G_{\text{LIE}} = \alpha(U_{\text{bound}}^{\text{vdW}} - U_{\text{free}}^{\text{vdW}}) + \beta(U_{\text{bound}}^{\text{elec}} - U_{\text{free}}^{\text{elec}}) + \gamma$$

Figure 1.4: LIE binding free energy calculation. The binding free energy is computed from force field energy estimates of the differences in van der Waals and electrostatic energies for the ligand bound to the protein and free in solvent environment. The system dependent LIE parameters α and β are empirically determined and used to scale the non-polar and coulombic interaction energies to have minimal error with respect to available experimental data. The final term γ acts as an optional offset parameter to further tune the model. LIE requires no post-processing and can be completed from a single trajectory.

$$\Delta G_{bind}(lig) = \Delta G_{solv}^{bound}(lig) - \Delta G_{solv}^{free}(lig) \quad (1.17)$$

This process considers binding in terms of the van der Waals (vdW) energy from creating the cavity in the target environment for the ligand and the electrostatic energy between the molecule and the environment. With that objective, LIE estimates ΔG_{bind} by an ensemble approach where two MD simulations are performed, with the ligand bound in the solvated protein and ligand free in solution, and the difference in VDW and electrostatic interactions between the ligand and environment in each case is measured[21, 22, 101].

$$\Delta G_{bind} = \left(\Delta G_{bound}^{polar} - \Delta G_{free}^{polar} \right) + \left(\Delta G_{bound}^{non-polar} - \Delta G_{free}^{non-polar} \right) \quad (1.18)$$

$$= \Delta \Delta G_{bind}^{polar} + \Delta \Delta G_{bind}^{non-polar} \quad (1.19)$$

The molecular mechanics force field applied in MD provides potential energies (U) composed of polar and non-polar components that can be converted into free-energies. The linear response approximation where averages of the electrostatic interaction energies between the ligand and environment is utilized to determine the polar term. The second term $\langle U_{lig-env}^{elec} \rangle_{off}$ representing the potential electrostatic energy from conformations sampled with interactions between ligand and environment turned off is a negligible constant, and is generally ignored[24].

$$\Delta G_{solv}^{elec} = \frac{1}{2} \left\{ \langle U_{lig-env}^{elec} \rangle_{on} - \langle U_{lig-env}^{elec} \rangle_{off} \right\} \quad (1.20)$$

$$= \frac{1}{2} \langle U_{lig-env}^{elec} \rangle_{on} \quad (1.21)$$

The scaling factor $\frac{1}{2}$ is replaced with the variable β , and the polar component for LIE free-energy calculation considering bound and free ligand simulation is:

$$\Delta G_{bind}^{polar} = \beta \left(\langle U_{lig-env}^{elec} \rangle_{bound} - \langle U_{lig-env}^{elec} \rangle_{free} \right) \quad (1.22)$$

$$= \beta \Delta \langle U_{lig-env}^{elec} \rangle \quad (1.23)$$

Non-polar interactions including hydrophobic packing and van der Waals interactions are derived from the Lennard-Jones potential force field term. Due to the observed linear correlation of solvation free energies for non-polar compounds with solute size, and similar linear scaling for average van der Waals interaction energies with solute size, LIE assumes that average van der Waals energies can be directly employed to capture non-polar binding contributions with a similarly formed estimate as the polar component[21].

$$\Delta G_{bind}^{non-polar} = \alpha \left(\langle U_{lig-env}^{vdW} \rangle_{bound} - \langle U_{lig-env}^{vdW} \rangle_{free} \right) \quad (1.24)$$

$$= \alpha \Delta \langle U_{lig-env}^{vdW} \rangle + \gamma \quad (1.25)$$

The set of three empirical parameters: α to scale the vdW interaction energies[102], β to scale coulombic interaction energies[101, 103], and γ as an optional offset constant[104], are all freely tunable. These parameters are known to be system dependent and must be calibrated based on available experimental data[105, 106]. Scaling of the model parameters is assumed to account for factors known to impact ΔG_{bind} but that are not explicitly declared including intramolecular energies, entropic confinement, desolvation effects, etc. The completed LIE estimation is based on force-field averaged energies and enables calculation of binding free energies solely through sampling of potential energies between the ligand and solvent or protein environments without post-processing

$$\Delta G_{bind} = \alpha \Delta \langle U_{lig-env}^{vdW} \rangle + \beta \Delta \langle U_{lig-env}^{elec} \rangle \quad (1.26)$$

1.4.1 LIE developments and benchmarks

As the least computationally expensive method, LIE is uniquely suited for high-throughput screening and recent efforts are devoted toward the direction of improving predictive accuracy, even if the calibrated parameters are system dependent. To this end, multiple alterations to the base LIE protocol are proposed to more rigorously account for polar and entropic interactions by including additional terms, combining LIE results with PBSA[107], or alchemical calculations, and utilizing ensemble docking poses with iterative LIE models. The extended linear interaction energy method (ELIE) introduced by He et al. includes the PBSA terms for the polar solvation energy, non-polar solvation energy, and entropic contribution and individual scaling factors for each[108]. Performance of ELIE in the Cathepsin S D3R 2017 Grand Challenge is found to show improved RMSE (1.17 kcal/mol) compared to MM-PBSA (3.19 kcal/mol)[108].

Further benchmarking on 8 drug targets with a series of congeneric ligands to examine the application of ELIE to drug lead optimization demonstrates that ELIE (0.94 kcal/mol RMSE) can approach the accuracy of Free Energy Perturbation (FEP)/Thermodynamic Integration (TI) (1.08/0.96 kcal/mol RMSE) methods when using receptor-specific parameters. The authors find that 25 ns MD simulations show optimal accuracy as it generally decreases with longer simulation[109]. The performance of LIE in host-guest systems is also evaluated on 4 host families (cucurbiturils, octa acids, β -cyclodextrin) with an array of 49 chemically diverse guests. The base LIE is modified to include host strain energy, and parameters are found to be transferable between the different host systems, notably resulting in binding predictions with RMSE below 1.5 kcal/mol through only a few nanoseconds of simulation[110]. Ngo et al. estimate HIV-1 protease inhibitor binding affinities with a modified LIE that includes a polar interaction term obtained from PBE, training on 22 samples and testing on a set of 11 ligands demonstrates good performance with 1.25 kcal/mol RMSE and 0.83 Pearson correlation[111].

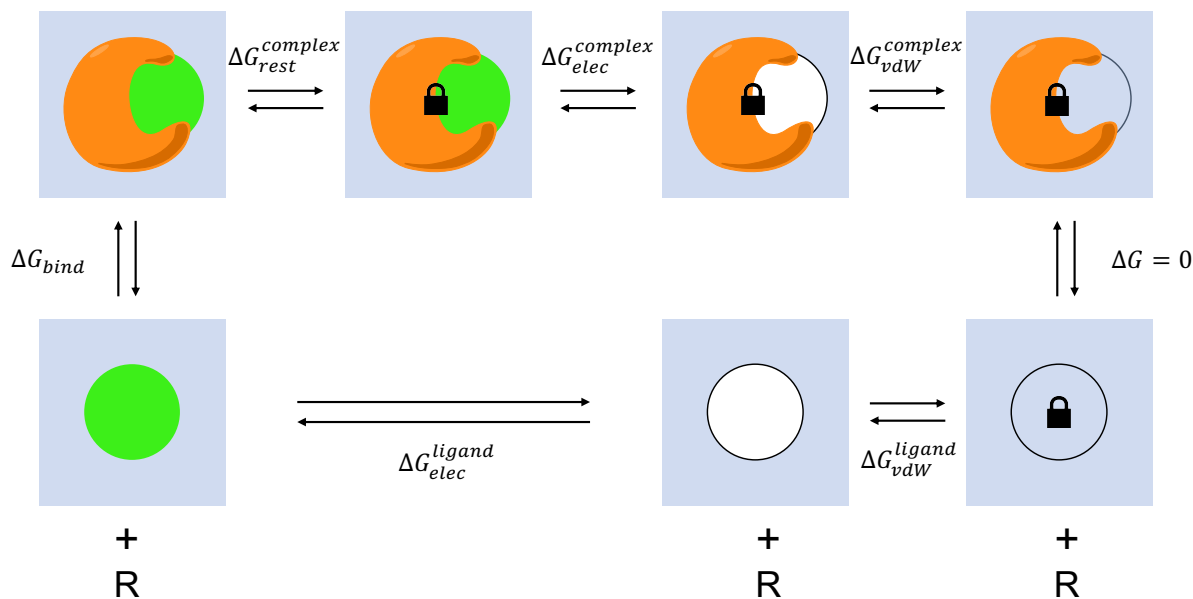
Proteins with flexible active sites may bind ligands in multiple orientations, this requires estimation of binding affinity from multiple poses weighted by their frequency to account for the contributions from each potential binding mode. Rifai et al. evaluate binding of inhibitors to malleable Cy-

tochrome P450s with an iterative weighing approach where each training compound is sampled with multiple simulations starting from different binding poses and LIE parameters are determined from Boltzmann weighing individual trajectory results[112]. Further accuracy is obtained by combining LIE with alchemical simulations to consider the ligand solvation free energies. Direct comparison of LIE with MM-PBSA on the SIRT1 system with a set of 27 inhibitors finds that both methods produce comparable Pearson correlations of 0.72 for LIE and 0.64 for MM-PBSA indicating good predictive value in ranking inhibitors, LIE is advantageous in requiring shorter simulation due to slow convergence of the MM-PBSA polar term[113]. The two-domain LIE (2D-LIE) approach is introduced to predict the binding free energy between protein domains and applied to computing cellulase kinetics[114].

1.5 Absolute alchemical simulations

End-point free energy prediction methods generally lack the ability to account for entropic and solvent effects, which play significant roles in protein-ligand interactions[115], except for methods that explicitly compute end-state free energies such as the Mining Minima method[116–121]. Capturing receptor conformation changes driven by ligand binding, water-mediated hydrogen-bonding, or solvent exchange that occurs as the ligand crowds the binding pocket are critical to rigorously estimate the free energy difference between the ligand bound and unbound states[122]. Pathway simulations tracking the MD trajectory of the ligand binding or unbinding event enable the computing of these effects, but come at high computational cost and increased simulation complexity[41, 123, 124]. The most direct approach to account for entropy and solvent effects in binding would be to simulate the receptor (R) and ligand (L) together and count the frequency of bound (RL) and unbound ($R + L$) conformations.





$$\Delta G_{bind} = (\Delta G_{elec}^{ligand} - \Delta G_{elec}^{complex}) + (\Delta G_{vdW}^{ligand} - \Delta G_{vdW}^{complex}) + \Delta G_{rest}^{complex}$$

Figure 1.5: Absolute alchemical simulation thermodynamic cycle. Two trajectories are completed to model the unbinding process. The simulations start from the complex of protein-ligand bound and end with receptor and unbound ligand (top track), and from ligand alone in solvent to ligand removed (bottom track). The ligand is transformed through a series of unphysical states to decouple electrostatic and van der Waals interactions with the surrounding environment to reach the final state where it no longer interacts with the initial system. The binding free energy prediction is the sum of the coulombic and non-polar energies involved in the transformation eliminating protein-ligand interactions. A restraint is typically included to prevent the ligand from exiting the active site while the binding interactions keeping the protein and ligand together are scaled off in order to aid convergence, this is corrected for with an additional transformation progressively turning on the restraints for the complex track and an analytical correction for the ligand track.

The ratio of bound to unbound states is an equilibrium constant (K_{eq}) that can be input into the Gibbs free energy equation where the Boltzmann constant (k_b) and temperature (T) are multiplied with the natural log of K_{eq} to calculate the binding free energy (G_{bind}).

$$K_{eq} = \frac{[RL]}{[R][L]} \tag{1.28}$$

$$\Delta G_{bind} = -k_b T \ln K_{eq} \tag{1.29}$$

In practice, it is not possible to estimate the equilibrium constant as the binding and unbinding events rarely occur within the timescales accessible with current simulation methods, leading to insufficient sampling. To bypass this sampling limitation, alchemical approaches modeling the gradual decoupling of electrostatic and van der Waals interactions between the ligand and receptor have been utilized to simulate the transition between ligand bound and unbound states without the need to physically capture the process[26]. The basis of this calculation is the thermodynamic cycle describing in one leg the removal of ligand from the complex, and in a parallel leg the removal of the ligand from solvent[30]. The end states with receptor alone and solvent alone interconvert with zero free energy difference as the ligand is absent from both systems, leaving the last transition between ligand in solvent to ligand bound to receptor solvable with knowledge of the free energy costs in transferring the ligand out of the receptor and out of solvent. This is typically performed through the Zwanzig equation also known as Exponential Averaging (EXP) or Free Energy Perturbation (FEP).

$$\Delta G_{AB} = -k_b T \ln \left\langle -\frac{1}{k_b T} (U_B - U_A) \right\rangle_A \tag{1.30}$$

EXP calculates the difference in potential of the end states using the ensemble of one simulated end state; however, this method is susceptible to bias in the free energies estimated due to poor phase space overlap of the end states[125].

Since free energy is a state function, its difference between states in the closed thermodynamic cycle is independent of the pathway taken, this includes non-physical intermediates that cannot be observed experimentally. The sampling of non-physical intermediate states is described by the parameter λ spanning from 0 where no perturbation has occurred to 1 where the ligand is fully decoupled from the environment and gives rise to the name alchemical. A drawback of the approach is the need for many intermediate states to guarantee accuracy of the simulation. The potential energies are computed for each intermediate state, and the free energy differences are calculated through thermodynamic integration by evaluating the integral of the ensemble averaged derivatives of potential energy with respect to λ [25, 126–129].

$$U(\lambda) = \lambda U_0 + (1 - \lambda) U_1 \tag{1.31}$$

$$\Delta G = \int_0^1 \left\langle \frac{dU_\lambda}{d\lambda} \right\rangle_\lambda d\lambda \tag{1.32}$$

Standard alchemical transformations are carried out in two stages, first with scaling ligand atom partial charges to model decoupling of electrostatics, and next with the van der Waals interactions[31?]. These two transformations are performed separately to avoid singularity artifacts that arise from atomic overlap created by strong attractive electrostatic interactions drawing atoms lacking steric bulk over others[130?]. It is also necessary to utilize an alternative “softcore” Lennard-Jones potential coupled to the λ window during the van der Waals scaling. Linear scaling with the standard Lennard-Jones potential leads to numerical instabilities at λ endpoints due to the severe repulsive forces calculated on overlapping atoms and contributes to poor phase space overlap with neighboring windows[131, 132]. An example “softcore” potential is illustrated as a function of the λ window and configuration (x), and contains the tunable parameters α, m, n and standard terms for the distance where the pair-wise potential is 0 (σ) and the distance separating the atoms (r)[131, 133, 134].

$$U(\lambda, x) = 4\varepsilon\lambda^n \left[\left(\alpha(1-\lambda)^m + \left(\frac{r}{\sigma}\right)^6 \right)^{-2} - \left(\alpha(1-\lambda)^m + \left(\frac{r}{\sigma}\right)^6 \right)^{-1} \right] \quad (1.33)$$

Further considerations involving the direction of the alchemical transformation, the utilization of restraints, the treatment of charge neutralization, λ window scheduling, procedure to select data samples that are both uncorrelated and equilibrated, and method to calculate free energy differences between the intermediate states must be made to ensure simulation stability and minimize variance in final free energy determination with the alchemical perturbation. The above factors all play some roles in the accuracy of the simulated free energies but are often not easy to decide *a priori*. Sampling of the physiologically relevant binding pose is essential to obtaining accurate values, initializing the alchemical transformation from an experimentally determined complex and modeling ligand decoupling generally maintains the ligand in the most applicable configurations[?]. Theoretically there should be no difference beginning from the opposite end point state with an empty active site and having the ligand grown in; however, this may require longer simulation time as the ligand can easily get trapped in local minima away from the true binding pose and sample irrelevant states. The ligand may leave the binding pocket as the interactions with the receptor are scaled, hindering convergence[135].

This is prevented by attaching restraints, which are later corrected for with an additional penalty term, to hold the ligand in the binding pocket. Two types of restraint schemes are common, the first involves imposing a single virtual bond between the ligand and receptor which is analytically corrected for by the formula

$$\Delta G_{restraint} = -k_b T \ln \left[\frac{8\pi V^0 K_r^{1/2}}{(2\pi k_b T)^{1/2}} \right] \quad (1.34)$$

where V^0 is the standard state volume and K_r the force constant[29, 136–138]. An alternative restraining approach, the 6DOF method introduced by Boresch et al.[30], enforces stricter adherence to a defined pose through one distance, two angular, and three dihedral restraints. Restraining the ligand to a single orientation expedites convergence, but may frustrate sampling of appropriate

conformations not directly captured in the crystal structure, leading to overestimation of binding affinities[139]. The 6DOF restraint correction is calculated with the following equation

$$\Delta G_{restraint}^{6DOF} = -k_b T \ln \left[\frac{8\pi V^0 (K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C})^{1/2}}{r_{a,A,0}^2 \sin \theta_{A,0} \sin \theta_{B,0} (2\pi k_b T)^3} \right] \quad (1.35)$$

where $r_{a,A,0}$ is the restrained distance, $\theta_{A,0}$ and $\theta_{B,0}$ are the two restrained angles, and K 's are the force constants[30].

The transformation of charged ligands demands corrections to maintain neutrality in the simulation box as the ligand partial charges are scaled[140]. Due to the usage of periodic boundary conditions, excess charges are propagated through all cells and cause errors in charge distribution[141–143]. This issue can be managed by performing the partial charge scaling simultaneously on a specified counter-ion[87, 144, 145], or through the correction scheme introduced by Rocklin et al.[146] based on an additional PB calculation to account for periodic finite-size effects.

The number and length of λ windows governs the variability of the free energy calculation[126, 147]. Increased sampling reduces the variance, but may not be worthwhile due to the added simulation costs. Rather than equally spacing the λ windows, a better strategy would be to more densely sample regions where transitions are non-linear near the end points of the van der Waals scaling stage and reduce sampling in more linear regions such as the electrostatic scaling. Datapoints from the beginning of each λ window are not yet equilibrated and sequential datapoints are autocorrelated, contamination with these energy values will distort the final free energy prediction[148]. Straightforward solutions to these problems would be to discard all data from the first half of the λ window and to only process energy values with large intervals. More sophisticated methods that aim to conserve as many datapoints as possible include the usage of automated equilibration detection based on reverse cumulative averaging[149] and subsampling of energies based on the calculated statistical inefficiency[148], this can be performed with the `pymbar`[150] package written by the Chodera group.

Lastly, thermodynamic integration is known to produce results with high variability due to the

numerical integration over highly non-linear functions. The Bennett Acceptance Ratio (BAR)[27] approach minimizes variance in the calculation of free energy by accounting for energies in neighboring states[125]. The BAR calculation self-consistently solves for the free energy (C) that satisfies the relations where i and j are consecutive states and U is the potential energy from a selected state.

$$\Delta G = \ln \frac{\sum_j f(U_i - U_j + C)}{\sum_j f(U_j - U_i + C)} + C \quad (1.36)$$

$$\Delta G = C \quad (1.37)$$

$$f(x) = \frac{1}{1 + e^x} \quad (1.38)$$

However, this method can face the same issues as EXP/FEP if there is no overlap between neighboring states. This has been extended to the Multistate Bennett Acceptance Ratio (MBAR)[150] method addressing the critical issues in BAR and produces the lowest variance of all free energy estimators by using energy differences from all λ windows[151].

1.5.1 Absolute alchemical developments and benchmarks

A major impediment to the usage of alchemical simulations is their complicated setup and data processing for ligand decharging and vdW removal stages. Updates to the popular molecular dynamics packages NAMD[152] and AMBER[153, 154] enable GPU accelerated calculation of the $\frac{dU_\lambda}{d\lambda}$ term necessary for thermodynamic integration or energy cross terms for sampling conformations at different lambda values for MBAR computation. To support high-throughput alchemical screening and improved reproducibility, a number of software packages automate the experimental setup in preparing the simulation files with appropriately decoupled ligand topologies and output the final binding free energy prediction after processing the trajectories. These include the VMD plugin BFEE[155], the python tool BAT.py[156] for AMBER, the CHARMM-GUI Free En-

ergy Calculator[157], the web platform Biomolecular Reaction and Interaction Dynamics Global Environment[158] (BRIDGE) for GROMACS, and Flare[159].

Improvements in simulation efficiency have allowed faster sampling of protein-ligand binding conformations and exploration of longer timescales to more comprehensively capture the significant perturbations that occur from ligand decoupling in absolute alchemical simulations. Giese et al.[134] utilize the simple but effective parameter interpolated thermodynamic integration (PI-TI) scheme where intermediate lambda states are defined by scaling the ligand molecular mechanic parameters, this allows taking full advantage of the standard GPU accelerated MD integrators and existing Hamiltonian replica exchange methods (HREMD) without the need to implement any alchemical specific code. Validation of this study examined pKa predictions on a double strand RNA system resulting in an error within 1.2 pKa units. Monte Carlo methods based on making unphysical, Boltzmann weighed rotamer and torsion moves lead to greater conformational sampling and crossing of energy barriers that would necessitate substantial simulation time in MD. Pure MC[160, 161] and the hybrid MC/MD method Binding modes of Ligands Using Enhanced Sampling (BLUES) involving random ligand rotations, relaxation with MD, and final acceptance or rejection through nonequilibrium Monte Carlo are demonstrated to have greater binding mode sampling efficiency than standard MD. Hamiltonian replicas parallelize sampling backbone torsions of T4 lysozyme[162] and solvent exchange in the cytochrome P450 binding site[163] to speed convergence within 1 ns in the latter study.

In cases where no reliable experimental structure with ligand bound is available, the generalized replica exchange with solute tempering (gREST) + FEP[164] approach where protein-ligand interactions are weakened through simulation at high temperature to force refinement of ligand binding orientation or Alchemical Grid Dock[165] method can be performed to obtain high quality binding poses. Alternative lambda schedules aimed at reducing the number of intermediate windows to simulate without sacrificing low variance are introduced by Konig et al.[166] with enveloping distribution sampling and addition of a restraint energy distribution function in the screening of SARS-CoV-2 protease inhibitors[167].

Metadynamics methods utilizing a history dependent bias potential to drive sampling of unexplored conformations are used for the theophylline-RNA complex to get within 0.02 kcal/mol of experiment[168]. The Gaussian algorithm enhanced FEP (GA-FEP) method is used to guide the design of Phosphodiesterase-10 inhibitors and overcomes poor sampling by fitting the observed energies to a multivariate Gaussian distribution to extrapolate better converged energy values for downstream BAR calculation[3]. Dual resolution models where the active site portions of the protein are modeled with full atom representation and other regions as coarse grained showed significant speedup with only minor loss in accuracy compared to the all-atom model for the lysozyme system binding with di-N-acetylchitotriose[169]. Sakae et al.[170] demonstrate a modified alchemical approach starting with unrestrained ligand for broader sampling of binding poses and bypass the need to exhaustively enumerate all potential binding modes. The DeepBAR method applies generative modeling to construct sample conformations of the cucurbit[7]uril host-guest system for the BAR analysis without the need for intermediate state sampling to achieve higher computational efficiency[171].

Advances in finite size and charge treatment schemes have improved accuracy in computing decharging energies, and new formulations for the evaluation of “soft-core” atoms lead to greater numerical stability and reduced variability in vdW removal. The poor representation of electronic polarization in molecular simulation makes binding affinity prediction for charged and titratable molecules challenging. Standard MD simulation is unable to model dielectric screening effects that alter the strength of ligand partial charges as it transitions between the polar solvent environment to the non-polar protein active site[139]. We demonstrate that scaling the dielectric constant with the MBAR/PBSA continuum solvent model provides a convenient method to reproduce the effects of charge polarization without requiring any modification to the MD integrator. RMSE for the predicted binding affinities of inhibitors for urokinase plasminogen activator is reduced from 3.2 kcal/mol with standard alchemical simulation to 0.89 kcal/mol with MBAR/PBSA[139]. The AMOEBA polarizable force field that incorporates electronic polarization through induced dipoles, atomic dipoles, and quadrupole terms is applied to the lead optimization of the MELK inhibitor IN17[138]. In the SAMPL7 TrimerTrip host-guest blind challenge, utilization of the AMOEBA

force field shows excellent results with 7/8 samples having errors within 2 kcal/mol[172–174].

The commonly used approach to maintain charge neutrality through co-alchemical ions is shown not to fully eliminate charge artifacts in periodic simulation boxes due to localized differences in electrostatic potentials and solvent densities for the distant ion and bound ligand[175]. Continuum-electrostatics calculations[176] and the “Warp-Drive”[177] method of simultaneously perturbing the protein-ligand complex and a distant unbound ligand are proposed to more accurately correct for finite-size effects. Difficulty in modeling the extraction of charged ligands from deeply buried binding sites with potential of mean force (PMF) methods is addressed with the AlchemPMF protocol where steric obstructions along the physical pathway are alchemically removed, resulting in improved binding free energy estimates on HIV-1 integrase and telomeric DNA G-quadruplex[178].

Li et al.[179] develop the Gaussian repulsive soft-core potential to produce a linear hybrid Hamiltonian with respect to lambda to allow improved simulation efficiency over the standard separation-shifted potential that generates non-linear Hamiltonians. Extension of smooth-step soft-core potentials that are composed of monotonically increasing polynomial functions that have the desirable end-point values enable one-step alchemical transformations by overcoming the issues of end-point catastrophe, particle collapse, and large gradient jumps[153].

Benchmarks of alchemical simulations demonstrate their utility and high accuracies. The SAMPL6 and SAMPL7 challenges[180] feature several entries examining alchemical approaches for CB[8] and tetra-methylated octa-acids host-guest systems with comparison to umbrella sampling[181, 182], TrimerTrip host-guest system with comparison of AM1-BCC and RESP charge schemes[183], and evaluation of GAFF and CGenFF force fields[184]. Novel applications of alchemical simulation include the estimation of binding affinity change upon protein mutation through the ensemble thermodynamic integration with enhanced sampling (TIES) approach on the fibroblast growth factor receptor 3 (FGFR3), notably simulations without enhanced sampling are unable to capture conformational changes driven by protein mutation in the binding site[4]. PMF methods based on utilizing restraints to physically pull the ligand out of the binding site are directly compared to absolute alchemical approaches on the HIV-1 integrase system by Deng et al.[185], the final results

show similar performance with absolute errors in the range of 1.6–4.3 kcal/mol for alchemical and 1.5–3.4 kcal/mol for PMF. The authors add that the alchemical approach supports simpler setup as they do not need to geometrically define the pathway for the ligand to exit the binding site.

Loeffler et al.[186] validate alchemical simulation results from different software packages in the calculation of hydration free energies and determine that the tested packages (AMBER, CHARMM, GROMACS, and SOMD) produce consistent free energies. The scale of alchemical simulations is growing dramatically by harnessing cloud computing[187]. The report of massive-scale simulation of 301 HIV-1 integrase inhibitors on the IBM World Community Grid[188] highlights how the availability of distributed computing is enabling high-throughput FEP screening.

Bibliography

- [1] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.*, 47:20–33, May 2016.
- [2] Shubhangi Kaushik, Sérgio M Marques, Prashant Khirsariya, Kamil Paruch, Lenka Libichova, Jan Brezovsky, Zbynek Prokop, Radka Chaloupkova, and Jiri Damborsky. Impact of the access tunnel engineering on catalysis is strictly ligand-specific. *FEBS J.*, 285(8):1456–1476, April 2018.
- [3] Zhe Li, Yiyong Huang, Yinuo Wu, Jingyi Chen, Deyan Wu, Chang-Guo Zhan, and Hai-Bin Luo. Absolute binding free energy calculation and design of a subnanomolar inhibitor of phosphodiesterase-10. *J. Med. Chem.*, 62(4):2099–2111, February 2019.
- [4] Agastya P Bhati, Shunzhou Wan, and Peter V Coveney. Ensemble-Based replica exchange alchemical free energy methods: The effect of protein mutations on inhibitor binding. *J. Chem. Theory Comput.*, 15(2):1265–1277, February 2019.
- [5] Fumie Ono, Shuntaro Chiba, Yuta Isaka, Shigeyuki Matsumoto, Biao Ma, Ryohei Katayama, Mitsugu Araki, and Yasushi Okuno. Improvement in predicting drug sensitivity changes associated with protein mutations using a molecular dynamics based alchemical mutation method. *Sci. Rep.*, 10(1):1–10, February 2020.
- [6] Zhi Chen, Hualei Wang, Lin Yang, Shuiqing Jiang, and Dongzhi Wei. Significantly enhancing the stereoselectivity of a regioselective nitrilase for the production of (s)-3-cyano-5-

- methylhexanoic acid using an MM/PBSA method. *Chem. Commun.*, 57(7):931–934, January 2021.
- [7] Matteo Aldeghi, Vytautas Gapsys, and Bert L de Groot. Accurate estimation of ligand binding affinity changes upon protein mutation. *ACS Cent. Sci.*, December 2018.
- [8] Zuzana Jandova, Daniel Fast, Martina Setz, Maria Pechlaner, and Chris Oostenbrink. Saturation mutagenesis by efficient Free-Energy calculation. *J. Chem. Theory Comput.*, 14(2): 894–904, February 2018.
- [9] Danial Pourjafar-Dehkordi, Sophie Vieweg, Aymelt Itzen, and Martin Zacharias. Phosphorylation of ser111 in rab8a modulates Rabin8-Dependent activation by perturbation of side chain interaction networks. *Biochemistry*, 58(33):3546–3554, August 2019.
- [10] William R Martin, Felice C Lightstone, and Feixiong Cheng. In silico insights into protein-protein interaction disruptive mutations in the PCSK9-LDLR complex. *Int. J. Mol. Sci.*, 21(5), February 2020.
- [11] Jing Xue, Xiaoqiang Huang, and Yushan Zhu. Using molecular dynamics simulations to evaluate active designs of cephradine hydrolase by molecular mechanics/poisson boltzmann surface area and molecular mechanics/generalized born surface area methods. *RSC Adv.*, 9(24):13868–13877, April 2019.
- [12] Kexin Wang, Qiuxia Huang, Hanxin Li, and Xihua Zhao. Co-evolution of beta-glucosidase activity and product tolerance for increasing cellulosic ethanol yield. *Biotechnol. Lett.*, 42(11):2239–2250, November 2020.
- [13] T E Cheatham, 3rd, J Srinivasan, D A Case, and P A Kollman. Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. *J. Biomol. Struct. Dyn.*, 16(2):265–280, October 1998.
- [14] Jayashree Srinivasan, Thomas E Cheatham, Piotr Cieplak, Peter A Kollman, and David A Case. Continuum solvent studies of the stability of DNA, RNA, and Phosphoramidate–DNA helices. *J. Am. Chem. Soc.*, 120(37):9401–9409, September 1998.

- [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, 3rd. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33(12):889–897, December 2000.
- [16] Holger Gohlke and David A Case. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.*, 25(2):238–250, January 2004.
- [17] Tianyi Yang, Johnny C Wu, Chunli Yan, Yuanfeng Wang, Ray Luo, Michael B Gonzales, Kevin N Dalby, and Pengyu Ren. Virtual screening using molecular simulations. *Proteins*, 79(6):1940–1951, June 2011.
- [18] Bill R Miller, 3rd, T Dwight McGee, Jr, Jason M Swails, Nadine Homeyer, Holger Gohlke, and Adrian E Roitberg. MMPBSA.py: An efficient program for End-State free energy calculations. *J. Chem. Theory Comput.*, 8(9):3314–3321, September 2012.
- [19] Changhao Wang, Peter H Nguyen, Kevin Pham, Danielle Huynh, Thanh-Binh Nancy Le, Hongli Wang, Pengyu Ren, and Ray Luo. Calculating protein-ligand binding affinities with MMPBSA: Method and error analysis. *J. Comput. Chem.*, 37(27):2436–2446, October 2016.
- [20] Changhao Wang, D’artagnan Greene, Li Xiao, Ruxi Qi, and Ray Luo. Recent developments and applications of the MMPBSA method. *Front Mol Biosci*, 4:87, 2017.
- [21] Johan Åqvist, Carmen Medina, and Jan-Erik Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng. Des. Sel.*, 7(3):385–391, March 1994.
- [22] J Aqvist and J Marelius. The linear interaction energy method for predicting ligand binding free energies. *Comb. Chem. High Throughput Screen.*, 4(8):613–626, December 2001.
- [23] Johan Åqvist, Victor B Luzhkov, and Bjørn O Brandsdal. Ligand binding affinities from MD simulations. *Acc. Chem. Res.*, 35(6):358–365, June 2002.
- [24] Hugo Gutiérrez-de Terán and Johan Aqvist. Linear interaction energy: method and applications in drug design. *Methods Mol. Biol.*, 819:305–323, 2012.

- [25] John G Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3(5):300–313, May 1935.
- [26] Robert W Zwanzig. High-Temperature equation of state by a perturbation method. i. non-polar gases. *J. Chem. Phys.*, 22(8):1420–1426, August 1954.
- [27] Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *J. Comput. Phys.*, 22(2):245–268, October 1976.
- [28] T P Straatsma and J A McCammon. Multiconfiguration thermodynamic integration. *J. Chem. Phys.*, 95(2):1175–1188, July 1991.
- [29] 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(3):1047–1069, March 1997.
- [30] Stefan Boresch, Franz Tettinger, Martin Leitgeb, and Martin Karplus. Absolute binding free energies: A quantitative approach for their calculation. *J. Phys. Chem. B*, 107(35):9535–9551, September 2003.
- [31] Michael R Shirts. Best practices in free energy calculations for drug design. *Methods Mol. Biol.*, 819:425–467, 2012.
- [32] Zoe Cournia, Bryce Allen, and Woody Sherman. Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. *J. Chem. Inf. Model.*, 57(12):2911–2937, December 2017.
- [33] Lin Frank Song and Kenneth M Merz, Jr. Evolution of alchemical free energy methods in drug discovery. *J. Chem. Inf. Model.*, 60(11):5308–5318, November 2020.
- [34] Kim A Sharp and Barry Honig. Calculating total electrostatic energies with the nonlinear poisson boltzmann equation. *J. Phys. Chem.*, 94(19):7684–7692, September 1990.
- [35] Samuel Genheden and Ulf Ryde. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.*, 10(5):449–461, May 2015.

- [36] Ercheng Wang, Huiyong Sun, Junmei Wang, Zhe Wang, Hui Liu, John Z H Zhang, and Tingjun Hou. End-Point binding free energy calculation with MM/PBSA and MM/GBSA: Strategies and applications in drug design. *Chem. Rev.*, 119(16):9478–9508, August 2019.
- [37] Giulio Rastelli, Alberto Del Rio, Gianluca Degliesposti, and Miriam Sgobba. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comput. Chem.*, 31(4):797–810, March 2010.
- [38] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.*, 51(1):69–82, January 2011.
- [39] Inna Slynko, Michael Scharfe, Tobias Rumpf, Julia Eib, Eric Metzger, Roland Schüle, Manfred Jung, and Wolfgang Sippl. Virtual screening of PRK1 inhibitors: ensemble docking, rescoring using binding free energy calculation and QSAR model development. *J. Chem. Inf. Model.*, 54(1):138–150, January 2014.
- [40] Huiyong Sun, Youyong Li, Mingyun Shen, Sheng Tian, Lei Xu, Peichen Pan, Yan Guan, and Tingjun Hou. Assessing the performance of MM/PBSA and MM/GBSA methods. 5. improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Phys. Chem. Chem. Phys.*, 16(40):22035–22045, 2014.
- [41] Michael S Lee and Mark A Olson. Calculation of absolute protein-ligand binding affinity using path and endpoint approaches. *Biophys. J.*, 90(3):864–877, February 2006.
- [42] Jessica M J Swanson, Richard H Henchman, and J Andrew McCammon. Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.*, 86(1 Pt 1):67–74, January 2004.
- [43] Summer Kassem, Marawan Ahmed, Salah El-Sheikh, and Khaled H Barakat. Entropy in bimolecular simulations: A comprehensive review of atomic fluctuations-based methods. *J. Mol. Graph. Model.*, 62:105–117, November 2015.
- [44] M F Perutz. Electrostatic effects in proteins. *Science*, 201(4362):1187–1191, September 1978.

- [45] J Warwicker and H C Watson. Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. *J. Mol. Biol.*, 157(4):671–679, June 1982.
- [46] D Bashford and M Karplus. pka’s of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry*, 29(44):10219–10225, November 1990.
- [47] Malcolm E Davis and J Andrew McCammon. Electrostatics in biomolecular structure and dynamics. *Chem. Rev.*, 90(3):509–521, May 1990.
- [48] Arald Jean-Charles, Anthony Nicholls, Kim Sharp, Barry Honig, Anna Tempczyk, Thomas F Hendrickson, and W Clark Still. Electrostatic contributions to solvation energies: comparison of free energy perturbation and continuum calculations. *J. Am. Chem. Soc.*, 113(4):1454–1455, February 1991.
- [49] M K Gilson. Theory of electrostatic interactions in macromolecules. *Curr. Opin. Struct. Biol.*, 5(2):216–223, April 1995.
- [50] B Honig and A Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, May 1995.
- [51] Shlomit R Edinger, Christian Cortis, Peter S Shenkin, and Richard A Friesner. Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of the Poisson–Boltzmann equation. *J. Phys. Chem. B*, 101(7):1190–1197, February 1997.
- [52] Rui Luo, John Moulton, and Michael K Gilson. Dielectric screening treatment of electrostatic solvation. *J. Phys. Chem. B*, 101(51):11226–11236, December 1997.
- [53] Ray Luo, Laurent David, and Michael K Gilson. Accelerated poisson boltzmann calculations for static and dynamic systems. *J. Comput. Chem.*, 23(13):1244–1253, October 2002.
- [54] Qiang Lu and Ray Luo. A Poisson–Boltzmann dynamics method with nonperiodic boundary condition. *J. Chem. Phys.*, 119(21):11035–11047, December 2003.

- [55] Chunhu Tan, Lijiang Yang, and Ray Luo. How well does poisson boltzmann implicit solvent agree with explicit solvent? a quantitative analysis. *J. Phys. Chem. B*, 110(37):18680–18687, September 2006.
- [56] Qin Cai, Jun Wang, Hong-Kai Zhao, and Ray Luo. On removal of charge singularity in Poisson–Boltzmann equation. *J. Chem. Phys.*, 130(14):145101, April 2009.
- [57] Jun Wang, Qin Cai, Zhi-Lin Li, Hong-Kai Zhao, and Ray Luo. Achieving energy conservation in Poisson–Boltzmann molecular dynamics: Accuracy and precision with finite-difference algorithms. *Chem. Phys. Lett.*, 468(4):112–118, January 2009.
- [58] Xiang Ye, Qin Cai, Wei Yang, and Ray Luo. Roles of boundary conditions in DNA simulations: analysis of ion distributions with the finite-difference poisson boltzmann method. *Biophys. J.*, 97(2):554–562, July 2009.
- [59] Qin Cai, Meng-Juei Hsieh, Jun Wang, and Ray Luo. Performance of nonlinear finite difference poisson boltzmann solvers. *J. Chem. Theory Comput.*, 6(1):203–211, January 2010.
- [60] Jun Wang, Chunhu Tan, Emmanuel Chanco, and Ray Luo. Quantitative analysis of Poisson–Boltzmann implicit solvent in molecular dynamics. *Phys. Chem. Chem. Phys.*, 12(5):1194–1202, 2010.
- [61] Jun Wang and Ray Luo. Assessment of linear finite-difference poisson boltzmann solvers. *J. Comput. Chem.*, 31(8):1689–1698, June 2010.
- [62] Xiang Ye, Jun Wang, and Ray Luo. A revised density function for molecular surface calculation in continuum solvent models. *J. Chem. Theory Comput.*, 6(4):1157–1169, April 2010.
- [63] Qin Cai, Xiang Ye, Jun Wang, and Ray Luo. On-the-Fly numerical surface integration for Finite-Difference Poisson–Boltzmann methods. *J. Chem. Theory Comput.*, 7(11):3608–3619, November 2011.
- [64] Meng-Juei Hsieh and Ray Luo. Exploring a coarse-grained distributive strategy for finite-difference Poisson–Boltzmann calculations. *J. Mol. Model.*, 17(8):1985–1996, August 2011.

- [65] Wesley M Botello-Smith, Xingping Liu, Qin Cai, Zhilin Li, Hongkai Zhao, and Ray Luo. Numerical Poisson–Boltzmann model for continuum membrane systems. *Chem. Phys. Lett.*, 555:274–281, January 2013.
- [66] Jason A Wagoner and Nathan A Baker. Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Proc. Natl. Acad. Sci. U. S. A.*, 103(22):8331–8336, May 2006.
- [67] Chunhu Tan, Yu-Hong Tan, and Ray Luo. Implicit nonpolar solvent models. *J. Phys. Chem. B*, 111(42):12263–12274, October 2007.
- [68] Fu Chen, Hui Liu, Huiyong Sun, Peichen Pan, Youyong Li, Dan Li, and Tingjun Hou. Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. capability to predict protein–protein binding free energies and re-rank binding poses generated by protein–protein docking. *Phys. Chem. Chem. Phys.*, 18(32):22129–22139, 2016.
- [69] David A Case, Thomas E Cheatham, 3rd, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, December 2005.
- [70] D Eisenberg and A D McLachlan. Solvation energy in protein folding and binding. *Nature*, 319(6050):199–203, 1986.
- [71] T Ooi, M Oobatake, G Némethy, and H A Scheraga. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. U. S. A.*, 84(10):3086–3090, May 1987.
- [72] Xiao Hu and Alessandro Contini. Rescoring virtual screening results with the MM-PBSA methods: Beware of internal dielectric constants. *J. Chem. Inf. Model.*, 59(6):2714–2728, June 2019.
- [73] D’artagnan Greene, Ruxi Qi, Remy Nguyen, Tianyin Qiu, and Ray Luo. Heterogeneous dielectric implicit membrane model for the calculation of MMPBSA binding free energies. *J. Chem. Inf. Model.*, 59(6):3041–3056, June 2019.

- [74] Tania Hazra, Sheik Ahmed Ullah, Siwen Wang, Emil Alexov, and Shan Zhao. A super-gaussian poisson boltzmann model for electrostatic free energy calculation: smooth dielectric distribution for protein cavities and in both water and vacuum states. *J. Math. Biol.*, 79(2): 631–672, July 2019.
- [75] Arghya Chakravorty, Emilio Gallicchio, and Emil Alexov. A grid-based algorithm in conjunction with a gaussian-based model of atoms for describing molecular geometry. *J. Comput. Chem.*, 40(12):1290–1304, May 2019.
- [76] Christopher D Cooper. A boundary-integral approach for the Poisson-Boltzmann equation with polarizable force fields. *J. Comput. Chem.*, 40(18):1680–1692, July 2019.
- [77] Alexey Aleksandrov, Benoît Roux, and Alexander D MacKerell. pka calculations with the polarizable drude force field and Poisson–Boltzmann solvation model. *J. Chem. Theory Comput.*, 16(7):4655–4668, July 2020.
- [78] Pedro B P S Reis, Diogo Vila-Viçosa, Walter Rocchia, and Miguel Machuqueiro. PypKa: A flexible python module for Poisson–Boltzmann-Based pka calculations. *J. Chem. Inf. Model.*, 60(10):4442–4448, October 2020.
- [79] James C Womack, Lucian Anton, Jacek Dziedzic, Phil J Hasnip, Matt I J Probert, and Chris-Kriton Skylaris. DL_MG: A parallel multigrid poisson and Poisson-Boltzmann solver for electronic structure calculations in vacuum and solution. *J. Chem. Theory Comput.*, 14(3):1412–1432, March 2018.
- [80] Ruxi Qi and Ray Luo. Robustness and efficiency of Poisson–Boltzmann modeling on graphics processing units. *J. Chem. Inf. Model.*, 59(1):409–420, January 2019.
- [81] Haixin Wei, Ray Luo, and Ruxi Qi. An efficient second-order poisson-boltzmann method. *J. Comput. Chem.*, 40(12):1257–1269, May 2019.
- [82] Arum Lee, Weihua Geng, and Shan Zhao. Regularization methods for the Poisson-Boltzmann equation: Comparison and accuracy recovery. *J. Comput. Phys.*, page 109958, October 2020.

- [83] David W Wright, Shunzhou Wan, Christophe Meyer, Herman van Vlijmen, Gary Tresadern, and Peter V Coveney. Application of ESMACS binding free energy protocols to diverse datasets: Bromodomain-containing protein 4. *Sci. Rep.*, 9(1):6017, April 2019.
- [84] William M Menzer, Chen Li, Wenji Sun, Bing Xie, and David D L Minh. Simple entropy terms for End-Point binding free energy calculations. *J. Chem. Theory Comput.*, 14(11):6035–6049, November 2018.
- [85] Mei Qian Yau, Abigail L Emtage, Nathaniel J Y Chan, Stephen W Doughty, and Jason S E Loo. Evaluating the performance of MM/PBSA for binding affinity prediction using class a GPCR crystal structures. *J. Comput. Aided Mol. Des.*, 33(5):487–496, May 2019.
- [86] Mei Qian Yau, Abigail L Emtage, and Jason S E Loo. Benchmarking the performance of MM/PBSA in virtual screening enrichment using the GPCR-Bench dataset. *J. Comput. Aided Mol. Des.*, 34(11):1133–1145, November 2020.
- [87] Wei Chen, Yuqing Deng, Ellery Russell, Yujie Wu, Robert Abel, and Lingle Wang. Accurate calculation of relative binding free energies between ligands with different net charges. *J. Chem. Theory Comput.*, November 2018.
- [88] Preeti Pandey, Rakesh Srivastava, and Pradipta Bandyopadhyay. Comparison of molecular mechanics-Poisson-Boltzmann surface area (MM-PBSA) and molecular mechanics-three-dimensional reference interaction site model (MM-3D-RISM) method to calculate the binding free energy of protein-ligand complexes: Effect of metal ion and advance statistical test. *Chem. Phys. Lett.*, 695:69–78, 2018.
- [89] Sushil K Mishra and Jaroslav Koča. Assessing the performance of MM/PBSA, MM/GBSA, and QM-MM/GBSA approaches on Protein/Carbohydrate complexes: Effect of implicit solvent models, QM methods, and entropic contributions. *J. Phys. Chem. B*, 122(34):8113–8121, August 2018.
- [90] Huiyong Sun, Lili Duan, Fu Chen, Hui Liu, Zhe Wang, Peichen Pan, Feng Zhu, John Z H Zhang, and Tingjun Hou. Assessing the performance of MM/PBSA and MM/GBSA methods.

7. entropy effects on the performance of end-point binding free energy calculation approaches. *Phys. Chem. Chem. Phys.*, 20(21):14450–14460, May 2018.
- [91] Ercheng Wang, Gaoqi Weng, Huiyong Sun, Hongyan Du, Feng Zhu, Fu Chen, Zhe Wang, and Tingjun Hou. Assessing the performance of the MM/PBSA and MM/GBSA methods. 10. impacts of enhanced sampling and variable dielectric model on protein–protein interactions. *Phys. Chem. Chem. Phys.*, 21(35):18958–18969, September 2019.
- [92] Matteo Aldeghi, Michael J Bodkin, Stefan Knapp, and Philip C Biggin. Statistical analysis on the performance of molecular mechanics Poisson-Boltzmann surface area versus absolute binding free energy calculations: Bromodomains as a case study. *J. Chem. Inf. Model.*, 57(9):2203–2221, September 2017.
- [93] Zhe Wang, Xuwen Wang, Youyong Li, Tailong Lei, Ercheng Wang, Dan Li, Yu Kang, Feng Zhu, and Tingjun Hou. farPPI: a webserver for accurate prediction of protein-ligand binding structures for small-molecule PPI inhibitors by MM/PB(GB)SA methods. *Bioinformatics*, 35(10):1777–1779, May 2019.
- [94] Yunhui Peng, Lexuan Sun, Zhe Jia, Lin Li, and Emil Alexov. Predicting protein-DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver. *Bioinformatics*, 34(5):779–786, March 2018.
- [95] Vladimir P Berishvili, Valentin O Perkin, Andrew E Voronkov, Eugene V Radchenko, Riyaz Syed, Chittireddy Venkata Ramana Reddy, Viness Pillay, Pradeep Kumar, Yahya E Choonara, Ahmed Kamal, and Vladimir A Palyulin. Time-Domain analysis of molecular dynamics trajectories using deep neural networks: Application to activity ranking of tankyrase inhibitors. *J. Chem. Inf. Model.*, 59(8):3519–3532, August 2019.
- [96] Kei Terayama, Hiroaki Iwata, Mitsugu Araki, Yasushi Okuno, and Koji Tsuda. Machine learning accelerates MD-based binding pose prediction between ligands and proteins. *Bioinformatics*, 34(5):770–778, March 2018.
- [97] Yaqian Wang, Jinfeng Liu, Jinjin Li, and Xiao He. Fragment-based quantum mechanical

- calculation of protein-protein binding affinities. *J. Comput. Chem.*, 39(21):1617–1628, August 2018.
- [98] Noriaki Okimoto, Takao Otsuka, Yoshinori Hirano, and Makoto Taiji. Use of the multi-layer fragment molecular orbital method to predict the rank order of Protein–Ligand binding affinities: A case study using tankyrase 2 inhibitors. *ACS Omega*, 3(4):4475–4485, April 2018.
- [99] Yoshio Okiyama, Tatsuya Nakano, Chiduru Watanabe, Kaori Fukuzawa, Yuji Mochizuki, and Shigenori Tanaka. Fragment molecular orbital calculations with implicit solvent based on the Poisson–Boltzmann equation: Implementation and DNA study. *J. Phys. Chem. B*, 122(16):4457–4471, April 2018.
- [100] Yoshio Okiyama, Chiduru Watanabe, Kaori Fukuzawa, Yuji Mochizuki, Tatsuya Nakano, and Shigenori Tanaka. Fragment molecular orbital calculations with implicit solvent based on the Poisson–Boltzmann equation: II. protein and its Ligand-Binding system studies. *J. Phys. Chem. B*, 123(5):957–973, February 2019.
- [101] T Hansson, J Marelius, and J Aqvist. Ligand binding affinity prediction by linear interaction energy methods. *J. Comput. Aided Mol. Des.*, 12(1):27–35, January 1998.
- [102] Wei Wang, Jian Wang, and Peter A Kollman. What determines the van der waals coefficient ϵ in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations? *Proteins*, 34(3):395–402, February 1999.
- [103] Johan Åqvist and Tomas Hansson. On the validity of electrostatic linear response in polar solvents. *J. Phys. Chem.*, 100(22):9512–9521, January 1996.
- [104] Martin Almlöf, Bjørn O Brandsdal, and Johan Aqvist. Binding affinity prediction with different force fields: examination of the linear interaction energy method. *J. Comput. Chem.*, 25(10):1242–1254, July 2004.
- [105] Martin Almlöf, Jens Carlsson, and Johan Åqvist. Improving the accuracy of the linear interaction energy method for solvation free energies. *J. Chem. Theory Comput.*, 3(6):2162–2175, November 2007.

- [106] Marc van Dijk, Antonius M Ter Laak, Jörg D Wichard, Luigi Capoferri, Nico P E Vermeulen, and Daan P Geerke. Comprehensive and automated linear interaction energy based Binding-Affinity prediction for multifarious cytochrome P450 aromatase inhibitors. *J. Chem. Inf. Model.*, 57(9):2294–2308, September 2017.
- [107] Kaifang Huang, Song Luo, Yalong Cong, Susu Zhong, John Z H Zhang, and Lili Duan. An accurate free energy estimator: based on MM/PBSA combined with interaction entropy for protein–ligand binding affinity. *Nanoscale*, 12(19):10737–10750, 2020.
- [108] Xibing He, Viet H Man, Beihong Ji, Xiang-Qun Xie, and Junmei Wang. Calculate protein–ligand binding affinities with the extended linear interaction energy method: application on the cathepsin S set in the D3R grand challenge 3. *J. Comput. Aided Mol. Des.*, 33(1):105–117, January 2019.
- [109] Dongxiao Hao, Xibing He, Beihong Ji, Shengli Zhang, and Junmei Wang. How well does the extended linear interaction energy method perform in accurate binding free energy calculations? *J. Chem. Inf. Model.*, November 2020.
- [110] Joel José Montalvo-Acosta, Paulina Pacak, Diego Enry Barreto Gomes, and Marco Cecchini. A linear interaction energy model for cavitand Host–Guest binding affinities. *J. Phys. Chem. B*, 122(26):6810–6814, July 2018.
- [111] Son Tung Ngo, Nam Dao Hong, Anh Le Huu Quynh, Dinh Minh Hiep, and Nguyen Thanh Tung. Effective estimation of the inhibitor affinity of HIV-1 protease via a modified LIE approach. *RSC Adv.*, 10(13):7732–7739, February 2020.
- [112] Eko Aditya Rifai, Valerio Ferrario, Jürgen Pleiss, and Daan P Geerke. Combined linear interaction energy and alchemical solvation Free-Energy approach for Protein-Binding affinity computation. *J. Chem. Theory Comput.*, 16(2):1300–1310, February 2020.
- [113] Eko Aditya Rifai, Marc van Dijk, Nico P E Vermeulen, Arry Yanuar, and Daan P Geerke. A comparative linear interaction energy and MM/PBSA study on SIRT1-Ligand binding free energy calculation. *J. Chem. Inf. Model.*, 59(9):4018–4033, September 2019.

- [114] Kay S Schaller, Jeppe Kari, Gustavo A Molina, Kasper D Tidemand, Kim Borch, Günther H J Peters, and Peter Westh. Computing cellulase kinetics with a Two-Domain linear interaction energy approach. *ACS Omega*, 6(2):1547–1555, January 2021.
- [115] David L Mobley and Ken A Dill. Binding of Small-Molecule ligands to proteins: “what you see” is not always “what you get”. *Structure*, 17(4):489–498, April 2009.
- [116] R Luo, M S Head, J A Given, and M K Gilson. Nucleic acid base-pairing and n-methylacetamide self-association in chloroform: affinity and conformation. *Biophys. Chem.*, 78(1-2):183–193, April 1999.
- [117] Martha S Head, James A Given, and Michael K Gilson. “mining minima”: Direct computation of conformational free energy. *J. Phys. Chem. A*, 101(8):1609–1618, February 1997.
- [118] Ray Luo and Michael K Gilson. Synthetic adenine receptors: Direct calculation of binding affinity and entropy. *J. Am. Chem. Soc.*, 122(12):2934–2937, March 2000.
- [119] K L Mardis, R Luo, and M K Gilson. Interpreting trends in the binding of cyclic ureas to HIV-1 protease. *J. Mol. Biol.*, 309(2):507–517, June 2001.
- [120] Wei Chen, Chia-En Chang, and Michael K Gilson. Calculation of cyclodextrin binding affinities: energy, entropy, and implications for drug design. *Biophys. J.*, 87(5):3035–3049, November 2004.
- [121] Sarvin Moghaddam, Cheng Yang, Mikhail Rekharsky, Young Ho Ko, Kimoon Kim, Yoshihisa Inoue, and Michael K Gilson. New ultrahigh affinity host-guest complexes of cucurbit[7]uril with bicyclo[2.2.2]octane and adamantane guests: thermodynamic analysis and evaluation of M2 affinity calculations. *J. Am. Chem. Soc.*, 133(10):3570–3581, March 2011.
- [122] David L Mobley, Alan P Graves, John D Chodera, Andrea C McReynolds, Brian K Shoichet, and Ken A Dill. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, 371(4):1118–1134, August 2007.
- [123] Hyung-June Woo and Benoît Roux. Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 102(19):6825–6830, May 2005.

- [124] Wenxun Gan and Benoît Roux. Binding specificity of SH2 domains: insight from free energy simulations. *Proteins*, 74(4):996–1007, March 2009.
- [125] Nandou Lu, Jayant K Singh, and David A Kofke. Appropriate methods to combine forward and reverse free-energy perturbation averages. *J. Chem. Phys.*, 118(7):2977–2984, February 2003.
- [126] Michael R Shirts and Vijay S Pande. Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.*, 122(14):144107, April 2005.
- [127] Stefan Bruckner and Stefan Boresch. Efficiency of alchemical free energy simulations. i. a practical comparison of the exponential formula, thermodynamic integration, and bennett’s acceptance ratio method. *J. Comput. Chem.*, 32(7):1303–1319, May 2011.
- [128] Stefan Bruckner and Stefan Boresch. Efficiency of alchemical free energy simulations. II. improvements for thermodynamic integration. *J. Comput. Chem.*, 32(7):1320–1333, May 2011.
- [129] Anita de Ruiter, Stefan Boresch, and Chris Oostenbrink. Comparison of thermodynamic integration and bennett acceptance ratio for calculating relative protein-ligand binding free energies. *J. Comput. Chem.*, 34(12):1024–1034, May 2013.
- [130] Thomas C Beutler, Alan E Mark, René C van Schaik, Paul R Gerber, and Wilfred F van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.*, 222(6):529–539, June 1994.
- [131] T Steinbrecher, I S Joung, and others. Soft-core potentials in thermodynamic integration: Comparing one-and two-step transformations. *Journal of computational*, 2011.
- [132] Thomas Steinbrecher, David L Mobley, and David A Case. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J. Chem. Phys.*, 127(21):214108, December 2007.

- [133] Viktor Hornak and Carlos Simmerling. Development of softcore potential functions for overcoming steric barriers in molecular dynamics simulations. *J. Mol. Graph. Model.*, 22(5):405–413, May 2004.
- [134] Timothy J Giese and Darrin M York. A GPU-Accelerated parameter interpolation thermodynamic integration free energy method. *J. Chem. Theory Comput.*, February 2018.
- [135] David L Mobley, John D Chodera, and Ken A Dill. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *J. Chem. Phys.*, 125(8):084902, August 2006.
- [136] B Roux, M Nina, R Pomès, and J C Smith. Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study. *Biophys. J.*, 71(2):670–681, August 1996.
- [137] G Mann and J Hermans. Modeling protein-small molecule interactions: structure and thermodynamics of noble gases binding in a cavity in mutant phage T4 lysozyme L99A. *J. Mol. Biol.*, 302(4):979–989, September 2000.
- [138] Matthew Harger, Daniel Li, Zhi Wang, Kevin Dalby, Louis Lagardère, Jean-Philip Piquemal, Jay Ponder, and Pengyu Ren. Tinker-OpenMM: Absolute and relative alchemical free energies using AMOEBA on GPUs. *J. Comput. Chem.*, 38(23):2047–2055, September 2017.
- [139] Edward King, Ruxi Qi, Han Li, Ray Luo, and Erick Aitchison. Estimating the roles of protonation and electronic polarization in absolute binding affinity simulations. *J. Chem. Theory Comput.*, 17(4):2541–2555, April 2021.
- [140] Yen-Lin Lin, Alexey Aleksandrov, Thomas Simonson, and Benoît Roux. An overview of electrostatic free energy computations for solutions and proteins. *J. Chem. Theory Comput.*, 10(7):2690–2709, July 2014.
- [141] P H Hünenberger and J A McCammon. Effect of artificial periodicity in simulations of biomolecules under ewald boundary conditions: a continuum electrostatics study. *Biophys. Chem.*, 78(1-2):69–88, April 1999.

- [142] Jamshed Anwar and David M Heyes. Robust and accurate method for free-energy calculation of charged molecular systems. *J. Chem. Phys.*, 122(22):224117, June 2005.
- [143] Jochen S Hub, Bert L de Groot, Helmut Grubmüller, and Gerrit Groenhof. Quantifying artifacts in ewald simulations of inhomogeneous systems with a net charge. *J. Chem. Theory Comput.*, 10(1):381–390, January 2014.
- [144] Surjit B Dixit and Christophe Chipot. Can absolute free energies of association be estimated from molecular mechanical simulations? the biotin streptavidin system revisited. *J. Phys. Chem. A*, 105(42):9795–9799, October 2001.
- [145] Jason A Wallace and Jana K Shen. Charge-leveling and proper treatment of long-range electrostatics in all-atom molecular dynamics at constant ph. *J. Chem. Phys.*, 137(18):184105, November 2012.
- [146] Gabriel J Rocklin, David L Mobley, Ken A Dill, and Philippe H Hünenberger. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: an accurate correction scheme for electrostatic finite-size effects. *J. Chem. Phys.*, 139(18):184103, November 2013.
- [147] Tri T Pham and Michael R Shirts. Identifying low variance pathways for free energy calculations of molecular transformations in solution phase. *J. Chem. Phys.*, 135(3):034114, July 2011.
- [148] John D Chodera. A simple method for automated equilibration detection in molecular simulations. *J. Chem. Theory Comput.*, 12(4):1799–1805, April 2016.
- [149] Wei Yang, Ryan Bitetti-Putzer, and Martin Karplus. Free energy simulations: use of reverse cumulative averaging to determine the equilibrated region and the time required for convergence. *J. Chem. Phys.*, 120(6):2618–2628, February 2004.
- [150] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12):124105, September 2008.

- [151] Himanshu Paliwal and Michael R Shirts. A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods. *J. Chem. Theory Comput.*, 7(12):4115–4134, December 2011.
- [152] Haochuan Chen, Julio D C Maia, Brian K Radak, David J Hardy, Wensheng Cai, Christophe Chipot, and Emad Tajkhorshid. Boosting Free-Energy perturbation calculations with GPU-Accelerated NAMD. *J. Chem. Inf. Model.*, 60(11):5301–5307, November 2020.
- [153] Tai-Sung Lee, Zhixiong Lin, Bryce K Allen, Charles Lin, Brian K Radak, Yujun Tao, Hsu-Chun Tsai, Woody Sherman, and Darrin M York. Improved alchemical free energy calculations with optimized smoothstep softcore potentials. *J. Chem. Theory Comput.*, 16(9):5512–5525, September 2020.
- [154] Xibing He, Shuhan Liu, Tai-Sung Lee, Beihong Ji, Viet H Man, Darrin M York, and Junmei Wang. Fast, accurate, and reliable protocols for routine calculations of Protein-Ligand binding affinities in drug design projects using AMBER GPU-TI with ff14SB/GAFF. *ACS Omega*, 5(9):4611–4619, March 2020.
- [155] Haohao Fu, James C Gumbart, Haochuan Chen, Xueguang Shao, Wensheng Cai, and Christophe Chipot. BFEE: A User-Friendly graphical interface facilitating absolute binding Free-Energy calculations. *J. Chem. Inf. Model.*, 58(3):556–560, March 2018.
- [156] Germano Heinzemann and Michael K Gilson. Automation of absolute protein-ligand binding free energy calculations for docking refinement and compound evaluation. *Sci. Rep.*, 11(1):1116, January 2021.
- [157] Seonghoon Kim, Hiraku Oshima, Han Zhang, Nathan R Kern, Suyong Re, Jumin Lee, Benoît Roux, Yuji Sugita, Wei Jiang, and Wonpil Im. CHARMM-GUI free energy calculator for absolute and relative ligand solvation and binding free energy simulations. *J. Chem. Theory Comput.*, 16(11):7207–7218, November 2020.
- [158] Tharindu Senapathi, Miroslav Suruzhon, Christopher B Barnett, Jonathan Essex, and

- Kevin J Naidoo. BRIDGE: An open platform for reproducible High-Throughput free energy simulations. *J. Chem. Inf. Model.*, 60(11):5290–5295, November 2020.
- [159] Maximilian Kuhn, Stuart Firth-Clark, Paolo Tosco, Antonia S J S Mey, Mark Mackey, and Julien Michel. Assessment of binding affinity via alchemical Free-Energy calculations. *J. Chem. Inf. Model.*, 60(6):3120–3130, June 2020.
- [160] Israel Cabeza de Vaca, Yue Qian, Jonah Z Vilseck, Julian Tirado-Rives, and William L Jorgensen. Enhanced monte carlo methods for modeling proteins including computation of absolute free energies of binding. *J. Chem. Theory Comput.*, 14(6):3279–3288, June 2018.
- [161] Yue Qian, Israel Cabeza de Vaca, Jonah Z Vilseck, Daniel J Cole, Julian Tirado-Rives, and William L Jorgensen. Absolute free energy of binding calculations for macrophage migration inhibitory factor in complex with a druglike inhibitor. *J. Phys. Chem. B*, 123(41):8675–8685, October 2019.
- [162] Wei Jiang, Jonathan Thirman, Sunhwan Jo, and Benoît Roux. Reduced free energy Perturbation/Hamiltonian replica exchange molecular dynamics method with unbiased alchemical thermodynamic axis. *J. Phys. Chem. B*, 122(41):9435–9442, October 2018.
- [163] Wei Jiang. Accelerating convergence of free energy computations with hamiltonian simulated annealing of solvent (HSAS). *J. Chem. Theory Comput.*, 15(4):2179–2186, April 2019.
- [164] Hiraku Oshima, Suyong Re, and Yuji Sugita. Prediction of protein-ligand binding pose and affinity using the gREST+FEP method. *J. Chem. Inf. Model.*, 60(11):5382–5394, November 2020.
- [165] David D L Minh. Alchemical grid dock (AlGDock): Binding free energy calculations between flexible ligands and rigid receptors. *J. Comput. Chem.*, 41(7):715–730, March 2020.
- [166] Gerhard König, Nina Glaser, Benjamin Schroeder, Alžbeta Kubincová, Philippe H Hünenberger, and Sereina Riniker. An alternative to conventional λ -intermediate states in alchemical free energy calculations: λ -enveloping distribution sampling. *J. Chem. Inf. Model.*, 60(11):5407–5423, November 2020.

- [167] Zhe Li, Xin Li, Yi-You Huang, Yaoxing Wu, Runduo Liu, Lingli Zhou, Yuxi Lin, Deyan Wu, Lei Zhang, Hao Liu, Ximing Xu, Kunqian Yu, Yuxia Zhang, Jun Cui, Chang-Guo Zhan, Xin Wang, and Hai-Bin Luo. Identify potent SARS-CoV-2 main protease inhibitors via accelerated free energy perturbation-based virtual screening of existing drugs. *Proc. Natl. Acad. Sci. U. S. A.*, 117(44):27381–27387, November 2020.
- [168] Yoshiaki Tanida and Azuma Matsuura. Alchemical free energy calculations via metadynamics: Application to the theophylline-RNA aptamer complex. *J. Comput. Chem.*, 41(20):1804–1819, July 2020.
- [169] Raffaele Fiorentini, Kurt Kremer, and Raffaello Potestio. Ligand-protein interactions in lysozyme investigated through a dual-resolution model. *Proteins*, 88(10):1351–1360, October 2020.
- [170] Yoshitake Sakae, Bin W Zhang, Ronald M Levy, and Nanjie Deng. Absolute protein binding free energy simulations for ligands with multiple poses, a thermodynamic path that avoids exhaustive enumeration of the poses. *J. Comput. Chem.*, 41(1):56–68, January 2020.
- [171] Xinqiang Ding and Bin Zhang. DeepBAR: A fast and exact method for binding free energy computation. *J. Phys. Chem. Lett.*, pages 2509–2515, March 2021.
- [172] Marie L Laury, Zhi Wang, Aaron S Gordon, and Jay W Ponder. Absolute binding free energies for the SAMPL6 cucurbit[8]uril host-guest challenge via the AMOEBA polarizable force field. *J. Comput. Aided Mol. Des.*, 32(10):1087–1095, October 2018.
- [173] Yuanjun Shi, Marie L Laury, Zhi Wang, and Jay W Ponder. AMOEBA binding free energies for the SAMPL7 TrimerTrip host-guest challenge. *J. Comput. Aided Mol. Des.*, November 2020.
- [174] Martin Amezcua, Léa El Khoury, and David L Mobley. SAMPL7 Host-Guest challenge overview: assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. *J. Comput. Aided Mol. Des.*, 35(1):1–35, January 2021.

- [175] Christoph Öhlknecht, Jan Walther Perthold, Bettina Lier, and Chris Oostenbrink. Charge-Changing perturbations and path sampling via classical molecular dynamic simulations of simple Guest–Host systems. *J. Chem. Theory Comput.*, 16(12):7721–7734, December 2020.
- [176] Christoph Öhlknecht, Bettina Lier, Drazen Petrov, Julian Fuchs, and Chris Oostenbrink. Correcting electrostatic artifacts due to net-charge changes in the calculation of ligand binding free energies. *J. Comput. Chem.*, 41(10):986–999, April 2020.
- [177] Toru Ekimoto, Tsutomu Yamane, and Mitsunori Ikeguchi. Elimination of Finite-Size effects on binding free energies via the Warp-Drive method. *J. Chem. Theory Comput.*, 14(12):6544–6559, December 2018.
- [178] Jeffrey Cruz, Lauren Wickstrom, Danzhou Yang, Emilio Gallicchio, and Nanjie Deng. Combining alchemical transformation with a physical pathway to accelerate absolute binding free energy calculations of charged ligands to enclosed binding sites. *J. Chem. Theory Comput.*, 16(4):2803–2813, April 2020.
- [179] Yaozong Li and Kwangho Nam. Repulsive Soft-Core potentials for efficient alchemical free energy calculations. *J. Chem. Theory Comput.*, 16(8):4776–4789, August 2020.
- [180] Andrea Rizzi, Travis Jensen, David R Slochower, Matteo Aldeghi, Vytautas Gapsys, Dimitris Ntekoumes, Stefano Bosisio, Michail Papadourakis, Niel M Henriksen, Bert L de Groot, Zoe Cournia, Alex Dickson, Julien Michel, Michael K Gilson, Michael R Shirts, David L Mobley, and John D Chodera. The SAMPL6 SAMPLing challenge: Assessing the reliability and efficiency of binding free energy calculations. October 2019.
- [181] Kyungreem Han, Phillip S Hudson, Michael R Jones, Naohiro Nishikawa, Florentina Tofoleanu, and Bernard R Brooks. Prediction of CB[8] host-guest binding free energies in SAMPL6 using the double-decoupling method. *J. Comput. Aided Mol. Des.*, 32(10):1059–1073, October 2018.
- [182] Naohiro Nishikawa, Kyungreem Han, Xiongwu Wu, Florentina Tofoleanu, and Bernard R Brooks. Comparison of the umbrella sampling and the double decoupling method in binding

- free energy predictions for SAMPL6 octa-acid host-guest challenges. *J. Comput. Aided Mol. Des.*, 32(10):1075–1086, October 2018.
- [183] Zhe Huai, Huaiyu Yang, Xiao Li, and Zhaoxi Sun. SAMPL7 TrimerTrip host-guest binding affinities from extensive alchemical and end-point free energy calculations. *J. Comput. Aided Mol. Des.*, 35(1):117–129, January 2021.
- [184] Yuriy Khalak, Gary Tresadern, Bert L de Groot, and Vytautas Gapsys. Non-equilibrium approach for binding free energies in cyclodextrins in SAMPL7: force fields and software. *J. Comput. Aided Mol. Des.*, 35(1):49–61, January 2021.
- [185] Nanjie Deng, Di Cui, Bin W Zhang, Junchao Xia, Jeffrey Cruz, and Ronald Levy. Comparing alchemical and physical pathway methods for computing the absolute binding free energy of charged ligands. *Phys. Chem. Chem. Phys.*, 20(25):17081–17092, June 2018.
- [186] Hannes H Loeffler, Stefano Bosisio, Guilherme Duarte Ramos Matos, Donghyuk Suh, Benoit Roux, David L Mobley, and Julien Michel. Reproducibility of free energy calculations across different molecular simulation software packages. *J. Chem. Theory Comput.*, 14(11):5567–5582, November 2018.
- [187] S J Zasada, D W Wright, and P V Coveney. Large-scale binding affinity calculations on commodity compute clouds. *Interface Focus*, 10(6):20190133, December 2020.
- [188] Junchao Xia, William Flynn, Emilio Gallicchio, Keith Uplinger, Jonathan D Armstrong, Stefano Forli, Arthur J Olson, and Ronald M Levy. Massive-scale binding free energy simulations of HIV integrase complexes using asynchronous replica exchange framework implemented on the IBM WCG distributed network. *J. Chem. Inf. Model.*, 59(4):1382–1397, April 2019.

Chapter 2

Estimating the roles of protonation and electronic polarization in absolute binding affinity simulations

Authors: Edward King, Ruxi Qi, Han Li, Ray Luo, Erick Aitchison

J Chem Theory Comput. 2021;17: 2541–2555.

doi: [10.1021/acs.jctc.0c01305](https://doi.org/10.1021/acs.jctc.0c01305)

Publication Date (Web): March 25, 2021

2.1 Abstract

Accurate prediction of binding free energies is critical to streamlining the drug development and protein design process. With the advent of GPU acceleration, absolute alchemical methods, which simulate the removal of ligand electrostatics and van der Waals interactions with the protein, have become routinely accessible and provide a physically rigorous approach that enables full consideration of flexibility and solvent interaction. However, standard explicit solvent simulations are

unable to model protonation or electronic polarization changes upon ligand transfer from water to the protein interior, leading to inaccurate prediction of binding affinities for charged molecules. Here, we perform extensive simulation totaling $\sim 540 \mu\text{s}$ to benchmark the impact of modeling conditions on predictive accuracy for absolute alchemical simulations. Binding to urokinase plasminogen activator (UPA), a protein frequently overexpressed in metastatic tumors, is evaluated for a set of 10 inhibitors with extended flexibility, highly charged character, and titratable properties. We demonstrate that the alchemical simulations can be adapted to utilize the MBAR/PBSA method to improve the accuracy upon incorporating electronic polarization, highlighting the importance of polarization in alchemical simulations of binding affinities. Comparison of binding energy prediction at various protonation states indicates that proper electrostatic setup is also crucial in binding affinity prediction of charged systems, prompting us to propose an alternative binding mode with protonated ligand phenol and Hid-46 at the binding site, a testable hypothesis for future experimental validation.

2.2 Introduction

Electrostatics and polarization effects are critical to the study of biomolecular processes such as dynamics, recognition, and enzymatic catalysis. The success of computational simulation in sampling physiologically apt biomolecular structures involved in enzyme activity is dependent on both efficient calculations, to enable consideration of atomic interactions at long time scales, and accurate treatment of those interactions to maximize predictive capability. Current simulation efforts often ignore the impact of electronic polarization due to their complexity and high computational costs, leading to errors such as the overestimation of gas-phase water dimer interaction energy by greater than 30% with the nonpolarizable TIP5P model[1, 2]. The standard nonpolarizable point-charge model allows analysis of electrostatics through straightforward application of the Coulombic potential but is unable to capture the effect of exposure to different electrostatic environments such as between the protein interior and solvent that is essential to biomolecular processes. Furthermore, reference parameters for nonpolarizable models are typically derived from gas-phase quantum me-

chanical calculations, resulting in spurious “pre-polarization” when used in an aqueous environment due to the inclusion of average bulk polarization effects inconsistent with the liquid phase. Improving the treatment of electrostatics and polarization would significantly enhance efforts to study the biomolecular processes of ion-dependent interactions, proton and electron transfer in enzyme catalysis, order–disorder transitions in intrinsically disordered regions, pKa effects in titration, etc.

A number of polarizable models have been developed to address the accurate representation of electrostatic interactions for biomolecular simulation including the OPLS-AA fluctuating charge model[3, 4], Drude oscillator[5–7] with CHARMM, and AMOEBA with multipole expansion and increased force field components[8–10]. Recent developments with AMBER include the polarizable Gaussian Multipole (pGM)[11, 12] model that improves over the previous induced dipole implementation based on Thole models[13–17]. pGM represents each atom’s multipole as a single Gaussian function or its derivatives, speeding electrostatic calculations over alternative Gaussian-based models. By screening short-range interactions in a physically consistent manner, pGM enables the stable charge-fitting necessary to describe molecular anisotropy that is difficult to achieve with Thole models[11, 12].

Regardless of which model to use, an important application of molecular simulations is the accurate prediction of binding affinities to accelerate the drug discovery process as recently reviewed[18]. Accurate virtual screening is necessary to reduce the excessive time and costs associated with drug development, which are estimated to be over 10 years and \$2.8 billion for an approved drug[19]. Methods based on geometric docking to optimize the shape and electrostatic complementarity between binding partners[20–24], end-point MD simulations with either the linear interaction energy method[25], or the Molecular Mechanics Poisson–Boltzmann Surface Area method[12, 16, 26–33], end-point MC simulations with the Mining Minima method[34–36], alchemical pathway simulations with full sampling of conformational flexibility in explicit solvent[21, 37–43], and machine learning based on correlation of structural features and protein–ligand interactions[44–46] have shown promise, but have not achieved the generalizable accuracy required or come at a too high computational cost for practical application to drug discovery.

Alchemical simulations measure the free energy difference between two states, so that it can be used to determine the free energy change between the complex state with protein and ligand bound and the unbound state with protein and ligand separated[47]. Alchemical simulations progress through a closed thermodynamic cycle, utilizing transformations through unphysical intermediate states modeling the gradual decoupling of ligand electrostatic and van der Waals (VDW) interactions with the protein environment, and provide a computational advantage over brute-force simulations of unbinding or binding processes[48]. Previous work has highlighted the utility of alchemical simulations in the computation of small molecule distribution coefficients between solvent phases[49], protein stability upon amino acid mutation[50], binding affinity through relative transformation growing or deleting functional groups off a reference structure[51–53], and absolute transformation where the larger perturbation of ligand transfer to gas phase is modeled[54–58]. Absolute alchemical transformations, which permit direct prediction of binding energy and do not require initialization from a reference structure with high similarity to the target as relative calculations, have only recently become practical with the development of high-performance computer hardware, such as graphical processing units (GPUs).

Structure-based drug design coupled to alchemical simulations has served as the foundation for drug development campaigns[54]; however, limitations due to heterogeneity in protocols and model setups, limited accuracies in molecular force fields, and insufficient sampling of the protein and ligand conformations still impede prediction accuracy. Furthermore, standard alchemical simulations are unable to model protonation or electronic polarization changes upon ligand transfer from water to the protein interior, leading to inaccurate prediction of binding affinities for charged molecules. In this study, we benchmarked the absolute alchemical transformation methods on the urokinase plasminogen activator (UPA) system to estimate the impacts of protonation and closely related polarization effects during the protein–ligand binding process. UPA is a serine protease that activates plasmin which is involved in the degradation of blood clots and extracellular matrix[59]. UPA has been found to be overexpressed in several types of metastatic tumors; this upregulation has been proposed to drive the tissue degradation required for cancer invasion and metastatic growth, making UPA a desirable target for anticancer therapeutics. The tested models are a set of high-resolution

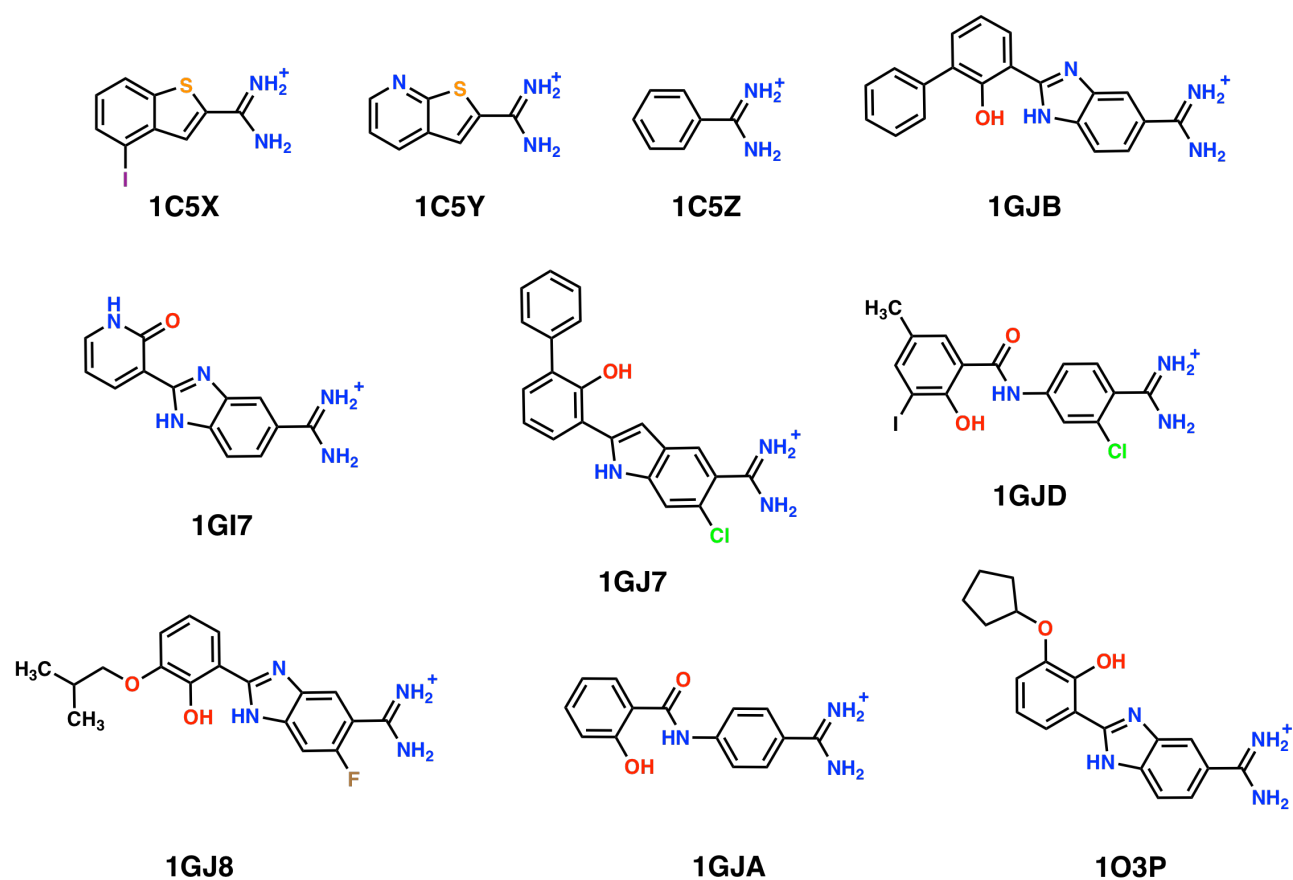


Figure 2.1: Chemical structures of the 10 evaluated UPA inhibitors. The molecules share a benzamidine-like scaffold with characteristic amidine group carrying positive charge, and extended tails comprised of a phenol group and other functional modifiers. The hydroxyl on the phenol is proposed to be titratable and samples deprotonated and protonated states during binding, altering the hydrogen bonding capability of the ligands. The inhibitors are categorized as small (those without the phenol group)—1C5X, 1C5Y, 1C5Z, and 1GI7—and big (for those with potentially charged phenols)—1GJ7, 1GJ8, 1GJA, 1GJB, 1GJD, 1O3P.

crystal structures collected by Katz et al. with 10 different competitive inhibitors of varying sizes, charges, and chemical groups[60–62] (Figure: 2.1). The inhibition constant (K_i) of each ligand has been experimentally determined, allowing for the validation of our computational protocols. This set of ligands represents a diverse and challenging test case with inhibitors bearing a large number of torsion angles that require lengthy simulation to sample the available conformation space, highly charged character amplifying inaccuracy in the treatment of electronic polarization, and multiple potential protonation states due to ionizability and tautomerization.

To address the heterogeneity in the alchemical protocol, we studied the effects of various simulation

setups/conditions including ligand force field choices, salt concentrations, alternative protonation states, and ligand restraints through a single harmonic distance (1DOF) or with the more involved 6 degrees of freedom (6DOF) restraints to improve convergence by maintaining the ligand in the binding pose. We further developed a new strategy to utilize the PBSA continuum solvent model coupled with the Multistate Bennet Acceptance Ratio (MBAR) approach to estimate the effect of electronic polarization in this challenging set of highly charged ligands. We demonstrate that the application of the MBAR/PBSA method with optimized solute dielectric constant permits more properly modeled electronic polarization, leading to superior accuracy in the absolute binding free energy prediction for this set of highly charged ligands. This allows us to assess alternative protonation states for the ligands and titratable residues in the binding pose, offering a testable hypothesis for future experimental validation.

2.3 Methods

2.3.1 Structure preparation for molecular simulations

Crystal structures for the inhibitor bound urokinase plasminogen activator (PDB: 1C5X, 1C5Y, 1C5Z, 1GI7, 1GJ7, 1GJ8, 1GJA, 1GJB, 1GJD, and 1O3P)[60–62] were obtained from the RCSB PDB database[63]. Experimentally determined binding free energies were obtained from the PDB-bind database[64]. The structures were prepared for simulation by removal of all water molecules greater than 5 Å away from the active site, removal of all cocrystallized ligands that were not the target inhibitor, and truncation of all structures to 245 amino acids by deletion of disordered C-terminal residues that were not resolved in all crystal structures (a maximum of 3 residues were deleted). Disulfide bonds were added as in the crystal structures.

As there are no pKa measurements available, protonation states of titratable residues were determined at pH 7.4 through the H++ Web server[65] except those in the binding pocket, where the complex crystal structures were used to infer the likely protonation states. In the binding pocket,

His-94 and Asp-97 are modeled as default neutral HIE and charged ASP, respectively, as there is no unusual polar interaction with the ligand molecule. For His-46 and the ligand phenol group, there are two possible protonation states that satisfy the steric constraint in the crystal structures as discussed in detail in Results and Discussion. The first possibility is to set His-46 as protonated HIP and the ligand phenol as deprotonated as suggested in ref [62]. The second possibility is to set His-46 as deprotonated HID and the ligand phenol as protonated. Given these protonation states, 1C5X, 1C5Y, 1C5Z, and 1GI7 are treated as +1 net charge due to protonation at the amidine group, and all other ligands are treated as +0 net charge zwitterions with a +1 charge of the amidine group and -1 charge on the deprotonated phenol hydroxyl or as +1 net charge ions with 0 charge on the protonated phenol hydroxyl. The predictive accuracy of both protonation states was compared to a baseline model with all ligands treated as +1 net charge due to default protonation at the amidine and phenol hydroxyl with Hip-46. All tested conditions and protonation states are summarized in Table: 2.1.

Condition	HIS46	Protonated Ligands (+1 charge)	Deprotonated Ligands (+0 charge)	Salt	Restraint Potential
Baseline	HIP	1C5X, 1C5Y, 1C5Z, 1GI7, 1GJ7, 1GJ8, 1GJA, 1GJB, 1GJD, 1O3P	-	Counter-ions only	1DOF
All-HIP	HIP	1C5X, 1C5Y, 1C5Z, 1GI7	1GJ7, 1GJ8, 1GJA, 1GJB, 1GJD, 1O3P	150 mM	1DOF/6DOF

All-HID	HID	1C5X, 1C5Y, 1C5Z, 1GI7, 1GJ7, 1GJ8, 1GJA, 1GJB, 1GJD, 1O3P	-		150 mM	1DOF
Small-HIP	Mixed	1C5X, 1C5Y, 1C5Z, 1GI7 (HIP)	-		150 mM	1DOF
		1GJ7, 1GJ8, 1GJA, 1GJB, 1GJD, 1O3P (HID)				
Small-HID	Mixed	1C5X, 1C5Y, 1C5Z, 1GI7 (HID)	1GJ7, 1GJ8, 1GJA, 1GJB, 1GJD, 1O3P (HIP)		150 mM	1DOF

Table 2.1: Summary of simulation conditions. The baseline corresponds to a default setup with full ligand and protein protonation, salt concentration at charge neutralizing amount, and 1DOF restraint. Singular condition changes to the baseline: 150 mM salt concentration, and deprotonated ligand phenol. Alternative protonation states are tested with variable ionization at the ligand phenol and His-46 to model the effect of hydrogen bonding potential on binding free energy prediction.

Ligand partial charges were determined with the Restrainted Electrostatic Potential (RESP) method[66] at the HF/6-31G* level using Gaussian09[67], except HF/CEP-31G was used for ligands with iodine. Other ligand parameters were taken from the General Amber Force Field (GAFF)[68] or GAFF2. The protein was modeled with the ff14sb[69] force field. Systems were solvated in TIP3P[70] water, in a truncated octahedron with 10 Å buffer, and charge neutralized with Na⁺/Cl⁻ ions. Additional ions were also added to reach 150 mM salt concentration under the high salt condition tested.

Molecular dynamics simulations were performed with pmemd.cuda[71] from the Amber18 package with an 8 Å Particle Mesh Ewald[72] cutoff and otherwise default settings.

2.3.2 Alchemical simulation protocol

Computation of binding free energies was conducted through a four-step process: equilibration, restraint sampling, decharging, and softcore van der Waals removal[73, 74]. Imposition of restraints and each of the two inhibitor transformation steps (decharging and VDW removal) proceeded through a series of alchemical intermediates described by the coupling parameter lambda increasing from 0 (starting state) to 1 (fully transformed ending state). Final simulation data are an aggregate of ensemble MD of five independent replicates started from the minimized crystal structures with randomized initial velocities. The free energy differences between states were calculated with MBAR[75, 76] through the pymbar[75] package and required the calculation of energy cross-terms for each trajectory at each restraint, charge, and VDW lambda step. Only data produced from frames in the last half of each trajectory were included in energy calculations to ensure well-equilibrated results.

2.3.3 Minimization and equilibration

The UPA systems were minimized in two steps: first with 2,500 steps of steepest descent and 2,500 steps of conjugate gradient where all non-hydrogen solute atoms were restrained with a 20 kcal mol⁻¹ Å⁻² force to relieve steric clash. The second minimization to remove solute steric clashes was run with the same cycle settings and restraints removed. Heating from 0 to 298 K was performed over 0.5 ns with 10 kcal mol⁻¹ Å⁻² restraints on all non-hydrogen solute atoms. Solvent density equilibration under the NPT condition and the Langevin thermostat with a collision frequency of 2 ps⁻¹ was carried out over 0.4 ns with 2 kcal mol⁻¹ Å⁻² restraints on all non-hydrogen solute atoms to stably reach 1 atm of pressure. Next, an unrestrained 100 ns NVT equilibration with the Langevin thermostat and collision frequency 1 ps⁻¹ was completed to clear remaining structural

artifacts from the initial crystal structure. Separate simulations for the unrestrained inhibitor alone and the protein–inhibitor complex were run for the decharging and VDW removal process. The inhibitor alone was extracted from the equilibrated complex and solvated in the TIP3P truncated octahedron box with 20 Å buffer and neutralized with Na⁺/Cl⁻ counterions or an up to 150 mM salt concentration. Trajectory data was analyzed with the cpptraj program[77] and the NumPy[78] packages.

2.3.4 Imposing restraints

As electrostatic and VDW interactions are decoupled, the ligand has the ability to escape the active site and sample states irrelevant to binding, hindering convergence. Standard practice is to apply a restraint on the ligand which requires calculating the free energy contribution of the restraints,

$$\Delta A_r = -kT \ln \frac{Z_P Z_L}{Z_{CL}} \tag{2.1}$$

where Z_P , Z_L , and Z_{CL} are the configurational partition functions of the protein, ligand, and the cross-linked state[34, 79–82]. The derivation of the restraint free energy depends on the external degrees of freedom restrained on the ligand relative to the protein, which defines the cross-linked state or virtual bond. Since the ligand position and/or orientation is restrained to the protein, the protein external degrees of freedom can be separated out leaving the integration of the internal and external degrees of freedom for the ligand. The restraint free energy can be simplified into the difference between the term from the integration of all external degrees of freedom of a nonlinear ligand, $8\pi^2 V$, and the term of a Gaussian integral for each degree of freedom used to restrain the ligand, $\sqrt{\frac{2\pi k_b T}{K\xi}}$, where $K\xi$ is the harmonic restraint force constant.

In the 1DOF restraint, a single harmonic distance restraint with a 20 kcal mol⁻¹ Å⁻² force constant was utilized as a virtual bond between the Asp-192 α-carbon and the ligand amidine carbon. The final analytical correction for the single distance restraint is as follows for restraining the ligand in

the unbound state[34, 81],

$$-k_bT \ln \left[\frac{8\pi^2 V^0 K_r^{1/2}}{(2\pi k_b T)^{1/2}} \right] \quad (2.2)$$

where K_r is the force constant of the distance restraint and V^0 is the standard state volume (SI Figure: 2.7). This virtual bond restraint is relative to the protein. This is different from the Cartesian position restraint from Roux et al.[79], which uses a point in three-dimensional Cartesian space to restrain the ligand, resulting in an integral of $\left(\frac{2\pi k_b T}{K\xi}\right)^{2/3}$.

To study the effects of the restraining protocol, an independent set of simulations was also run with the set of 6 degrees of freedom (6DOF) orientational restraints proposed by Boresch et al.[81] based on a single distance, two angular, and three dihedral parameters, all with 10 kcal mol⁻¹ Å⁻² force constants. For the 6DOF restraint, the final analytical correction for restraining the ligand in the unbound state is[81],

$$-k_bT \ln \left[\frac{8\pi^2 V^0 (K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C})^{1/2}}{r_{a,A,0}^2 \sin \theta_{A,0} \sin \theta_{B,0} (2\pi k_b T)^3} \right] \quad (2.3)$$

where $r_{a,A,0}$ is the restrained distance, $\theta_{A,0}$ and $\theta_{B,0}$ are the two restrained angles, and K 's are the force constants (SI Figure: 2.7).

All restraint bounds were selected based on the final positions of the ligands at the end of the equilibration stage. Restraint sampling from off to full strength was performed over 6 equally spaced lambda values (0, 0.2, 0.4, 0.6, 0.8, 1.0), each with 10 ns. A separate analytical correction is calculated to determine the penalty for restraining the ligand in the unbound state.

2.3.5 Alchemical simulation parameters

Decharging through parameter-interpolation of the inhibitors' partial charges to the sampled lambda window was performed to gradually decouple all electrostatic interactions between the inhibitor and environment and was separately run prior to VDW removal to avert the possibility of attractive atom overlap singularities. Decharging for both ligands alone and complex was performed linearly over 11 equally spaced lambda values (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0), each for 40 ns with full restraints. System neutrality was maintained with charged ligands by simultaneously decharging a counterion alongside the ligand. Energies from lambda dependent VDW removal were calculated with the softcore potential to avoid numerical instability at end point lambdas observed with linear scaling due to atomic overlap[73, 74]. VDW removal was completed over 16 lambda values (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.0) each for 20 ns, with denser sampling of lambda values at the later stages to more smoothly decouple VDW interactions. The dummy counterions present with charged inhibitor systems are VDW decoupled concurrently with the ligand. All free energy simulations were conducted with the pmemd.GTI[40] program in Amber18.

Alchemical simulation results for each ligand were aggregated from five individual trajectories with randomized starting velocities to ensure robust conformational sampling. Energy values from the last half of each lambda window for the replicates are concatenated together to combine equilibrated data for final MBAR analyses. Examination of convergence involves calculating the difference in final free energy with the addition of each replicate trajectory; the analysis shows that cumulative free energies with five replicates leads to less than 0.5 kcal/mol deviations. Achieving reasonable convergence in absolute binding affinities for the systems studied here was not trivial, and the total MD simulation time including equilibration, restraint sampling, decharging, and VDW removal was 1.7 μ s for a single sample. The total cumulative MD simulation time including all tested conditions and replicates was \sim 540 μ s. Raw traces of the changes in free energy with each lambda window illustrate the linearity of the decharging process, and the high variation of the VDW removal process in the simulation of the complex (SI Figures: 2.8, 2.9, 2.10, 2.11).

2.3.6 Estimation of electronic polarization with MBAR/PBSA

The Poisson–Boltzmann Surface Area (PBSA) method differs from the standard MD approach in that solvent molecules are modeled implicitly as a continuum in a mean-field manner rather than as explicit molecules, which offers significant simulation efficiency[83–104]. PBSA coupled with the MBAR protocol (i.e., MBAR/PBSA) was developed as an alternative to computing the decharging free energy for alchemical simulations in explicit solvent[105]. The original explicit solvent trajectories for all lambda windows used for decharging were prepared by stripping the waters and ions, except for the counterion used to maintain the ligand charge neutrality.

MBAR/PBSA energy evaluation was performed via the linear Poisson–Boltzmann (LPB) method with the Amber18 sander module[106] by postprocessing solvent-stripped snapshots from alchemical simulations. Nonpolar solvation free energies[107] were turned off, as only electrostatic interactions with and without polarization were compared. Following calculation of electrostatic free energies from the individual snapshots, the MBAR method was used to determine the composite free energy change for the complete decharging process as in the explicit solvent model. The PBSA parameters were set to 0.5 Å grid spacing with different interior dielectric constants ranging from the default of 1 to 2 and solvent dielectric constant 80. Periodic boundary conditions were used, and the box size was set to twice the size of the complex dimension or four times the size of the ligand dimension. The incomplete Cholesky conjugate gradient numerical LPB solver was utilized, and the iteration convergence criterion was set as 10^{-3} [99, 103, 108]. Atomic radii were based on the default mbondi parameters in the Amber package[106]. The solvent probe radius was set to the default 1.4 Å and the mobile ion probe radius for the ion accessible surface was also set to the default 2.0 Å. The short-range pairwise charge-based interactions were cutoff at 7 Å, and long-range interactions were calculated from the LPB numerical solution[109]. Ionic strength was set to match the value from the explicit solvent MD simulations.

The solvation free energies computed from the PBSA model are critically dependent on the atomic radii. The Amber default mbondi radii parameters are revised from the Bondi radius set, and do not reproduce the solvation free energies with the TIP3P water as used in this study. Thus,

the binding free energies from the explicit solvent trajectories were first utilized to calibrate the PBSA model through scaling of the ligand and protein radii at solute dielectric constant 1 to match the explicit solvent simulations as previously developed for free energy simulations of ionic systems[105]. First, ligand radii were uniformly scaled by the “Radiscale” PBSA input value and were tuned to minimize the absolute deviation between PBSA and explicit-solvent electrostatic free energies for the ligand alchemical simulations. Next given optimized “Radiscale”, the protein radii were then uniformly scaled by the “Protscale” input value and were tuned to minimize the absolute deviation between PBSA and explicit-solvent electrostatic free energies for the complex alchemical simulations. Following calibration of the atomic radii, the PBSA model can be appropriately utilized for the investigation of electronic polarization by varying the solute dielectric constant.

2.4 Results and discussion

2.4.1 Structural agreement between simulation and experiment

Errors or deficiencies in sampling experimentally relevant conformations are attributed to standard MD protocols/force fields and highlight sources for inaccuracy in the downstream alchemical process that is sensitive to sampled conformations. Thus, we first analyzed the effects of force field choices on the quality of sampled conformations of UPA inhibitors prior to alchemical simulations. Here GAFF and GAFF2 were both studied.

The 10 cocrystallized inhibitors share a common amidine group attached to an aromatic ring (Figure: 2.1). The larger ligands maintain the benzamidine scaffold of the small ligands linked to a phenol-like ring and additional functional groups including methyl and cyclic structures, including 1GJ7, 1GJ8, 1GJA, 1GJB, 1GJD, 1O3P (termed big ligands below). The rest of the ligands, 1C5X, 1C5Y, 1C5Z, and 1GI7, are categorized as small ligands (1GI7 is large in size but lacks the characteristic phenol-like ring of the larger ligands and so is grouped here).

Binding is mediated by two sets of polar interactions. One is from the positively charged amidine

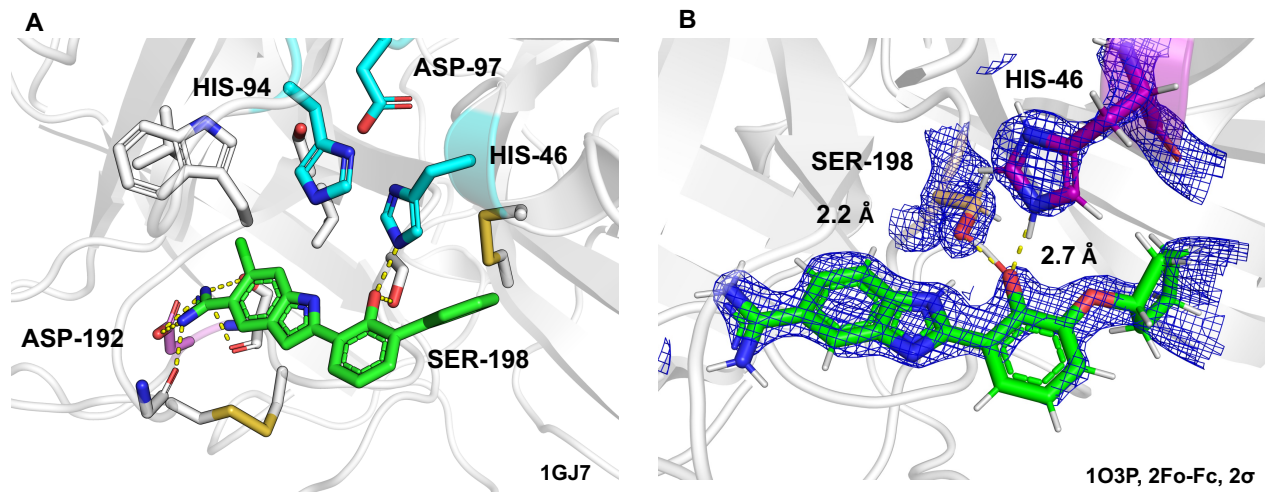


Figure 2.2: Example inhibitor binding poses. **(A)** The protein and ligand form a network of polar interactions at two locations, at the base of the active site between the negatively charged Asp-192 and the positively charged amidine, and near the phenol hydroxyl with Ser-198 and His-46. **(B)** Electron density supports the positioning of the ligand hydroxyl unusually close to Ser-198. An exceptionally short hydrogen bond is formed between the phenol hydroxyl and Ser-198 hydroxyl with a distance of ~ 2.2 Å; this interaction may not be captured accordingly with typical force fields due to van der Waals repulsion.

group, which is common among these inhibitors and makes a dense network of stabilizing polar contacts to a buried Asp-192 and Ser-193 in the active site (Figure: 2.2A). The phenol hydroxyl makes an additional group of hydrogen bonds centered on Ser-198; worth noting is its short hydrogen bond to Ser-198 with a distance ~ 2.2 Å, the lower bound of a hydrogen bond length[62] (Figure: 2.2B). Due to the short distance, the phenol hydroxyl is inferred to act as an acid and be deprotonated in the bound state, and His-46 is interpreted to be a fully protonated Hip-46 to function as a hydrogen bond donor for the ligand phenol[62]. Interestingly, the hydrogen bond between the phenol hydroxyl and His-46 is longer at ~ 2.7 Å even if donor and acceptor are both charged when inferred this way. An alternative solution that satisfies the similar steric constraint in the crystal structure is for the protonated phenol hydroxyl to form the hydrogen bond with the Hid-46. In doing so, both groups are neutral. It should be pointed out that there is no direct pKa measurement of these residues/functional groups. Several of the inhibitors contain halogens (1C5X, 1GJ7, 1GJD, 1GJ8), which are not parametrized comprehensively in current force fields, possibly leading to inaccurate treatment of these ligands.

2.4.2 Crystal structure analysis

The stability of binding sites and ligands in these crystal structures are evaluated through B-factor and electron density analyses. The binding pocket is defined to include any residue with atoms within 6 Å of the inhibitor (SI Figure: 2.12). Crystal B-factors describing flexibility are not directly comparable between structures since they are a function of the crystalline disorder and resolution. Thus, the B-factors are normalized within each structure and Z-scores are compared (SI Figure: 2.13). The binding pockets exhibit roughly equivalent stability with median B-factor Z-scores around -0.6, and the ligands mostly fall into the range of -0.5 to 0. The most stable ligand is 1GJB, possibly due to hydrophobic packing of the highly nonpolar and compact benzene tail, and the most flexible is 1GI7, which is large but lacks the phenol hydroxyl that enables the hydrogen bonding array at Ser-198, and instead has the hydroxyl pointing out toward solvent. Visualization of the density maps supports the close contact between Ser-198 and the phenol hydroxyl (Figure: 2.2). It was noted that atoms involved in the interaction were ignored during structure refinement due to incompatibility of the short hydrogen bonds with the force field used during refinement[62].

2.4.3 Effect of force field choices on ligand binding modes

The positions of the inhibitors in the equilibrated models were first compared with those in the crystal structures (SI Figure: 2.14). It is clear that the inhibitors move further into the active site and assume binding poses with phenol turned slightly outward. The distribution of distances sampled between the ligand phenol and Ser-198 shows that the ligands move further away to relieve the steric clash with both tested force fields, and largely maintain the hydrogen bond except for 1GJD (Figure: 2.3A). The average distances are still within hydrogen bonding range even though there is sampling of unbonded conformations. 1GJD diverges due to rotation of the phenol group; full rotation causes the hydroxyl to point outward toward solvent and the original hydrogen bond is replaced by interaction with the carbonyl oxygen that links the phenol ring to the benzamidine scaffold (Figure: 2.3B). This alternative binding pose is observed with a higher frequency with GAFF2.

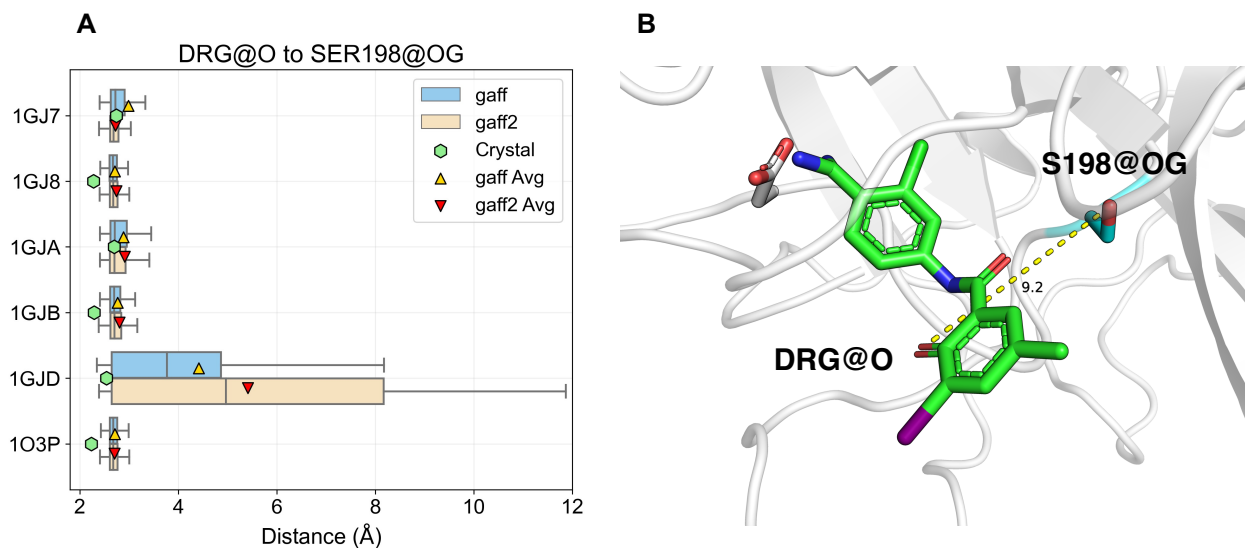


Figure 2.3: Relieving steric clash between the ligand phenol and Ser-198. **(A)** The distance between the ligand phenol oxygen and Ser-198 hydroxyl oxygen is recorded over the last 10 ns of equilibration to analyze sampled conformations and compared to the distance observed in the crystal structures. The trend observed is identical for both GAFF and GAFF2 force fields, all ligands except 1GJD twist away due to repulsive steric interactions but remain in hydrogen bonding range. 1GJD samples broad distances, indicating the initial hydrogen bond is detached. **(B)** Sample frame from the 1GJD simulation illustrates that the phenol hydroxyl rotates outward away from the protein, and the starting hydrogen bond is replaced with one between the peptide bond-like carbonyl and Ser-198. The inhibitor is colored green and labeled DRG.

Next, evaluation of the time evolution of backbone α -carbon RMSD to crystal, binding pocket RMSD, ligand heavy atom RMSD, and distance from Asp-192 $C\gamma$ to ligand amidine is performed (SI Figures: 2.15, 2.16, 2.17). These values are discretized into 10 ns bins and averaged together from the five replicate trajectories. 1GJD stands out with a 0.75 Å binding pocket RMSD with GAFF2 compared to a 0.57 Å RMSD with GAFF at the end of equilibration. This major rearrangement is an indication that an alternative binding pose is sampled and agrees with the phenol distance data, showing substantial rotation of the phenol ring. Ligand heavy atom RMSD shows no difference between GAFF and GAFF2. The small set of ligands with fewer torsions cluster together with a low RMSD, while the highest RMSD values are observed with 1GJ8 and 1GJD. 1GJD is explained by the phenol rotation. For 1GJ8 the ligand moves away from the crystal pose by sliding more deeply into the binding pocket. The movement into the binding pocket is also observed to a lesser degree with 1O3P, 1GJB, 1GI7, and 1GJA. With both GAFF and GAFF2, the favorable polar interactions between the negatively charged Asp-192 and positively charged amidine draw the ligands into the binding pocket, signaling overestimation of electrostatic interactions that is typical of point charge models. In summary, it is clear that a large discrepancy between the crystal structure and equilibrium binding pose is observed with 1GJD and to a lesser extent with 1GJ8, while the remaining models show close agreement, suggesting that the current force field treatment of 1GJD may not sufficiently characterize the important binding interactions observed in the crystal structure.

2.4.4 Benchmarking the effects of simulation conditions on predictive accuracy

We analyzed a range of factors including salt concentration, alternative ligand protonation states, and restraint potential that are known to impact alchemical simulation accuracy. These elements play critical roles in highly charged ligand binding interactions, and their effects on predictive accuracy have not been thoroughly characterized in absolute alchemical simulations. Salt concentration plays a role in screening the strength of electrostatic interactions, yet consideration of physiologi-

cally relevant salt conditions is often ignored, and counterions are generally added only up to the amount necessary to neutralize charge to prevent artifacts arising from periodic boundary conditions. In standard MD simulations, the protonation states of the ligands are fixed and potential changes due to tautomerization or pKa shifts from differences in the solvent and protein environments are not accounted for, which leads to inaccuracy when considering ligands that undergo protonation changes during the binding process. Finally, two types of restraint potentials, 1DOF and 6DOF, have been utilized to prevent the ligand from drifting out of the active site as binding interactions are decoupled; the purpose of these restraints is to focus conformational sampling on configurations most relevant to the binding pose and aid convergence.

2.4.5 Default setup leads to no correlation with experiment

The benchmarks begin with a baseline binding free energy prediction to determine the accuracy with a simple and widely accepted model setup, based on a single harmonic distance restraint between the C α on Asp-192 to the amidine carbon (i.e., 1DOF), with counterions added only to the amount to neutralize the system charge, and full ligand protonation without consideration of the experimental data. The Root Mean Square Error (RMSE) for the baseline prediction is 3.2 kcal/mol, and the Pearson correlation coefficient is -0.15, indicating no linear correlation between the experimentally determined binding affinities and those predicted from simulation (Figure: 2.4A). The ligands with 0 net charge form a cluster of samples with underestimated binding free energies, while the charged ligands are predicted to have overestimated free energies indicating excessively favorable binding.

2.4.6 Use of salt and consistent protonation state improves predicted affinities

Many automated setups neglect setting simulation parameters to match the physiologically relevant conditions, either due to lack of information or to simplify the protocol for higher computation throughput. One often overlooked condition is the salt concentration. The oversight may not

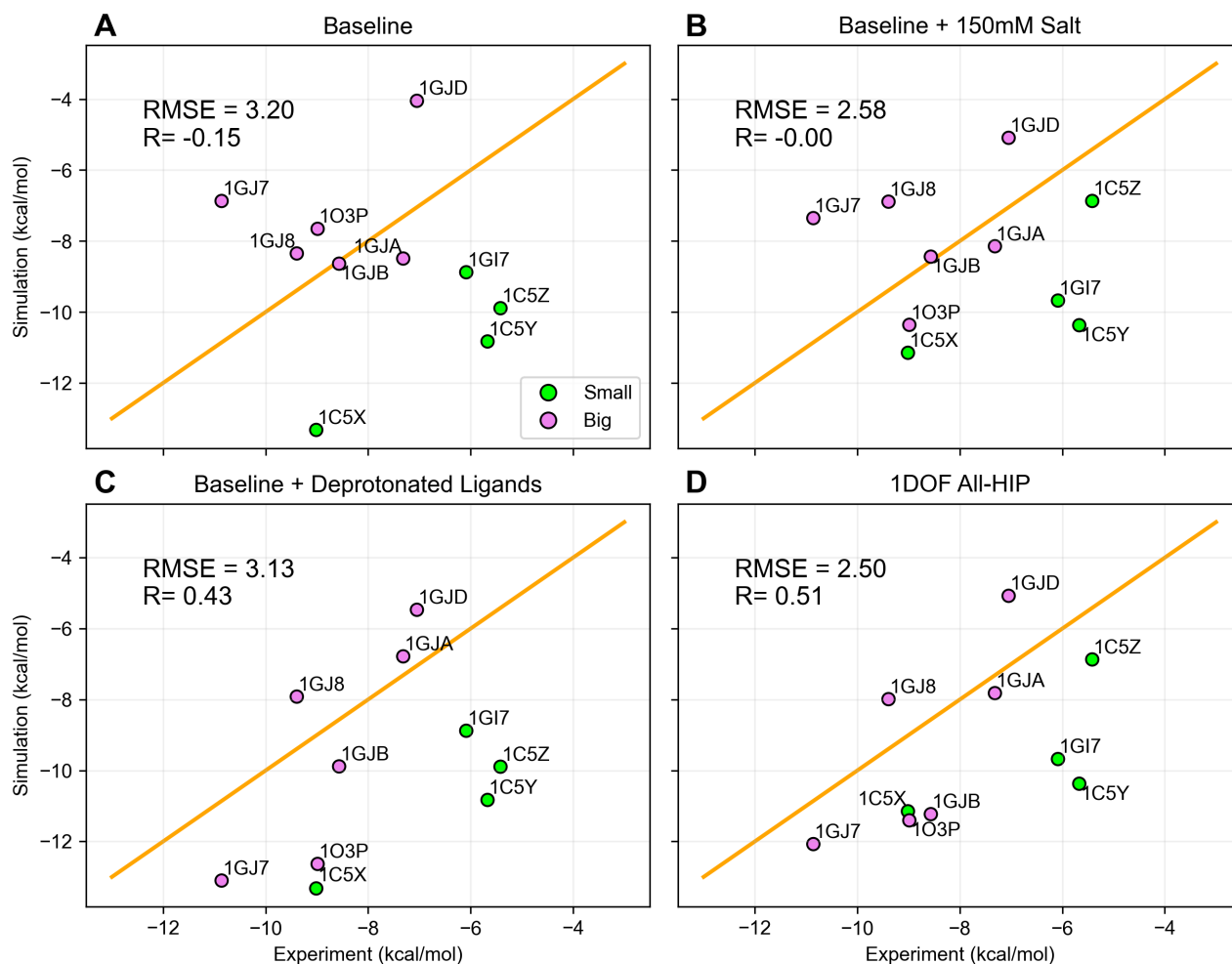


Figure 2.4: Baseline absolute alchemical binding predictions for UPA inhibitors. Evaluating the effects of simulation with 150 mM salt alone, deprotonated ligands alone, and with 150 mM salt and deprotonated ligands combined (1DOF All-HIP) on the baseline condition (fully protonated ligands, counterions added only up to neutralize system charge, and 1DOF restraints). The highest performance is observed with the 1DOF All-HIP condition with RMSE 2.50 kcal/mol and Pearson correlation 0.51.

be an issue for most neutral or hydrophobic ligands but becomes an important issue for charged ligands due to the impact of ions on electrostatic screening. Given an otherwise identical setup and identical restraint as the baseline, the use of 150 mM salt concentration reduced RMSE to 2.58 kcal/mol and improved the Pearson correlation from negative to 0 (Figure: 2.4B). Another discounted issue is the treatment of the protonation state for the ligands and amino acid side chains in the binding pocket, which is critical for defining the polar interactions that retain charged ligands. Consistent protonation states maintain hydrogen bond donor and acceptor pairing and/or charge complementary. Modification from the baseline using the deprotonated ligands was found to have minimal improvement in RMSE to 3.13 kcal/mol and significant improvement on the Pearson correlation to 0.43 (Figure: 2.4C). Further, when the 150 mM salt concentration and deprotonated ligands are combined, RMSE is noticeably reduced to 2.50 kcal/mol, and the Pearson correlation is further boosted to 0.51 (Figure: 2.4D). These comparisons highlight a consistent beneficial effect in improving both accuracy and correlation by matching the physiologically relevant salt conditions, and a greatly improved correlation when maintaining consistent protonation states.

2.4.7 6DOF versus 1DOF in predicted affinities

Both 1DOF and 6DOF restraints are widely utilized, but direct comparison has been lacking. The single distance restraint is simpler to implement and enables broader sampling of the binding pocket volume available but has been noted to require longer simulation to reach convergence and may be contaminated by erroneously high energies when the ligand is trapped in a local minimum[82]. The 6DOF approach more tightly locks the ligand into a predefined conformation with limited translational and rotational mobility to more readily achieve convergence and is dependent on securely holding the ligand in the pose that is physically relevant to binding. The binding energy predictions from the 6DOF simulation were found to have an RMSE of 5.59 kcal/mol and a Pearson correlation of 0.74 (SI Figure: 2.18). The higher Pearson correlation observed with the 6DOF restraints enables a more accurate ranking of the binding energies and may be due to restricting the ligand conformational sampling to a small number of dominant and energetically favorable poses. Indeed, the predicted binding affinities for the 6DOF runs are all more negative than the experi-

mentally determined values, consistent with the ligand being trapped in an excessively favorable binding mode with hindered sampling of higher energy states that are relevant to binding. This reflects how the entropic component is improperly estimated due to the more intensive restriction on sampling. It should also be pointed out that the higher correlation here is likely due to the more negative binding affinities spanning a larger range, indicating use of correlation alone may be insufficient in evaluating the performance.

2.4.8 Possible protonation states at active sites

We next evaluated the effects of varying the ligand and binding pocket protonation, with deprotonated ligand phenol and Hip-46 or protonated ligand phenol and Hid-46 on predictive accuracy as both satisfy the steric constraint in the crystal structures. Assignment of hydrogens is typically not resolved with structure determination by X-ray crystallography. The issue is further complicated by the absence of direct pKa measurement for the system. Nevertheless, based on the close distance between the ligand phenol hydroxyl and Ser-198, Katz et al. inferred that the ligand phenol binds as an acid and is deprotonated to minimize steric clash with surrounding atoms[62]. The free oxygen then acts as a hydrogen bond acceptor interacting with Hip-46. However, the typical pKa of a phenol hydroxyl is approximately 10 and those on the ligands range between 8 and 9[62, 110], which suggests that maintenance of the hydroxyl proton is favored under physiological conditions. His-46 would more likely assume the neutral HID form allowing hydrogen bonding to occur at Ne on Hid-46. To investigate both possibilities, trials were conducted in four groups as all-HID (all ligands interacting with Hid-46), all-HIP (all ligands interacting with Hip-46), small-HID (larger phenol ligands interacting with Hip-46 and smaller nonphenol ligands interacting with Hid-46), and small-HIP (larger phenol ligands interacting with Hid-46 and smaller nonphenol ligands interacting with Hip-46).

Utilization of deprotonated ligands and all Hip-46 (all-HIP) led to an RMSE of 2.50 kcal/mol and a Pearson correlation of 0.51 as previously shown in Figure: 2.4. In contrast, the alternative with protonated ligands and all Hid-46 (all-HID) resulted in a worse RMSE of 3.91 kcal/mol and slightly

reduced Pearson correlation of 0.47 (SI Figure: 2.19). Since the smaller and nonphenol ligands are not expected to form hydrogen-bonding contact with His-46, the protonation state of His-46 may not match that for the larger ligands. Thus, a more appropriate comparison is between small-HID versus all-HID. Interestingly the small-HID condition leads to RMSE of 3.40 kcal/mol and Pearson correlation of 0.21 (SI Figure: 2.19). For the fourth small-HIP condition, the RMSE was calculated to be 3.16 kcal/mol with the highest Pearson correlation of 0.69 (SI Figure: 2.19).

Notably, 1GJD is an outlier in all conditions, separated from the cluster of other ligands and is predicted to have a higher binding free energy than measured in experiment for both HIP and HID conditions. This is potentially due to force field imperfections as discussed above: the 1GJD phenol pivots away from Ser-198 observed during the equilibration phase (Figure: 2.3B). All other ligands adopted poses with the phenol shifted away from Ser-198 slightly to alleviate steric clash but maintained hydrogen bonding range. Thus, 1GJD is excluded from further binding analysis. Removal of 1GJD from aggregate calculations does not improve the RMSE as it increased to 2.55, 4.06, 3.52, and 3.25 kcal/mol for the all-HIP, all-HID, small-HID, and small-HIP conditions, respectively, but its omission increases Pearson correlations for all-HIP to 0.55, all-HID to 0.81, reduced small-HID to 0.14, and brings small-HIP to 0.85 (SI Figure: 2.19). These simulations demonstrate the impact of protonation state on the binding free energy prediction. Our standard alchemical simulations suggest that the all-HIP condition with the lowest RMSE and all-HID and small-HIP conditions with over 80% correlation may all explain some aspects of the experimental binding affinities. However, the absolute errors are all quite large, over 2.5 kcal/mol which is above the chemically accurate threshold of 1.0 kcal/mol. Therefore, it is still uncertain which binding mode best describes these challenging systems.

2.4.9 Estimation of electronic polarization by MBAR/PBSA

Following optimization of protein and ligand radii for all alchemical conditions tested (SI Table: 2.4), we evaluated the effect of solute dielectric on the accuracy of binding affinities to assess the impact of incorporating polarization into the computational models. Evaluation of the solute di-

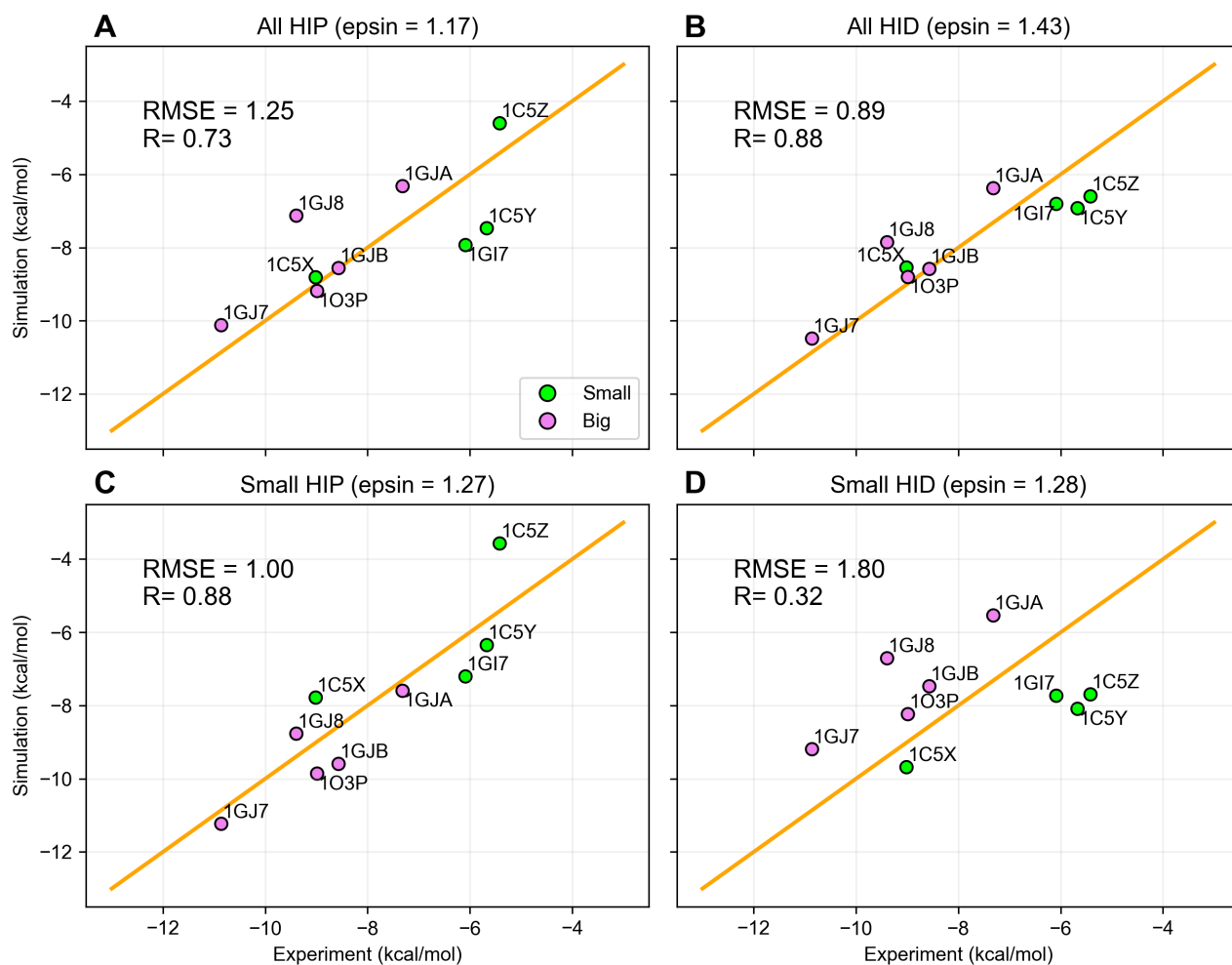


Figure 2.5: MBAR/PBSA binding affinity calculations. The All-HID condition shows the best agreement to experiment with consideration of polarization effects through solute dielectric scaling. In comparison to values from the standard alchemical transformation, RMSEs are reduced and Pearson correlations are improved for all conditions.

electric at the theoretical value of 2 responsible for electronic polarization[111–114] shows improved Pearson correlation to as high as 0.81, highlighting its applicability to correctly ranking candidate inhibitors by offsetting charge polarization errors. However, the RMSE increases to as high as 4.84 kcal/mol (SI Table: 2.5). All samples are predicted to have more positive binding free energies with the increasing solute dielectrics, demonstrating that screening charged effects increase the predicted free energies.

The standard Amber force fields were developed with effective partial charges to model electrostatics and include polarization responses to the environment (mostly in water), though only in an averaged, mean-field manner. They are not fully compatible with the theoretical dielectric constant of 2 because polarization is already partially accounted for in the effective partial charges. Thus, a further scanning procedure to find the optimal solute dielectrics is necessary. In doing so, the RMSE to experimental affinities was found to be reduced to as low as 0.89 kcal/mol and the Pearson correlation is increased to as high as 0.88 for the all-HID condition (Figure: 2.5, SI Table: 2.5). Both metrics are dramatically improved compared to the explicit solvent simulation, enabling more accurate binding free energy prediction with only postprocessing of existing trajectory data and minimal modification to current protocols.

It is interesting to see how accounting for polarization affects the prediction of binding free energy at the tested alternative protonation states. Among the three viable candidates from standard all-chemical simulations, all-HIP, all-HID, and small-HIP conditions, the all-HIP condition is improved to a 1.25 kcal/mol RMSE and 0.74 Pearson correlation, the all-HID condition is calculated to have a 0.89 kcal/mol RMSE and 0.88 Pearson correlation, and the small-HIP condition is changed to a 1.0 kcal/mol RMSE and 0.88 Pearson correlation (Table: 2.2). The last tested condition, small-HID, has a 1.80 kcal/mol RMSE and 0.32 Pearson correlation.

This suggests that the environment of the binding pocket and charged nature of the ligands may not fully support the originally proposed binding mode with deprotonated phenol and Hip-46. Instead, the alternative hypothesis of the ligands with protonated phenol as a hydrogen bond donor and Hid-46 assuming the role of hydrogen bond acceptor may better describe the protein–ligand

interactions. Both protonation states satisfy the steric constraint in the crystal structures. However, the differences in errors are within kT between different protonation states. Our analysis here points to the need for more definite NMR pKa measurement to resolve the issue[115, 116].

Condition	Method	RMSE (kcal/mol)	R
Baseline	Standard alchemical	3.22 (3.20)	-0.34 (-0.15)
Baseline + 150 mM salt	Standard alchemical	2.64 (2.58)	-0.12 (0.00)
Baseline + Deprotonated Ligands	Standard alchemical	3.25 (3.13)	0.42 (0.43)
6DOF (All-HIP)	Standard alchemical	5.85 (5.59)	0.75 (0.74)
1DOF (All-HIP)	Standard alchemical	2.55 (2.50)	0.55 (0.51)
All-HID	Standard alchemical	4.06 (3.91)	0.81 (0.47)
Small-HIP	Standard alchemical	3.25 (3.16)	0.85 (0.69)
Small-HID	Standard alchemical	3.52 (3.40)	0.14 (0.21)
All-HIP	MBAR/PBSA	1.25 (1.61)	0.73 (0.65)
All-HID	MBAR/PBSA	0.89 (1.53)	0.88 (0.67)
Small-HIP	MBAR/PBSA	1.00 (1.48)	0.88 (0.81)
Small-HID	MBAR/PBSA	1.80 (2.18)	0.32 (0.31)

Table 2.2: Summary of error and correlation statistics. Binding free energy prediction metrics with outlier 1GJD removed. Values in parentheses represent inclusion of the outlier. Conditions examining binding pocket protonation include the simulation with 150 mM salt and deprotonated ligands (1DOF All-HIP). The Baseline condition is described by inclusion of only neutralizing counterions and with fully protonated ligand phenol groups.

2.5 Conclusion

The current study aims to understand the impact of simulation conditions for absolute binding calculations in the UPA system, introduces the MBAR/PBSA continuum solvent approach in the calculation of decharging free energies to capture electronic polarization effects absent in standard explicit solvent models, and evaluates the effect of varying protonation states of titratable ligands

and protein residues in binding free energy prediction. Extensive simulations of UPA with a broad set of inhibitors were performed to benchmark the performance of absolute alchemical simulations, which have been sparsely studied due to their demanding calculation, allowing us to identify factors pivotal to increasing predictive accuracy.

The force field description of the ligands plays a significant role in maintaining important interactions and poses for binding, and issues with the ligand force field parameters can cause inaccuracies in the binding calculation as seen in the case of 1GJD. Here, difficulty maintaining the short hydrogen bond between the ligand phenol and Ser-198 caused excessive rotation of the phenol ring and also overly positive binding energy prediction. Furthermore, the setup of simulation systems contributes significantly to predictive accuracy as seen in the baseline condition, which does not account for salt concentration, protonation state, or polarization effects. These oversights lead to poor performance in prediction with a 3.2 kcal/mol RMSE and -0.15 Pearson correlation. As the simulation conditions are modified to be consistent with physiologically relevant conditions, notable improvements in the accuracy are observed with the RMSE decreasing to 2.5 kcal/mol and an increase in Pearson correlation to 0.51. The more restrictive 6DOF ligand restraints were found to overestimate binding affinity by keeping the ligand in a singular binding conformation, preventing exploration of relevant higher energy conformations, and resulting in larger error, but improved Pearson correlation compared to 1DOF restraints.

Importantly, simulation conditions that affect electrostatic interactions are observed to have a major contribution to binding prediction accuracy, augmenting results from previous studies[53, 117]. Standard MD simulation utilizing explicit solvation and point-charge models lack the capability to account for electronic polarization effects that undoubtedly occur as the ligands transition from the high-dielectric water environment to the low-dielectric protein interior. Polarization effects can be captured through ab initio quantum calculations that evaluate the electron densities surrounding each atom, but their usage is limited by steep computational costs and typically require that the system is separated into coupled QM/MM regions where the choice of boundary and level of QM theory entangle accurate treatment[22, 118]. The MBAR/PBSA calculation allows the assignment of different dielectric values to solute and solvent, enabling us to measure the impact

of including polarization effects on binding affinity prediction. The interior dielectric constant in PBSA parametrizes the strength of charge screening in the protein environment. At the default value 1, atom charges are not shielded resulting in exaggerated attractive and repulsive interactions as the atom partial charges, typically assigned for the ligand in gas phase, cannot be adjusted in standard MD simulation. This overestimation of the electrostatic potential can be offset by finely increasing the solute dielectric value to imitate the effect of electronic polarization that masks electrostatics. When the active site protonation state is defined and the polarization effects are modeled in the MBAR/PBSA calculation, significant enhancement of prediction accuracy is observed. This method is a mean-field approach demonstrated here for its ease of implementation and inspires the utilization of more explicit calculations of electronic polarization such as with polarizable multipole electrostatics[8, 11, 119].

The all-HIP condition was first inferred to be a likely protonation state that satisfies the crystal steric constraint. It was found to have a 1.25 kcal/mol RMSE and 0.73 Pearson correlation, but is not necessarily the definitive state, as the alternative protonation state, all-HID, shows higher prediction accuracy, though the differences in errors are within kT. Conclusive protonation assignment requires further experimental validation such as pKa determination via NMR spectroscopy[115, 116]. This is significant when considering how simulation protocols and algorithms deal with aspects of electrostatic interactions in defining protonation states and handling polarization effects.

Complete examination of protonation changes is limited with existing simulation protocols. Exploring alternative protonation states becomes an important and complicating process if the proton transfer events of the whole system are coupled to the binding process. In the Supporting Information, the calculation of the contributions of a single titratable group coupled to the binding process is discussed. However, this simple model is inadequate for most protein systems, as they often have multiple titratable groups that are coupled directly or through long-range allosteric interactions[120, 121]. In particular, the current UPA system involves titratable residues in the active site at His-46, His-94, and Asp-97 and titratable functional groups on several ligands such as the phenol hydroxyl. Direct interaction may shift the pKa of the titratable groups involved. The

investigation of how to approach these coupled processes has been explored by several groups[122–125] and includes approaches ranging from corrections as discussed in the SI[123, 124], the explicit enumeration of protonation states for the binding simulations, and techniques such as constant pH molecular dynamics[125].

2.6 Acknowledgements

This work was supported by the National Institute of Health/NIGMS (Grants GM093040 and GM130367).

2.7 Supplementary information

2.7.1 Apparent binding free energies with one titratable group in the active site

The presence of titratable groups in the active site poses an additional challenge for the calculation of binding free energies. The experimentally resolved apparent binding free energies encapsulate the whole physical process, which may not distinguish the contributions of coupled processes to the protein-ligand binding event. Titratable groups in the active site are susceptible to the system pH, and this susceptibility is observed in the pH dependence of receptor-ligand binding[126, 127] and enzymatic catalysis[128]. This can be further complicated when the interaction of the binding ligand shifts the pKa of those titratable groups which can alter protonation states. These interactions and coupled processes need to be considered for the binding free energy calculations. Using a single titratable group as a model coupled process, one can derive the separable contributions from the coupled processes. The coupled binding process can be separated into four separate processes: binding in the protonated form, binding in the deprotonated form, and protonation/deprotonation processes in the free and complex states defined in the illustrated thermodynamic cycle where one

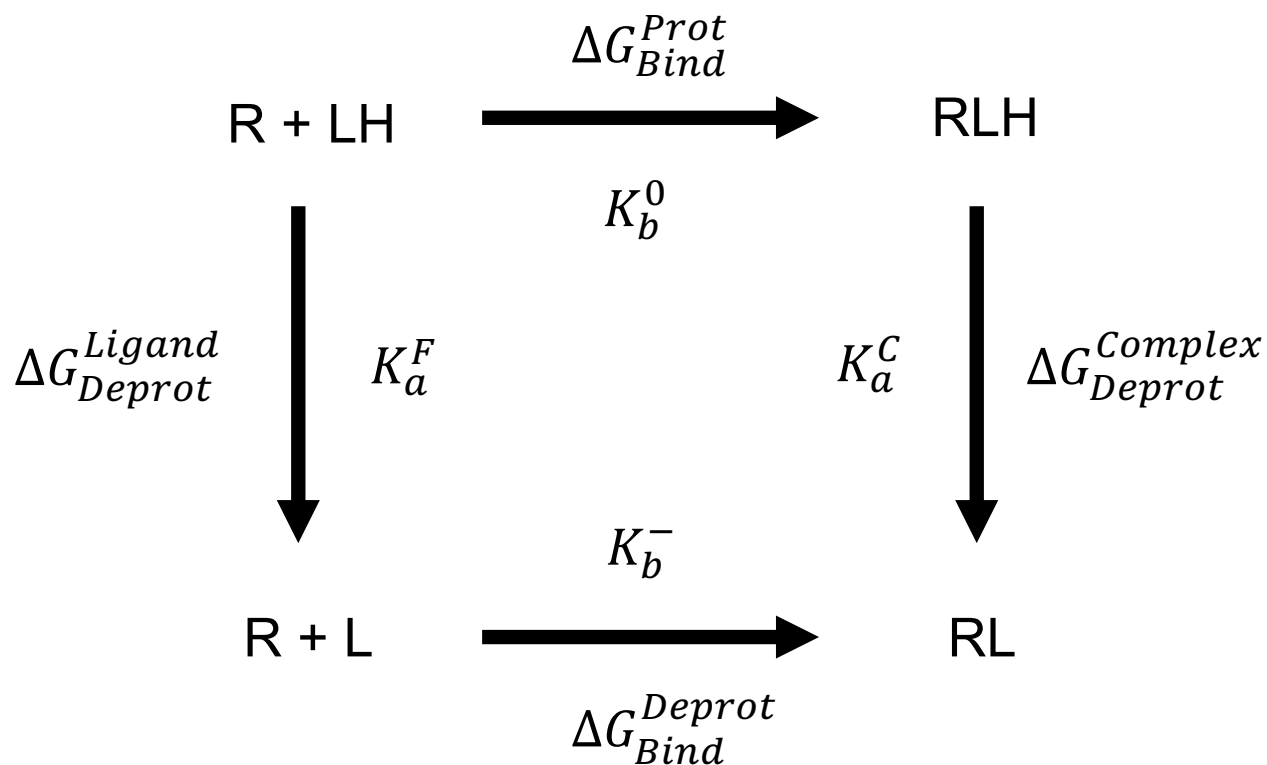


Figure 2.6: Ligand binding pKa process.

of the binding processes is directly calculated.

The apparent binding constant can be expressed with a proton dissociation process for both the complex and free states,

$$K_{app} = \frac{[RLH] + [RL^-]}{[R][LH] + [R][L^-]} \quad (2.4)$$

This can be further simplified by substitution to give,

$$K_{app} = K_b^0 \frac{(1 + (K_a^C)^{-1}[H])}{(1 + (K_a^F)^{-1}[H])} \quad (2.5)$$

where $K_b^0 = \frac{[RLH]}{[R][LH]}$ is the protonated binding equilibrium constant, $K_a^C = \frac{[RL^-][H]}{[RLH]}$ and $K_a^F =$

$\frac{[L^-][H]}{[LH]}$ are the proton dissociation constants in the complex and free states, respectively. The proton dissociation constants can be expressed in pKa and pH units, and the change in free energy can be calculated using

$$\Delta G^0(pH) = -k_b T \left[\ln K_b^0 + \ln \frac{(1 + 10^{pH - pKa^C})}{(1 + 10^{pH - pKa^F})} \right] \quad (2.6)$$

[129, 130]

The charged binding equilibrium constant can be converted to the binding free energy which can be explicitly calculated, where $\Delta G_{Bind}^{Prot} = -k_b T \ln K_b^0$, resulting in

$$\Delta G^0(pH) = \Delta G_{Bind}^{Prot} - k_b T \ln \frac{(1 + 10^{pH - pKa^C})}{(1 + 10^{pH - pKa^F})} \quad (2.7)$$

Additionally, this equation then becomes process dependent, where the equation for binding coupled to a proton association process is,

$$\Delta G^0(pH) = \Delta G_{Bind}^{Deprot} - k_b T \ln \frac{(1 + 10^{pKa^C - pH})}{(1 + 10^{pKa^F - pH})} \quad (2.8)$$

Application of the above equation shows that computation of the apparent binding free energy requires the pKa's of the ligand in both the free and bound states in addition to the simulated binding affinity of the ligand in either of the states. However, the application of this simplified single titratable group coupled binding process equation is apparently inadequate in describing the complete binding process for most protein systems where many residues in the active site are also titratable and require proper modeling.

2.7.2 SI Tables

Condition	Epsin	RMSE (kcal/mol)	R
-----------	-------	-----------------	---

All HIP	1	2.31	0.6
All HID	1	3.75	0.75
Small HIP	1	2.89	0.85
Small HID	1	3.32	0.17

Table 2.3: MBAR/PBSA binding affinity accuracy with optimized Radiscale and Protscale parameters. Radiscale and Protscale values were scaled to minimize mean absolute error between MBAR/PBSA free energies and explicit solvent free energies.

Condition	Epsin	RMSE (kcal/mol)	R
All-HIP	1.17	1.25	0.74
All-HIP	2	4.84	0.81
All-HID	1.43	0.89	0.88
All-HID	2	2.65	0.88
Small-HIP	1.27	1	0.88
Small-HIP	2	3.91	0.87
Small-HID	1.28	1.8	0.32
Small-HID	2	3.9	0.52

Table 2.4: Binding affinity prediction accuracy versus solute interior dielectric (Epsin) with MBAR/PBSA. Commonly used Epsin of 2.0 and Epsin resulting in the lowest RMSE to experiment are reported. All-HID condition shows the lowest RMSE and highest Pearson correlation at optimized Epsin. Epsin 2.0 results in improved Pearson correlations, but also higher RMSE's.

Sample	Baseline	Baseline + 150mM salt	Baseline + de-protodated ligands	1DOF	6DOF	All	Small	Small	Small	PBSA	PBSA	PBSA	PBSA	PBSA	PBSA	Experiment
				All- Hip	All- Hip	All HID	HIP	HIP	HID	All HIP	All HID	Small HIP	Small HIP	Small HID	Small HID	
1C5X	-13.32	-11.15	-13.32	-11.15	-14.29	-12.68	-11.15	-11.15	-12.68	-8.81	-8.55	-7.78	-9.68	-9.01		
1C5Y	-10.83	-10.37	-10.83	-10.37	-13.91	-11.66	-10.37	-10.37	-11.66	-7.47	-6.93	-6.35	-8.09	-5.67		
1C5Z	-9.89	-6.87	-9.89	-6.87	-8.95	-10.74	-6.87	-6.87	-10.74	-4.6	-6.6	-3.57	-7.7	-5.42		
1G17	-8.88	-9.68	-8.88	-9.68	-11.39	-10.26	-9.68	-9.68	-10.26	-7.93	-6.81	-7.21	-7.73	-6.09		
1G17	-6.87	-7.35	-13.1	-12.08	-17.74	-13.41	-13.41	-13.41	-12.08	-10.12	-10.49	-11.23	-9.19	-10.86		
1G18	-8.35	-6.89	-7.92	-7.99	-13.66	-11.51	-11.51	-11.51	-7.99	-7.13	-7.86	-8.77	-6.71	-9.39		
1G1A	-8.49	-8.15	-6.78	-7.82	-9.15	-10.87	-10.87	-10.87	-7.82	-6.32	-6.38	-7.6	-5.54	-7.32		
1G1B	-8.64	-8.44	-9.89	-11.23	-15.75	-12.5	-12.5	-12.5	-11.23	-8.56	-8.58	-9.59	-7.47	-8.57		
1G1D	-4.05	-5.09	-5.47	-5.08	-9.14	-4.85	-4.85	-4.85	-5.08	-3.6	-3.03	-3.48	-2.73	-7.05		
1O3P	-7.66	-10.36	-12.63	-11.41	-16.18	-12.73	-12.73	-12.73	-11.41	-9.19	-8.81	-9.86	-8.24	-8.99		

Table 2.5: Full binding predictions at all conditions tested compared to experimental values. Absolute binding free energy calculations aggregated from 5 independent replicates with randomized starting velocities. All units reported in kcal/mol.

2.7.3 SI Figures

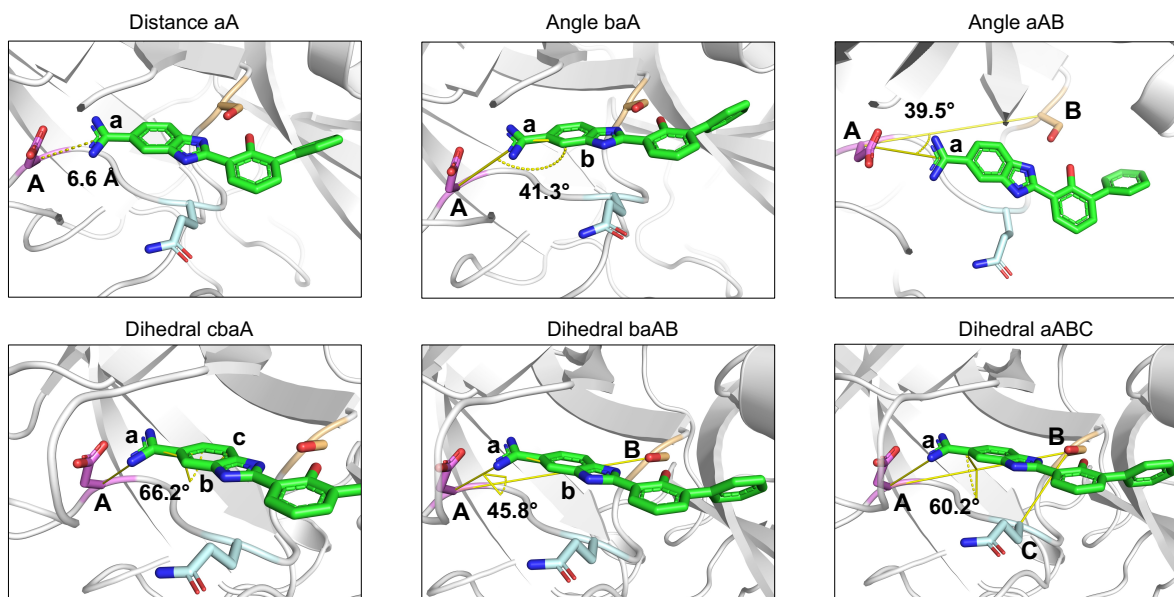


Figure 2.7: Illustration of Boresch 6DOF orientational restraints. The ligand is constrained by a single distance, two angles, and three dihedrals selected from the end of the equilibration phase to lock the ligand into a target conformation. 1DOF condition involves only the distance restraint, which allows greater exploration of conformational states at the cost of slower convergence.

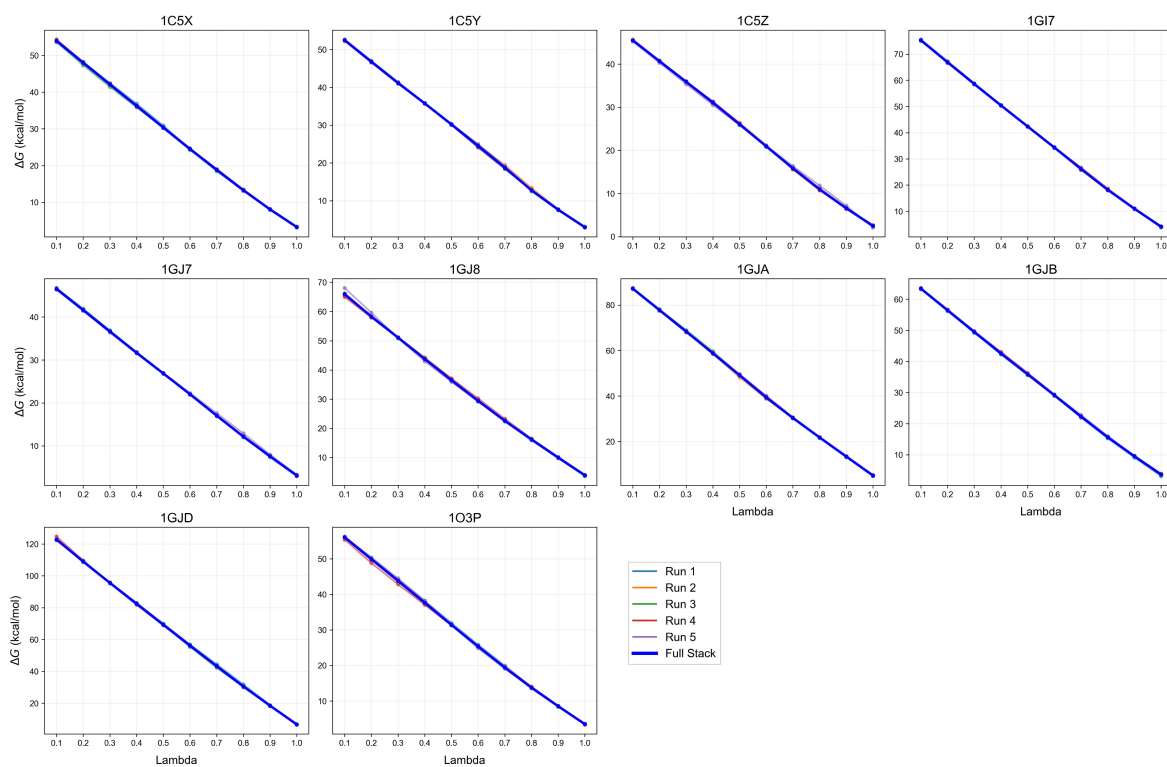


Figure 2.8: . Free energy transitions during the decharging phase for the complex trajectories in the baseline simulation. Individual replicates show only small variation, the aggregated energies show almost complete overlap and smooth, nearly linear transition from full ligand partial charges to zero.

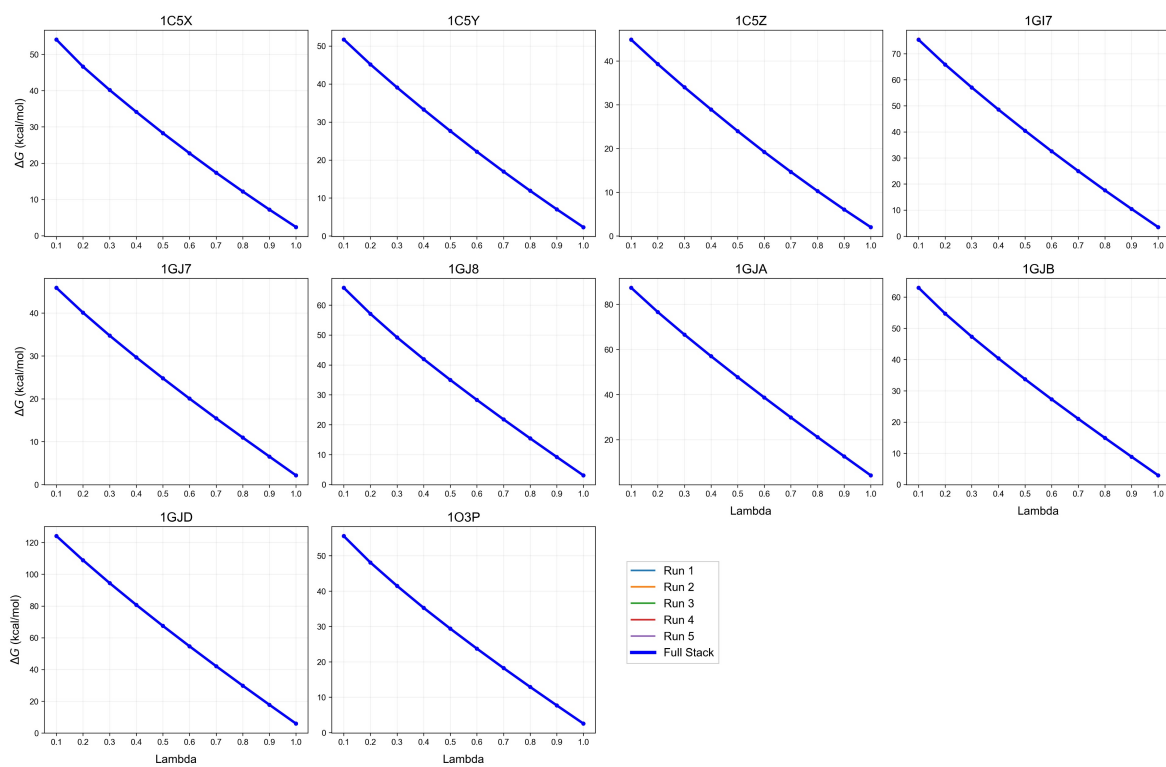


Figure 2.9: Free energy transition during the decharging phase for the ligand trajectories in the baseline alchemical simulation. The same pattern of small variation and linear transition from full ligand partial charges to zero as the complex is observed.

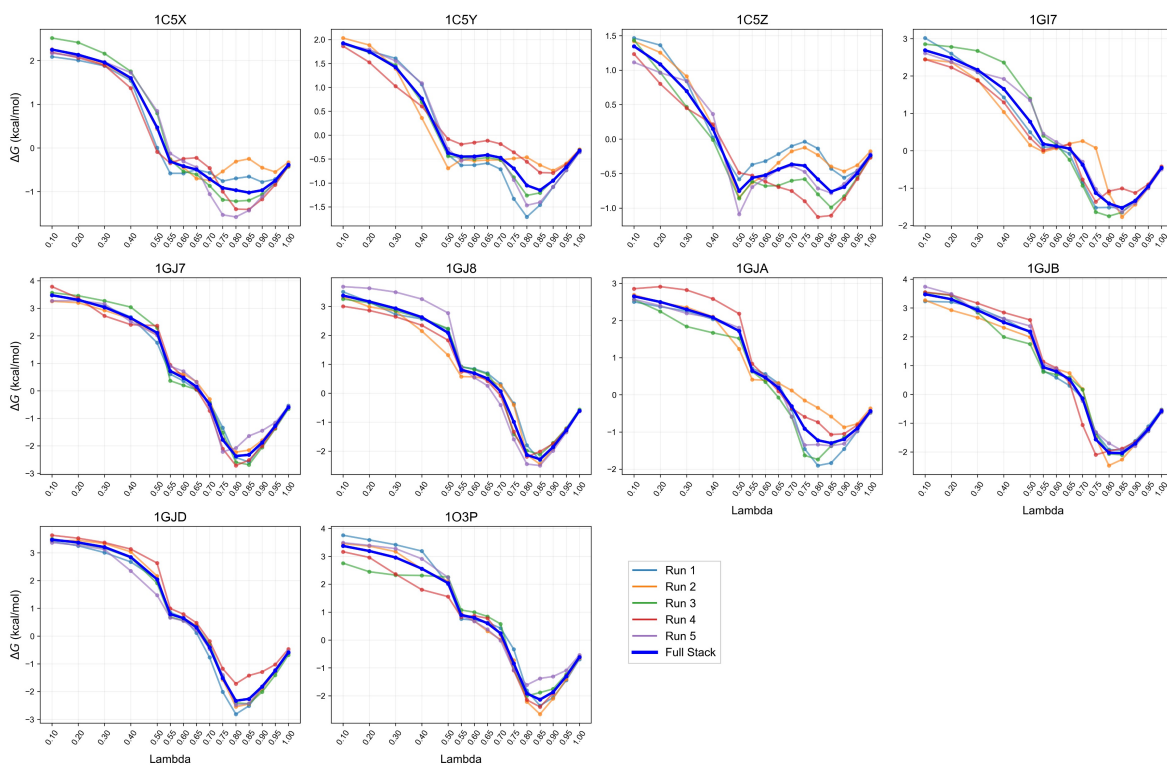


Figure 2.10: Free energy transition during the VDW phase for the complex trajectories in the baseline simulation. High variance is observed between replicates, highlighting the sampling difficulties associated with decoupling VDW interactions.

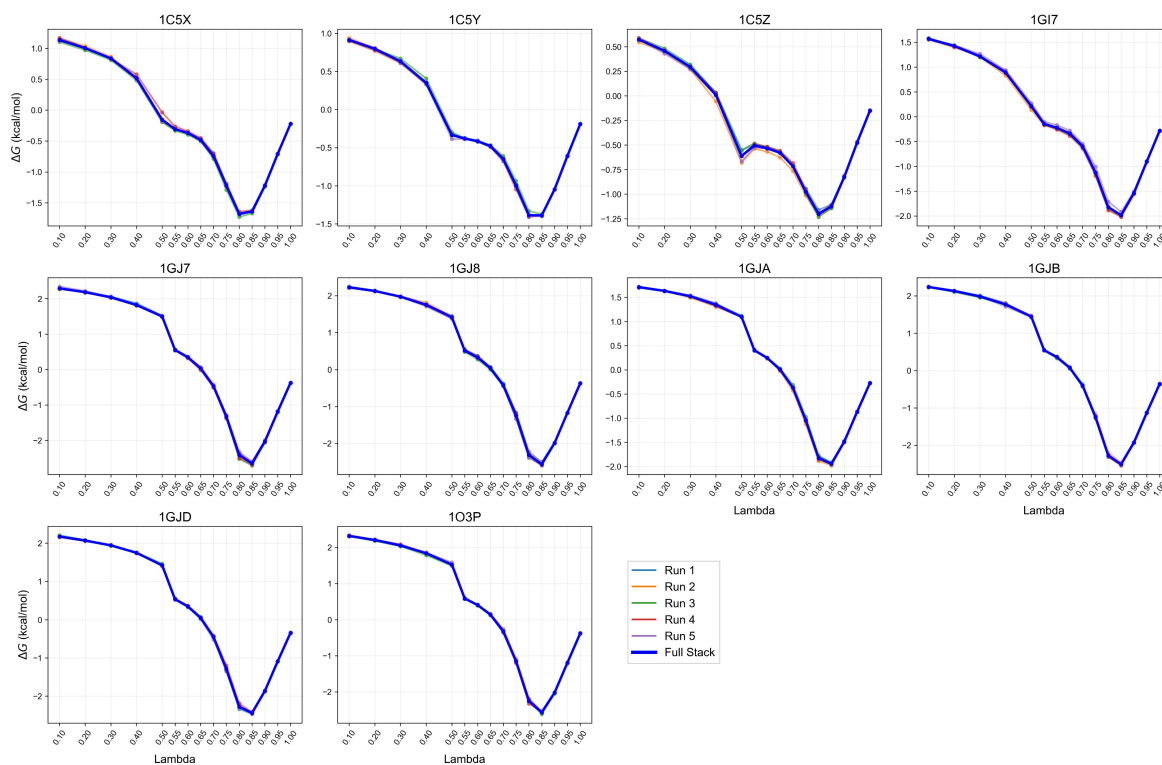


Figure 2.11: Free energy transition during the VDW phase for the ligand trajectories in the baseline simulation. Replicates show high agreement over the course of the highly non-linear transitions.

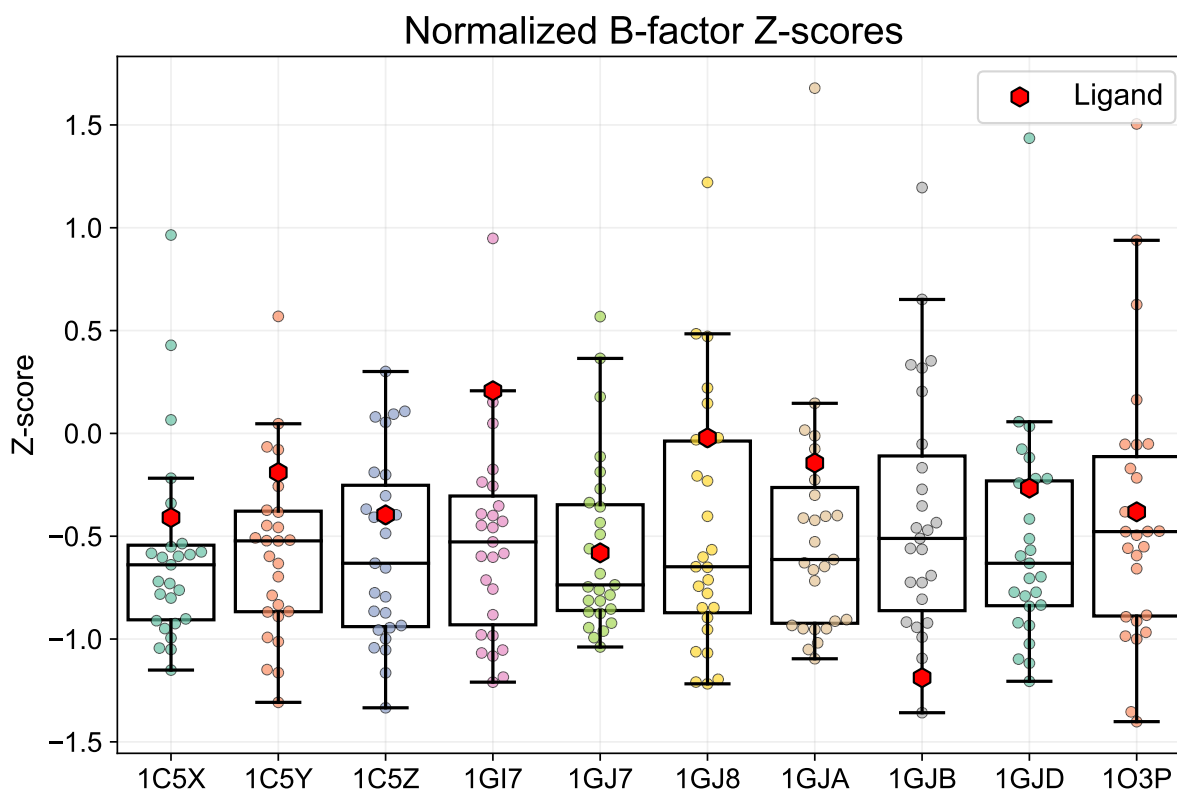


Figure 2.13: Analysis of binding pocket flexibility through normalized B-factor Z-scores. All structures show similar binding pocket flexibility, with higher than average rigidity relative to the rest of the protein. Ligands show varying levels of displacement, notably 1GI7 shows the highest flexibility, which is larger in size but unable to form a hydrogen bond to Ser-198. 1GJB shows the highest stability, potentially due to its hydrophobic benzene groups and internal hydrogen bond between the ligand phenol and nitrogen. Each marker represents the Z-score per residue with all atoms averaged.

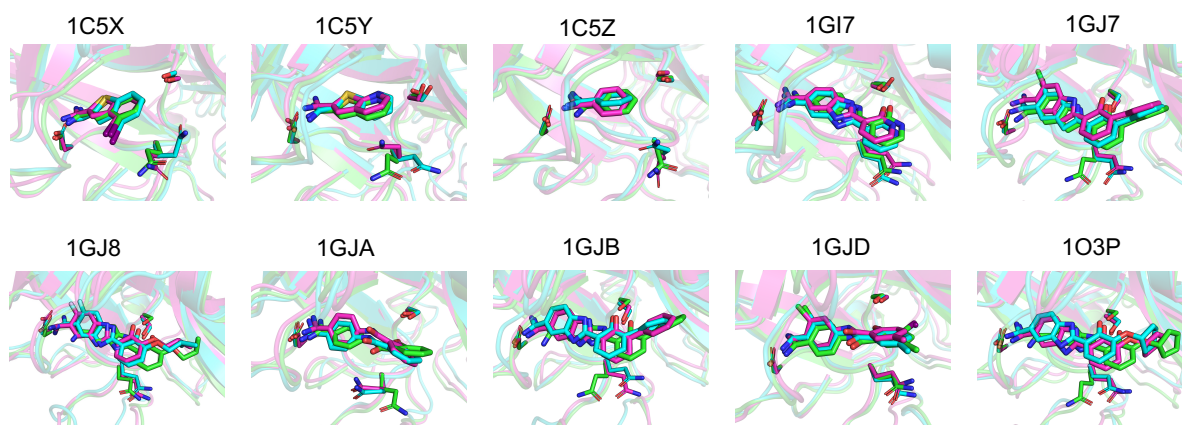


Figure 2.14: Inhibitor equilibration poses from GAFF and GAFF2 compared to starting crystal poses. GAFF and GAFF2 trajectories show similar trends, with the ligands moving further into the binding pocket to more tightly interact with Asp-192, and outward twisting of the phenol tail to relieve steric clash. Structures were generated from identifying the frame with the lowest RMSD to the average structure from the last 10 ns of equilibration. The starting crystal structure models are colored green, GAFF samples are colored cyan, and GAFF2 samples are colored purple.

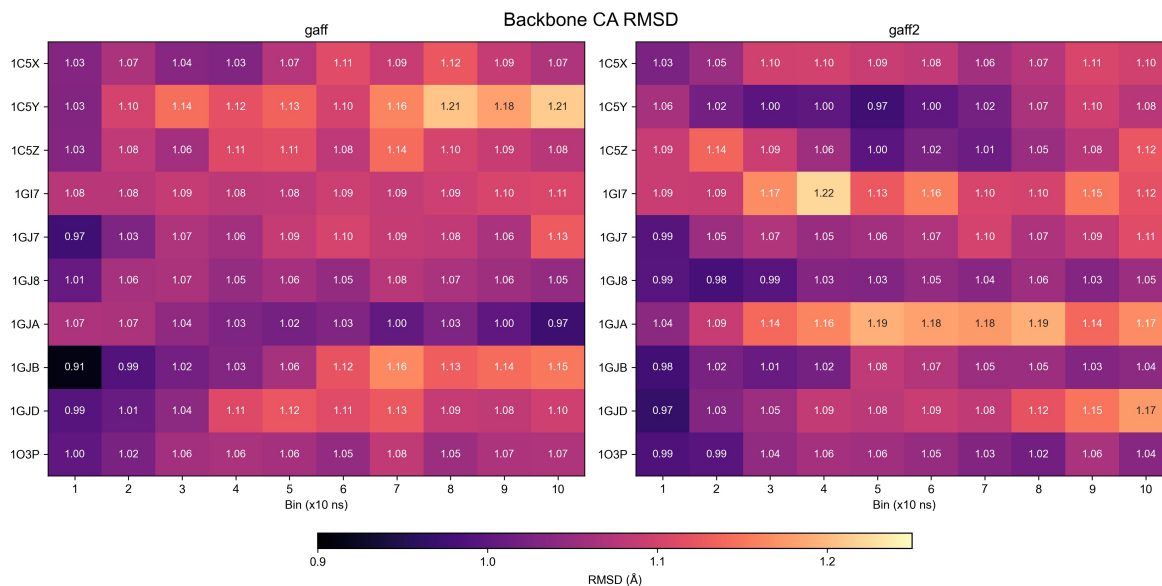


Figure 2.15: Backbone CA RMSD development over equilibration with GAFF and GAFF2 force fields. No clear pattern emerges, all proteins drift away from the starting ligand pose and show a maximum divergence of ~ 1.2 Å RMSD, indicating that minor conformational changes occur.

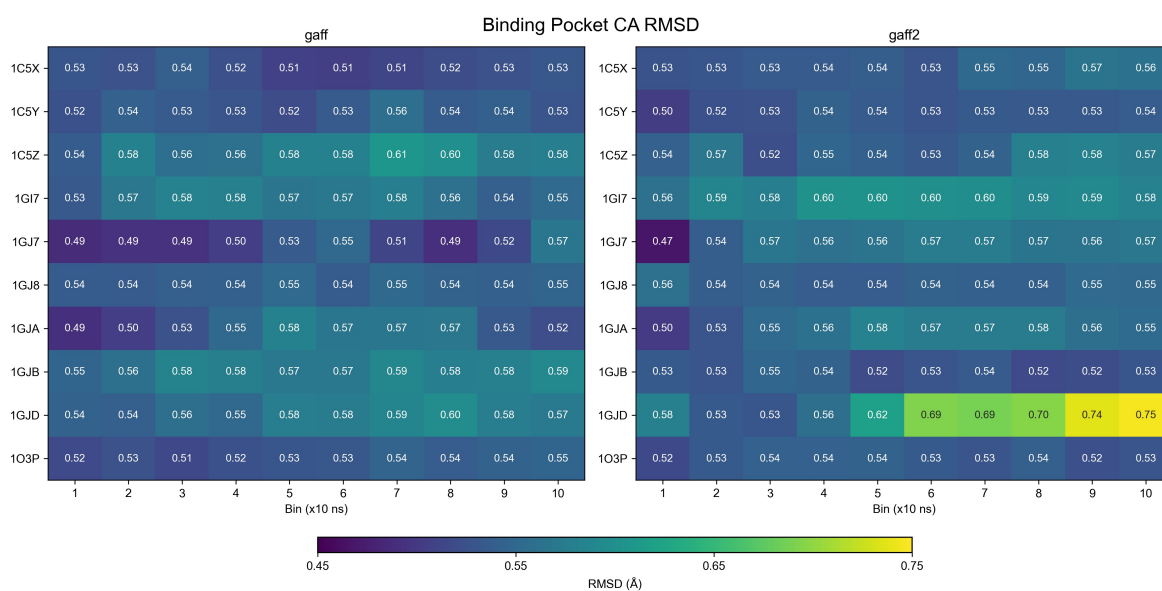


Figure 2.16: Binding pocket CA RMSD development over equilibration with GAFF and GAFF2 force fields. All GAFF samples show stability and do not change noticeably from the crystal over the course of equilibration. In GAFF2, 1GJD shows larger divergence from the crystal pose reaching 0.75 Å RMSD.

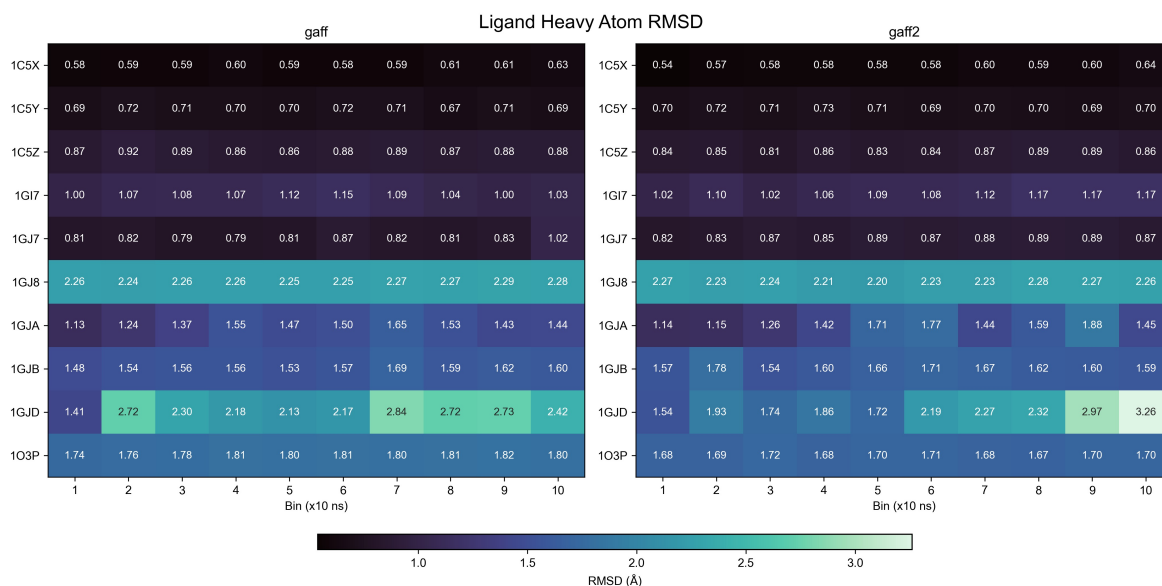


Figure 2.17: Ligand heavy atom RMSD development over the equilibration with GAFF and GAFF2 force fields. Small ligands (1C5X, 1C5Y, 1C5Z, and 1GI7) show minimal changes in positioning. 1GJ8 shows consistent departure from the crystal pose, the ligand moves further into the binding pocket to maximize hydrophobic interactions and polar interactions with Asp-192. 1GJD shows dissimilarity with crystal as well, from the rotation of the phenol group outward leading to the loss of the hydrogen bond. The aberration is more substantial with GAFF2.

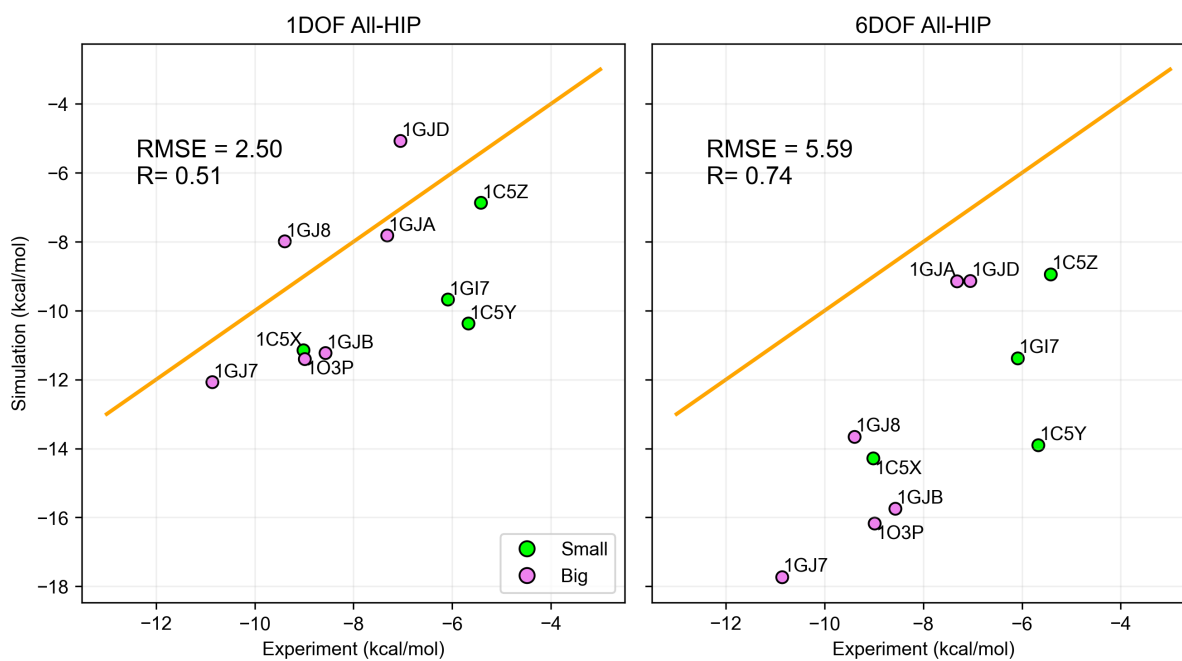


Figure 2.18: Comparison of 1DOF and 6DOF restraint schemes. The 1DOF single distance restraint showed lower error, but worse Pearson correlation than the 6DOF (Borresch) method. Samples with the 6DOF restraint showed excessively negative free energy predictions, indicating potential over-stabilization in a favorable pose.

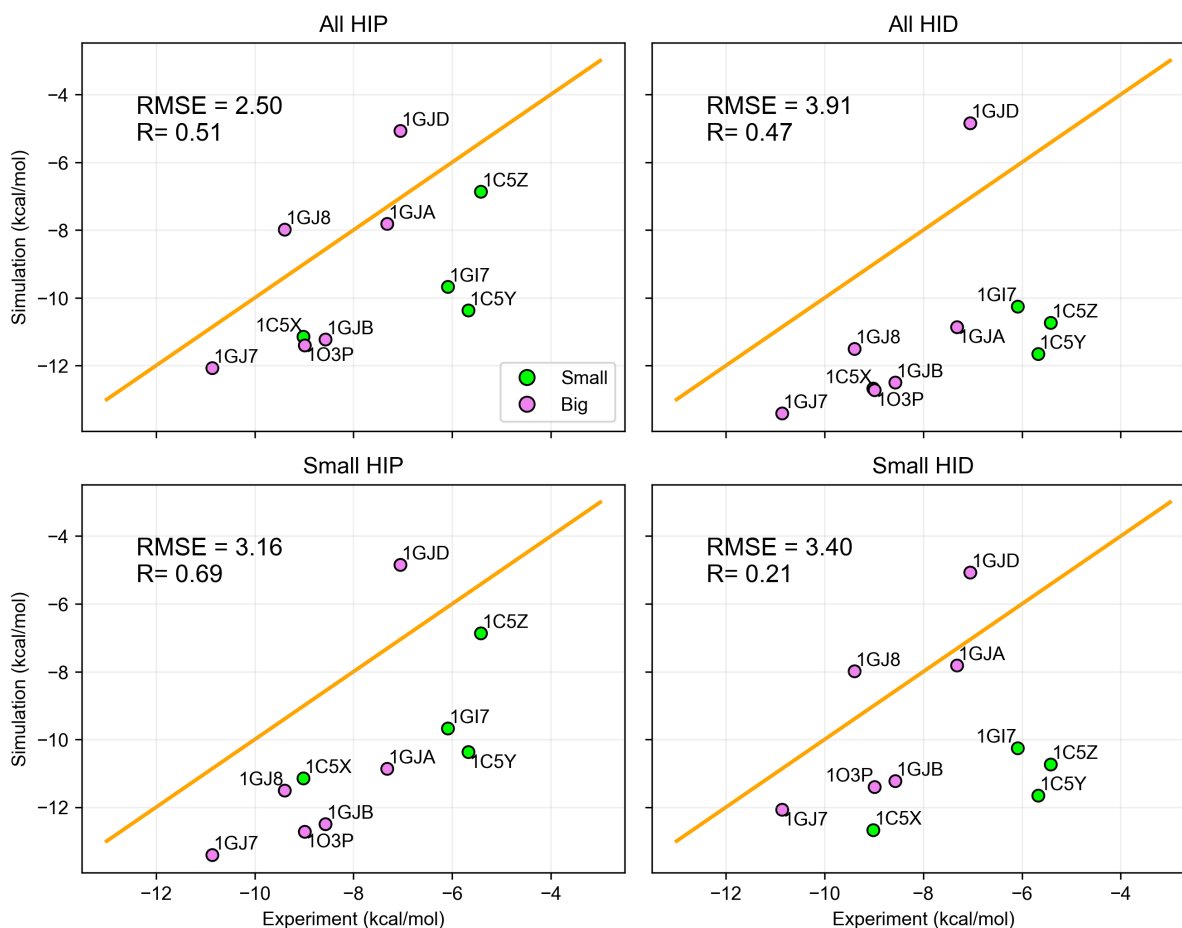


Figure 2.19: Binding affinity predictions with standard alchemical simulation with different protonation states. In general, binding affinities are predicted to be more negative than expected, possibly due to exaggeration of favorable charge-charge interactions typical of the point-charge models used. 1GJD is shown to be an outlier, with free energies far more positive than the cluster of other tested ligands, this is likely related to the issues in sampling incorrect binding poses recognized during equilibration where the phenol swings outward such that the native hydrogen bond to Ser-198 is not maintained.

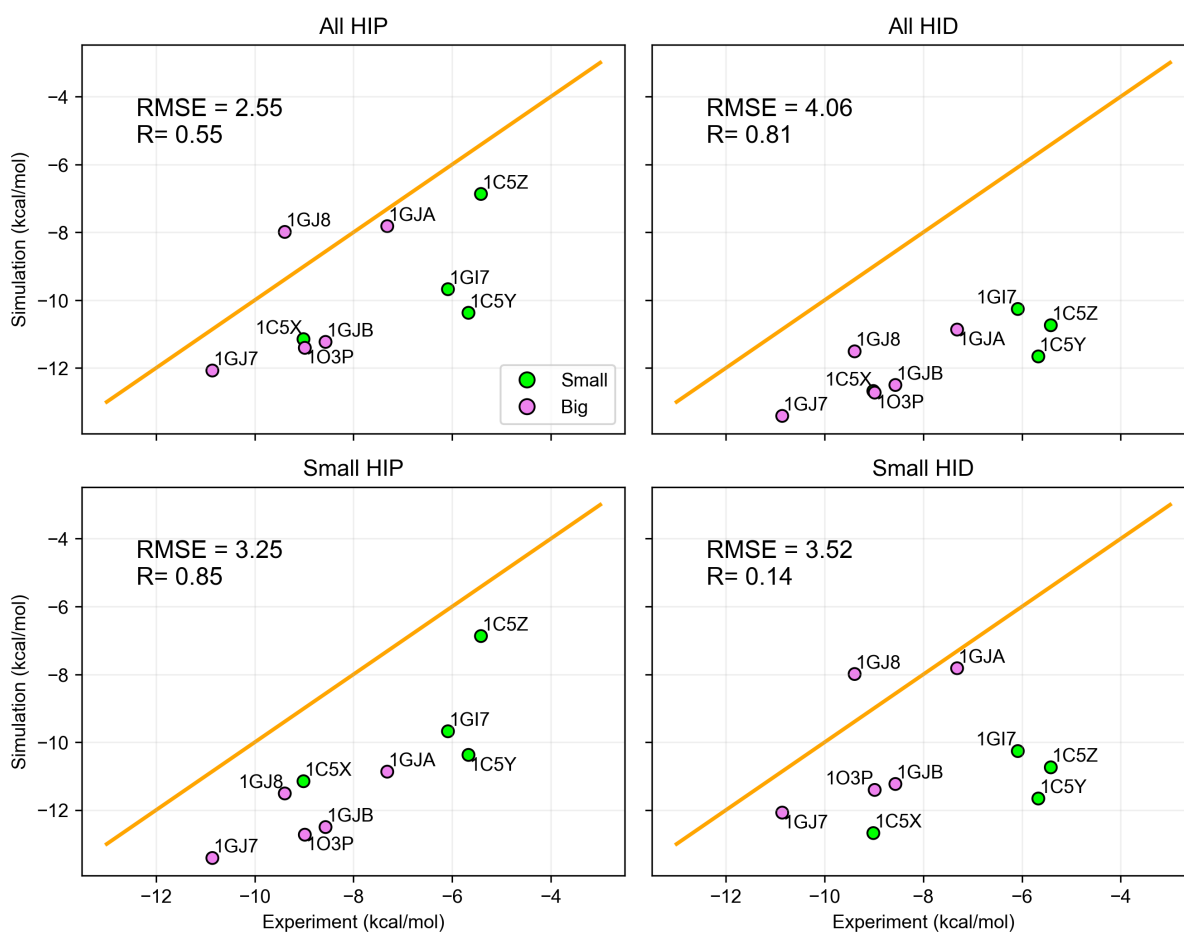


Figure 2.20: Binding affinity predictions with outlier 1GJD removed for standard alchemical simulation with different protonation states. In the standard alchemical simulation, minimal change is seen in RMSE for all conditions. However, Pearson correlation is found to improve dramatically for both All-HID and Small-HIP conditions.

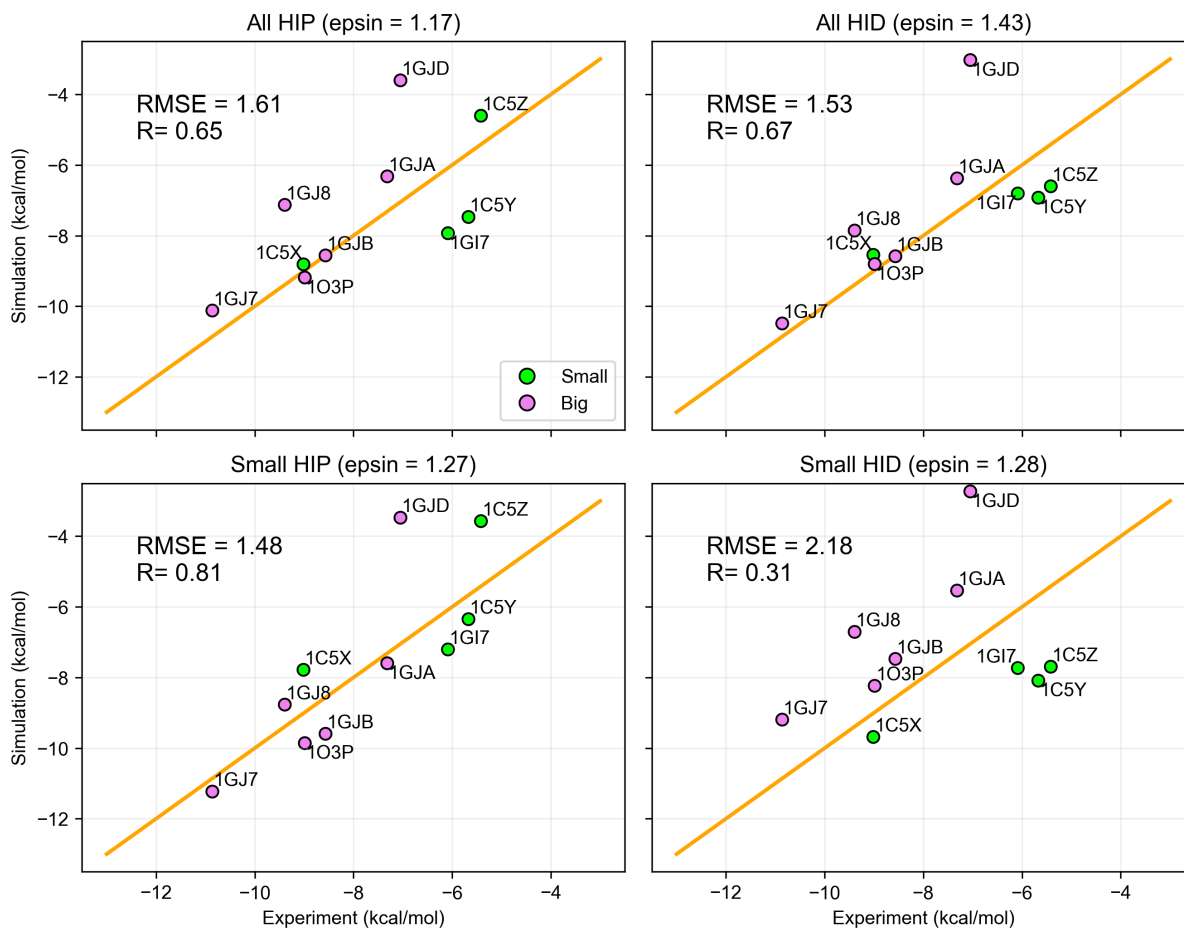


Figure 2.21: MBAR/PBSA binding affinity calculations including the outlier 1GJD. All metrics are found to worsen with the outlier pushing the trend toward overly positive values.

Bibliography

- [1] Enrico Clementi, Hans Kistenmacher, Włodzimierz Kołos, and Silvano Romano. Non-additivity in water-ion-water interactions. *Theor. Chim. Acta*, 55(4):257–266, December 1980.
- [2] Pengyu Ren and Jay W Ponder. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B*, 107(24):5933–5947, June 2003.
- [3] George A Kaminski, Harry A Stern, B J Berne, Richard A Friesner, Yixiang X Cao, Robert B Murphy, Ruhong Zhou, and Thomas A Halgren. Development of a polarizable force field for proteins via ab initio quantum chemistry: first generation model and gas phase tests. *J. Comput. Chem.*, 23(16):1515–1531, December 2002.
- [4] Richard A Friesner. Modeling polarization in proteins and protein-ligand complexes: Methods and preliminary results. *Adv. Protein Chem.*, 72:79–104, 2005.
- [5] Pedro E M Lopes, Guillaume Lamoureux, Benoît Roux, and Alexander D MacKerell. Polarizable empirical force field for aromatic compounds based on the classical drude oscillator. *J. Phys. Chem. B*, 111(11):2873–2885, March 2007.
- [6] Sandeep Patel, Alexander D Mackerell, Jr, and Charles L Brooks, 3rd. CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.*, 25(12):1504–1514, September 2004.
- [7] Wei Jiang, David J Hardy, James C Phillips, Alexander D MacKerell, Klaus Schulten, and Benoît Roux. High-Performance scalable molecular dynamics simulations of a polarizable

- force field based on classical drude oscillators in NAMD. *J. Phys. Chem. Lett.*, 2(2):87–92, January 2011.
- [8] Jay W Ponder, Chuanjie Wu, Pengyu Ren, Vijay S Pande, John D Chodera, Michael J Schnieders, Imran Haque, David L Mobley, Daniel S Lambrecht, Robert A DiStasio, Martin Head-Gordon, Gary N I Clark, Margaret E Johnson, and Teresa Head-Gordon. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B*, 114(8):2549–2564, March 2010.
- [9] Yue Shi, Zhen Xia, Jiajing Zhang, Robert Best, Chuanjie Wu, Jay W Ponder, and Pengyu Ren. Polarizable atomic Multipole-Based AMOEBA force field for proteins. *J. Chem. Theory Comput.*, 9(9):4046–4063, September 2013.
- [10] Changsheng Zhang, Chao Lu, Zhifeng Jing, Chuanjie Wu, Jean-Philip Piquemal, Jay W Ponder, and Pengyu Ren. AMOEBA polarizable atomic multipole force field for nucleic acids. *J. Chem. Theory Comput.*, 14(4):2084–2108, April 2018.
- [11] Haixin Wei, Ruxi Qi, Junmei Wang, Piotr Cieplak, Yong Duan, and Ray Luo. Efficient formulation of polarizable gaussian multipole electrostatics for biomolecular simulations. *J. Chem. Phys.*, 153(11):114116, September 2020.
- [12] Junmei Wang, Piotr Cieplak, Ray Luo, and Yong Duan. Development of polarizable gaussian model for molecular mechanical calculations i: Atomic polarizability parameterization to reproduce ab initio anisotropy. *J. Chem. Theory Comput.*, 15(2):1146–1158, February 2019.
- [13] Piotr Cieplak, James Caldwell, and Peter Kollman. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and n-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *J. Comput. Chem.*, 22(10):1048–1057, July 2001.
- [14] Junmei Wang, Piotr Cieplak, Jie Li, Tingjun Hou, Ray Luo, and Yong Duan. Development

- of polarizable models for molecular mechanical calculations i: Parameterization of atomic polarizability. *J. Phys. Chem. B*, 115(12):3091–3099, March 2011.
- [15] Junmei Wang, Piotr Cieplak, Jie Li, Jun Wang, Qin Cai, Mengjuei Hsieh, Hongxing Lei, Ray Luo, and Yong Duan. Development of polarizable models for molecular mechanical calculations II: induced dipole models significantly improve accuracy of intermolecular interaction energies. *J. Phys. Chem. B*, 115(12):3100–3111, March 2011.
- [16] Jun Wang, Piotr Cieplak, Qin Cai, Meng-Juei Hsieh, Junmei Wang, Yong Duan, and Ray Luo. Development of polarizable models for molecular mechanical calculations. 3. polarizable water models conforming to thole polarization screening schemes. *J. Phys. Chem. B*, 116(28):7999–8008, July 2012.
- [17] Junmei Wang, Piotr Cieplak, Jie Li, Qin Cai, Meng-Juei Hsieh, Ray Luo, and Yong Duan. Development of polarizable models for molecular mechanical calculations. 4. van der waals parametrization. *J. Phys. Chem. B*, 116(24):7088–7101, June 2012.
- [18] Michael K Gilson and Huan-Xiang Zhou. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36:21–42, 2007.
- [19] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.*, 47:20–33, May 2016.
- [20] Samuel DeLuca, Karen Khar, and Jens Meiler. Fully flexible docking of medium sized ligand libraries with RosettaLigand. *PLoS One*, 10(7):e0132508, July 2015.
- [21] Anthony J Clark, Pratyush Tiwary, Ken Borrelli, Shulu Feng, Edward B Miller, Robert Abel, Richard A Friesner, and B J Berne. Prediction of Protein–Ligand binding poses via a combination of induced fit docking and metadynamics simulations. *J. Chem. Theory Comput.*, 12(6):2990–2998, June 2016.
- [22] Martin A Olsson and Ulf Ryde. Comparison of QM/MM methods to obtain Ligand-Binding free energies. *J. Chem. Theory Comput.*, 13(5):2245–2253, May 2017.

- [23] Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O’Meara, Tao Che, Enkhjargal Alгаа, Kateryna Tolmachova, Andrey A Tolmachev, Brian K Shoichet, Bryan L Roth, and John J Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, February 2019.
- [24] L David, R Luo, and M K Gilson. Ligand-receptor docking with the mining minima optimizer. *J. Comput. Aided Mol. Des.*, 15(2):157–171, February 2001.
- [25] Johan Åqvist, Carmen Medina, and Jan-Erik Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng. Des. Sel.*, 7(3):385–391, March 1994.
- [26] Bill R Miller, 3rd, T Dwight McGee, Jr, Jason M Swails, Nadine Homeyer, Holger Gohlke, and Adrian E Roitberg. MMPBSA.py: An efficient program for End-State free energy calculations. *J. Chem. Theory Comput.*, 8(9):3314–3321, September 2012.
- [27] Changhao Wang, Peter H Nguyen, Kevin Pham, Danielle Huynh, Thanh-Binh Nancy Le, Hongli Wang, Pengyu Ren, and Ray Luo. Calculating protein-ligand binding affinities with MMPBSA: Method and error analysis. *J. Comput. Chem.*, 37(27):2436–2446, October 2016.
- [28] Li Xiao, Jianxiong Diao, D’artagnan Greene, Junmei Wang, and Ray Luo. A continuum Poisson-Boltzmann model for membrane channel proteins. *J. Chem. Theory Comput.*, 13(7): 3398–3412, July 2017.
- [29] Ruxi Qi, Wesley M Botello-Smith, and Ray Luo. Acceleration of linear Finite-Difference Poisson-Boltzmann methods on graphics processing units. *J. Chem. Theory Comput.*, 13(7): 3378–3387, July 2017.
- [30] D’artagnan Greene, Ruxi Qi, Remy Nguyen, Tianyin Qiu, and Ray Luo. A heterogeneous dielectric implicit membrane model for the calculation of MMPBSA binding free energies. *J. Chem. Inf. Model.*, May 2019.
- [31] Ruxi Qi and Ray Luo. Robustness and efficiency of Poisson-Boltzmann modeling on graphics processing units. *J. Chem. Inf. Model.*, 59(1):409–420, January 2019.

- [32] Haixin Wei, Aaron Luo, Tianyin Qiu, Ray Luo, and Ruxi Qi. Improved Poisson-Boltzmann methods for High-Performance computing. *J. Chem. Theory Comput.*, 15(11):6190–6202, November 2019.
- [33] Haixin Wei, Ray Luo, and Ruxi Qi. An efficient second-order poisson-boltzmann method. *J. Comput. Chem.*, 40(12):1257–1269, May 2019.
- [34] 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(3):1047–1069, March 1997.
- [35] Rui Luo, Martha S Head, John Moult, and Michael K Gilson. pka shifts in small molecules and HIV protease: Electrostatics and conformation. *J. Am. Chem. Soc.*, 120(24):6138–6146, June 1998.
- [36] R Luo, M S Head, J A Given, and M K Gilson. Nucleic acid base-pairing and n-methylacetamide self-association in chloroform: affinity and conformation. *Biophys. Chem.*, 78(1-2):183–193, April 1999.
- [37] Junjie Zou, Chuan Tian, and Carlos Simmerling. Blinded prediction of protein–ligand binding affinity using amber thermodynamic integration for the 2018 D3R grand challenge 4. *J. Comput. Aided Mol. Des.*, September 2019.
- [38] M Olivia Kim, Patrick G Blachly, and J Andrew McCammon. Conformational dynamics and binding free energies of inhibitors of BACE-1: From the perspective of protonation equilibria. *PLoS Comput. Biol.*, 11(10):e1004341, October 2015.
- [39] Timothy J Giese and Darrin M York. A GPU-Accelerated parameter interpolation thermodynamic integration free energy method. *J. Chem. Theory Comput.*, February 2018.
- [40] Tai-Sung Lee, Yuan Hu, Brad Sherborne, Zhuyan Guo, and Darrin M York. Toward fast and accurate binding affinity prediction with pmemdGTI: An efficient implementation of GPU-Accelerated thermodynamic integration. *J. Chem. Theory Comput.*, 13(7):3077–3084, July 2017.

- [41] Matteo Aldeghi, Alexander Heifetz, Michael J Bodkin, Stefan Knapp, and Philip C Biggin. Predictions of ligand selectivity from absolute binding free energy calculations. *J. Am. Chem. Soc.*, 139(2):946–957, January 2017.
- [42] Mauro Lapelosa, Emilio Gallicchio, and Ronald M Levy. Conformational transitions and convergence of absolute binding free energy calculations. *J. Chem. Theory Comput.*, 8(1):47–60, January 2012.
- [43] Wei Jiang and Benoît Roux. Free energy perturbation hamiltonian Replica-Exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations. *J. Chem. Theory Comput.*, 6(9):2559–2565, July 2010.
- [44] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein-Ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.*, 57(4):942–957, April 2017.
- [45] Spencer S Ericksen, Haozhen Wu, Huikun Zhang, Lauren A Michael, Michael A Newton, F Michael Hoffmann, and Scott A Wildman. Machine learning consensus scoring improves performance across targets in Structure-Based virtual screening. *J. Chem. Inf. Model.*, 57(7):1579–1590, July 2017.
- [46] Janaina Cruz Pereira, Ernesto Raúl Caffarena, and Cicero Nogueira Dos Santos. Boosting Docking-Based virtual screening with deep learning. *J. Chem. Inf. Model.*, 56(12):2495–2506, December 2016.
- [47] Pavel V Klimovich, Michael R Shirts, and David L Mobley. Guidelines for the analysis of free energy calculations. *J. Comput. Aided Mol. Des.*, 29(5):397–411, May 2015.
- [48] Michael R Shirts, David L Mobley, and John D Chodera. Chapter 4 alchemical free energy calculations: Ready for prime time? In D C Spellmeyer and R Wheeler, editors, *Annual Reports in Computational Chemistry*, volume 3, pages 41–59. Elsevier, January 2007.
- [49] Ariën S Rustenburg, Justin Dancer, Baiwei Lin, Jianwen A Feng, Daniel F Ortwine, David L Mobley, and John D Chodera. Measuring experimental cyclohexane-water distribution coef-

- ficients for the SAMPL5 challenge. *J. Comput. Aided Mol. Des.*, 30(11):945–958, November 2016.
- [50] Vytautas Gapsys, Servaas Michielssens, Daniel Seeliger, and Bert L de Groot. Accurate and rigorous prediction of the changes in protein free energies in a large-scale mutation scan. *Angew. Chem. Int. Ed.*, 55(26):7364–7368, 2016.
- [51] César de Oliveira, Haoyu S Yu, Wei Chen, Robert Abel, and Lingle Wang. Rigorous free energy perturbation approach to estimating relative binding affinities between ligands with multiple protonation and tautomeric states. *J. Chem. Theory Comput.*, 15(1):424–435, January 2019.
- [52] Zoe Cournia, Bryce Allen, and Woody Sherman. Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. *J. Chem. Inf. Model.*, 57(12):2911–2937, December 2017.
- [53] Wei Chen, Yuqing Deng, Ellery Russell, Yujie Wu, Robert Abel, and Lingle Wang. Accurate calculation of relative binding free energies between ligands with different net charges. *J. Chem. Theory Comput.*, November 2018.
- [54] Zhe Li, Yiyu Huang, Yinuo Wu, Jingyi Chen, Deyan Wu, Chang-Guo Zhan, and Hai-Bin Luo. Absolute binding free energy calculation and design of a subnanomolar inhibitor of phosphodiesterase-10. *J. Med. Chem.*, 62(4):2099–2111, February 2019.
- [55] Matteo Aldeghi, Alexander Heifetz, Michael J Bodkin, Stefan Knapp, and Philip C Biggin. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.*, 7(1):207–218, January 2016.
- [56] Yue Qian, Israel Cabeza de Vaca, Jonah Z Vilseck, Daniel J Cole, Julian Tirado-Rives, and William L Jorgensen. Absolute free energy of binding calculations for macrophage migration inhibitory factor in complex with a druglike inhibitor. *J. Phys. Chem. B*, 123(41):8675–8685, October 2019.

- [57] Yuko Okamoto, Hironori Kokubo, and Toshimasa Tanaka. Prediction of ligand binding affinity by the combination of Replica-Exchange method and Double-Decoupling method. *J. Chem. Theory Comput.*, 10(8):3563–3569, August 2014.
- [58] Nanjie Deng, Lauren Wickstrom, Piotr Cieplak, Clement Lin, and Danzhou Yang. Resolving the Ligand-Binding specificity in c-MYC G-Quadruplex DNA: Absolute binding free energy calculations and SPR experiment. *J. Phys. Chem. B*, 121(46):10484–10497, November 2017.
- [59] Niaz Mahmood, Catalin Mihalciou, and Shafaat A Rabbani. Multifaceted role of the Urokinase-Type plasminogen activator (uPA) and its receptor (uPAR): Diagnostic, prognostic, and therapeutic applications. *Front. Oncol.*, 8:24, February 2018.
- [60] B A Katz, R Mackman, C Luong, K Radika, A Martelli, P A Sprengeler, J Wang, H Chan, and L Wong. Structural basis for selectivity of a small molecule, s1-binding, submicromolar inhibitor of urokinase-type plasminogen activator. *Chem. Biol.*, 7(4):299–312, April 2000.
- [61] B A Katz, K Elrod, C Luong, M J Rice, R L Mackman, P A Sprengeler, J Spencer, J Hataye, J Janc, J Link, J Litvak, R Rai, K Rice, S Sideris, E Verner, and W Young. A novel serine protease inhibition motif involving a multi-centered short hydrogen bonding network at the active site. *J. Mol. Biol.*, 307(5):1451–1486, April 2001.
- [62] Bradley A Katz, Kyle Elrod, Erik Verner, Richard L Mackman, Christine Luong, William D Shrader, Martin Sendzik, Jeffrey R Spencer, Paul A Sprengeler, Aleks Kolesnikov, Vincent W-F Tai, Hon C Hui, J Guy Breitenbucher, Darin Allen, and James W Janc. Elaborate manifold of short hydrogen bond arrays mediating binding of active site-directed serine protease inhibitors. *J. Mol. Biol.*, 329(1):93–120, May 2003.
- [63] 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(1):235–242, January 2000.
- [64] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The PDB-bind database: methodologies and updates. *J. Med. Chem.*, 48(12):4111–4119, June 2005.

- [65] Ramu Anandakrishnan, Boris Aguilar, and Alexey V Onufriev. H++ 3.0: automating pk prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.*, 40(Web Server issue):W537–41, July 2012.
- [66] Christopher I Bayly, Piotr Cieplak, Wendy Cornell, and Peter A Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.*, 97(40):10269–10280, October 1993.
- [67] Mjhea Frisch, G W Trucks, Hs B Schlegel, G E Scuseria, M A Robb, J R Cheeseman, G Scalmani, V Barone, B Mennucci, Gaeo Petersson, and Others. Gaussian 09, revision a. 02, gaussian. Inc. , Wallingford, CT, 200:28, 2009.
- [68] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, July 2004.
- [69] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, August 2015.
- [70] William L Jorgensen, Jayaraman Chandrasekhar, Jeffrey D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926, July 1983.
- [71] Romelia Salomon-Ferrer, Andreas W Götz, Duncan Poole, Scott Le Grand, and Ross C Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. *J. Chem. Theory Comput.*, 9(9):3878–3888, September 2013.
- [72] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103(19):8577–8593, November 1995.

- [73] Thomas Steinbrecher, David L Mobley, and David A Case. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J. Chem. Phys.*, 127(21):214108, December 2007.
- [74] T Steinbrecher, I S Joung, and others. Soft-core potentials in thermodynamic integration: Comparing one-and two-step transformations. *Journal of computational*, 2011.
- [75] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129(12):124105, September 2008.
- [76] Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *J. Comput. Phys.*, 22(2):245–268, October 1976.
- [77] Daniel R Roe and Thomas E Cheatham, 3rd. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.*, 9(7):3084–3095, July 2013.
- [78] S van der Walt, S C Colbert, and G Varoquaux. The NumPy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2):22–30, March 2011.
- [79] B Roux, M Nina, R Pomès, and J C Smith. Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study. *Biophys. J.*, 71(2):670–681, August 1996.
- [80] G Mann and J Hermans. Modeling protein-small molecule interactions: structure and thermodynamics of noble gases binding in a cavity in mutant phage T4 lysozyme L99A. *J. Mol. Biol.*, 302(4):979–989, September 2000.
- [81] Stefan Boresch, Franz Tetteringer, Martin Leitgeb, and Martin Karplus. Absolute binding free energies: A quantitative approach for their calculation. *J. Phys. Chem. B*, 107(35):9535–9551, September 2003.
- [82] David L Mobley, John D Chodera, and Ken A Dill. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *J. Chem. Phys.*, 125(8):084902, August 2006.

- [83] Malcolm E Davis and J Andrew McCammon. Electrostatics in biomolecular structure and dynamics. *Chem. Rev.*, 90(3):509–521, May 1990.
- [84] K A Sharp and B Honig. Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Biophys. Chem.*, 19:301–332, 1990.
- [85] Arald Jean-Charles, Anthony Nicholls, Kim Sharp, Barry Honig, Anna Tempczyk, Thomas F Hendrickson, and W Clark Still. Electrostatic contributions to solvation energies: comparison of free energy perturbation and continuum calculations. *J. Am. Chem. Soc.*, 113(4):1454–1455, February 1991.
- [86] B Honig and A Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, May 1995.
- [87] M K Gilson. Theory of electrostatic interactions in macromolecules. *Curr. Opin. Struct. Biol.*, 5(2):216–223, April 1995.
- [88] Dmitrii Beglov and Benoît Roux. Solvation of complex molecules in a polar liquid: An integral equation theory. *J. Chem. Phys.*, 104(21):8678–8689, June 1996.
- [89] Christopher J Cramer and Donald G Truhlar. Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.*, 99(8):2161–2200, August 1999.
- [90] D Bashford and D A Case. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, 51:129–152, 2000.
- [91] Nathan A Baker. Improving implicit solvent simulations: a poisson-centric view. *Curr. Opin. Struct. Biol.*, 15(2):137–143, April 2005.
- [92] Jianhan Chen, Wonpil Im, and Charles L Brooks, 3rd. Balancing solvation and intramolecular interactions: toward a consistent generalized born force field. *J. Am. Chem. Soc.*, 128(11):3728–3736, March 2006.

- [93] Michael Feig, Jana Chocholoušová, and Seiichiro Tanizaki. Extending the horizon: towards the efficient modeling of large biomolecular complexes in atomic detail. *Theor. Chem. Acc.*, 116(1):194–205, August 2006.
- [94] Wonpil Im, Jianhan Chen, and Charles L Brooks, 3rd. Peptide and protein folding and conformational equilibria: theoretical treatment of electrostatics and hydrogen bonding with implicit solvent models. *Adv. Protein Chem.*, 72:173–198, 2005.
- [95] B Z Lu, Y C Zhou, M J Holst, and J A McCammon. Recent progress in numerical methods for the Poisson-Boltzmann equation in biophysical applications. *Commun. Comput. Phys.*, 3(5):973–1009, 2008.
- [96] Jun Wang, Chunhu Tan, Yu-Hong Tan, Qiang Lu, and Ray Luo. Poisson-Boltzmann solvents in molecular dynamics simulations. *Commun. Comput. Phys.*, 3(5):1010–1031, 2008.
- [97] Michael D Altman, Jaydeep P Bardhan, Jacob K White, and Bruce Tidor. Accurate solution of multi-region continuum biomolecule electrostatic problems using the linearized Poisson-Boltzmann equation with curved boundary elements. *J. Comput. Chem.*, 30(1):132–153, January 2009.
- [98] Qin Cai, Jun Wang, Meng-Juei Hsieh, Xiang Ye, and Ray Luo. Annual reports in computational chemistry, 2012.
- [99] Wesley M Botello-Smith and Ray Luo. Applications of MMPBSA to membrane proteins i: Efficient numerical solutions of periodic Poisson–Boltzmann equation. *J. Chem. Inf. Model.*, 55(10):2187–2199, October 2015.
- [100] Li Xiao, Changhao Wang, and Ray Luo. Recent progress in adapting Poisson–Boltzmann methods to molecular simulations. *J. Theor. Comput. Chem.*, 13(03):1430001, May 2014.
- [101] Edward Z Wen, Meng-Juei Hsieh, Peter A Kollman, and Ray Luo. Enhanced ab initio protein folding simulations in Poisson–Boltzmann molecular dynamics with self-guiding forces. *J. Mol. Graph. Model.*, 22(5):415–424, May 2004.

- [102] Thu Zar Lwin, Ruhong Zhou, and Ray Luo. Is Poisson-Boltzmann theory insufficient for protein folding simulations? *J. Chem. Phys.*, 124(3):034902, January 2006.
- [103] Jun Wang, Chunhu Tan, Emmanuel Chanco, and Ray Luo. Quantitative analysis of Poisson-Boltzmann implicit solvent in molecular dynamics. *Phys. Chem. Chem. Phys.*, 12(5):1194–1202, 2010.
- [104] Changhao Wang, Jun Wang, Qin Cai, Zhilin Li, Hong-Kai Zhao, and Ray Luo. Exploring accurate Poisson-Boltzmann methods for biomolecular simulations. *Computational and Theoretical Chemistry*, 1024:34–44, November 2013.
- [105] Changhao Wang, Pengyu Ren, and Ray Luo. Ionic solution: What goes right and wrong with continuum solvation modeling. *J. Phys. Chem. B*, 121(49):11169–11179, December 2017.
- [106] David A Case, Thomas E Cheatham, 3rd, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, December 2005.
- [107] Chunhu Tan, Yu-Hong Tan, and Ray Luo. Implicit nonpolar solvent models. *J. Phys. Chem. B*, 111(42):12263–12274, October 2007.
- [108] Qin Cai, Meng-Juei Hsieh, Jun Wang, and Ray Luo. Performance of nonlinear finite-difference Poisson-Boltzmann solvers. *J. Chem. Theory Comput.*, 6(1):203–211, January 2010.
- [109] Qiang Lu and Ray Luo. A Poisson-Boltzmann dynamics method with nonperiodic boundary condition. *J. Chem. Phys.*, 119(21):11035–11047, December 2003.
- [110] Matthew D Liptak, Kevin C Gross, Paul G Seybold, Steven Feldgus, and George C Shields. Absolute pka determinations for substituted phenols. *J. Am. Chem. Soc.*, 124(22):6421–6427, June 2002.
- [111] A Warshel and A Papazyan. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol.*, 8(2):211–217, April 1998.

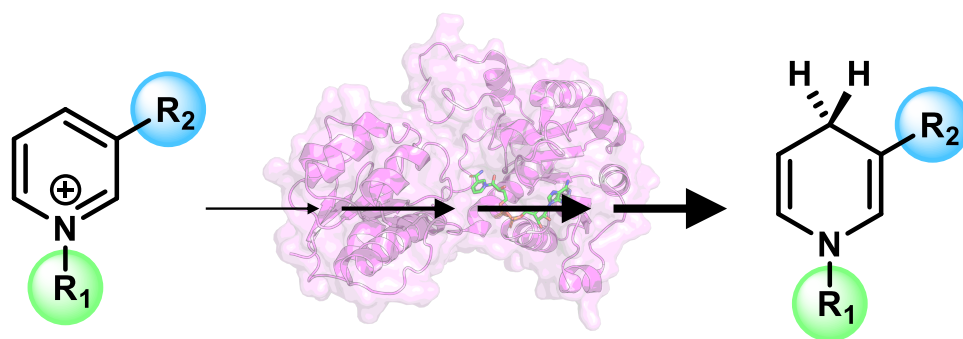
- [112] C N Schutz and A Warshel. What are the dielectric “constants” of proteins and how to validate electrostatic models? *Proteins: Struct. Funct. Bioinf.*, 2001.
- [113] H Gouda, I D Kuntz, D A Case, and others. Free energy calculations for theophylline binding to an RNA aptamer: comparison of MM-PBSA and thermodynamic integration methods. : *Original Research on . . .*, 2003.
- [114] 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, 3rd. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33(12):889–897, December 2000.
- [115] Jacqueline Bezençon, Matthias B Wittwer, Brian Cutting, Martin Smieško, Bjoern Wagner, Manfred Kansy, and Beat Ernst. pka determination by ^1H NMR spectroscopy—an old methodology revisited. *J. Pharm. Biomed. Anal.*, 93:147–155, 2014.
- [116] D Khare, P Alexander, J Antosiewicz, P Bryan, M Gilson, and J Orban. pka measurements from nuclear magnetic resonance for the B1 and B2 immunoglobulin g-binding domains of protein g: comparison with calculated values for nuclear magnetic resonance and x-ray structures. *Biochemistry*, 36(12):3580–3589, March 1997.
- [117] Christos P Papanephytou, Asterios I Grigoroudis, Campbell McInnes, and George Kontopidis. Quantification of the effects of ionic strength, viscosity, and hydrophobicity on Protein–Ligand binding affinity. *ACS Med. Chem. Lett.*, 5(8):931–936, August 2014.
- [118] Frank R Beierlein, Julien Michel, and Jonathan W Essex. A simple QM/MM approach for capturing polarization effects in protein-ligand binding free energy calculations. *J. Phys. Chem. B*, 115(17):4911–4926, May 2011.
- [119] Guillaume Lamoureux and Benoit Roux. Modeling induced polarization with classical drude oscillators: Theory and molecular dynamics simulation algorithm. *J. Chem. Phys.*, 119(6): 3025–3039, August 2003.

- [120] Boris Aguilar, Ramu Anandakrishnan, Jory Z Ruscio, and Alexey V Onufriev. Statistics and physical origins of pK and ionization state changes upon protein-ligand binding. *Biophys. J.*, 98(5):872–880, March 2010.
- [121] Delphine C Bas, David M Rogers, and Jan H Jensen. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins*, 73(3):765–783, November 2008.
- [122] A D MacKerell, Jr, M S Sommer, and M Karplus. pH dependence of binding reactions from free energy simulations and macroscopic continuum electrostatic calculations: application to 2 GMP/3 GMP *J. Mol. Biol.*, 1995.
- [123] Alexey V Onufriev and Emil Alexov. Protonation and pK changes in protein–ligand binding. *Q. Rev. Biophys.*, 46(2):181–209, May 2013.
- [124] M Olivia Kim and J Andrew McCammon. Computation of pH-dependent binding free energies. *Biopolymers*, 105(1):43–49, January 2016.
- [125] Wei Chen, Brian H Morrow, Chuanyin Shi, and Jana K Shen. Recent development and application of constant pH molecular dynamics. *Mol. Simul.*, 40(10-11):830–838, January 2014.
- [126] Atanu Maiti and Alexander C Drohat. Dependence of substrate binding and catalysis on pH, ionic strength, and temperature for thymine DNA glycosylase: Insights into recognition and processing of GT mismatches. *DNA Repair*, 10(5):545–553, May 2011.
- [127] Mathilde J Kaas Hansen, Johan G Olsen, Sophie Bernichtein, Charlotte O’Shea, Bent W Sigurskjold, Vincent Goffin, and Birthe B Kragelund. Development of prolactin receptor antagonists with reduced pH-dependence of receptor binding. *J. Mol. Recognit.*, 24(4):533–547, July 2011.
- [128] Kemper Talley and Emil Alexov. On the pH-optimum of activity and stability of proteins. *Proteins*, 78(12):2699–2706, September 2010.
- [129] Jan H Jensen. Calculating pH and salt dependence of protein-protein binding. *Curr. Pharm. Biotechnol.*, 9(2):96–102, April 2008.

- [130] Aaron C Mason and Jan H Jensen. Protein-protein binding is often associated with changes in protonation state. *Proteins*, 71(1):81–91, April 2008.

Chapter 3

Engineering natural and noncanonical nicotinamide cofactor-dependent enzymes: design principles and technology development



Natural and noncanonical nicotinamide cofactor-dependent enzymes

Figure 3.1: Cofactor engineering graphical abstract

Authors: Edward King*, Sarah Maxel*, Han Li

Curr Opin Biotechnol. 2020;66: 217–226.

doi: [10.1016/j.copbio.2020.08.005](https://doi.org/10.1016/j.copbio.2020.08.005)

Publication Date (Web): September 18, 2020

3.1 Abstract

Nicotinamide cofactors enable oxidoreductases to catalyze a myriad of important reactions in biomanufacturing. Decades of research has focused on optimizing enzymes which utilize natural nicotinamide cofactors, namely nicotinamide adenine dinucleotide (phosphate) (NAD(P)⁺). Recent findings reignite the interest in engineering enzymes to utilize noncanonical cofactors, the mimetics of NAD⁺ (mNADs), which exhibit superior industrial properties *in vitro* and enable specific electron delivery *in vivo*. We compare recent advances in engineering natural versus non-canonical cofactor-utilizing enzymes, discuss design principles discovered, and survey emerging high-throughput platforms beyond the traditional 96-well plate-based methods. Obtaining mNAD-dependent enzymes remains challenging with a limited toolkit. To this end, we highlight design principles and technologies which can potentially be translated from engineering natural to non-canonical cofactor-dependent enzymes.

3.2 Introduction

Nicotinamide cofactor-utilizing enzymes are versatile catalysts for both *in vitro* chemical synthesis and *in vivo* metabolic engineering. Although more than 15,000 sequences have been confirmed or predicted to encode NAD(P)⁺ or NAD(P)H utilizing enzymes[1], natural enzymes frequently do not meet the catalytic needs of compatibility with the metabolism of a chassis host *in vivo* and viability at large scales *in vitro*, and often require engineering of the enzyme's natural cofactor specificity, substrate scope, and robustness.

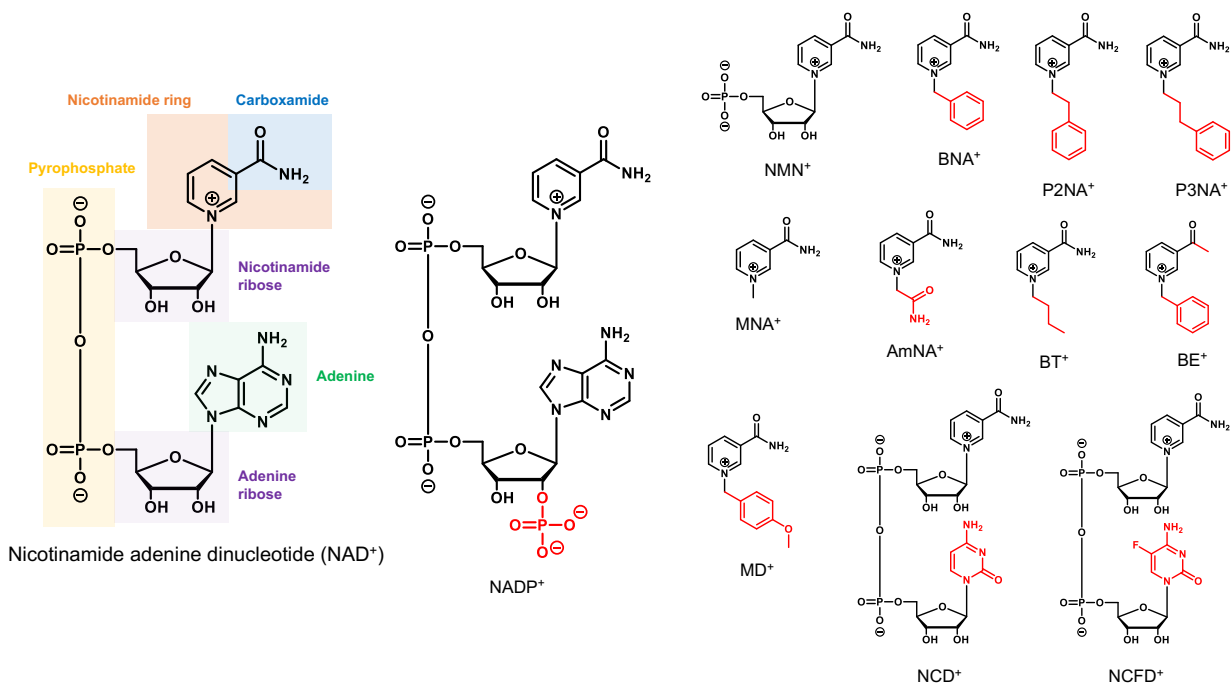


Figure 3.2: Chemical structures of natural nicotinamide redox cofactors and mNADs. The natural nicotinamide redox cofactor NAD⁺ is composed of the catalytic nicotinamide ring attached to a ribose, pyrophosphate, a second ribose, and adenine base. NADP⁺ differs in that a phosphate group replaces a hydroxyl on the 2'-carbon of the adenine ribose. mNADs maintain the central nicotinamide ring, but are truncated or incorporate alternative functional groups compared to NAD⁺. The cofactors are illustrated in their oxidized form, complete cofactor names are listed under Table 3.1, Table 3.2.

Recent studies highlight the value of engineering these enzymes to use noncanonical nicotinamide cofactors, which are mimetics of NAD^+ (mNADs). The nicotinamide ring is the only fragment required for a small molecule to function as a redox cofactor[2–4]. Molecules with alternative functional groups replacing the NAD(P)/H carboxamide[2], the adenine base[5–7], the nicotinamide ribose[8], and mimics truncated at different atoms have been explored as artificial redox cofactors[9, 10] (Figure 3.2). These mimics have industrial value for lowering feedstock costs as mNADs are often simpler to synthesize[11] and have greater stability than native cofactors[12], permit access to new chemistries with altered redox potential[13], reduce oxygenase decoupling[14, 15], and importantly enable specific delivery of electrons in both whole cells and crude cell lysates[9, 11, 16]. Metabolic pathways engineered to specifically utilize mNADs are orthogonal from the host’s metabolism, as they do not cross-talk with native pathways which only use natural cofactors. This allows precise control of chemical reactions in the cells without interference and has been demonstrated in vivo for the production of malate from the carboxylation of pyruvate via nicotinamide cytosine dinucleotide (NCD^+)[6, 16], and selective generation of the pharmaceutical intermediate levodione by nicotinamide mononucleotide (NMN^+) mediated reduction[9].

A number of approaches have been used to design proteins, but all show limited success rates and ultimately reduce to repeated trial and error experiments. A widely applied method for protein design is directed evolution (DE)[17], which combines steps of: 1) generating diverse variants, 2) assaying the pool of mutants for activity, and 3) selecting the most fit samples for subsequent rounds of design to mimic natural evolution of proteins toward a defined function. DE can be performed as a blind search with minimal knowledge of the protein. This is in contrast to rational design methods[18] where focused changes are made based on bioinformatic and structural data to minimize the experimental effort required in screening.

3.3 Design principles in engineering natural cofactor-dependent enzymes

Many general design principles are derived from decades of research in engineering NAD(P)/H-dependent enzymes. For example, altering specificity between natural cofactors commonly relies on the mutagenesis of binding pocket amino acids interacting with the signature 2'-phosphate or 2'-hydroxyl groups that differentiate NADP(H) and NAD(H), respectively. Fundamental semi-rational design rules have been captured by an easy-to-use web tool Cofactor Specificity Reversal-Structural Analysis and Library Design (CSR-SALAD)[19]. This computational method incorporates structural activity and genetic information to automatically design focused libraries[20]; however, it has met with limited success for enzymes that utilize cofactors in complex reaction mechanisms[19, 21]. Flexible loop grafting has emerged as another design principle in engineering the cofactor preference of TIM barrel oxidoreductases. Swapping of cofactor binding loops between homologous ene reductases (ERs) with NADPH and NADH preference[22] shows promise as a means of generating flexibility in cofactor preferences. Additional studies on aldo-keto reductase (AKR) inverted cofactor preference from NAD^+ to NADP^+ by inserting either additional residues[23] or a calcium controllable repeats-in-toxin (RTX) domain[24] into substrate binding loops.

Beyond cofactor specificity, complex traits such as stability and conformational dynamics are a challenging task for rational design. Recent reports revealed the effects of modulating the microenvironment surrounding oxidoreductases, which can potentially be a universal design principle in engineering both NAD(P)/H and mNAD-dependent enzymes. For example, fusing a variant of superfolder green fluorescent protein (sfGFP) with extreme surface charges enhanced the activity of AKR, possibly by influencing the apparent ionic strength of the active site[25]. Furthermore, increased cofactor availability has been explored with DNA-enzyme nanostructures[26] acting as local reservoirs of cofactors, fusing of redox cycling partners by co-expression[27–31], and directly tethering NAD(H) to proteins with polyethylene glycol chains[32].

3.4 Design principles in engineering noncanonical cofactor-dependent enzymes

We summarize efforts in enhancing mNAD catalysis and evaluate the extent of success through the following metrics[33] (Tables: 3.1, 3.2).

(1) Coenzyme Specificity Ratio (CSR) (Equation: 3.1), a measure of preference for the mNAD over natural cofactors. While most wild type enzymes use mNADs very poorly, as reflected by near zero CSR, many flavoenzymes including enoate reductases, nitroreductases, and para-hydroxybenzoate hydroxylase exhibit promising activities[9, 15, 34] (Table: 3.1). In particular, the xenobiotic reductase from *Pseudomonas putida* (*P. putida* XenA) utilizes a range of mNADs more efficiently than natural cofactors[34] (Table: 3.1). High CSR is desirable for creating orthogonal redox circuitry[6, 9, 16]; however, while most studies only consider NAD^+ , it is important to measure CSR for both NAD^+ and NADP^+ [9] when determining orthogonality *in vivo*.

$$CSR = \frac{\left(\frac{k_{cat}}{K_m}\right)_{mNAD}}{\left(\frac{k_{cat}}{K_m}\right)_{NAD(P)}} \quad (3.1)$$

(2) Relative Catalytic Efficiency (RCE) (Equation 3.2), is the ratio of the mutant’s catalytic efficiency with mNAD compared to wild type with native cofactor. Since wild type enzymes have been optimized by Nature with its native cofactor, RCE essentially indicates how effective the engineering approaches are compared to natural evolution. RCEs for mNADs are extremely low for most engineered enzymes, indicating that catalytic activities are a small fraction of the wild type with native cofactor. We note that P450-BM3 R966D-W1046S reported by Lo *et al.*[35] had an exceptional RCE of ~ 96 for 1-benzyl-1,4-dihydronicotinamide (BNAH) and ~ 60 for N-4-methoxybenzyl-1,4-dihydronicotinamide (MDH) (Table: 3.2). In comparison, RCEs of >1 are frequently achieved in switching NAD^+ and NADP^+ specificity[33].

$$RCE = \frac{\left(\frac{k_{cat}}{K_m}\right)_{mNAD}^{mut}}{\left(\frac{k_{cat}}{K_m}\right)_{NAD(P)}^{WT}} \quad (3.2)$$

(3) Relative Specificity (RS) (Equation: 3.3), the CSR of a variant compared to that of the wild type, which is often referred to as the fold of cofactor specificity switch toward noncanonical cofactors. This parameter is useful for comparing the effectiveness of different engineering approaches in general, independently of the specific enzymes targeted (Table: 3.2).

$$RS = \frac{\left(\frac{\left(\frac{k_{cat}}{K_m}\right)_{mNAD}}{\left(\frac{k_{cat}}{K_m}\right)_{NAD(P)}}\right)^{mut}}{\left(\frac{\left(\frac{k_{cat}}{K_m}\right)_{mNAD}}{\left(\frac{k_{cat}}{K_m}\right)_{NAD(P)}}\right)^{WT}} \quad (3.3)$$

Enzyme	Uniprot	Reaction	Native Cofactor	noncanonical Cofactor ^a	CSR ^b	Source
D-lactate dehydrogenase (<i>L. helveticus</i>)	P30901	Red	NAD	NCD	0.05	Liu et al, 2020[5]
Formate dehydrogenase (<i>Pseudomonas sp.</i> 101)	P33160	Red	NAD	NCD	0.05	Guo et al, 2020[6]
Glucose Dehydrogenase (<i>B. subtilis</i>)	A0A1B2ATD9	Red	NADP	NMN	0	Black et al, 2019[9]
			NAD	NMN	0	
6-phosphogluconate dehydrogenase (<i>T. maritima</i>)	A0A2N5RRL69	Red	NADP	NMN	0	Huang et al, 2019[30]
Phosphite dehydrogenase (<i>Ralstonia sp.</i> 4506)	G4XDR8	Red	NAD	NCD	0.01	Liu et al, 2019[5]

3-hydroxy benzoate 6-hydroxylase (<i>R. jostii</i>)	Q0SFK6	Ox		NADH	BNAH	0	Guarneri et al., 2019[15]
					AmNAH	0	
				NADPH	BNAH	0	
					AmNAH	0	
para-Hydroxybenzoate hydroxylase (<i>P. fluorescens</i>)	P20586	Ox		NADH	AmNAH	0	
				NADPH	AmNAH	2.09	
Salicylate hydroxylase (<i>P. putida</i>)	-	Ox		NADH	BNAH	0	
				NADPH	BNAH	0.88	
Glucose dehydrogenase (<i>S. solfataricus</i>)	O93715	Red		NAD	BNA	0.01	Nowak et al., 2017[8]
					P2NA	0.02	

Pentaerythritol tetranitrate reductase (<i>E. cloacae</i>)	P71278	Ox	NADPH	BNAH	1.43	
				BTCH	5.22	
				BTEH	0.1	
Thermophilic old yellow enzyme (<i>T. pseudethanolicus</i>)	B0KAH1	Ox	NADH	BNAH	0.82	
				BTCH	1.2	
				BTEH	0.33	
			NADPH	BNAH	0.02	
				BTCH	0.03	
				BTEH	0	
Styrene monooxygenase (<i>R. opacus</i>)	C7ACG0	Ox	NADH	BNAH	-	Paul et al., 2015[36]

Alcohol dehydrogenase (<i>P. furiosus</i>)	-	Red	NAD	NMN	0	Campbell et al., 2012[37]
Malic enzyme (<i>E. coli</i>)	P26616	Red	NAD	NFCD	0.01	Ji et al., 2011[38]
				NCD	0.01	

Table 3.1: Performance of wild type enzymes with noncanonical cofactors.

^aFull names of the noncanonical cofactors: AmNA⁺, 1-(2-carbamoylmethyl)-1,4-dihydronicotinamide; BNA⁺, 1-benzyl-1,4-dihydronicotinamide; BT⁺, 1-butyl-1,4-dihydronicotinamide; BE⁺, 1-(1-benzyl-1,4-dihydro-3-yl) ethanone; MD⁺, N-4-methoxybenzyl-1,4-dihydronicotinamide; MNA⁺, 1-methyl-1,4-dihydropyridine-3-carboxamide; NCD⁺, Nicotinamide cytosine dinucleotide; NCFD⁺, Nicotinamide flucytosine dinucleotide; NMN⁺, Nicotinamide mononucleotide; P2NA⁺, 1-phenethyl-1,4-dihydropyridine-3-carboxamide; P3NA⁺, 1-(3-phenylpropyl)-1,4-dihydropyridine-3-carboxamide. Reduced cofactor ends with ‘H’.

^bCSR, Cofactor Specificity Ratio (Equation: 3.1)

Enzyme	Uniprot	Strategy	Reaction	Native Cofactor	Noncanonical Cofactor ^d	Mutations ^b	CSR ^b	RCE ^c	Log RS ^d	Source
D-lactate dehydrogenase (<i>L. helveticus</i>)	P30901	Bibliography Saturation, Structure	Red	NAD	NCD	V152R- N213E	42.86	0.21	2.96	Liu et al., 2020 [7]
						V152R- I177K- N213I	41.89	0.31	2.95	
Formate dehydrogenase (<i>Pseudomonas sp.</i> 101)	P33160	Bibliography Saturation, Structure	Red	NAD	NCD	V198I- C256I- P260S- E261P- S381N- S383F	165.15	0.04	3.55	Guo et al., 2020 [6]

Glucose Dehydrogenase (<i>B. subtilis</i>)	A0A1B2ATD0	Structure	RefSeq	NADP	NMN	Y34Q-A93K-I195R	0.07	0	4.61	Black et al., 2019[9]
						S17E-Y34Q-A93K-I195R	19.09	0	7.06	
				NAD	NMN	Y34Q-A93K-I195R	4.64	0	6.25	
						S17E-Y34Q-A93K-I195R	55.26	0	7.33	

6-phosphogluconate dehydrogenase (<i>T. maritima</i>)	A0A2N5R166	Random, Saturation, Structure	Computation	NADP	NMN	Mut 5-1 ^e	0.01	0	3.66	Huang et al., 2019 [10]
Glucose-6-phosphate dehydrogenase (<i>T. maritima</i>)	A0A2N5R167	Random, Saturation, Structure	Computation	NADP	NMN	Mut 6-1 ^f	0.01	0	3.64	
						A64S-R65I-T66I	-	-	-	

Phosphite dehydrogenase (<i>Ralstonia</i> sp. 4506)	G4XDR8	Bibliography	Computational, Saturation, Structure	Red	NAD	NCD	I151R-P176R	19.79	0.13	2.53	Liu et al., 2019[5]
							I151R-P176R-M207A	45.33	0.01	2.89	
Glucose dehydrogenase (<i>S. solfataricus</i>)	O93715	Bibliography	Saturation, Structure	Red	NAD	BNA	I192T-V306G	5.47	0.07	2.74	Nowak et al., 2017[8]
							I192T-V306I	0.29	0.1	1.46	
						P2NA	I192T-V306G	3.05	0.04	2.31	

P450-BM3 (<i>B. megaterium</i>)	P14779	Bibliography	NADH	BNA	R966D-W1064S	0.1	96.05	-	Lo et al., 2017[35]
		Structure		MDH	R966D-W1064S	0.06	60.47	-	
Alcohol dehydrogenase (<i>P. furiosus</i>)	-	MSA, Structure	NAD	NMN	K249G-H255R	0	0	2.26	Campbell et al., 2012[37]
Malic enzyme (<i>E. coli</i>)	P26616	Computational MSA, Saturation, Structure	NAD	NFCD	L310R	98.22	0.41	4.02	Ji et al., 2011[38]
					L310R-Q401C	268.61	0.45	4.46	
				NCD	L310R	188.78	0.79	4.16	

D-lactate dehydrogenase (<i>L. helveticus</i>)	P30901	MSA, Structure	Red	NAD	NCFD	V152R	-	429.44	0.72	4.52	-	Flores et al., 2005 [39]
Malate dehydrogenase (<i>E. coli</i>)	P61889	MSA, Structure	Red	NAD	NCFD	L6R	-	-	-	-	-	-
Lactate dehydrogenase (<i>B. stearothermophilus</i>)	P00344	MSA, Saturation, Structure	Red	NAD	NMN	C81S-N85R-F16Q	-	-	-	-	-	-

Table 3.2: Performance of engineered enzymes with noncanonical cofactors.

^aFull names of the noncanonical cofactors: AmNA⁺, 1-(2-carbamoylmethyl)-1,4-dihydronicotinamide; BNA⁺, 1-benzyl-1,4-dihydronicotinamide; BT⁺, 1-butyl-1,4-dihydronicotinamide; BE⁺, 1-(1-benzyl-1,4-dihydro-3-yl) ethanone; MD⁺, N-4-methoxybenzyl-1,4-dihydronicotinamide; MNA⁺, 1-methyl-1,4-dihydropyridine-3-carboxamide; NCD⁺, Nicotinamide cytosine dinucleotide; NCFD⁺, Nicotinamide flucytosine dinucleotide; NMN⁺, Nicotinamide mononucleotide; P2NA⁺, 1-phenethyl-1,4-dihydropyridine-3-carboxamide; P3NA⁺, 1-(3-phenylpropyl)-1,4-dihydropyridine-3-carboxamide. Reduced cofactor ends with ‘H’.

^bCSR, Cofactor Specificity Ratio (Equation: 3.1)

^cRCE, Relative Catalytic Efficiency (Equation: 3.2)

^dRS, Relative Specificity (Equation: 3.3)

^e Mut 5-1 contains A11G-K27R-R33I-T34I-F60Y-D82L-T83L-Q86L-K118N-I120F-D294V-F326S-Y383C-N387S-A447V.

^d Mut 6-1 contains A11G-K27R-R33I-T34I-F60Y-D82L-T83L-Q86L-K118N-I120F-D251E-D294V-F326S-F329Y-Y383C-N387S-V390G-A447V.

Because the number of successful cases is still relatively small, core design principles for switching cofactor specificity toward noncanonical cofactors have yet to clearly emerge. The field still largely relies on semi-rational and random engineering which often yields beneficial mutations with unknown mechanisms. Gaining fundamental understanding on enzyme-mNAD interaction through structural and kinetic studies is crucial to deriving design principles to streamline engineering. Nevertheless, the following trends are notable:

First, relaxation of cofactor specificity is linked to enhanced activity with mNADs. *Bacillus stearothermophilus* lactate dehydrogenase F16Q-C81S-N85R with specificity switched from NAD^+ to NADP^+ was found to reduce NMN^+ with trace activity[39]. The K249G-H255R variant of *Pyrococcus furiosus* alcohol dehydrogenase designed to increase the volume of the active site for NADP^+ binding unexpectedly gained the ability to utilize NMN^+ (Table: 3.2), and showed a 40% increase in maximum current density when used in a biofuel cell, postulated to be due to improved mass transfer of NMN^+ compared to NAD^+ [37]. The P450-BM3 mutant R966D-W1046S (Table: 3.2) capable of using both NADPH and NADH was also able to utilize BNAH for the reduction of cytochrome c with a catalytic efficiency of $41.3 \text{ min}^{-1} \text{ uM}^{-1}$, while the wild type had no detectable activity[35, 40]. A similar variant P450-BM3 W1046S also gained activity for utilizing both natural cofactors and reduced NMN^+ (NMNH)[9].

Second, size reduction of the cofactor binding pocket to improve packing often affords increased activity toward mNADs. For example, the phosphite dehydrogenase from *Ralstonia sp.* 4506 harboring I151R-P176R-M207A mutations had significantly enhanced activity toward NCD^+ . Crystallography suggested activity was achieved through compression of the binding pocket around the smaller cytosine[7]. Interestingly, natural flavoenzymes that efficiently utilize mNADs also employ this strategy. The bulky Trp302 residue in *P. putida* XenA active site adopts a different conformation when smaller mNADs are bound to pack more tightly against the cofactors[34].

Third, design to install polar interactions, which in principle contribute more strongly to binding affinity than hydrophobic packing, is effective for achieving stringent binding of mNADs. We recently engineered a highly orthogonal *Bacillus subtilis* glucose dehydrogenase S17E-Y34Q-A93K-

I195R to use NMN⁺[9] which showed the highest RS (Equation: 3.3) reported to date of 1.1×10^7 for NADP⁺ and 2.1×10^7 for NAD⁺. We first utilized Rosetta modeling to identify the positively charged I195R mutation which is predicted to form a salt bridge with the highly negative NMN⁺ phosphate. Next, we achieved exclusive specificity for NMN⁺ by introducing S17E which is modeled to repel the phosphate in the adenosine monophosphate (AMP) moiety that is only present in the natural cofactors but not in NMN⁺. Because of the high conservation of residues lining the cofactor binding pocket, we hypothesize that these mutations should be readily transferable and support NMN⁺ binding in homologs.

3.5 Technology development for engineering natural cofactor-dependent enzymes

Limited throughput has driven the use of semi-rational strategies to minimize the number of variants screened and to maximize the likelihood of isolating promising candidates. Many of these focused libraries have been screened based on readouts that can be determined by a microplate reader[41, 42] or visualized on an agar plate[43, 44]. Application of a 4-nitrophenylacetonitrile microplate assay provided a colorimetric screen to isolate cytochrome P450-BM3 variants for hydroquinone production with 70-fold improvement over wild type activity[41] (Figure: 3.3). In another example, an agar screen leveraged the solubility difference of the substrate and product to evolve the substrate scope of a cyclohexanone monooxygenase (CHMO) for pilot-scale applications[44, 45]. For enzymes that do not produce color or absorbance change during catalysis, a mass spectrometry-based screening platform was developed (Figure: 3.3) to use ‘click’ chemistry to enhance throughput[46] This mass spectrometry-based platform may be readily applicable to engineering mNAD-dependent enzymes. Despite success, throughput remains limiting (10^3 – 10^5); furthermore, reduced library sizes may miss potential cooperative effects critical for dramatic improvements[9].

Recent campaigns apply ‘ultra-throughput’ ($>10^6$) methods using reactions that can be detected by fluorescence sorting[47–49]. For example, *Brevibacterium oxydans* cyclohexylamine oxidase (*Bo*

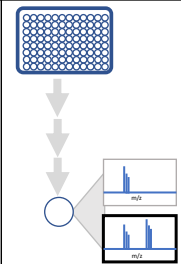
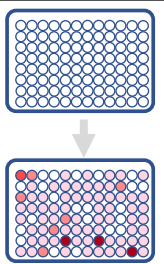
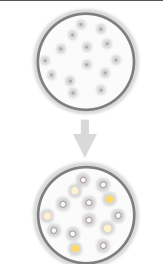
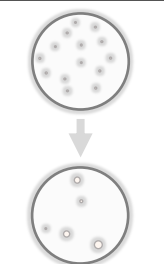
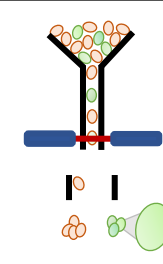
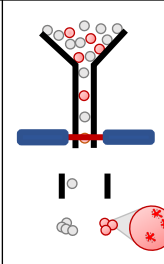
	PECAN (Mass Spectrometry)	NpCN Assay (Microplate Plate)	NESPA Assay (Agar Plate Screen)	Redox Balance Growth Assay (Plate Selection)	Redox Sensor SoxR (FACS)	HRP Assay (FADS)
Enzyme	P450 _{BM3}	P450 _{BM3}	<i>Tm</i> 6PGDH	<i>Ld</i> ldh	<i>Lb</i> adh	<i>Bo</i> CHAO
Throughput	~10 ³	10 ² -10 ⁴	10 ⁴ -10 ⁵	>10 ⁶	>10 ⁶	>10 ⁶
Readout	m/z	UV-vis	Digital Imaging	Growth	Fluorescence (GFP)	Fluorescence (Amplex Red)
Screening Process						
Benefits	Improves Mass Spectrometry Throughput	Simple and common	Sensitive to non-canonical cofactors	Low-cost, broad application	Rapid, broad application	Rapid
Limitations	Multi-step processing and substrate analog required	Requires colorimetric substrate analog	Multi-step Processing	Substrate permeability and toxicity	Substrate permeability and toxicity, high background	Requires H ₂ O ₂ production
Reference	de Rond et al. (2019)	Weingartner et al. (2018)	Huang et al. (2019)	Zhang et al. (2018)	Spielmann et al. (2020)	Debon et al. (2019)

Figure 3.3: Representative screening methods used to facilitate the directed evolution of oxidoreductases. PECAN (probing enzymes with click-assisted NIMS), NpCN (4-nitrophenylacetonitrile), NESPA (NAD(P)-eliminated solid-phase assay), soxR (Redox-sensitive transcriptional activator), FACS (Fluorescence Activated Cell Sorting), HRP (Horse radish peroxidase), and FADS (Fluorescence Activated Droplet Sorting). Targeted Oxidoreductases: P450_{BM3} (NADPH-dependent Cytochrome P450 BM3), *Tm* 6PGDH (NADP⁺-dependent *Thermotoga maritima* 6-phosphogluconate dehydrogenase), *Ld* ldh (NADH-dependent *Lactobacillus delbrueckii* d-lactate dehydrogenase), *Lb* adh (NADPH-dependent *Lactobacillus brevis* alcohol dehydrogenase), *Bo* CHAO (FADH₂-dependent *Brevibacterium oxydans* cyclohexamine oxidase).

CHAO) variants were compartmentalized in droplets and screened for their activity towards a non-natural substrate using fluorescent activated droplet sorting (FADS) (Figure: 3.3), which yielded a mutant with 960-fold increased catalytic efficiency[49]. However, this method is only applicable to enzymes that produce H₂O₂ which is detected by a fluorescent dye, Amplex-Ultra Red. To overcome this limitation, an *Escherichia coli* strain harboring SoxR-regulated GFP cassette to report the intracellular NADPH/NADP⁺ ratio was developed to screen NADPH-dependent enzymes via fluorescent activated cell sorting (FACS)[47, 48]. This system enabled screening of a random library and isolated a *Lactobacillus brevis* alcohol dehydrogenase variant with improved activity

for the reduction of 2,5-hexanedione to (2R,5R)-hexanediol[48] (Table: 3.3). Advanced sorting techniques offer rapid screening of the library being explored, but are often hindered by narrow dynamic ranges and high background signal. Selections, as opposed to screens, do not rely on special instrumentation and automatically eliminate undesirable candidates.

In vivo selection platforms modulate cell growth by disrupting intracellular cofactor cycling within engineered *E. coli* strains. These platforms were pioneered in early work aiming to accumulate NADH in anaerobic condition by disrupting the host’s native fermentative pathways, for example in strain JCL166 ($\Delta adhE \Delta ldhA \Delta fr$). In this strain, anaerobic growth is only restored when an NADH-recycling enzyme is present. This system has identified endogenous *E. coli* enzymes which form a 2,3-butanediol production pathway[50]. The same principle of cofactor recycling is the foundation for a variety of ultra-high throughput ($>10^6$) growth-based selections of nicotinamide-dependent oxidoreductases in directed evolution[51, 51–54]. A recent growth-based selection strain has an engineered NADPH-dependent glycolysis, and therefore required a NADPH-consuming ‘fermentative’ reaction to grow anaerobically. This platform enabled the selection of $\sim 6.2 \times 10^7$ variants in one round, and produced a *Lactobacillus delbrueckii* d-lactate dehydrogenase with a 470-fold increase in activity with NADPH[52] (Figure: 3.3).

This selection strategy has since been expanded to include both NADPH and NADH-dependent selections in aerobic conditions, to be compatible with engineering oxygenases such as p-hydroxybenzoate hydroxylase[55] and cyclohexanone monooxygenase[56]. These results highlight the usefulness of *in vivo* growth platforms for oxidoreductase selections.

Growth selection has not been applied in engineering noncanonical cofactor-dependent enzymes. However, our recent work where *E. coli* growth was obligately linked to the cycling of the non-canonical cofactor NMN^+ presents a platform for future studies. This was achieved by disrupting standard glycolysis networks and directing glucose entry into the life-essential carbon metabolism through our NMN^+ -specific glucose dehydrogenase (GDH)[9]. Cell growth was only restored when the NMN^+ -cycling partner of GDH was present to complete the NMN^+ based redox cycle and prevent cofactor depletion. The specific function of the partner is not linked to cell survival and

we anticipate that the complementary partner can be exchanged.

3.6 Technology development for engineering noncanonical cofactor-dependent enzymes

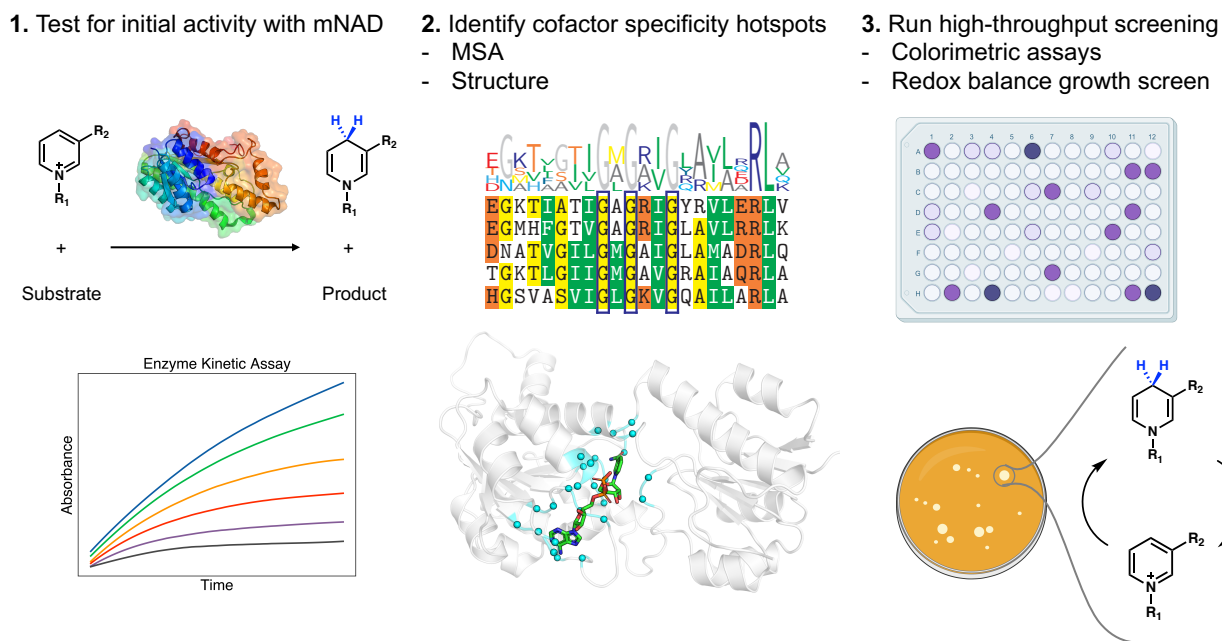


Figure 3.4: Outline of protocol to engineer enzymes for mNAD activity. An initial screen is performed with the mNAD of interest and wild type enzyme to determine the baseline performance. Positions surrounding the cofactor binding pocket and those that contribute to cofactor specificity found through sequence alignment are chosen for mutagenesis. Variants are screened through colorimetric assay measuring activity through color development reflecting production of reduced cofactor. Future tool developments to improve throughput will include growth-based selection assays where the ability of the cell to regenerate mNAD is linked to survival.

In general, efforts in engineering mNAD-dependent enzymes follow three steps (Figure: 3.4): First, wild type enzymes from different organisms are screened with the mNAD of interest to identify a starting template. Second, sequence alignment or computational models predict positions surrounding the cofactor binding pocket. Third, identified positions are targeted by mutagenesis,

often in combinatorial fashion. To achieve high diversity, site-saturation mutagenesis is typically performed with degenerate primers, and variants are screened with 96-well plate-based absorbance assays detecting reduced cofactor[7, 8] or colorimetric assays detecting reactions of reduced cofactor with nitroblue tetrazolium and phenazine methosulfate producing the purple dye formazan[5, 57]. Future tool developments will include growth-based selection where the ability of the cell to cycle the target mNAD is linked to life-essential functions such as carbon metabolism. Variants with more active mNAD cycling will readily outcompete those with lower fitness resulting in facile, high-throughput selection of mNAD-dependent enzymes through readout of cell growth.

We highlight two recent reports departing from the standard saturation mutagenesis and 96-well plate based screening approach. Huang et al.[10] developed the NAD(P)-eliminated solid phase assay (NESPA) (Figure: 3.3), a colorimetric screen performed with colonies grown on agar plate advancing throughput to over 10^5 samples per round while managing low background noise. A heat treatment step is performed to permeabilize cells, followed by washing to remove endogenous NAD(P)⁺. Rounds of saturation mutagenesis at the cofactor binding site, followed by error prone PCR to raise diversity in more distal regions, resulted in a *Thermotoga maritima* 6-phosphogluconate dehydrogenase variant with 50-fold enhanced NMN⁺-dependent activity. However, heat treatment limits the assay to screening thermostable enzymes, and manual washing steps may lead to high variance. In our recent study[9], *in silico* screening was performed in lieu of experimental screening. Bioinformatic analysis was used to identify positions with high plasticity to tolerate mutations. Then, by simulating the effects of mutations on the mNAD binding pose using Rosetta, we greatly narrowed down candidates that warranted experimental testing and eliminated the need to broadly sample with site-saturation mutagenesis. The best mutant *B. subtilis* glucose dehydrogenase S17E-Y34Q-A93K-I195R was obtained from just experimentally testing <20 candidates.

3.7 Conclusion

When engineering enzymes to utilize noncanonical cofactors, even the most developed variants often show low relative catalytic efficiencies with mNADs. The sampling cap from utilizing 96-well plate-based screens greatly restricts our ability to identify rare, highly functional variants. Future directions to expand the mNAD evolution toolbox will involve adapting principles and methods currently used for natural cofactors.

In addition, computational methods will be highly instrumental in engineering mNAD-dependent enzymes. Without crystal structures of the target enzymes with noncanonical cofactors bound, molecular modeling tools are essential for visualizing enzyme-cofactor interaction. Furthermore, homology modeling tools and sequence alignment facilitate the translation of successful mutations between different enzymes. For example, *E. coli* malic enzyme L310R gained the ability to utilize nicotinamide flucytosine dinucleotide (NCFD⁺) and NCD⁺[38]. High sequence conservation at L310 inspired the rational design of *E. coli* malate dehydrogenase L6R for NCFD⁺ binding[38], *Lactobacillus helveticus* D-lactate dehydrogenase V152R[7, 38], and *Ralstonia sp.* 4506 phosphite dehydrogenase I151R for NCD⁺ binding[16].

Bibliography

- [1] Lara Sellés Vidal, Ciarán L Kelly, Paweł M Mordaka, and John T Heap. Review of NAD(P)H-dependent oxidoreductases: Properties, engineering and application. *Biochim. Biophys. Acta: Proteins Proteomics*, 1866(2):327–347, February 2018.
- [2] B M Anderson and N O Kaplan. Enzymatic studies with analogues of diphosphopyridine nucleotide. *J. Biol. Chem.*, 234(5):1226–1232, May 1959.
- [3] S Sicsic, P Durand, S Langrene, and F le Goffic. A new approach for using cofactor dependent enzymes: example of alcohol dehydrogenase. *FEBS Lett.*, 176(2):321–330, October 1984.
- [4] S Sicsic, P Durand, S Langrené, and F Le Goffic. Activity of NMN⁺, nicotinamide ribose and analogs in alcohol oxidation promoted by horse-liver alcohol dehydrogenase. improvement of this activity and structural requirements of the pyridine nucleotide part of the NAD⁺ coenzyme. *Eur. J. Biochem.*, 155(2):403–407, March 1986.
- [5] Yuxue Liu, Yanbin Feng, Lei Wang, Xiaojia Guo, Wujun Liu, Qing Li, Xueying Wang, Song Xue, and Zongbao Kent Zhao. Structural insights into phosphite dehydrogenase variants favoring a non-natural redox cofactor. *ACS Catal.*, 9(3):1883–1887, March 2019.
- [6] Xiaojia Guo, Yuxue Liu, Qian Wang, Xueying Wang, Qing Li, Wujun Liu, and Zongbao K Zhao. Non-natural cofactor and Formate-Driven reductive carboxylation of pyruvate. *Angew. Chem. Int. Ed Engl.*, 59(8):3143–3146, February 2020.
- [7] Yuxue Liu, Qing Li, Lei Wang, Xiaojia Guo, Junting Wang, Qian Wang, and Zongbao K Zhao.

- Engineering d-lactate dehydrogenase to favor an non-natural cofactor nicotinamide cytosine dinucleotide. *Chembiochem*, 21(14):1972–1975, July 2020.
- [8] Claudia Nowak, André Pick, Petra Lommes, and Volker Sieber. Enzymatic reduction of nicotinamide biomimetic cofactors using an engineered glucose dehydrogenase: Providing a regeneration system for artificial cofactors. *ACS Catal.*, 7(8):5202–5208, August 2017.
- [9] William B Black, Linyue Zhang, Wai Shun Mak, Sarah Maxel, Youtian Cui, Edward King, Bonnie Fong, Alicia Sanchez Martinez, Justin B Siegel, and Han Li. Engineering a nicotinamide mononucleotide redox cofactor system for biocatalysis. *Nat. Chem. Biol.*, 16(1):87–94, January 2020.
- [10] Rui Huang, Hui Chen, David M Upp, Jared C Lewis, and Yi-Heng P Job Zhang. A High-Throughput method for directed evolution of NAD(P)+-Dependent dehydrogenases for the reduction of biomimetic nicotinamide analogues. *ACS Catal.*, 9(12):11709–11719, December 2019.
- [11] Caroline E Paul, Serena Gargiulo, Diederik J Opperman, Ivan Lavandera, Vicente Gotor-Fernandez, Vicente Gotor, Andreas Taglieber, Isabel W C E Arends, and Frank Hollmann. Mimicking nature: synthetic nicotinamide cofactors for C=C bioreduction using enoate reductases. *Org. Lett.*, 15(1):180–183, January 2013.
- [12] Claudia Nowak, André Pick, Lénárd-István Csepei, and Volker Sieber. Characterization of biomimetic cofactors according to stability, redox potentials, and enzymatic conversion by NADH oxidase from lactobacillus pentosus. *Chembiochem*, 18(19):1944–1949, October 2017.
- [13] Sebastian A Löw, Isabell M Löw, Martin J Weissenborn, and Bernhard Hauer. Enhanced Ene-Reductase activity through alteration of artificial nicotinamide cofactor substituents. *ChemCatChem*, 8(5):911–915, March 2016.
- [14] Claudia Nowak, Barbara Beer, André Pick, Teresa Roth, Petra Lommes, and Volker Sieber. A water-forming NADH oxidase from lactobacillus pentosus suitable for the regeneration of synthetic biomimetic cofactors. *Front. Microbiol.*, 6:957, September 2015.

- [15] Alice Guarneri, Adrie H Westphal, Jos Leertouwer, Joy Lunsonga, Maurice C R Franssen, Diederik J Opperman, Frank Hollmann, Willem J H Berkel, and Caroline E Paul. Flavoenzyme-mediated regioselective aromatic hydroxylation with coenzyme biomimetics. *ChemCatChem*, 12(5):1368–1375, March 2020.
- [16] Lei Wang, Debin Ji, Yuxue Liu, Qian Wang, Xueying Wang, Yongjin J Zhou, Yixin Zhang, Wujun Liu, and Zongbao K Zhao. Synthetic Cofactor-Linked metabolic circuits for selective energy transfer. *ACS Catal.*, 7(3):1977–1983, March 2017.
- [17] Michael S Packer and David R Liu. Methods for the directed evolution of proteins. *Nat. Rev. Genet.*, 16(7):379–394, July 2015.
- [18] Brian Kuhlman and Philip Bradley. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.*, 20(11):681–697, November 2019.
- [19] Jackson K B Cahn, Sabine Brinkmann-Chen, and Frances H Arnold. Enzyme nicotinamide cofactor specificity reversal guided by automated structural analysis and library design. In Michael Krogh Jensen and Jay D Keasling, editors, *Synthetic Metabolic Pathways: Methods and Protocols*, pages 15–26. Springer New York, New York, NY, 2018.
- [20] Tobias J Gmelch, Josef M Sperl, and Volker Sieber. Molecular dynamics analysis of a rationally designed aldehyde dehydrogenase gives insights into improved activity for the Non-Native cofactor NAD. *ACS Synth. Biol.*, 9(4):920–929, April 2020.
- [21] Andy Beier, Sven Bordewick, Maika Genz, Sandy Schmidt, Tom van den Bergh, Christin Peters, Henk-Jan Joosten, and Uwe T Bornscheuer. Switch in cofactor specificity of a Baeyer-Villiger monooxygenase. *Chembiochem*, 17(24):2312–2315, December 2016.
- [22] Christoph Mahler, Franziska Kratzl, Melina Vogel, Stefan Vinnenberg, Dirk Weuster-Botz, and Kathrin Castiglione. Loop swapping as a potent approach to increase ene reductase activity with nicotinamide adenine dinucleotide (NADH). *Adv. Synth. Catal.*, 106:946, April 2019.
- [23] Kusum Solanki, Walaa Abdallah, and Scott Banta. Engineering the cofactor specificity of an

- alcohol dehydrogenase via single mutations or insertions distal to the 2'-phosphate group of NADP(H). *Protein Eng. Des. Sel.*, 30(5):373–380, May 2017.
- [24] Walaa Abdallah, Kusum Solanki, and Scott Banta. Insertion of a Calcium-Responsive β -Roll domain into a thermostable alcohol dehydrogenase enables tunable control over cofactor selectivity. *ACS Catal.*, 8(2):1602–1613, February 2018.
- [25] Walaa Abdallah, Vanessa Chirino, Ian Wheeldon, and Scott Banta. Catalysis of thermostable alcohol dehydrogenase improved by engineering the microenvironment through fusion with supercharged proteins. *Chembiochem*, 20(14):1827–1837, July 2019.
- [26] Yingning Gao, Christopher C Roberts, Jie Zhu, Jyun-Liang Lin, Chia-En A Chang, and Ian Wheeldon. Tuning enzyme kinetics through designed intermolecular interactions far from the active site. *ACS Catal.*, 5(4):2149–2153, April 2015.
- [27] Chen Lu, Fenglin Shen, Shuaibo Wang, Yuyang Wang, Juan Liu, Wen-Ju Bai, and Xiqing Wang. An engineered Self-Sufficient biocatalyst enables scalable production of linear α -Olefins from carboxylic acids. *ACS Catal.*, 8(7):5794–5798, July 2018.
- [28] Nina Beyer, Justyna K Kulig, Anette Bartsch, Martin A Hayes, Dick B Janssen, and Marco W Fraaije. P450BM3 fused to phosphite dehydrogenase allows phosphite-driven selective oxidations. *Appl. Microbiol. Biotechnol.*, 101(6):2319–2331, March 2017.
- [29] Maria L Corrado, Tanja Knaus, and Francesco G Mutti. A chimeric styrene monooxygenase with increased efficiency in asymmetric biocatalytic epoxidation. *Chembiochem*, 19(7):679–686, April 2018.
- [30] Lei Huang, Friso S Aalbers, Wei Tang, Robert Röllig, Marco W Fraaije, and Selin Kara. Convergent cascade catalyzed by Monooxygenase–Alcohol dehydrogenase fusion applied in organic media. *Chembiochem*, 20(13):1653–1658, July 2019.
- [31] Friso S Aalbers and Marco W Fraaije. Design of artificial alcohol oxidases: Alcohol Dehydrogenase–NADPH oxidase fusions for continuous oxidations. *Chembiochem*, 20(1):51–56, January 2019.

- [32] Harun F Ozbakir, Kristen E Garcia, and Scott Banta. Creation of a formate: malate oxidoreductase by fusion of dehydrogenase enzymes with PEGylated cofactor swing arms. *Protein Eng. Des. Sel.*, 31(4):103–108, April 2018.
- [33] Andrea M Chánique and Loreto P Parra. Protein engineering for nicotinamide coenzyme specificity in oxidoreductases: Attempts and challenges. *Front. Microbiol.*, 9:194, February 2018.
- [34] Tanja Knaus, Caroline E Paul, Colin W Levy, Simon de Vries, Francesco G Mutti, Frank Hollmann, and Nigel S Scrutton. Better than nature: Nicotinamide biomimetics that outperform natural coenzymes. *J. Am. Chem. Soc.*, 138(3):1033–1039, January 2016.
- [35] H Christine Lo, Jessica D Ryan, John B Kerr, Douglas S Clark, and Richard H Fish. Bioorganometallic chemistry: Co-factor regeneration, enzyme recognition of biomimetic 1,4-NADH analogs, and organic synthesis; tandem catalyzed regioselective formation of n-substituted-1,4-dihydronicotinamide derivatives with $[\text{Cp}^*\text{Rh}(\text{bpy})\text{H}]^+$, coupled to chiral s-alcohol formation with HLADH, and engineered cytochrome p450s, for selective C-H oxidation reactions. *J. Organomet. Chem.*, 839:38–52, June 2017.
- [36] Caroline E Paul, Dirk Tischler, Anika Riedel, Thomas Heine, Nobuya Itoh, and Frank Hollmann. Nonenzymatic regeneration of styrene monooxygenase for catalysis. *ACS Catal.*, 5(5):2961–2965, May 2015.
- [37] Elliot Campbell, Matthew Meredith, Shelley D Minter, and Scott Banta. Enzymatic biofuel cells utilizing a biomimetic cofactor. *Chem. Commun.*, 48(13):1898–1900, February 2012.
- [38] Debin Ji, Lei Wang, Shuhua Hou, Wujun Liu, Jinxia Wang, Qian Wang, and Zongbao K Zhao. Creation of bioorthogonal redox systems depending on nicotinamide flucytosine dinucleotide. *J. Am. Chem. Soc.*, 133(51):20857–20862, December 2011.
- [39] Humberto Flores and Andrew D Ellington. A modified consensus approach to mutagenesis inverts the cofactor specificity of bacillus stearothermophilus lactate dehydrogenase. *Protein Eng. Des. Sel.*, 18(8):369–377, August 2005.

- [40] Jessica D Ryan, Richard H Fish, and Douglas S Clark. Engineering cytochrome P450 enzymes for improved activity towards biomimetic 1,4-NADH cofactors. *Chembiochem*, 9(16):2579–2582, November 2008.
- [41] Alexandra M Weingartner, Daniel F Sauer, Gaurao V Dhoke, Mehdi D Davari, Anna Joëlle Ruff, and Ulrich Schwaneberg. A hydroquinone-specific screening system for directed P450 evolution. *Appl. Microbiol. Biotechnol.*, 102(22):9657–9667, November 2018.
- [42] Oliver F Brandenburg, Kai Chen, and Frances H Arnold. Directed evolution of a cytochrome P450 carbene transferase for selective functionalization of cyclic compounds. *J. Am. Chem. Soc.*, 141(22):8989–8995, June 2019.
- [43] Vida Časaitė, Mikas Sadauskas, Justas Vaitekūnas, Renata Gasparavičiūtė, Rita Meškienė, Izabelė Skikaitė, Mantas Sakalauskas, Jevgenija Jakubovska, Daiva Tauraitė, and Rolandas Meškys. Engineering of a chromogenic enzyme screening system based on an auxiliary indole-3-carboxylic acid monooxygenase. *Microbiologyopen*, 8(8):e00795, August 2019.
- [44] Yan Zhang, Yin-Qi Wu, Na Xu, Qian Zhao, Hui-Lei Yu, and Jian-He Xu. Engineering of cyclohexanone monooxygenase for the enantioselective synthesis of (S)-Omeprazole. *ACS Sustainable Chem. Eng.*, 7(7):7218–7226, April 2019.
- [45] Na Xu, Jun Zhu, Yin-Qi Wu, Yan Zhang, Jian-Ye Xia, Qian Zhao, Guo-Qiang Lin, Hui-Lei Yu, and Jian-He Xu. Enzymatic preparation of the chiral (S)-Sulfoxide drug esomeprazole at Pilot-Scale levels. *Org. Process Res. Dev.*, 24(6):1124–1130, June 2020.
- [46] Tristan de Rond, Jian Gao, Amin Zargar, Markus de Raad, Jack Cunha, Trent R Northen, and Jay D Keasling. A High-Throughput mass spectrometric enzyme activity assay enabling the discovery of cytochrome P450 biocatalysts. *Angew. Chem. Int. Ed.*, 58(30):10114–10119, July 2019.
- [47] Solvej Siedler, Georg Schendzielorz, Stephan Binder, Lothar Eggeling, Stephanie Bringer, and Michael Bott. SoxR as a single-cell biosensor for NADPH-consuming enzymes in *escherichia coli*. *ACS Synth. Biol.*, 3(1):41–47, January 2014.

- [48] Alina Spielmann, Yannik Brack, Hugo van Beek, Lion Flachbart, Lea Sundermeyer, Meike Baumgart, and Michael Bott. NADPH biosensor-based identification of an alcohol dehydrogenase variant with improved catalytic properties caused by a single charge reversal at the protein surface. *AMB Express*, 10(1):14, January 2020.
- [49] Aaron Debon, Moritz Pott, Richard Obexer, Anthony P Green, Lukas Friedrich, Andrew D Griffiths, and Donald Hilvert. Ultrahigh-throughput screening enables efficient single-round oxidase remodelling. *Nature Catalysis*, 2(9):740–747, September 2019.
- [50] Keming Liang and Claire R Shen. Selection of an endogenous 2,3-butanediol pathway in *escherichia coli* by fermentative redox balance. *Metab. Eng.*, 39:181–191, January 2017.
- [51] Sarah Maxel, Samer Saleh, Edward King, Derek Aspacio, Linyue Zhang, Ray Luo, and Han Li. Growth-Based, High-Throughput selection for NADH preference in an Oxygen-Dependent biocatalyst. *ACS Synth. Biol.*, September 2021.
- [52] Linyue Zhang, Edward King, Ray Luo, and Han Li. Development of a High-Throughput, in vivo selection platform for NADPH-Dependent reactions based on redox balance principles. *ACS Synth. Biol.*, 7(7):1715–1721, July 2018.
- [53] Liliana Calzadiaz-Ramirez, Carla Calvó-Tusell, Gabriele M M Stoffel, Steffen N Lindner, Sílvia Osuna, Tobias J Erb, Marc Garcia-Borràs, Arren Bar-Even, and Carlos G Acevedo-Rocha. In vivo selection for formate dehydrogenases with high efficiency and specificity toward NADP. *ACS Catal.*, 10(14):7512–7525, July 2020.
- [54] Steffen N Lindner, Liliana Calzad Iacute Az Ramirez, Jan Krüsemann, Oren Yishai, Sophia Belkhelfa, Hai He, Madeleine Bouzon, Volker Döring, and Arren Bar-Even. NADPH-auxotrophic *e. coli*: a sensor strain for testing in vivo regeneration of NADPH. *ACS Synth. Biol.*, November 2018.
- [55] Sarah Maxel, Derek Aspacio, Edward King, Linyue Zhang, Ana Paula Acosta, and Han Li. A Growth-Based, High-Throughput selection platform enables remodeling of 4-hydroxybenzoate hydroxylase active site. *ACS Catal.*, 10(12):6969–6974, June 2020.

- [56] Sarah Maxel, Linyue Zhang, Edward King, Ana Paula Acosta, Ray Luo, and Han Li. In vivo, High-Throughput selection of thermostable cyclohexanone monooxygenase (CHMO). *Catalysts*, 10(8):935, August 2020.
- [57] Xueying Wang, Yongjin J Zhou, Lei Wang, Wujun Liu, Yuxue Liu, Chang Peng, and Zongbao K Zhao. Engineering escherichia coli nicotinic acid mononucleotide adenylyltransferase for fully active amidated NAD biosynthesis. *Appl. Environ. Microbiol.*, 83(13), July 2017.

Chapter 4

Semi-rational design of *E. coli* gapA to utilize the artificial redox cofactor NMN⁺

4.1 Abstract

Nicotinamide cofactors shuttle the electron energy required for enzymatic redox transformations. The native cofactors NAD/H and NADP/H are consumed with remarkable specificity determined by the amino acid sequence of the cofactor binding site. However, the mapping of sequence to cofactor preference and activity levels are not well understood, hindering efforts to expand enzyme cofactor scopes to biomimetic nicotinamide cofactors. Here we utilize *E. coli* gapA to build upon previous efforts designing proteins with enhanced activity for the artificial redox cofactor NMN⁺ through two approaches, computationally guided enzyme design and high throughput screening of semi-rational variants through crude-lysate based colorimetric assay. Through computationally guided enzyme design, we identify the variant A180S with ~6-fold increase in NMN⁺ catalytic efficiency and ~10-fold cofactor specificity switch compared to wildtype gapA (WT), and further

refine orthogonality by developing the double mutant A180S-G10R with ~ 7 -fold greater NMN⁺ catalytic efficiency and ~ 200 -fold cofactor specificity switch from NAD⁺ to NMN⁺. From high-throughput screening of semi-rational mutants on A180S to augment NMN⁺ activity we discover the variant A180S-G187K-P188A that has ~ 32 -fold increase in NMN⁺ catalytic efficiency and ~ 50 -fold cofactor specificity switch over WT. Molecular modeling suggests that the improved NMN⁺ specificity is driven by the formation of novel polar contacts to the NMN⁺ phosphate group and cooperative reshaping of the binding pocket with increased loop flexibility at the subunit binding interface to exclude native cofactor binding. Overall, we demonstrate two parallel strategies to engineer enzymes that improve utilization of NMN⁺ and will be essential tools in the development of designer metabolic pathways that utilize orthogonal cofactor systems.

4.2 Introduction

Metabolic redox reactions are dependent on electron transfer via the cofactors nicotinamide adenine dinucleotide (NAD/H) and nicotinamide adenine dinucleotide phosphate (NADP/H)[1]. These two molecules share a nearly identical scaffold of a nicotinamide head group attached to an adenosine monophosphate (AMP) tail. The nicotinamide ring is mandatory in catalyzing the hydride transfer for redox reactions, while the AMP functions as a handle for the enzyme to recognize and latch on to secure binding with the cofactor. NADP/H differs from NAD/H through addition of a single phosphate group at the 2'-OH of the AMP ribose. Although both molecules are highly similar and both bind to the same conserved Rossmann fold motif, enzymes are able to precisely distinguish the cofactors and typically evolve specificity for one of the two[2, 3]. The Rossmann fold sequence signatures that determine cofactor preference and degree of activity are not well characterized, and efforts to swap cofactor specificity typically show low success rates, leading to demands for more effective methods for cofactor engineering[4].

The high costs and low stability of native nicotinamide cofactors limits the scalability of biomanufacturing processes[5]. We propose to solve this problem with artificial redox cofactors, which

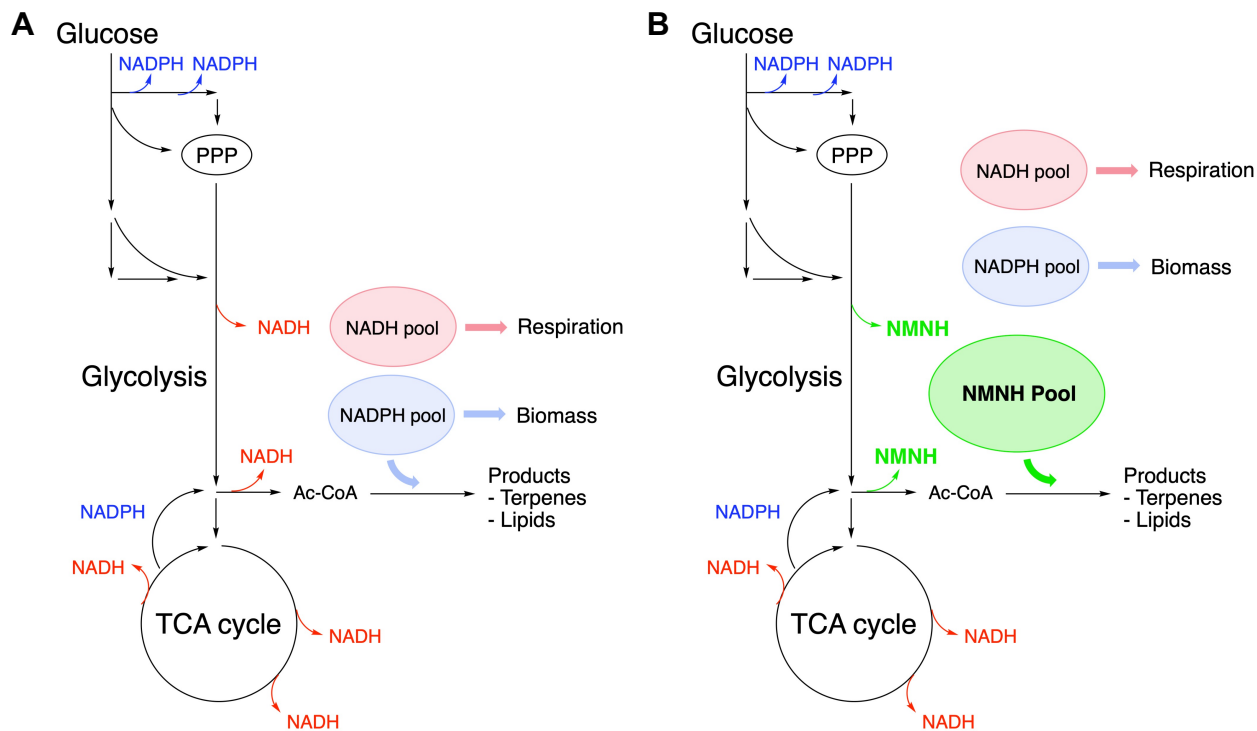


Figure 4.1: Redox cofactor generation with native and engineered glycolysis. **(A)** Glucose processing results in production of the reduced cofactors NADH and NADPH through enzymatic transformations in the pentose phosphate pathway, glycolysis, and TCA cycle. The functions of the cofactors are partitioned, NADH is utilized in respiration and NADPH in biomass synthesis. **(B)** A separate cofactor pool based on the artificial redox cofactor NMNH enables access to its electron energy with minimal disruption to natural processes and specific delivery to designer pathways.

are cheaply synthesized due to their simpler structures and which have been demonstrated to enable specific electron delivery in engineered metabolic pathways for more cost-efficient product generation[6, 7] (Figure: 4.1). Our proof of concept molecule is nicotinamide mononucleotide (NMN/H), a truncated version of the native nicotinamide cofactors that maintains the catalytic nicotinamide ring, ribose, and a single phosphate group (Figure: 4.2A). No known enzyme natively utilizes NMN⁺, a natural metabolite in NAD⁺ biosynthesis, as a redox cofactor, but there have been several reports of engineered enzymes binding NMN⁺ or other artificial redox cofactor[6, 8–16]. Here we demonstrate how parallel approaches of computationally guided enzyme design and high-throughput screening of combinatorial variants through plate based colorimetric assay can be utilized to engineer NMN⁺ consuming enzymes with nominal screening effort.

E. coli glyceraldehyde-3 phosphate dehydrogenase (gapA) is a central enzyme in glycolysis functioning to convert glyceraldehyde-3 phosphate to D-glycerate 1,3-bisphosphate, in the process reducing NAD⁺ to NADH[17, 18]. We select gapA to engineer for NMN⁺ utilization due to its capacity for high catalytic turnover rates and challenging functional profile arising from strong sequence conservation due to its essential role in glycolysis. Previous reports have highlighted the difficulty of achieving highly active variants with cofactor preference switched to NADP⁺[19]. We have previously demonstrated that a small number of mutations are necessary to enable practical levels of catalysis with NMN⁺ as the functional portion involved in hydride transfer, the nicotinamide ring, is kept intact, while the lack of the AMP on NMN⁺ results in the loss of a substantial number of polar contacts and surface area for hydrophobic packing with the protein[6]. Our approach is based on the rationale that low NMN⁺ activity with WT enzymes is due to the lack of binding contacts to the Rossmann fold that has been optimized by evolution to explicitly recognize the larger native cofactors, and alterations to the binding site to compensate for the lost AMP contacts need to be made to realize sufficient binding affinity[20, 21]. Furthermore, to advance orthogonality the activity with the native cofactor NAD⁺ must be ablated by impeding binding of the dinucleotide cofactor. Due to the limited number of successful cases of enzymes engineered to utilize NMN⁺, there are no known design principles to follow making the engineering process especially challenging.

The successful design of enzymes for a target function relies on either having the capability to

screen with high-throughput and sensitivity to discover rare variants with greater fitness, or the ability to accurately predict the effect of mutations on the function of a protein such that the required experimental validation is minimized[22]. We combine both approaches for engineering gapA to bind NMN⁺ with computationally guided selection of residues to mutate and a high-throughput colorimetric activity assay measuring cofactor reduction in crude lysate. Our results lay the blueprint for designing NMN⁺ consuming enzymes by introducing structurally predicted novel hydrogen bonds to the NMN⁺ phosphate group and greater NMN⁺ specificity by occluding the AMP portion of native cofactors from binding through steric hindrance. We further performed high-throughput screening with a library of combinatorial mutants carrying randomly sampled substitutions at four hotspots involved in cofactor binding and protein flexibility to discover the variant A180S-G187K-P188A (HT-9). Based on molecular modeling, HT-9 is proposed to form an additional salt bridge from G187K to the NMN⁺ phosphate that is able to reach the ligand due to increased loop flexibility through mutation of the rigid P188 to the smaller Ala, and achieves a ~32-fold increase in NMN⁺ catalytic efficiency over WT.

4.3 Methods

4.3.1 Plasmid and strain construction

Utilized plasmids and strains are listed in Supplementary Tables 4.3 and 4.2. Plasmid construction was completed with the Gibson isothermal DNA assembly method. Site directed mutagenesis was performed via PCR with PrimeSTAR Max DNA Polymerase (TaKaRa) and mutagenic primers carrying the target codon substitutions to amplify DNA fragments for ligation. Cloning steps were run with *E. coli* XLI-Blue (Stratagene).

The *E. coli* gapA gene was amplified from *E. coli* BW25113 genomic DNA through PCR, cleaned through gel extraction, and inserted into the pQE vector backbone (N-terminal 6x His-tag, ColE1 ori, Amp^R) through Gibson assembly to generate pEK-28.

4.3.2 Protein expression and purification

Proteins were expressed with a N-terminal 6x His-tag for affinity purification with the His-Spin Protein Miniprep kit (Zymo Research Corporation). Plasmids were transformed into *E. coli* BL21 (DE3) for overexpression, transformant colonies were inoculated in 2XYT media with 100 $\mu\text{g}/\text{mL}$ ampicillin for overnight expansion, sub-cultured the next day at 1% volume with 100 $\mu\text{g}/\text{mL}$ ampicillin, induced with 0.5 mM IPTG at OD600 0.5, and incubated at 30°C for 24 hrs with 250 rpm shaking for protein production. Pelleted cells were disrupted with bead-beating and purification from the cell lysate was performed with Ni-NTA resin according to manufacturer protocols. Isolated proteins were quantified with Bradford assay comparing to BSA standard curve and stored with 20% glycerol at -80°C.

4.3.3 gapA enzymatic assays and kinetics study

The gapA enzyme assay protocol was adapted from previous work[19]. The reactions to measure specific activities were initiated by addition of purified enzyme into the assay mixture containing 50 mM Tris-Cl pH 8.5, 0.2 mM EDTA, 50 mM Na_2PO_4 , 3 mM DL-G3P, and 4 mM cofactor at 25°C. Production of reduced cofactor was detected with spectrophotometer by absorbance at 340 nm. Final specific activities were corrected for cofactor carry-over during purification by subtracting the background activity measured from reaction with no cofactor added.

Determination of the Michaelis-Menten kinetic parameters k_{cat} describing the turnover rate and Michaelis constant K_m was completed with similar master mix where DL-G3P was replaced with 1.5 mM D-G3P and cofactor concentration was varied. Initial reaction rates were recorded and fit to the Michaelis-Menten equation where v_0 is the initial velocity, E_t is the total enzyme concentration, and S is the cofactor concentration.

$$v_0 = \frac{E_t \cdot k_{cat} \cdot S}{K_m + S} \quad (4.1)$$

Under conditions where the enzyme could not be saturated with cofactor ($K_m \gg S$), the initial velocities were fit to the linear Michaelis-Menten equation to solve for the catalytic efficiency k_{cat}/K_m .

$$v_0 = \frac{E_t \cdot k_{cat} \cdot S}{K_m} \quad (4.2)$$

4.3.4 Rosetta ligand docking and enzyme design

The crystal structure of *E.coli* gapA 1GAD is in the non-functional dimeric form[17]. We first structurally align single chains of 1GAD to each subunit of the homologous gapA 1J0X from rabbit muscle (*O. cuniculus*), which is resolved to be in the functional tetrameric form[23]. The Ec gapA models with tetrameric symmetry exhibit a binding mode with the cofactor in position to form inter-subunit polar contacts between neighboring monomers. All simulations are built upon the resultant structure. The NMN⁺ conformer library was built and optimized using Spartan, then used for a docking and design simulation with RosettaDesign[24] using distance and angle constraints to maintain catalytic geometry. A total of 5,000 simulations were run for each round of design and the top 20 best scoring outputs sorted based on protein-ligand interface energy and Rosetta total system energy were selected for analysis and visually checked through Foldit. During the design simulations, all side chains within 6 Å of the NMN⁺ ligand were allowed to be designed and any residues within 8 Å of the ligand were relaxed with backbone movements enabled. For docking and designing with NAD⁺, the docking protocol was nearly the same except with a conformer library of NAD⁺ generated previously[6]. For the computationally guided library, all side chains within 10 Å of the NMN⁺ ligand were allowed to be designed and mutated to all other 19 residues to maximize the diversity of mutants and comprehensively explore sequence space. Mutants with Rosetta interface energy and total energy greater than that of the WT protein were discarded. The remaining substitutions were sorted by energy, and a total of 240 variants arising from combinatorial substitutions at 4 positions were selected for library construction and experimental screening.

4.3.5 High-throughput library screening

Positions for semi-rational mutagenesis were selected with Rosetta. Primers encoding mutations for the selected substitutions were pooled together for site-directed mutagenesis and DNA fragments were generated with PCR using pEK-32 (gapA A180S) as template and gel purified. The backbone fragment was amplified separately and digested with DpnI overnight at 37°C to remove residual template plasmid, then gel purified. The gene inserts and backbone were ligated together through Gibson assembly with 5:1 insert to backbone ratio and transformed into XLI-Blue through electroporation with recovery for 1 hr in 2XYT at 37°C with 250 rpm shaking to produce the library of randomly sampled, combinatorial variants on pQE vector for testing. A small volume of the culture was plated to verify correct and diverse assembly through Sanger sequencing of randomly selected colonies, and transformation efficiency from colony counts greatly exceeded the library size indicating deep coverage of the possible variants. The remaining culture volume was grown overnight and minipreped to isolate the library plasmids for storage and later transformation.

The library plasmids were transformed into *E. coli* DS113[25] (Δ gapA12::Cm; obtained from the Yale *E. coli* Genetic Stock Center) to minimize background gapA activity that could interfere with the colorimetric assay. Transformed cells were grown on 2XYT agar with 100 μ g/mL ampicillin, 25 μ g/mL chloramphenicol, 12.5 mM sodium succinate, and 0.05% glycerol media at 37°C. Individual colonies were picked and inoculated in 300 μ L 2XYT-Amp-Cm-Suc-Gly in 96 deep well plates for overnight growth at 37°C. The overnight growth was sub-cultured the next day with 10 μ L transferred to fresh 300 μ L 2XYT-Amp-Cm-Suc-Gly in another 96 deep well plate, immediately induced with 0.5 mM IPTG, then sealed with breathable membrane for 24 hrs of growth at 30°C with 250 rpm shaking.

Purification from 96 deep well plates started from pelleting the cells with centrifugation at 3,500 rcf for 20 min. The supernatant was discarded, and a mixture of 200 μ L BugBuster Protein Extraction Reagent (Millipore-Sigma) and 0.2 μ L Lysonase Bioprocessing Reagent (Millipore-Sigma) was added to each sample. The pellets were resuspended and incubated at room temperature for 20 min, then pelleted again through centrifugation at 4°C for 30 min at 3,500 rcf. The supernatant

containing the protein lysate was directly utilized in the colorimetric assay.

The colorimetric assay detecting production of reduced cofactor via purple formazan development[26] utilized the same master mix as the specific activity assays with 3 mM NMN⁺ and the addition of 0.1 mM nitroblue tetrazolium and 25 μ M phenazine methosulfate. 160 μ L of the reaction master mix was aliquoted into 96 well plates, then 40 μ L of the protein lysate was added to initiate the reaction. The plate was briefly mixed at 500 rpm for 20 sec, and color development was monitored by spectrophotometer readings at 580 nm for 1 hr. Samples showing greater color development than the included control gapA A180S were saved from the initial overnight plate for validation with specific activity assay.

4.4 Results

4.4.1 Rational design of NMN⁺ binding gapA

Based on the expectation that it is more challenging to improve binding affinity for a nonnative substrate compared to lowering binding affinity for the natural redox cofactor, since reducing affinity can be readily performed through insertion of bulky residues throughout the binding site to introduce steric clash with the ligand, we begin by engineering for enhanced NMN⁺ binding. The successful design of enzymes for a target function relies on either having the capability to screen with high-throughput and sensitivity to discover rare variants with greater fitness, or the ability to accurately predict the effect of mutations on the function of a protein such that the required experimental validation is minimized. We aim to combine both approaches for engineering proteins to bind NMN⁺ with Rosetta guided selection of residues to mutate and a high-throughput colorimetric activity assay measuring cofactor reduction in crude lysate. Our strategy starts with placing polar residues in the first shell of the cofactor binding site that are predicted to generate attractive hydrogen-bonding or salt bridge interactions with the NMN⁺ phosphate group for stronger electrostatic complementarity.

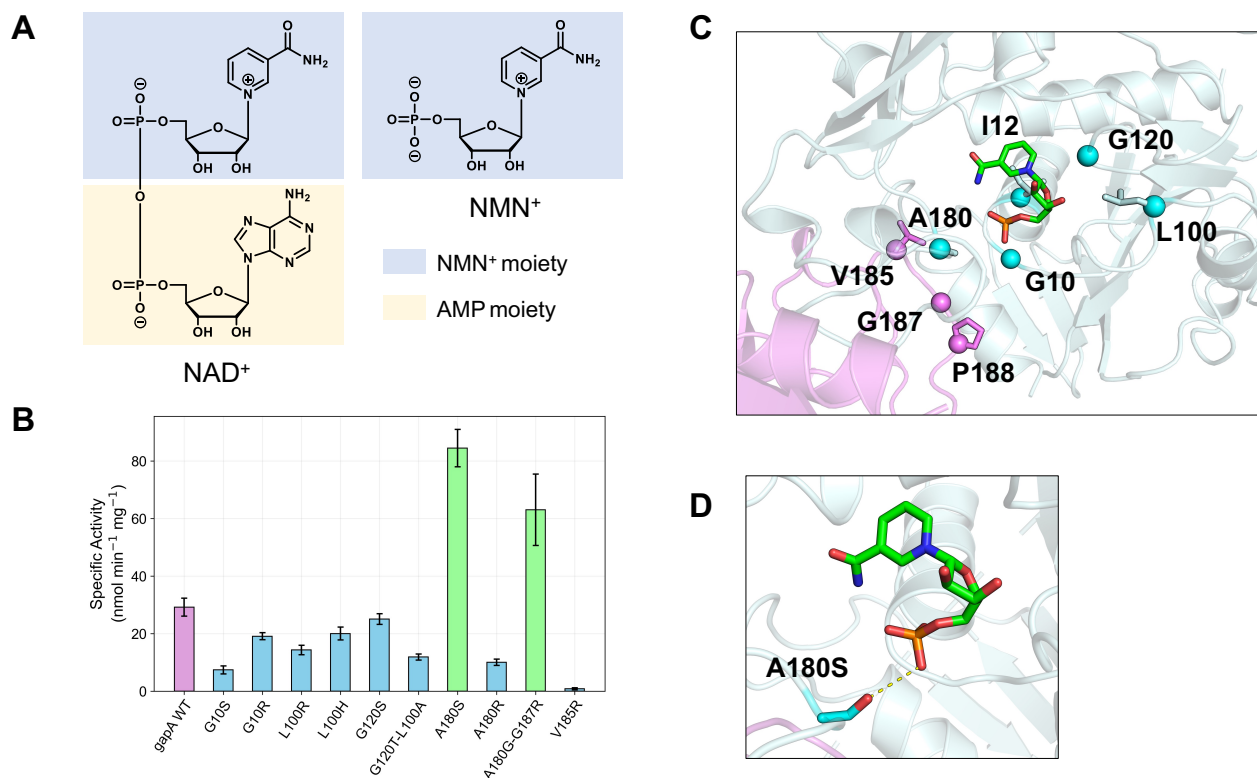


Figure 4.2: Modeled NMN^+ binding pose and first round gapA variant activities (**A**) NMN^+ is a truncated version of the natural cofactor NAD^+ without the adenosine monophosphate (AMP) tail. Hydride transfer capability is maintained at the nicotinamide ring, while the loss of the AMP handle results in diminished binding affinity. (**B**) WT gapA (purple) showed low specific activities with NMN^+ . Two variants, A180S and A180G-G187R (green), were discovered to have over 2-fold increase in specific activity. (**C**) Rosetta model of NMN^+ docked into the gapA Rossmann fold. The cofactor sits at the interface between neighboring subunits and is able to participate in inter-subunit interactions. The individual monomers are colored purple and cyan, and residues selected for mutagenesis are highlighted as spheres. (**D**) A180S is proposed to improve activity by forming a hydrogen bond with the NMN^+ phosphate group.

Wild-type Ec gapA displays low specific activity of $29.2 \pm 3.0 \text{ nmol min}^{-1} \text{ mg}^{-1}$ with NMN⁺, nearly 4,000-fold less than with the native cofactor NAD⁺ (Figure: 4.2B). We attempted to measure kinetic parameters with NMN⁺, and were unable to determine individual k_{cat} or K_m values as the reaction could not be saturated with substrate; the catalytic efficiency found from fitting the reaction rates to the linear Michaelis-Menten equation was estimated to be $2.0 \mu\text{M}^{-1} \text{ s}^{-1}$ (Table: 4.1). To build on the low baseline activity, we utilized molecular modeling with Rosetta to systematically simulate the effects of amino acid substitutions in the cofactor binding site on the predicted NMN⁺ binding pose and affinity. A conformer library of NMN⁺ was constructed and docked into the structure of gapA to optimize the protein-ligand interactions. During the simulation, all side chains within 6 Å of NMN⁺ were allowed to be designed and any residues within 8 Å of the ligand were relaxed with backbone movements enabled. The obtained poses are sorted based on protein-ligand interface energy and total system energy, and the top 20 best scoring outputs were selected for further inspection and design using Foldit[27]. We selected 10 candidates forming novel polar contacts to the NMN⁺ phosphate or ribose hydroxyls to experimentally construct and test (Figure: 4.2C). G10 occurs at the tip of the Rossman alpha helix, residues at this position would immediately contact the NMN⁺ phosphate group. A180 appears on a loop running parallel to the extended cofactor binding pose and is located across from the Rossman alpha helix. G120 and L100 are located on separate loops near the nicotinamide ribose, and may support contact to the hydroxyls. gapA is functionally a tetrameric enzyme, the crystal structure 1GAD[17] used as the template for molecular modeling only captures the dimeric form. However, a homologous gapA 1J0X[23] from rabbit muscle (*O. cuniculus*), was solved in the full tetrameric model and observed to have the cofactor binding site situated at the interface between adjoining subunits. Structural alignment of the Ec gapA subunits with the symmetry observed in Oc gapA resulted in a binding mode with the cofactor in position to form inter-subunit polar contacts with the neighboring monomer via a loop with V185 or G187 (Figure: 4.2C).

We discovered two variants with enhanced NMN⁺ activity from the first round of rational design, A180S and double mutant A180G-G187R (Figure: 4.2B). A180S is predicted to switch the polarity of the side chain by elongating from a non-polar methyl to a polar hydroxyl group, and establishes

a novel hydrogen bond to the NMN⁺ phosphate (Figure: 4.2D). The specific activity is measured to be 84.5 ± 6.0 nmol min⁻¹ mg⁻¹ with NMN⁺, roughly 2.8-fold increased over WT, with k_{cat} 0.18 ± 0.01 s⁻¹, K_m 15.32 ± 0.9 mM⁻¹, and catalytic efficiency 2.6×10^{-4} mM⁻¹ s⁻¹. A180S was noted to additionally improve NAD⁺ activity, increasing specific activity from $(1.20 \pm 0.03) \times 10^5$ nmol min⁻¹ mg⁻¹ in the WT to $(1.37 \pm 0.09) \times 10^5$ nmol min⁻¹ mg⁻¹. A180G-G187R is modeled first to reduce steric hindrance around the NMN⁺ phosphate by removing the Ala methyl with the A180G mutation, this enables G187R to extend without volume restriction to form a salt bridge with the NMN⁺ phosphate group. A180G-G187R was measured to have lower specific activity than A180S with NMN⁺ at 63.1 ± 12 nmol min⁻¹ mg⁻¹, and decreased NAD⁺ activity compared to WT with $(6.0 \pm 0.09) \times 10^4$ nmol min⁻¹ mg⁻¹. The results demonstrate that targeting the NMN⁺ phosphate for polar contacts is a viable strategy to improve NMN⁺ binding; however, the obtained variants still display low activity levels. Obtaining catalytic activity that approaches native levels likely requires greater perturbation to the enzyme sequence for optimal shape and electrostatic complementarity. We continue engineering mutations on A180S, the most effective variant, to further improve activity and specificity for NMN⁺.

4.4.2 Advancing orthogonality by disrupting native cofactor binding

To enforce specificity for NMN⁺ and orthogonality from native metabolism, we next designed mutations that block binding of the native cofactor NAD⁺ in gapA, while minimally disturbing NMN⁺ binding. We applied the two approaches of electrostatic repulsion and steric clash to reduce binding affinity for NAD⁺. The AMP handle on NAD⁺ is observed to fit into a cleft formed between the alpha helix and second beta strand of the Rossman fold. The introduction of bulky residues into this region would occlude the adenosine from fitting, while the NMN⁺ binding pose would be nominally affected as NMN⁺ makes no interaction in that region. The AMP phosphate group carries a strong negative charge that is not present on NMN⁺, this portion can be explicitly targeted for electrostatic repulsion by placing a negatively charged residue such as aspartate or glutamate

nearby to make binding unfavorable. Since the NBT-PMS saturation library screening is not amenable to negative screening, the assay is not able to distinguish low activity due to no binding affinity from completely non-functional protein, we evaluated these variants individually through rational design. Mutants harboring introduced bulky and negatively charged residues concentrated in the adenine cleft were systematically built *in silico*, docked with a conformer library of NAD⁺, and visually inspected for realistic geometries. We focused on selecting variants that possessed the most disturbed NAD⁺ binding mode, where NAD⁺ is restricted to binding as a non-native conformer or is dislocated from the position in crystal structure.

We constructed mutations on top of gapA A180S, and identified the double mutant A180S-G10R with greater orthogonality (Figure: 4.3). The specific activity for NMN⁺ was improved compared to WT, but lower than the template A180S at $52.3 \pm 0.3 \text{ nmol min}^{-1} \text{ mg}^{-1}$, with kinetic parameters $k_{cat} 0.05 \pm 0.001 \text{ s}^{-1}$, $K_m 3.88 \pm 0.3 \text{ mM}^{-1}$, and catalytic efficiency $0.014 \pm 0.001 \text{ mM}^{-1} \text{ s}^{-1}$. Improved orthogonality is demonstrated by the lowered NAD⁺ specific activity levels (3.3 ± 0.08) $\times 10^3 \text{ nmol min}^{-1} \text{ mg}^{-1}$, $k_{cat} 2.5 \pm 0.6 \text{ s}^{-1}$, $K_m 0.041 \pm 0.001 \text{ mM}^{-1}$, and catalytic efficiency $61.7 \pm 9 \text{ mM}^{-1} \text{ s}^{-1}$. Modeling suggests that the decrease in NAD⁺ activity is driven by G10R steric blockage. The Arg side chain extends from the tip of the Rossman helix to fill in the void where the AMP would typically bind, displacing the AMP and forcing it to suspend out freely in solvent instead of packing tightly against the protein (Figure: 4.3D). Since the native contacts stabilizing the AMP tail are not adhered to, the nicotinamide ring at the head of the cofactor cannot reliably achieve the naturally optimized binding geometry necessary for hydride transfer, resulting in reduced catalytic activity.

4.4.3 High-throughput library screening for NMN⁺ activity

Given the sparsity of functional mutants and vast extent of the sequence search space, we must take advantage of high-throughput methods to overcome low success rates by testing larger numbers of samples. We generate semi-rational variants through site directed mutagenesis with mixed codons, this results in a pool of mutants with combinatorial amino acid substitutions at focused positions

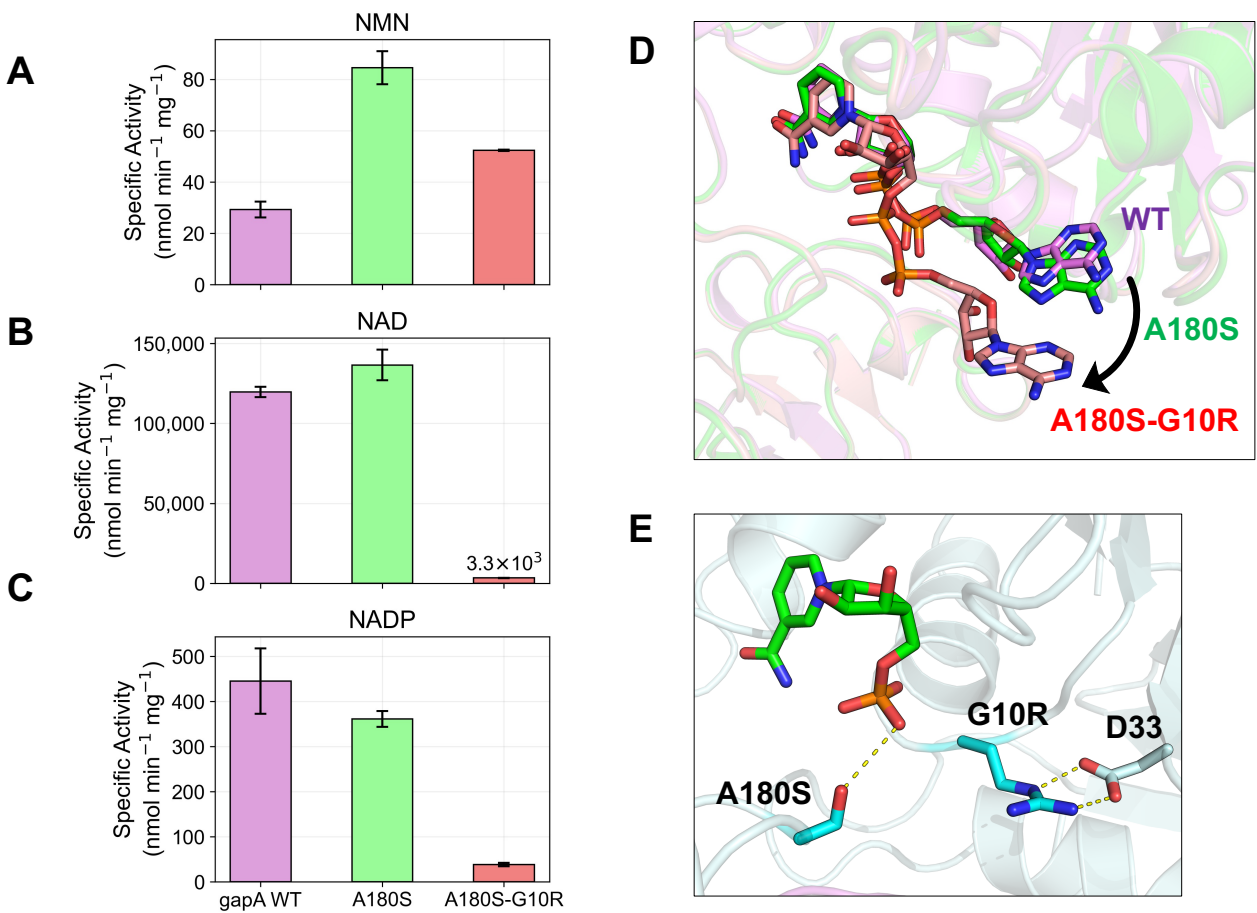


Figure 4.3: Orthogonal gapA specific activities and cofactor binding poses. **(A)** Specific activities with NMN⁺ and the two natural cofactors NAD(P)⁺. gapA (purple) natively utilizes NAD⁺, A180S (green) enhanced activity with both NMN⁺ and NAD⁺. A180S-G10R (red) substantially reduced specific activity with NAD⁺ while having small reduction in NMN⁺ compared to A180S alone. **(B)** G10R extends into the AMP binding pocket to create steric clash. This excludes the NAD from binding with high affinity and forces the cofactor to twist out toward solvent. **(D)** A180S maintains the polar contact to the NMN⁺ phosphate group, and G10R forms a salt bridge with D33 to occupy the AMP cleft.

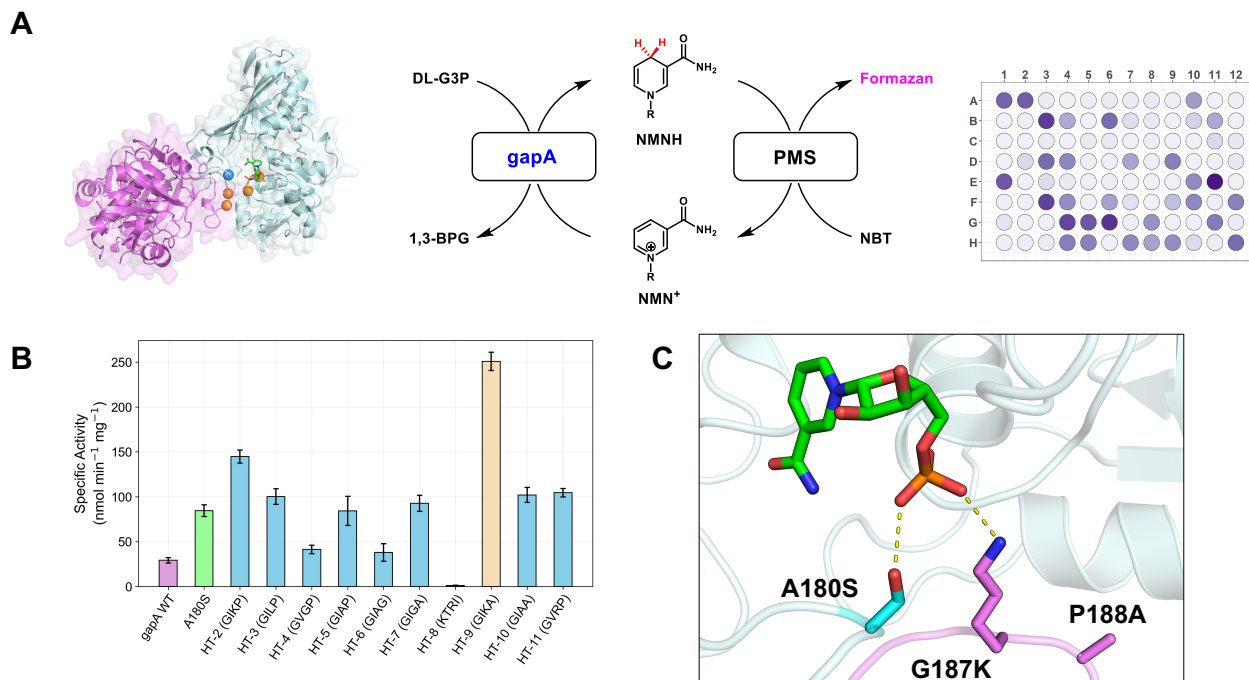


Figure 4.4: High-throughput colorimetric screening of *gapA* for NMN^+ activity. **(A)** Model of *Ec gapA* with positions selected for saturation mutagenesis highlighted as spheres, individual chains colored purple and cyan. The colorimetric assay is based on the reaction of the generated reduced cofactor NMNH with NBT and PMS to form the purple dye formazan. Activity from crude lysate in 96-well plate is measured by the intensity of color development. **(B)** NMN^+ specific activities from variants identified from the colorimetric assay. Baselines WT *gapA* (purple) and A180S (green) are compared to the high-throughput samples (blue and tan). HT-9 (GIKA) was found with $250.8 \pm 10.0 \text{ nmol min}^{-1} \text{ mg}^{-1}$ specific activity. **(C)** HT-9 is modeled to form an additional salt bridge from G187K to the NMN^+ phosphate group. This is made possible by P188A increasing loop flexibility for optimal positioning.

critical for cofactor binding. In this way, we limit the library search space to be experimentally tractable, yet random enough to allow for unpredictable cooperative interactions and major reorganization of the binding pocket. Based on Rosetta prediction of cofactor binding hotspots, we selected G10 (G, S, T, K), I12 (I, V, T), G187 (G, A, L, K, R), and P188 (P, I, G, A) for mutagenesis to enhance NMN^+ activity. I12, one of the initial residues on the Rossman alpha helix, natively packs alongside the face of the nicotinamide ring, and mutation could potentially allow the formation of hydrogen bonds with the ribose oxygen or phosphate group, or alter the flexibility of the alpha helix to shift the cofactor binding pose. The consecutive mutations at G187 and P188 reduce rigidity at the Phe base of the loop such that substitutions at G187, especially positively

charged residues, can move closer to the cofactor. In total our screening library covers 240 possible variants, the small library size permits screening by 96-well plate based assays.

Following construction of the library and transformation into cells with knocked out native gapA to minimize background noise, we cultured individual transformants with each carrying a separate gapA variant. Growth was completed in deep well plates, and following overnight development the cultures were transferred to another plate for induction and protein purification. The cells were lysed with detergent and the crude lysate was separated from the cellular debris through centrifugation. The lysate is mixed with the enzyme reaction buffer containing NMN⁺, the substrate DL-G3P, and the reagents nitroblue tetrazolium (NBT) and phenazine methosulfate (PMS). gapA production of NMNH reacts with NBT-PMS to produce the purple dye formazan, which can be easily readout with greater purple color intensity corresponding to higher total enzyme activity in the sample (Figure: 4.4A).

With screening over six 96 well plates to oversample and ensure complete coverage, we identified 33 samples with improved color development over the template A180S. Sequencing of these resulted in 10 unique variants, the sequence discovered with the highest frequency was the A180S template which was found in 20 of the samples (Figure: 4.4B) (SI Table: 4.4). G10 showed almost no variability, one sample was G10K, while the rest maintained the WT residue, highlighting the importance of conserving Gly at this position. I12 was also highly maintained with 2 samples showing I12V, and 1 sample showing I12T. G187 sampled the greatest range of substitutions, with 4 G187K, 3 G187A, 1 G187L, and 2 G187R with the rest staying Gly. Moderate variation was observed at P188 with 4 P188A, 1 P188G, 1 P188A, and 1 P188I. The unique variants were assayed for NMN⁺ specific activity, 7 showed increased NMN⁺ activity compared to the A180S template, and 3 had reduced activity. The most active variant, A180S-G187K-P188A, which we label HT9, has NMN⁺ specific activity of $250.8 \pm 10.0 \text{ nmol min}^{-1} \text{ mg}^{-1}$. Interestingly, the second most active mutant A180S-G187K differed only in maintaining the native P188 and had $144.8 \pm 7.0 \text{ nmol min}^{-1} \text{ mg}^{-1}$ specific activity with NMN⁺, emphasizing the key role of G187K in advancing NMN⁺ activity. This is suggested by Rosetta modeling to form a salt bridge with the NMN⁺ phosphate, and the beneficial influence of increasing loop flexibility by mutating the rigid Phe

residue to permit optimal positioning of G187K (Figure: 4.4C). The kinetic parameters for HT9 with NMN⁺ were measured to be k_{cat} $0.51 \pm 0.001 \text{ s}^{-1}$, K_m $8.16 \pm 0.02 \text{ mM}^{-1}$, and catalytic efficiency $0.06 \pm 0.001 \text{ mM}^{-1} \text{ s}^{-1}$ (Table: 4.1). Compared to the template A180S, HT9 has both superior turnover and binding affinity.

Interestingly, HT9 also gained the ability to utilize NADP⁺. An unintended effect of engineering for improved NMN⁺ binding ability was enhancing catalytic activity with NADP⁺ as HT9 showed specific activity of $1829.2 \pm 109 \text{ nmol min}^{-1} \text{ mg}^{-1}$ compared to A180S and gapA WT which showed similar NADP⁺ activity levels at $361.1 \pm 17 \text{ nmol min}^{-1} \text{ mg}^{-1}$ and $445.0 \pm 72 \text{ nmol min}^{-1} \text{ mg}^{-1}$ respectively (SI Figure: 4.6). Modeling suggests that P188A reduces crowding in the binding pocket to allow space for the extra phosphate group on the 2'-OH to fit, furthermore the Lys from G187K is able to form a salt bridge with the phosphate group for greater attractive binding interactions. The results indicate that strategies relaxing cofactor specificity by reshaping the contours of the binding interface are broadly applicable to native and non-native cofactors, and that binding modes designed to accommodate the smaller NMN⁺ through polar contacts at the pyrophosphate group may inadvertently support promiscuous cofactor binding. We further tested HT9-G10R to evaluate if the AMP blocking mutation would function similarly on HT9 in enhancing orthogonality (SI Table: 4.5). HT9-G10R showed decreased NAD⁺ activity as expected, but also showed dramatic loss in the ability to utilize NMN⁺ as specific activity dropped from $250.8 \pm 10 \text{ nmol min}^{-1} \text{ mg}^{-1}$ in HT9 to $11.5 \pm 0.4 \text{ nmol min}^{-1} \text{ mg}^{-1}$ in HT9-G10R. The G10R mutation is incompatible with HT9 likely due to unfavorable steric and electrostatic repulsion between the outstretched Arg and G187K near the NMN⁺ phosphate group (SI Figure: 4.6).

Enzyme	Cofactor	k_{cat} (s^{-1})	K_m (mM)	k_{cat}/K_m ($\text{mM}^{-1} \text{ s}^{-1}$)
gapA WT	NAD ⁺	51.8 ± 1.7	$(3.2 \pm 0.3) \times 10^{-2}$	$(1.6 \pm 0.1) \times 10^3$
gapA WT	NMN ⁺	ND	ND	$(2.0 \pm 0.1) \times 10^{-3}$
A180S	NAD ⁺	53.4 ± 0.8	$(6.1 \pm 0.4) \times 10^{-2}$	$(8.7 \pm 0.5) \times 10^2$
A180S	NMN ⁺	$(1.7 \pm 0.1) \times 10^{-1}$	15.3 ± 0.9	$(1.1 \pm 0.1) \times 10^{-2}$
A180S-G10R	NAD ⁺	2.5 ± 0.6	$(4.1 \pm 0.3) \times 10^{-2}$	61.7 ± 9.4

A180S-G10R	NMN ⁺	$(5.4 \pm 0.1) \times 10^{-2}$	3.8 ± 0.3	$(1.4 \pm 0.1) \times 10^{-2}$
HT9 (A180S-G187K-P188A)	NAD ⁺	43.2 ± 2.8	$(3.9 \pm 0.5) \times 10^{-2}$	$(1.1 \pm 0.1) \times 10^3$
HT9 (A180S-G187K-P188A)	NMN ⁺	$(5.1 \pm 0.1) \times 10^{-1}$	8.1 ± 0.1	$(6.3 \pm 0.1) \times 10^{-2}$

Table 4.1: gapA Michaelis-Menten kinetic parameters. ND indicates the value could not be determined due to high K_m preventing reaction saturation.

4.5 Discussion

We engineered NAD-dependent *E. coli* gapA to utilize the artificial redox cofactor NMN⁺ with improved catalytic activity and specificity. Through a single round of rational design guided by Rosetta molecular simulation, we discovered the variant A180S with ~ 6 -fold enhancement in NMN⁺ catalytic efficiency and ~ 10 -fold cofactor specificity switch. We carried out a second round of rational design with the goal of blocking native NAD⁺ activity to increase orthogonality, this resulted in the variant A180S-G10R with ~ 7 -fold improved NMN⁺ catalytic efficiency and ~ 200 -fold cofactor specificity switch from the WT. To more deeply navigate sequence space and identify variants with unpredictable cooperative effects, we turned to high-throughput screening. We utilized a 96-well plate based assay with facile readout of color development that measured the production of reduced cofactor, NMNH, from crude lysate through reaction generating the purple dye formazan. After screening a semi-rational library covering four cofactor binding hotspot positions, we found the triple mutant A180S-G187K-P188A that displayed ~ 32 -fold increase in NMN⁺ catalytic efficiency while maintaining ~ 50 -fold cofactor specificity switch over the WT.

Through combination of molecular simulation predicting the NMN⁺ binding pose of mutants and experimental validation, we uncover general principles that can be applied to designing enzymes to utilize NMN⁺ and other artificial redox cofactors. Since NMN⁺ lacks the AMP portion of native cofactors critical to binding interactions, we must introduce compensatory mutations able

to form novel polar contacts to the remaining phosphate and ribose hydroxyl groups and reshape the binding pocket to more tightly contour the smaller NMN⁺. Although polar contacts can be formed with the carboxamide portion of NMN⁺, we avoid perturbing this region as it is directly involved in positioning the nicotinamide ring for the catalytic hydride transfer and is acutely sensitive to the surrounding environment. By exploring the placement of polar residues around the NMN⁺ phosphate, we found A180S that formed an inter-subunit hydrogen bond across the binding interface of separate monomers. Next, we sought to block binding of the native cofactor NAD⁺ by introducing bulky residues in the active site cleft where the AMP would typically bind. The G10R mutation extends off the loop between the first Rossmann beta strand and alpha helix, and forms a salt bridge with the conserved D33 that follows the second Rossmann beta strand. Typically the D33 participates in a bidentate hydrogen bonding interaction with the AMP ribose groups, this is a well-known interaction that determines the specificity for NAD⁺ over NADP⁺ as the acidic residue creates electrostatic repulsion with the negatively charged 2' phosphate group. NAD⁺ is excluded from binding here due to the steric clash formed by G10R occupying the pocket and the unavailability of D33 to form the conserved bidentate hydrogen bonds. The last variant A180S-G187K-P188A showed the highest NMN⁺ activity due to the additional salt bridge from G187K to the NMN⁺ phosphate group. Notably, this improved activity was observed to a lesser degree with A180S-G187K alone, highlighting the key role of P188A in facilitating greater loop flexibility for G187K to position optimally to bind NMN⁺.

We demonstrate two complementary approaches to the design and screening of enzymes for NMN⁺ activity. By using Rosetta to filter rational designs on only the first shell residues able to form polar contacts, we considerably reduce the number of variants to test. Previous work has relied heavily on chemical intuition to drive rational design, leading to long, laborious rounds of trial-and-error as human analysis of candidates has far lower throughput and less quantitative assessment than computational evaluation. Works based on purely random mutagenesis have also seen success, but these approaches rely on complicated screening approaches with stiff limitations such as only being amenable to engineering thermostable enzymes, and the mechanistic contributions of the accumulated mutations are difficult to interpret leading to low generalizability[8]. By exploring

computationally guided mutations focused on hotspot positions with limited residue substitutions at each, we balance the library search space to be experimentally tractable, yet random enough to allow for unpredictable cooperative interactions and major re-organization of the binding pocket. Both computational simulation and the colorimetric screen are extendable to other artificial redox cofactors and enzyme systems.

Future directions to advance this work will involve adapting the protocol to achieve higher throughput. Plate based assays are still limited to sampling several hundreds of samples per round. Design of enzymes with cofactor specificity switch between NAD⁺ and NADP⁺ has been performed with growth selection schemes where engineered bacterial strains experience cofactor imbalance and impeded growth[28–31]. By transforming the bacteria with an enzyme able to restore redox balance, growth is rescued. Growth selections enable the highest throughput of over 10⁶ variants per round and easy readout of growth where the most fit mutant will outcompete the less active variants in a mixed culture. With higher throughput, we will be able to more broadly explore sequence space and can investigate mutations predicted to have indirect, allosteric effects further away from the active site.

4.6 Supplementary information

4.6.1 SI Tables

Strains	Description	Reference
XL-1 Blue	Cloning strain	Stratagene
BL21 (DE3)	Protein expression strain	Invitrogen
BW25113	<i>E. coli</i> F-, DE(araD-araB)567, lacZ4787(del)::rrnB-3, LAM-, rph-1, DE(rhaD-rhaB)568, hsdR514	Datsenko et al.[32]

DS113	<i>E. coli</i> MG1655 Δ gapA::Cm Δ gapB::Erm	Seta et al.[25]
-------	---	-----------------

Table 4.2: gapA strains table. Ec, *Escherichia coli*.

Plasmids	Description	Reference
pQElac	Amp ^R ; ColE1 ori; <i>P_{LlacO1}</i> Expression vector	Li et al.[33]
pEK-28	pQElac 6xHis Ec gapA, Amp ^R	This study
pEK-30	pQElac 6xHis Ec gapA G10S, Amp ^R	This study
pEK-31	pQElac 6xHis Ec gapA G10R, Amp ^R	This study
pEK-32	pQElac 6xHis Ec gapA A180S, Amp ^R	This study
pEK-33	pQElac 6xHis Ec gapA A180R, Amp ^R	This study
pEK-34	pQElac 6xHis Ec gapA A180G-G187R, Amp ^R	This study
pEK-35	pQElac 6xHis Ec gapA V185R, Amp ^R	This study
pEK-36	pQElac 6xHis Ec gapA G120S, Amp ^R	This study
pEK-37	pQElac 6xHis Ec gapA G120T-L100A, Amp ^R	This study
pEK-38	pQElac 6xHis Ec gapA L100R, Amp ^R	This study
pEK-39	pQElac 6xHis Ec gapA L100H, Amp ^R	This study
pEK-52	pQElac 6xHis Ec gapA A180S-G10R, Amp ^R	This study
pEK-74	pQElac 6xHis Ec gapA A180S-G10R-G187K-P188A, Amp ^R	This study
HT-2	pQElac 6xHis Ec gapA A180S-G187K, Amp ^R	This study
HT-3	pQElac 6xHis Ec gapA A180S-G187L, Amp ^R	This study
HT-4	pQElac 6xHis Ec gapA A180S-I12V, Amp ^R	This study
HT-5	pQElac 6xHis Ec gapA A180S-G187A, Amp ^R	This study
HT-6	pQElac 6xHis Ec gapA A180S-G187A-P188G, Amp ^R	This study
HT-7	pQElac 6xHis Ec gapA A180S-P188A, Amp ^R	This study
HT-8	pQElac 6xHis Ec gapA A180S-G10K-I12T-G187R-P188I, Amp ^R	This study
HT-9	pQElac 6xHis Ec gapA A180S-G187K-P188A, Amp ^R	This study

HT-10	pQElac 6xHis Ec gapA A180S-G187A-P188A, Amp ^R	This study
HT-11	pQElac 6xHis Ec gapA A180S-I12V-G187R, Amp ^R	This study

Table 4.3: gapA plasmid table. Ec, *Escherichia coli*.

Name	Sequence	Count
HT-2	GIKP	3
HT-3	GILP	1
HT-4	GVGP	1
HT-5	GIAP	1
HT-6	GIAG	1
HT-7	GIGA	2
HT-8	KTRI	1
HT-9	GIKA	1
HT-10	GIAA	1
HT-11	GVRP	1
A180S	GIGP	20

Table 4.4: Counts of gapA variants obtained from high-throughput screen. Sequence indicates the amino acid observed at G10, I12, G187, and P188. Most samples found were the template A180S.

Enzyme	Cofactor	k_{cat} (s ⁻¹)	K_m (mM)	k_{cat}/K_m (mM ⁻¹ s ⁻¹)
HT9-G10R	NAD ⁺	$(5.1 \pm 0.1) \times 10^{-1}$	$(1.2 \pm 0.1) \times 10^{-2}$	42.7 ± 3.2
HT9-G10R	NMN ⁺	$(1.9 \pm 0.1) \times 10^{-2}$	9.4 ± 1.2	$(2.0 \pm 0.1) \times 10^{-3}$

Table 4.5: Michaelis-Menten parameters for HT9-G10R. HT9 is A180S-G187K-P188A and loses activity with addition of G10R.

4.6.2 SI Figures

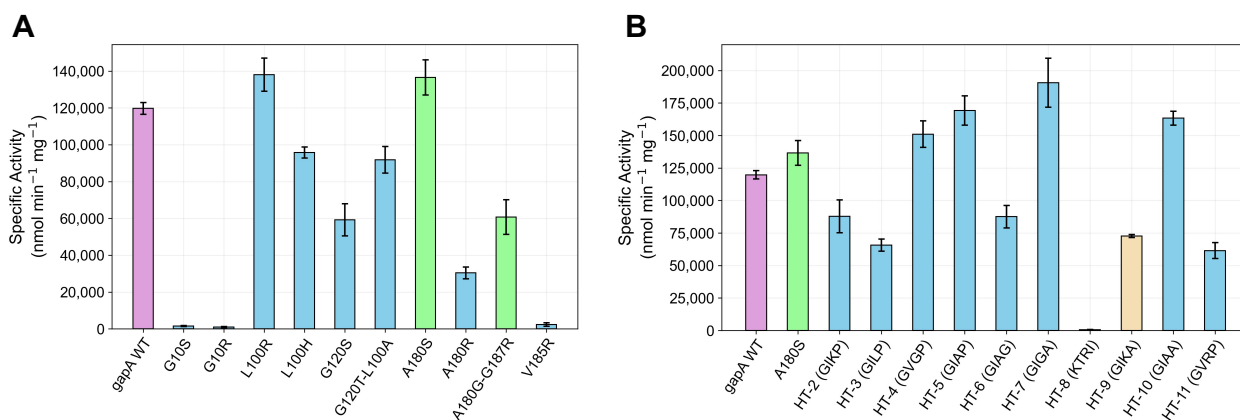


Figure 4.5: gapA specific activities with the native cofactor NAD⁺. **(A)** Activities from first round variants. A180S showed increased activity with both NAD⁺ and NMN⁺. **(B)** Samples from high-throughput colormetric screen.

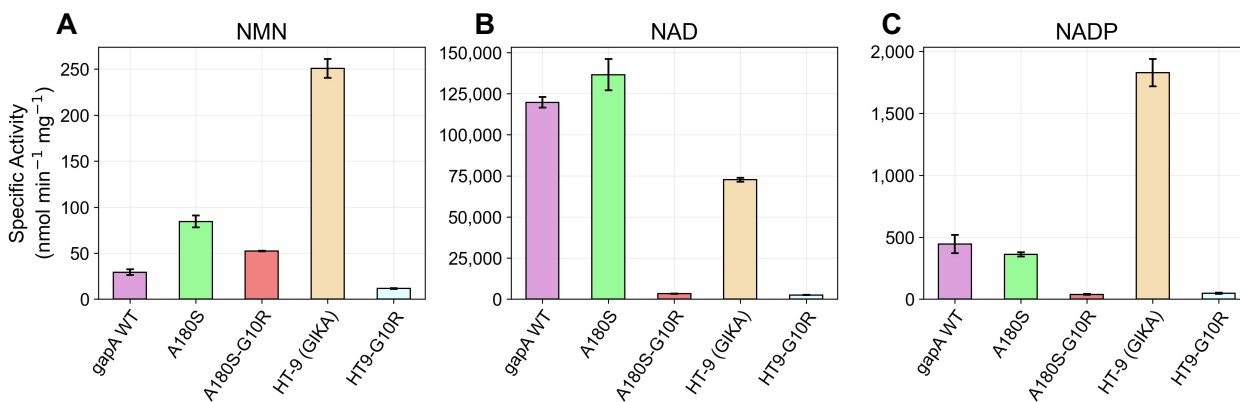


Figure 4.6: Designed gapA specific activities with NMN⁺, NAD⁺, and NADP⁺. **(A)** NMN⁺ specific activities. HT-9 was identified from high-throughput colorimetric assay. **(B)** NAD⁺ specific activities. The orthogonal variant A180S-G10R has substantial loss in NAD⁺ activity. **(C)** NADP⁺ specific activities. HT-9 unexpectedly gained increased ability to utilize NADP⁺.

Bibliography

- [1] Andrea M Chánique and Loreto P Parra. Protein engineering for nicotinamide coenzyme specificity in oxidoreductases: Attempts and challenges. *Front. Microbiol.*, 9:194, February 2018.
- [2] Paola Laurino, Ágnes Tóth-Petróczy, Rubén Meana-Pañeda, Wei Lin, Donald G Truhlar, and Dan S Tawfik. An ancient fingerprint indicates the common ancestry of Rossmann-Fold enzymes utilizing different Ribose-Based cofactors. *PLoS Biol.*, 14(3):e1002396, March 2016.
- [3] Meng Wang, Biqiang Chen, Yunming Fang, and Tianwei Tan. Cofactor engineering for more efficient production of chemicals and biofuels. *Biotechnol. Adv.*, 35(8):1032–1039, December 2017.
- [4] Edward King, Sarah Maxel, and Han Li. Engineering natural and noncanonical nicotinamide cofactor-dependent enzymes: design principles and technology development. *Curr. Opin. Biotechnol.*, 66:217–226, September 2020.
- [5] Tanja Knaus, Caroline E Paul, Colin W Levy, Simon de Vries, Francesco G Mutti, Frank Hollmann, and Nigel S Scrutton. Better than nature: Nicotinamide biomimetics that outperform natural coenzymes. *J. Am. Chem. Soc.*, 138(3):1033–1039, January 2016.
- [6] William B Black, Linyue Zhang, Wai Shun Mak, Sarah Maxel, Youtian Cui, Edward King, Bonnie Fong, Alicia Sanchez Martinez, Justin B Siegel, and Han Li. Engineering a nicotinamide mononucleotide redox cofactor system for biocatalysis. *Nat. Chem. Biol.*, 16(1):87–94, January 2020.

- [7] Lei Wang, Debin Ji, Yuxue Liu, Qian Wang, Xueying Wang, Yongjin J Zhou, Yixin Zhang, Wujun Liu, and Zongbao K Zhao. Synthetic Cofactor-Linked metabolic circuits for selective energy transfer. *ACS Catal.*, 7(3):1977–1983, March 2017.
- [8] Rui Huang, Hui Chen, David M Upp, Jared C Lewis, and Yi-Heng P Job Zhang. A High-Throughput method for directed evolution of NAD(P)+-Dependent dehydrogenases for the reduction of biomimetic nicotinamide analogues. *ACS Catal.*, 9(12):11709–11719, December 2019.
- [9] Yao Liu, Yalong Cong, Chuanxi Zhang, Bohuan Fang, Yue Pan, Qiangzi Li, Chun You, Bei Gao, John Z H Zhang, Tong Zhu, and Lujia Zhang. Engineering the biomimetic cofactors of NMNH for cytochrome P450 BM3 based on binding conformation refinement. *RSC Adv.*, 11(20):12036–12042, March 2021.
- [10] Yuxue Liu, Yanbin Feng, Lei Wang, Xiaojia Guo, Wujun Liu, Qing Li, Xueying Wang, Song Xue, and Zongbao Kent Zhao. Structural insights into phosphite dehydrogenase variants favoring a non-natural redox cofactor. *ACS Catal.*, 9(3):1883–1887, March 2019.
- [11] Claudia Nowak, André Pick, Petra Lommès, and Volker Sieber. Enzymatic reduction of nicotinamide biomimetic cofactors using an engineered glucose dehydrogenase: Providing a regeneration system for artificial cofactors. *ACS Catal.*, 7(8):5202–5208, August 2017.
- [12] Yuxue Liu, Xiaojia Guo, Wujun Liu, Junting Wang, and Zongbao Kent Zhao. Structural insights into malic enzyme variants favoring an unnatural redox cofactor. *ChemBiochem*, 22(10):1765–1768, May 2021.
- [13] Xueying Wang, Yanbin Feng, Xiaojia Guo, Qian Wang, Siyang Ning, Qing Li, Junting Wang, Lei Wang, and Zongbao K Zhao. Creating enzymes and self-sufficient cells for biosynthesis of the non-natural cofactor nicotinamide cytosine dinucleotide. *Nat. Commun.*, 12(1):2116, April 2021.
- [14] Yuxue Liu, Qing Li, Lei Wang, Xiaojia Guo, Junting Wang, Qian Wang, and Zongbao K Zhao.

- Engineering d-lactate dehydrogenase to favor an non-natural cofactor nicotinamide cytosine dinucleotide. *Chembiochem*, 21(14):1972–1975, July 2020.
- [15] Debin Ji, Lei Wang, Shuhua Hou, Wujun Liu, Jinxia Wang, Qian Wang, and Zongbao K Zhao. Creation of bioorthogonal redox systems depending on nicotinamide flucytosine dinucleotide. *J. Am. Chem. Soc.*, 133(51):20857–20862, December 2011.
- [16] Xiaojia Guo, Yuxue Liu, Qian Wang, Xueying Wang, Qing Li, Wujun Liu, and Zongbao K Zhao. Non-natural cofactor and Formate-Driven reductive carboxylation of pyruvate. *Angew. Chem. Int. Ed Engl.*, 59(8):3143–3146, February 2020.
- [17] E Duée, L Olivier-Deyris, E Fanchon, C Corbier, G Branlant, and O Dideberg. Comparison of the structures of wild-type and a N313T mutant of escherichia coli glyceraldehyde 3-phosphate dehydrogenases: implication for NAD binding and cooperativity. *J. Mol. Biol.*, 257(4):814–838, April 1996.
- [18] Mikyung Yun, Cheon-Gil Park, Ji-Yeon Kim, and Hee-Won Park. Structural analysis of glyceraldehyde 3-phosphate dehydrogenase from escherichia coli: Direct evidence of substrate binding and Cofactor-Induced conformational changes,. *Biochemistry*, 39(35):10702–10710, September 2000.
- [19] S Clermont, C Corbier, Y Mely, D Gerard, A Wonacott, and G Branlant. Determinants of coenzyme specificity in glyceraldehyde-3-phosphate dehydrogenase: role of the acidic residue in the fingerprint region of the nucleotide binding fold. *Biochemistry*, 32(38):10178–10184, September 1993.
- [20] Kirill E Medvedev, Lisa N Kinch, R Dustin Schaeffer, and Nick V Grishin. Functional analysis of rossmann-like domains reveals convergent evolution of topology and reaction pathways. *PLoS Comput. Biol.*, 15(12):e1007569, December 2019.
- [21] Liam M Longo, Jagoda Jabłońska, Pratik Vyas, Manil Kanade, Rachel Kolodny, Nir Ben-Tal, and Dan S Tawfik. On the emergence of P-Loop NTPase and rossmann enzymes from a Beta-Alpha-Beta ancestral fragment. *Elife*, 9, December 2020.

- [22] Michael S Packer and David R Liu. Methods for the directed evolution of proteins. *Nat. Rev. Genet.*, 16(7):379–394, July 2015.
- [23] Sandra W Cowan-Jacob, Markus Kaufmann, Anthony N Anselmo, Wilhelm Stark, and Markus G Grütter. Structure of rabbit-muscle glyceraldehyde-3-phosphate dehydrogenase. *Acta Crystallogr. D Biol. Crystallogr.*, 59(Pt 12):2218–2227, December 2003.
- [24] Sarel J Fleishman, Andrew Leaver-Fay, Jacob E Corn, Eva-Maria Strauch, Sagar D Khare, Nobuyasu Koga, Justin Ashworth, Paul Murphy, Florian Richter, Gordon Lemmon, Jens Meiler, and David Baker. RosettaScripts: A scripting language interface to the rosetta macromolecular modeling suite. *PLoS One*, 6(6):e20161, June 2011.
- [25] F D Seta, S Boschi-Muller, M L Vignais, and G Branlant. Characterization of escherichia coli strains with gapa and gapb genes deleted. *J. Bacteriol.*, 179(16):5218–5221, August 1997.
- [26] Kimberly M Mayer and Frances H Arnold. A colorimetric assay to quantify dehydrogenase activity in crude cell lysates. *J. Biomol. Screen.*, 7(2):135–140, April 2002.
- [27] Brian Koepnick, Jeff Flatten, Tamir Husain, Alex Ford, Daniel-Adriano Silva, Matthew J Bick, Aaron Bauer, Gaohua Liu, Yojiro Ishida, Alexander Boykov, Roger D Estep, Susan Kleinfelter, Toke Nørgård-Solano, Linda Wei, Foldit Players, Gaetano T Montelione, Frank DiMaio, Zoran Popović, Firas Khatib, Seth Cooper, and David Baker. De novo protein design by citizen scientists. *Nature*, June 2019.
- [28] Linyue Zhang, Edward King, Ray Luo, and Han Li. Development of a High-Throughput, in vivo selection platform for NADPH-Dependent reactions based on redox balance principles. *ACS Synth. Biol.*, 7(7):1715–1721, July 2018.
- [29] Sarah Maxel, Samer Saleh, Edward King, Derek Aspacio, Linyue Zhang, Ray Luo, and Han Li. Growth-Based, High-Throughput selection for NADH preference in an Oxygen-Dependent biocatalyst. *ACS Synth. Biol.*, September 2021.
- [30] Sarah Maxel, Edward King, Yulai Zhang, Ray Luo, and Han Li. Leveraging oxidative stress

- to regulate redox Balance-Based, in vivo growth selections for oxygenase engineering. *ACS Synth. Biol.*, October 2020.
- [31] Sarah Maxel, Derek Aspacio, Edward King, Linyue Zhang, Ana Paula Acosta, and Han Li. A Growth-Based, High-Throughput selection platform enables remodeling of 4-hydroxybenzoate hydroxylase active site. *ACS Catal.*, 10(12):6969–6974, June 2020.
- [32] Kirill A Datsenko and Barry L Wanner. One-step inactivation of chromosomal genes in escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.*, 97(12):6640–6645, June 2000.
- [33] Han Li and James C Liao. Engineering a cyanobacterium as the catalyst for the photosynthetic conversion of CO₂ to 1,2-propanediol. *Microb. Cell Fact.*, 12:4, January 2013.

Chapter 5

Analysis of mutations altering oxygenase conformational dynamics and substrate specificity

Adapted from:

Growth-Based, High-Throughput Selection for NADH Preference in an Oxygen-Dependent Biocatalyst

Authors: Sarah Maxel*, Samer Saleh*, Edward King*, Derek Aspacio, Linyue Zhang, Ray Luo, Han Li

ACS Synth Biol. 2021;10: 2359–2370.

doi: [10.1021/acssynbio.1c00258](https://doi.org/10.1021/acssynbio.1c00258)

Publication Date (Web): September 1, 2021

and

Leveraging Oxidative Stress to Regulate Redox Balance-Based, *In Vivo* Growth Selections for Oxygenase Engineering

Authors: Sarah Maxel*, Edward King*, Yulai Zhang, Ray Luo, Han Li

ACS Synth Biol. 2020;9: 3124–3133.

doi: [10.1021/acssynbio.0c00380](https://doi.org/10.1021/acssynbio.0c00380)

Publication Date (Web): September 23, 2020

5.1 Abstract

Directed evolution methods based on high-throughput growth selection enable efficient discovery of enzymes with improved function *in vivo*. High-throughput selection is particularly useful when engineering oxygenases, which are sensitive to structural perturbations and prone to uncoupled activity. Through redox balance-based growth selection of variants generated through site-saturation mutagenesis, we engineered: 1) P450-BM3 to degrade acenaphthene (ACN), a recalcitrant environmental pollutant, and 2) cyclohexanone monooxygenase (CHMO) to favor NADH over the more expensive NADPH. The P450-BM3 variants GVQ-AL (A74G-F87V-L188Q-V78A-A328L) and GVQ-D222N (A74G-F87V-L188Q-D222N), which have both improved coupling efficiency and catalytic activity compared to the starting variant, were discovered. Computational modeling indicates that the discovered mutations cooperatively optimize binding pocket shape complementarity to ACN, and shift the protein's conformational dynamics to favor the lid-closed, catalytically competent state. The CHMO variant DTNP (S209D-K326T-K349N-L143P) with a \sim 1,200-fold relative cofactor specificity switch from NADPH to NADH was identified and rationalized to be due to concerted fine-tuning of cofactor contacts.

5.2 Introduction

Biooxygenation provides a viable alternative to traditional means of synthetic chemistry for the selective activation of C–H bonds. Members of the diverse oxygenase families such as two-component flavin hydroxylases[1, 2], Cytochrome P450s[3], and Baeyer–Villiger monooxygenases (BVMOs)[4]

show considerable potential as industrial catalysts. Directed evolution has been widely exploited to tailor these oxygenases for desired reactions; however, their full potential is limited by the relatively low throughput of downstream selection technologies. To address this issue, advancements in designing efficient libraries with smaller theoretical sizes[5] and and ultrahigh-throughput screening methods utilizing microfluidic devices and fluorescence sorting[6–9] have been developed. Although these processes have facilitated the successful directed evolution of a number of enzymes, there is a need for more general and accessible methods that do not require specialized reagents or expensive equipment.

Growth-based selection methods are high-throughput (10^6 candidates per round) and use growth as a facile readout. Importantly, they directly yield enzymes that are active *in vivo*. However, the selection platforms currently do not account for the unique engineering requirements of oxygenases. Oxygenases execute a complex orchestration of reaction mechanisms to couple NAD(P)H consumption to substrate conversion. This coupling is often disrupted in engineered oxygenases, resulting in futile NAD(P)H consumption to reduce O_2 in lieu of substrate conversion[10, 11]. The frequency of the completed reaction cycle can be described by the product formed relative to the NADPH consumed and is reported as the coupling efficiency. In existing platforms, it is unclear whether the growth-based selection methods can accommodate the need for improving coupling efficiency.

Cytochrome P450s are a promiscuous class of heme-containing monooxygenases that are able to incorporate molecular oxygen with high regio- and stereoselectivity onto inert C–H bonds. The naturally chimeric, NADPH-dependent P450 BM3 is prolific as an engineering target because it is self-sufficient, readily soluble, and possesses the highest turnover rate of any known P450[12]. Extensive studies aimed at engineering BM3 to modify substrate scope beyond its native fatty acid preference have faced a reoccurring challenge to control electron coupling of the catalytic cycle[12]. During the reaction cycle, the enzyme is activated by the transfer of electrons from NAD(P)H through a FAD and subsequent FMN coenzyme on the reductase domain. Following this initial transfer, electrons are mediated through a series of residues serving as electron transport pathways to the heme center of the P450 domain. Upon activation, the reactive heme is able to bind oxygen

and generate an iron–oxygen complex that proceeds through a series of reductive steps. When disrupted through engineering, the precise electron transfer dynamics can be shifted out of tune with substrate binding, and the activated oxygen can decay into reactive oxygen species (ROS). Despite their broad substrate scope, engineered BM3 variants with non-native substrates frequently demonstrate extremely low coupling between NADPH consumption and product formation[12]. BM3 engineering strategies frequently target bulky active site residues close to the internal heme such as F87 for mutagenesis to smaller residues to allow substrate access to the reactive center. Although this approach is effective, these preliminary mutants do not always provide the expected promiscuity[13], and frequently display low coupling with the desired non-native substrates. Subsequent engineering typically aims to reshape the binding pocket with targeted mutations to establish improved coupling through complementary steric packing. Because excessive uncoupling wastes reducing power and produces reactive oxygen species (ROS) that can inactivate the enzyme, it is essential to pursue both high activity and high coupling efficiency simultaneously in oxygenase engineering.

We simultaneously improved both the activity and the coupling efficiency of an engineered P450-BM3 variant for an unnatural substrate, acenaphthene (ACN). The discovered P450-BM3 variant may be applied in environmental remediation to facilitate the degradation of this persistent pollutant[14, 15]. Computational modeling indicates that we obtained synergistic mutations reshaping the substrate binding pocket and identified mutations distal to the active site shifting the global conformational dynamics of P450-BM3. This highlights the power of high-throughput selection methods for directed evolution in discovering mutations that optimize multiple criteria in concert.

Acinetobacter sp. cyclohexanone monooxygenase (*Ac* CHMO), is a NADPH-dependent Baeyer–Villiger monooxygenase (BVMO) with broad applications[16–18]. A major limitation of *Ac* CHMO is that it strictly prefers NADPH[19]. At large scale *in vitro*, NADPH is less stable and more costly than NADH[20], and *in vivo*, the rate of NADPH regeneration relative to NADH is lower in ubiquitous microbial chassis such as *Escherichia coli* and *Bacillus subtilis*[5, 20]. CHMO has been shown to be notoriously averse to cofactor specificity switching with existing methods[19, 21, 22], and a high-

throughput selection is advantageous for exploring alternative engineering strategies. Through the development and utilization of an aerobic, NADH-dependent selection platform, we obtained *Ac* CHMO DTNP (S208D-K326T-K349N-L143P with $\sim 1,200$ -fold improvement in catalytic efficiency (k_{cat}/K_M) for NADH over NADPH in comparison to CHMO WT. Molecular dynamics (MD) simulation suggest that the selected mutations may function by tuning the conformational dynamics of the protein and the cofactors, which would not be readily predicted in structure-guided protein design. For example, the key mutation L143P in CHMO DTNP emerged as a spontaneous mutation outside the three rationally picked positions (S208, K326, K349) for site-saturated mutagenesis. Although L143P does not directly interact with NADH, MD analysis suggests that it tunes the conformation of the flavin adenine dinucleotide (FAD) in CHMO DTNP, allowing more efficient hydride transfer from NADH. These effects would not be evident through analysis of static models, illustrating the difficulty of engineering functionally innovative variants through rational design and the critical role of high-throughput selections.

5.3 Methods

5.3.1 Docking acenaphthene into P450-BM3

The substrate acenaphthene (ACN) was docked into the P450-BM3 binding pocket with Rosetta[23]. The crystal structure 1ZO9[24] of P450-BM3 with N-palmitoylmethionine (EPM) bound in the closed conformation was used as the starting template. The heme was modeled in the catalytically active Fe(IV)-oxo compound 1 state. The coordinates for EPM were removed from the structure, the ACN model was downloaded from the PubChem database[25], and ACN was initially placed in the open region above the heme oxygen. The Rosetta docking protocol consisted of mutation from the WT structure, perturbation of the ACN binding pose through random translation and rotation, and optimization of active site rotamers through Monte Carlo evaluation. A distance restraint was imposed between the heme oxygen and ACN reactive carbons (due to ligand symmetry two carbons have the potential to be hydroxylated) to focus sampling on the protein–ligand intermediate state

preceding catalysis, and full flexibility for protein backbone torsions was allowed. A total of 1,000 docking trials were completed for each variant, the top 100 models filtered on total Rosetta energy were further sorted based on ACN interface energy, and the model with the most favorable interface energy was selected as the reference.

5.3.2 *Ac* CHMO homology modeling

The model of WT *Ac* CHMO with cofactors FAD and NADPH bound was generated with Rosetta CM[23, 26]. Threading templates with high sequence identity and cocrystallized cofactors were identified through BLASTP search of the Protein Data Bank with *Ac* CHMO as the query[27, 28]. Crystal structures of CHMO from *Rhodococcus sp.* (PDB: 4RG3, 3GWD, 3GWF, 3UCL), which shares 57.8% sequence identity to *Ac* CHMO, were selected as input models[29–31]. The Rosetta CM protocol consisted of repeated rounds where the target *Ac* CHMO sequence was threaded onto the template structure based on MAFFT sequence alignment, segments of the protein structure were constructed through insertion of fragments drawn from the library provided by the templates through Monte Carlo evaluation, followed by minimization to relax the final output[32]. 1,500 homology modeling trajectories were completed, the output structure with the most favorable total Rosetta energy was selected as the representative model for all further analysis. Point mutations for the *Ac* CHMO variants were produced through 1,000 further Rosetta docking simulations on the homology model with backbone flexibility, side chain repacking, and ligand minimizations; final models were selected by lowest total Rosetta energies. The NADH binding poses were prepared by deleting atoms of the ribose 2' phosphate group on the existing NADPH models prior to the Rosetta Design trials.

5.3.3 Molecular dynamics simulations

MD simulations were completed with PMEMD[33] from the AMBER 18[34] package utilizing the ff14sb force field[35] and 8 Å Particle Mesh Ewald real space cutoff[36]. Protonation states of

titratable residues were determined with the H++ web server[37]. The TLEAP program was utilized to solvate the complexes with TIP3P water molecules in a truncated octahedron with 10 Å buffer and neutralizing Na⁺/Cl⁻ counterions. Minimization was performed in two stages, first with 2,500 steps of steepest descent and 2,500 steps of conjugate gradient with non-hydrogen solute atoms restrained with a 20 kcal mol⁻¹ Å⁻² force to relieve solvent clash. The second stage minimization to remove solute steric clashes was run with the same cycle settings and restraints removed. Heating from 0 to 298 K was performed over 0.5 ns with 10 kcal mol⁻¹ Å⁻² restraints on all non-hydrogen solute atoms under NPT conditions at 1 atm pressure with Langevin thermostat and 1 fs time step. Structural artifacts from the heating step were cleared with solvent density equilibration over 5 ns with 5 kcal mol⁻¹ Å⁻² restraints on all solute atoms and an unrestrained 10 ns equilibration using 2 fs time step. Production MD trajectories were carried out with SHAKE restraints on hydrogens, NVT ensemble, Langevin thermostat with collision frequency 1.0 ps⁻¹, and periodic boundary conditions.

P450-BM3 simulation were run with Compound 1 heme parameters obtained from Shahrokh et al.[38] for 250 ns with 2 fs time step. The apo models for P450-BM3 variants were initiated from the open state with 2HPD[39] as the template, amino acid substitutions were introduced with Rosetta, and structures were relaxed to relieve unfavorable contacts. The holostructures for P450-BM3 variants with ACN bound started from the representative Rosetta docking models. Average structures based on α carbon coordinates were calculated for each apo trajectory by aligning all snapshots to the starting frame and averaging the α carbon positions. Direct coordinate averaging distorts bond angles and lengths, to depict a realistic model we identified the frame with the minimum α carbon RMSD to the calculated average model as an instance of the mean structure. Lid opening distances describing the positioning of the G helix were recorded as the length between PRO146 and PRO45 alpha carbons. Root mean square fluctuation (RMSF) over α carbon atoms was measured for trajectories of the holomodels and aggregated over secondary structure elements to compare flexibility. *Ac* CHMO production MD simulations were run for 400 ns with 2 fs timestep.

5.3.4 *Ac* CHMO cofactor binding analysis

Cofactor binding comparisons were completed between WT *Ac* CHMO, DTN CHMO, and DTNP CHMO with either NADH or NADPH bound. Protein backbone flexibility over the trajectories was measured through α -carbon RMSF. Hydride transfer potential was recorded as the distance between the nicotinamide C4 to FAD N5. Metastable FAD binding conformations were established through featurization on minimum heavy atom distance to residues with any atom within 4 Å contact and PCA dimensionality reduction with K-means clustering. The distance array was standardized to zero mean and unit variance and transformed to lower dimensional space with components maintaining maximal variance through PCA. K-means clustering was performed to discretize the sampled conformations projected onto the free energy landscape of the first two PCA components into metastable states. The optimal number of clusters was selected by the elbow heuristic where clustering over a range of K values, from one to nine here, is completed and the sum of squared distances from the sample points to their assigned cluster center is computed. The value of K where the sum of squared distances decrease becomes linear is selected as optimal and indicates that increasing K further will result in overfitting. The residue positions neighboring FAD include: 12, 13, 15, 16, 17, 36, 37, 44, 45, 46, 48, 49, 51, 56, 57, 58, 63, 109, 110, 140, 141, 142, 390, 426, 434, 435, 436, and 439.

5.4 Results and discussion

5.4.1 P450-BM3 GVQ-AL shows improved binding pocket complementarity to ACN

We hypothesized that P450-BM3 GVQ-AL (A74G-F87V-L188Q-V78A-A328L), the best variant obtained from substrate active site (SAS) engineering library, has improved active site shape complementarity to ACN (SI Figure: 5.6). To investigate this hypothesis, we performed computational docking with Rosetta to model the substrate binding pose.

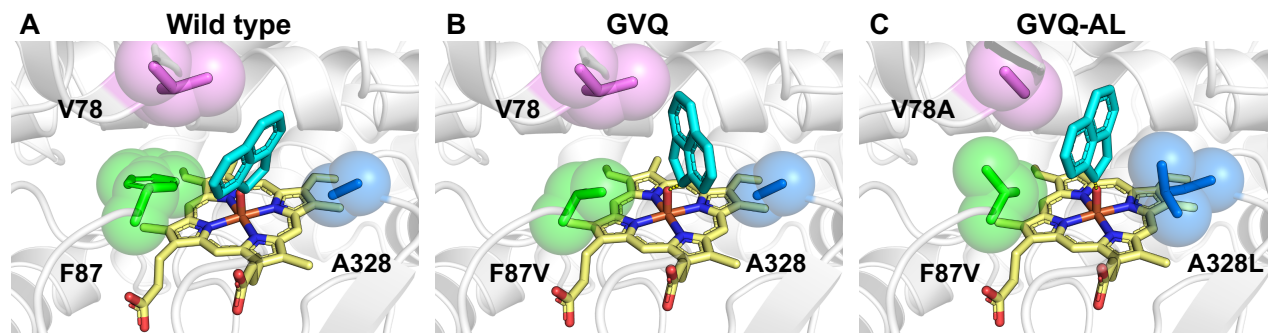


Figure 5.1: Docked models of the P450-BM3 ACN binding poses. (A) In the WT, steric clash from the bulky F87 and V78 block the ACN from positioning the catalytic carbon over the heme oxygen. (B) In GVQ, the mutation F87V results in greater volume for the ligand to maneuver over the heme oxygen, but the extension of V78 prevents the ACN left surface from reaching optimal hydrophobic packing against F87V. (C) GVQ-AL displays improved binding pocket shape complementarity for ACN binding, F87V and V78A reduce steric hindrance, while A328L provides increased nonpolar surface area to face the ACN rings.

Docking of ACN in the wild type P450-BM3 results in an unproductive binding pose, with the ACN catalytic carbons positioned too far (>3.0 Å) from the heme oxygen for catalysis (Figure: 5.1A). In P450-BM3 GVQ (A74G-F87V-L188Q), which was the starting point of engineering, the F87V mutation partially relieves steric hindrance, but the hydrophobic packing against ACN is not optimal (Figure: 5.1B). By targeting V78 and A328 for saturation mutagenesis, the active site is further contoured to narrowly enclose ACN and limit unfavorable solvent interactions or excessive ligand mobility, while creating headspace for the ligand to readily maneuver the catalytic carbon over the heme oxygen (Figure: 5.1C). V78A clears vertical space to accommodate ACN and allows F87V to adopt a different rotamer state compared to in the GVQ model, which creates even more volume. On the other side, A328L, with its increased bulk, packs tightly against the face of ACN.

5.4.2 P450-BM3 GVQ-D222N favors the catalytically active, lid-closed state

Since P450-BM3 GVQ-D222N (A74G-F87V-L188Q-D222N), the best variant obtained from electron transport (ET) pathway modulation library, has key mutations distal from the active site, we

hypothesized that their beneficial effect arises from altering the enzyme conformational dynamics to more frequently sample catalytically productive states. To evaluate this hypothesis, we performed molecular dynamics (MD) simulation(Figure: 5.2).

The flexible nature of P450-BM3 is reflected by crystal structures with the F helix, F/G loop, and G helix regions (Figure: 5.2A) , which are known to act as a lid that moves during catalytic cycle, in varying positions with the “closed” lid often associated with the substrate bound, catalytically active state[12] (Figure: 5.7). We compared representative models of wild type and GVQ-D222N apo enzymes and our simulations indicate that the mutations promote lid closing (Figure: 5.2A, B). The A74G-L188Q mutations function cooperatively, forming a novel hydrogen bond between the B' helix and the F helix. The function of the hydrogen bond is suggested to fasten the F helix, minimizing the mobility of the lid. The D222N mutation occurs at the base of the G helix and forms a novel polar contact with the backbone carbonyl of K218, which may function to anchor the G helix and reduce lid flexibility (Figure: 5.2A).

The effect of the mutations on P450-BM3 dynamics is first evaluated from trajectories started from the lid-open, no substrate bound states. We compare the distribution of lid distances in wild type (WT), GVQ, and GVQ-D222N (Figure: 5.2B). WT samples the largest lid distances with an average of 21.1 Å, indicating that it favors maintaining the unproductive open state; GVQ samples more intermediate distances averaging 20.0 Å, suggesting greater disposition to closing than the WT; and GVQ-D222N favors occupying the fully closed conformation with average lid distance 18.6 Å. These results are consistent with the hypothesis that novel hydrogen bonds formed by the evolved mutations are critical to decreasing the free energy barrier of transitioning from open to closed forms and stabilizes the closed state.

We then analyzed trajectories started from the lid-closed, ACN bound state. While all enzyme variants tend to maintain the closed state, the stability of the lid varies (Figure: 5.2C). The FG loop shows similar root mean square fluctuation (RMSF) for all samples, but the RMSF for the F helix declines from 1.34 Å in the WT, to 1.18 Å in GVQ, and finally to 0.93 Å in GVQ-D222N, while the G helix RMSF trends identically with 1.55 Å for WT, 1.44 Å for GVQ, and 1.28 Å for

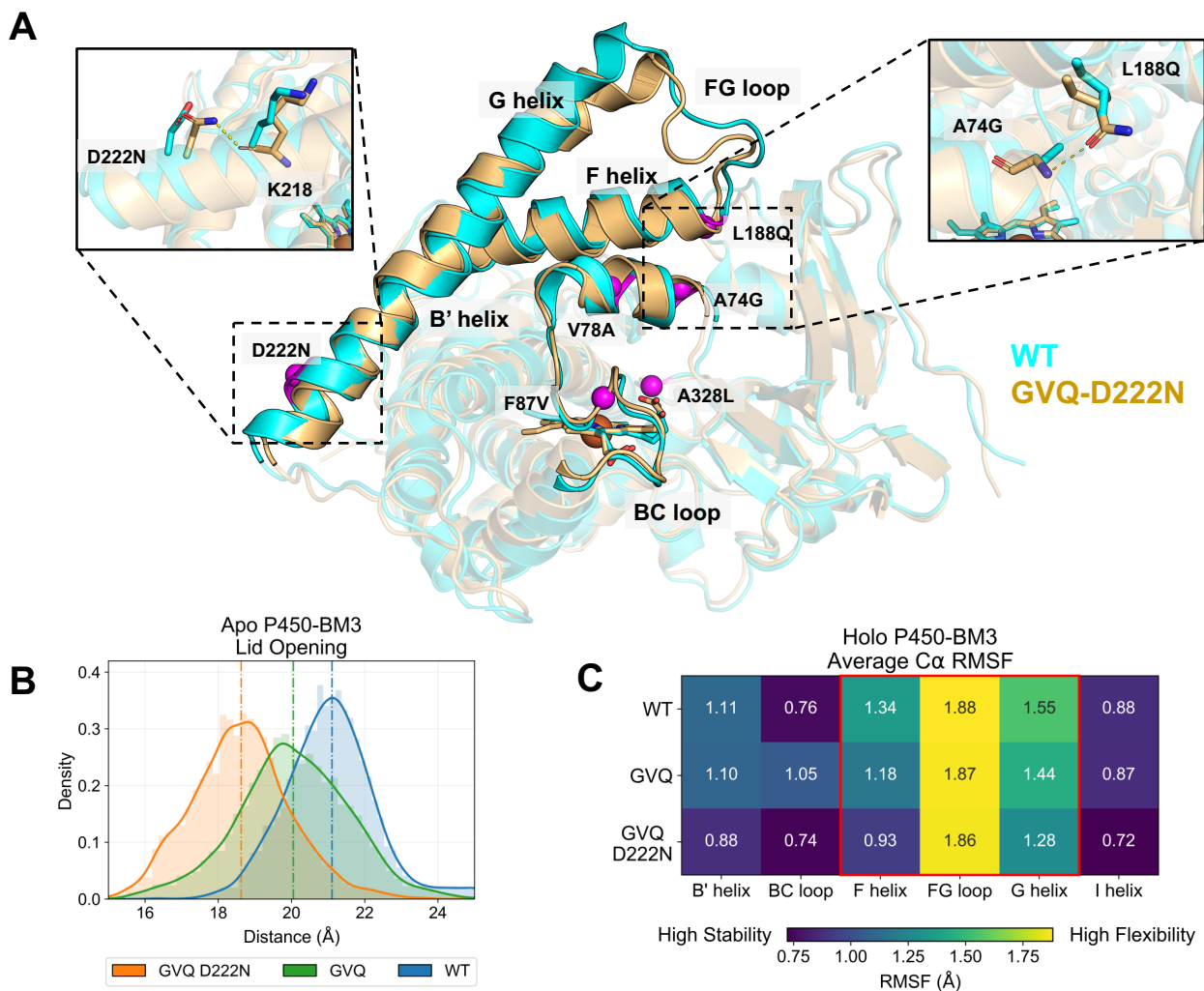


Figure 5.2: Selected mutations alter P450-BM3 conformational dynamics. **(A)** Overlay of WT (cyan) and GVVQ-D222N (orange) average structures from MD simulations of the apo structures with mutated positions highlighted as purple spheres. GVVQ-D222N favors adopting the catalytically active closed conformation with the G helix lowered while the WT maintains the open conformation. L188Q forms a hydrogen bond to A74G to fasten the F helix, and D222N potentially acts as an anchor to curb lid opening motions by establishing a backbone hydrogen bond with the K218 carbonyl to stabilize the base of the G helix. **(B)** The lid distance characterizing the substrate channel opening is defined as the length between PRO196 and PRO45 alpha carbons. GVVQ-D222N samples conformations resembling the active closed state, while GVVQ experiences intermediate states, and the WT tends to stay open. **(C)** The WT holoprotein shows the highest mobility at the F/G-helix as measured by α carbon RMSF. GVVQ has reduced flexibility at the lid regions, and GVVQ-D222N is the most stable throughout.

GVQ-D222N. The decrease in RMSF upon accumulation of the selected mutations supports the role of the novel hydrogen bonds in stabilizing the closed state and reducing excess flexibility of the lid region.

5.4.3 *Ac* CHMO cofactor binding pose

The *Ac* CHMO homology model shows that NADPH binds in an extended conformation with the nicotinamide ring tucked into a small binding pocket against the FAD flavin. The NADPH binding mode is characterized by the conserved Rossman fold with β - α - β secondary structure motifs enclosing the cofactor. The primary interactions predicted to hold the NADPH in place include: Q190 side-chain amide which contacts the NADPH carboxamide oxygen, W490 indole which forms a hydrogen bond to the nicotinamide ribose hydroxyl, backbone polar interactions at the N-terminus of the Rossman α -helix and hydrogen bonding from the T189 hydroxyl to the pyrophosphate, a salt-bridge from K326 extending from a loop across the substrate channel to the pyrophosphate, and R207 guanidinium forming a salt-bridge to the ribose 2' phosphate group and packing against the adenosine ring. The side of the NADPH facing away from the Rossman fold is marginally exposed, with the control loop defined as residues 489–505 lightly packing against the cofactor. Since NADH differs from NADPH only in the absence of the 2' phosphate group and is capable of establishing the same set of binding interactions along the pyrophosphate and nicotinamide ring, we postulate that *Ac* CHMO's strict specificity for NADPH is driven by the R207 contact spanning from the end of the second Rossman β -strand stabilizing the adenosine tail of the cofactor for optimal packing dynamics with the control loop. The FAD is held opposite of the nicotinamide cofactor with the flavin group facing the nicotinamide ring and is tightly bound through polar contacts throughout the adenosine tail, ribose, pyrophosphate, ribitol, and flavin carbonyls. The numerous hydrogen bonds restrict FAD mobility, position the flavin for efficient hydride transfer with the nicotinamide cofactor or substrate, and prevent the FAD release. The cyclohexanone binding pocket is adjacent to the nicotinamide ring, it is proposed that some degree of protein and NADPH flexibility is required to allow the substrate to move close enough to the FAD for electron transfer.

Rosetta modeling suggests that S208D and K349N function cooperatively to recognize NADH, with K349N supporting the S208D loop through backbone hydrogen bond and S208D forming a novel bidentate hydrogen bond to the NADH adenosine ribose that resembles interactions seen in native NADH specific proteins (Figure: 5.3A, B). The contribution of K326T and L143P are not evident from static structural analysis since they do not directly contact NADH (Figure: 5.3A, C, D), which motivated further analysis through MD simulations.

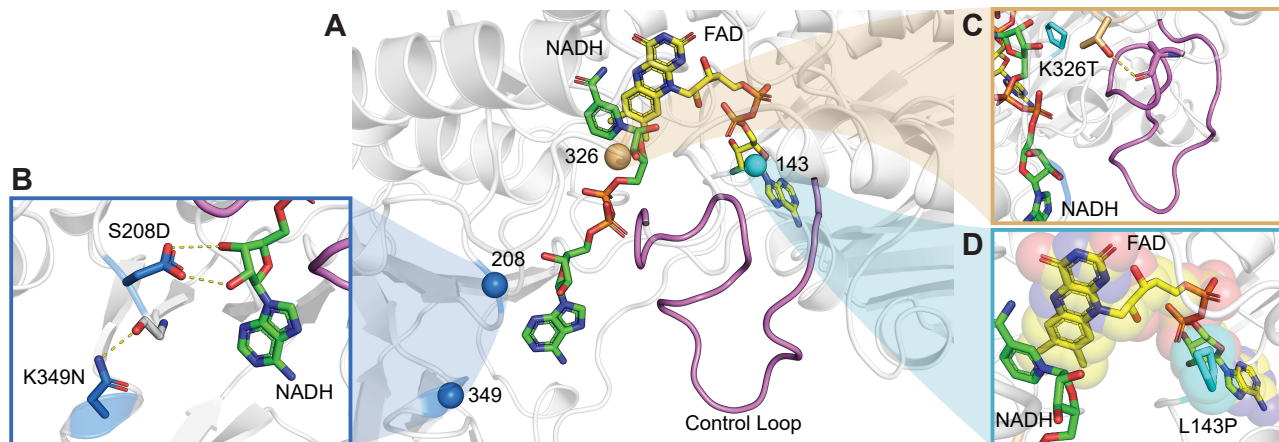


Figure 5.3: Homology model of CHMO DTNP. (A) Overview of the cofactor binding pocket, NADH (in green) is pressed on one end by the control loop (pink). FAD (yellow) is held opposite of the NADH, and positions of the selected mutations are represented as spheres. (B) K349N supports the loop that S208D sits on, and S208D makes a bidentate hydrogen bond to the NADH adenosine ribose. (C) K326T initiates a hydrogen bond to the control loop at the backbone carbonyl of A487. The additional hydrogen bond may cause the control loop to favor maintaining the catalytically relevant, closed conformation. (D) L143P impacts the conformations that FAD can adopt.

5.4.4 Mutations reshape *Ac* CHMO conformational landscape and hydride transfer potential

K326T contacts the “control loop” through A487 (Figure: 5.3A, C). The control loop is observed to be ordered in some CHMO crystal structures but disordered in others[29, 31, 40], suggesting that its flexibility varies depending on the enzyme’s position in catalytic cycle. Specifically, it is hypothesized that the control loop must become rigid to hold the NADPH cofactor and cyclohexanone during the hydride transfer stages[40]. We analyzed the flexibility of the control loop through

α -carbon root-mean-square fluctuation (RMSF), and plot the difference between CHMO DTNP (bound with NADH) and CHMO WT (bound with NADH or NADPH, respectively) (Figure: 5.4). The results show that CHMO DTNP with NADH bound maintains greater rigidity over the control loop in comparison to the WT with NADH bound and the level of control loop rigidity in CHMO DTNP with NADH bound is nearly identical with that of the CHMO WT binding NADPH. These results support the role of K326T in stabilizing the control loop which clamps on the otherwise loosely bound cofactor NADH.

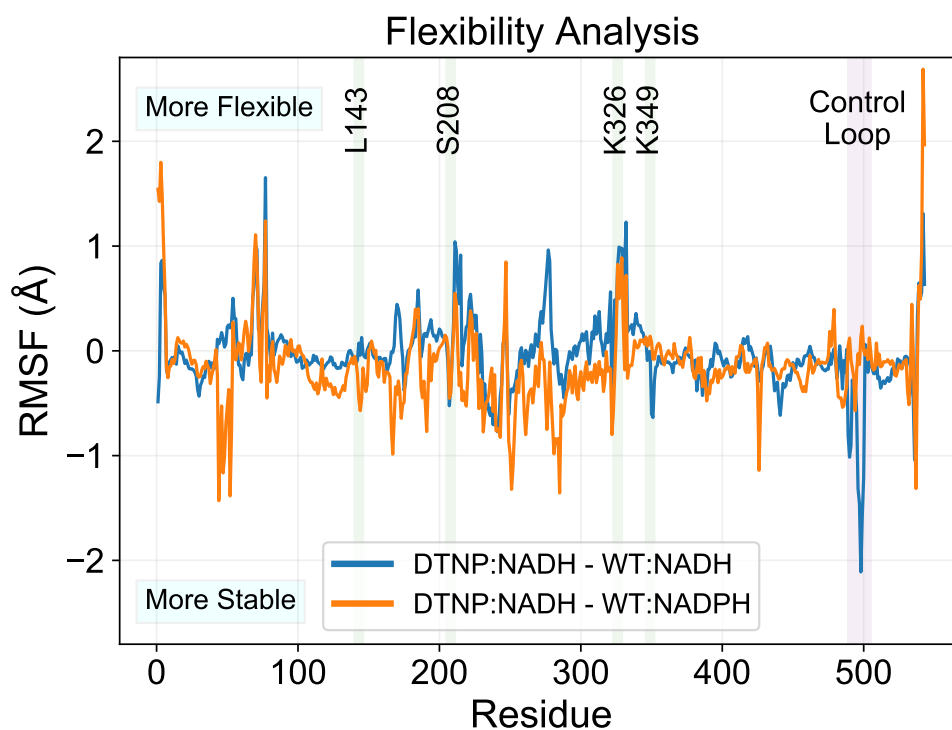


Figure 5.4: Root mean square fluctuation (RMSF) analysis of CHMO DTNP. Results are plotted as the difference in RMSF between DTNP (with NADH bound) and WT (with NADH or NADPH bound, respectively). The control loop is found to have greater stability with DTNP:NADH, comparable to the native condition of WT:NADPH. The inactive pairing of WT:NADH exhibits high flexibility at the control loop.

L143P contacts FAD and not NADH (Figure: 5.3D, 5.5) Therefore, we hypothesized that this mutation is involved in positioning FAD and in turn affects hydride transfer from the nicotinamide cofactors to FAD. We first compared the hydride transfer distances (the distance between nicotinamide C4 and FAD N5) when CHMO WT utilizes different cofactors, and confirmed that higher activity is linked to a shorter hydride transfer distance ($4.6 \pm 0.5 \text{ \AA}$ for NADPH versus 6.5 ± 1.0

Å for NADH[29] (Figure: 5.5A), in line with the logic that enzymes capable of sampling catalytic geometry more frequently are more active. Interestingly, the hydride transfer distance when using NADH as the cofactor was much shorter in CHMO DTNP compared to in CHMO DTN (Figure: 5.5B, C), which is consistent with our hypothesis that L143P facilitates more efficient hydride transfer from NADH.

To understand the role of L143P in influencing hydride transfer, we identified metastable FAD conformations in CHMO DTN (with NADH bound) and CHMO DTNP (with NADH bound) based on minimum heavy atom distance to residues within 4 Å by performing PCA and K-means clustering (Figure: 5.8). Conformations from the most populated cluster of each sample were compared and indicate that the native leucine packs underneath the flavin (Figure: 5.5D), limiting flavin's flexibility to move closer to the nicotinamide. This limited flexibility is likely advantageous under the native condition of NADPH binding to CHMO WT, resulting in precise organization of the flavin for hydride transfer. However, upon mutation at S208D and K326T which contact the adenosine end of the cofactor, the binding pose of the cofactor is altered, which requires the flavin to adjust its position accordingly. The L143P proline in CHMO DTNP appears to pack along the FAD ribitol instead of directly against flavin (Figure: 5.5E). This may support twisting motion of FAD to optimally orient the flavin toward the NADH nicotinamide ring.

5.5 Supplementary information

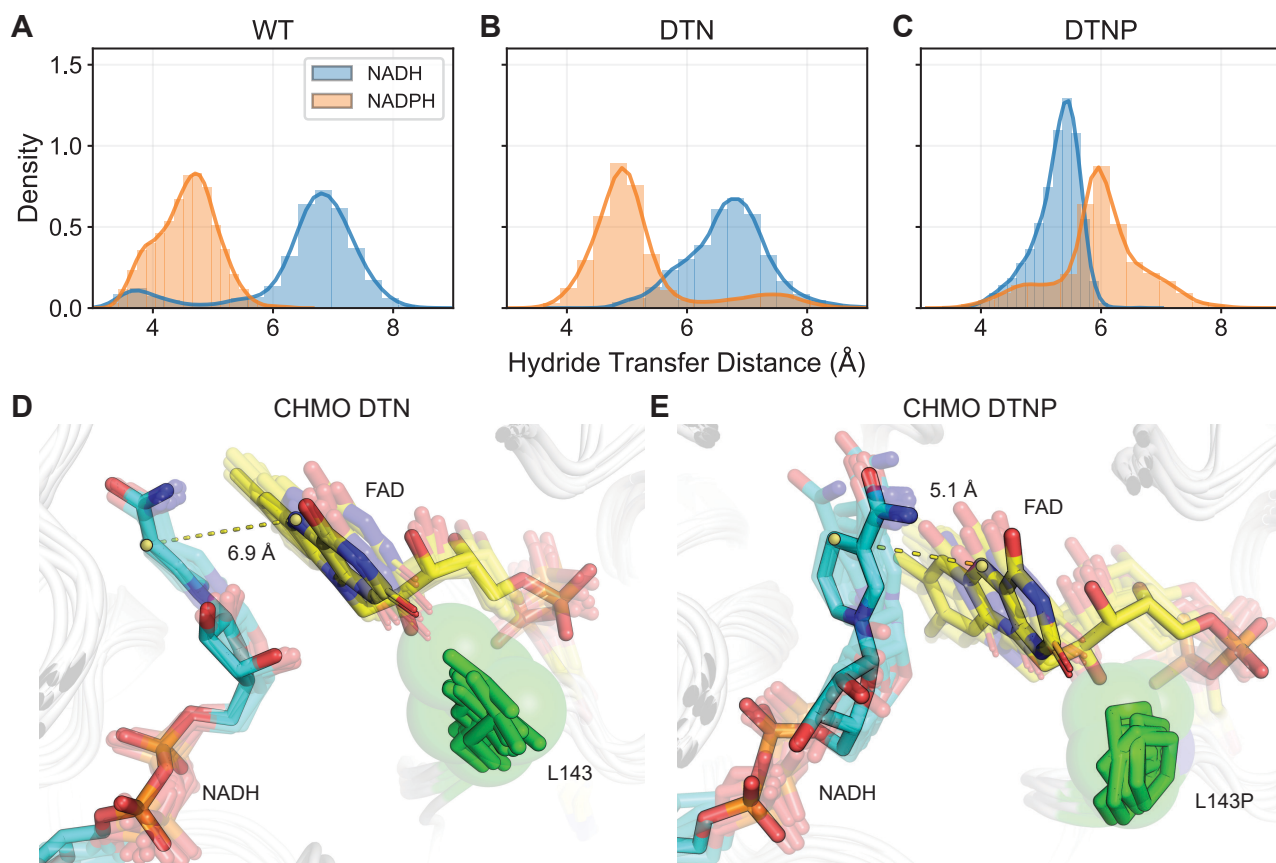


Figure 5.5: Evaluation of hydride transfer efficiency in CHMO variants. Shorter distances between NAD(P)H C4 and FAD N5 contribute to enhanced catalysis. **(A, B)** WT and DTN are both marked by NADPH sampling distances ~ 5 Å, while the less active NADH samples distances are >6 Å and are too remote to engage in hydride transfer. **(C)** DTNP displays the opposite arrangement, with NADH sampling closer distances than NADPH. **(D)** L143 in DTN firmly packs under the flavin ring, blocking the flavin from moving closer to nicotinamide, resulting in suboptimal hydride transfer distance. **(E)** L143P presses against the FAD ribitol rather than contacting the flavin. This anchors the FAD core and allows the flavin head to rotate in response to changes in the nicotinamide positioning to sustain closer contact. The dashed lines show representative distances between the nicotinamide C4 and FAD N5.

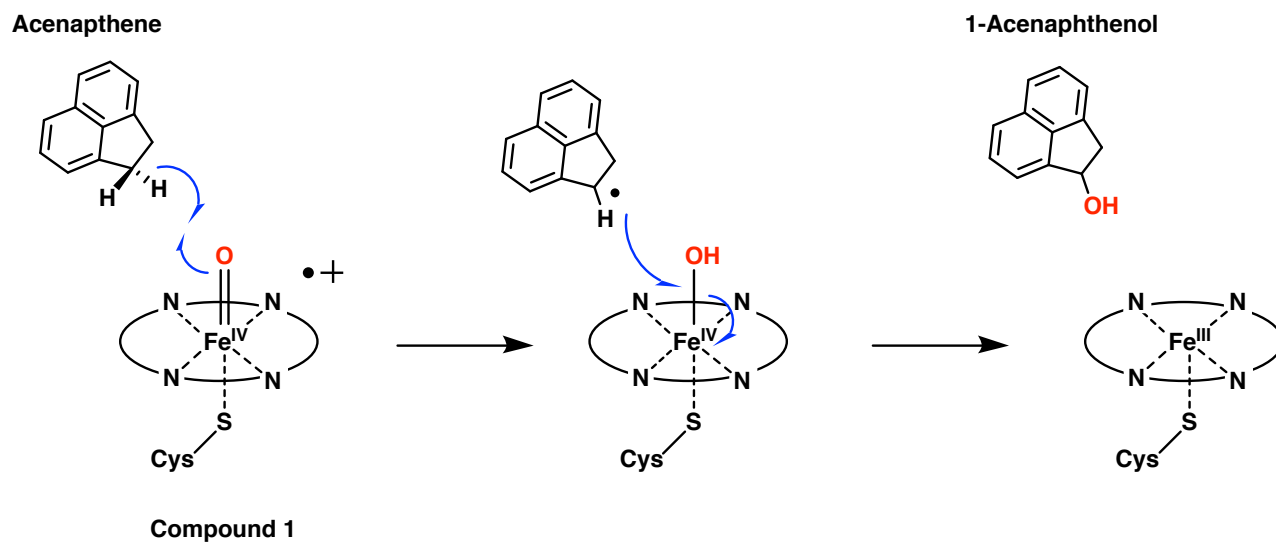


Figure 5.6: Proposed P450-BM3 ACN reaction mechanism. Compound I Fe(IV)=O abstracts a hydrogen from the substrate to form Compound II Fe(IV)-OH. This is followed by the "rebound" of the substrate and hydroxy radical to form the final product.)

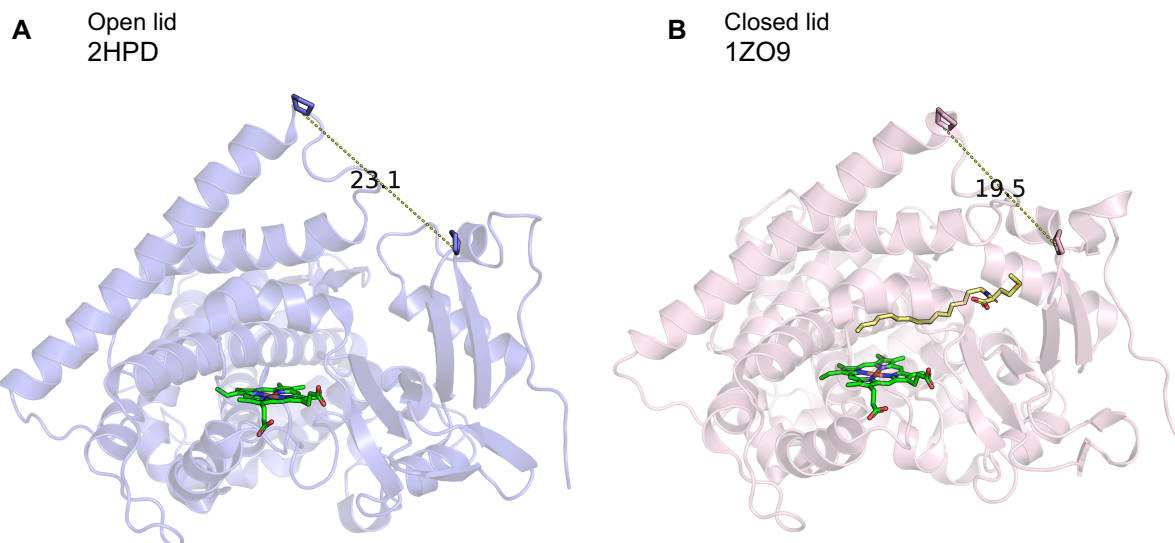


Figure 5.7: P450-BM3 structural flexibility. The lid opening motion of P450-BM3 is described by the alpha carbon distance from P196 located on the F/G loop to P45 across the substrate channel. The heme group is highlighted in green. **(A)** PDB 2HPD captures P450-BM3 in the inactive, open state with raised F and G helices. The lid opening distance is measured to be 23.1 Å. **(B)** PDB 1ZO9 illustrates P450-BM3 in the substrate bound, closed form that is catalytically active. The bound substrate N-palmitoylmethionine is colored yellow, and the lid opening distance is measured to be 19.5 Å, indicating downward movement of the F and G helices by roughly 3.6 Å.

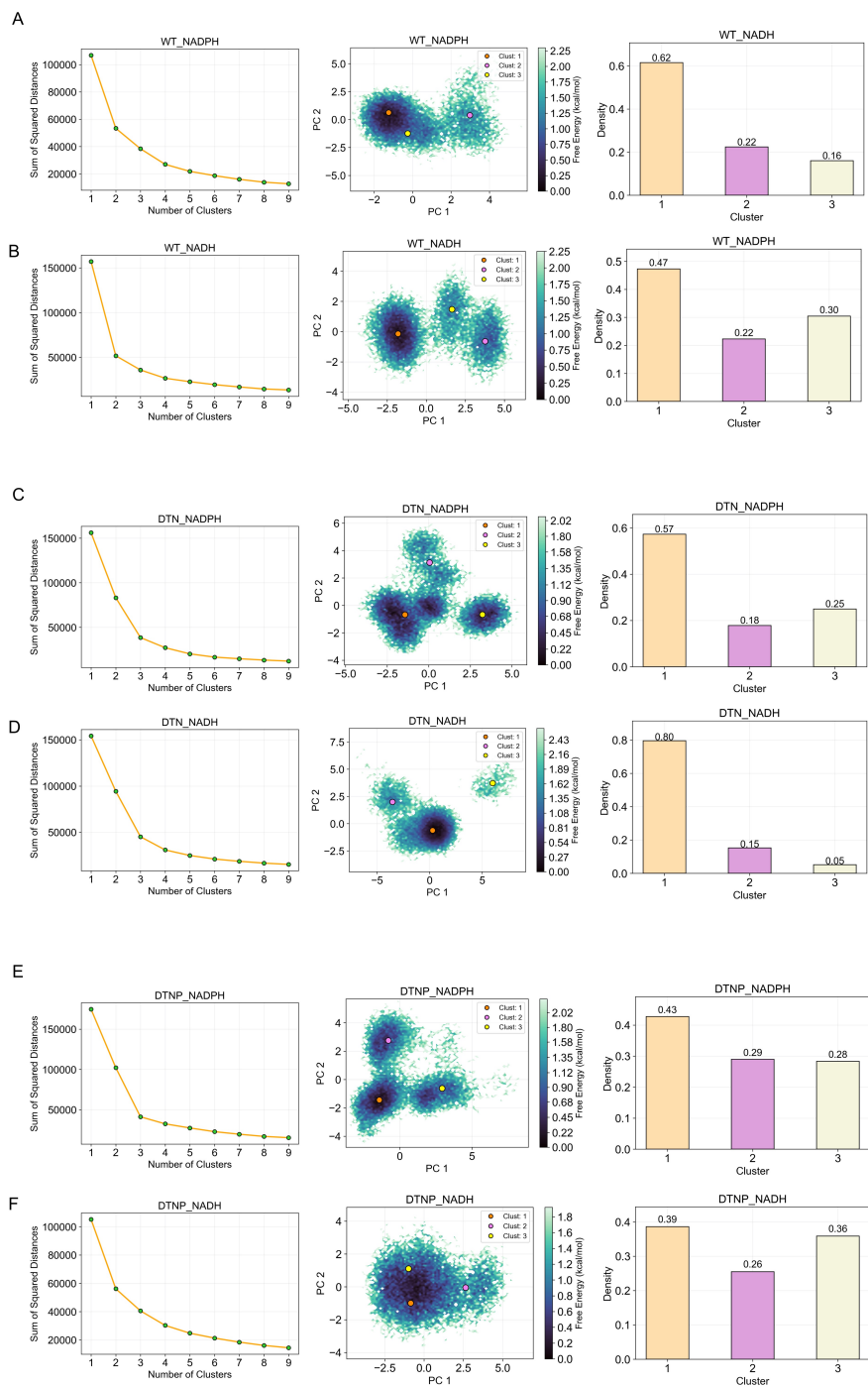


Figure 5.8: Free energy landscapes for DTN, DTNP, and WT CHMO with NADH or NADPH bound. Each row illustrates the K-means clustering elbow heuristic to determine the optimal number of clusters, free energy landscapes projected on first 2 principal components with cluster centers, and cluster populations. **(A)** WT with NADPH bound **(B)** WT with NADH bound **(C)** DTN with NADPH bound **(D)** DTN with NADH bound **(E)** DTNP with NADPH bound **(F)** DTNP with NADH bound.

Bibliography

- [1] Yuheng Lin and Yajun Yan. Biotechnological production of plant-specific hydroxylated phenylpropanoids. *Biotechnol. Bioeng.*, 111(9):1895–1899, September 2014.
- [2] Toshiki Furuya and Kuniki Kino. Catalytic activity of the two-component flavin-dependent monooxygenase from *Pseudomonas aeruginosa* toward cinnamic acid derivatives. *Appl. Microbiol. Biotechnol.*, 98(3):1145–1154, February 2014.
- [3] Vlada B Urlacher and Sabine Eiben. Cytochrome P450 monooxygenases: perspectives for synthetic application. *Trends Biotechnol.*, 24(7):324–330, July 2006.
- [4] Sandy Schmidt, Christian Scherkus, Jan Muschiol, Ulf Menyes, Till Winkler, Werner Hummel, Harald Gröger, Andreas Liese, Hans-Georg Herz, and Uwe T Bornscheuer. An enzyme cascade synthesis of ϵ -caprolactone and its oligomers. *Angew. Chem. Int. Ed Engl.*, 54(9):2784–2787, February 2015.
- [5] Jackson K B Cahn, Caroline A Werlang, Armin Baumschlager, Sabine Brinkmann-Chen, Stephen L Mayo, and Frances H Arnold. A general tool for engineering the NAD/NADP cofactor preference of oxidoreductases. *ACS Synth. Biol.*, 6(2):326–333, February 2017.
- [6] Tristan de Rond, Jian Gao, Amin Zargar, Markus de Raad, Jack Cunha, Trent R Northen, and Jay D Keasling. A High-Throughput mass spectrometric enzyme activity assay enabling the discovery of cytochrome P450 biocatalysts. *Angew. Chem. Int. Ed.*, 58(30):10114–10119, July 2019.
- [7] Fabrice Gielen, Raphaëlle Hours, Stéphane Emond, Martin Fischlechner, Ursula Schell, and

- Florian Hollfelder. Ultrahigh-throughput-directed enzyme evolution by absorbance-activated droplet sorting (AADS). *Proc. Natl. Acad. Sci. U. S. A.*, 113(47):E7383–E7389, November 2016.
- [8] Aaron Debon, Moritz Pott, Richard Obexer, Anthony P Green, Lukas Friedrich, Andrew D Griffiths, and Donald Hilvert. Ultrahigh-throughput screening enables efficient single-round oxidase remodelling. *Nature Catalysis*, 2(9):740–747, September 2019.
- [9] Solvej Siedler, Georg Schendzielorz, Stephan Binder, Lothar Eggeling, Stephanie Bringer, and Michael Bott. SoxR as a single-cell biosensor for NADPH-consuming enzymes in *escherichia coli*. *ACS Synth. Biol.*, 3(1):41–47, January 2014.
- [10] Hiroshi Suzuki, Kanako Inabe, Yoshinori Shirakawa, Naoki Umezawa, Nobuki Kato, and Tsunehiko Higuchi. Role of thiolate ligand in spin state and redox switching in the cytochrome P450 catalytic cycle. *Inorg. Chem.*, 56(8):4245–4248, April 2017.
- [11] Lisa K Morlock, Dominique Böttcher, and Uwe T Bornscheuer. Simultaneous detection of NADPH consumption and H₂O₂ production using the ampliflu™ red assay for screening of P450 activities and uncoupling. *Appl. Microbiol. Biotechnol.*, 102(2):985–994, January 2018.
- [12] Christopher J C Whitehouse, Stephen G Bell, and Luet-Lok Wong. P450(BM3) (CYP102A1): connecting the dots. *Chem. Soc. Rev.*, 41(3):1218–1260, February 2012.
- [13] Joelle N Pelletier, Olivier Rousseau, Maximilian Ccjc Ebert, Daniela Quaglia, Ali Fendri, Adem H Parisien, Jonathan N Besna, and Saathanan Iyathurai. Indigo formation and rapid NADPH consumption provide robust prediction of raspberry ketone synthesis by engineered cytochrome P450 BM3. *ChemCatChem*, 12(3).
- [14] A B Carmichael and L L Wong. Protein engineering of *bacillus megaterium* CYP102. the oxidation of polycyclic aromatic hydrocarbons. *Eur. J. Biochem.*, 268(10):3117–3125, May 2001.
- [15] Q S Li, J Ogawa, R D Schmid, and S Shimizu. Engineering cytochrome P450 BM-3 for

- oxidation of polycyclic aromatic hydrocarbons. *Appl. Environ. Microbiol.*, 67(12):5735–5739, December 2001.
- [16] Yan Zhang, Yin-Qi Wu, Na Xu, Qian Zhao, Hui-Lei Yu, and Jian-He Xu. Engineering of cyclohexanone monooxygenase for the enantioselective synthesis of (S)-Omeprazole. *ACS Sustainable Chem. Eng.*, 7(7):7218–7226, April 2019.
- [17] Na Xu, Jun Zhu, Yin-Qi Wu, Yan Zhang, Jian-Ye Xia, Qian Zhao, Guo-Qiang Lin, Hui-Lei Yu, and Jian-He Xu. Enzymatic preparation of the chiral (S)-Sulfoxide drug esomeprazole at Pilot-Scale levels. *Org. Process Res. Dev.*, 24(6):1124–1130, June 2020.
- [18] Gonzalo de Gonzalo and Andrés R Alcántara. Multienzymatic processes involving Baeyer–Villiger monooxygenases. *Catalysts*, 11(5):605, May 2021.
- [19] Andy Beier, Sven Bordewick, Maika Genz, Sandy Schmidt, Tom van den Bergh, Christin Peters, Henk-Jan Joosten, and Uwe T Bornscheuer. Switch in cofactor specificity of a Baeyer–Villiger monooxygenase. *Chembiochem*, 17(24):2312–2315, December 2016.
- [20] Andrea M Chánique and Loreto P Parra. Protein engineering for nicotinamide coenzyme specificity in oxidoreductases: Attempts and challenges. *Front. Microbiol.*, 9:194, February 2018.
- [21] Silja Mordhorst and Jennifer N Andexer. Round, round we go - strategies for enzymatic cofactor regeneration. *Nat. Prod. Rep.*, 37(10):1316–1333, October 2020.
- [22] André Pick, Wolfgang Ott, Thomas Howe, Jochen Schmid, and Volker Sieber. Improving the NADH-cofactor specificity of the highly active AdhZ3 and AdhZ2 from escherichia coli K-12. *J. Biotechnol.*, 189:157–165, November 2014.
- [23] Sarel J Fleishman, Andrew Leaver-Fay, Jacob E Corn, Eva-Maria Strauch, Sagar D Khare, Nobuyasu Koga, Justin Ashworth, Paul Murphy, Florian Richter, Gordon Lemmon, Jens Meiler, and David Baker. RosettaScripts: A scripting language interface to the rosetta macromolecular modeling suite. *PLoS One*, 6(6):e20161, June 2011.

- [24] Amita Hegde, Donovan C Haines, Muralidhar Bondlela, Baozhi Chen, Nathaniel Schaffer, Diana R Tomchick, Mischa Machius, Hien Nguyen, Puneet K Chowdhary, Larissa Stewart, Claudia Lopez, and Julian A Peterson. Interactions of substrates at the surface of p450s can greatly enhance substrate potency. *Biochemistry*, 46(49):14010–14017, December 2007.
- [25] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, 47(D1):D1102–D1109, January 2019.
- [26] Yifan Song, Frank DiMaio, Ray Yu-Ruei Wang, David Kim, Chris Miles, Tj Brunette, James Thompson, and David Baker. High-resolution comparative modeling with RosettaCM. *Structure*, 21(10):1735–1742, October 2013.
- [27] Tom Madden. The NCBI handbook [internet]. *National Center for Biotechnology Information (US)*, 2013.
- [28] 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(1):235–242, January 2000.
- [29] I Ahmad Mirza, Brahm J Yachnin, Shaozhao Wang, Stephan Grosse, H el ene Bergeron, Akihiro Imura, Hiroaki Iwaki, Yoshie Hasegawa, Peter C K Lau, and Albert M Berghuis. Crystal structures of cyclohexanone monooxygenase reveal complex domain movements and a sliding cofactor. *J. Am. Chem. Soc.*, 131(25):8848–8854, July 2009.
- [30] Brahm J Yachnin, Tara Sprules, Michelle B McEvoy, Peter C K Lau, and Albert M Berghuis. The Substrate-Bound crystal structure of a Baeyer–Villiger monooxygenase exhibits a criegee-like conformation. *J. Am. Chem. Soc.*, 134(18):7788–7795, May 2012.
- [31] Brahm J Yachnin, Michelle B McEvoy, Roderick J D MacCuish, Krista L Morley, Peter C K Lau, and Albert M Berghuis. Lactone-bound structures of cyclohexanone monooxygenase provide insight into the stereochemistry of catalysis. *ACS Chem. Biol.*, 9(12):2843–2851, December 2014.

- [32] Kazutaka Katoh, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.*, 30(14):3059–3066, July 2002.
- [33] Romelia Salomon-Ferrer, Andreas W Götz, Duncan Poole, Scott Le Grand, and Ross C Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. *J. Chem. Theory Comput.*, 9(9):3878–3888, September 2013.
- [34] David A Case, Thomas E Cheatham, 3rd, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, December 2005.
- [35] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, August 2015.
- [36] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103(19):8577–8593, November 1995.
- [37] Ramu Anandakrishnan, Boris Aguilar, and Alexey V Onufriev. H++ 3.0: automating pk prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.*, 40(Web Server issue):W537–41, July 2012.
- [38] Kiumars Shahrokh, Anita Orendt, Garold S Yost, and Thomas E Cheatham, 3rd. Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. *J. Comput. Chem.*, 33(2):119–133, January 2012.
- [39] K G Ravichandran, S S Boddupalli, C A Hasermann, J A Peterson, and J Deisenhofer. Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal p450’s. *Science*, 261(5122):731–736, August 1993.
- [40] Brahm J Yachnin, Peter C K Lau, and Albert M Berghuis. The role of conformational flexibility

in Baeyer-Villiger monooxygenase catalysis and structure. *Biochim. Biophys. Acta*, 1864(12): 1641–1648, December 2016.

Chapter 6

Conclusions and future directions

Applying molecular simulation to accurately capture protein-ligand binding interactions is critical to accelerating efforts in rational protein design and drug discovery.

Chapter 1 reviews the mathematical foundations and steps for commonly utilized free energy methods including: MM-PBSA, LIE, and absolute alchemical simulations. We next demonstrated how absolute alchemical free energy simulations based on removal of electrostatic and van der Waals interactions between the protein and ligand are able to achieve predictive accuracies below 1 kcal/mol with polarization corrections on the UPA system (Chapter 2). Previous approaches treating atoms as fixed point charges in explicit solvent are unable to model the polarization changes that occur as the ligand moves between the solvent and non-polar protein interior, leading to overly favorable (negative) binding free energies. By benchmarking binding free energy predictions with 10 UPA inhibitors, we found that standard alchemical approaches reach 3.2 kcal/mol RMSE and -0.15 Pearson correlation, indicating poor predictive value. With scaling of the dielectric constant in PBSA continuum solvent to screen charged effects, the performance was improved to reach 0.89 kcal/mol RMSE and 0.67 Pearson correlation. We further explored the effects of simulation parameters including counter-ion concentration, restraint choice, forcefield, and protonation states of titratable residues in the binding pocket. Each of the parameters had consequential effects on con-

formational sampling and predictive outcome, and we conclude that the best practice for absolute alchemical simulation involves simulating with experimental salt conditions, with 1DOF restraints to allow some degree of ligand mobility, and that examination of protonation is limited with existing protocols.

The MBAR/PBSA method demonstrated is a mean-field approach that is easily implemented through post-processing of existing MD trajectories. More sophisticated and computationally expensive methods that explicitly calculate the effects of electronic polarization such as through polarizable forcefields[1, 2] or Gaussian multipoles[3] will lead to further improvements as they continue to mature. The reliability of any molecular simulation is dependent on convergence in sampling of the configurations available. Although our simulations total over 500 μ s, these timescales are insufficient to draw definitive conclusions. Hardware improvements will inevitably allow longer simulations, and algorithmic developments to overcome these sampling limitations include the use of Markov state models[4, 5] by composing configurations from multiple shorter trajectories together or with enhanced sampling approaches[6], where small boost potentials are dynamically added to accelerate motion in configuration regions that have been previously sampled, thereby reducing the depth of free energy basins to push the system forward along the reaction coordinate.

In Chapter 3 we review methods to swap cofactor specificity between the natural redox cofactors NAD/H and NADP/H, and examine how these approaches can be extended to design proteins utilizing artificial redox cofactors. We applied semi-rational design with *Ec* gapA to engineer the protein to utilize the artificial redox cofactor NMN⁺ (Chapter 4). Our strategy is based on: 1) boosting binding affinity by introducing mutations that can form novel polar interactions with the NMN⁺ phosphate group. and 2) enforcing NMN⁺ specificity by sterically occluding the native cofactor from binding through insertion of bulky residues into the adenosine pocket, With Rosetta simulation where residues lining the binding pocket were systematically mutated to sample all possible residues, we identified the variant A180S that had the potential to form an inter-subunit hydrogen bond with the phosphate group. Experimental testing showed A180S had \sim 6-fold increase in NMN⁺ catalytic efficiency in comparison to the WT. An orthogonal gapA variant with \sim 200-fold cofactor specificity switch was discovered by introducing G10R on top of A180S. G10R is

modeled to extend into the adenosine cleft and form salt bridges with D33, preventing the larger NAD^+ from fitting and making the conserved polar contact. To broadly explore gapA sequence space and find cooperative mutations that would not be evident through molecular modeling, we developed a high-throughput colorimetric assay measuring NMNH production from crude lysate. High-throughput screening led to the best performing gapA HT-9 (A180S-G187K-P188A) that had ~ 32 -fold increase in NMN^+ catalytic efficiency over the WT.

Our design strategy was successful, but the gapA mutants still utilize NMN^+ at levels significantly lower than the WT with NAD^+ , and far below the levels necessary for industrial scale processes. Multiple sequence alignment, comparison of crystal structures with the cofactors NAD(P)^+ in bound pose, and existing mutagenesis data describing cofactor specificity switching shows that even minor changes to the Rossman fold residues can have a large impact on binding specificity and kinetics, and that these effects are generally transferrable[7, 8]. Based on these observations, we hypothesize that there exists a general set of mutations that will allow universal conversion of enzymes with the Rossman fold architecture to have altered cofactor preference for NMN^+ . We have examples of engineered NMN^+ binding proteins, but with so few active samples we cannot draw any broader conclusions about what sequence profile may be optimal for NMN^+ activity. To better understand the contributions of mutations to NMN^+ binding and identify a general sequence motif for conversion of enzymes to prefer NMN^+ , we should obtain crystal structures of our NMN^+ binders with cofactor bound and apply deep mutational scanning[9, 10] methods to systematically test the effect of all possible single residue mutations in the Rossman fold on NMN^+ binding. Future work can also incorporate continuous growth selections[11, 12], where cell survival is linked to the enzymes ability to regenerate NMN/H , or test out other artificial redox cofactors[13, 14].

Lastly in Chapter 5, we describe how molecular simulations can be used to rationalize the mechanistic effects of mutations on conformational dynamics. The previous chapters focused on applying MD or Rosetta as predictive tools to select variants with binding potential, here we begin with experimentally characterized mutants and investigate how the discovered mutations lead to the target activities. Oxygenases have complex reaction mechanisms dependent on concerted, global changes in structure that are not apparent through static models. Docking of ACN into P450-BM3

GVQ-AL illustrated that the mutations cooperatively reshaped the binding pocket for improved shape complementarity to the planar ACN ligand, and MD simulation of GVQ-D222N revealed that the mutations shifted the protein to favor the lid closed, catalytically competent state with ACN over the open state where the reaction could not proceed. Homology modeling and simulation of *Ac* CHMO DTNP, which had $\sim 1,200$ -fold cofactor specificity switch from NADPH to NADH, identified changes in control loop flexibility and cofactor positioning affecting hydride transfer potential. Flexibility of the control loop measured through RMSF decreased with NADH in CHMO DTNP compared to the WT, suggesting that the mutations resulted in a stronger clamp on the cofactor. Hydride transfer efficiency between NAD(P)H and FAD was evaluated by recording the distances sampled between the NAD(P)H C4 and FAD N5. CHMO DTNP with NADH bound was found to average shorter distances $< 6 \text{ \AA}$ resembling the behaviour observed with WT CHMO and the native cofactor NADPH. Modeling suggests that the L143P mutation presses against the FAD ribitol to support rotation of the flavin head for better positioning with the NADH for hydride transfer. Further insight can be obtained through graph analysis comparing how residue dynamic cross correlation is affected[15], or with alchemical simulation transforming the WT to mutated amino acids[16].

Bibliography

- [1] Jay W Ponder, Chuanjie Wu, Pengyu Ren, Vijay S Pande, John D Chodera, Michael J Schnieders, Imran Haque, David L Mobley, Daniel S Lambrecht, Robert A DiStasio, Martin Head-Gordon, Gary N I Clark, Margaret E Johnson, and Teresa Head-Gordon. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B*, 114(8):2549–2564, March 2010.
- [2] Yue Shi, Zhen Xia, Jiajing Zhang, Robert Best, Chuanjie Wu, Jay W Ponder, and Pengyu Ren. Polarizable atomic Multipole-Based AMOEBA force field for proteins. *J. Chem. Theory Comput.*, 9(9):4046–4063, September 2013.
- [3] Haixin Wei, Ruxi Qi, Junmei Wang, Piotr Cieplak, Yong Duan, and Ray Luo. Efficient formulation of polarizable gaussian multipole electrostatics for biomolecular simulations. *J. Chem. Phys.*, 153(11):114116, September 2020.
- [4] Brooke E Husic and Vijay S Pande. Markov state models: From an art to a science. *J. Am. Chem. Soc.*, 140(7):2386–2396, February 2018.
- [5] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.*, 25:135–144, April 2014.
- [6] Yinglong Miao and J Andrew McCammon. Gaussian accelerated molecular dynamics: Theory, implementation, and applications. *Annu. Rep. Comput. Chem.*, 13:231–278, August 2017.
- [7] Andrea M Chánique and Loreto P Parra. Protein engineering for nicotinamide coenzyme

- specificity in oxidoreductases: Attempts and challenges. *Front. Microbiol.*, 9:194, February 2018.
- [8] Jackson K B Cahn, Caroline A Werlang, Armin Baumschlager, Sabine Brinkmann-Chen, Stephen L Mayo, and Frances H Arnold. A general tool for engineering the NAD/NADP cofactor preference of oxidoreductases. *ACS Synth. Biol.*, 6(2):326–333, February 2017.
- [9] Ali Nikoomanzar, Derek Vallejo, and John C Chaput. Elucidating the determinants of polymerase specificity by Microfluidic-Based deep mutational scanning. *ACS Synth. Biol.*, 8(6):1421–1429, June 2019.
- [10] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nat. Methods*, 11(8):801–807, August 2014.
- [11] Arjun Ravikumar, Garri A Arzumanyan, Muaeen K A Obadi, Alex A Javanpour, and Chang C Liu. Scalable, continuous evolution of genes at mutation rates above genomic error thresholds. *Cell*, 175(7):1946–1957.e13, December 2018.
- [12] Gordon Rix and Chang C Liu. Systems for in vivo hypermutation: a quest for scale and depth in directed evolution. *Curr. Opin. Chem. Biol.*, 64:20–26, October 2021.
- [13] Yuxue Liu, Xiaojia Guo, Wujun Liu, Junting Wang, and Zongbao Kent Zhao. Structural insights into malic enzyme variants favoring an unnatural redox cofactor. *Chembiochem*, 22(10):1765–1768, May 2021.
- [14] Tanja Knaus, Caroline E Paul, Colin W Levy, Simon de Vries, Francesco G Mutti, Frank Hollmann, and Nigel S Scrutton. Better than nature: Nicotinamide biomimetics that outperform natural coenzymes. *J. Am. Chem. Soc.*, 138(3):1033–1039, January 2016.
- [15] Adrian Romero-Rivera, Marc Garcia-Borràs, and Sílvia Osuna. Role of conformational dynamics in the evolution of Retro-Aldolase activity. *ACS Catal.*, pages 8524–8532, November 2017.
- [16] Matteo Aldeghi, Vytautas Gapsys, and Bert L de Groot. Accurate estimation of ligand binding affinity changes upon protein mutation. *ACS Cent. Sci.*, December 2018.