2021

The Design and Application of Enzyme Inter-residue Interaction Networks Towards Quantum Mechanical Modeling

Thomas Summers

THE DESIGN AND APPLICATION OF ENZYME INTER-RESIDUE INTERACTION NETWORKS TOWARDS QUANTUM MECHANICAL MODELING

by

Thomas Jeffrey Summers

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Chemistry

The University of Memphis

August 2021

Acknowledgements

First and foremost, I want to thank my parents, Jeff and Mary Grace, my sisters, Emily and Grace, and the rest of my family for the never-ending love, support, and laughter I have received from them. I also want to thank my dearest friends for continually reminding me that there is more to life than research.

I am grateful for the time and effort Dr. Nathan DeYonker has devoted to me these past years, along with his persistence in directing me towards new opportunities I otherwise would have not undertaken. I also want to thank the current and previous members of the DeYonker research group. In particular, I thank Dr. Qianyi Cheng, as I would have certainly not completed this journey without her compassion, patience, and guidance.

Lastly, I acknowledge the resources and funding provided by the University of Memphis High Performance Computing facilities, the University of Memphis Department of Chemistry, and the National Science Foundation, without which the research detailed in this work would not have been possible.

Preface

This dissertation discusses the development and application of residue interaction networks towards the design of enzyme cluster models. By using these cheminformatic graphs as rationale for model creation, we seek to provide a protocol that is rational, reproducible, and practical for the simulation of various biosystems beyond those detailed within this work.

Several chapters are adapted from published works. Chapter 2 is adapted from the journal article “A transition state ‘trapped’? QM-cluster models of engineered threonyl-tRNA synthetase” published in *Organic & Biomolecular Chemistry* (T. J. Summers, Q. Cheng, and N. J. DeYonker. *Org. Biomol. Chem.* 2018, 16, 4090-4100). Chapter 4 and part of Chapter 1 is adapted from the journal article “Cheminformatic quantum mechanical enzyme model design: a catechol-O-methyltransferase case study” under review by *Biophysical Journal* (T. J. Summers, Q. Cheng, M. A. Palma, D.-T. Pham, D. K. Kelso III, C. E. Webster, and N. J. DeYonker). Chapter 5 is adapted from the journal article “Quantifying inter-residue contacts through interaction energies” published in *Journal of Chemical Information and Modeling* (T. J. Summers, B. P. Daniel, Q. Cheng, and N. J. DeYonker. *J. Chem. Inf. Model.* 2019, 59, 5034-5044). For consistency, the tables and figures of these journal articles have been renumbered, and references are formatted following *Journal of Chemical Information and Modeling* guidelines.

Abstract

In order to accurately simulate the inner workings of an enzyme active site with quantum mechanics (QM), not only must the reactive species be included in the model, but also any important surrounding residues, solvent, ions, and coenzymes involved in crafting the microenvironment. The Residue Interaction Network ResidUe Selector (*RINRUS*) toolkit was designed to utilize interatomic contact network information for automated, rational residue selection and QM-cluster model generation. An X-ray crystal structure of a protein is translated into a two-dimensional network which may be then used to discern residues with significant interactions with the enzyme substrates. The rest of the protein is trimmed away following a defined protocol to create QM-cluster models suitable for simulation.

Three QM-cluster enzyme case studies demonstrating the capability of network-based models are presented in this work. First, models of six bioengineered threonyl-tRNA synthetase enzymes are simulated to reveal the impact residue mutations have towards creation of a transition state analogue structure within a protein pocket. Second, models of the zinc-native enzyme human carbonic anhydrase II with various transition state ions in the active site are shown to provide insight into the reduced catalytic activity of the metallovariants, along with predicting the potential viability of the iron-substituted variant. Third, over 500 *RINRUS*-designed models of the enzyme catechol-O-methyltransferase are analyzed to identify cheminformatic features that might be foundational for efficient, accurate model designs.

There is the possibility to incorporate machine learning into the *RINRUS* workflow to enable the transformation of simple qualitative/semi-quantitative chemical characteristics into descriptors suitable for more quantitative network designs. This is illustrated in the final piece of this work where random forest models constructed from the chemical information of four proteins were able to accurately predict quantitative inter-residue interaction energies for an untested protein only using several structural, network, and chemical descriptors. Collectively, the studies illustrate the value of the *RINRUS* toolkit in creating practical, accurate models of enzyme active sites, and they provide direction for future improvement with the methodology.

Table of Contents

Chapter **Page**

List of Tables

List of Figures

Figure **Page**

List of Abbreviations

BBI	Backbone-Backbone Interaction
BiPhe	*p*-Biphenylalanine
CA	Carbonic Anhydrase
CAT	Catecholate
COMT	Catechol-O-methyltransferase
CPCM	Conductor-like Polarizable Continuum Model
DFT	Density Functional Theory
ES	Enzyme-Substrate complex
E + S	Separate Enzyme and Substrate
F-SAPT	Functional Group Symmetry Adapted Perturbation Theory
GD3BJ	Grimme D3 Becke-Johnson Dispersion Correction
HCAII	Human Carbonic Anhydrase II
INT	Intermediate
I-SAPT	Intramolecular Symmetry Adapted Perturbation Theory
k_B	Boltzmann constant
k_{pos}	Harmonic Positional Restraint
MM	Molecular Mechanics
MD	Molecular Dynamics
ONIOM Mechanics	Our Own n-Layered Integrated Molecular Orbital and Molecular
PDB	Protein Data Bank
QM	Quantum Mechanics
RIC	Residue Interaction Count
RIN	Residue Interaction Network
RINRUS	Residue Interaction Network ResidUe Selector
RMSD	Root Mean Square Deviation
RMSE	Root Mean Square Error
SAH	*S*-adenosylhomocysteine

SAM	*S*-adenosylmethionine
SAPT	Symmetry Adapted Perturbation Theory
SIFt	Structural Interaction Fingerprint
SSI	Side Chain-Side Chain Interaction
ThrRS	Threonyl-tRNA Synthetase
TS	Transition State
ZPE	Zero-point Energy

Chapter 1: Introduction

For nearly two centuries, the structure, function, and catalytic power of enzymes have fascinated scientists, with countless studies seeking to understand their underlying mechanisms and biological function. With the advancement of computers, atomic-scale computer modeling of enzymes has become a necessary part of the global multibillion-dollar research effort that aids the design of new pharmaceuticals, helps to investigate and engineer novel protein structures and functions, and advances our understanding of the molecular basis of disease.[1,2] The importance of atomic-level simulation of enzyme-catalyzed reactions was publicly acknowledged with the 2013 Chemistry Nobel Prize being awarded to Warshel, Levitt, and Karplus, who developed methods to treat the active site of an enzyme with quantum mechanics (QM) and the periphery with classical or molecular mechanics (MM).[3]

QM-only (also called QM-cluster), QM/MM, and ONIOM (Our own n-layered Integrated molecular Orbital and Molecular Mechanics) are alternative approaches that have leveraged advancements in quantum mechanical theory and molecular dynamics (MD) to continually increase the ubiquity of computational enzymology.[4–8] As with all forms of modeling, the comparative accuracy of a model to reality is limited by the design of the model and relevant/reliable experimental data. For simulating the active site of enzymes, it is crucial to ensure not only the amino acids directly involved with the reaction are modeled at the QM-level but also any residues, water molecules, ions, and coenzymes sterically and/or electrostatically crafting the active site microenvironment.[4–6,9] While this is a simple idea in principle, it is far harder in practice to identify rationally which residues must be partitioned into the QM-level.

While *ad hoc* protocols exist for selecting residues for inclusion in QM-level modeling, recommendations are typically ambiguous, inefficient, or challenging to implement.[4,5] One of the most common practices is to simply include all residues within a certain radial distance from a point, perhaps the center of mass of substrate(s) or an active-site metal center. While suitable models could be constructed this way, calibration studies have confirmed large spheres (and consequently large models) are needed for convergence of simulated enzyme thermodynamics/kinetics.[9–18] These results are perhaps unsurprising as nature does not enforce any geometric requirement to the design of an enzyme active site. Published "big-QM" models further add distant charged residues within the protein to generate 500-1000 atom models; however, inclusion of less important residues unnecessarily increases the computational cost of any model.[11,19,20] Attempts to quantify the importance of residues have been performed via *a posteriori* computations such as QM/MM thermodynamic cycle perturbations,[21,22] linear response functions,[23] or Fukui/Charge Shift Analysis.[14,24] However, such methods essentially require computational effort and thorough analysis of the constructed enzyme models in order to decide on an optimal system. Iterating an undirected residue selection process to self-consistency via QM or QM/MM computations is even more expensive.

Ideally, there would be a computationally inexpensive, *a priori* approach to enzyme model construction that utilizes structural and chemical data to rationally select residues (or parts of residues) for QM-cluster modeling. As a potential solution for this model creation problem, our lab has been developing the software *Residue Interaction Network ResidUe Selector* (*RINRUS*) for automating QM-cluster model design and construction.

The general workflow of *RINRUS* (Figure 1) begins with the user uploading the protein structure of interest and specifying the substrate(s)/residue(s)/cofactor(s) directly involved in the reaction (termed the "seed"). The protein structure may undergo pre-processing if needed (e.g., adding missing hydrogens to an X-ray crystal structure or removing alternate residue conformations) before having its three-dimensional structure analyzed and translated into a two-dimensional inter-residue interaction network.[25,26] In its current form, the interaction network is generated from interatomic contact interactions as computed from the program *Probe*.[27] In short, *Probe* rolls a sphere along the van der Waals surface of all atoms and indicates with "contact dots" where the sphere comes into contact or overlaps with nearby atoms. This interatomic contact information is compiled for each of the residues and translated into a network graph where residues

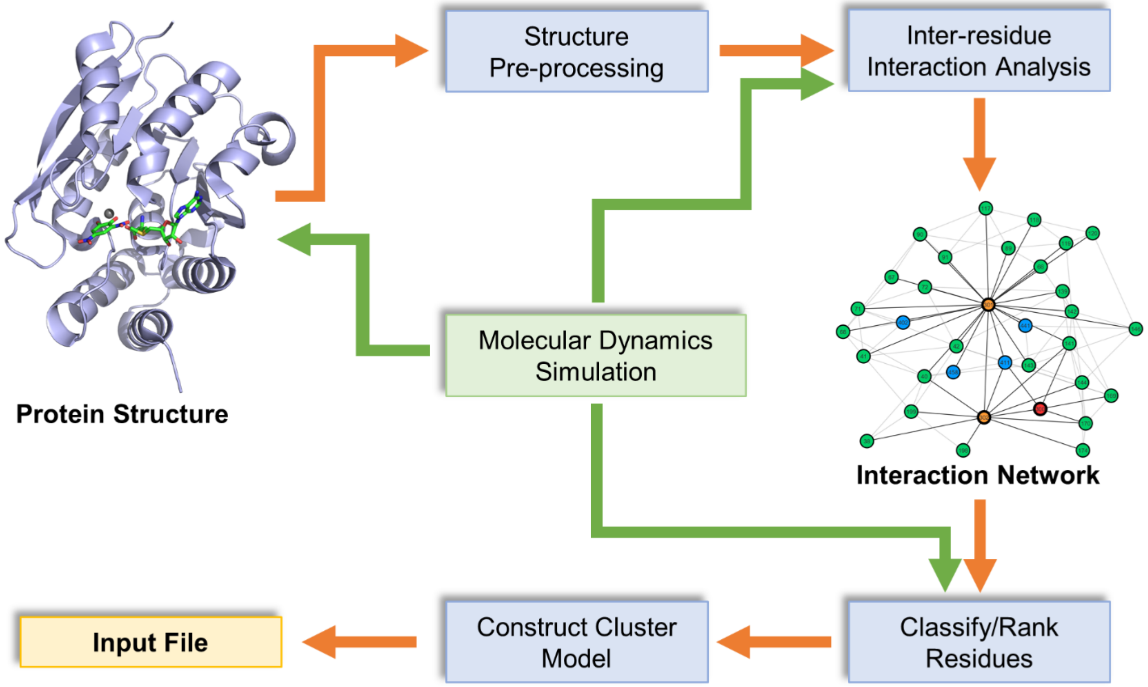


Figure 1. General workflow of the *RINRUS* toolkit (orange) where a user-specified protein structure is analyzed and processed into a QM-cluster model suitable for further QM computation. There are also opportunities for molecular dynamics cheminformatics to be incorporated into the workflow (green).

(represented as “nodes” in the graph) with interatomic interactions are interconnected with lines (termed “edges” in the graph). As the focus for constructing QM-models is on the species directly interacting with the seed, *RINRUS* isolates the subgraph composed of the user-specified seed and its neighboring residues (nodes). The interactions (edges) of the subgraph may be weighted by a property such as the number of interatomic contacts each residue has with the seed, favoring residues with many interatomic contacts with the seed over residues with few contacts. In constructing models, *RINRUS* currently has two modes of model design: one, models may be formed iteratively, adding residues to the seed based upon their edge weights (a ranking scheme); or two, models are formed by adding groups of residues with similar properties (a classification scheme). With the residues to be included in the cluster model identified, the rest of the protein is systematically trimmed away following a pre-defined protocol. To maintain the valence state of the atoms, hydrogens are added where covalent bonds are broken. The resulting cluster model may then be translated into an input file for QM-treatment by external software.

The chapters of this work detail the development of the *RINRUS* methodology alongside its application in the case studies of four different enzymes. Chapters 2 and 3 cover the earliest applications of this methodology at a time when the procedure was not yet automated and a detailed protocol for model trimming was not established. Even without the extra rigor of the current form of *RINRUS*, the manually crafted models successfully provided insight into the flexibility of a noncanonical residue side chain within an inner protein pocket (Chapter 2) and the impact of substituting a native active site metal ion with other metals on a catalyzed reaction mechanism (Chapter 3). By

Chapter 4, a model trimming protocol based upon network chemical information had been constructed, and the steps for *RINRUS* model building had been encoded into a Python toolkit. In automating the workflow, hundreds of QM-models are able to be created within mere minutes. With this newfound capability, we simulated a single enzyme reaction with over 500 unique models generated from several different building schemes in order to identify a scheme capable of building accurate, efficient QM-models (Chapter 4). Chapter 5 covers recent efforts to improve and expand upon the cheminformatics with which *RINRUS* operates by investigating the ability of simple random forest algorithms to translate inter-residue contacts into quantitative interaction energy values. Collectively, these works highlight the capability of the *RINRUS* framework and indicate directions for further advancement of this toolkit.

Chapter 2: Threonyl-tRNA Synthetase

Introduction

Most research on enzymes centers on studying their extraordinary capability of catalyzing biochemical reactions at high kinetic rates and specificity. Over the past several decades, interest in enzymes has greatly expanded towards using them as bioengineering tools for scientific inquiry. A recent example of this features the enzyme threonyl-tRNA synthetase (ThrRS), a member of the class-II aminoacyl-tRNA synthetase family primarily known for its function in protein biosynthesis. Through a two-step process, ThrRS activates the amino acid threonine by catalyzing the esterification of the amino acid to its cognate tRNA.[28,29] The newly charged tRNA may then be used by cellular ribosomes to construct genetically encoded proteins via translation. Although most research has focused on examining the substrate enantiomeric selectivity of ThrRS,[30–32] Schultz and co-workers investigated using ThrRS as a platform for protein engineering.[33] Starting from the highly thermostable ThrRS enzyme found in the thermophile *Pyrococcus abyssi*,[34] Schultz and coworkers used Rosetta software[35,36] to computationally redesign the interior of the ThrRS enzyme. Using results from the Rosetta package for suggested point mutations, they experimentally created a microenvironment that favors the stabilization of the planar conformation of the biphenyl sidechain on the noncanonical amino acid *p*-biphenylalanine. After an iterative procedure of amino acid mutations (PDB entries = 4S02, 4S0J, 4S0L, 4S0I, 4S0K), their group successfully obtained the X-ray crystal structure of a ThrRS containing the *p*-biphenylalanine (BiPhe) side chain in the coplanar conformation (PDB entry = 4S03, Figure 2).

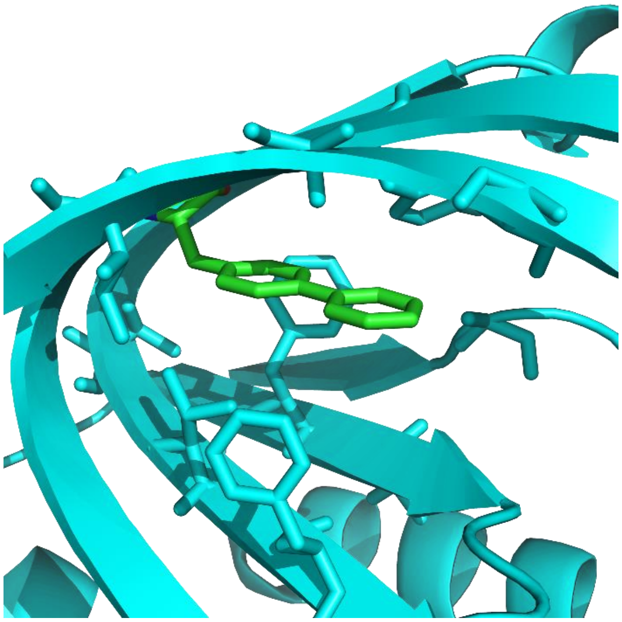


Figure 2. The X-ray crystal structure of 4S03, detailing the planar rings of the *p*-biphenylalanine residue.

The feat of Schultz and coworkers centers on the fact that the coplanar conformation (Φ = 0°) of free biphenyl is one of two rotational transition states (TSs) for the molecule, the other being at Φ = 90°. Electron diffraction studies have shown gas-phase biphenyl to have a central dihedral angle of 44.4 ± 1.2° at equilibrium.[37] This "staggered" global minimum (Figure 3) is commonly explained to be the result of energetic-steric competition whereby inter-ring π-conjugation favors the two rings to be coplanar but inter-ring hyperconjugation and steric repulsion between adjacent hydrogen atoms at the *ortho* positions favor a nonplanar conformation,[38,39] though this interpretation remains under debate.[40–42] Intramolecularly controlling the biphenyl conformation by inserting substituents or complexing with metals remains an active area of research, particularly towards the development of microscopic electron transport systems.[42–44] Alternatively, the stabilization of the coplanar BiPhe side chain in PDB 4S03 demonstrates how varying favorable (π-stacking) or unfavorable (steric/hydro-

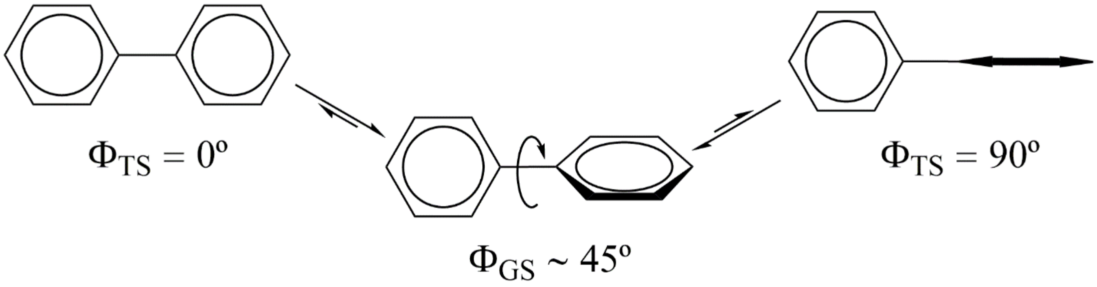


Figure 3. Rotation of biphenyl about its central dihedral angle Φ.

phobic) intermolecular forces may be used to promote the structurally "frustrated" conformation within a protein "active site".[33]

In addition to the progress the Schultz work brings in investigative bioengineering, their results raise questions about the atomic-level forces at play within proteins. For the last two decades, quantum mechanical (QM) computations have played a crucial role in investigating the structure, function and mechanism of biomolecules at the atomic level.[45–51] Certainly, the developers of multiscale enzyme modeling (QM/MM, QM/QM, ONIOM) have received accolades in the scientific community and the public at large with Warshel, Karplus, and Levitt being awarded the 2013 Nobel Prize in Chemistry.[3] Because of improvements in both computational methodology and efficiency, QM-only (also often called "QM-cluster") enzyme modeling has also advanced into a dependable tool within enzymology and biomolecular engineering to study metalloenzyme active sites.[46,47,56,48–55] As an example, previous work in our lab has shown the reliability of QM-cluster models in accurately characterizing details of the phosphoryl transfer mechanism within the Phospholipase D[57] and Tyrosyl-DNA Phosphodiesterase I[58] active sites. Although cluster models are typically used for modeling bioinorganic systems, there is not expected to be any issue with using a fully QM model to study the purely organic ThRS protein pocket. With the conformation of the rings predominantly influenced by the mutated “first shell” residues immediately surrounding BiPhe, cluster modeling becomes an efficient method to examine this system at the atom-level.

In this work, QM-cluster models are employed to computationally investigate the energetic profiles of the biphenyl dihedral angle rotation within the ThrRS cores. Several details of the Schultz work pose interesting unanswered questions. Primarily, does a Φ= 0° transition state of *p*-biphenylalanine exist in any iteration of the ThRS mutated proteins, particularly 4S03? If the TS does exist and is located at Φ= 0°, there is great likelihood that the activation energy of the coplanar TS is negligible. Then the local energy curve around Φ= 0° would show an extremely shallow double-well potential. Overcoming an existing barrier of coplanarity is expected to be thermally facile at physiological/experimental conditions. The X-ray crystal structure 4S03 would thus represent a “trapped” transition state in the sense that it would be an ensemble average of the minima on both sides of the double-well. However, if there is no computed TS at Φ= 0°, then the coplanar sidechain of *p*-biphenylalanine in the 4S03 X-ray crystal structure simply represents an energetic minimum on the potential energy landscape. The 4S03 X-ray crystal structure would then be the transition state analogue of the free biphenyl, but it cannot be labeled a TS analogue of any known enzyme mechanism. Additionally, how do the rotational energy profiles for *p*-biphenylalanine within the different protein cores compare to the rotation of free biphenyl? How much rotational flexibility is structurally and energetically permitted for *p*-biphenyl-alanine within the cores? Beyond examining the ThrRS microenvironments, this work also serves as a demonstration that properly and rationally designed QM-cluster models accurately describe non-metalloenzyme biochemistry. Our systematic method for QM cluster model creation produces protein models that can successfully emulate structural, thermodynamic, and kinetic features of the parent X-ray crystal structures on the atom-level.

Computational Methods and Model Building

Construction of the QM-cluster models began from their respective X-ray crystal structures available in the Protein Data Bank (PDB codes = 4S02, 4S0J, 4S0L, 4S0I, 4S0K, 4S03). Using a methodology currently in development by our lab to systematically construct reproducible enzyme models, we used the *Reduce*[59] and *Probe*[27] utilities and a modified version of the *RINalyzer*[26,60] code to map the topology of the X-ray crystal structures and identify the active site based upon interactions between the *p*-biphenylalanine amino acid and surrounding local protein structure. The residues determined to have important interatomic contacts with BiPhe were consistent among the six different protein models with the exception of V38, A115, and W81. From our systematic model creation scheme, residues V38 and A115 were flagged to be included in all protein models except 4S03. For consistency, V38 and A115 were still included in the 4S03 cluster model. Conversely, a BiPhe–W81 interaction was only detected within the 4S0K and 4S0L X-ray crystal structures. Two additional cluster models of 4S03 and 4S0I were constructed to include the W81 residue (labelled 4S03_W81 and 4S0I_W81). Residues included in all QM-cluster models were trimmed, with peripheral residue backbone or sidechains replaced with C–H bonds to further reduce the size of the models (see Appendix A: Table 1). As BiPhe is located within a very hydrophobic protein core, and as water molecules were not observed nearby in the crystal structures, no explicit water molecules were expected to significantly inter-act with BiPhe to warrant inclusion in the cluster models. This was reaffirmed ex post facto by molecular dynamics (MD) simulations of the enzymes (see Appendix A). Through this method, final active site models composed of 19–20 amino acids and 267–300 atoms were generated for the six proteins (see Appendix A). To retain the general shape of the active site and mimic the constrained behavior of the protein tertiary structure, C_α and select C_β atoms were frozen at their crystallographic positions, a technique that has performed reliably in other studies (Figure 4).[57,58,61] A total of 31 backbone atoms were frozen for the 4S02, 4S0J, 4S0I, and 4S03 cluster models; 33 atoms were frozen for the models containing residue W81 (4S0L, 4S0I_W81, 4S0K, and 4S03_W81 models). In addition to freezing the backbone atoms and the central biphenyl dihedral angles for desired measurements, two other parameters were constrained. In all energy scans, the β-carbon and one $H–C_\beta–C_\gamma–C_\delta$ dihedral of the *p*-biphenylalanine (Figure 5) were frozen using generalized redundant internal coordinates to limit translation of the biphenylalanine residue. A short description of how to reproduce these scans is included in Appendix A.

QM computations were performed using the Gaussian09 software program.[62] All QM computations utilized density functional theory (DFT) with the hybrid B3LYP exchange–correlation functional.[63,64] The 6-31G(d') basis set was used for N, O, and S atoms[65,66] and the 6-31G basis set was used for C and H atoms.[67] Models of free biphenyl, 4S02, 4S03, and 4S03_W81 were optimized with and without inclusion of the Grimme D3 (Becke–Johnson) dispersion correction (GD3BJ) and/or a conductor-like polarizable continuum model (CPCM)[68,69] with UAKS sets of atomic radii, a non-default electrostatic scaling factor of 1.2, and a dielectric constant of $\varepsilon= 4$, a value previously determined as appropriate for simulating the less-polarized environment within an

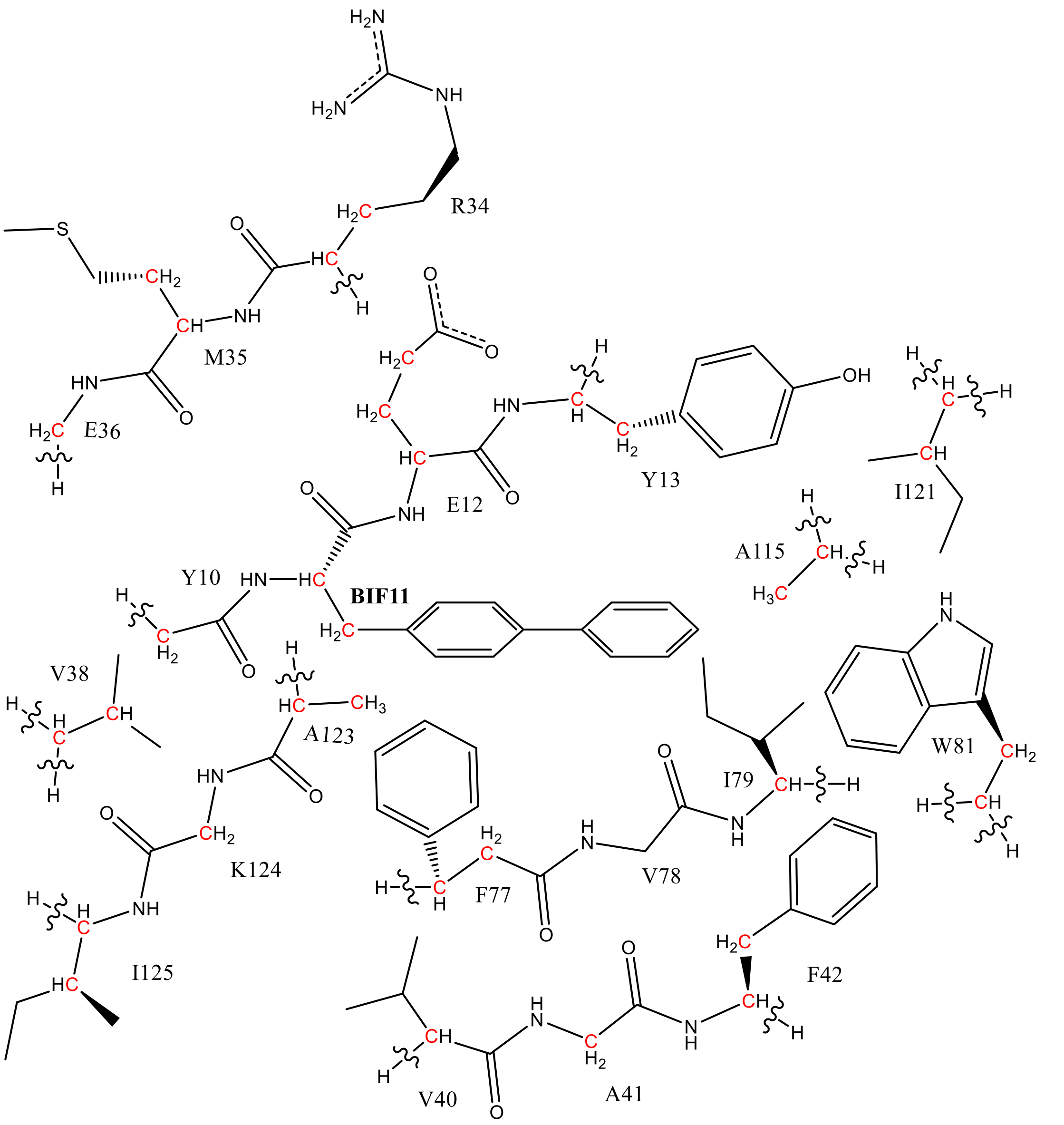


Figure 4. 2D representation of the 4S03_W81 cluster model. C_α and C_β atoms depicted in red are frozen, a total of 33 atoms (31 atoms in models without W81).

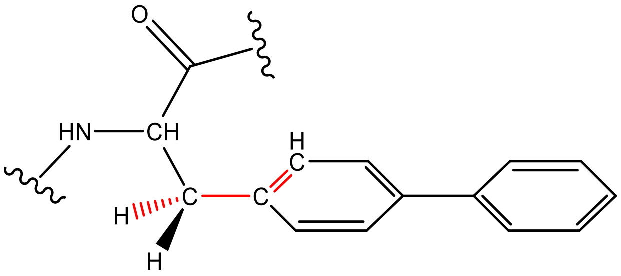


Figure 5. Structure of the non-canonical *p*-biphenylalanine residues. Red is used to indicate the frozen H–C_β–C_γ–C_δ dihedral.

enzyme active site.[49,50] Computations involving models of 4S0I, 4S0I_W81,4S0J, 4S0K, and 4S0L were performed with both GD3BJ and CPCM. Unscaled harmonic vibrational frequency calculations were used to identify all stationary points as either minima or transition states. Zero-point energies (ZPE) and thermal enthalpy/free energy corrections were computed at 1 atm and 298.15 K.

Molecular dynamics (MD) simulations were performed using the AMBER14 MD package[70] for initial structure relaxation. For all of the proteins, MD simulations were carried out using the AMBER force field ff14SB[71] and an explicit solvent model of TIP3P.[72] The proteins were solvated in a truncated octahedron water box with a 15 Å cutoff to the box edge, and Na^+ and Cl^- ions[73] were added to each system to achieve a total neutral charge. The systems were simulated using periodic boundary conditions and a cutoff value for non-bonded interactions of 8 Å. The simulations were performed using Langevin dynamics under the constant-temperature, constant-pressure (NPT) condition at 300 K and 1 atm. The SHAKE algorithm was used to constrain all bonds involved with hydrogen atoms.

All proteins were subjected to four minimization procedures followed by one relaxation procedure. All four minimizations ran 100 steps with the force constants of the harmonic positional restraints (k_{pos}) set at 20, 10, 5, and 2 kcal $mol^{-1}Å^{-2}$, applied to all heavy atoms. The relaxation procedure was run for 500 ps with k_{pos} set at 2.0 kcal $mol^{-1}Å^{-2}$ on heavy atoms before the MD simulation was run for 10 ns with k_{pos} set at 1.0 kcal $mol^{-1}Å^{-2}$. The *cpptraj* program of AMBER was used alongside Visual Molecular Dynamics[74] for analysis of the simulation trajectories.

Results and Discussion

Examination of Free Biphenyl

Previous studies of the biphenyl torsional profile have noted the challenge in obtaining accurate theoretical results for torsional activation energies.[75–77] Despite this, some calibration of the B3LYP/6-31G(d') level of theory is necessary to validate semi-quantitative accuracy in the protein cluster model energy curves. Experimental work on biphenyl in the gaseous state[37] showed $\Delta E^{\ddagger}_{\Phi=0}$ = 1.4 ± 0.5 kcal/mol and $\Delta E^{\ddagger}_{\Phi=90}$ = 1.6 ± 0.5 kcal/mol with an equilibrium dihedral angle of 44.4 ± 1.2°. Benchmark computations done by Johansson and Olsen[77] used coupled cluster theory with a Goodson continued-fraction approach and the cc-pVTZ basis set to obtain gas phase activation energies of $\Delta E^{\ddagger}_{\Phi=0}$ = 1.91 kcal/mol and $\Delta E^{\ddagger}_{\Phi=90}$ = 1.98 kcal/mol and an equilibrium dihedral angle of Φ = 38.8°.

Using B3LYP/6-31G(d') we computed gas phase biphenyl to have $\Delta E^{\ddagger}_{\Phi=0}$ = 1.94 kcal/mol and $\Delta E^{\ddagger}_{\Phi=90}$ = 2.63 kcal/mol with the equilibrium dihedral angle of Φ= 37.6°. Inclusion of CPCM reduces $\Delta E^{\ddagger}_{\Phi=0}$ and increases $\Delta E^{\ddagger}_{\Phi=90}$ values. In the gas phase, inclusion of GD3BJ noticeably increases $\Delta E^{\ddagger}_{\Phi=90}$ (Table 1). Computations using both

Table 1. Experimental and electronic energy calculations for the torsional barriers of free biphenyl at Φ = 0° and 90°.

	ΔE_0 (kcal/mol)	ΔE_{90} (kcal/mol)
Experimental (gas phase)	1.4 ± 0.5	1.6 ± 0.5
Continued fraction CCSD(T)/cc-pVTZ	1.91	1.98
B3LYP/6-31G(d')	1.94	2.37
B3LYP/6-31G(d')+GD3BJ	1.93	2.61
B3LYP/6-31G(d')+CPCM	1.55	2.67
B3LYP/6-31G(d')+GD3BJ+CPCM	1.67	2.85

GD3BJ and implicit solvation with CPCM show $\Delta E^{\ddagger}_{\Phi=0}$ = 1.95 kcal/mol and $\Delta E^{\ddagger}_{\Phi=90}$ = 2.90 kcal/mol with the equilibrium torsional angle of Φ = 35.5°. The rotational energy profiles for free biphenyl (Appendix A: Figure 1) are provided. While there is a notable overestimation of the $\Delta E^{\ddagger}_{\Phi=90}$ energy barrier when using B3LYP/6-31G(d'), the $\Delta E^{\ddagger}_{\Phi=0}$ barrier and equilibrium dihedral angle are comparable to that of Johansson and Olsen. This method will be sufficient for examining the difference in the biphenyl energetic rotation profile within the much larger QM-models of the ThrRS active site, where expensive *ab initio* methods like coupled cluster theory would be clearly intractable. Similar computations were performed on a BiPhe derivative (Appendix A: Figure 2) at the B3LYP/6-31G(d')+GD3BJ+CPCM level of theory. The rotational energy curve for the BiPhe derivative is qualitatively identical to the curve for free biphenyl, indicating the torsional rotation of the BiPhe rings is not influenced by the amino acid backbone.

Examination of Biphenyl Rotation Within the 4S02 Cluster Model

To begin the examination of the rotational energy profile of biphenylalanine within the protein cores, cluster models of the 4S02 protein pocket were constructed from the X-ray crystal structure with 10° increments in the biphenylalanine central dihedral angle in both directions. Finer 1° increments were additionally conducted near the global minimum to better determine the dihedral angle (Figure 6). Unlike free biphenyl, the QM-cluster dihedral rotational energy curves are not expected to be symmetric around the global minimum due to the various steric constraints provided by the other amino acid residues surrounding BiPhe. The effects of implicit solvation and empirical dispersion corrections on the BiPhe rotational profile were tested both individually and conjunctively on the cluster model derived from the 4S02 X-ray crystal structure. Among

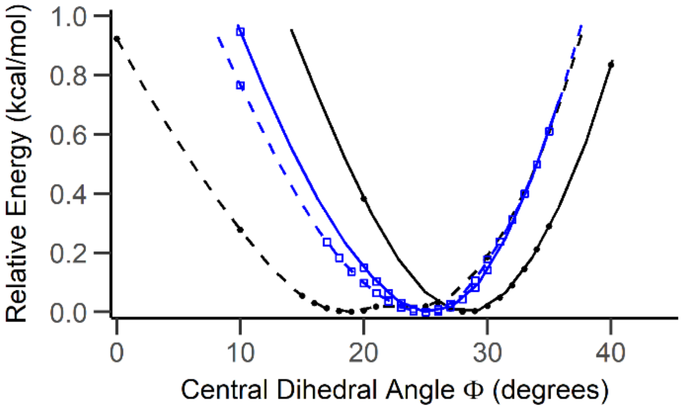


Figure 6. Potential energy curves near equilibrium for the torsional rotation of *p*-biphenylalanine within the 4S02 protein cluster model. Gas phase (black circle, solid line); gas phase with GD3BJ (black circle, dashed line); CPCM (blue square; solid line); CPCM with GD3BJ (blue square, dashed line).

the four variants of the methodology used on the 4S02 model, the computed dihedral angle at the minimum (Appendix A: Table 2) better resembled the experimentally observed dihedral angle of Φ= 26° when implicit solvation was included (Φ_{CPCM} = 25.3°; $\Phi_{\mathrm{CPCM+GD3BJ}}$ = 24.8°) compared to gas phase computations (Φ_{gas} = 28.4°; $\Phi_{\mathrm{gas+GD3BJ}}$ = 18.8°). To examine the impact of the four additional constraints placed on BiPhe in the dihedral energy scans, the previously mentioned dihedral/atom position constraints (Figure 5) were removed, and the models were re-optimized. This increased mobility of the proximal biphenylalanine rings permitted additional geometric relaxation within the protein pocket. The computed dihedrals for the less constrained models at their equilibrium geometries[1] are Φ_{gas} = 29.0°, $\Phi_{\mathrm{gas+GD3BJ}}$ = 36.5°,Φ_{CPCM} = 25.6°, and $\Phi_{\mathrm{CPCM+GD3BJ}}$ = 31.5°. The energy difference from releasing the four extra constrains in the 4S02 models is ΔE_{gas} = 0.84 kcal/mol, $\Delta E_{\mathrm{gas+GD3BJ}}$ =2.2 kcal/mol, ΔE_{CPCM} =0.83 kcal/mol,

[1] Please note that the computed ground state for $4S02_{\mathrm{CPCM+GD3BJ}}$ contained one imaginary frequency at −16.2 cm^{-1} attributable to the entire I121 residue rocking away from the BiPhe. In this entire study, this is the only occurrence of an imaginary vibrational mode observed in the QM protein minimization when one is not observed in the constrained BiPhe dihedral curve scans. This should not qualitatively affect the results, as we typically are reporting relative electronic energies.

and $\Delta E_{CPCM+GD3BJ}$ = 0.90 kcal/mol (Appendix A: Table 2). With the exception of $4S02_{gas+GD3BJ}$, the results indicate freezing the additional dihedral angle and atom position accounts for less than 1 kcal/mol difference between the models. The root mean square deviation (RMSD) between the crystal structure atomic positions and the fully optimized model (not including atoms frozen to their crystallographic coordinates or hydrogens) was evaluated for each the four 4S02 variants. The RMSD values between the 4S02 model and the original crystal structure were 0.729 Å (gas), 0.869 Å (gas+GD3BJ), 0.952 Å (CPCM), and 0.782 Å (CPCM+GD3BJ), well within the atomic resolution of the 4S02 X-ray crystal structure reported by Schultz (1.95 Å). The optimized $4S02_{CPCM+GD3BJ}$ model is overlaid with the trimmed geometry from the 4S02 X-ray crystal structure in Figure 7a. As shown, the optimized cluster model retains nearly all of its structural similarity to the 4S02 X-ray crystal structure.

There are two general features seen in the various dihedral scans of the four 4S02 potential energy curves (Appendix A: Figure 4a). First (and as expected), there is a

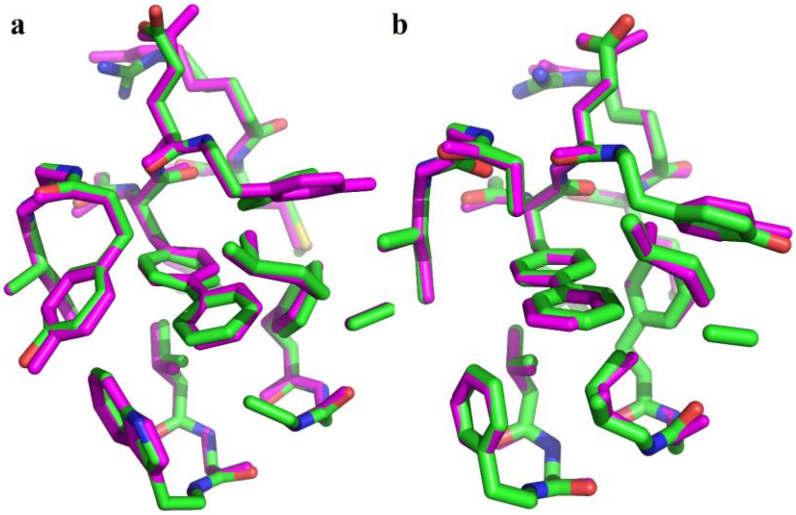


Figure 7. Overlay of the (a) 4S02 and (b) 4S03 cluster models optimized at the B3LYP/6-31G(d')+CPCM+GD3BJ level of theory (carbons colored green) compared to their respective, experimentally determined x-ray crystal structures (magenta).

drastic increase in the energy required for biphenyl to fully rotate within the protein core due to steric clashing between *p*-biphenylalanine and nearby side chains. In the computations using the 4S02 models, the measured maximum ΔE of the curve compared to the constrained minimum ranges from 11.4 kcal/mol at $\Phi_{gas} = -90°$ to 16.2 kcal/mol at $\Phi_{CPCM+GD3BJ} = -83°$. These energies for biphenyl within the sterically hindered protein microenvironment are comparable to the 6.0 to 45 kcal/mol rotational barriers of substituted biphenyls.[78] It is important to note that these "maxima" within the scans and their abrupt discontinuities (see Appendix A: Figure 4a) are indicative of the amino acid residues around the BiPhe undergoing structural relaxation in order to relieve steric strain and are not indicative of true transition states. It is also expected that the maxima of these curves have a much lower relative energy than the true activation energy; cluster models lack the many thousands of degrees of freedom affected by such a massive steric repulsion within the active site. It is certain that full 180° rotation of the BiPhe dihedral would be thermally impossible.

Second, there are discontinuities in the four 4S02 potential energy curves observed when Φ is less than −50°. Examination of those structures indicates that the unexpected reduction in relative energy seen in all four curves results from the *p*-biphenyl-alanine conformation sterically forcing the side chain of A79 to rotate from facing inside the core to outside the model (void solvent continuum), an action that would not occur in the intact protein due to rigidity of the surrounding amino acids not included in these cluster models. On the timescale of BiPhe dihedral rotation, it is unlikely that steric relaxation of surrounding residues like A79 would be a facile process. Based on a Gaussian distribution of Φ angles in the MD snapshots of 4S02 (Appendix A: Figure 8a) and all ThrRS protein cores investigated, destabilizing biphenyl ring distortion would preferentially occur, significantly increasing the energy of the model at extreme values of Φ (Appendix A: Figure 9d).

Examination of Biphenyl Rotation Within the 4S0J, 4S0L, 4S0I, 4S0I_W81, and 4S0K Cluster Models

Rotational energy curves were computed with the 4S0J, 4S0L,4S0I, 4S0I_W81, and 4S0K cluster models. Considering the previous calibrations of free biphenyl and 4S02 cluster models, the computations were performed only using the B3LYP/6-31G(d')+CPCM+GD3BJ method. The resulting energy profiles are shown in Figure 8 (except 4S0I_W81 – Appendix A: Figure 3a). Equilibrium biphenyl dihedral angles (Table 2) were computed to be within 3° – 4° degrees of the experimentally observed dihedrals, except for the 4S0I model, where a difference between the X-ray crystal structure Φ and that obtained with DFT was 16.9°. Removing the additional dihedral/atom constraints and re-optimizing the models allowed further relaxation of the BiPhe moiety similar to the 4S02 models. The ground state dihedrals for these less constrained models (Table 2) were marginally closer to the crystallographically-determined dihedrals for the 4S0J and 4S0I models, with the difference in dihedral angle between the more constrained and the less constrained models being 1.1° and 0.6°, respectively. Conversely, removing the constraints for the 4S0L, 4S0I_W81, and 4S0K cluster models permitted greater relaxation of the BiPhe dihedral. The relaxation energies from releasing the four additional constraints for the models (Table 2) are less than 1 kcal/mol except for 4S0L (2.3 kcal/mol) and 4S0I_W81 (1.4 kcal/mol). The RMSD of

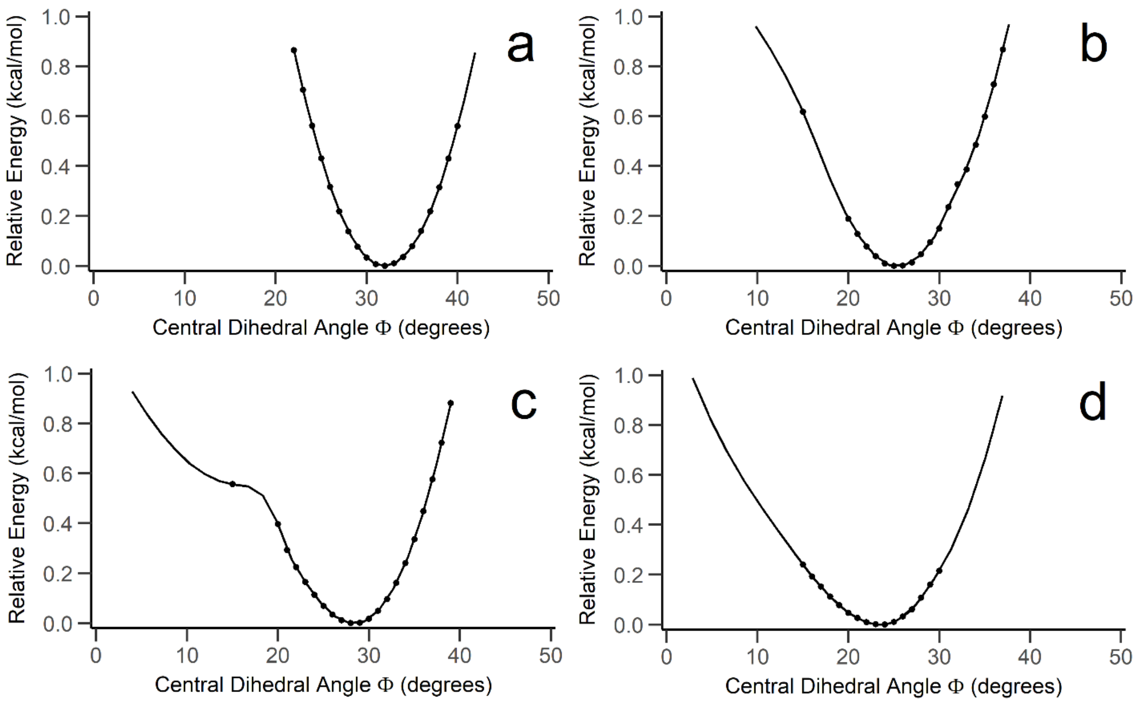


Figure 8. Potential energy curves near equilibrium for the torsional rotation of *p*-biphenylalanine within the (a) 4S0J, (b) 4S0L, (c) 4S0I, and (d) 4S0K protein cluster models, calculated at the B3LYP/6-31G(d')+CPCM+GD3BJ level of theory.

Table 2. Comparison of calculated Φ among constrained and unconstrained cluster models, computed at the B3LYP/6-31G(d')+GD3BJ+CPCM method.

Model	Experimental Φ (°)	Φ at dE/dΦ = 0 (°)	Unconstrained model Φ (°)	Relaxation Energy (kcal/mol)
4S02	26	24.8	31.9	0.9
4S0J	35	32.1	33.2	0.4
4S0L	21	25.2	32.0	2.3
4S0I	15	28.0	27.4	0.9
4S0I_W81	15	28.7	29.6	1.4
4S0K	20	23.0	26.7	0.4
4S03	0	−2.1	−6.7	0.3
4S03_W81	0	−2.5	−10.9	0.5

heavy unfrozen atoms in each model and the respective X-ray crystal structure are provided in Table 3 and are all notably within the reported atomic resolution.

In general, the qualitative trends for these rotational scans are quite similar to those seen for 4S02. Large energetic barriers of rotation ranging from 12.3 kcal/mol for 4S0J to 28.7 kcal/mol for 4S0K (Appendix A: Figure 4) are seen resulting from steric repulsion between the surrounding residue side chains and BiPhe. Regions where the models deviate onto a separate energy curve are seen in the plots for 4S0J when the dihedral angle is less than −55° and greater than 76° and for 4S0I and 4S0I_W81 when the angle is between −75° to −60°. The separate energy curves seen in the 4S0J, 4S0I, and 4S0I_W81 models all result from the BiPhe sterically forcing the side chain of S79 to rotate from facing inside the core to outside, a phenomenon seen in the 4S02 scans and unfeasible within the engineered proteins.

Table 3. Thermal flexibility of biphenyl within the protein clusters at 310K, computed at the B3LYP/6-31G(d')+GD3BJ+CPCM method, compared to dihedral values observed in MD simulations of the enzymes. Root mean square deviation (RMSD) values between the trimmed x-ray crystal structure and its respective optimized unconstrained model.

Model	Thermally Allowed Displacement from Φ_{min} (°)		Thermal Range (°)	RMSD of Cluster Model (Å)	Average Φ in MD Simulation (°)	Standard Deviation
4S02	−13.0	+10.5	23.5	0.78	15.0	7.3
4S0J	−8.6	+8.7	17.3	0.47	19.1	6.9
4S0L	−10.1	+9.8	19.9	0.87	13.0	7.3
4S0I	−12.1	+10.2	22.3	0.57	7.2	9.3
4S0I_W81	−10.4	+9.6	20.0	0.53		
4S0K	−16.2	+9.9	26.1	0.53	8.0	8.9
4S03	−12.1	+13.3	25.4	0.37	−0.3	8.3
4S03_W81	−11.5	+14.0	25.5	0.47		

The rotational energy profiles of 4S0I and 4S0I_W81 are nearly identical. Both 4S0I and 4S0I_W81 have a similar minimum dihedral angle, with a difference of only 0.7°, though this computed angle is also atypically larger than the experimental angle by ~13°. The minima computed when the proximal dihedral constraints are released are also similar. The dihedral angles differ by only 2.2° and the relaxation energy differs by 0.5 kcal/mol. The lack of a substantial difference between the 4S0I and 4S0I_W81 results reaffirms the expectation formed during model construction of a negligible BiPhe–W81 interaction.

Examination of Biphenyl Rotation Within the 4S03 Cluster Models

The 4S03 and 4S03_W81 optimizations and rotational curves were also computed with and without CPCM and/or GD3BJcorrections. Among the four variants for the cluster models, the computed dihedral angle at the minimum was closer to the experimental dihedral of Φ = 0° when implicit solvation and empirical dispersion was included in the models (Figure 9 and Appendix A: Figure 3b) with $\Phi_{CPCM+GD3BJ} = -2.1°$

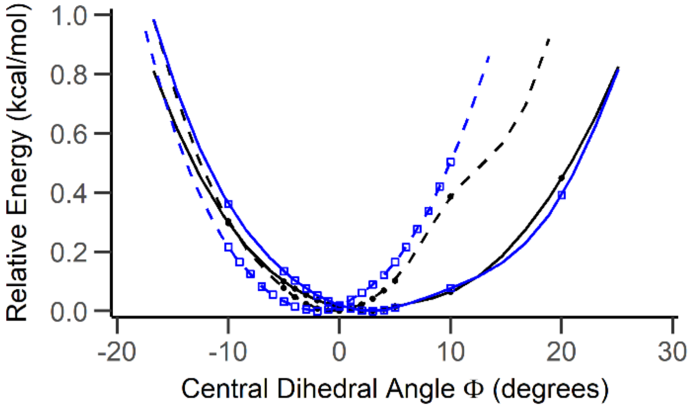


Figure 9. Potential energy curves near equilibrium for the torsional rotation of *p*-biphenylalanine within the 4S03 protein cluster model. Gas phase (black circle, solid line); gas phase with GD3BJ (black circle, dashed line); CPCM (blue square, solid line); CPCM with GD3BJ (blue square, dashed line).

for 4S03 and −2.5° for 4S03_W81. Re-optimizing the models without the proximal dihedral constraints resulted in equilibrium dihedral values (Appendix A: Table 2) significantly different from the experimental value when empirical dispersion corrections were not included (4S03: Φ_{gas} = −27.5°,Φ_{CPCM} = −28.9°) compared to when they were included (4S03: $\Phi_{gas+GD3BJ}$ = 9.4°, $\Phi_{CPCM+GD3BJ}$ = −6.7°). In comparing the relaxed 4S03 and 4S03_W81 models, there is a dihedral angle difference of only 2.1° and a relaxation energy difference of only 0.2 kcal/mol (Table 2). Similar to the comparison between 4S0I and 4S0I_W81, there is no substantial difference between the results by including the W81 residue in the 4S03 cluster model. This is expected, as a BiPhe–W81interaction was not detected as a necessary residue for the 4S03 cluster model. As with the previous protein models, the RMSD values (Table 3) for 4S03 and 4S03_W81 are all within the reported atomic resolution of 2.05 Å for the 4S03 X-ray crystal structure. The $4S03_{CPCM+GD3BJ}$ optimized model is overlaid with the trimmed geometry from the 4S03 X-ray crystal structure in Figure 7b, demonstrating how the optimized cluster model retains the positioning of the residues.

Thermal Flexibility and Transition State Searches

While the computed energy profiles indicate that complete torsional rotation of BiPhe would be impossible under physiological conditions, it is important to consider the range of rotational flexibility energetically permitted within the different ThRS protein models under investigation. To semi-quantitatively examine this property, we approximate the available rotational energy at physiological temperature (310 K) via the Boltzmann expression $E = k_B T$ = 0.62 kcal/mol. Fitting a second-order polynomial to the computed electronic energies for each of the models, dihedral angles with a relative energy of +0.62 kcal/mol from the minima were defined as the range of Φ where BiPhe may freely fluctuate at 310 K (Table 3). Our computations suggest that the biphenylalanine rings have a reasonable amount of flexibility within the protein cores ranging from a total of 17.3° to 26.1°. The computed thermal ranges also indicate the BiPhe rings may rotate from the equilibrium geometry both in positive and negative directions. The X-ray crystal structures represent an ensemble average of thermally allowed rotational states of the BiPhe dihedral. It may be hypothesized that greater flexibility of the BiPhe may correlate with an increase in X-ray crystallographic resolution. However, comparison of the computed thermal range of BiPhe to the respective crystallographic resolution of the ThRS proteins engineered by Schultz (Table 3) indicated no significant correlation between the two factors (Appendix A: Figure 5, R^2= 0.149).

MD simulations were performed on the six ThrRS proteins in the interest of providing additional insight into the flexibility of BiPhe and its surroundings. Comparing the MD simulation snapshots to their respective X-ray crystal structure, the average root-mean square deviation (RMSD) of the non-hydrogen atoms for the proteins ranged from 0.33 to 0.52 Å (Appendix A: Table 4 and Figure 6). Considering only the non-hydrogen atoms of the residues included in our cluster model, the average RMSD between the crystal structure and MD snapshots ranged from 0.28 to 0.31 Å. The absence of substantial structural changes between the simulated and PDB crystal structures indicates the protein effectively retains its tertiary structure without noticeable unfolding. Likewise, the BiPhe core maintains its structural integrity. Most of the BiPhe core residues tend to exhibit little change among the different protein models (Appendix A:

Figure 7) compared to X-ray crystal structures. Unsurprisingly the largest differences are observed with residues that undergo point mutation during the progression from 4S02 to 4S03.

Examination of the BiPhe central dihedral angle in the MD simulation snapshots presents results that are considerably different from the experimental and cluster model-based results. The average BiPhe dihedral angle for each of the protein simulations (Table 3) is shown to be significantly lower than the experimental value, with the exception of 4S03. The histogram of the dihedral values for 4S03 (Figure 10) demonstrates an anticipated result: a relatively Gaussian distribution of the data centered on/near the experimentally observed dihedral in the 4S03 X-ray crystal structure. Visualization of these 4S03 structures with the largest, smallest, and average dihedral angle is provided in Appendix A: Figure 9. The histograms of the biphenyl dihedral values in the other enzymes (Appendix A: Figure 8) demonstrate distributions shifted to favor smaller dihedral angles of BiPhe. This tendency towards more coplanar conformations is likely due to the generated AMBER force field parameters for BiPhe

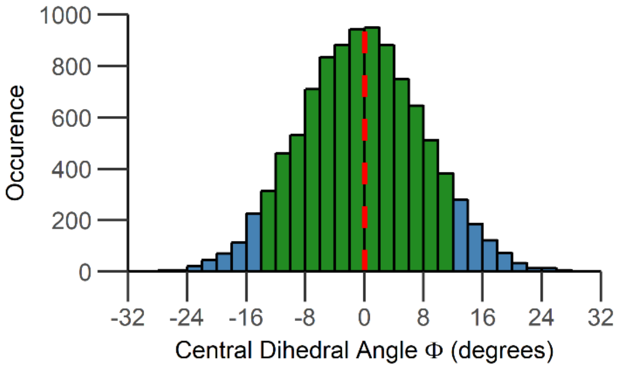


Figure 10. Frequency of BiPhe central dihedral angles in MD simulation snapshots of 4S03. The red dashed line represents the X-ray crystallographically determined value. Green is used to represent dihedral angle values within the thermal range calculated by our cluster models.

being insufficient to account for the intricate steric/electronic competition between the rings. A more fine-tuned parameterization of BiPhe is thus needed to more adequately model this residue outside the constrained 4S03 protein, a feat beyond the scope of this work. As an additional comparison between the MD simulation and the QM-cluster model results, the ranges of the histograms with values within the thermal range computed by the cluster models were colored green. Focusing on the 4S03 results (Figure 10), 87% of the dihedral values in the snapshots were within the thermal range, with the remaining 5% and 8% outside the left and right boundaries, respectively. In this manner, the MD and cluster model results are in close agreement with the relative flexibility of the dihedral angle within the 4S03 protein core. Interestingly, snapshots farther outside of the thermal range display severe biphenyl ring distortion rather than steric accommodation of the ThRS residues around the BiPhe (Appendix A: Figure 9). Similar behavior is observed by Masson in the study of dihedral rotational profiles of substituted biphenyls.[78] As the minimization of BiPhe to a large value of Φ can be considered a “rare event” in the 10 ns MD simulations of ThRS, we could estimate a rate constant based on one complete dihedral rotation of BiPhe every 10 ns. At 310 K, this would correspond to an extremely conservative $\Delta G^{\ddagger}$ of 6.5 kcal/mol, but actual values are likely closer to the barrier heights shown in Appendix A: Figure 4.

Lastly, to determine if a coplanar transition state persists in any of the protein models, transition state searches near Φ=0° were conducted for each of the models. Transition states with a definitive imaginary vibrational frequency attributable to the rotation of the biphenylalanine central dihedral were identified for 4S02 at $\Phi_{gas} = 9.3°$, 4S0I at $\Phi_{CPCM+GD3BJ} = -3.2°$, and 4S0K at $\Phi_{CPCM+GD3BJ} = 2.4°$. Activation free energies were computed as the difference between the transition state conformations and their unconstrained ground state conformations and were 1.28 kcal/mol, 1.55 kcal/mol, and 0.66 kcal/mol, respectively. Of particular significance is the fact that these energies are comparable to the rotational energy for free biphenyl (Figure 11). The similar energy barriers support the notion that the detected near-coplanar TSs within 4S0I and 4S0K are directly attributable to the BiPhe rings rotating through the unfavorable coplanar conformation. Collectively, this evidence strongly suggests the coplanar conformation of biphenylalanine exists within the 4S0I and 4S0K protein models as a computationally detectable local maximum on their potential energy curves. It is important to note that while these two transition states are observed using DFT, they will have no qualitative impact on the potential energy surface. As the energy barrier for the reverse rotation is negligible, it is expected that the BiPhe rings will undergo facile relaxation to the global minimum conformation.

No transition states near Φ = 0° were found for the 4S03 or 4S03_W81 models, indicating the surrounding steric forces of the hydrophobic residues acting on

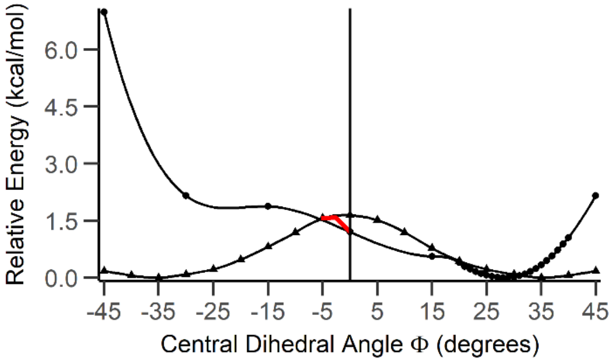


Figure 11. Potential energy curves for the torsional rotation of *p*-biphenylalanine within the 4S0I protein (circles) with the proposed transition state maxima and local minimum (red) in comparison to the potential energy curve for free biphenyl (triangles).

biphenylalanine effectively counteract the intramolecular H–H steric repulsion of the coplanar conformation. With this juxtaposition of opposing forces, the coplanar conformation exists as an energetic minimum within the designed 4S03 protein model rather than a “trapped” local transition state.

The concept of the 4S03 structure sterically compacting the BiPhe side chain into the coplanar conformation is further reflected in the number and type of residue–BiPhe interactions computed by *probe* (Appendix A: Table 5) during cluster model construction. The last three ThRS synthetase proteins (4S0I, 4S0K, and 4S03) each differ by a single amino acid substitution of Y79S, Y79V, or Y79I, respectively. As each residue substitution is characterized by an increase in the size of the side chain and in hydrophobicity, it may be expected that the number of computed BiPhe ↔ residue interaction counts (RICs) will increase going from 4S0I to 4S03. A distinct increase in the number of RICs is indeed seen for residue 79 as *probe* computes 4 RICs for Y79S (4S0I), 14 RICs for Y79V (4S0K), and 17 RICs for Y79I (4S03). Additionally, the RICs for residue A123 in all three proteins are noted to also increase from 3 RICs (4S0I) to 9 RICs (4S0K) and 10 RICs (4S03). These increased RICs result from increased steric clashing between the BiPhe and A123 as BiPhe shifted to accommodate the larger Y79I side chain. With each mutation of Y79, the number of BiPhe RICs progressively increased with the final Y79I mutation in 4S03 providing sufficient steric and hydrophobic interactions to force the BiPhe to be coplanar. Thus, the RICs reflect how the designed 4S03 core effectively compacts the coplanar BiPhe side chain. As seen in our work, there is potential in using RICs (or other chemical mapping methods) to qualitatively assess protein structure at the residue level. As these interaction networks may be obtained with negligible computational cost, we speculate that they may prove to be a novel qualitative measure for prediction of mutant protein thermostability and rational bioengineering.

Conclusions

QM-only cluster models ranging from 267 – 300 atoms of several bioengineered threonyl-tRNA synthetase proteins were constructed to examine the energetics of the torsional rotation of the noncanonical *p*-biphenylalanine residue. We successfully computed resting dihedral angles for most of the models within 1–4° of the experimental X-ray crystal structures. Unlike free biphenyl, complete rotation of the biphenylalanine rings within the proteins requires overcoming substantial energetic barriers of at least 5 – 15 kcal/mol, which are likely much higher for the actual ThRS potential energy surfaces. These barriers are noted to be similar to the rotational activation energies of substituted biphenyl molecules. While complete rotation of the biphenylalanine rings is not facile, the rings are also not rigidly constrained and may fluctuate at 310 K by 17.3° – 26.1°. Transition state searches near $\Phi = 0°$ were conducted to determine whether a coplanar biphenylalanine transition state exists within the various models. We identified coplanar TSs for the 4S02 (gas phase), 4S0I, and 4S0K cluster models, and the activation energies for these are noted to be close to the barrier height for free biphenyl. The evidence of these detected transition states suggests an intramolecular steric transition state of biphenylalanine exists within the 4S0I and 4S0K models; however, this transition state has a negligible energetic barrier and so the transition state is not expected to impact the qualitative potential energy surface of the proteins.

Transition states near $\Phi = 0°$ were not identified for 4S03 or 4S03_W81 models, suggesting that the surrounding steric forces acting on biphenylalanine effectively counteract the intramolecular H–H steric repulsion of the coplanar conformation, resulting in the coplanar conformation being an energetic minimum. While the computational evidence does not support the idea that the 4S03 X-ray crystal structure represents a "trapped" transition state alluded to in the work of Schultz and coworkers, their use of rational computational bioengineering generated a protein capable of stabilizing an energetically unfavorable conformation, which is a remarkable and highly commendable accomplishment. Indeed, their general method allows new insight into the function and mechanism of proteins, along with the potential to design proteins with new properties, which is work in progress in the Schultz laboratory.[79] Future investigations could pursue enzyme modifications to stabilize/"trap" TS analogues of bond breaking/forming reactions, and this study validates a supporting role for quantitative QM-cluster model computations.

This work reiterates computationally what is expected to chemically occur during the process of protein design and bioengineering. The interaction-based cluster models demonstrate the impact inter-residue steric repulsions play in determining the conformation and orientation of nearby residues. This work further demonstrates how protein models constructed based upon inter-residue interactions provide semi-quantitatively accurate results without using QM/MM or ONIOM models. Just as our QM-cluster model creation scheme was able to provide valid results in this proof-of-concept study, we anticipate similar reliability in its application to many useful topics in biochemistry and biophysics, such as elucidation of enzyme mechanisms, molecular basis of disease, *in silico* protein engineering, and drug design.

Chapter 3: Human Carbonic Anhydrase II

Introduction

Metalloenzymes have crucial roles in living organisms, ranging from facilitating cellular signal transduction pathways to functionalizing substrates. Many of these proteins utilize the trace element zinc, with up to 10% of the human genome potentially encoding zinc-binding proteins.[80] Carbonic anhydrases[81] (CA) are a family of these zinc-dependent enzymes that rely upon a relatively simple zinc-bound active site to activate water to hydrolyze carbon dioxide through the reaction:

$$CO_2 + H_2O \rightleftharpoons HCO_3^- + H^+ \tag{1}$$

With this functionality, CAs are found widely throughout organisms within all three domains and play important roles in cellular respiration, pH and fluid homeostasis, and carbon dioxide fixation.[81,82]

Although there are several different isoforms of CA, the most well studied is human carbonic anhydrase II (HCAII), which is a 32 kDa monomeric α-class protein that supports a four-coordinate Zn^{2+} center (Figure 12).[81,82] Over the past twenty years, it has been observed that this protein is cambialistic in which the native Zn^{2+} may be substituted with other divalent metal ions and the enzyme still retains some catalytic activity. In general, the metal binding affinity of CAs follow the Irving-Williams series ($Mg^{2+} < Fe^{2+} < Co^{2+} < Ni^{2+} < Cu^{2+} > Zn^{2+}$) with the exceptions of the native Zn^{2+} having a greater affinity than Cu^{2+} and Fe^{2+} having a smaller affinity than Mg^{2+}.[83] Although HCAII is capable of binding to different divalent transition metals, the enzymatic activity of these HCAII variants is significantly reduced, with only Zn^{2+} having a high catalytic activity.

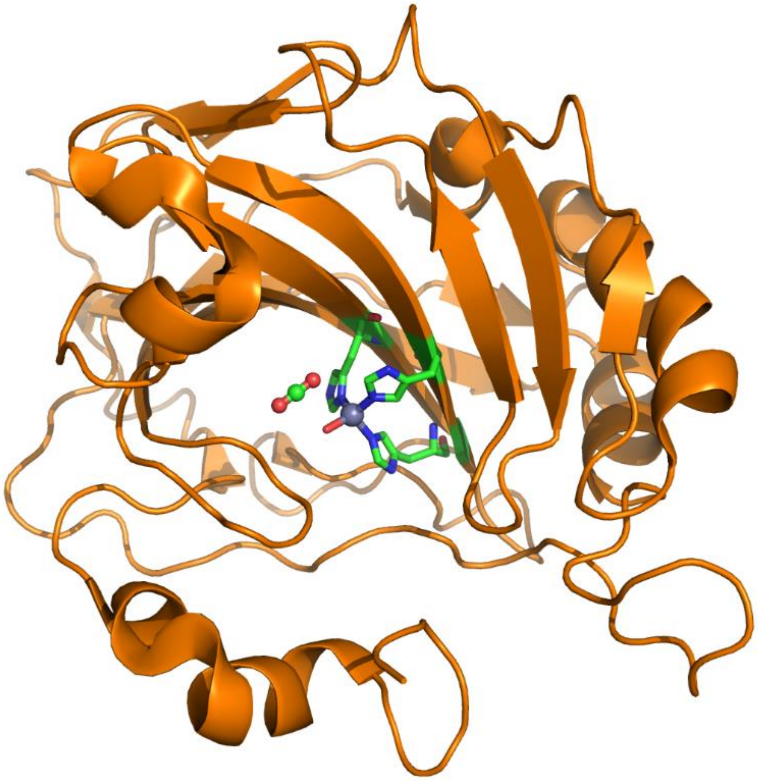


Figure 12. Cartoon representation of HCAII (PDB: 3D92) with carbon dioxide in the active site pocket and the zinc bound to a water and three histidines (stick representation with green carbon atoms; hydrogens are omitted).

HCAII with Co^{2+}, Mn^{2+}, Fe^{2+}, Ni^{2+}, and Cd^{2+}-bound showed approximately 50%, 8%, 4%, 2%, and 2% activity respectively (compared to Zn^{2+}) while HCAII with Cd^{2+} did not show any detectable activity.[81,82,84–86]

It has been reported[87] that the γ-class CA of the anaerobic archaeon *Methanosarcina thermophilia* may be successfully reconstituted with Fe^{2+} in an anaerobic environment to yield an enzyme with a catalytic efficiency greater than Zn^{2+}-reconstituted CA. Subsequently exposing the Fe^{2+}-reconstituted enzyme to hydrogen peroxide oxidized the Fe^{2+} to Fe^{3+} and inactivated the enzyme by dissociating Fe^{3+} from the enzyme active site. These results suggested that previous reports of the poor activity of Fe-substituted γ-class CA may have arisen from Fe^{2+} being oxidized to Fe^{3+} during aerobic purification. Although there are substantial structural differences between the γ-class CA of *Methanosarcina thermophilia* and α-class HCAII, the aforementioned finding leads to questions into whether similar effects may impact the activity[86] of Fe-substituted HCAII. In light of this information, we seek to use a theoretical approach to investigate the mechanistic details of Fe- and other metal-substituted HCAII variants.

There are two commonly proposed reaction mechanisms for native HCAII.[81] Both begin with a metal-bound hydroxide performing a nucleophilic attack on carbon dioxide to form bicarbonate with two oxygens bound to the metal. The Lipscomb mechanism proposes that dual proton transfers occur between the bicarbonate and the nearby, evolutionarily conserved threonine and glutamic acid residues before the bicarbonate detaches from the metal. Alternatively, the Lindskog pathway proposes the bicarbonate breaks away without additional steps (Figure 13). A water molecule then replaces bicarbonate, and the catalytic hydroxide is regenerated via a proton dissociation shuttle transferring a water proton to the protein surface and subsequent solvent. For decades, there has been debate over which pathway is preferred, but recent studies tend to suggest both mechanisms are competitive with each other.[88,89,98,99,90–97] This historic inability to identify the preferred pathway(s) computationally is in part based on the wide variability in models used to simulate this biosystem. Models range from minimal QM-cluster models (e.g. a model composed of only a CO_2 and $[(NH_3)_3Zn{-}OH]^+$) to more recent QM/MM simulations.[88,89,96–100]

In this work, we seek to investigate the viability of Fe-substituted HCAII by quantum mechanically simulating both the Lipscomb and Lindskog reaction mechanisms. For improved context and comparison, reaction pathways for models with Zn^{2+}, Co^{2+}, Mn^{2+}, Ni^{2+}, and Cd^{2+} as the active site ions will also be computed. This work will use

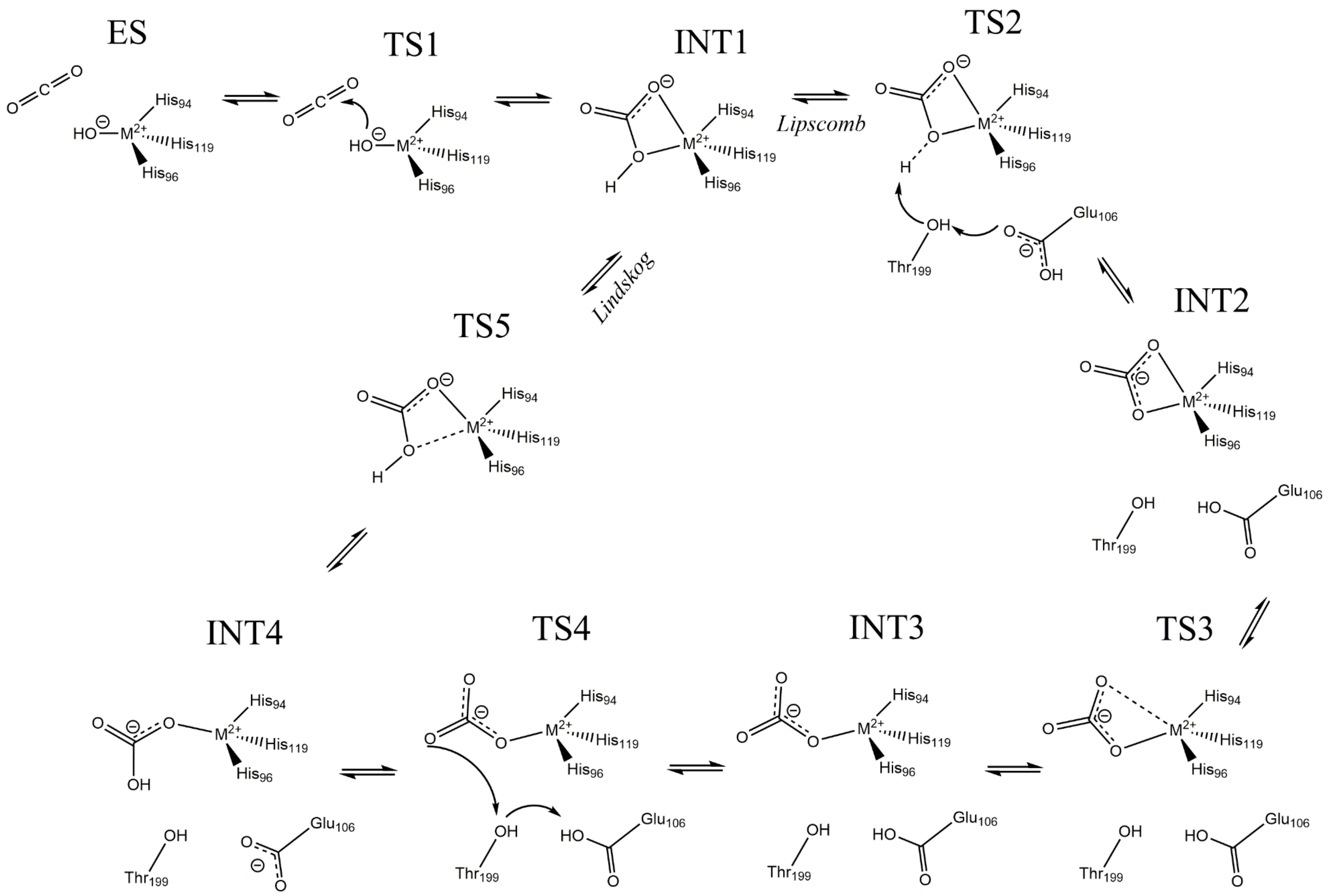


Figure 13. Part of the proposed carbonic anhydrase reaction pathway illustrating the differences between the Lipscomb and Lindskog mechanisms. Steps shown include the enzyme-substrate complex (ES), transition state (TS) and intermediate (INT) structures. Labels correspond to coordinates for the reactions where the active site metal (M) is Zn^{2+} or high-spin Fe^{2+}.

information from protein residue interaction networks to construct models large enough to mimic the active site microenvironment better than previously reported minimal QM-cluster models while avoiding simulating the entire protein via QM/MM simulation. Through this work, we seek to efficiently (and accurately) obtain atomic-level insight into the HCAII active site and its cambialistic properties.

Methods

The QM-cluster models of the CA active site were constructed using the X-ray crystallographic coordinates of a cobalt-substituted derivative of HCAII (PDB: 3KOI).[101] Hydrogen atoms were added to the structure using the program *Reduce*.[59] To identify the residues that craft the active site microenvironment, the inter-residue topologies for three HCAII X-ray crystal structures (PDBs 3KOI, 3D92, 1CAH) were mapped by using the *Probe* software[27] to identify inter-residue contact interactions. The computed inter-residue contact interactions were used to construct a residue interaction network, a graph that translates the three-dimensional protein structure into a two-dimensional network of residues (called “nodes”) interconnected by their contact interactions (called “edges”).[25] The three X-ray crystal structures examined each have unique characteristics which may help ensure crucial interactions are captured in modeling the various metal-substituted systems. The cobalt-substituted structure 3KOI has three waters coordinated to the metal to give an octahedral geometry;[101] the 3D92 structure is obtained from CO_2-pressurized, cryo-cooled crystals, capturing the location of CO_2 within the active site;[102] and the cobalt-substituted structure 1CAH has bicarbonate complexed to the metal.[103] The contact maps of these three proteins were analyzed to identify residues interacting with either the metal-coordinated waters, the three metal-coordinated histidine side chains

(His94, His96, His119), the unbound CO_2 molecule (within 3D92) and the metal-bound bicarbonate molecule (within 1CAH). A total of 15 residues were identified as having contact interactions with the aforementioned species. Three interacting residues were excluded from the final model (Asn67 and Thr200 had very few contact interactions and Phe66 only had minor non-hydrogen bonding backbone contacts with His94), allowing the final list of directly interacting residues modeled to be Gln92, Phe93, Phe95, Glu106, Glu117, Leu118, Val121, Val143, Leu198, Thr199, Trp209, and Asn244.

The rest of the protein was trimmed away (see Appendix B: Table 1) and places where bonds were broken were capped with C–H bonds to satisfy valency. A water molecule was also included in the model positioned at the crystallographic coordinates of Wat592 of PDB entry 3D92. The final model for the Zn-, Fe-, Cd-, and Co-composed active sites is composed of 224 atoms. An additional metal-coordinated water is added to the Mn- and Ni-substituted model to yield a 227-atom model. To mimic the generally rigid nature of the protein backbone, a total of 22 C_α and select C_β atoms were frozen to their crystallographic coordinates (Figure 14).

QM computations were conducted using the Gaussian16 software[62] using density functional theory. The hybrid B3LYP exchange-correlation functional[63,64] was used with 6-31G(d') basis set for N and O atoms,[66] 6-31G basis set for C and H atoms,[67] and LANL2DZ basis and effective core potentials for the metals.[104] The models were simulated with the inclusion of the GD3BJ dispersion correction and a CPCM environment using UFF sets of atomic radii, a non-default electrostatic scaling factor of 1.2, and a dielectric constant of ε=4.[68,69,105] Unscaled harmonic vibrational frequency

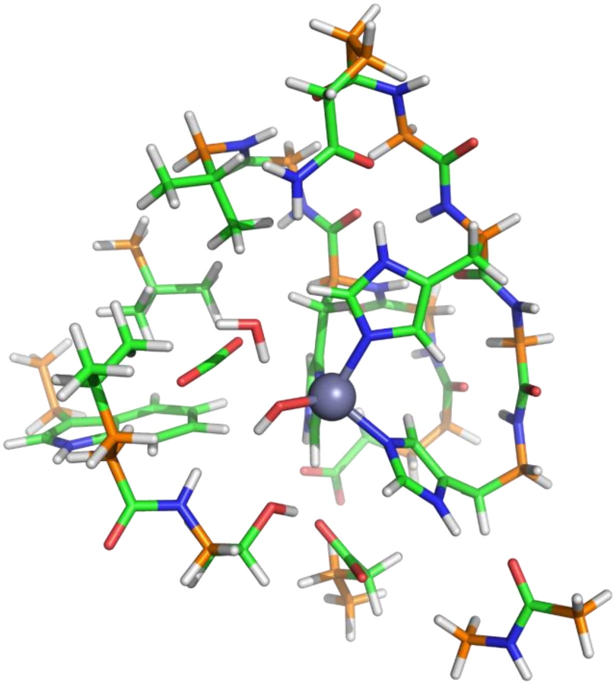


Figure 14. The 224-atom QM-cluster model of the CA active site. The C_α and select C_β atoms which were frozen to their crystallographic coordinates are indicated in orange.

calculations were used to confirm all stationary points as either minima or transition states.

Results and Discussion

Proposed Mechanism for Zn-HCAII

The relative Gibbs free energies for stationary points along both Lipscomb and Lindskog reaction pathways (Figure 13) were computed using the designed QM-cluster models, and the resulting energy profile for the mechanisms are shown in Figure 15. The results indicate that at the point where the mechanisms deviate, the dual hydrogen transfer of the Lipscomb mechanism is barrierless (TS2) and leads to a 4.5 kcal/mol more stable intermediate (INT2) while the Lindskog mechanism requires 9.3 kcal/mol to break the Zn-O bond (TS5). However, the energy for the carbonate to break its Zn-O bond (TS3, $\Delta\Delta G$ = 15.1 kcal/mol) is greater than that for the bicarbonate Zn-O bond to break

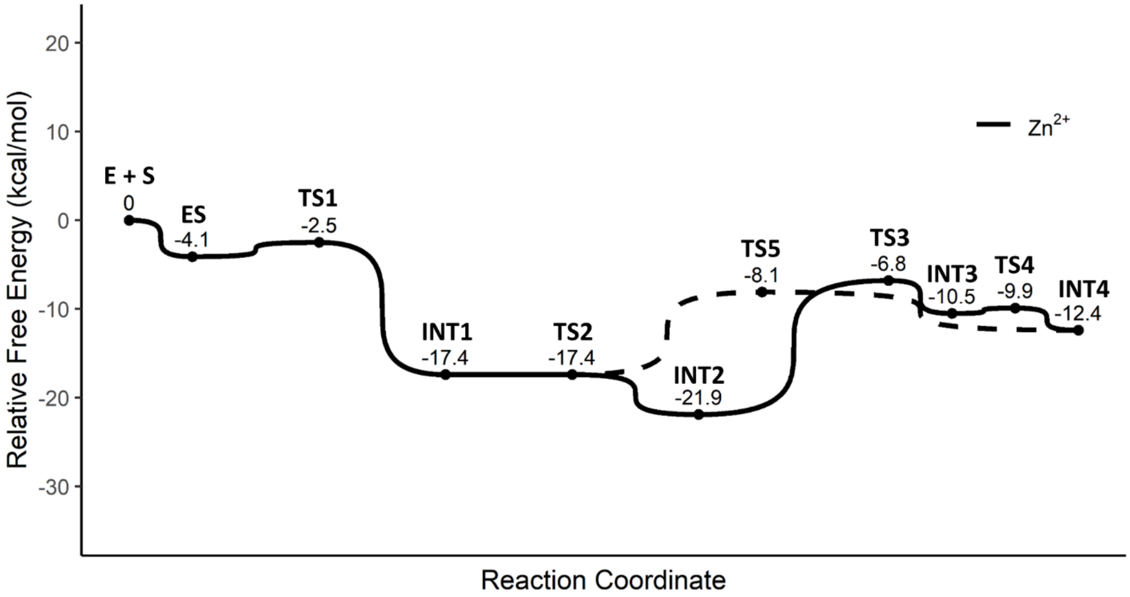


Figure 15. Free energy profile for the Lipscomb (solid) and Lindskog (dashed) reaction mechanisms for HCAII with Zn^{2+} as the active site metal.

(TS5, ΔΔG = 9.3 kcal/mol). Reversibility of INT2 to INT1 may be expected to occur given the low activation energy for the reverse reaction, enabling the reaction to follow the slightly more energetically favorable Lindskog pathway. However, both Lipscomb and Lindskog rate determining steps are feasible to overcome and both pathways may be expected to be competitive.

In comparing these results to other free energy profiles in the literature, the thermodynamics of our computed mechanisms have characteristics reportedly attributable to the absence of extensive water networks within the active model (e.g. within QM/MM or QM-only models with additional water clusters).[88,90] For example, the TS1 activation energy (ΔΔG = 1.6 kcal/mol) is lower than the ~7 kcal/mol activation energy reported for water-packed models as the hydroxide and CO_2 substrate do not have to break through a "wall" of waters to bind. The calculated value is instead comparable to the ~1 kcal/mol activation energy reported in QM-models with few waters.[88,89,106] The issue of how to properly ensure waters are identified and modeled appropriately in active site models remains a complicated topic, compounded by the mobility of waters and their typically poor resolution in X-ray crystal structures. Efforts to design solvent-accessible active sites that better account for the presence of waters though MM or brief MD enzyme solvation simulations are underway by our lab. Apart from the water differences, the results reported here are comparable to those from reported "larger" QM-models, providing good evidence to support the accuracy of this computational method. These results with Zn in the active site will serve as a reference of comparison against the other metal- substituted HCAII models.

Proposed Mechanism for Co-HCAII

The Co^{2+}-substituted metallovariant of HCAII was reported as having the next highest enzymatic activity after Zn^{2+}. Like Zn^{2+}, the metal coordination geometry is tetrahedral for the lone enzyme (E+S) and substrate-bound starting structures (ES). The computed reaction energy profile (Figure 16) indicates that upon nucleophilic attack of the Co^{2+}-bound hydroxide to the CO_2 (TS1), the structure rapidly relaxes to the bidentate carbonate within the Lipscomb pathway (INT2) where the hydrogen has been transferred to the Thr199; an INT1 stationary point structure was not able to be isolated. Breaking of one of the bidentate carbonate bonds in concert with the transfer of the hydrogen from Thr199 to the carbonate may occur (TS5) to lead to the monodentate bicarbonate (INT4), but this process is computed to require 18 kcal/mol to occur. Given the significant stability of the bidentate carbonate structure, our model is unable to isolate the Lindskog pathway structures.

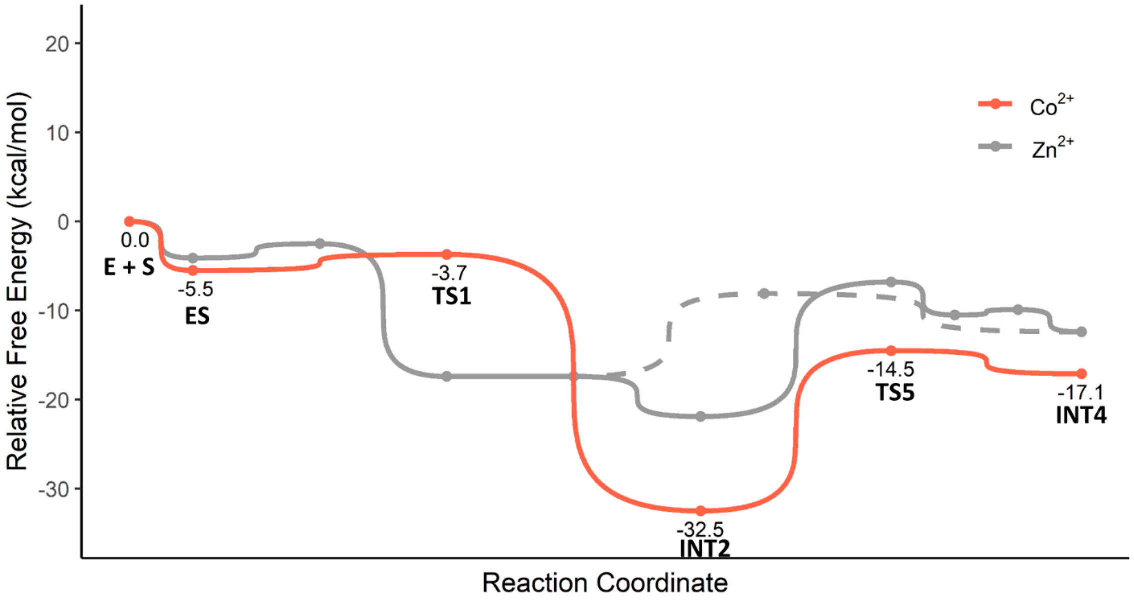


Figure 16. Free energy profile for the modified Lipscomb reaction mechanism or HCAII with Co^{2+} as the active site metal compared to Zn^{2+} (grey).

Since the computation of this free energy profile, recent X-ray crystallographic studies have provided additional insight into the Co^{2+}-HCAII mechanism of action.[107] These crystal structures reveal that while a tetrahedral coordination is confirmed for E+S and ES states, the coordination is unusually expanded into an octahedral coordination upon formation of the bicarbonate whereby the Co^{2+} is coordinated to the three protein histidines, a bidentate bicarbonate structure, and an additional water molecule. The authors of this work[107] theorize that the binding mode of bicarbonate to the Co^{2+} will be stronger than that of Zn^{2+} and that deprotonation of the additional Co^{2+}-bound water molecule may facilitate dissociation of bicarbonate due to charge-charge repulsion between the resulting hydroxide and bicarbonate. The QM-cluster models support the idea of the Co^{2+} having tighter product metal-binding than Zn^{2+} (TS5, ΔΔG = 18.0 kcal/mol), but given the absence of additional waters within this model, they are not capable of giving insight into the impact the additional coordinated water and its deprotonation may have on the reaction mechanism. This inquiry will be investigated in the future by adding several more waters to the model and recomputing the stationary points to include examination of the octahedrally coordinated structures.

Proposed Mechanisms for Mn-, Ni-, and Cd-HCAII

The Mn^{2+}- and Ni^{2+}-substituted metallovariants of HCAII were reported as having poor enzymatic activity (8% and 2% respectively) compared to the activity of Zn^{2+}. Both QM-models include an additional water coordinated to the metal, beginning with a square pyramidal starting geometry. This difference in metal and coordination alters the predicted mechanism (Figure 17) and subsequent reaction energy profiles (Figures 18 and 19). Both Mn^{2+}- and Ni^{2+}-HCAII display a remarkable change in activation energy for the beginning step, increasing from <2 kcal/mol for both Zn^{2+} and Co^{2+} to ≥11.2 kcal/mol (TS6). This substantial increase in activation energy suggests the beginning coordination geometry is not as conducive towards reaction initiation as the tetrahedral geometry and may partly explain why enzymatic activity for the Mn^{2+} and Ni^{2+} metallovariants are poor. The TS6 activation energy for Mn^{2+}-HCAII is 6.6 kcal/mol greater than that for Ni^{2+}-HCAII, suggesting reduced enzymatic activity for the former than the latter which is not observed experimentally; other phenomena not captured in this model are likely involved. Besides the increase in activation energy for TS6, it is noted that the other transition state and intermediate structures are not as stabilized within the Mn^{2+}- and Ni^{2+}-active sites compared to the Zn^{2+}- and Co^{2+}-equivalents and the net reactions are less exergonic.

X-ray crystal structures reveal hexacoordinate geometries for Mn^{2+}- and Ni^{2+}-HCAII, which will have to have one of their waters displaced for the reaction to

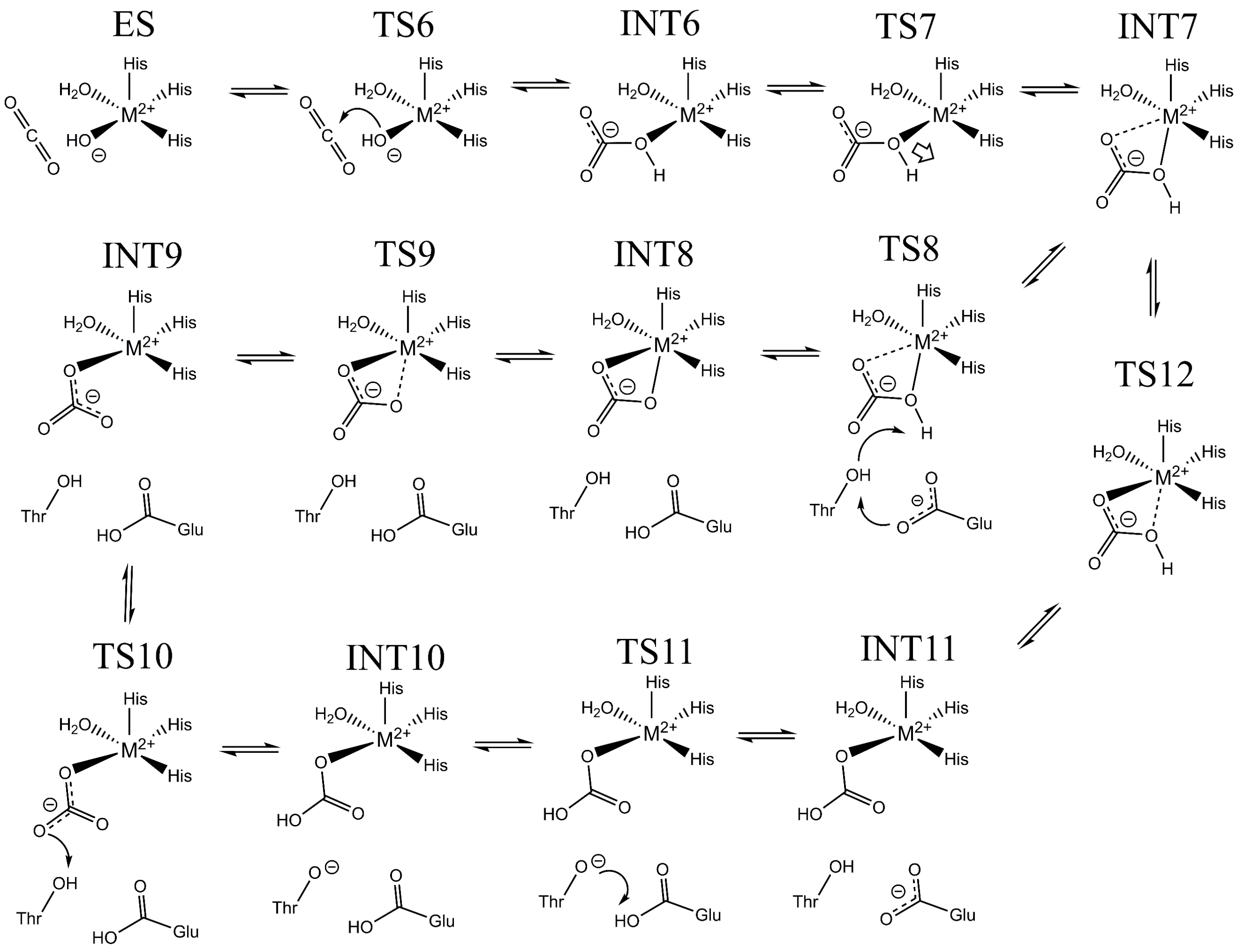


Figure 17. Part of the proposed carbonic anhydrase reaction pathway illustrating the differences between the Lipscomb and Lindskog mechanisms. Steps shown include the enzyme-substrate complex (ES), transition state (TS) and intermediate (INT) structures. Labels correspond to coordinates along for the reactions where the active site metal (M) is Mn^{2+} or Ni^{2+}.

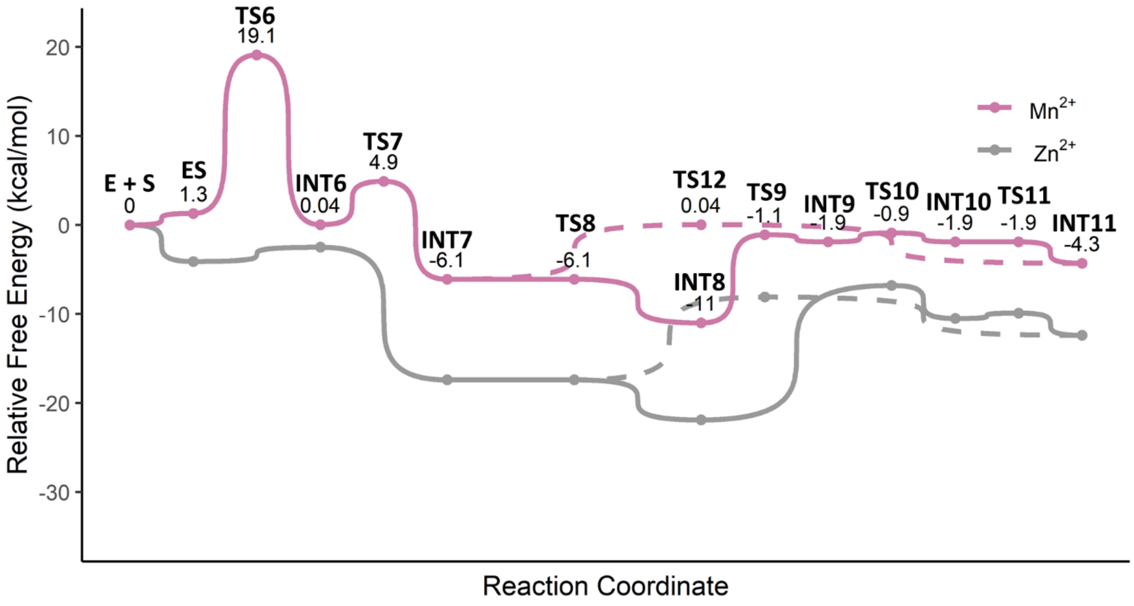


Figure 18. Free energy profile for the modified Lipscomb (solid) and Lindskog (dashed) reaction mechanisms for HCAII with Mn^{2+} as the active site metal compared to Zn^{2+} (grey).

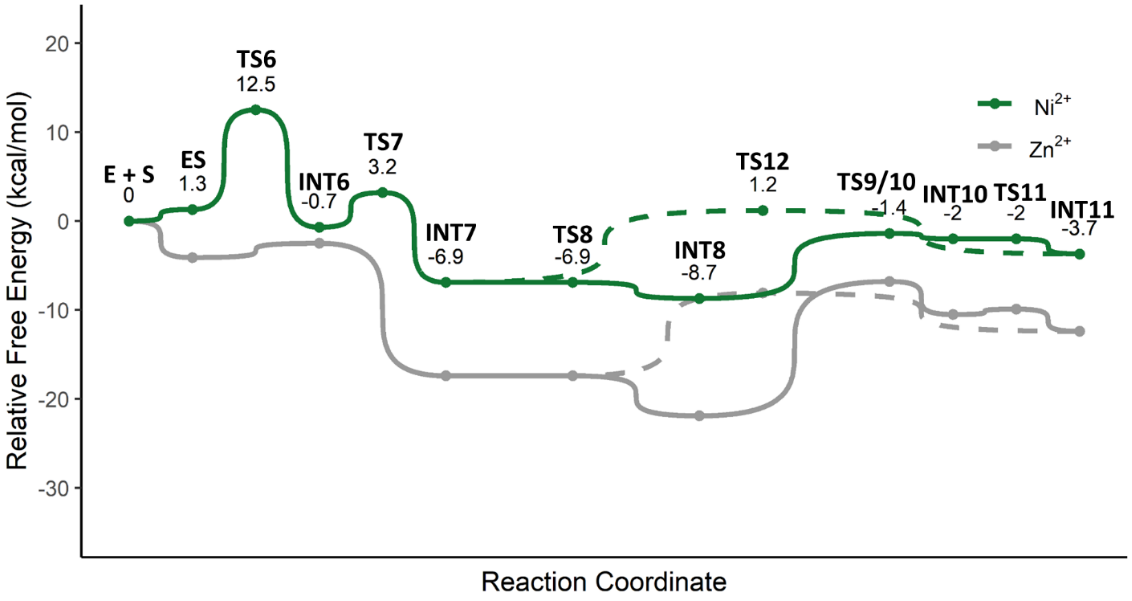


Figure 19. Free energy profile for the modified Lipscomb (solid) and Lindskog (dashed) reaction mechanisms for HCAII with Ni^{2+} as the active site metal compared to Zn^{2+} (grey).

commence.[107,108] CO_2-pressurized, cryo-cooled crystals of Ni^{2+}-HCAII reveal there will be steric hindrance between the CO_2 and metal-bound waters as CO_2 enters the active site pocket, distorting final orientation of CO_2 within the cavity compared to within Co^{2+}- and Zn^{2+}-substituted active sites. Although our water-sparse QM-model is not designed to capture the true magnitude of these steric effects when CO_2 enters this active site, it should be noted that the enzyme-substrate complex (ES) is computed to be less stable within the Mn^{2+} and Ni^{2+}-model than the separate enzyme and substrate species (E + S), which is the opposite of what is observed with the Zn^{2+}- and Co^{2+}-substituted models. The evidence from the QM-models support the theory[107] that the reduced enzymatic activity of Mn^{2+}- and Ni^{2+}-substituted HCAII is attributable to the non-tetrahedral metal coordination geometries hindering the ease for CO2 to orient within the active site and efficiently initiate the reaction.

The Cd^{2+}-substituted HCAII was similarly reported as having poor enzymatic activity (~2% compared to Zn^{2+-}-HCAII). The energy profile for the reaction (Figure 20) was computed using the tetrahedral metal starting geometry, and intermediates and transition state structures similar to those along the Zn^{2+}-HCAII reaction pathways (Figure 13) were identified. Although a significantly reduced enzymatic activity is reported experimentally, the energy profile indicates the reaction catalyzed by Cd^{2+}-HCAII is slightly more favorable thermodynamically and kinetically. The comparability between the computed Zn^{2+}- and Cd^{2+}-HCAII reaction thermodynamics is similar to a previous QM-cluster model study.[91]

This discrepancy between our model and experiment may be due to the fact that the reaction pathways computed in this and the aforementioned QM-cluster modeling

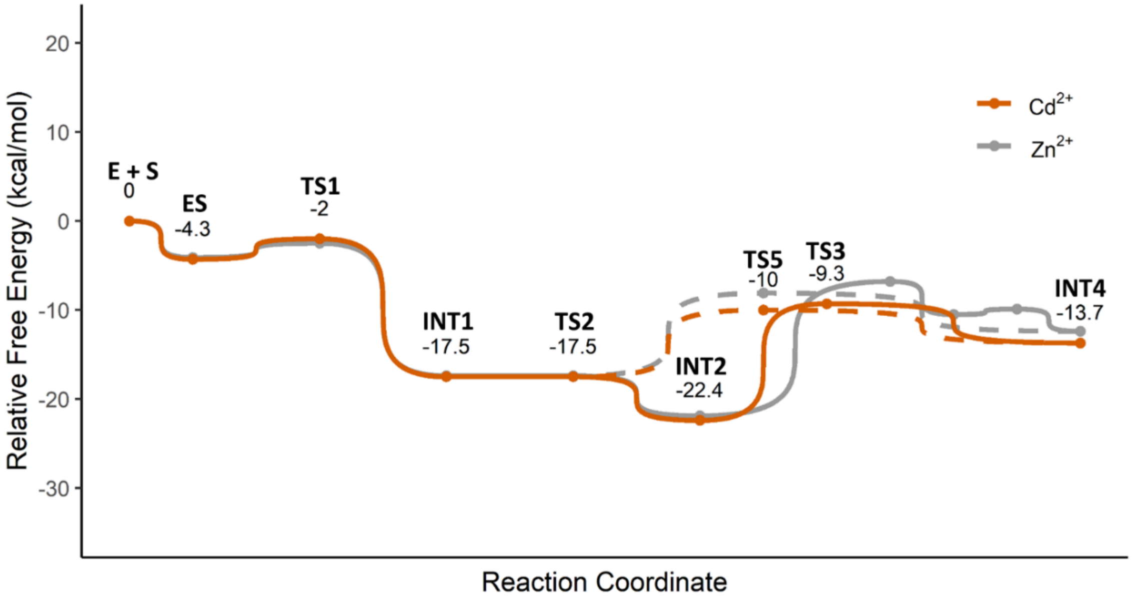


Figure 20. Free energy profile for the Lipscomb (solid) and Lindskog (dashed) reaction mechanisms for HCAII with Cd^{2+} as the active site metal compared to Zn^{2+} (grey).

studies focus on the CO_2 hydration mechanism and begin with the metal-hydroxide structure (ES) without simulating the steps required to generate this starting structure. Water deprotonation occurs via a histidine-directed water shuttle which drives the proton from within the protein pocket into the bulk solvent,[81,109–113] a mechanism not able to be properly simulated in this current QM-cluster model. Given that Cd^{2+} is a weaker Lewis acid than Zn^{2+}, it may be that the predominant form of the ligand bound to Cd^{2+} is water rather than hydroxide.[91] This is supported by experimental evidence that the activity of Cd^{2+}-HCAII is induced at higher pH with an activity profile corresponding to the ionization of a Cd^{2+}-bound water molecule.[114] Additional modeling beyond the scope of this current work would need to be conducted to examine this hypothesis.

Proposed Mechanism for Fe-HCAII

Low enzymatic activity (~4% compared to Zn^{2+}-HCAII) was similarly reported for Fe-HCAII. The Fe^{2+}-HCAII model was simulated at both low and high-spin configurations, with previous experiments on γ-class CA reporting a high-spin state for its Fe^{2+}-bound CA.[87] Structures for the low-spin Fe^{2+}-HCAII were computed using a model with the additional metal-bound water, and the resulting mechanism (Figure 21) and free energy profile (Figure 22) are shown to be similar to the Mn^{2+} and Ni^{2+}-HCAII pathways. The reaction is shown to be slightly more thermodynamically and kinetically favorable with low-spin Fe^{2+} compared to Mn^{2+} and Ni^{2+}, but there is still a substantial activation energy required for reaction initiation with this coordination geometry.

Structures for the high-spin Fe^{2+}-HCAII were computed with mechanisms (Figure 17) and a free energy profile similar to the Zn^{2+}-HCAII pathway (Figure 23). The intermediates are shown to be more thermodynamically stable, and the reaction is slightly thermodynamically favorable compared to Zn^{2+}-HCAII. Fe^{2+}-HCAII also has a reduced activation energy required for the Lipscomb pathway (TS3; $\Delta\Delta G_{Zn} = 15.1$, $\Delta\Delta G_{Fe} = 11.7$), and slightly greater activation energy required for the Lindskog pathway (TS3; $\Delta\Delta G_{Zn} = 9.3$, $\Delta\Delta G_{Fe} = 10.6$). Both metal active sites kinetically favor the Lindskog pathway over the Lipscomb pathway, although the advantage is reduced in Fe^{2+}-HCAII, making the pathways more competitive. Based upon these results, the hydration of CO_2 is predicted to be thermodynamically and kinetically feasible for Fe^{2+}-HCAII.

The models were also used to simulate the CO_2 hydration mechanism for Fe^{3+}-HCAII. The reaction mechanism (Figure 13) was computed for the low-spin state Fe^{3+} (Figure 24), and it is shown to be similar to the high-spin Fe^{2+}-HCAII with the exception of the very thermostable Lipscomb intermediate INT2. This increases the activation energy for the Lipscomb reaction pathway to 21.1 kcal/mol; the Lindskog pathway remains kinetically favored with a $\Delta\Delta G = 11.5$ kcal/mol. These results support the

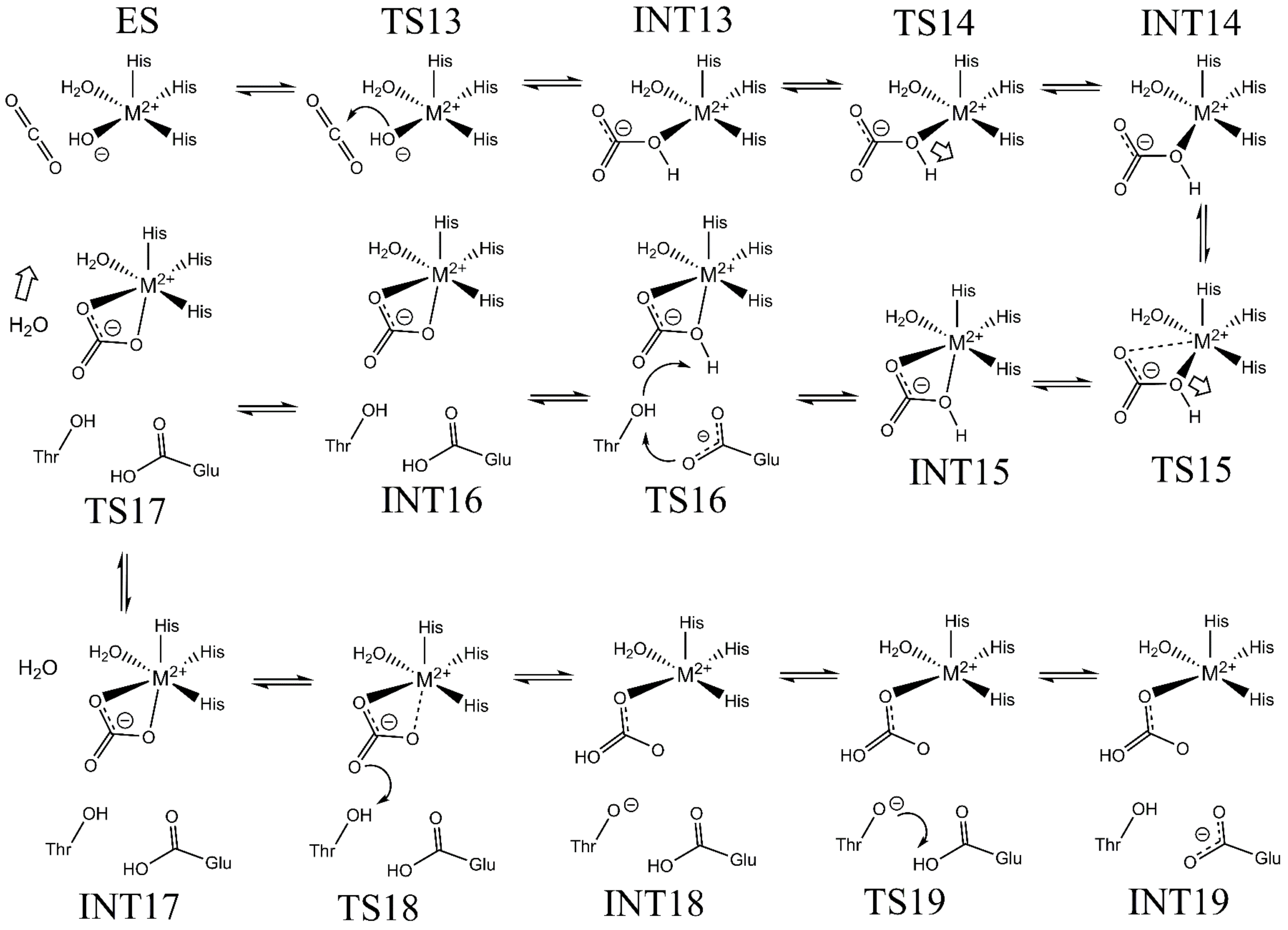


Figure 21. Part of the proposed carbonic anhydrase reaction pathway illustrating the Lipscomb mechanism where active site metal (M) is low-spin Fe^{2+}. Steps shown include the enzyme-substrate complex (ES), transition state (TS) and intermediate (INT) structures.

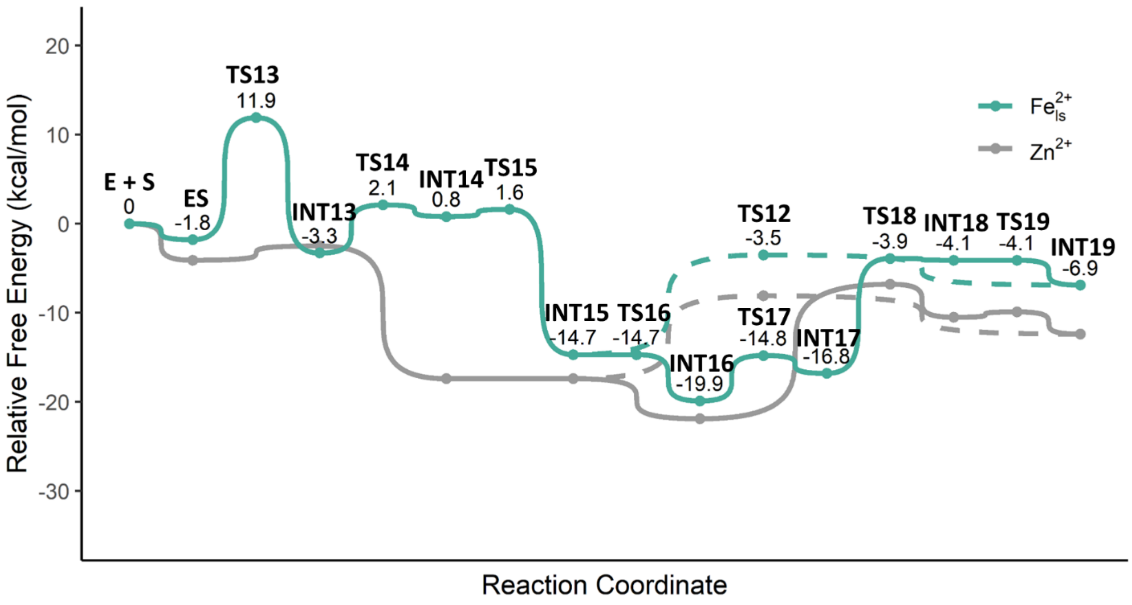


Figure 22. Free energy profile for the modified Lipscomb (solid) and Lindskog (dashed) reaction mechanisms for HCAII with low-spin Fe^{2+} as the active site metal compared to Zn^{2+} (grey).

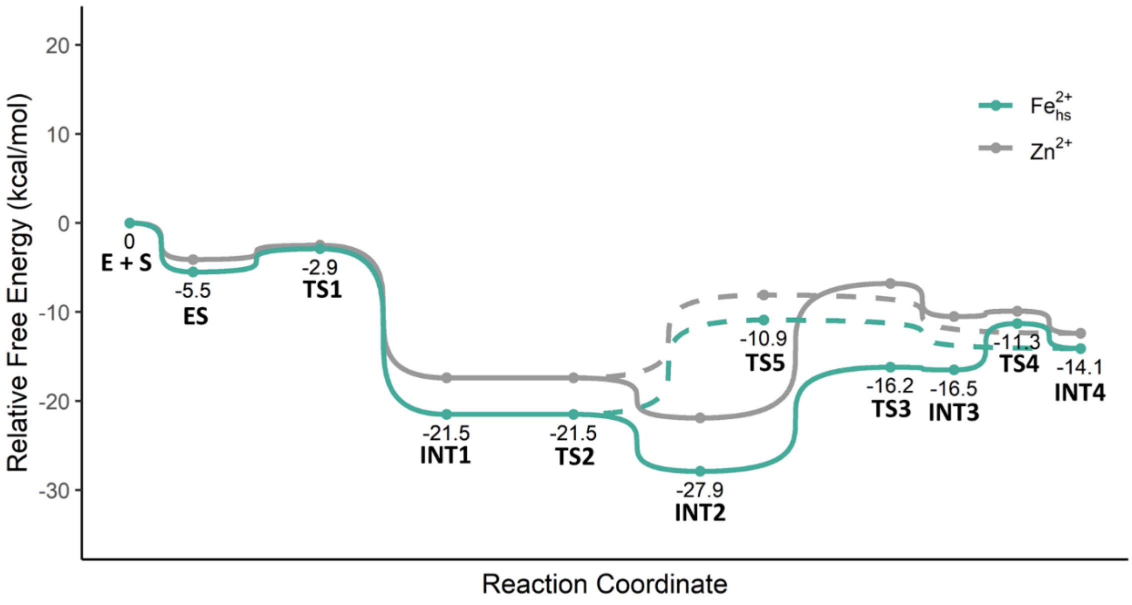


Figure 23. Free energy profile for the Lipscomb (solid) and Lindskog (dashed) reaction mechanisms for HCAII with high-spin Fe^{2+} as the active site metal compared to Zn^{2+} (grey).

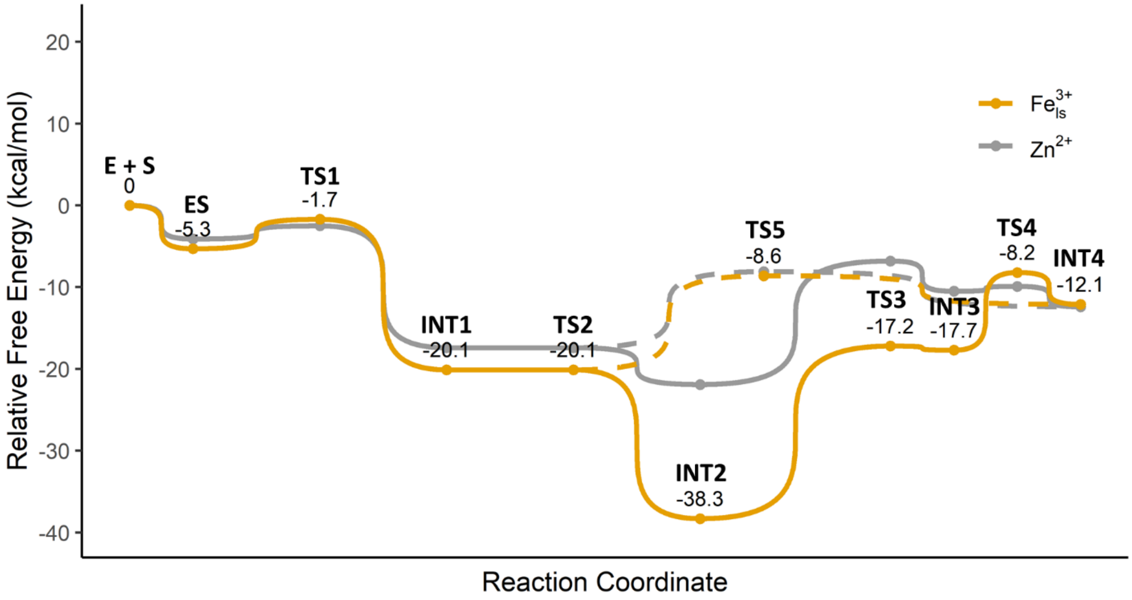


Figure 24. Free energy profile for the Lipscomb (solid) and Lindskog (dashed) reaction mechanisms for HCAII with low-spin Fe^{3+} as the active site metal compared to Zn^{2+} (grey).

feasibility of Fe^{3+}-HCAII to catalyze the CO_2 hydration reaction. However, this assumes the starting metal-bound hydroxide structure is readily formed, and experiments report the poor metal binding affinity of HCAII for Fe^{3+}.[86,87]

Conclusions

In this work, the viability of multiple transition metal-substituted HCAII enzymes were investigated by quantum mechanically modeling the Lipscomb and Lindskog reaction mechanisms. Using Zn^{2+}-HCAII as a point of reference, the models demonstrated Co^{2+}-substituted HCAII is catalytically feasible but is limited by the energy needed to break the Co-O bond of a tightly bound, bidentate carbonate intermediate structure. Additional computations are needed to examine recent experimental evidence finding an additional water bound to the intermediate metal structure, a possibility not accounted for in the current model but easily rectified by adding additional waters to the active site model. Models of the Mn^{2+} and Ni^{2+}-substituted active site reveal that square pyramidal coordination substantially increases the activation energy required for the hydroxide to bind to CO_2 and initiate the reaction. Models of Cd^{2+}-HCAII suggest the reaction is readily catalyzed, in agreement with previous modeling studies but in disagreement with the poor experimentally reported activity levels. It is hypothesized this discrepancy arises from Cd^{2+}-HCAII not readily deprotonating the metal-bound water to form the metal-bound hydroxide, a mechanistic step not modeled in this work.

Lastly, the mechanisms for Fe^{2+}-HCAII in low and high metal spin states were examined along with Fe^{3+}-HCAII in the low spin state. The high spin state Fe^{2+} is expected to be the predominant form, and the reaction pathways computed are thermodynamically more favored than Zn^{2+}-HCAII along with being kinetically comparable. These results suggest that, in an anaerobic environment where the Fe^{2+} is not able to be oxidized to Fe^{3+}, the hydration of CO_2 by Fe^{2+}-HCAII is theoretically feasible. This feasibility is notably limited to the steps of the reaction cycle modeled, as steps from the overall catalytic cycle not addressed by these models (e.g. generation of the metal-hydroxide or dissociation of the product) may inhibit the drive of the reaction. Nevertheless, these results give hope in the ability to synthesize an active, anaerobic Fe^{2+}-HCAII. When the Fe^{2+} is oxidized to Fe^{3+}, the models suggest the reaction is still catalytically feasible, which is supported by the poor enzyme activity reported for Fe-HCAII.

In conjunction with this theoretical work, experiments involving the synthesis of Fe^{2+}-HCAII within an anaerobic environment and measurement of its activity and metal spin state are currently being carried out by the lab of Dr. Joseph Emerson at Mississippi State University.

Chapter 4: Catechol-O-methyltransferase

Introduction

With the advancement of computers, the modeling and simulation of enzymes have become invaluable tools for insight into atomic-scale protein properties. Enzyme simulations typically apply quantum mechanics (QM), molecular mechanics (MM), or a hybrid of the theories depending on whether the question of interest requires simulating the entire protein or only the enzyme active site.[1,2,4–6] When creating QM-cluster model simulations of an enzyme active site, it is crucial to include any residues, solvent, ions, and coenzymes sterically and/or electrostatically crafting the active site microenvironment to ensure the results reflect reality, while also excluding less important residues to ensure computational feasibility and efficiency.[4–6,9] While this is relatively simple, much remains to be done to establish a rational, computationally inexpensive protocol for identifying these chemically important residues.

Ideally, there would be a computationally inexpensive, *a priori* approach to enzyme model construction that utilizes structural and chemical data to rationally select residues (or parts of residues) for QM-cluster modeling. As a potential solution for this model creation problem, our lab has been developing the software Residue Interaction Network ResidUe Selector (*RINRUS*) which computes a contact-based residue interaction network[25,26] and uses the data to identify and rank residues for subsequent modeling. Further, *RINRUS* automatically trims and caps the residues via a rules-based criterion to form appropriate models and generates formatted input files for several popular electronic structure theory packages (see Methods and Appendix C for details). The success of incorporating interaction and rules-based rationale into model design has been reported for QM-only models[115] and recently implemented into a QM/MM modeling API;[116] however, there continues to be no definitive protocol for generalized QM-cluster enzyme model creation. Through establishing an automated and rigorous workflow, we envision solutions to several community-wide problems including standardization of enzyme QM-model creation, reducing learning curves for new users, and minimizing trial and error using poorly or incorrectly designed models. Implementing the *RINRUS* toolkit may also improve reproducibility of workflows and published results, a scientific community-wide need which was most recently emphasized by the 2019 consensus study report *Reproducibility and Replicability in Science* released by The National Academies of Sciences, Engineering, and Medicine.[117] To informally highlight the reproducibility problem within the QM/MM and QM-cluster modeling communities, we surveyed 58 QM/MM or QM-cluster model papers published between Jan 1 – Mar 31 of 2015 and Jan 1 – Mar 31 2019, evaluating whether the models could be directly reproduced via reporting of Cartesian coordinates (see Appendix C for details). Only 20 papers (34%) reported Cartesian coordinates to the extent that reproduction is possible. Given the absence of consistent community reporting, embedding reproducibility via a systematic model design workflow would be a large step towards research standards in computational enzymology.

Ideally, the *RINRUS* workflow would be capable of identifying a singular model or handful of models that best capture the balance between maximizing the number of key residues included to simulate the active site while minimizing the size of the QM-region for computational efficiency. This leads to questions such as what makes the enzyme model “good”? What easily obtainable metrics might be universal in computational biochemistry for ranking the importance of interatomic/inter-residue interactions? We begin to answer these questions within the context of contact-based residue interaction networks.[25,26]

The protein of interest for this case study is catechol-O-methyltransferase (COMT), a target enzyme of numerous QM-cluster and QM/MM studies.[9,18,124–133,21,22,118–123] The mechanism catalyzed by COMT is rather simple, involving only an S_N2 methyl transfer from *S*-adenosylmethionine (SAM) coenzyme to the oxygen of a Mg^{2+}-bound catecholate substrate (CAT, Figure 25A). Kinetic experiments on human COMT provide a free energy of activation ($\Delta G^{‡}$) of 18 - 19 kcal/mol at 310 K[134,135] and computational studies report the methyl transfer reaction to be exergonic.[9,122,123,131]

Previous computational studies have shown substantial variation in both $\Delta G^{‡}$ and free energies of reaction (ΔG_{rxn}) with respect to QM-cluster or QM/MM model size. Recent results from QM/MM calibration studies using radial distance-based QM-regions suggest that asymptotic convergence of thermodynamics/kinetics requires radial QM-regions of 400 - 600 atoms.[9,18,122] Unfortunately, conventional DFT calculations of 400 - 600 atom models are prohibitively expensive for many research groups. The large QM-region size required to study the COMT mechanism also defies conventional wisdom that kinetic/thermodynamic properties should converge quickly as the size of the QM-region grows in a QM/MM partition. Slow convergence behavior of COMT has been attributed to the non-spherical active site, requiring an accurate description of both the Mg^{2+}/catechol coordination chemistry and the electrostatic stabilization of the large SAM cofactor.[122]

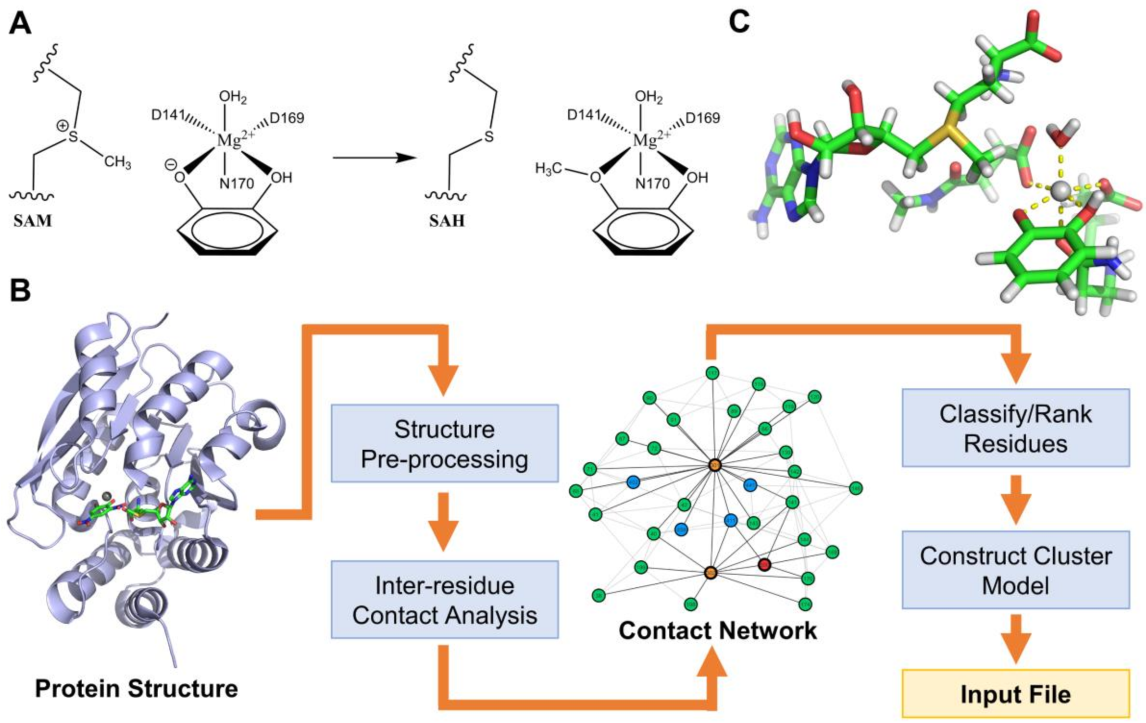


Figure 25. (A) COMT catalyzes the methyl-transfer reaction from *S*-adenosylmethionine (SAM) to the oxygen of a Mg^{2+}-bound catecholate substrate, forming *S*-adenosylhomocysteine (SAH) and guaiacol. (B) The *RINRUS* workflow begins by processing a protein structure (X-ray, NMR, or computational simulation in PDB file format) before computing inter-residue contacts to form a contact network. Residues (green) and solvent (blue) interacting with the species of interest (the "seed", orange and red) are identified. Systematic classification or ranking schemes are used to construct appropriate cluster models. *RINRUS* then writes these models into an input file format appropriate for simulation in a variety of quantum chemistry software packages. (C) The base model from which all COMT models were built-up. It is composed of the seed (SAM, CAT, Mg^{2+}), three residues, and one coordinating water completing the coordination of Mg^{2+} (D141, D169, N170, HOH411).

While the paradigm of calibrating expanding QM-regions in a radial distance-based fashion has been established to provide poor convergence for COMT, there is a surprising dearth of explored alternatives to distance-based active site models in the literature. In this work, we present the reaction thermodynamics and free energies of activation for hundreds of QM-cluster models of COMT constructed by *RINRUS* using several possible workflows. By tracing the final results back to how the models were constructed, we seek to identify a construction protocol that consistently constructs accurate and efficient QM-cluster models of COMT. Though this work will only involve one case study, the findings from surveying an immense range of models of the same enzyme will allow future studies to invert the focus towards assessing the benefits of a particular approach on enzymes with more diverse structure and function. This cheminformatics perspective will be a rigorous step towards establishing a translatable, generalized computational enzymology protocol.

Methods

The various structures and functions of proteins arise in part from the noncovalent interaction networks of their amino acid subunits. To highlight these networks, the complex three-dimensional structure of proteins may be simplified into a two-dimensional adjacency matrix or a graph mapping the residues to points (nodes) interconnected by lines (edges). Conventionally, each node represents an individual amino acid of the protein, and each edge represents a noncovalent interaction occurring between two amino acids. For more information on inter-residue contact networks and their design, properties, and applications within chemistry, the reader is directed to reviews by Giuliani[25] and Shen.[136]

In this work, the construction of inter-residue contact networks begins by following a procedure similar to that of the software *RINerator*.[26] First, hydrogens are added to the protein crystal structure (PDB ID: 3BWM) using the program *Reduce*.[59,137] As the 3BWM crystal structure has the inhibitor 3,5-dinitrocatechol coordinated to the active site metal, the two nitro-groups were replaced with hydrogens to form the catechol substrate. An additional hydrogen was also added to the 2-amino functional group of the S-adenosyl methionine substrate to bring it to a +1 charge, its expected protonation state. This modified crystal structure is the structure used for all subsequent network generation and model construction. The program *Probe*[27] is then used to identify non-covalent interactions throughout this structure. The program does this by rolling a small (0.25 Å radius) spherical probe over the van der Waals surface of the atoms and identifying both where the probe comes in contact with other non-covalently bound atoms and where van der Waals surfaces are clashing. The *Probe* output file details the contact/overlap "dots" for all of the atoms reflecting the distance of contacts or volume of overlaps. Wide contacts have an interatomic gap distance ≥0.25 Å; close contacts have an interatomic gap distance <0.25 Å; big overlaps have overlapping van der Waals radii ≥0.4 Å; small overlaps have overlapping van der Waals radii <0.4 Å; and hydrogen bonding are overlapping van der Waals radii between donor hydrogen and electronegative acceptor atoms.[27] All of the reported contact dots (places where an interatomic contact/overlap occurs) are then collated for each residue to indicate which residues are interacting. The network illustrating all Probe-predicted contact interactions within 3BWM is shown in Appendix C: Figure 1.

The chemically reactive species for this enzyme include the two substrates SAM and CAT along with the Mg^{2+} that CAT binds. One rationale for building-up models of the active site would be to first focus on including residues immediately interacting with these reactive species. The network indicates this list includes 27 amino acids and 4 crystallographic waters. The specific parts of the residues having contact interactions with the reactive species (main chain or side chain) and the number of each contact type is provided in Appendix C: Table 1.

The base for building-up all models described in this work is composed of the substrates SAM and CAT, Mg^{2+}, and the four species completing the coordination of Mg^{2+} (D141, D169, N170, HOH411; Figure 25). Residues are added to this base model by either assigning each residue an ordered rank or by adding groups of residues classified by a common feature. Models were automatically generated using the *RINRUS* software, trimming the models based upon a residue amino, carboxyl, or side chain having interatomic contacts with the seed. Places where covalent bonds are broken in trimming the model have hydrogens added to satisfy valency via the program *PyMol* v2.3.a0.[138] To maintain the general shape and semi-rigid character of the protein tertiary structure, all C_α atoms, along with the C_β atoms of Arg, Lys, Glu, Gln, Met, Trp, Tyr, and Phe side chains, were frozen to their crystallographic positions. Further details about residue selection and model trimming are provided in Appendix C. Though other research groups who employ QM-cluster models may have developed internal research protocols for trimming residues/fragments and freezing backbone atoms, we intend RINRUS to be the first enzyme model design toolkit to publicly codify these reproducible workflows (and also encourage hypothesis-driven testing of variations to our model building decision trees).

All QM computations were performed using the Gaussian16 software package.[62] The models were geometrically optimized using density functional theory (DFT) with the hybrid B3LYP exchange-correlation functional.[63,64] The computations used the 6-31G(d') basis set for N, O, and S;[66] the 6-31G basis set for C and H atoms;[67] and the LANL2DZ effective core potential and basis set combination for Mg.[104] The Grimme D3 (Becke-Johnson) dispersion correction (GD3BJ) was also included[105] along with a conductor-like polarizable continuum model (CPCM) using UAKS sets of atomic radii, a nondefault electronic scaling factor of 1.2, and a dielectric constant of $\varepsilon = 4$.[68,69] Unscaled harmonic vibrational frequency calculations were used to confirm all stationary points as either minima or transition states. Stationary points were found by first pre-optimizing the model to the reactant structure. This pre-optimized structure was then used to construct an approximate transition state structure by translating the methyl midway between the sulfur of SAM and the oxygen of CAT and flattening the methyl to a planar configuration. The transition state was optimized, and intrinsic reaction coordinate computations were used to confirm the formal reactant and product minima and calculate reaction free energies. Whether this procedure biases the simulated active site to more strongly stabilize the reactant structure (and whether such a bias would be of any significance) is unknown and an uninvestigated facet of computational enzymology.

The *k*-means clustering analysis[139] was run through *RStudio* v.3.6.3[140] using seed 3163 for replication purposes. Elbow and gap statistics (Appendix C: Figure 6) were run using the *factoextra* package.[141] For the gap statistic, the number of “bootstrap” Monte

Carlo samples used was 50. Both elbow and gap statistics suggest using a *k* near *k* = 6 for the cluster analysis (Appendix C: Figure 6). A *k* = 6 was ultimately used for further analysis as the clusters with *k* = 6 are reasonably partitioned into distinct groupings where the range of free energies predicted by models within a cluster are not too broad (would happen with small k-clusters) and the interpretation of the clusters are not so narrow as to fail to be generalizable (would happen with large k-clusters). To identify the appropriate clusters, the Hartigan and Wong *k*-means clustering algorithm was used starting from a total of 50 different random starts.[142]

Results and Discussion

We began by computing a contact-based residue interaction network (Figure 25B) for an X-ray crystal structure of human COMT (Protein Data Bank ID 3BWM), where residues, substrates, and solvent are illustrated as circles (termed "nodes" in standard graph theory nomenclature) interconnected by lines (termed "edges") when there are interatomic contacts between two residues/fragments. Though the construction and analysis of these graphs are already known to provide insight into allosteric regulation, protein folding and stability, and structure-function relationships,[25,136] we repurpose the networks towards QM-cluster model design. The network indicated 27 protein residues and 4 crystallographic waters had contact interactions with any fragments central to the catalytic reaction (termed the "seed": SAM, CAT or Mg^{2+}). The residue contacts with the seed were classified into five different types: wide contacts, close contacts, small overlaps, big overlaps, and hydrogen bonding. All QM-cluster models of COMT were constructed using the crystallographic coordinates of these residues and, unless otherwise indicated, trimmed according to the *RINRUS* workflow (refer to Appendix C). Models were expanded from the seed by one of two general ways: residues were incrementally added based upon a ranking criterion (e.g., distance from the seed, number of contacts with the seed) or groups of residues were added to the seed based upon similar residue features (e.g. type of interatomic contacts). The models constructed solely from the *RINRUS* contact information expand to a 485-atom model representing a “first-interaction shell” maximal model that includes all residues with quantified contacts with any of the seed fragments. This maximal model is ellipsoidal in shape, reflective of the non-spherical geometry of the COMT active site. Further details on the model building schemes beyond what will be outlined in the discussion are provided in Appendix C. In total, the methyl transfer transition state and connecting reactants/products for 550 unique QM-cluster models were computed. A total of 1650 DFT-optimized stationary points were analyzed in this work.

Expansion of QM-cluster Models by Ranking of Residues

We will first detail several ways COMT QM-cluster models were incrementally built-up by ranking residues. The first metric is the current paradigm of ranking residues based on their distance to the active site. Though a simple distance metric may seem straightforward, this method can be ambiguous and tricky to replicate without reporting very precise definitions of the radial origin and the thresholds for adding residue fragments or entire residues. Subtle variances in definitions might qualitatively affect which residues or atoms are captured within varying radially expanding models. For this work, 25 models were constructed with *RINRUS* by incrementally adding residues ranked by the shortest distance from the position of any atom (including hydrogens) of the seed to the position of any atom of the surrounding residues. The models were expanded until all residues predicted by the contact network were incorporated, encompassing a 3.10 Å expansion from any atom of the seed. Two residues (K46 and N92) with no *RINRUS*-predicted contact interactions with the seed but fall within the 3.10 Å distance threshold were necessarily included in these distance-based models.

Computed values of $\Delta G^{\ddagger}$ and ΔG_{rxn} are plotted against the distance-based expansion from the seed (Figure 26A). As the size of the model increases, the predicted $\Delta G^{\ddagger}$ converges (the $\Delta G^{\ddagger}$ is within ±2 kcal/mol of the largest distance-based model) with QM-cluster models containing >342 atoms, but the predicted ΔG_{rxn} does not similarly converge even with the largest distance-based models. Some of the largest distance-based models computed in this work (containing 444 and 447 atoms) incorrectly predict an endergonic reaction.

The surprising appearance of qualitatively incorrect reaction free energies in the largest distance-based models brings up some crucial pitfalls in designing QM-cluster models, but also ways that *RINRUS* can be used by the QM-cluster modeling community to circumvent these pitfalls. The convergence of the reaction free energy is disrupted by addition of the charged residue K46, which as previously noted, does not have direct contact interactions with the seed. Such a qualitative shift in thermodynamic properties contradicts intuition that a larger QM-cluster model will always be "better" than a smaller model. At best, the addition of peripheral residues with no quantifiable interaction with seed residues/fragments adds unnecessary time to the DFT simulations, as observed with the addition of the uncharged N92 residue (not present in RINRUS-constructed models) changing ΔG^7 and ΔG_{rxn} by < 0.2 kcal/mol in the 486-atom

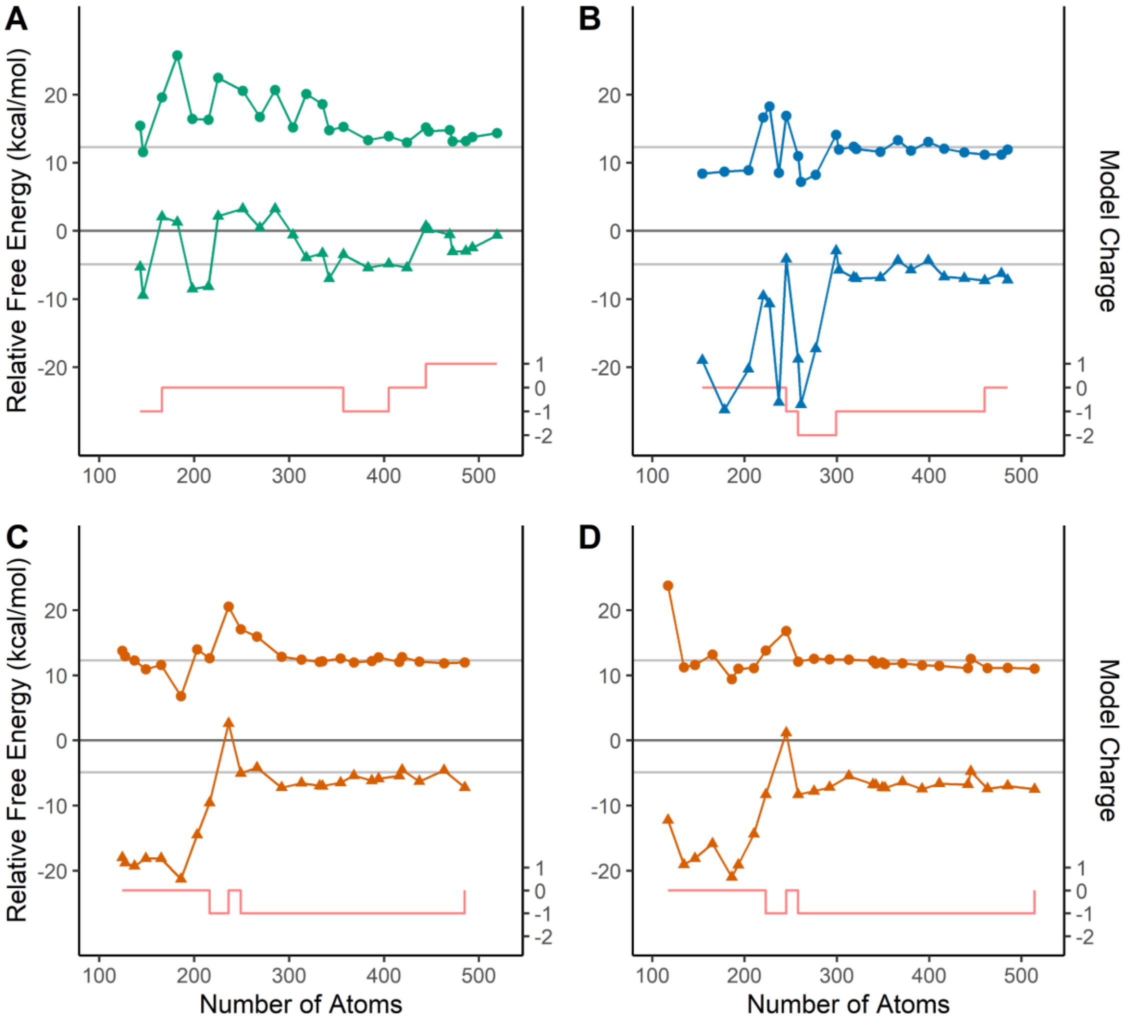


Figure 26. Computed methyl transfer $\Delta G^{‡}$ (circle) and ΔG_{rxn} (triangle) free energies as models are systematically built-up through different methods of ranking residues including distance from the seed (A), total number of contacts with the seed (B), frequency of residue in Combinatoric Scheme 2 sets (C), and a by-hand reconstruction of models by frequency of residue in Combinatoric Scheme 2 sets (D). Grey lines indicate the reference convergence values.

scheme does not address the enzyme active site chemistry in a physically meaningful way. It may be fortuitous that the maximal COMT model generated by *RINRUS* does not include any boundary residues that are part of an unrequited charged pair. If the maximal model is thought of as a "first interaction shell" that encapsulates all residues that influence the active site chemistry, regardless of distance from the seed fragments, then the *RINRUS* source code can be easily adapted to include residues in the "second shell" that are necessary for charge balancing of larger-sized models.

As a step towards identifying a chemically-directed way to expand models, we next considered the convergence of QM-cluster models constructed by ranking based on the number of contacts each residue has with the seed and incrementally building models from residues with the most contacts to fewest contacts with the seed. We define "convergence" in this study as being within ±2 kcal/mol of the convergence reference values and remaining so as the model size is increased one residue at a time. The convergence reference values are defined as average relative free energies of the five largest models designed solely using *RINRUS* contact interactions: 12.3 kcal/mol for $\Delta G^{\ddagger}$ and −4.9 kcal/mol for ΔG_{rxn}. The converged reference value for $\Delta G^{\ddagger}$ is lower than the experimentally derived value but this is expected considering the marginal level of theory used in this case study. The accuracy of *RINRUS*-derived models will be a subject of several future studies in our groups, by varying level of theory, treatment of solvation, and approaches for freezing atoms. With an improved ranking scheme using number of residue-seed contacts, $\Delta G^{\ddagger}$ and ΔG_{rxn} both converge by the 302-atom model (Figure 26B). While an interaction-based ranking fares better at prioritizing residues than distance-based expansion, there are some inherent limitations. Namely, larger residues with more surface area (e.g., lysine or tryptophan) are more likely to have more contacts with the seed and may bias the ranking compared to smaller residues. Ranking by number of contacts with the seed also does not weight or quantify the magnitude of electrostatic influences (e.g., charge, hydrogen bonding, and polarity). Nevertheless, even with this nonoptimal metric, constructing models by contact count still yields impressively small, converged models.

Below, we will detail two combinatoric workflows for building models where residues are classified into sets by common contact type. The third method for ranking residues involves ordering residues by the number of times each residue appears in a unique model from the Combinatoric Scheme 2 model sets (see below and Appendix C for details). This ranking inherently favors residues with more than one type of contact interaction. In using this residue ordering, $\Delta G^{‡}$ and ΔG_{rxn} are converged when QM-cluster model size is greater than ~300 atoms (Figure 26C), similar to the models designed through ranking residues by total contacts with the seed. The model with the greatest overestimation of $\Delta G^{‡}$ and endergonic ΔG_{rxn} (236 atoms) corresponds to the addition of the positively charged residue, K144. The subsequent inclusion of the negatively charged E199 residue places the predicted free energies within qualitative accuracy, re-emphasizing the point that particular care in model design must be given towards charged residues and nearby residues that counter their effective charges.

Automation Versus Constructing QM-cluster Models Manually

The *RINRUS* package is still undergoing rapid development and needs further testing to address broader QM-cluster model design issues such as residue/substrate protonation states, orientation of explicit solvent molecules, and conformational sampling.[5,6] While these factors may be manually addressed by the user, doing so places a potential bottleneck in the throughput of QM-cluster model applications.

In consideration of possible differences between manual and automated model building, models built by ranking residues via their frequency of appearance in Combinatoric Scheme 2 models (Figure 26C) were reconstructed by-hand by the PI. The models were designed without any guidance from *RINRUS* beyond the identity of the specific residues in contact with the seed and their ranked order. The results of these “bespoke” models are presented in Figure 26D and are shown to be comparable to the models built by RINRUS (Figure 26C). There is reduced fluctuation in the $\Delta G^{\ddagger}$ for the smaller bespoke models versus comparably-sized *RINRUS*-generated models, likely attributable to manual sampling of residue orientations, a treatment not done for any of the *RINRUS*-derived models. However, for the models greater than 300 atoms, there is no qualitative difference between the automated and the "by-hand” approach. These results demonstrate how *RINRUS*, even without carefully attending to residue protonation and conformational sampling, can construct QM-cluster models in a way similar to that by an experienced scientist, but which is founded on a traceable cheminformatic basis and a reproducible, rational workflow.

Expansion of QM-cluster Models by Residue Interaction Features

The remaining models were built up from the seed by combining residues with common features, specifically by inter-residue contact type. The contact types contain two pieces of information used in QM-cluster model construction: the section of the residue contacting the seed (classified as either residue main chain, residue side chain, or explicit water molecule) and the contact type (wide contact, close contact, small overlap, big overlap, hydrogen bonding). Models were constructed by taking all combinations of the contact types and, for each combination, building a QM-cluster model using all residues with the specific contact types of that combination. These models represent a combinatoric approach to building-up models by adding groups of residues by common features to the seed (Combinatoric Scheme 1, see Appendix C for details). To further increase the number of models and dataset size, the sets of residues classified by contact types were repartitioned into a second combinatoric approach (Combinatoric Scheme 2, see Appendix C for details), though the generation of these sets is not rigorous or necessarily applicable to other biosystems. Given the limitations of time and resources, 114 (of 204 possible) models of Combinatoric Scheme 1 and 357 (of 736 possible) models of Combinatoric Scheme 2 have been simulated, representing all unique combination-based models up to at least 320 atoms (Appendix C: Figure 5). As the goal is identifying small, yet accurate, QM-cluster models, the cost of expanding the dataset to include hundreds of additional large models is not expected to lead to substantial improvements in analysis.

In plotting $\Delta G^{‡}$ and ΔG_{rxn} of QM-cluster models built through the two combinatoric schemes (Figure 27A and B), a wide range of computed kinetic and thermodynamic values were exhibited. Variation in $\Delta G^{‡}$ and ΔG_{rxn} originates from differences in model composition rather than models optimizing into unnatural orientations, since the root mean square deviation (RMSD) of unconstrained residue heavy atoms of the geometry optimized reactant state compared to the X-ray crystal structure is on average only 0.53 Å for all models (Appendix C: Figure 4; standard deviation, 0.17 Å). Similar to the models built by ranking residues, there are models with

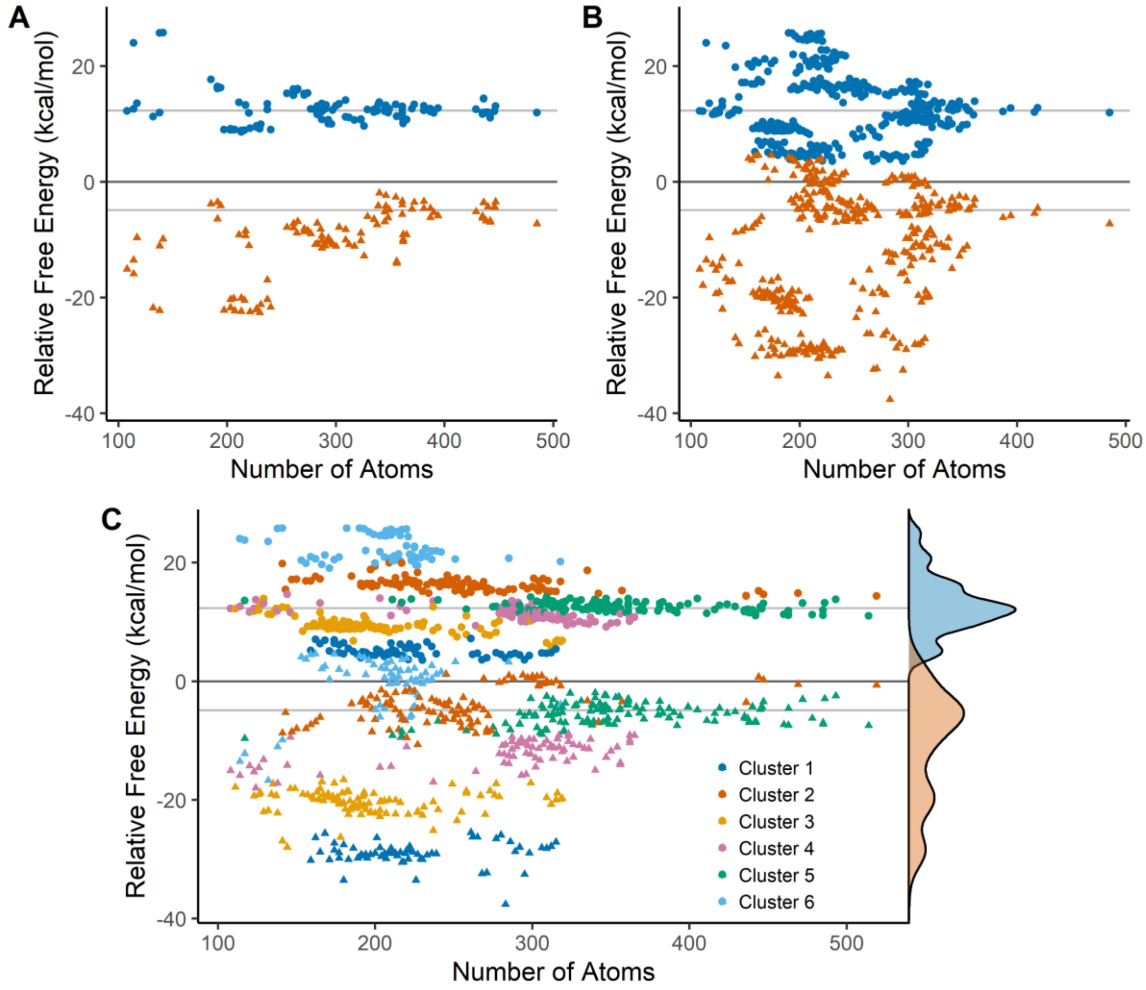


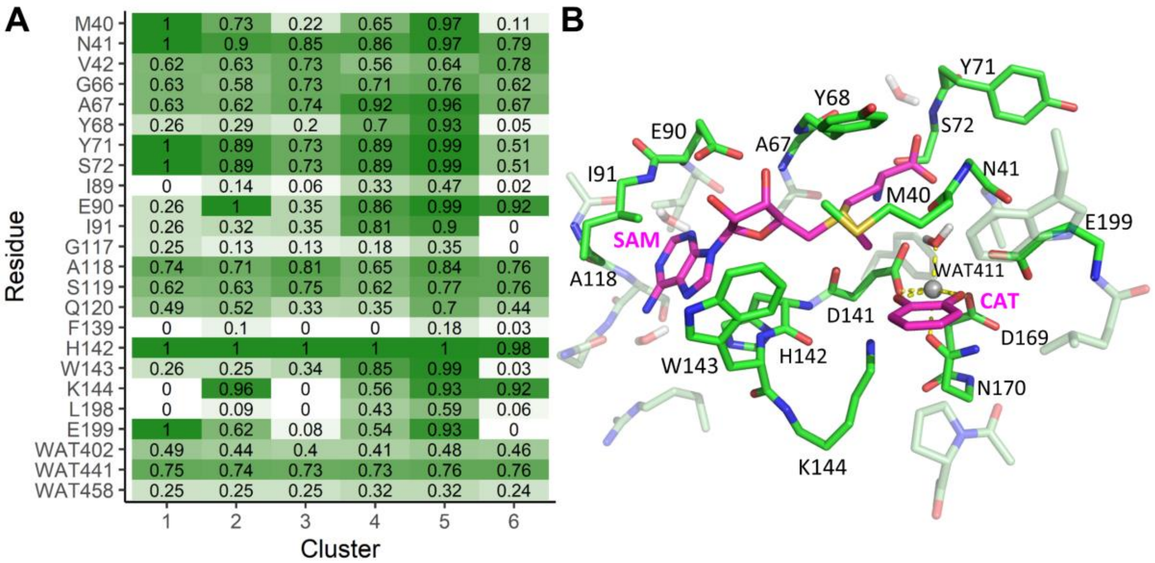
Figure 27. Computed methyl transfer $\Delta G^{‡}$ (circle) and ΔG_{rxn} (triangle) as models are constructed through either the Combinatoric Scheme 1 (A) and Combinatoric Scheme 2 (B). (C) Scatter and density plot of $\Delta G^{‡}$ (blue density) and ΔG_{rxn} (tan density) for all simulated models. Six clusters identified by *k*-means clustering of similar $\Delta G^{‡}$ and ΔG_{rxn} are differentially colored. Grey lines indicate the reference convergence values.

fewer than 300 atoms that yield accurate predictions, affirming that QM-cluster model convergence for COMT does not require > 400 atom models.

Identifying Important Residues

A general grouping of COMT QM-cluster models that predict similar (though not necessarily accurate) free energies is observed in Figure 27 for both combinatoric schemes. This leads to the question of which residues are required to form an accurate model? To more clearly distinguish the grouping of unique models that predict similar kinetic/thermodynamic properties, the *k*-means clustering algorithm was used to partition the entire dataset of unique QM-cluster models into six groups (Figure 27C) based upon their predicted $\Delta G^{\ddagger}$ and ΔG_{rxn}.[139] Though an unsupervised method was used to group the models, the identified clusters are reasonable and properly differentiate the models with both converged $\Delta G^{\ddagger}$ and ΔG_{rxn} (Cluster 5) from markedly inaccurate models (Clusters 1 and 6), as well as models with converged values for either $\Delta G^{\ddagger}$ or ΔG_{rxn}, but not both (Clusters 2, 3 and 4).

The residues that differ among the clusters give insight into which residues have a comparably strong influence on convergence. Tabulating the percent occurrence of each residue within the COMT models of each cluster (Figures 28 and Appendix C: Figure 7 and Table 2), nine residues present in >90% of the Cluster 5 models are absent or have a greatly reduced presence in other clusters. For example, in the models of Cluster 6, which systematically overestimate $\Delta G^{\ddagger}$ and 65% of which incorrectly predict an endergonic reaction, none contain E199 and only 11% contain M40. Without these residues, the QM-cluster models are missing 1) the stabilizing hydrogen bonding interactions between



Residue	1	2	3	4	5	6
M40	1	0.73	0.22	0.65	0.97	0.11
N41	1	0.9	0.85	0.86	0.97	0.79
V42	0.62	0.63	0.73	0.56	0.64	0.78
G66	0.63	0.58	0.73	0.71	0.76	0.62
A67	0.63	0.62	0.74	0.92	0.96	0.67
Y68	0.26	0.29	0.2	0.7	0.93	0.05
Y71	1	0.89	0.73	0.89	0.99	0.51
S72	1	0.89	0.73	0.89	0.99	0.51
I89	0	0.14	0.06	0.33	0.47	0.02
E90	0.26	1	0.35	0.86	0.99	0.92
I91	0.26	0.32	0.35	0.81	0.9	0
G117	0.25	0.13	0.13	0.18	0.35	0
A118	0.74	0.71	0.81	0.65	0.84	0.76
S119	0.62	0.63	0.75	0.62	0.77	0.76
Q120	0.49	0.52	0.33	0.35	0.7	0.44
F139	0	0.1	0	0	0.18	0.03
H142	1	1	1	1	1	0.98
W143	0.26	0.25	0.34	0.85	0.99	0.03
K144	0	0.96	0	0.56	0.93	0.92
L198	0	0.09	0	0.43	0.59	0.06
E199	1	0.62	0.08	0.54	0.93	0
WAT402	0.49	0.44	0.4	0.41	0.48	0.46
WAT441	0.75	0.74	0.73	0.73	0.76	0.76
WAT458	0.25	0.25	0.25	0.32	0.32	0.24

Figure 28. A) Relative frequency for each residue being present in the models of a *k*-cluster. Values are proportionally shaded to emphasize differences in residue composition among *k*-clusters. B) Visualization of the maximal 485-atom model highlighting the residues that occur in >80% of Cluster 5 models. The carbon atoms of the substrates are colored magenta.

E199 and the catechol and 2) the hydrophobic interactions between M40 and the SAM, resulting in consistently large deviations with respect to the converged free energies.

Surprisingly, residues identified as particularly important for convergence are not always localized around the atoms directly involved in the methyl transfer. For instance, E90 (which is present in 99% of the models in Cluster 5 but only in < 35% of the models in Clusters 1 and 3) is ~10 Å from the catechol, but plays a role in stabilizing and properly orienting the SAM. Other residues such as I91, A118, S119, and H142 are present in >70% of the models in Cluster 2 and appear to play important roles in crafting the active site microenvironment.

With residues crucial for accurate QM-cluster modeling of COMT identified, the next step is to examine contact and classification metrics to see if any were particularly suitable for predicting the relative importance of residues. For the contact classifications, there is unfortunately no consistent combination of contact types among the Cluster 5 models for yielding converged models. Using the total contacts between the seed and each residue (Figure 26B) as a ranking system proves modestly successful as 9 of the 13 residues present in > 80% of the Cluster 5 models have a high frequency of contacts with the seed and would be correctly prioritized. The four residues with low contacts (N41, A67, Y71, A118) are adjacent to high-contact residues and largely have main chain interactions with the seed, explaining the fewer contacts. The general success of using total contacts as a ranking scheme was previously shown in Figure 26B where converged models had 302 atoms as a lower bound. Improvements to this ranking method are warranted (and are under current investigation by our lab), ranging from incorporating additional chemical descriptors to the interatomic contacts (e.g., through *Arpeggio*),[143] to developing a weighting system to favor certain contact interactions (e.g., hydrogen bonding, polar, aromatic).

Expansion of QM-cluster Models Using *Arpeggio* as an Interaction Feature

The previous sections of this work are all founded on a *Probe*-based interatomic contact network, which is the default network generator for *RINRUS*. To supplement the previous results and give insight into the possible utility of alternate network-creation schemes, 78 additional COMT models were constructed based upon the residue interaction feature scheme using the residue interaction grouping defined by the *Arpeggio* program.[143] In short, *Arpeggio* identifies 15 different inter-residue interaction types: steric clash, covalent, van der Waals clash, van der Waals interaction, proximal interaction, hydrogen bond, weak hydrogen bond, halogen bond, ionic, metal complex, aromatic ring interaction, hydrophobic, carbonyl, polar, and weak polar. Residues were grouped based upon these interaction types, and combinations of the different interactions led to the creation and simulation of 78 unique model residue compositions.

The computed reaction thermodynamics/kinetics of these *Arpeggio*-based models are similar to those computed for the *Probe*-based models. Mapping the data back to the previously computed *k*-cluster centroids allows insight into which clusters the data might have been grouped (Figure 29). Of the 78 models, 29 (37%) group into Cluster 5 and 26 (33%) group into Cluster 2, placing over two-thirds of the new models close to or at the reference convergence values rather than the significantly incorrect models within the other clusters. Although there are fewer models compared to the *Probe*-based models, the residue composition (Figure 30) largely reflects the trends previously observed. Mapping the converged models back to the *Arpeggio*-based interactions, it is observed that the

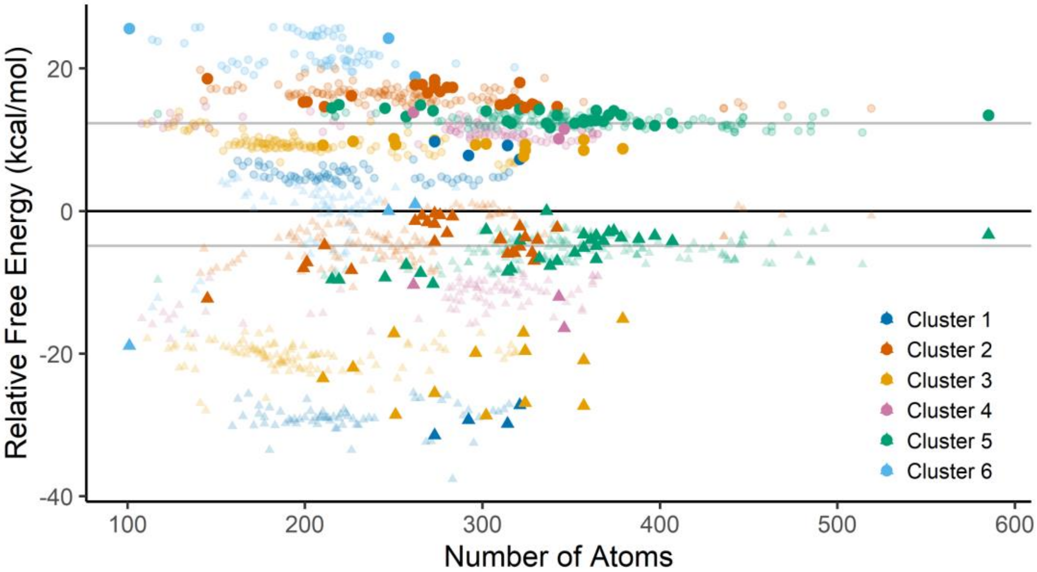


Figure 29. Computed methyl transfer $\Delta G^{\ddagger}$ (circle) and ΔG_{rxn} (triangle) as models are constructed through either the *Probe*-based contact network (transparent) or the *Arpeggio*-based interaction network. Six clusters identified by *k*-means clustering of similar $\Delta G^{\ddagger}$ and ΔG_{rxn} are differentially colored. Grey lines indicate the reference convergence values.

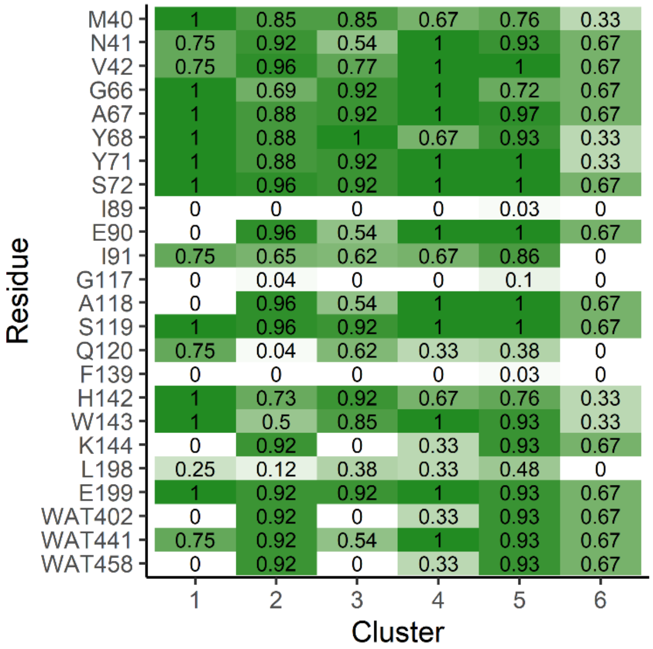


Figure 30. Relative frequency for each residue being present in the *Arpeggio*-based models of a *k*-cluster. Values are proportionally shaded to emphasize differences in residue composition among *k*-clusters.

models that included hydrogen bonding interactions, polar interactions, and van der Waals interactions (in addition to the metal complex) were consistently converged. Based upon these results, the residue interaction scheme employed by *Arpeggio* appears to be an better interaction feature classifier compared to *Probe*. Further investigation into how this chemical information may be used alongside the *Probe* information to yield a more finely-tuned improved model is underway by our lab.

Conclusions

Computational enzymology has made incredible impacts on understanding the atomic-level intricacies of enzyme function. While computational resources and scaling limitations of quantum chemistry are among factors limiting progress in this field, little attention has been given towards how poor or irreproducible model design might be hampering scientific progress. Many publication-quality enzyme models have been founded on rationale not necessarily suited for modeling non-spherical active sites (e.g., radial distance criterion) or via rationale prone to fallibility (a researcher's chemical intuition). Techniques addressing this problem by identifying important residues *a posteriori* have been useful but fail to meet the need for a computationally inexpensive *a priori* method for designing enzyme models.

As a step towards addressing community-wide problems in computational enzymology, we have been developing the *RINRUS* toolkit to automate the residue selection and construction of QM-cluster models. *RINRUS* utilizes the cheminformatics of interatomic contact networks as the rationale for identifying active site residues and ranking/classifying them. The catalytic methyl transfer reaction of the human COMT enzyme was simulated with a total of 550 unique models, illustrating how information from *RINRUS* was used to build models up from a base structure by either adding residues incrementally via a ranking scheme (e.g., total contacts with the seed) or by adding combinations of groups of residues (e.g., type of contacts). Clusters of models with common predictions of reaction and transition state free energies were compared to identify residues important for accurate simulations of COMT. Tracing the converged models and important residues back to how the models were constructed revealed that ranking residues by the frequency of their contacts with the seed was a particularly useful method, with QM-cluster models with 210 – 300 atoms yielding converged thermodynamic and kinetic properties. Additionally, 78 models built using chemical information from the *Arpeggio* program were evaluated to consider the potential benefits of a more defined chemical interaction type classifier. Chemical interaction types crucial for convergence were successfully identified, giving direction towards future improvements in how *RINRUS* designs and classifies its network interactions.

The major focus of this work has been to quickly converge energetic properties of smaller QM-cluster models to those of a maximally sized QM-cluster model. Further testing of the QM-cluster modeling methodology for accuracy to other well-defined experimentally known quantities (e.g., NMR chemical shifts) is an obvious next step for our lab to take. However, proper calibration of QM-based computational enzymology is contingent upon first developing a rational and reproducible scheme for building, QM-cluster models. Particular avenues of study include calibration of Density Functional Theory, one-electron basis set, implicit solvation parameters, empirical dispersion corrections, and other variables of electronic structure theory to truly assess the accuracy of QM-cluster modeling beyond a metric of internal consistency. Recent developments in linear scaling coupled cluster theory suggest ways to incorporate more rigorous “black box” electronic structure theories into the realm of computational enzymology. Investigating the structural and cheminformatic variation from constructing models using X-ray crystal structures versus conformational sampling frames from molecular dynamics simulations are also underway. These studies are in concert with investigations by our lab on improving the chemical descriptors and ranking schemes, integrating machine learning into the workflow, and examining how to best account for the impact that charged residues have on modeling the active site. In the future, we also seek to expand functionality into automating QM/MM modeling construction. A forthcoming publication will describe the *RINRUS* software package and include thorough tutorials. Public availability and adoption of *RINRUS* will substantially reduce learning curves for new practitioners of QM-cluster modeling and initiate a feedback loop for improving the generalizability of *RINRUS* for constructing QM-models of proteins beyond COMT and the enzymes studied within our lab.

Though model design and reproducibility questions have been largely ignored within the greater computational enzymology community, we hope this work will foster self-reflection on the underlying assumptions behind how atomic-level enzyme simulations are derived. The current practices often require unnecessarily large models to obtain accurate or internally converged results, which is limiting progress and is undoubtedly daunting to inexperienced chemists/biochemists interested in contributing to the field. Through the automated workflows provided by *RINRUS* and its successful results demonstrated in this work, we present the first steps towards discovering and implementing a computationally inexpensive, cheminformatic-based means for constructing reproducible, rational, and rigorous enzyme models. Admittedly, this case study of a single enzyme does not fully address all parameters of QM-cluster enzyme model construction. Nevertheless, reproducible workflows in computational enzymology, supported by *RINRUS* development, will improve openness, data sharing, and facilitate novel cyber- and software infrastructure in biochemistry and biology.

Chapter 5: Residue Interaction Networks and Machine Learning

Introduction

The specific structures and functions of proteins arise from the intricate networks of interacting amino acids. Numerous methods have been developed to characterize and quantify amino acid interactions to improve our understanding of biological processes. Examples include, but are not limited to, experimental NMR spectroscopy,[144] site-directed mutagenesis,[145] and quantum mechanical (QM) and molecular mechanical (MM) computations.[53,56,146] An example of qualitative methods to visualize protein interactions involves the application of graph theory to construct residue interaction networks (RINs) of a given protein. These graphs translate a protein structure into a set of nodes (defined as a single amino acid residue) interconnected by edges (defined as an electronic or steric interaction between two amino acid residues).[25] Edges are generally established by properties such as interatomic distances, hydrogen bonding, and interaction strength computed at the MM-level of theory.[25,147] Analyzing the topologies of RINs has already provided insight into structure−function features including protein stability,[148,149] allosteric regulation,[150,151] protein folding and dynamics,[152,153] and active site identification.[154–160] Building up from RINs, edges may also include metadata such as structural, chemical, or evolutionary properties.[161] This forms a structural interaction fingerprint (SIFt or SIF) for each edge, and the analysis of SIFts has proven valuable in the domains of drug design and virtual screening.[161–163]

Previous case studies by our lab have demonstrated that RINs may serve as a practical tool for designing rational models for QM-only (and potentially QM/MM) computations of enzyme active sites and protein functional sites.[57,58,61,164] In those studies, *Probe*[27] software was used to generate "contact dots" at coordinates where the van der Waals radii of two noncovalently bound atoms are in close contact. The *RINerator*[26] software package was then used to convert the interatomic contact data into a RIN with weights proportional to an estimated interaction strength and interaction type. Using these networks, atomic-level QM-cluster models of the active site were rationally constructed by including chemically important neighboring residues with edges linking their nodes to those of the substrate, catalytic residues, and/or cofactors.

Overall, these contact networks provide a more chemically reasoned basis for shaping QM-models compared to the popular method of radially expanding from a geometrically defined point, and it can be significantly less expensive than performing "back-end" model validation using charge shift or free energy perturbation analyses.[9,10,14,20,131] Our research group is developing a flexible Python-based software toolkit, *RINRUS* (Residue Interaction Network ResidUe Selector), to create reliable and reproducible atomic-level biological models. Starting from PDB-formatted structural data, prototype *RINRUS* can generate robust input files for electronic structure packages such as *Gaussian 16* and *PSI4*. Manuscripts detailing the *RINRUS* toolkit and its application in a large-scale testing of automated enzyme modeling are currently in preparation.

There is not yet evidence that a given RIN has a quantitative correspondence to the actual residue–residue interaction energies. It may be further reasoned that interatomic contact information alone should be insufficient to accurately predict interaction strength as it fails to account for residue charges, polarization, the strength of hydrogen bonding, and their dynamic conformations. As such, it is of interest to evaluate contact dot-based networks against a quantitative interaction energy network. Interaction energy-based networks have been previously constructed in the literature, generally using forcefield data averaged over molecular dynamic simulations.[165,166] This work will instead compute interaction energies using symmetry-adapted perturbation theory (SAPT), a quantum mechanical and nonempirical energy decomposition analysis method useful for partitioning the interaction energy into physical components (e.g., electrostatic, exchange, inductive, and dispersion energies).[167]

The main goal of this work is to create and evaluate the contact networks and interaction energy networks for five proteins to attempt to determine a statistical relationship between structural contact dot data and quantitative non-covalent interaction energies. We have trained an appropriate random forest model to predict SAPT-computed noncovalent interaction energies using minimal, readily available molecular descriptors. The trained forest is validated on an untrained protein network and is demonstrated to be suitable for predicting interaction energies from the structural contact information on similar untested networks. As this work is conducted with the design and construction of QM-cluster models of proteins in mind, we deviate from conventional protein network analysis and define our intraprotein RINs using chemical functional groups as our nodes rather than amino acid subunits.

Methodology

Protein Selection Criteria

Five model proteins from the Protein Data Bank[168] (PDB) were analyzed in this work. The PDB entries were selected randomly from a list of PDBs having all of the following criteria: (1) All five proteins have high X-ray crystallographic resolution (<2.0 Å resolution), (2) they have all or nearly all of their amino acid sequence identified in the deposited X-ray crystal structure, (3) residues missing in the crystal structure are terminal, and (4) there are no substrates, ligands, or metal cofactors in the crystal structures. The five model proteins are a bacteriophage T4 lysozyme (PDB entry =265L),[169] an alginate lyase from *Corynebacterium* sp. strain ALY-1 (PDB ID: 1UAI),[170] a peptidyl-prolyl isomerase from *Candida albicans*(PDB ID: 1YW5),[171] a chymotrypsin from *Cellulomonas borgoriensis* (PDB ID: 2EA3),[172] and a serine protease from *Anthrobacter nicotinovorans* (PDB ID: 3WY8).[173] Based on CATH Protein Structure Classification,[174] protein 256L has a mainly alpha-helical secondary structure, 1UAI is mainly beta-strand, and the remaining contain a mix of alpha and beta motifs.

Network Construction

Protein RINs are typically constructed in terms of the edge-interactions of their monomeric amino acid nodes (Figure 31A).[136,175] Chemically, this partitioning scheme muddles the distinct interactions occurring between the side chain and two residue−residue amide main chain groups for each amino acid. The *RINRUS* toolkit

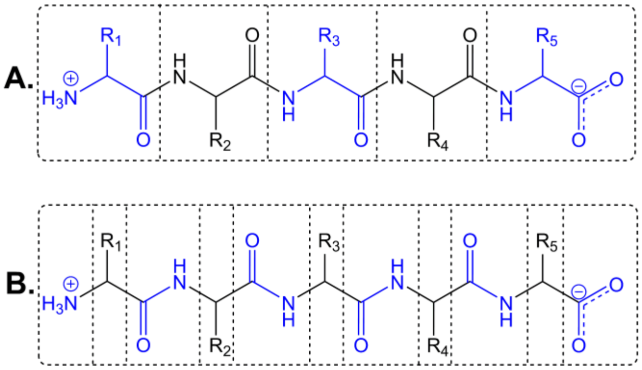


Figure 31. Atomic partitioning of a five amino-acid peptide in terms of amino acids (A) and chemical functional groups (B).

follows this reasoning and designs QM-models based on whether the main or side chain groups of a residue are interacting with the ligand or other moieties of interest. This work will deviate from the conventional usage of amino acid−based nodes and instead use the more chemically relevant main chain and side chain units for nodes (Figure 31B). In this work, "main chain" unit refers to the peptide functional group formed between two neighboring residues, and "side chain" unit refers to the functional group formed by a residue side chain, C_α, and H_α atoms.

Construction of these functional group-based networks with *RINRUS* is similar to the procedure for generating amino acid−based RINs by *RINerator*.[26] Hydrogens are first added to the crystallographic structures using the software *Reduce*.[59] The software *Probe*[27] rolls a virtual, small (0.25 Å radius) spherical probe along the van der Waals surface of each atom and generates either a contact "dot" interaction if the probe touches a noncovalently bound atom or a contact "clash" interaction if the probe encounters overlapping van der Waals surfaces. *Probe* quantifies each contact with a score based upon an error-function weighting of the volume of overlap between the spherical probe and the van der Waals surface, and then sums the score for each atom pair. The network is constructed from these results. Nodes represent the main chain and side chain functional groups of the protein, and they are connected by one or more edges representing the *Probe*-detected noncovalent interactions. Edges possess information pertaining to whether the interaction is from interatomic contact dots, "bad overlaps" (chemically defined as a steric clash), or between hydrogen bonding atoms. Each edge is weighted by the summed interatomic contact scores. In short, this procedure generates an undirected contact network of main/side chain nodes interconnected by edges weighted by a contact score. Graph visualization and analyses were performed using Cytoscapev3.7.1.[176]

Computation of Interaction Energies

Given the desire to investigate the interactions of residue functional groups and the need for a robust ab initio method to handle noncovalent interactions, SAPT was used to compute the interaction energies.[167,177–179] For the simplest SAPT method, SAPT0, the interaction energy decomposition can be described by the equation:

$$E_{int}^{SAPT0} = E_{elec}^{(1)} + E_{exch}^{(1)} + [E_{ind}^{(2)} + E_{exch-ind}^{(2)} + \delta E_{HF}^{(2)}]_{ind} + [E_{disp}^{(2)} + E_{exch-disp}^{(2)}]_{disp} \quad (2)$$

where the interaction energy is broken into components of electrostatic, exchange-repulsion, induction, and dispersion terms. Extensions of this technique include functional group SAPT (F-SAPT),[180] which provides an effective two-body partition of the SAPT terms to localized functional groups, and intramolecular SAPT (I-SAPT),[181,182] which computes the intramolecular interaction between two moieties within the embedding field of a third body.

As this work seeks to establish a link between the contact networks and inter-residue interaction energies, functional group-based networks served as the basis for identifying interacting residues. If two nodes (main/side chains) had an interlinking edge (interaction) present in the contact network, a fragment model of the interacting pair is constructed. Starting from the hydrogen-added X-ray crystal structures used for the contact networks, the two interacting functional groups are isolated along with select atoms of neighboring residues to maintain the local chemical environment. Determining which neighboring atoms to include is based on the identities of the interacting functional groups and sequence distance between them (Figure 32). Interacting pairs sequentially "distant" with 3 or more main/side chain units between them are modeled as two noncovalently bound fragments. For each fragment, if an interacting unit is a side chain, the fragment is constructed to include the adjacent main chain units and is capped with methyl groups (Figure 32A). If an interacting unit is a main chain, the fragment is constructed to include the C_α atoms of neighboring side chain units along with adjacent sequential main chain units (Figure 32B). Interacting functional groups sequentially "close" with 1 or 2 main/side chain units between each other are modeled as a single fragment. Design of the single fragment followed similar neighboring atom-selection rules as the aforementioned "distant" pair rules and is visually presented in Figure 32C−H. Additional details and treatment of unique cases (e.g., prolines and cystines) are provided in Appendix E.

Hydrogens are added to the model fragment(s) to satisfy the valency where bonds were trimmed using *PyMol* v2.3.0a0.[138] To ensure the hydrogens added by both *Reduce* and *PyMol* are in optimal positions with minimal steric effects, all hydrogens were geometrically optimized using density functional theory (DFT) with the hybrid B3LYP exchange-correlation functional[63,64] using the 6-31G(d') basis set for N, O, and S atoms and the 6-31G basis set for C and H atoms.[67] The Grimme D3 (Becke-Johnson) dispersion correction (GD3BJ) and conductor-like polarizable continuum model (CPCM) with UAKS sets of atomic radii, a nondefault electrostatic scaling factor of 1.2, and a dielectric constant of ε= 4 were also used.[68,69,105] The heavy atoms remained frozen to their crystallographic coordinates. All QM geometry optimizations were done with the Gaussian 16.B01 software package.[62]

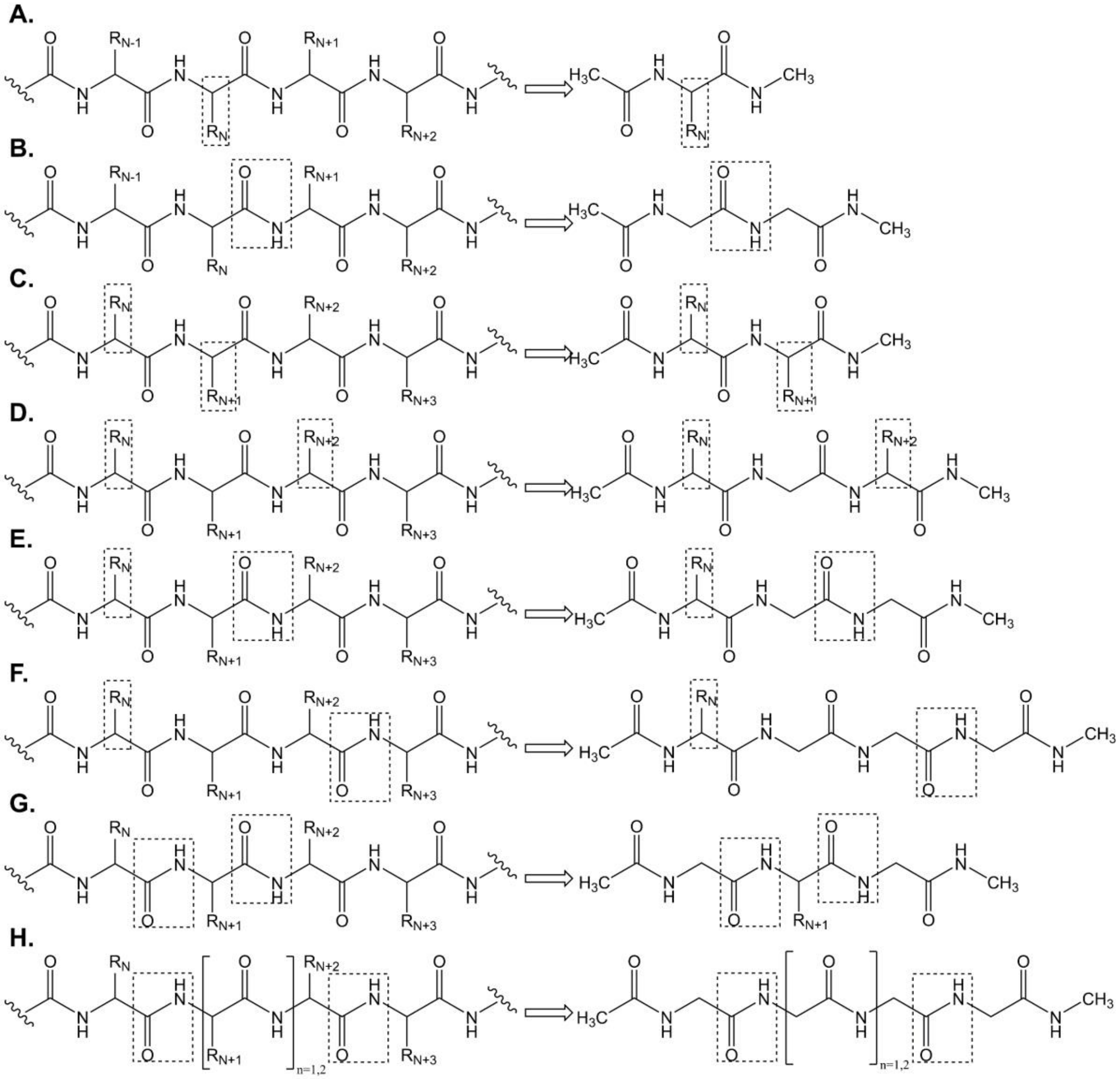


Figure 32. Schemes for translating protein functional groups into fragments. Distant interacting groups are trimmed into two separate fragments based on whether the group is a side chain (A) or main chain (B). Close interacting groups are trimmed into a single fragment based on whether the groups are both side chains (C, D), a side chain and a main chain (E, F), or both main chains (G, H).

The final geometries after hydrogen optimization were used for computing the SAPT interaction energies. The interaction energy between two sequentially distant functional groups modeled as two noncovalently bound fragments was computed using the functional group F-SAPT method. The interaction energy between two sequentially close functional groups modeled as a single fragment was computed using the intramolecular I-SAPT method. Energies were computed at the SAPT0 level of theory using the jun-cc-pVDZ basisset.[177,180–182] The BioFragment Database, which archives the structures and energies of 3380 side chain-side chain (SSI dataset) and 100 backbone−backbone interactions (BBI data set), demonstrates the SAPT0/jun-cc-pVDZ method as an inexpensive, reliable level of theory for computing residue side chain and main chain interactions.[183] The mean signed error of this method compared to "silver standard"[183] DW-CCSD(T**)-F12/aug-cc-pV(D+d)Z reference energies is 0.51 kcal/mol (0.53 kcal/mol standard deviation) for the SSI, and −0.10 kcal/mol (0.74 kcal/mol standard deviation) for the BBI data sets. All SAPT computations were done with the F/I-SAPT module of *PSI4* v1.3.[184]

Statistical Testing

All statistical methods were conducted using the statistical computing environment *R* version 3.6.0.[140] Random forest regression modeling was performed to construct a predictive regression model suitable for predicting interaction energies from contact network information and qualitative descriptors using the randomForest library.[185] Random forest regression involves an ensemble of regression decision trees from which the prediction of a continuous variable is computed as the average of the predictions of all the trees within the forest.[186] The predictive ability of the forests is evaluated using both a test set and out-of-bag validation. The out-of-bag error represents the mean prediction error of the bagged subsample of data not used for tree growth. The importance of the descriptors in the random forest models were also evaluated by measuring the change in mean squared error for the out-of-bag validation as each descriptor is permuted.[185] For this method, larger changes in mean squared error reflect a greater importance of the descriptor in the random forest.

Results and Discussion

Protein Network Analysis

In this work, the networks of five proteins (Protein Data Bank IDs 1UAI,[170] 1YW5,[171] 256L,[169] 2EA3,[172] and 3WY8[173]) were constructed by partitioning the protein into residue side chain and main chain units. By defining the network nodes in terms of residue side chain and main chain units, the number of nodes nearly doubles compared to conventional protein networks. Although the atomic size of the four-atom MC unit is smaller than nearly all SC units, there is not a diminutive number of MC interactions, reinforcing the idea that this functional group partitioning does not distort or inappropriately distribute interactions between the two different node types (Table 4).

Table 4. General Network Information of the Tested Protein Models

PDB	Number of Residues	Number of Nodes	Number of Edges			
			MC-MC	MC-SC	SC-SC	Total
1UAI	223	430	233	369	408	1010
1YW5	177	336	234	273	287	794
256L	164	319	315	356	262	833
2EA3	183	347	200	281	299	780
3WY8	219	423	252	338	389	979

These results are expected as the chemical importance of hydrogen bonding protein backbones is clearly established, though as far as we are aware, this protein partitioning scheme is novel within the domain of protein network analysis.

As these networks are based on the contact information within the protein structure, the predicted interactions of the network are highly dependent on the positions of the atoms in the X-ray crystal structure. This fact, compounded with the finer-grained partitioning of residues into functional groups, allows the interesting opportunity for a network to form small "tidal islands" where two or more functional groups have noncovalent interactions with each other but no noncovalent interactions linking them to the main network body. One island is observed in each of the networks for 1UAI, 1YW5, and 2EA3 (Figure 33 and Appendix E: Figures 9 and 10). In all three instances, the islands correspond to main/side chains located on the surface of the protein that are oriented away from neighboring residues and outward into the solvent, likely to improve protein solubility. These islands reflect the minimal roles these functional groups have in

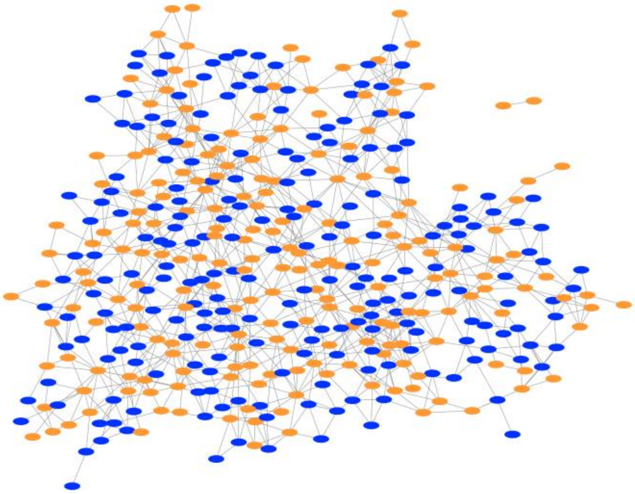


Figure 33. Contact network of PDB 1UAI where nodes are colored by interacting main chain (blue) and side chain (orange). An "island" of nodes is observed in the upper right region. A higher resolution image is provided in Appendix D.

intraprotein stability, but it also demonstrates a limitation of predicting interactions for mobile residues from a single crystallographic snapshot. For example, if the networks of multiple snapshots of the proteins throughout a molecular dynamics simulation were compiled, it is likely that a selection of the snapshot networks would show the tidal island connected to one or more nodes of the main network body. Determining the best protocol for formulating networks more representative of mobile and solvent interacting functional groups is beyond the scope of this work but is an interesting point for future investigation in our laboratory.

Molecular Descriptors and Interaction Energies

As it is the goal of this work to utilize only the immediately available contact and structural descriptors of the protein networks, the interactions were characterized by the following network, structural, and molecular descriptors. *Position* is the sequence ID of the two interacting functional groups. *Sequence Distance* is the distance in sequential number between the interacting functional groups. *Functional Group Type* is distinguishing whether the interacting unit is a main chain or a side chain. *Functional Group Name* is the side chain identity. *Contact Types* is the total number of wide contacts, close contacts, small overlaps, bad overlaps, and hydrogen bonding contacts computed by *Probe*.[27] *Score* is the interaction strength score computed by *Probe* as a function of the overlap between contact probes and the van der Waals radii.[26,27] *Center of Mass Distance* is the distance between the center of mass for the two functional groups. *Interaction Charge* classifies the interaction based on the individual charge of each species (positive, negative, neutral) with further separation into two categories for either a neutral side chain or a main chain. *Chemical Type* classifies the interaction based on the functional group’s chemical character, specifically as being either a main chain, aliphatic (Gly, Ala, Val, Leu, Ile, and Pro), aromatic (Phe, Tyr, Trp, and His), polar (Ser,Thr, Cys, Met, Asn, and Gln), negatively charged (Glu and Asp), or positively charged (Arg and Lys) group.

Collectively, 4381 pair interactions were computed for the five proteins. There is generally abundant representation of the different interaction types among the test proteins with the only dearth of data being for interactions where both side chains are charged (Appendix E: Table 3). This is expected as charged amino acid sidechains are usually located at the surface of water-soluble proteins and would have predominantly side chain-solvent interactions (an interaction not examined in this work). The sparse number of interactions between similarly charged, solvent-exposed sidechains (POS−POS and NEG-NEG) is also noted but rationalized as resulting from factors including the absence of a complete hydration shell in the X-ray crystal structure, the fact that protonation equilibrium is not considered to allow fluctuation between charged/neutral states, and the consequence of only evaluating the interaction energies at the nuclear positions within the X-ray crystal structure compared to a set of molecular dynamics simulation snapshots. The distribution of the SAPT interaction energies among the five proteins is also consistent, where similar types of residue pairs consistently yield similar interaction energy strengths (Figure 34, and Appendix E: Figure 14). This fundamental chemical consistency among functionally and evolutionarily different proteins lends support to our belief that the prediction model yielded from this work should be generalizable to many proteins.

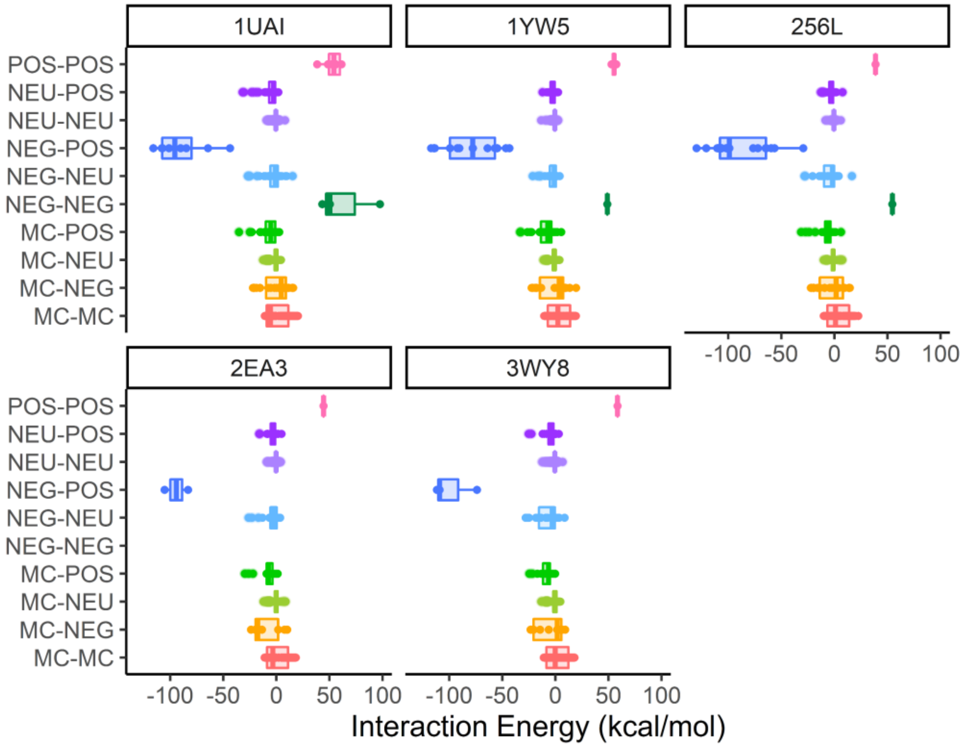


Figure 34. Box and whisker plots of the range of interaction energy values among the test set. The data is separated based upon the interaction charge of the two species. *MC* refers to main chains, and *POS*, *NEU*, and *NEG* refer to positive, neutral, and negative side chains, respectively.

For the computed SAPT energies, the identity of whether electrostatic or dispersion forces dominate the interaction energy is consistent in our pair models with expected chemical intuition (Figure 35). The grid in Figure 35 plots each computed main/side chain interaction as a box proportionally sized to the number of corresponding pairs (i.e., more data points for a given interaction type is displayed as more boxes) and colored according to the contribution of electrostatics/dispersion. As may be expected, the interactions between aliphatic or aromatic residues are predominantly influenced by dispersion forces, the interactions between charged residues are predominantly influenced by electrostatic forces, and the interactions between polar residues and between combinations of nonsimilarly typed residues are comparably influenced by both forces.

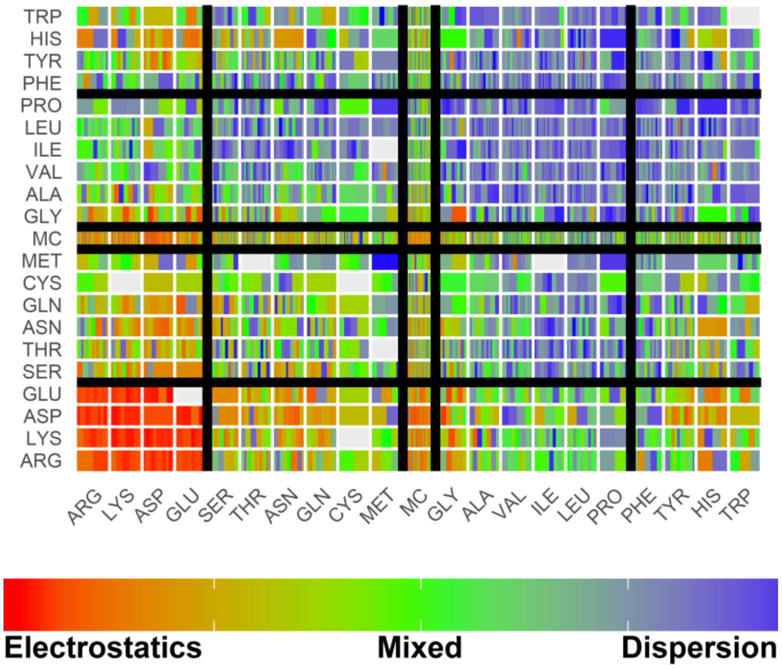


Figure 35. Grid of computed main and side chain pair interactions colored according to the proportion electrostatics and dispersion SAPT decomposition terms contribute to the interaction energy.

These trends are consistent with previously published reports and data from the BioFragment Database.[183,187]

Comparing Interaction Energies to Probe Descriptors

It may be reasoned that interatomic contact data alone should be insufficient to accurately predict interaction strength due to the lack of information regarding residue charges, polarization, or hydrogen bonding strength. To affirm that there is indeed no simple correlation between the descriptors output by *Probe* (i.e., number of contacts and interaction score) and the computed or relative interaction energy strength, correlation plots are provided (Figure 36). There is no direct linear, polynomial, or exponential correlation between the two descriptors and the interaction energy. It may be argued that descriptors are more appropriate for comparison among similarly charged residue pairs as

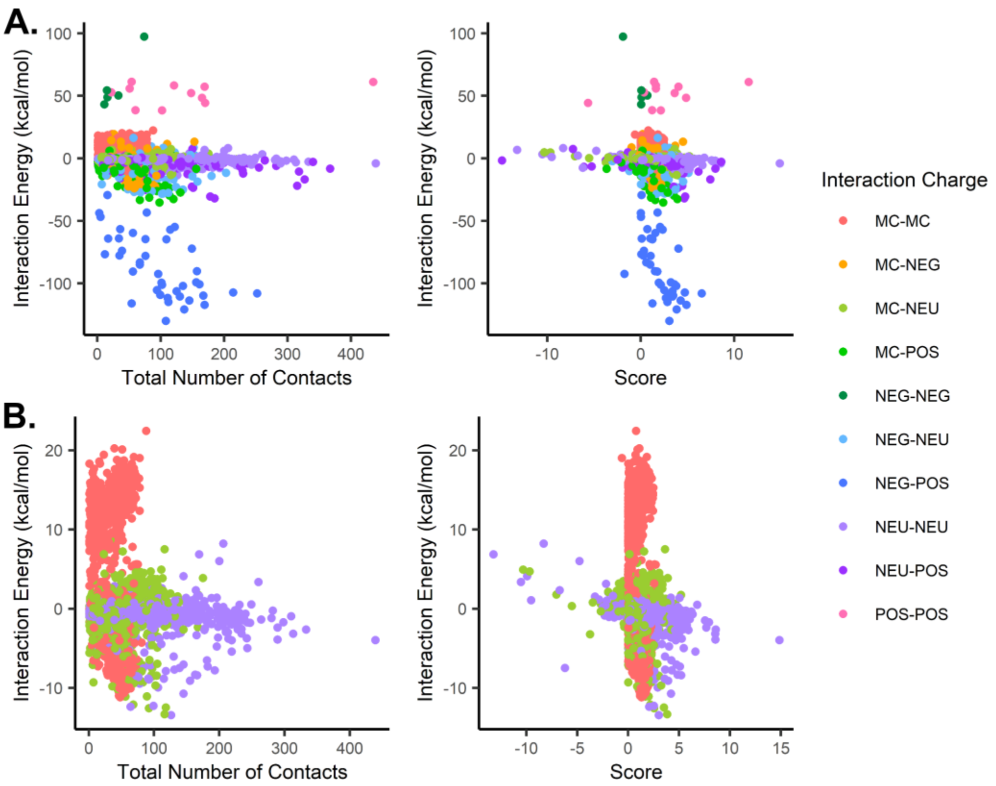


Figure 36. A) Correlation plots comparing the *Probe*-computed total number of contacts and score against SAPT-computed interaction energies. B) Correlation plots showing only the neutral-charged interactions.

the dominating electrostatic interactions of charged functional groups will disproportionally spread the results. However, this is shown to not be the case (Figure 36B). The distribution of neutral pairs of functional groups remains scattered without any distinct relationship between the contact descriptors and interaction energy strength. Additionally, it may be noted that both the number of contacts and the interaction score do not qualitatively correlate with whether a particular interacting pair would have a favorable (negative) or unfavorable (positive) interaction energy value. Overall, these results demonstrate that it would be inappropriate to use the contact score and count alone in approximating the relative interaction energy between two residue side/main chains. As additional information is required to effectively predict functional group pair interaction strength, we turn to using random forest modeling on the previously defined, minimal descriptor set.

Random Forest Regression

Random forest regression modeling was used to identify descriptors important for determining the interaction energy between two main/sidechain units and to construct a predictive model. The training set was formed from the interactions computed for the networks of 1UAI, 256L, 2EA3, and 3WY8 (3589 data points); the test set was from the network for PDB 1YW5 (792 data points). 1YW5was selected as the test set for having a complete set of all interaction types.

The parameter for the number of predictors sampled for splitting at each node in the forest was initially tuned on the training set using mean squared error results to select the optimized parameter. The model was tested with up to 16 variables at each node and with 500 trees in the forest. The results (Figure 37) demonstrate the appropriate number of predictors to sample at each split should be 6, which is in agreement with the general

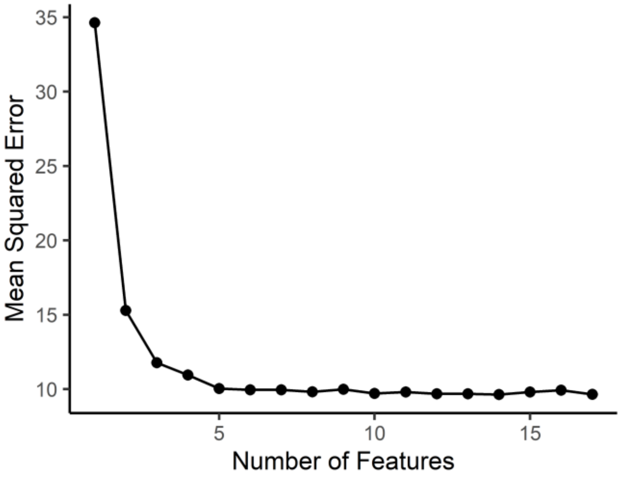


Figure 37. Convergence of mean squared error as the number of features tested at each node is increased (500 trees used in each forest test).

rule of thumb that the number tested should be approximately the total number of descriptors possible (in this set, 17) divided by 3.[185] Literature has suggested that the number of trees in the forest does not need to be optimized given a large enough number of forests.[188] As such, the random forests are run using the optimized node sampling of 6 and a total number of trees of 1000.

For the training set, the random forest using all descriptors was able to account for 91.9% of the variance in interaction energies and a root-mean-square error (RMSE) of 3.2 kcal/mol. The frequency that each descriptor is used in the trees of the random forest and the number of training data points affected by the inclusion of the descriptor are described by the relative importance of the variable. The importance of the descriptors in this forest, as measured by the increase in mean squared error of predictions estimated through out-of-bag error, indicate that the most important five descriptors are (in decreasing importance) the Chemical Type, Interaction Charge, Center of Mass Distance, Number of Hydrogen Bonding Contacts, and Sequence Distance (Appendix E: Table 4). Functional Group Positions and the number of Bad Overlap Contacts were observed to have a insignificant impact on the model (<0.2% increase in mean squared error) and were excluded from descriptor selection in subsequent random forests.

The fit of the model to training data is not representative of how accurate the model will predict values from new data. In consideration of this, the random forest model was tested against the 1YW5 validation set and was shown to account for 94.3% of the variance in the validation set and have a RMSE of 3.2 kcal/mol and a mean absolute error (MAE) of 1.6 kcal/mol. The forest constructed from the combined training and test sets was concurrently run and it was able to account for 92.2% of the variance and have an RMSE of 3.1 kcal/mol and a MAE of 1.6 kcal/mol. These forests were repeated 10 times and the cumulative range of each descriptor's importance is shown in Figure 38. The relative rank of descriptor importance is notably consistent between validation and test set.

In predicting the interaction energies of the test set, most of the values are concentrated around the line of equality between the computed and predicted value (Figure 39A), showing that the model is able to appropriately estimate the property with good accuracy (94.3% variance explained). The distribution of actual and relative errors between the SAPT-computed and random forest-predicted energies for the validation set is plotted in density plots (Figure 39B and Appendix E: Figure 20). The error outside the range of ±1.6 kcal/mol is largely in pair models involving one or more charged residues (Table 5 and Appendix E: Figure 21), an expected result as there is both a smaller

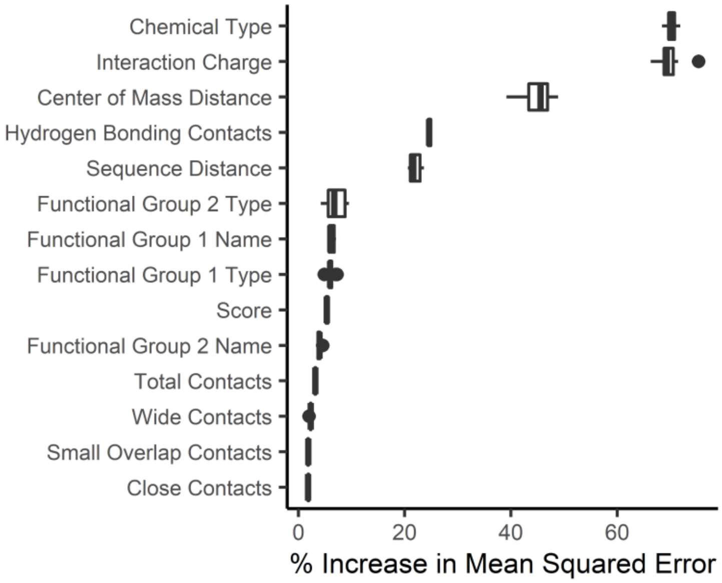


Figure 38. Range of the importance of descriptors for ten random forest models of the validation set. Importance is measured by the percent increase in mean square error where high values of percent increase in mean square error indicate more important descriptors in the random forests.

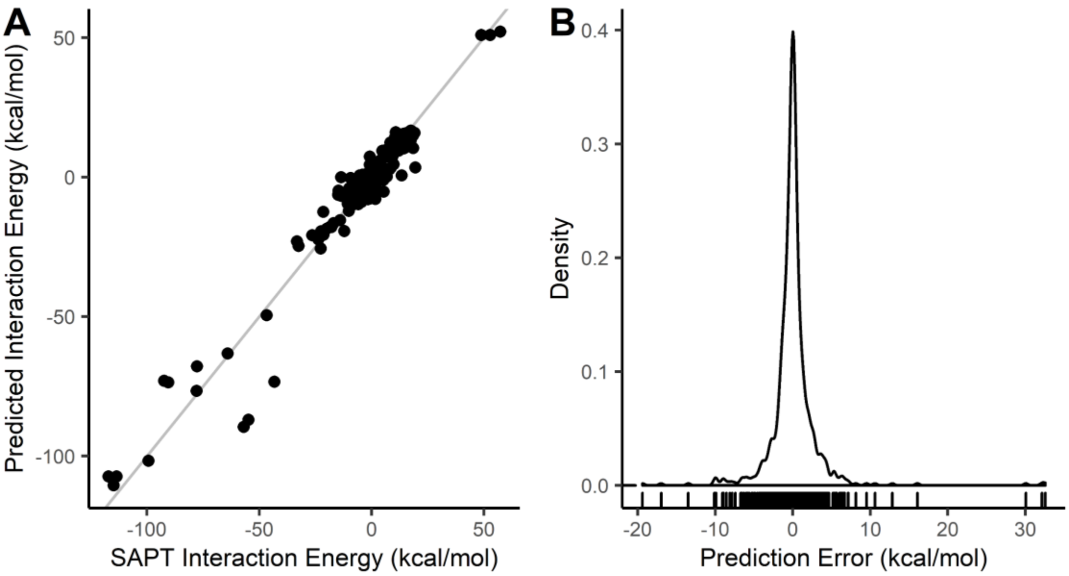


Figure 39. A) Plot of SAPT-computed vs random forest-predicted interaction energy. The grey line represents the line of equality where RF-predicted energies would equal SAPT-computed energies. B) Density plot of differences between SAPT-computed and RF-predicted energies

Table 5. Distribution of Predicted Error for 1YW5 by Model Charge

Model Charge Type	Mean Error (kcal/mol)	Standard Deviation (kcal/mol)	Mean Absolute Error (kcal/mol)	Number of Interaction Energies with Incorrect Sign
MC-MC	−0.18	1.9	1.3	11
MC-NEG	−1.0	6.2	4.4	2
MC-NEU	0.14	1.7	1.2	56
MC-POS	1.3	5.3	4.4	2
NEG-NEG[1]	2.0	--	--	0
NEG-NEU	0.37	3.2	2.3	12
NEG-POS	−2.4	17.9	13.0	0
NEU-NEU	0.05	1.1	0.62	19
NEU-POS	−0.54	2.8	2.1	10
POS-POS[1]	−3.6	--	--	0
Entire Data Set	−0.05	3.2	1.6	112

[1] NEG-NEG consists of one data point and POS-POS consists of only two data points, thus mean absolute error and standard deviation is inappropriate to report. The NEG-NEG and POS-POS mean errors are reported for reference, though there is no statistical significance of these values compared to the other subsets.

sample size of the charged interaction types and these models often have interaction energies an order of magnitude greater than neutral models. This difference is represented most in the pair models involving oppositely charged side chains (classification NEG-POS). In this subset, there is a large standard deviation of error (±17.9 kcal/mol) which would appear to suggest a poor ability to predict the interaction energy, especially compared to the neutral interacting side chains (NEU-NEU standard deviation is ±1.1 kcal/mol). This is rationalized by the fact that the range of energies for NEG-POS spans −43 to −117 kcal/mol, a substantially larger range compared to NEU-NEU (−7 to 2 kcal/mol), and so the RF-predictive model does provide a relatively accurate prediction of the diverse charge-based interactions.

Lastly, it is important to examine the qualitative accuracy of the random forest modeling toward predicting whether the interaction is stabilizing (negative in value) or destabilizing (positive). The results for the number of qualitatively incorrect predictions for the 1YW5 test set are presented in Table 5, showing that 14% of the predicted results were of the incorrect sign. Of the incorrect predictions, 69% are typing destabilizing interactions as stabilizing interactions, indicating a bias in this random forest model toward predicting favorable interactions. Reduction of both this bias and the frequency of mistyped pair interactions is expected to occur with additional data points and chemical descriptors, which is currently being explored by our laboratory.

Conclusions

With the continued growth of computational enzymology, there is need for a relatively inexpensive and rational methodology for determining which residues are to be included in the QM-region of QM-cluster and QM/MM enzyme models. The *RINRUS* program in development by our lab seeks to provide an automated solution to this by utilizing the information from protein contact networks. In this work, we sought to evaluate the relationship between the qualitative and semiquantitative data of contact networks and quantitative interaction energies. The contact networks of five proteins were constructed in a novel way by defining the nodes in terms of chemical functional groups (main chains and side chains) rather than as conventional amino acid residues. Through this partitioning, the network is crafted to more appropriately represent the unique chemical inter-molecular interaction types. The noncovalent interaction energies for the edges in the five networks were computed using the ab initio SAPT method, totaling 4381 main/side chain interaction energies.

As our results showed no direct correlation between the immediate information from the contact networks (*Probe* score and contacts) and quantitative interaction energies, we constructed a predictive random forest model capable of predicting interaction energies from minimal descriptors, namely the contact network information (number of contacts, types of contacts, contact score), sequence information, a general interaction type classification scheme, and center of mass distances. When tested against a test set, the random forest was able to account for 94.3% of the variance in the data with a root mean squared error of 3.2 kcal/mol and mean error of 1.6 kcal/mol. Most of the variance arising from models involves charged functional groups. The data used in this work is provided in the Supporting Information in the interest of serving both as training data for predictive random forests for other works and as a benchmark for future improved statistical modeling. As this work utilizes only a minimal set of chemical descriptors immediately available from contact mapping methods, we anticipate significant improvement in qualitative and quantitative accuracy by including rationally selected 1D-, 2D-, and 3D- molecular descriptors. Further expanding the dataset allows for the opportunity to utilize unsupervised machine learning methods, such as neural networks, for improved model quality.

In summary, this work demonstrates the ability to use random forests to predict interaction energies among residue functional groups from readily available contact network descriptors. In addition to the immediate impact these results have in improving the development of the *RINRUS* software, this work may serve as a basis for functional group network-based investigations into fields examining protein−protein interactions, noncanonical amino acids, and the impact of point mutations.

Chapter 6: Conclusions

QM-cluster models have proven to be a reliable modeling technique for obtaining atomic-level insight into the inner workings of proteins and enzymes. Although early studies were originally limited to modeling QM-systems of <100 atoms, the accelerated advancement of computer hardware and software has now enabled simulation of QM-models with several hundred atoms at ever-increasing accuracy, allowing a more thorough modeling of the protein active site. Despite having been utilized for several decades, there has not yet been an efficient, systematic protocol developed for rationally designing enzyme QM-cluster models. This work detailed the development of the cheminformatics-based toolkit *RINRUS* and its application towards multiple different biosystems.

In Chapter 2, QM-cluster models of the inner pocket of six bioengineered threonyl-tRNA synthetase enzymes were used to investigate the energetic profiles of BiPhe dihedral rotation. The models were used to demonstrate how, after several iterations of protein engineering, the final protein synthesized had an inner pocket able to compact the staggered BiPhe dihedral angle into a coplanar conformation, creating a transition state analogue structure. In Chapter 3, QM-cluster models of the active site of human carbonic anhydrase II were simulated with the native Zn^{2+} ion alongside other transition metal ions. The models gave insight into the theoretical viability for the Fe^{2+}-substituted metallovariant to catalyze CO_2 hydration, and the findings are soon to be followed up by experimental studies. In Chapter 4, hundreds of QM-cluster models of the catechol-O-methyltransferase active site were constructed to explore what cheminformatics might be particularly useful for creating reliable, converged models.

The interatomic contact metrics currently employed by *RINRUS* were validated, and avenues for improvement were highlighted. In Chapter 5, it was shown that there is no correspondence between interatomic contacts and quantitative inter-residue interaction energies; however, random forest algorithms using the interatomic contacts and several easily accessible chemical descriptors are capable of predicting the interaction energies with considerable accuracy.

The studies in this work affirm the capability of the *RINRUS* workflow to model enzymes with varying active site sizes. However, much remains to be done to further fine tune and improve upon *RINRUS* to ensure it is able to be easily applied to the wide range of enzymes. Chapter 3 highlighted the need to account for waters in solvent-accessible active sites, especially since X-ray crystal structures may not have well-resolved hydration spheres. Chapter 4 touched on the importance of handling charged residues and how alternative interaction classification schemes (e.g. *Arpeggio*) may be useful in differentiating the interactions important in converged models. Chapter 5 demonstrated the potential behind using machine learning algorithms to transform easily computed qualitative/semi-quantitative descriptors into quantitative metrics. Additional investigations into these details, in addition to examining the benefits of using molecular dynamics structures and networks, are underway. Nevertheless, *RINRUS* currently stands as a strong, cheminformatics-based tool whose further development and adoption by the enzymology community will improve QM-cluster modeling accuracy and facilitate novel insights behind the inner workings of enzymes.

References

(1) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. Computational Enzyme Design. *Angew. Chem. Int. Ed*. **2013**, *52* (22), 5700–5725. https://doi.org/10.1002/anie.201204077.

(2) Kollman, P. A.; Kuhn, B.; Peräkylä, M. Computational Studies of Enzyme-Catalyzed Reactions: Where Are We in Predicting Mechanisms and in Understanding the Nature of Enzyme Catalysis? *J. Phys. Chem. B*. **2002**, *106* (7), 1537–1542. https://doi.org/10.1021/jp012017p.

(3) *The Nobel Prize in Chemistry 2013* https://www.nobelprize.org/prizes/chemistry/2013/summary/ (accessed Jun 9, 2020).

(4) Ahmadi, S.; Barrios Herrera, L.; Chehelamirani, M.; Hostaš, J.; Jalife, S.; Salahub, D. R. Multiscale Modeling of Enzymes: QM-Cluster, QM/MM, and QM/MM/MD: A Tutorial Review. *Int. J. Quantum Chem.* **2018**, *118* (9), e25558. https://doi.org/10.1002/qua.25558.

(5) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. A Practical Guide to Modelling Enzyme-Catalysed Reactions. *Chem. Soc. Rev.* **2012**, *41* (8), 3025–3038. https://doi.org/10.1039/c2cs15297e.

(6) Borowski, T.; Quesne, M.; Szaleniec, M. QM and QM/MM Methods Compared: Case Studies on Reaction Mechanisms of Metalloenzymes. In *Advances in Protein Chemistry and Structural Biology*; Academic Press Inc., 2015; Vol. 100, pp 187–224. https://doi.org/10.1016/bs.apcsb.2015.06.005.

(7) Kmita, K.; Wirth, C.; Warnau, J.; Guerrero-Castillo, S.; Hunte, C.; Hummer, G.; Kaila, V. R. I.; Zwicker, K.; Brandt, U.; Zickermann, V. Accessory NUMM (NDUFS6) Subunit Harbors a Zn-Binding Site and Is Essential for Biogenesis of Mitochondrial Complex I. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (18), 5685–5690. https://doi.org/10.1073/pnas.1424353112.

(8) Li, X.; Siegbahn, P. E. M.; Ryde, U. Simulation of the Isotropic EXAFS Spectra for the S2 and S3 Structures of the Oxygen Evolving Complex in Photosystem II. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (13), 3979–3984. https://doi.org/10.1073/pnas.1422058112.

(9) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martínez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol O-Methyltransferase. *J. Phys. Chem. B* **2016**, *120* (44), 11381–11394. https://doi.org/10.1021/acs.jpcb.6b07814.

(10) Sumner, S.; Söderhjelm, P.; Ryde, U. Effect of Geometry Optimizations on QM-Cluster and QM/MM Studies of Reaction Energies in Proteins. *J. Chem. Theory Comput.* **2013**, *9* (9), 4205–4214. https://doi.org/10.1021/ct400339c.

(11) Hu, L.; Söderhjelm, P.; Ryde, U. Accurate Reaction Energies in Proteins Obtained

by Combining QM/MM and Large QM Calculations. *J. Chem. Theory Comput.* **2013**, *9* (1), 640–649. https://doi.org/10.1021/ct3005003.

(12) Hu, L.; Söderhjelm, P.; Ryde, U. On the Convergence of QM/MM Energies. *J. Chem. Theory Comput.* **2011**, *7* (3), 761–777. https://doi.org/10.1021/ct100530r.

(13) Sumowski, C. V.; Ochsenfeld, C. A Convergence Study of QM/MM Isomerization Energies with the Selected Size of the QM Region for Peptidic Systems. *J. Phys. Chem. A* **2009**, *113* (43), 11734–11741. https://doi.org/10.1021/jp902876n.

(14) Liao, R. Z.; Thiel, W. Convergence in the QM-Only and QM/MM Modeling of Enzymatic Reactions: A Case Study for Acetylene Hydratase. *J. Comput. Chem.* **2013**, *34* (27), 2389–2397. https://doi.org/10.1002/jcc.23403.

(15) Solt, I.; Kulhánek, P.; Simon, I.; Winfield, S.; Payne, M. C.; Csányi, G.; Fuxreiter, M. Evaluating Boundary Dependent Errors in QM/MM Simulations. *J. Phys. Chem. B* **2009**, *113* (17), 5728–5735. https://doi.org/10.1021/jp807277r.

(16) Vanpoucke, D. E. P.; Oláh, J.; De Proft, F.; Van Speybroeck, V.; Roos, G. Convergence of Atomic Charges with the Size of the Enzymatic Environment. *J. Chem. Inf. Model.* **2015**, *55* (3), 564–571. https://doi.org/10.1021/ci5006417.

(17) Morgenstern, A.; Jaszai, M.; Eberhart, M. E.; Alexandrova, A. N. Quantified Electrostatic Preorganization in Enzymes Using the Geometry of the Electron Charge Density. *Chem. Sci.* **2017**, *8* (7), 5010–5018. https://doi.org/10.1039/c7sc01301a.

(18) Kulik, H. J. Large-Scale QM/MM Free Energy Simulations of Enzyme Catalysis Reveal the Influence of Charge Transfer. *Phys. Chem. Chem. Phys.* **2018**, *20* (31), 20650–20660. https://doi.org/10.1039/c8cp03871f.

(19) Alavi, F. S.; Gheidi, M.; Zahedi, M.; Safari, N.; Ryde, U. A Novel Mechanism of Heme Degradation to Biliverdin Studied by QM/MM and QM Calculations. *Dalt. Trans.* **2018**, *47* (25), 8283–8291. https://doi.org/10.1039/c8dt00064f.

(20) Hu, L.; Eliasson, J.; Heimdal, J.; Ryde, U. Do Quantum Mechanical Energies Calculated for Small Models of Protein-Active Sites Converge. *J. Phys. Chem. A* **2009**, *113* (43), 11793–11800. https://doi.org/10.1021/jp9029024.

(21) Rod, T. H.; Ryde, U. Quantum Mechanical Free Energy Barrier for an Enzymatic Reaction. *Phys. Rev. Lett.* **2005**, *94* (13), 1–4. https://doi.org/10.1103/PhysRevLett.94.138302.

(22) Rod, T. H.; Ryde, U. Accurate QM/MM Free Energy Calculations of Enzyme Reactions: Methylation by Catechol O-Methyltransferase. *J. Chem. Theory Comput.* **2005**, *1* (6), 1240–1251. https://doi.org/10.1021/ct0501102.

(23) Sharir-Ivry, A.; Varatharaj, R.; Shurki, A. Challenges within the Linear Response Approximation When Studying Enzyme Catalysis and Effects of Mutations. *J. Chem. Theory Comput.* **2015**, *11* (1), 293–302. https://doi.org/10.1021/ct500751f.

(24) Karelina, M.; Kulik, H. J. Systematic Quantum Mechanical Region Determination

in QM/MM Simulation. *J. Chem. Theory Comput.* **2017**, *13* (2), 563–576. https://doi.org/10.1021/acs.jctc.6b01049.

(25) Di Paola, L.; De Ruvo, M.; Paci, P.; Santoni, D.; Giuliani, A. Protein Contact Networks: An Emerging Paradigm in Chemistry. *Chem. Rev.* **2013**, *113* (3), 1598–1613. https://doi.org/10.1021/cr3002356.

(26) Doncheva, N. T.; Klein, K.; Domingues, F. S.; Albrecht, M. Analyzing and Visualizing Residue Networks of Protein Structures. *Trends Biochem. Sci.* **2011**, *36* (4), 179–182. https://doi.org/10.1016/j.tibs.2011.01.002.

(27) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. Visualizing and Quantifying Molecular Goodness-of-Fit: Small-Probe Contact Dots with Explicit Hydrogen Atoms. *J. Mol. Biol.* **1999**, *285* (4), 1711–1733. https://doi.org/10.1006/jmbi.1998.2400.

(28) Safro, M. G.; Moor, N. A. Codases: 50 Years After. *Mol. Biol.* **2009**, *43* (2), 211–222. https://doi.org/10.1134/S0026893309020046.

(29) Gottlieb, A.; Frenkel-Morgenstern, M.; Safro, M.; Horn, D. Common Peptides Study of Aminoacyl-TRNA Synthetases. *PLoS One* **2011**, *6* (5), e20361. https://doi.org/10.1371/journal.pone.0020361.

(30) Hussain, T.; Kamarthapu, V.; Kruparani, S. P.; Deshmukh, M. V; Sankaranarayanan, R. Mechanistic Insights into Cognate Substrate Discrimination during Proofreading in Translation. *Proc Natl Acad Sci U. S. A.* **2010**, *107* (51), 22117–22121. https://doi.org/1014299107 [pii]\r10.1073/pnas.1014299107.

(31) Malde, A. K.; Mark, A. E. Binding and Enantiomeric Selectivity of Threonyl-TRNA Synthetase. *J. Am. Chem. Soc.* **2009**, *131* (11), 3848–3849. https://doi.org/10.1021/ja9002124.

(32) Bushnell, E. A. C.; Huang, W.; Llano, J.; Gauld, J. W. Molecular Dynamics Investigation into Substrate Binding and Identity of the Catalytic Base in the Mechanism of Threonyl-TRNA Synthetase. *J. Phys. Chem. B* **2012**, *116* (17), 5205–5212. https://doi.org/10.1021/jp302556e.

(33) Pearson, A. D.; Mills, J. H.; Song, Y.; Nasertorabi, F.; Han, G. W.; Baker, D.; Stevens, R. C.; Schultz, P. G. Trapping a Transition State in a Computationally Designed Protein Bottle. *Science (80-.).* **2015**, *347* (6224), 863–867. https://doi.org/10.1126/science.aaa2424.

(34) Dwivedi, S.; Kruparani, S. P.; Sankaranarayanan, R. A D-Amino Acid Editing Module Coupled to the Translational Apparatus in Archaea. *Nat. Struct. Mol. Biol.* **2005**, *12* (6), 556–557. https://doi.org/10.1038/nsmb943.

(35) Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Röthlisberger, D.; Baker, D. New Algorithms and an in Silico Benchmark for Computational Enzyme Design. *Protein Sci.* **2006**, *15* (12), 2785–2794. https://doi.org/10.1110/ps.062353106.

(36) Kuhlman, B.; Baker, D. Native Protein Sequences Are Close to Optimal for Their

Structures. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (19), 10383–10388. https://doi.org/10.1073/PNAS.97.19.10383.

(37) Almenningen, A.; Bastiansen, O.; Fernholt, L.; Cyvin, B. N.; Cyvin, S. J.; Samdal, S. Structure and Barrier of Internal Rotation of Biphenyl Derivatives in the Gaseous State. Part 1. The Molecular Structure and Normal Coordinate Analysis of Normal Biphenyl and Pedeuterated Biphenyl. *J. Mol. Struct.* **1985**, *128* (1–3), 59–76. https://doi.org/10.1016/0022-2860(85)85041-9.

(38) Poater, J.; Solà, M.; Bickelhaupt, F. M. Hydrogen-Hydrogen Bonding in Planar Biphenyl, Predicted by Atoms-in-Molecules Theory, Does Not Exist. *Chem. Eur. J.* **2006**, *12* (10), 2889–2895. https://doi.org/10.1002/chem.200500850.

(39) Wu, J. I.-C.; Schleyer, P. von R. Hyperconjugation in Hydrocarbons: Not Just a “Mild Sort of Conjugation.” *Pure Appl. Chem.* **2013**, *85* (5), 921–940. https://doi.org/10.1351/pac-con-13-01-03.

(40) Matta, C. F.; Hernández-Trujillo, J.; Tang, T.-H.; Bader, R. F. W. Hydrogen–Hydrogen Bonding: A Stabilizing Interaction in Molecules and Crystals. *Chem. Eur. J.* **2003**, *9* (9), 1940–1951. https://doi.org/10.1002/chem.200204626.

(41) Jenkins, S.; Maza, J. R.; Xu, T.; Jiajun, D.; Kirk, S. R. Biphenyl: A Stress Tensor and Vector-Based Perspective Explored within the Quantum Theory of Atoms in Molecules. *Int. J. Quantum Chem.* **2015**, *115* (23), 1678–1690. https://doi.org/10.1002/qua.25006.

(42) Gómez-Gallego, M.; Martín-Ortiz, M.; Sierra, M. A. Concerning the Electronic Control of Torsion Angles in Biphenyls. *European J. Org. Chem.* **2011**, *2011* (32), 6502–6506. https://doi.org/10.1002/ejoc.201100874.

(43) Esteruelas, M. A.; Fernández, I.; Herrera, A.; Martín-Ortiz, M.; Martínez-Álvarez, R.; Oliván, M.; Oñate, E.; Sierra, M. A.; Valencia, M. Multiple C−H Bond Activation of Phenyl-Substituted Pyrimidines and Triazines Promoted by an Osmium Polyhydride: Formation of Osmapolycycles with Three, Five, and Eight Fused Rings. *Organometallics* **2010**, *29* (4), 976–986. https://doi.org/10.1021/om901030q.

(44) Kong, D.-D.; Xue, L.-S.; Jang, R.; Liu, B.; Meng, X.-G.; Jin, S.; Ou, Y.-P.; Hao, X.; Liu, S.-H. Conformational Tuning of the Intramolecular Electronic Coupling in Molecular-Wire Biruthenium Complexes Bridged by Biphenyl Derivatives. *Chem. - A Eur. J.* **2015**, *21* (27), 9895–9904. https://doi.org/10.1002/chem.201500509.

(45) Himo, F. Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *J. Am. Chem. Soc.* **2017**, *139* (20), 6780–6786. https://doi.org/10.1021/jacs.7b02671.

(46) Siegbahn, P. E. M.; Himo, F. Recent Developments of the Quantum Chemical Cluster Approach for Modeling Enzyme Reactions. *J. Biol. Inorg. Chem.* **2009**, *14* (5), 643–651. https://doi.org/10.1007/s00775-009-0511-y.

(47) Blomberg, M. R. A. How Quantum Chemistry Can Solve Fundamental Problems

in Bioenergetics. *Int. J. Quantum Chem.* **2015**, *115* (18), 1197–1201. https://doi.org/10.1002/qua.24868.

(48) Siegbahn, P. E. M.; Blomberg, M. R. A. Quantum Chemical Studies of Proton-Coupled Electron Transfer in Metalloenzymes. *Chem. Rev.* **2010**, *110* (12), 7040–7061. https://doi.org/10.1021/cr100070p.

(49) Siegbahn, P. E. M.; Blomberg, M. R. A. Transition-Metal Systems in Biochemistry Studied by High-Accuracy Quantum Chemical Methods. *Chem. Rev.* **2000**, *100* (2), 421–438. https://doi.org/10.1021/cr980390w.

(50) Blomberg, M. R. A.; Borowski, T.; Himo, F.; Liao, R.; Siegbahn, P. E. M. Quantum Chemical Studies of Mechanisms for Metalloenzymes. *Chem. Rev.* **2014**, *114* (7), 3601–3658. https://doi.org/10.1021/cr400388t.

(51) Gherib, R.; Dokainish, H. M.; Gauld, J. W. Multi-Scale Computational Enzymology: Enhancing Our Understanding of Enzymatic Catalysis. *Int. J. Mol. Sci.* **2013**, *15* (1), 401–422. https://doi.org/10.3390/ijms15010401.

(52) Roos, G.; Geerlings, P.; Messens, J. Enzymatic Catalysis: The Emerging Role of Conceptual Density Functional Theory. *J. Phys. Chem. B* **2009**, *113* (41), 13465–13475. https://doi.org/10.1021/jp9034584.

(53) Van Der Kamp, M. W.; Mulholland, A. J. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013**, *52* (16), 2708–2728. https://doi.org/10.1021/bi400215w.

(54) Gao, J.; Ma, S.; Major, D. T.; Nam, K.; Pu, J.; Truhlar, D. G. Mechanisms and Free Energies of Enzymatic Reactions. *Chem. Rev.* **2006**, *106* (8), 3188–3209. https://doi.org/10.1021/cr050293k.

(55) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. Theoretical Perspective on the Structure and Mechanism of Cytochrome P450 Enzymes. *Chemical Reviews*. **2005**, *105* (6), 2279–2328. https://doi.org/10.1021/cr030722j.

(56) Cole, D. J.; Hine, N. D. M. Applications of Large-Scale Density Functional Theory in Biology. *J. Phys. Condens. Matter* **2016**, *28* (39), 393001. https://doi.org/10.1088/0953-8984/28/39/393001.

(57) DeYonker, N. J.; Webster, C. E. Phosphoryl Transfers of the Phospholipase D Superfamily: A Quantum Mechanical Theoretical Study. *J. Am. Chem. Soc.* **2013**, *135* (37), 13764–13774. https://doi.org/10.1021/ja4042753.

(58) DeYonker, N. J.; Webster, C. E. A Theoretical Study of Phosphoryl Transfers of Tyrosyl-DNA Phosphodiesterase i (Tdp1) and the Possibility of a “Dead-End” Phosphohistidine Intermediate. *Biochemistry* **2015**, *54* (27), 4236–4247. https://doi.org/10.1021/acs.biochem.5b00396.

(59) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-Chain Amide Orientation. *J. Mol. Biol.* **1999**, *285* (4), 1735–1747.

https://doi.org/10.1006/jmbi.1998.2401.

(60) Doncheva, N. T.; Assenov, Y.; Domingues, F. S.; Albrecht, M. Topological Analysis and Interactive Visualization of Biological Networks and Protein Structures. *Nat. Protoc.* **2012**, *7* (4), 670–685. https://doi.org/10.1038/nprot.2012.004.

(61) Griffin, J. L.; Bowler, M. W.; Baxter, N. J.; Leigh, K. N.; Dannatt, H. R. W.; Hounslow, A. M.; Blackburn, G. M.; Webster, C. E.; Cliff, M. J.; Waltho, J. P. Near Attack Conformers Dominate -Phosphoglucomutase Complexes Where Geometry and Charge Distribution Reflect Those of Substrate. *Proc. Natl. Acad. Sci.* **2012**, *109* (18), 6910–6915. https://doi.org/10.1073/pnas.1116855109.

(62) *Gaussian16 (Revision B.01).*; Gaussian, Inc.: Wallingford CT, 2016. https://gaussian.com/gaussian16/ (accessed 2021-06-16).

(63) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648. https://doi.org/10.1063/1.464913.

(64) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789. https://doi.org/10.1103/PhysRevB.37.785.

(65) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28* (3), 213–222. https://doi.org/10.1007/BF00533485.

(66) Petersson, G. A.; Al-Laham, M. A. A Complete Basis Set Model Chemistry. II. Open-shell Systems and the Total Energies of the First-row Atoms. *J. Chem. Phys.* **1991**, *94* (9), 6081–6090. https://doi.org/10.1063/1.460447.

(67) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56* (5), 2257–2261. https://doi.org/10.1063/1.1677527.

(68) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102* (11), 1995–2001. https://doi.org/10.1021/jp9716997.

(69) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, Structures, and Electronic Properties of Molecules in Solution with the C-PCM Solvation Model. *J. Comput. Chem.* **2003**, *24* (6), 669–681. https://doi.org/10.1002/jcc.10189.

(70) 1. Case, D.; Berryman, J.; Betz, R.; Cerutti, D.; T.E. Cheatham, I.; Darden, T. et al. AMBER 14. **2014**.

(71) dos Reis, M. A.; Aparicio, R.; Zhang, Y. Improving Protein Template Recognition by Using Small-Angle X-Ray Scattering Profiles. *Biophys. J.* **2011**, *101* (11), 2770–2781. https://doi.org/10.1016/j.bpj.2011.10.046.

(72) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L.

Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(73) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041. https://doi.org/10.1021/jp8001614.

(74) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Molec. Graphics* **1996,** *14* (1), 33-38.

(75) Grein, F. Twist Angles and Rotational Energy Barriers of Biphenyl and Substituted Biphenyls. *J. Phys. Chem. A* **2002**, *106* (15), 3823–3827. https://doi.org/10.1021/jp0122124.

(76) Grein, F. New Theoretical Studies on the Dihedral Angle and Energy Barriers of Biphenyl. *J. Mol. Struct.* **2003**, *624*, 23–28. https://doi.org/10.1016/S0166-1280(02)00590-0.

(77) Johansson, M. P.; Olsen, J. Torsional Barriers and Equilibrium Angle of Biphenyl: Reconciling Theory with Experiment. *J. Chem. Theory Comput.* **2008**, *4* (9), 1460–1471. https://doi.org/10.1021/ct800182e.

(78) Masson, E. Torsional Barriers of Substituted Biphenyls Calculated Using Density Functional Theory: A Benchmarking Study. *Org. Biomol. Chem.* **2013**, *11* (17), 2859. https://doi.org/10.1039/c3ob26704k.

(79) Xiao, H.; Schultz, P. G. At the Interface of Chemical and Biological Synthesis: An Expanded Genetic Code. *Cold Spring Harb. Perspect. Biol.* **2016**, *8* (9), a023945. https://doi.org/10.1101/cshperspect.a023945.

(80) Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. Counting the Zinc-Proteins Encoded in the Human Genome. *J. Proteome Res.* **2006**, *5* (1), 196–201. https://doi.org/10.1021/pr050361j.

(81) Lindskog, S. Structure and Mechanism of Carbonic Anhydrase. *Pharmacol. Ther.* **1997**, *74* (1), 1–20. https://doi.org/10.1016/S0163-7258(96)00198-2.

(82) Krishnamurthy, V. M.; Kaufman, G. K.; Urbach, A. R.; Gitlin, I.; Gudiksen, K. L.; Weibel, D. B.; Whitesides, G. M. Carbonic Anhydrase as a Model for Biophysical and Physical-Organic Studies of Proteins and Protein-Ligand Binding. *Chem. Rev.* **2008**, *108* (3), 946–1051. https://doi.org/10.1021/cr050262p.

(83) Hurst, T. K.; Wang, D.; Thompson, R. B.; Fierke, C. A. Carbonic Anhydrase II-Based Metal Ion Sensing: Advances and New Perspectives. *Biochim Biophys Acta Proteins Proteom*. **2010**, *1804* (2) 393–403. https://doi.org/10.1016/j.bbapap.2009.09.031.

(84) Lionetto, M. G.; Caricato, R.; Giordano, M. E.; Schettino, T. The Complex Relationship between Metals and Carbonic Anhydrase: New Insights and Perspectives. *Int. J. Mol. Sci*. **2016**, *17* (1). https://doi.org/10.3390/ijms17010127.

(85) Kogut, K. A.; Rowlett, R. S. A Comparison of the Mechanisms of CO2 Hydration

by Native and Co2+-Substituted Carbonic Anhydrase II. *J. Biol. Chem.* **1987**, *262* (34), 16417–16424. https://doi.org/10.1016/s0021-9258(18)49272-1.

(86) Lindskog, S.; Nyman, P. O. Metal-Binding Properties of Human Erythrocyte Carbonic Anhydrases. *BBA - Enzymol. Subj.* **1964**, *85* (3), 462–474. https://doi.org/10.1016/0926-6569(64)90310-4.

(87) Tripp, B. C.; Bell, C. B.; Cruz, F.; Krebs, C.; Ferry, J. G. A Role for Iron in an Ancient Carbonic Anhydrase. *J. Biol. Chem.* **2004**, *279* (8), 6683–6687. https://doi.org/10.1074/jbc.M311648200.

(88) Miscione, G. Pietro; Stenta, M.; Spinelli, D.; Anders, E.; Bottoni, A. New Computational Evidence for the Catalytic Mechanism of Carbonic Anhydrase. *Theor. Chem. Acc.* **2007**, *118* (1), 193–201. https://doi.org/10.1007/s00214-007-0274-x.

(89) Bottoni, A.; Lanza, C. Z.; Miscione, G. Pietro; Spinelli, D. New Model for a Theoretical Density Functional Theory Investigation of the Mechanism of the Carbonic Anhydrase: How Does the Internal Bicarbonate Rearrangement Occur? *J. Am. Chem. Soc.* **2004**, *126* (5), 1542–1550. https://doi.org/10.1021/ja030336j.

(90) Piazzetta, P.; Marino, T.; Russo, N.; Salahub, D. R. The Role of Metal Substitution in the Promiscuity of Natural and Artificial Carbonic Anhydrases. *Coord. Chem. Rev*. **2017**, *345* (15) 73–85. https://doi.org/10.1016/j.ccr.2016.12.014.

(91) Marino, T.; Russo, N.; Toscano, M. A Comparative Study of the Catalytic Mechanisms of the Zinc and Cadmium Containing Carbonic Anhydrase. *J. Am. Chem. Soc.* **2005**, *127* (12), 4242–4253. https://doi.org/10.1021/ja045546q.

(92) Bertini, I.; Luchinat, C. Cobalt(II) as a Probe of the Structure and Function of Carbonic Anhydrase. *Acc. Chem. Res.* **1983**, *16* (8), 272–279. https://doi.org/10.1021/ar00092a002.

(93) Silverman, D. N.; Lindskog, S. The Catalytic Mechanism of Carbonic Anhydrase: Implications of a Rate-Limiting Protolysis of Water. *Acc. Chem. Res.* **1988**, *21* (1), 30–36. https://doi.org/10.1021/ar00145a005.

(94) Woolley, P. Models for Metal Ion Function in Carbonic Anhydrase. *Nature* **1975**, *258* (5537), 677–682. https://doi.org/10.1038/258677a0.

(95) Håkansson, K.; Carlsson, M.; Svensson, L. A.; Liljas, A. Structure of Native and Apo Carbonic Anhydrase II and Structure of Some of Its Anion-Ligand Complexes. *J. Mol. Biol.* **1992**, *227* (4), 1192–1204. https://doi.org/10.1016/0022-2836(92)90531-N.

(96) Liang, J. Y.; Lipscomb, W. N. Theoretical Study of the Uncatalyzed Hydration of Carbon Dioxide in the Gas Phase. *J. Am. Chem. Soc.* **1986**, *108* (17), 5051–5058. https://doi.org/10.1021/ja00277a001.

(97) Merz, K. M.; Hoffmann, R.; Dewar, M. J. S. Mode of Action of Carbonic Anhydrase. *J. Am. Chem. Soc.* **1989**, *111* (15), 5636–5649. https://doi.org/10.1021/ja00197a021.

(98) Liang, J. -Y; Lipscomb, W. N. Theoretical Study of Carbonic Anhydrase-catalyzed Hydration of Co2: A Brief Review. *Int. J. Quantum Chem.* **1989**, *36* (3), 299–312. https://doi.org/10.1002/qua.560360313.

(99) Solá, M.; Lledós, A.; Duran, M.; Bertrán, J. Ab Initio Study of the Hydration of CO2 by Carbonic Anhydrase. A Comparison between the Lipscomb and Lindskog Mechanisms. *J. Am. Chem. Soc.* **1992**, *114* (3), 869–877. https://doi.org/10.1021/ja00029a010.

(100) Bräuer, M.; Pérez-Lustres, J. L.; Weston, J.; Anders, E. Quantitative Reactivity Model for the Hydration of Carbon Dioxide by Biomimetic Zinc Complexes. *Inorg. Chem.* **2002**, *41* (6), 1454–1463. https://doi.org/10.1021/ic0010510.

(101) Avvaru, B. S.; Arenas, D. J.; Tu, C.; Tanner, D. B.; McKenna, R.; Silverman, D. N. Comparison of Solution and Crystal Properties of Co(II)-Substituted Human Carbonic Anhydrase II. *Arch. Biochem. Biophys.* **2010**, *502* (1), 53–59. https://doi.org/10.1016/j.abb.2010.07.010.

(102) Domsic, J. F.; Avvaru, B. S.; Kim, C. U.; Gruner, S. M.; Agbandje-Mckenna, M.; Silverman, D. N.; Mckenna, R. Entrapment of Carbon Dioxide in the Active Site of Carbonic Anhydrase II□. *Publ. JBC Pap. Press* **2008**. https://doi.org/10.1074/jbc.M805353200.

(103) Håkansson, K.; Wehnert, A. Structure of Cobalt Carbonic Anhydrase Complexed with Bicarbonate. *J. Mol. Biol.* **1992**, *228* (4), 1212–1218. https://doi.org/10.1016/0022-2836(92)90327-G.

(104) Wadt, W. R.; Hay, P. J. Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for Main Group Elements Na to Bi. *J. Chem. Phys.* **1985**, *82* (1), 284–298. https://doi.org/10.1063/1.448800.

(105) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465. https://doi.org/10.1002/jcc.21759.

(106) Piazzetta, P.; Marino, T.; Russo, N. Promiscuous Ability of Human Carbonic Anhydrase: QM and QM/MM Investigation of Carbon Dioxide and Carbodiimide Hydration. *Inorg. Chem.* **2014**, *53* (7), 3488–3493. https://doi.org/10.1021/ic402932y.

(107) Kim, J. K.; Lee, C.; Lim, S. W.; Adhikari, A.; Andring, J. T.; Mckenna, R.; Ghim, C.-M.; Chae, &; Kim, U. Elucidating the Role of Metal Ions in Carbonic Anhydrase Catalysis. *Nat. Commun.* **2020**, *11* (4557). https://doi.org/10.1038/s41467-020-18425-5.

(108) Hakansson, K.; Wehnert, A.; Liljas, A. X-Ray Analysis of Metal-Substituted Human Carbonic Anhydrase II Derivatives. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1994**, *50* (1), 93–100. https://doi.org/10.1107/S0907444993008790.

(109) Cui, Q.; Karplus, M. Is a “Proton Wire” Concerted or Stepwise? A Model Study of Proton Transfer in Carbonic Anhydrase. *J. Phys. Chem. B* **2003**, *107* (4), 1071–

1078. https://doi.org/10.1021/jp021931v.

(110) Mikulski, R.; West, D.; Sippel, K. H.; Avvaru, B. S.; Aggarwal, M.; Tu, C.; McKenna, R.; Silverman, D. N. Water Networks in Fast Proton Transfer during Catalysis by Human Carbonic Anhydrase II. *Biochemistry* **2013**, *52* (1), 125–131. https://doi.org/10.1021/bi301099k.

(111) Jiao, D.; Rempe, S. B. Combined Density Functional Theory (DFT) and Continuum Calculations of p K a in Carbonic Anhydrase. *Biochemistry* **2012**, *51* (30), 5979–5989. https://doi.org/10.1021/bi201771q.

(112) Hakkim, V.; Subramanian, V. Role of Second Coordination Sphere Amino Acid Residues on the Proton Transfer Mechanism of Human Carbonic Anhydrase II (HCA II). *J. Phys. Chem. A* **2010**, *114* (30), 7952–7959. https://doi.org/10.1021/jp101515h.

(113) Maupin, C. M.; McKenna, R.; Silverman, D. N.; Voth, G. A. Elucidation of the Proton Transport Mechanism in Human Carbonic Anhydrase II. *J. Am. Chem. Soc.* **2009**, *131* (22), 7598–7608. https://doi.org/10.1021/ja8091938.

(114) Kimblin, C.; Parkin, G. Comparison of Zinc and Cadmium Coordination Environments in Synthetic Analogues of Carbonic Anhydrase: Synthesis and Structure of {[Pimpr1,But]Cd(OH2)(OClO3)}(ClO4). *Inorg. Chem.* **1996**, *35* (24), 6912–6913. https://doi.org/10.1021/ic961007d.

(115) Harris, T. V.; Szilagyi, R. K. Protein Environmental Effects on Iron-Sulfur Clusters a Set of Rules for Constructing Computational Models for Inner and Outer Coordination Spheres. *J. Comput. Chem.* **2016**, *37* (18), 1681–1696. https://doi.org/10.1002/jcc.24384.

(116) Zheng, M.; Waller, M. P. Yoink: An Interaction-Based Partitioning API. *J. Comput. Chem.* **2018**, *39* (13), 799–806. https://doi.org/10.1002/jcc.25146.

(117) National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*; Washington, DC, 2019. https://doi.org/10.17226/25303.

(118) Kanaan, N.; Ruiz Pernía, J. J.; Williams, I. H. QM/MM Simulations for Methyl Transfer in Solution and Catalysed by COMT: Ensemble-Averaging of Kinetic Isotope Effects. *Chem. Commun.* **2008**, No. 46, 6114–6116. https://doi.org/10.1039/b814212b.

(119) Rod, T. H.; Rydberg, P.; Ryde, U. Implicit versus Explicit Solvent in Free Energy Calculations of Enzyme Catalysis: Methyl Transfer Catalyzed by Catechol O - Methyltransferase. *J. Chem. Phys.* **2006**, *124* (17), 174503. https://doi.org/10.1063/1.2186635.

(120) Roca, M.; Moliner, V.; Ruiz-Pernía, J. J.; Silla, E.; Tuñón, I. Activation Free Energy of Catechol O-Methyltransferase. Corrections to the Potential of Mean Force. *J. Phys. Chem. A* **2006**, *110* (2), 503–509. https://doi.org/10.1021/jp0520953.

(121) Hatstat, A. K.; Morris, M.; Peterson, L. W.; Cafiero, M. Ab Initio Study of

Electronic Interaction Energies and Desolvation Energies for Dopaminergic Ligands in the Catechol-O-Methyltransferase Active Site. *Comput. Theor. Chem.* **2016**, *1078*, 146–162. https://doi.org/10.1016/j.comptc.2016.01.003.

(122) Yang, Z.; Mehmood, R.; Wang, M.; Qi, H. W.; Steeves, A. H.; Kulik, H. J. Revealing Quantum Mechanical Effects in Enzyme Catalysis with Large-Scale Electronic Structure Simulation. *React. Chem. Eng.* **2019**, *4* (2), 298–315. https://doi.org/10.1039/c8re00213d.

(123) Roca, M.; Martí, S.; Andrés, J.; Moliner, V.; Tuñón, I.; Bertrán, J.; Williams, I. H. Theoretical Modeling of Enzyme Catalytic Power: Analysis of “Cratic” and Electrostatic Factors in Catechol O-Methyltransferase. *J. Am. Chem. Soc.* **2003**, *125* (25), 7726–7737. https://doi.org/10.1021/ja0299497.

(124) Roca, M.; Andrés, J.; Moliner, V.; Tuñón, I.; Bertrán, J. On the Nature of the Transition State in Catechol O-Methyltransferase. A Complementary Study Based on Molecular Dynamics and Potential Energy Surface Explorations. *J. Am. Chem. Soc.* **2005**, *127* (30), 10648–10655. https://doi.org/10.1021/ja051503d.

(125) García-Meseguer, R.; Zinovjev, K.; Roca, M.; Ruiz-Pernía, J. J.; Tuñón, I. Linking Electrostatic Effects and Protein Motions in Enzymatic Catalysis. A Theoretical Analysis of Catechol O-Methyltransferase. *J. Phys. Chem. B* **2015**, *119* (3), 873–882. https://doi.org/10.1021/jp505746x.

(126) Chen, X.; Schwartz, S. D. Examining the Origin of Catalytic Power of Catechol O-Methyltransferase. *ACS Catal.* **2019**, *9* (11), 9870–9879. https://doi.org/10.1021/acscatal.9b02657.

(127) Patra, N.; Ioannidis, E. I.; Kulik, H. J. Computational Investigation of the Interplay of Substrate Positioning and Reactivity in Catechol O-Methyltransferase. *PLoS One* **2016**, *11* (8), e0161868. https://doi.org/10.1371/journal.pone.0161868.

(128) Lameira, J.; Bora, R. P.; Chu, Z. T.; Warshel, A. Methyltransferases Do Not Work by Compression, Cratic, or Desolvation Effects, but by Electrostatic Preorganization. *Proteins Struct. Funct. Bioinforma.* **2015**, *83* (2), 318–330. https://doi.org/10.1002/prot.24717.

(129) Roca, M.; Moliner, V.; Tuñón, I.; Hynes, J. T. Coupling between Protein and Reaction Dynamics in Enzymatic Processes: Application of Grote-Hynes Theory to Catechol O-Methyltransferase. *J. Am. Chem. Soc.* **2006**, *128* (18), 6186–6193. https://doi.org/10.1021/ja058826u.

(130) Saez, D. A.; Zinovjev, K.; Tuñón, I.; Vöhringer-Martinez, E. Catalytic Reaction Mechanism in Native and Mutant Catechol- O-Methyltransferase from the Adaptive String Method and Mean Reaction Force Analysis. *J. Phys. Chem. B* **2018**, *122* (38), 8861–8871. https://doi.org/10.1021/acs.jpcb.8b07339.

(131) Jindal, G.; Warshel, A. Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region. *J. Phys. Chem. B* **2016**, *120* (37), 9913–9921. https://doi.org/10.1021/acs.jpcb.6b07203.

(132) Ruggiero, G. D.; Williams, I. H.; Roca, M.; Moliner, V.; Tuñón, I. QM/MM Determination of Kinetic Isotope Effects for COMT-Catalyzed Methyl Transfer Does Not Support Compression Hypothesis. *J. Am. Chem. Soc.* **2004**, *126* (28), 8634–8635. https://doi.org/10.1021/ja048055e.

(133) Kuhn, B.; Kollman, P. A. QM-FE and Molecular Dynamics Calculations on Catechol O- Methyltransferase: Free Energy of Activation in the Enzyme and in Aqueous Solution and Regioselectivity of the Enzyme-Catalyzed Reaction. *J. Am. Chem. Soc.* **2000**, *122* (11), 2586–2596. https://doi.org/10.1021/ja992218v.

(134) Zhang, J.; Klinman, J. P. Enzymatic Methyl Transfer: Role of an Active Site Residue in Generating Active Site Compaction That Correlates with Catalytic Efficiency. *J. Am. Chem. Soc.* **2011**, *133* (43), 17134–17137. https://doi.org/10.1021/ja207467d.

(135) Lautala, P. Ulmanen, I; Taskinen, J. Molecular Mechanisms Controlling the Rate and Specificity of Catechol O-Methylation by Human Soluble Catechol O-Methyltransferase. *Mol Pharmacol*. **2001**. *59* (2), 393-402. https://doi.org/10.1124/mol.59.2.393.

(136) Yan, W.; Zhou, J.; Sun, M.; Chen, J.; Hu, G.; Shen, B. The Construction of an Amino Acid Network for Understanding Protein Structure and Function. *Amino Acids*. **2014**, *46* (6), 1419–1439. https://doi.org/10.1007/s00726-014-1710-6.

(137) Rutherford, K.; Le Trong, I.; Stenkamp, R. E.; Parson, W. W. Crystal Structures of Human 108V and 108M Catechol O-Methyltransferase. *J. Mol. Biol.* **2008**, *380* (1), 120–130. https://doi.org/10.1016/j.jmb.2008.04.040.

(138) *The {PyMOL} Molecular Graphics System, Version2.3*; Schrödinger, LLC: New York, 2015. https://pymol.org/2/ (accessed 2021-06-16).

(139) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979**, *28* (1), 100. https://doi.org/10.2307/2346830.

(140) *RStudio: Integrated Development Environment for R*. RStudio Team: Boston, MA 2019. https://www.rstudio.com/ (accessed 2021-06-16).

(141) *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. Kassambara, A. and Mundt, F. 2020. https://rpkgs.datanovia.com/factoextra/index.html (accessed 2020-06-16).

(142) Tibshirani, R.; Walther, G.; Hastie, T. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2001**, *63* (2), 411–423. https://doi.org/10.1111/1467-9868.00293.

(143) Jubb, H. C.; Higueruelo, A. P.; Ochoa-Montaño, B.; Pitt, W. R.; Ascher, D. B.; Blundell, T. L. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.* **2017**, *429* (3), 365–371. https://doi.org/10.1016/j.jmb.2016.12.004.

(144) Sugiki, T.; Kobayashi, N.; Fujiwara, T. Modern Technologies of Solution Nuclear Magnetic Resonance Spectroscopy for Three-Dimensional Structure

Determination of Proteins Open Avenues for Life Scientists. *Comput. Struct. Biotechnol. J*. **2017**, *13* (15), 328–339. https://doi.org/10.1016/j.csbj.2017.04.001.

(145) Cornish, V. W.; Mendel, D.; Schultz, P. G. Probing Protein Structure and Function with an Expanded Genetic Code. *Angew. Chem. Int. Ed.* **1995**, *34* (6), 621–633. https://doi.org/10.1002/anie.199506211.

(146) Van Der Kamp, M. W.; Shaw, K. E.; Woods, C. J.; Mulholland, A. J. Biomolecular Simulation and Modelling: Status, Progress and Prospects. *J. R. Soc. Interface.* **2008**, *5* (3). 173-90. https://doi.org/10.1098/rsif.2008.0105.focus.

(147) Baruah, K.; Sinha, S.; Hazarika, S.; Bhattacharyya, P. K. QM/MM Studies on Cyclodextrin-Alcohol Interaction. *J. Macromol. Sci. Part A Pure Appl. Chem.* **2015**, *52* (1), 64–68. https://doi.org/10.1080/10601325.2014.976754.

(148) Brinda, K. V; Vishveshwara, S. A Network Representation of Protein Structures: Implications for Protein Stability. *Biophys. J.* **2005**, *89* (6), 4159–4170. https://doi.org/10.1529/biophysj.105.064485.

(149) Gromiha, M. M.; Selvaraj, S. Inter-Residue Interactions in Protein Folding and Stability. *Prog. Biophys. Mol. Biol*. **2004**, *86* (2), 235–277. https://doi.org/10.1016/j.pbiomolbio.2003.09.003.

(150) Daily, M. D.; Upadhyaya, T. J.; Gray, J. J. Contact Rearrangements Form Coupled Networks from Local Motions in Allosteric Proteins. *Proteins Struct. Funct. Genet.* **2008**, *71* (1), 455–466. https://doi.org/10.1002/prot.21800.

(151) Blacklock, K.; Verkhivker, G. M. Computational Modeling of Allosteric Regulation in the Hsp90 Chaperones: A Statistical Ensemble Analysis of Protein Structure Networks and Allosteric Communications. *PLoS Comput. Biol.* **2014**, *10* (6), e1003679. https://doi.org/10.1371/journal.pcbi.1003679.

(152) Böde, C.; Kovács, I. A.; Szalay, M. S.; Palotai, R.; Korcsmáros, T.; Csermely, P. Network Analysis of Protein Dynamics. *FEBS Letters*. **2007**, *581* (15), 2776–2782. https://doi.org/10.1016/j.febslet.2007.05.021.

(153) Bagler, G.; Sinha, S. Assortative Mixing in Protein Contact Networks and Protein Folding Kinetics. *Bioinformatics* **2007**, *23* (14), 1760–1767. https://doi.org/10.1093/bioinformatics/btm257.

(154) Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. Network Analysis of Protein Structures Identifies Functional Residues. *J. Mol. Biol.* **2004**, *344* (4), 1135–1146. https://doi.org/10.1016/j.jmb.2004.10.055.

(155) Ben-Shimon, A.; Eisenstein, M. Looking at Enzymes from the inside out: The Proximity of Catalytic Residues to the Molecular Centroid Can Be Used for Detection of Active Sites and Enzyme-Ligand Interfaces. *J. Mol. Biol.* **2005**, *351* (2), 309–326. https://doi.org/10.1016/j.jmb.2005.06.047.

(156) del Sol, A.; Fujihashi, H.; Amoros, D.; Nussinov, R. Residue Centrality, Functionally Important Residues, and Active Site Shape: Analysis of Enzyme and

Non-Enzyme Families. *Protein Sci.* **2006**, *15* (9), 2120–2128. https://doi.org/10.1110/ps.062249106.

(157) Slama, P.; Filippis, I.; Lappe, M. Detection of Protein Catalytic Residues at High Precision Using Local Network Properties. *BMC Bioinformatics* **2008**, *9* (1), 517. https://doi.org/10.1186/1471-2105-9-517.

(158) Du, S.; Sakurai, M. Multivariate Analysis of Properties of Amino Acid Residues in Proteins from a Viewpoint of Functional Site Prediction. *Chem. Phys. Lett.* **2010**, *488* (1–3), 81–85. https://doi.org/10.1016/j.cplett.2010.02.006.

(159) Fajardo, J. E.; Fiser, A. Protein Structure Based Prediction of Catalytic Residues. *BMC Bioinformatics* **2013**, *14* (1), 63. https://doi.org/10.1186/1471-2105-14-63.

(160) Nosrati, G. R.; Houk, K. N. Using Catalytic Atom Maps to Predict the Catalytic Functions Present in Enzyme Active Sites. *Biochemistry* **2012**, *51* (37), 7321–7329. https://doi.org/10.1021/bi3008438.

(161) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47* (2), 337–344. https://doi.org/10.1021/jm030331x.

(162) Waszkowycz, B.; Clark, D. E.; Gancia, E. Outstanding Challenges in Protein-Ligand Docking and Structure-Based Virtual Screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (2), 229–259. https://doi.org/10.1002/wcms.18.

(163) Sydow, D.; Burggraaff, L.; Szengel, A.; Van Vlijmen, H. W. T.; Ijzerman, A. P.; Van Westen, G. J. P.; Volkamer, A. Advances and Challenges in Computational Target Prediction. *J. Chem. Inf. Model.* **2019**, *59* (5), 1728–1742. https://doi.org/10.1021/acs.jcim.8b00832.

(164) Summers, T. J.; Cheng, Q.; Deyonker, N. J. A Transition State "Trapped"? QM-Cluster Models of Engineered Threonyl-TRNA Synthetase. *Org. Biomol. Chem.* **2018**, *16* (22), 4090–4100. https://doi.org/10.1039/c8ob00540k.

(165) Vijayabaskar, M. S.; Vishveshwara, S. Interaction Energy Based Protein Structure Networks. *Biophys. J.* **2010**, *99* (11), 3704–3715. https://doi.org/10.1016/j.bpj.2010.08.079.

(166) Serçinoğlu, O.; Ozbek, P. GRINN: A Tool for Calculation of Residue Interaction Energies and Protein Energy Network Analysis of Molecular Dynamics Simulations. *Nucleic Acids Res.* **2018**, *46* (W1), W554–W562. https://doi.org/10.1093/nar/gky381.

(167) Szalewicz, K. Symmetry-Adapted Perturbation Theory of Intermolecular Forces. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2* (2), 254–272. https://doi.org/10.1002/wcms.86.

(168) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. In *Structural Bioinformatics*; Narnia, 2005; Vol. 28, pp 181–198. https://doi.org/10.1002/0471721204.ch9.

(169) Faber, H. R.; Matthews, B. W. A Mutant T4 Lysozyme Displays Five Different Crystal Conformations. *Nature* **1990**, *348* (6298), 263–266. https://doi.org/10.1038/348263a0.

(170) Osawa, T.; Matsubara, Y.; Muramatsu, T.; Kimura, M.; Kakuta, Y. Crystal Structure of the Alginate (Poly α-L-Guluronate) Lyase from Corynebacterium Sp. at 1.2 Å Resolution. *J. Mol. Biol.* **2005**, *345* (5), 1111–1118. https://doi.org/10.1016/j.jmb.2004.10.081.

(171) Li, Z.; Li, H.; Devasahayam, G.; Gemmill, T.; Chaturvedi, V.; Hanes, S. D.; Van Roey, P. The Structure of the Candida Albicans Ess1 Prolyl Isomerase Reveals a Well-Ordered Linker That Restricts Domain Mobility. *Biochemistry* **2005**, *44* (16), 6180–6189. https://doi.org/10.1021/bi050115l.

(172) Shaw, A.; Saldajeno, M. L.; Kolkman, M. A. B.; Jones, B. E.; Bott, R. Structure Determination and Analysis of a Bacterial Chymotrypsin from Cellulomonas Bogoriensis. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2007**, *63* (4), 266–269. https://doi.org/10.1107/S1744309107008937.

(173) Sone, T.; Haraguchi, Y.; Kuwahara, A.; Ose, T.; Takano, M.; Abe, A.; Tanaka, M.; Tanaka, I.; Asano, K. Structural Characterization Reveals the Keratinolytic Activity of an Arthrobacter Nicotinovorans Protease. *Protein Pept. Lett.* **2015**, *22* (1), 63–72. https://doi.org/10.2174/0929866521666140919100851.

(174) Dawson, N. L.; Lewis, T. E.; Das, S.; Lees, J. G.; Lee, D.; Ashford, P.; Orengo, C. A.; Sillitoe, I. CATH: An Expanded Resource to Predict Protein Function through Structure and Sequence. *Nucleic Acids Res.* **2017**, *45* (D1), D289–D295. https://doi.org/10.1093/nar/gkw1098.

(175) O'Rourke, K. F.; Gorman, S. D.; Boehr, D. D. Biophysical and Computational Methods to Analyze Amino Acid Interaction Networks in Proteins. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 245–251. https://doi.org/10.1016/j.csbj.2016.06.002.

(176) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504. https://doi.org/10.1101/gr.1239303.

(177) Parker, T. M.; Burns, L. A.; Parrish, R. M.; Ryno, A. G.; Sherrill, C. D. Levels of Symmetry Adapted Perturbation Theory (SAPT). I. Efficiency and Performance for Interaction Energies. *J. Chem. Phys.* **2014**, *140* (9), 094106. https://doi.org/10.1063/1.4867135.

(178) Jeziorski, B.; Moszynski, R.; Szalewicz, K. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van Der Waals Complexes. *Chem. Rev.* **1994**, *94* (7), 1887–1930. https://doi.org/10.1021/cr00031a008.

(179) Hohenstein, E. G.; Sherrill, C. D. Wavefunction Methods for Noncovalent Interactions. *Wiley Interdiscip. Rev. Comput. Mol. Sci*. **2012**, *2* (2), 304–326. https://doi.org/10.1002/wcms.84.

(180) Parrish, R. M.; Parker, T. M.; David Sherrill, C. Chemical Assignment of Symmetry-Adapted Perturbation Theory Interaction Energy Components: The Functional-Group SAPT Partition. *J. Chem. Theory Comput.* **2014**, *10* (10), 4417–4431. https://doi.org/10.1021/ct500724p.

(181) Parrish, R. M.; Gonthier, J. F.; Corminbœuf, C.; Sherrill, C. D. Communication: Practical Intramolecular Symmetry Adapted Perturbation Theory via Hartree-Fock Embedding. *J. Chem. Phys.* **2015**, *143* (5), 051103. https://doi.org/10.1063/1.4927575.

(182) Pastorczak, E.; Prlj, A.; Gonthier, J. F.; Corminboeuf, C. Intramolecular Symmetry-Adapted Perturbation Theory with a Single-Determinant Wavefunction. *J. Chem. Phys.* **2015**, *143* (22), 224107. https://doi.org/10.1063/1.4936830.

(183) Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G. A.; Vanommeslaeghe, K.; MacKerell, A. D.; Merz, K. M.; Sherrill, C. D. The BioFragment Database (BFDb): An Open-Data Platform for Computational Chemistry Analysis of Noncovalent Interactions. *J. Chem. Phys.* **2017**, *147* (16), 161727. https://doi.org/10.1063/1.5001028.

(184) M. Parrish, R.; A. Burns, L.; G. A. Smith, D.; C. Simmonett, A.; Eugene DePrince, A.; G. Hohenstein, E.; Bozkaya, U.; Yu. Sokolov, A.; Di Remigio, R.; M. Richard, R.; F. Gonthier, J.; M. James, A.; R. McAlexander, H.; Kumar, A.; Saitow, M.; Wang, X.; P. Pritchard, B.; Verma, P.; F. Schaefer, H.; Patkowski, K.; A. King, R.; F. Valeev, E.; A. Evangelista, F.; M. Turney, J.; Daniel Crawford, T.; David Sherrill, C. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**, *13* (7), 3185–3197. https://doi.org/10.1021/acs.jctc.7b00174.

(185) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2* (3), 18–22.

(186) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.

(187) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrášek, J. Representative Amino Acid Side Chain Interactions in Proteins. a Comparison of Highly Accurate Correlated Ab Initio Quantum Chemical and Empirical Potential Procedures. *J. Chem. Theory Comput.* **2009**, *5* (4), 982–992. https://doi.org/10.1021/ct800508v.

(188) Probst, P.; Boulesteix, A. L. To Tune or Not to Tune the Number of Trees in Random Forest. *J. Mach. Learn. Res.* **2018**, *18*, 1–8.

(189) Siegbahn, P. E. M.; Borowski, T. Modeling Enzymatic Reactions Involving Transition Metals. *Acc. Chem. Res.* **2006**, *39* (10), 729–738. https://doi.org/10.1021/ar050123u.

Appendices

Appendix A: Chapter 2 Supplementary Information

Cartesian coordinates of model structures are available at doi:10.1039/C8OB00540K

Potential Energy Scans

Cluster models of the torsional conformations were obtained by freezing C_α and select C_β atoms (Figure 4) to their x-ray crystallographic coordinates using Gaussian09 freeze codes, along with using generalized redundant internal coordinates to constrain the *p*-biphenyalanine C_β position, a select H-C_β-C_γ-H dihedral angle (Figure 5), and the two central C-C dihedral angles characteristic of the rotating biphenyl rings. To reproduce the constrained dihedral scans, the redundant internal coordinates below should be used with “opt(modred)” in Gaussian09:

```
115 114 121 120 F

113 114 121 122 F

268 31 111 112 F

111 F
```

Table 1. List of residues included in the QM cluster model. Total charge of model is neutral.

Species/residue label	Protonated R group?	Species/residue charge	Trim N-side	Trim R-group	Trim C-side
4S02					
Y^{10}	N/A	0	H	H	-
BIF^{11}	N/A	0	-	-	-
E^{12}	N	-1	-	-	-
Y^{13}	N/A	0	-	-	H
R^{34}	Y	+1	H	-	-
M^{35}	N/A	0	-	-	-
E^{36}	N/A	0	-	H	H
V^{38}	N/A	0	H	-	H
V^{40}	N/A	0	H	-	-
A^{41}	N/A	0	-	H	-
F^{77}	N/A	0	H	-	-
V^{78}	N/A	0	-	H	-
$Y^{79}A$	N/A	0	-	-	H
A^{115}	N/A	0	H	-	H
I^{121}	N/A	0	H	-	H
$F^{123}Y$	N/A	0	H	-	-
K^{124}	N/A	0	-	H	-
I^{125}	N/A	0	-	-	H
4S0J					
$F^{42}F$	N/A	0	-	-	H
$Y^{79}S$	N/A	0	-	-	H
$F^{123}V$	N/A	0	H	-	-
4S0L					
$F^{42}F$	N/A	0	-	-	H
$Y^{79}V$	N/A	0	-	-	H
W^{81}	N/A	0	H	-	H
$F^{123}V$	N/A	0	H	-	-
4S0I					
$F^{42}F$	N/A	0	-	-	H
$Y^{79}S$	N/A	0	-	-	H
$F^{123}A$	N/A	0	H	-	-
4S0I_W81					
$F^{42}F$	N/A	0	-	-	H
$Y^{79}S$	N/A	0	-	-	H
W^{81}	N/A	0	H	-	H
$F^{123}A$	N/A	0	H	-	-
4S0K					
$F^{42}F$	N/A	0	-	-	H
$Y^{79}V$	N/A	0	-	-	H
W^{81}	N/A	0	H	-	H

Table 1 (continued)

Species/residue label	Protonated R group?	Species/residue charge	Trim N-side	Trim R-group	Trim C-side
F^{123}A	N/A	0	H	-	-
4S03					
F^{42}F	N/A	0	-	-	H
Y^{79}I	N/A	0	-	-	H
F^{123}A	N/A	0	H	-	-
4S03_W81					
F^{42}F	N/A	0	-	-	H
Y^{79}I	N/A	0	-	-	H
W^{81}	N/A	0	H	-	H
F^{123}A	N/A	0	H	-	-

Table 2. Comparison of calculated Φ among constrained and unconstrained cluster models. Experimental Φ are from their respective PDB crystal structures.

Model	Conditions	Experimental Φ (degrees)	Φ at $dE/d\Phi = 0$ (degrees)	Unconstrained model Φ (degrees)	Relaxation Energy (kcal/mol)
4S02	Gas	26	28.4	29.0	0.84
	Gas+GD3BJ	26	18.8	31.1	2.2
	CPCM	26	25.3	24.9	0.83
	CPCM+GD3BJ	26	24.8	31.9	0.90
4S0J	CPCM+GD3BJ	35	32.1	33.2	0.38
4S0L	CPCM+GD3BJ	21	25.2	32.0	2.3
4S0I	CPCM+GD3BJ	15	28.0	27.4	0.92
4S0I_W81	CPCM+GD3BJ	15	28.7	29.6	1.4
4S0K	CPCM+GD3BJ	20	23.0	26.7	0.39
4S03	Gas	0	2.3	-27.5	1.4
	Gas+GD3BJ	0	-0.4	9.4	0.53
	CPCM	0	2.8	-28.9	2.4
	CPCM+GD3BJ	0	-2.1	-6.7	0.33
4S03_W81	Gas	0	10.5	-26.7	1.9
	Gas+GD3BJ	0	-7.6	-13.1	0.33
	CPCM	0	10.5	-28.7	2.5
	CPCM+GD3BJ	0	-2.5	-10.9	0.48

Table 3. Thermal flexibility of biphenyl within the protein clusters at 310K, and root mean square deviation (RMSD) values between the trimmed x-ray crystal structure and its respective optimized unconstrained model.

Model	Conditions	Thermally Allowed Displacement from Φ_{min} (degrees)		Thermal Range (degrees)	RMSD of Cluster Model (Angstroms)	Resolution of Crystallized Enzyme (Angstroms)
4S02	Gas	-12.6	+10.5	23.1	0.73	1.95
	Gas+GD3BJ	-15.0	+15.0	30.0	0.87	1.95
	CPCM	-12.4	+9.8	22.2	0.95	1.95
	CPCM+GD3BJ	-13.0	+10.5	23.5	0.78	1.95
4S0J	CPCM+GD3BJ	-8.6	+8.7	17.3	0.47	2.1
4S0L	CPCM+GD3BJ	-10.1	+9.8	19.9	0.87	2.5
4S0I	CPCM+GD3BJ	-12.1	+10.2	22.3	0.57	2.36
4S0I_W81	CPCM+GD3BJ	-10.4	+9.6	20.0	0.53	2.36
4S0K	CPCM+GD3BJ	-16.2	+9.9	26.1	0.53	2.1
4S03	Gas	-16.3	+19.7	36.0	0.62	2.05
	Gas+GD3BJ	-12.6	+15.2	27.8	0.52	2.05
	CPCM	-14.9	+19.8	34.7	0.58	2.05
	CPCM+GD3BJ	-12.1	+13.3	25.4	0.37	2.05
4S03_W81	Gas	-20.0	+14.4	34.4	0.78	2.05
	Gas+GD3BJ	-10.4	+15.7	26.1	0.66	2.05
	CPCM	-18.8	+15.8	34.6	0.85	2.05
	CPCM+GD3BJ	-11.5	+14.0	25.5	0.47	2.05

Table 4. Average RMSD calculations between the original PDB crystal structure and structures from the MD simulation

Enzyme	RMSD of Entire Enzyme (Å)	RMSD of Cluster Model Residues (Å)
4S02	0.52	0.30
4S0J	0.35	0.31
4s0L	0.36	0.30
4S0I	0.37	0.31
4S0K	0.36	0.30
4S03	0.33	0.28

BiPhe-Water Distances

The MD simulations reinforce the expectation that the BiPhe residue has no significant interaction with the solvent, as anticipated by the hydrophobic residues surrounding the side chain and the absence of nearby waters in the crystal structure. The distance between the oxygen of the nearest water to the nearest atom of BiPhe (which is consistently the solvent-exposed atom H9 or H10, two hydrogens on the terminal ring of BiPhe) was measured and averaged for each of the MD simulations. The average of the shortest distance ranges from 2.85 Å (s, or standard deviation, of 0.33Å) of 4S02 to 3.04Å (s = 0.39 Å) of 4S0J, indicating no strong solvent-BiPhe residue interaction.

Table 5. Residue interactions with BIF^{11} within the 4S0x enzymes. C_{MC} indicates a main chain contact with BIF^{11}. C_{SC} indicates a side chain contact with BIF^{11}. O_{MC} indicates a main chain overlap with BIF^{11}. O_{SC} indicates a side chain overlap with BIF^{11}. HB indicates the main chain hydrogen bonds with the BIF^{11} main chain. (*) indicates the residues were included in the model for inter-model consistency and no distinct BIF^{11}-residue interaction was detected. (**) indicates the residue main chain was included for structural integrity and no distinct BIF^{11}-residue interaction was detected.

Residue		4S02						4S0J					
Position	1Y2Q		CMC	CSC	OMC	OSC	HB		CMC	CSC	OMC	OSC	HB
10	Y	Y	*	*	*	*	*	Y		1			
12	E	E	1					E	1				
13	Y	Y		11				Y		9		2	
34	R	R	2	2	1			R	2	1	1		
35	M	M	3		1		1	M	3		1		1
36	E	E	**	**	**	**	**	E	**	**	**	**	**
38	V	V		3				V		3			
40	V	V		7				V		8			
41	A	A	**	**	**	**	**	A	**	**	**	**	**
42	F	W		6		1		F		5		1	
77	F	F		11		4		F		11		2	
78	V	V	**	**	**	**	**	V	**	**	**	**	**
79	Y	A		4				S		7		1	
81	F	W						W					
115	A	A		5		3		A		5		2	
121	K	I		10				I		7			
123	F	Y		10				V		7		3	
124	K	K	1		1			K	2				
125	I	I	1	6		1		I	1	7			

Table 5 (continued)

Residue		4S0L						4S0I					
Position	1Y2Q		CMC	CSC	OMC	OSC	HB		CMC	CSC	OMC	OSC	HB
10	Y	Y	1					Y	*	*	*	*	*
12	E	E	1					E	1				
13	Y	Y		8				Y		6			
34	R	R	2		1			R	1		1		
35	M	M	3		1		1	M	3		1		1
36	E	E	**	**	**	**	**	E	**	**	**	**	**
38	V	V		3				V		1			
40	V	V		7				V		7			
41	A	A	**	**	**	**	**	A	**	**	**	**	**
42	F	F		5				F		5			
77	F	F		4		1		F		8		1	
78	V	V	**	**	**	**	**	V	**	**	**	**	**
79	Y	V		13		4		S		4			
81	F	W		1				W					
115	A	A		5				A		6		1	
121	K	I		6				I		7			
123	F	V		10		1		A		3			
124	K	K	1					K	2		1		
125	I	I		7				I	1	7		1	

Table 5 (continued)

Residue		4S0K						4S03					
Position	1Y2Q		CMC	CSC	OMC	OSC	HB		CMC	CSC	OMC	OSC	HB
10	Y	Y	*	*	*	*	*	Y		1			
12	E	E	2					E	2				
13	Y	Y		6				Y		7			
34	R	R	2	1	1			R	2	1	1		
35	M	M	3		1		1	M	3				1
36	E	E	**	**	**	**	**	E	**	**	**	**	**
38	V	V		1				V	*	*	*	*	*
40	V	V		6				V		5			
41	A	A	**	**	**	**	**	A	**	**	**	**	**
42	F	F		5				F		4			
77	F	F		7		1		F		4		2	
78	V	V	**	**	**	**	**	V	**	**	**	**	**
79	Y	V		14				I		15		2	
81	F	W		2				W					
115	A	A		6				A	*	*	*	*	*
121	K	I		9		1		I		8			
123	F	A		8		1		A		9		1	
124	K	K	2					K	3				
125	I	I		6				I	1	8		1	

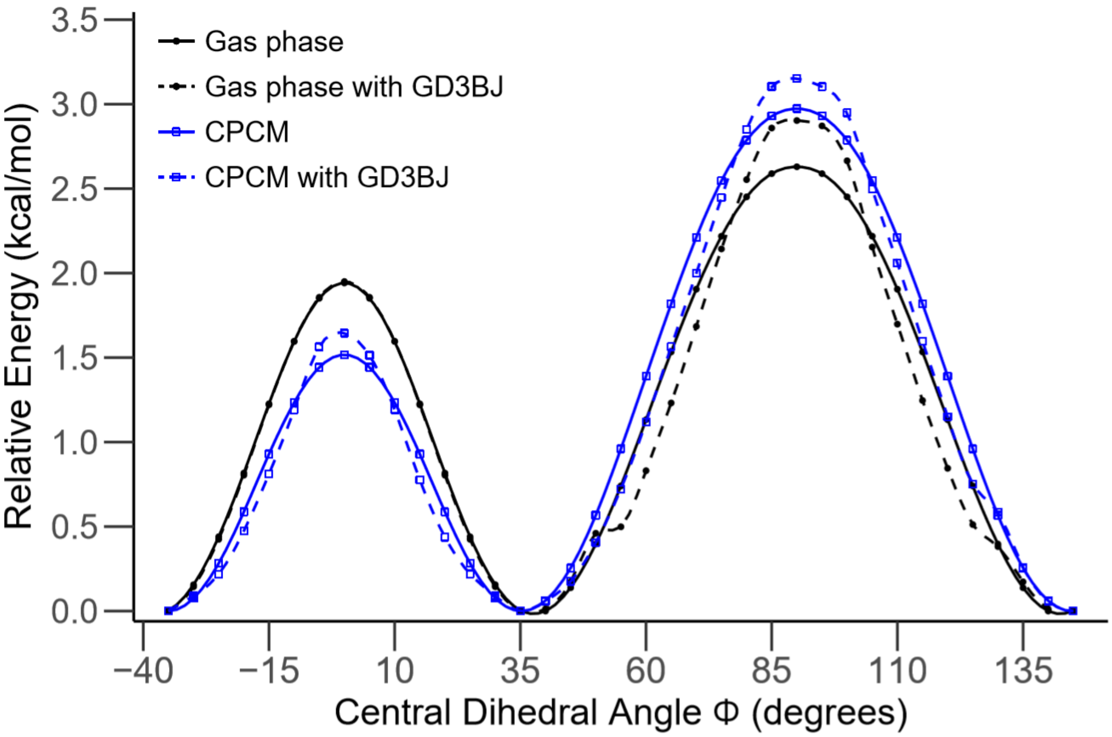


Figure 1. Potential energy curves for the torsional rotation of free biphenyl.

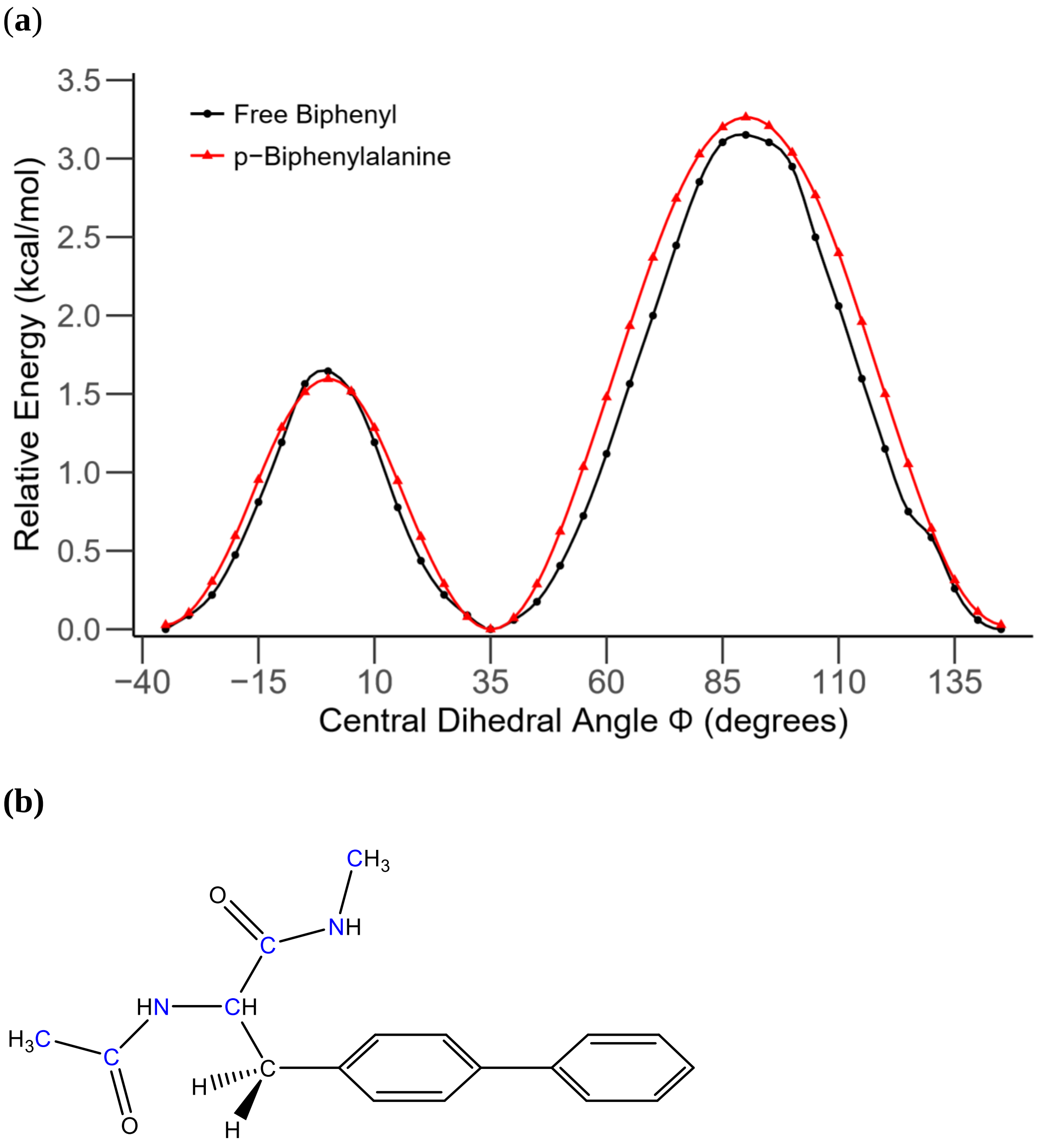
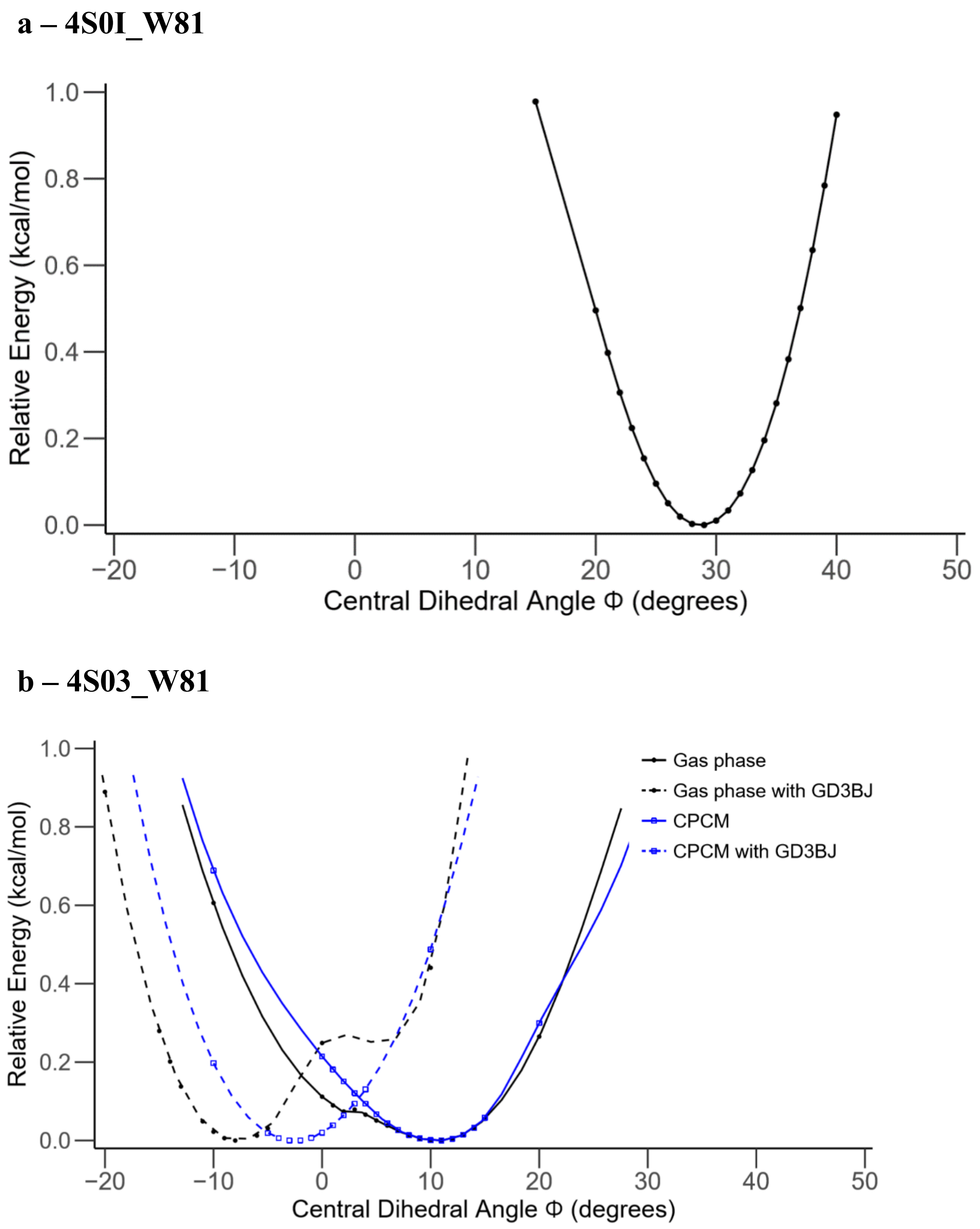


Figure 2. (a) Potential energy curves for a *p*-biphenylalanine derivative model and free biphenyl for the torsional rotation about the central C-C biphenyl bond, both calculated at the B3LYP/6-31G(d')+CPCM+GD3BJ level of theory. (b) Structure of the simulated *p*-biphenylalanine derivative. Blue is used to indicate the backbone atoms frozen to their respective 4S03 crystallographic coordinates.



a – 4S0I_W81

b – 4S03_W81

Figure 3. Potential energy curve near equilibrium for the torsional rotation of *p*-biphenylalanine within the (a) 4S0I_W81 and (b) 4S03_W81 protein cluster models. The 4S0I_W81 scan was only computed using the B3LYP/6-31G(d')+CPCM+GD3BJ level of theory. The 4S03_W81 scan was carried out in gas/aqueous phase, with and without GD3BJ.

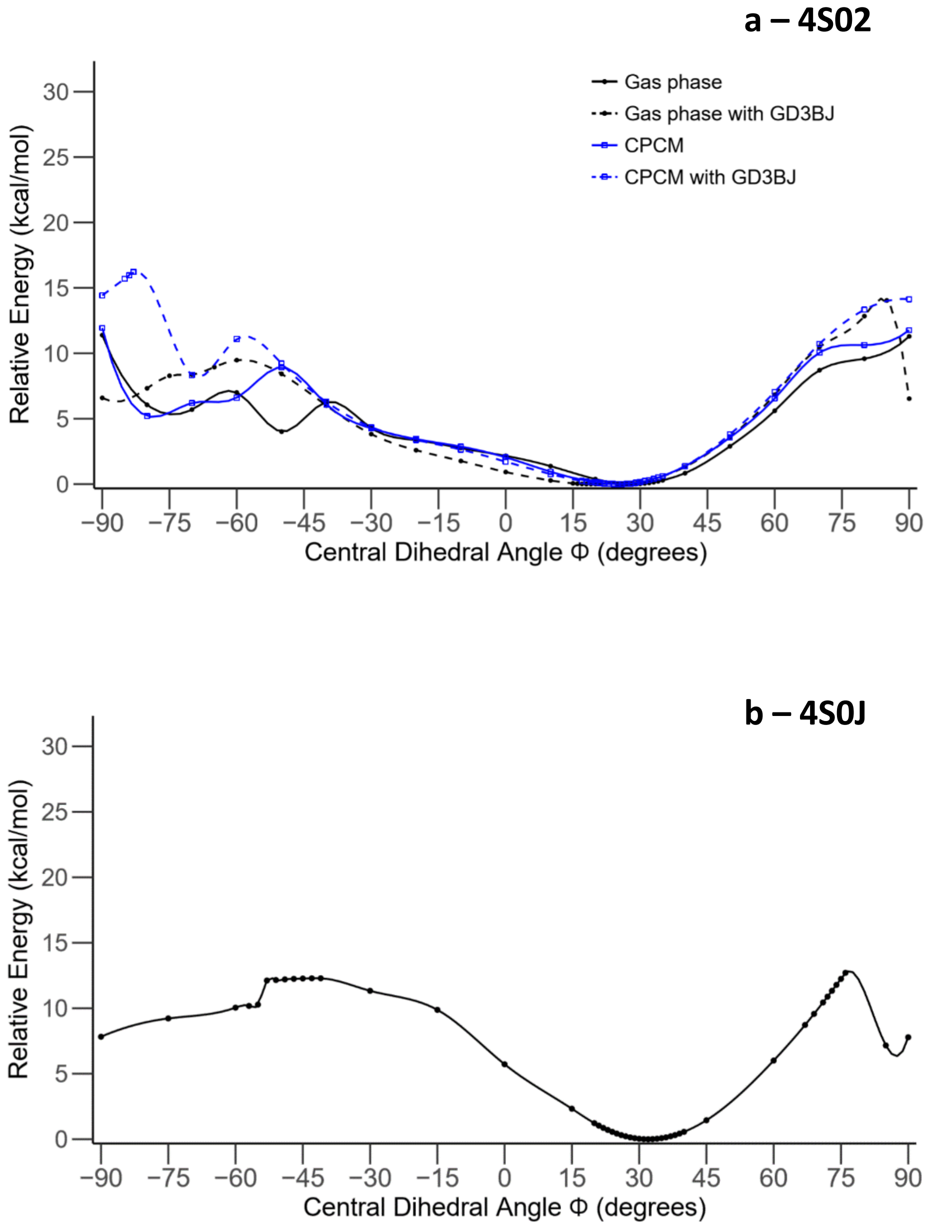


Figure 4. 180°-scan potential energy curves for the torsional rotation of p-biphenylalanine within the (a) 4S02, (b) 4S0J, (c) 4S0L, (d) 4S0I, (e) 4S0I_W81, (f) 4S0K, (g) 4S03, and (h) 4S03_W81 enzyme cluster models, calculated at the B3LYP/6-31G(d')+CPCM+GD3BJ level of theory unless otherwise labeled.

Figure 4 (continued)

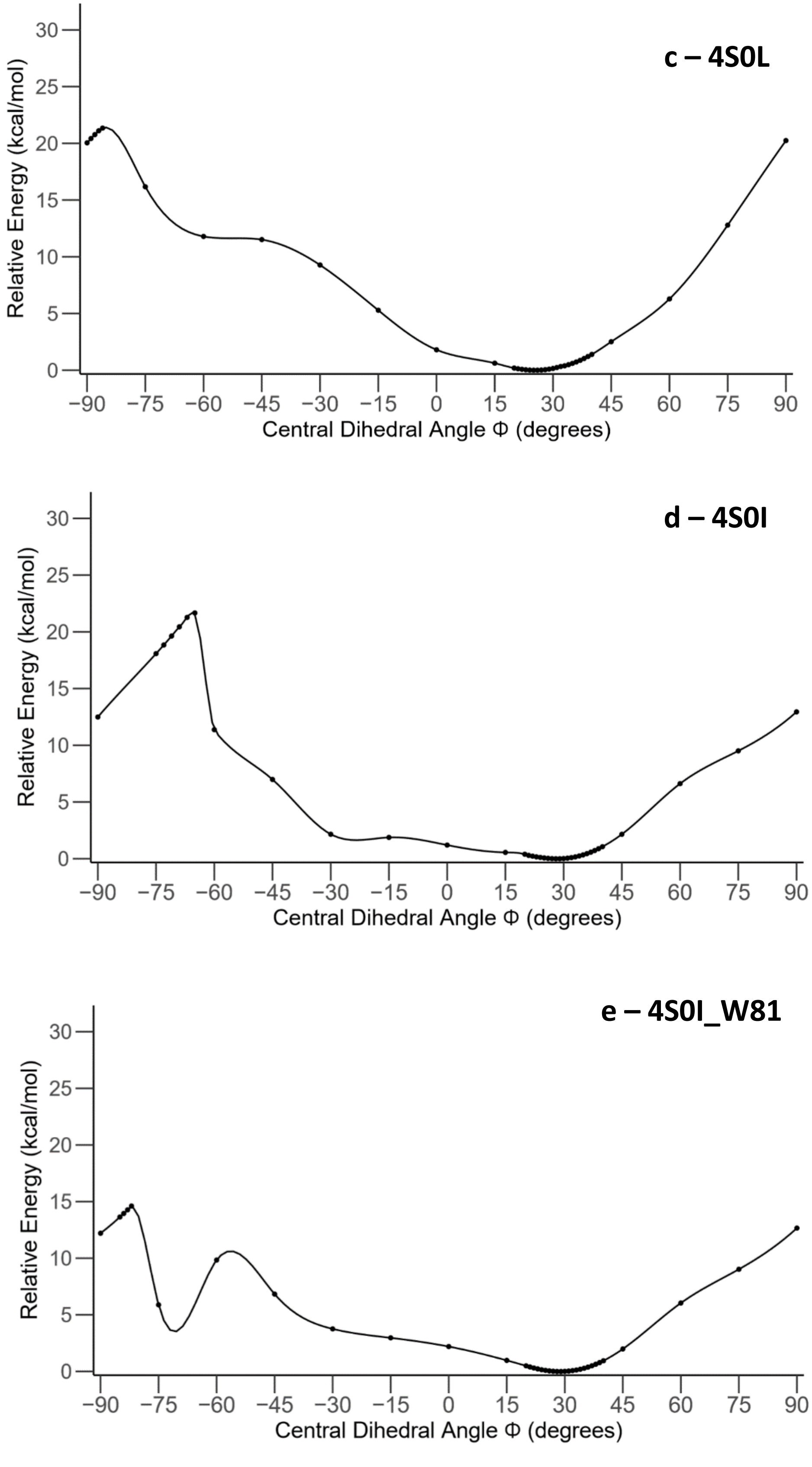
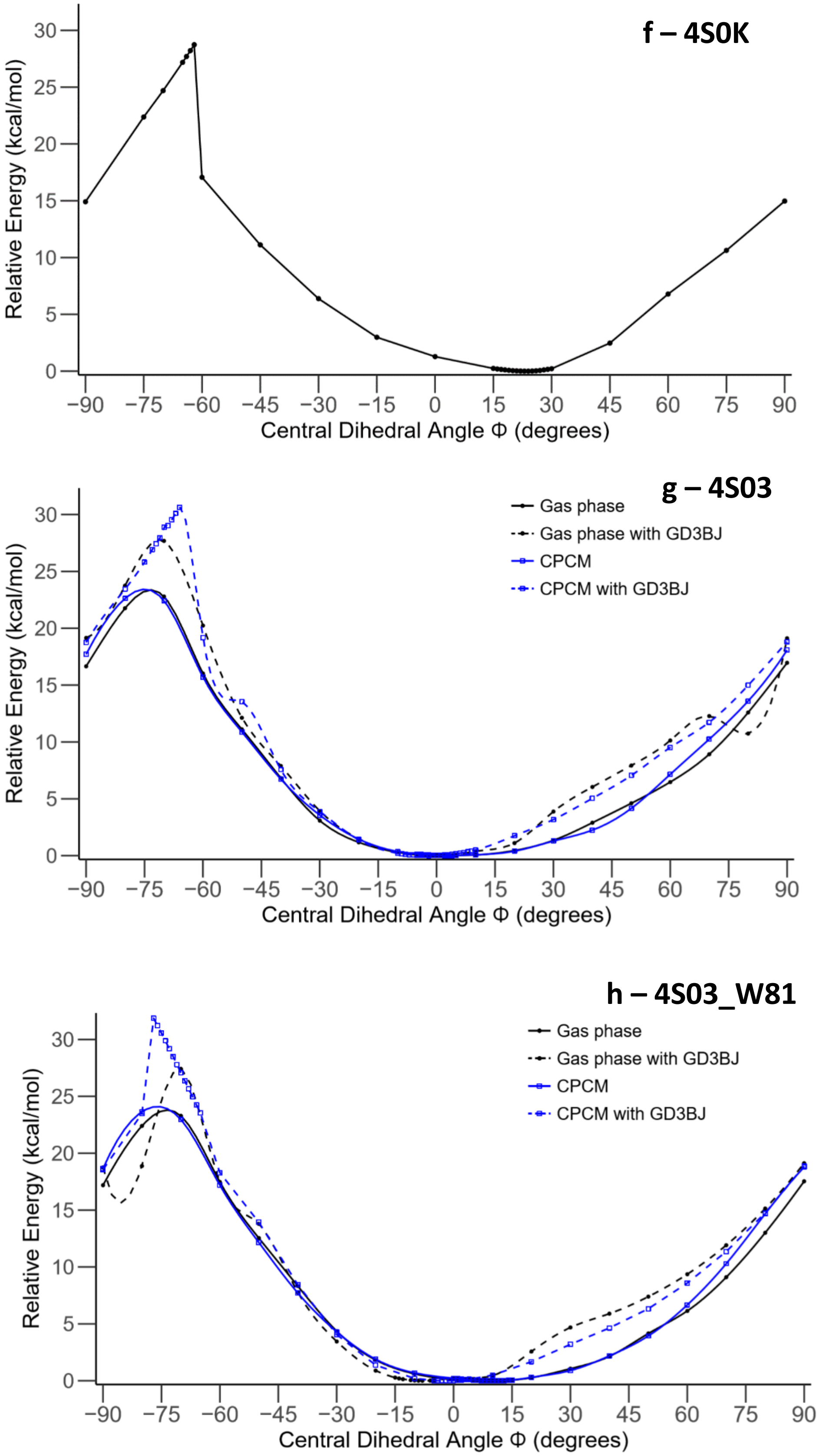


Figure 4 (continued)



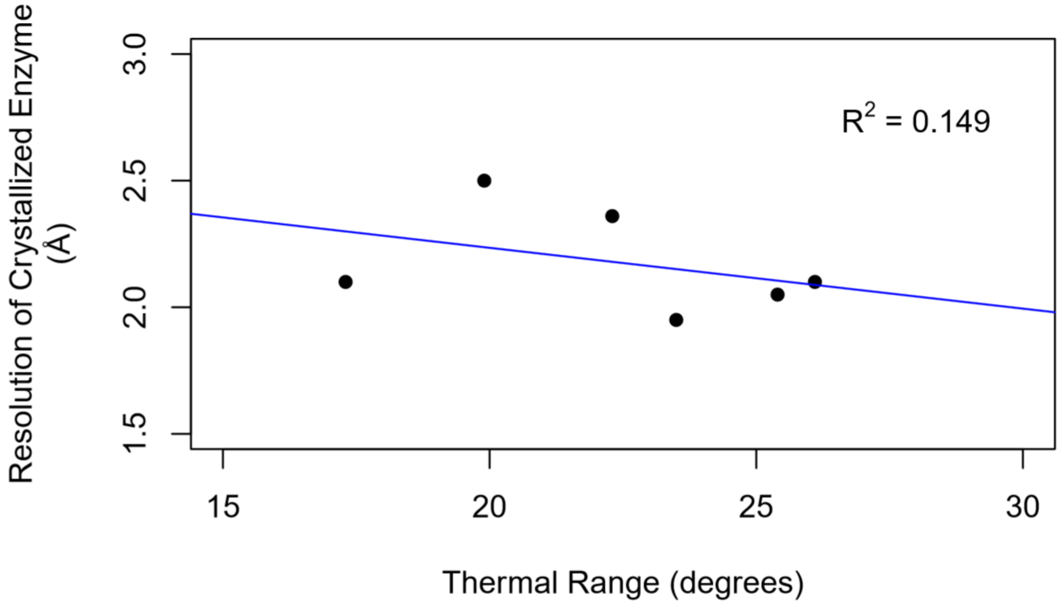


Figure 5. Linear model between the computed thermal range for *p*-biphenylalanine within a given protein core model and the x-ray crystallographic resolution.

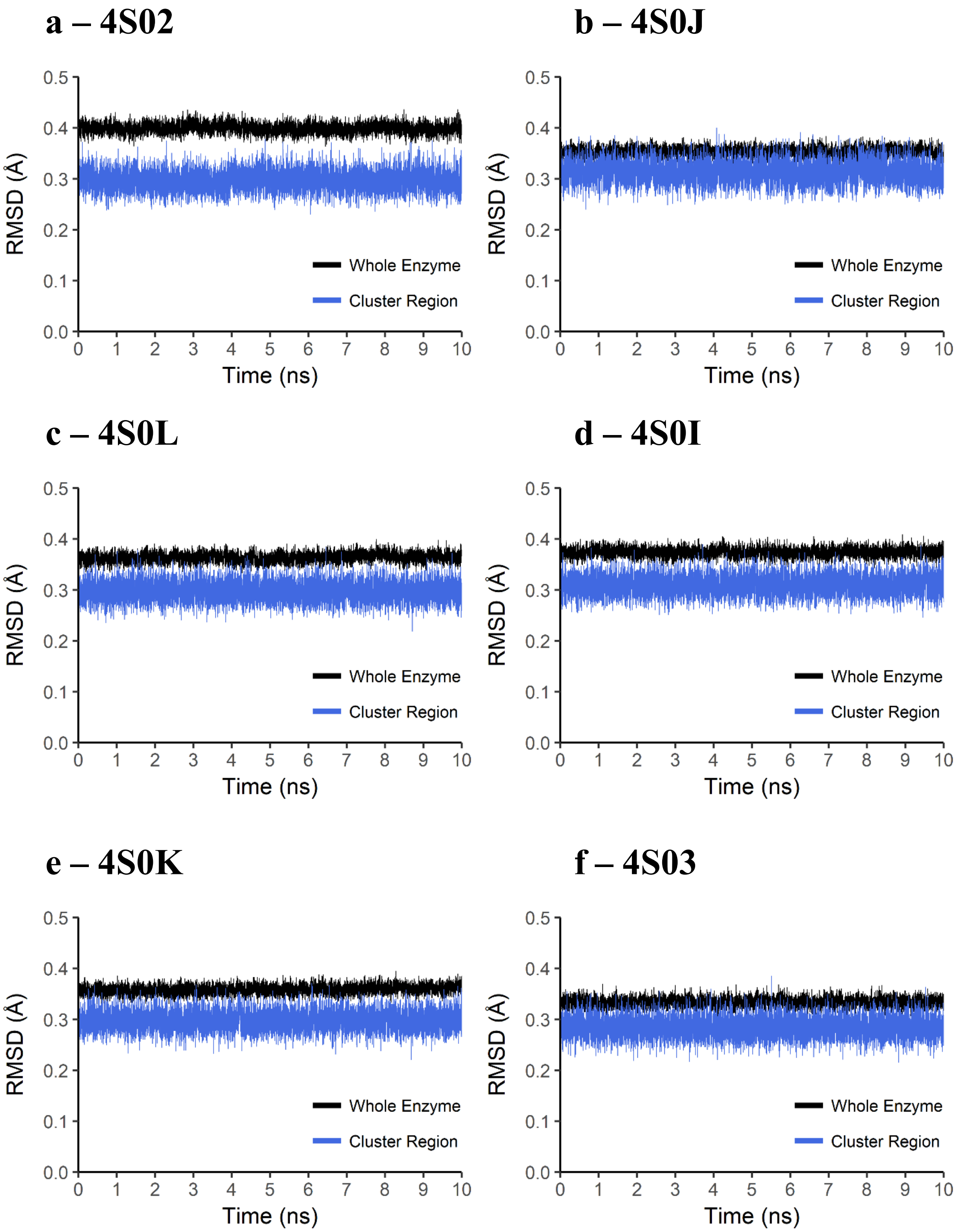


Figure 6. RMSD of non-hydrogen atoms of the whole protein and QM-cluster residues for MD simulations of (a) 4S02 (b) 4S0J (c) 4S0L (d) 4S0I (e) 4S0K and (f) 4S03 compared to their respective crystal structure with respect to a timescale of 10 ns.

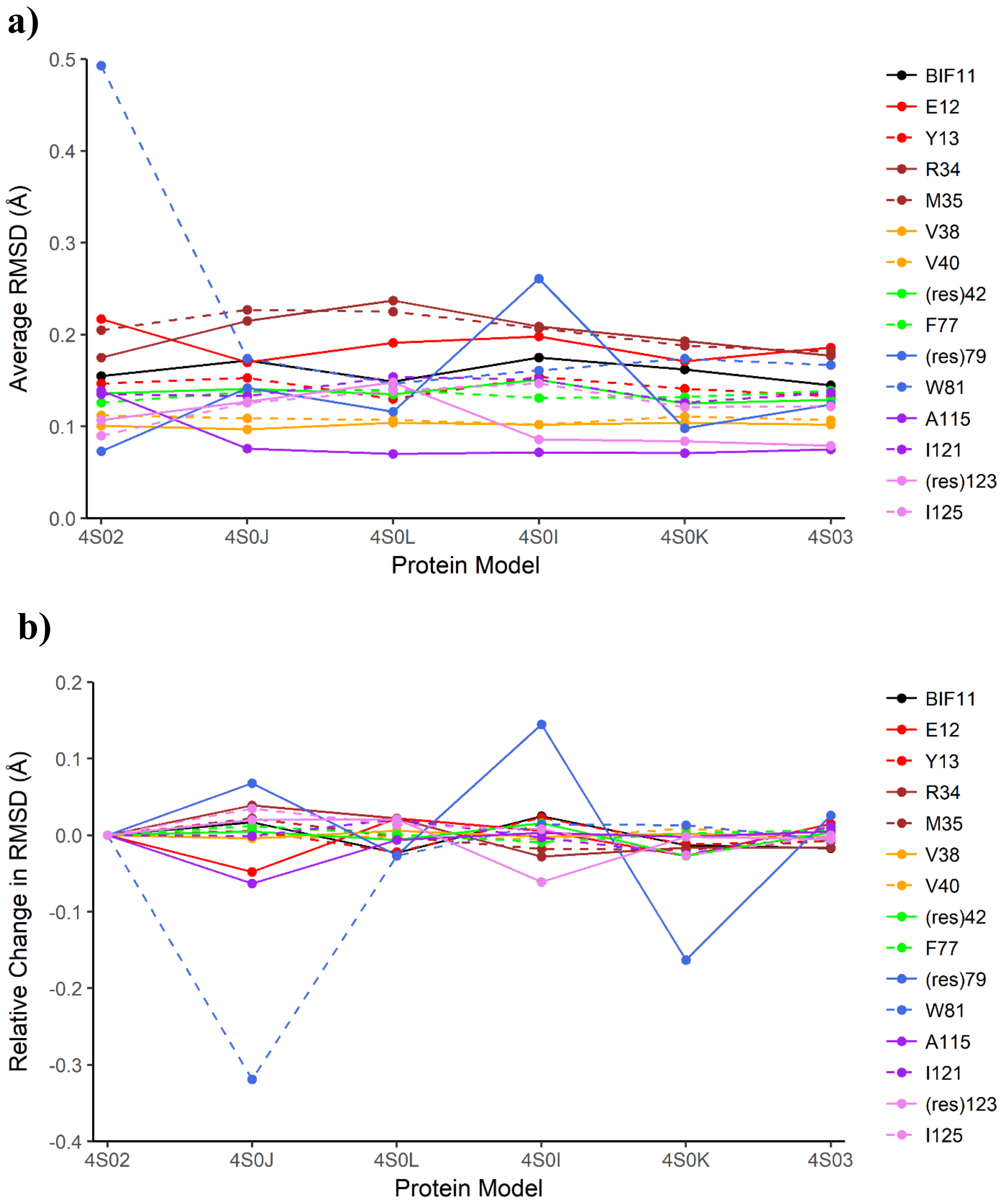


Figure 7. a) Average RMSD of non-hydrogen atoms of select residues within MD simulations compared to their respective crystal structure. b) Relative change in residue RMSD with respect to its value in the previous model.

a – 4S02

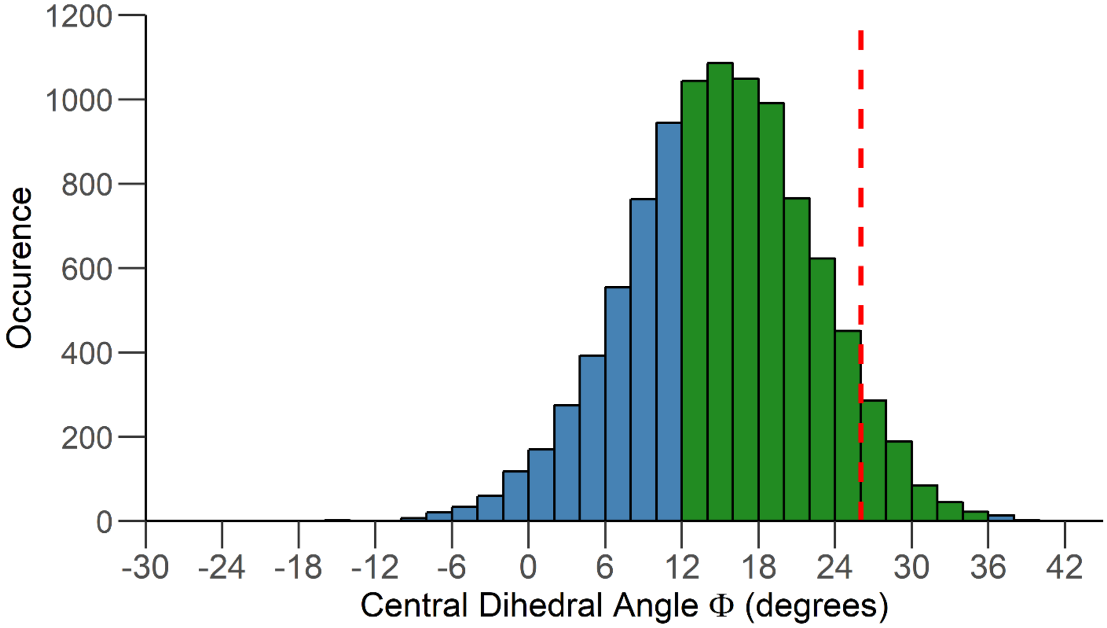
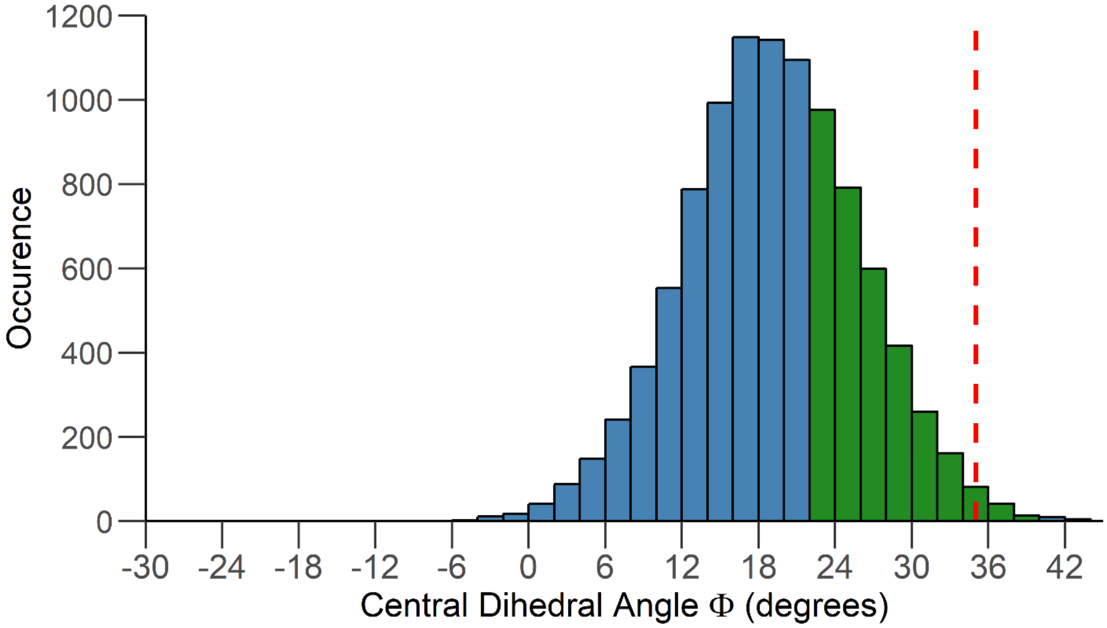


Figure 8. Distribution of the BiPhe central dihedral angles in MD simulation snapshots of (a) 4S02 (b) 4S0J (c) 4S0L (d) 4S0I (e) 4S0K and (f) 4S03. The red dashed line represents the x-ray crystallographically determined value. Green is used to represent dihedral angle values within the thermal range calculated by our cluster models.

Figure 8 (continued)

b – 4S0J



c – 4S0L

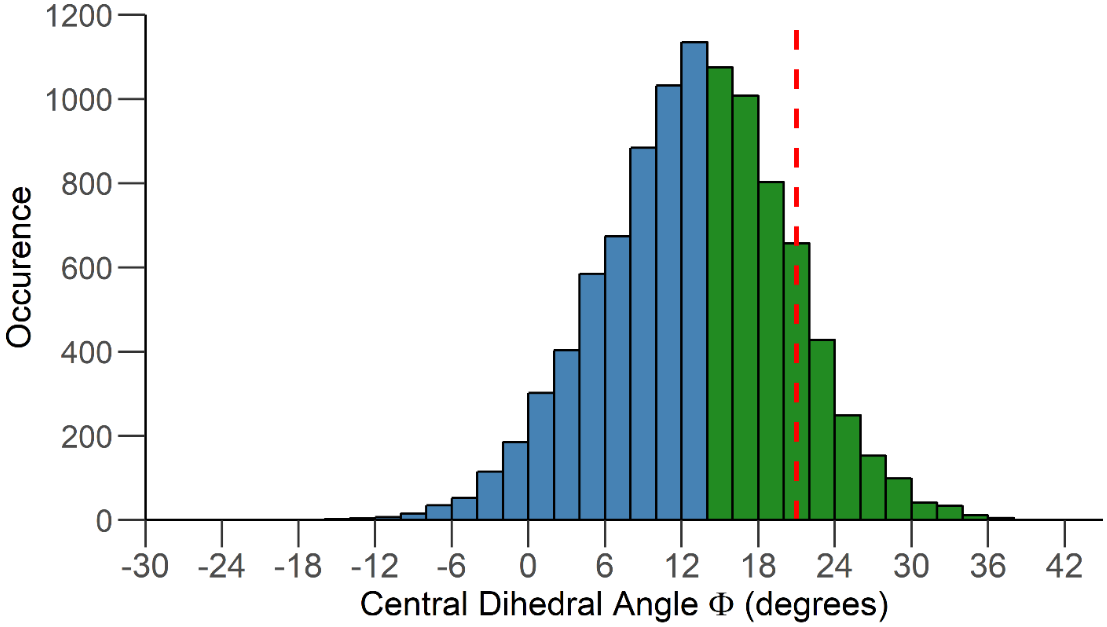
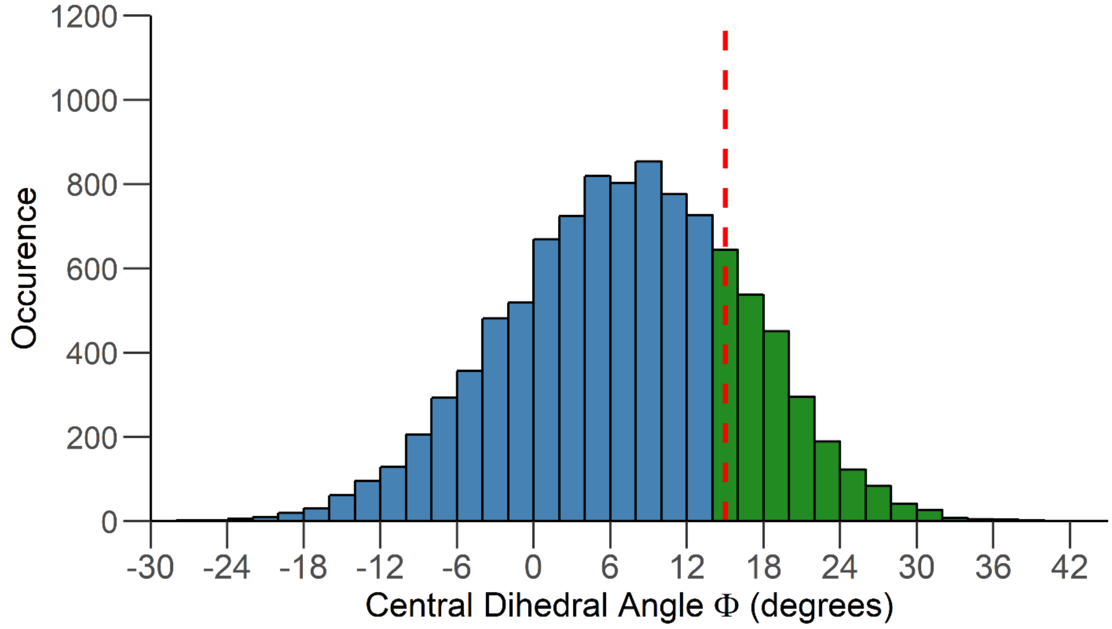


Figure 8 (continued)

d – 4S0I



e – 4S0K

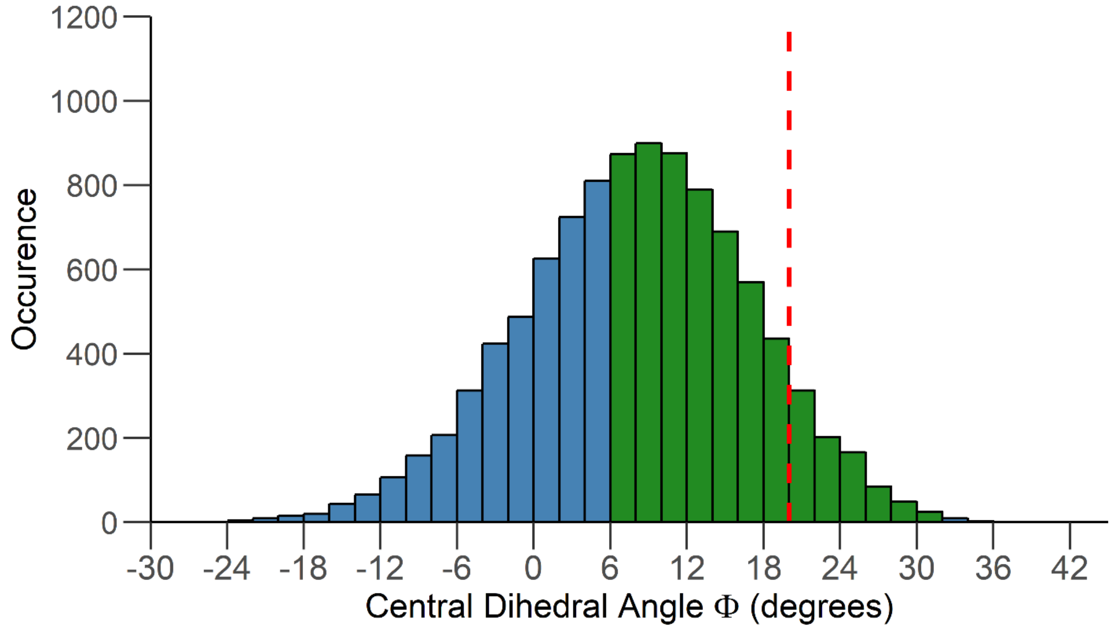
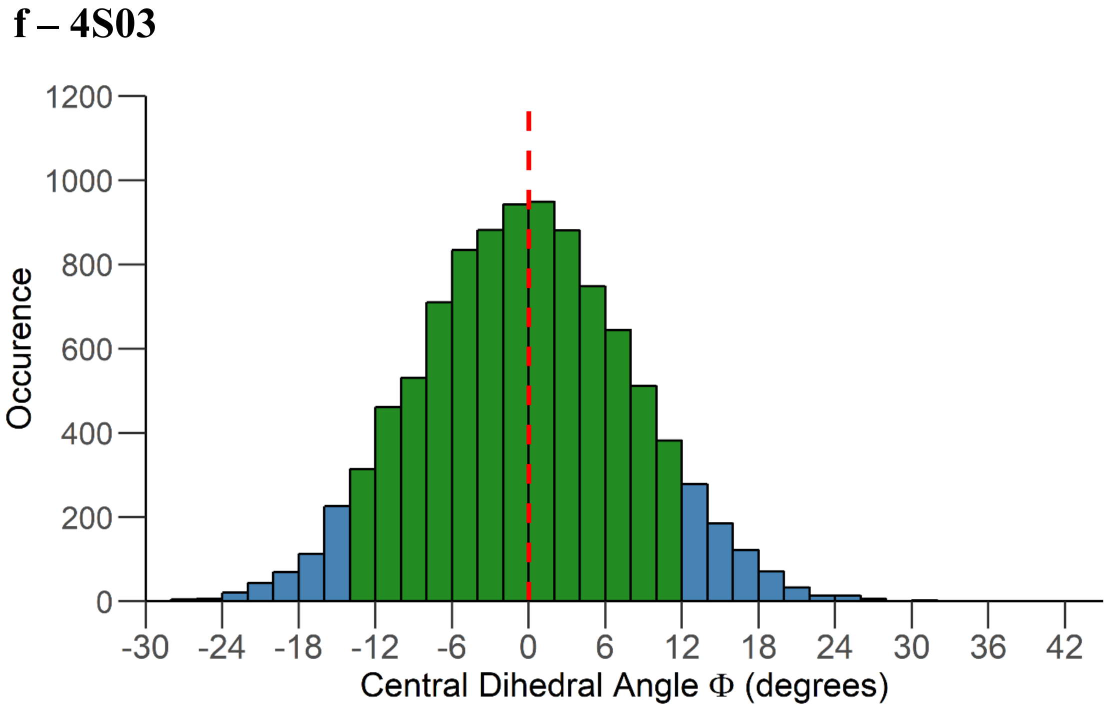


Figure 8 (continued)



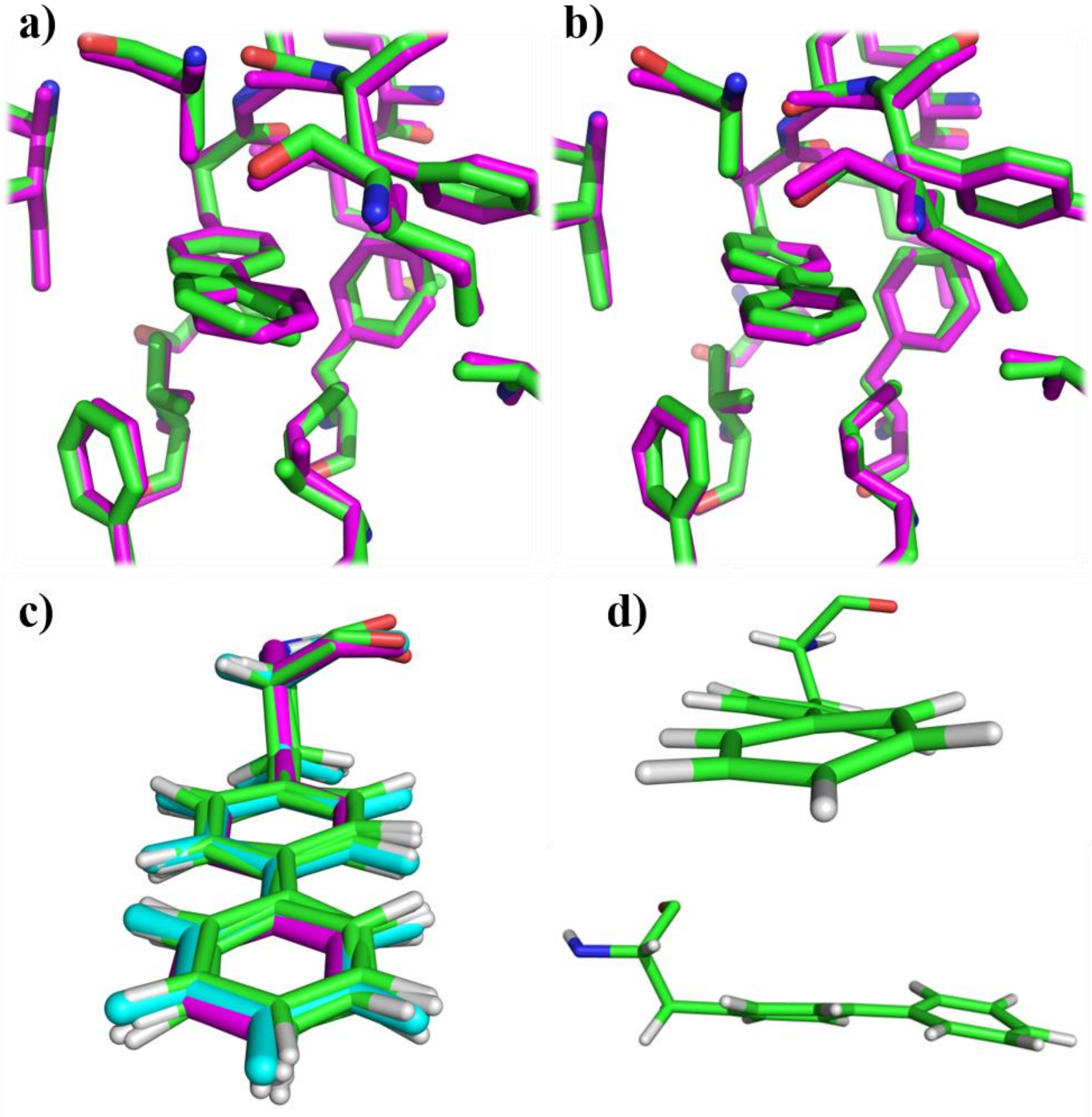


Figure 9. Overlay of select MD simulation snapshots of the 4S03 BiPhe core and the x-ray crystal structure (magenta) and average MD simulation structure (cyan). Models with the a) smallest and b) largest BiPhe dihedral angle are presented, along with five random intermediary angles (c). Viewing along the central BiPhe bond shows how snapshots outside the thermal range often have biphenyl ring distortion (d).

Appendix B: Chapter 3 Supplementary Information

Model cartesian coordinates and data are available on request.

Table 1. Residue trimming scheme for the residues indicating where the N-terminus, C-terminus, and Side Chains are trimmed away and capped with hydrogens.

Residue	N-terminus	Trim Side Chain	C-terminus
A 92 GLN	H	-	-
A 93 PHE	-	H	-
A 94 HIS	-	-	-
A 95 Phe	-	H	-
A 96 HIS	-	-	H
A 106 GLU	H	-	H
A 117 GLU	H	-	-
A 118 LEU	-	H	-
A 119 HIS	-	-	-
A 120 Leu	-	H	-
A 121 VAL	-	-	H
A 143 VAL	H	-	H
A 198 _LEU	H	-	-
A 199 THR	-	-	H
A 209 TRP	H	-	H
A 244 ASN	H	-	-
A 245 TRP	-	H	H
A 262 CO	-	-	-
A 267 HOH	-	-	-
A 272 HOH	-	-	-
A 375 HOH	-	-	-

Appendix C: Chapter 4 Supplementary Information

Model cartesian coordinates and additional data are available on request.

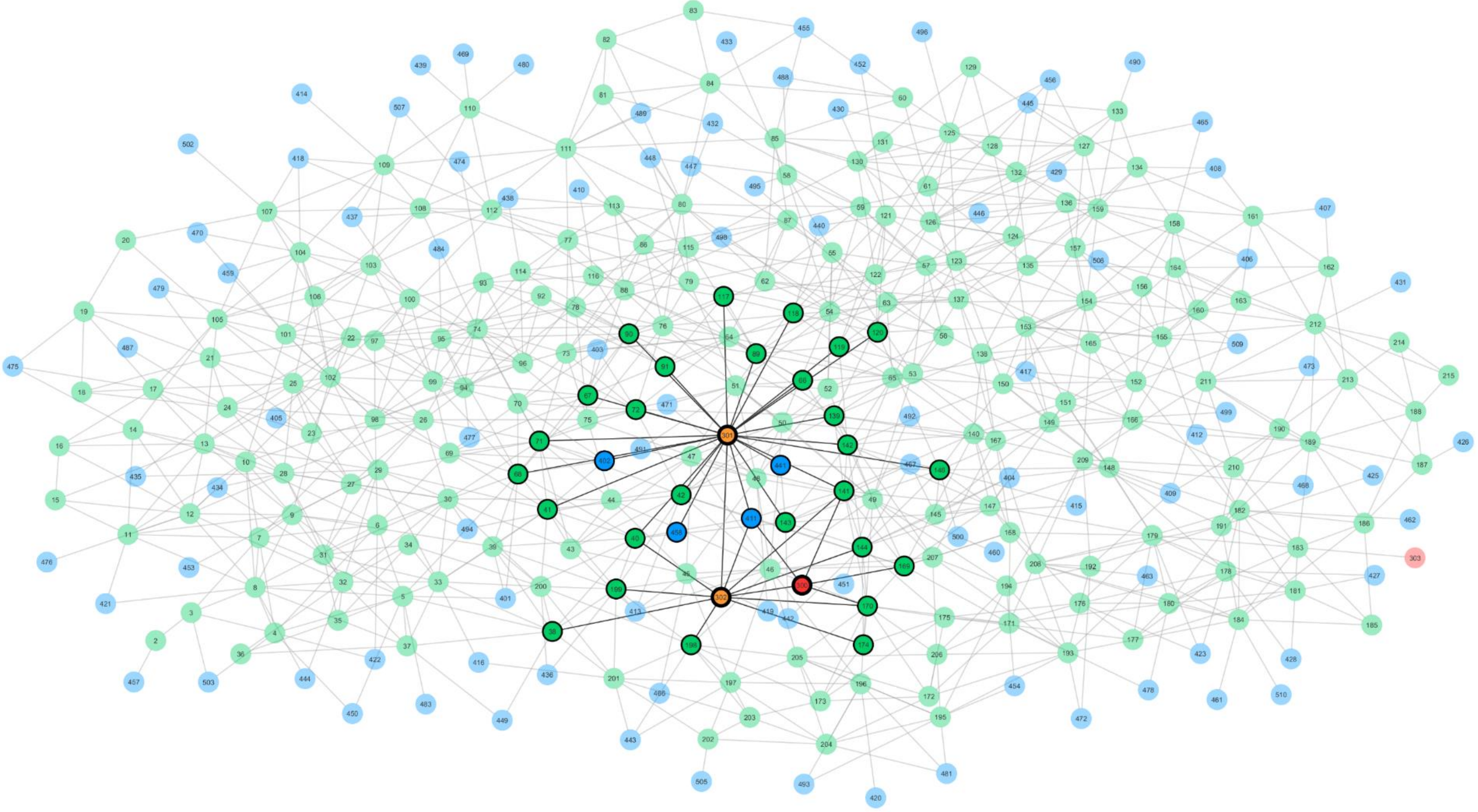


Figure 1. Interaction network for COMT (PDB: 3BWM). Nodes are labeled by their residue sequence number and colored by identity: green for amino acids, blue for waters, orange for substrates, red for metals). Nodes representing the chemically reactive species (nodes 300 [Mg^{2+}, red], 301 [SAM, orange] and 302 [CAT, orange]) and their first neighbor nodes are emphasized.

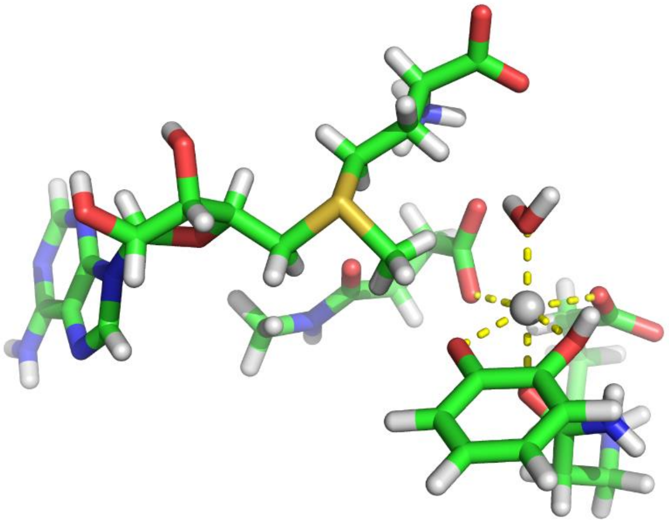


Figure 2. 3D structure of the baseline model, or “seed”, used for constructing larger QM-cluster models of the COMT active site.

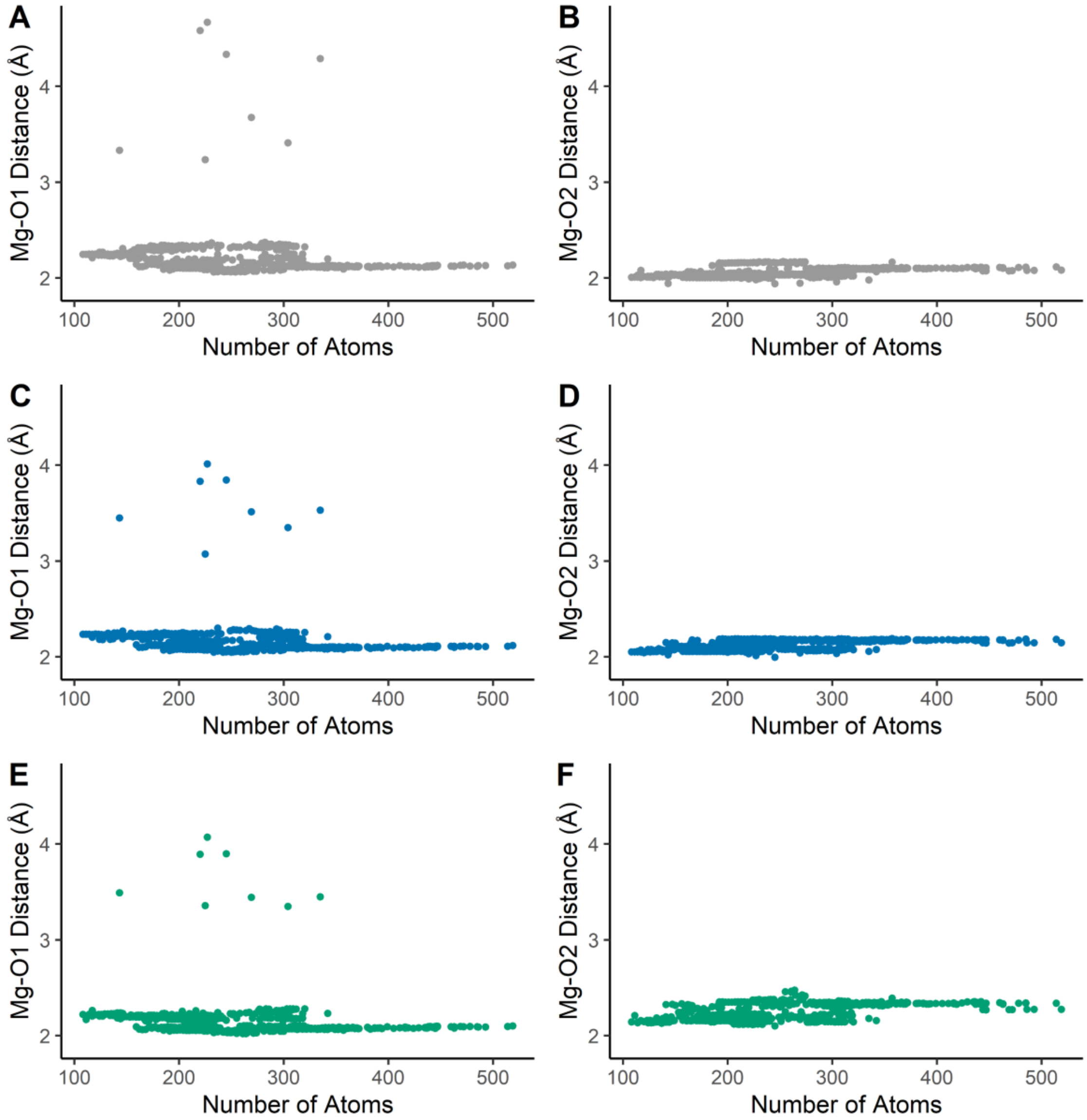


Figure 3. The denticity of the catecholate substrate in the reactant (A, B), transition state (C, D), and product (E, F) models. The O1 atom is not bound to the Mg (Mg-O1 distance > 3Å) in 8 models of the reactant (A), transition state (C), and product structures (E) while the O2 atom remains consistently bound.

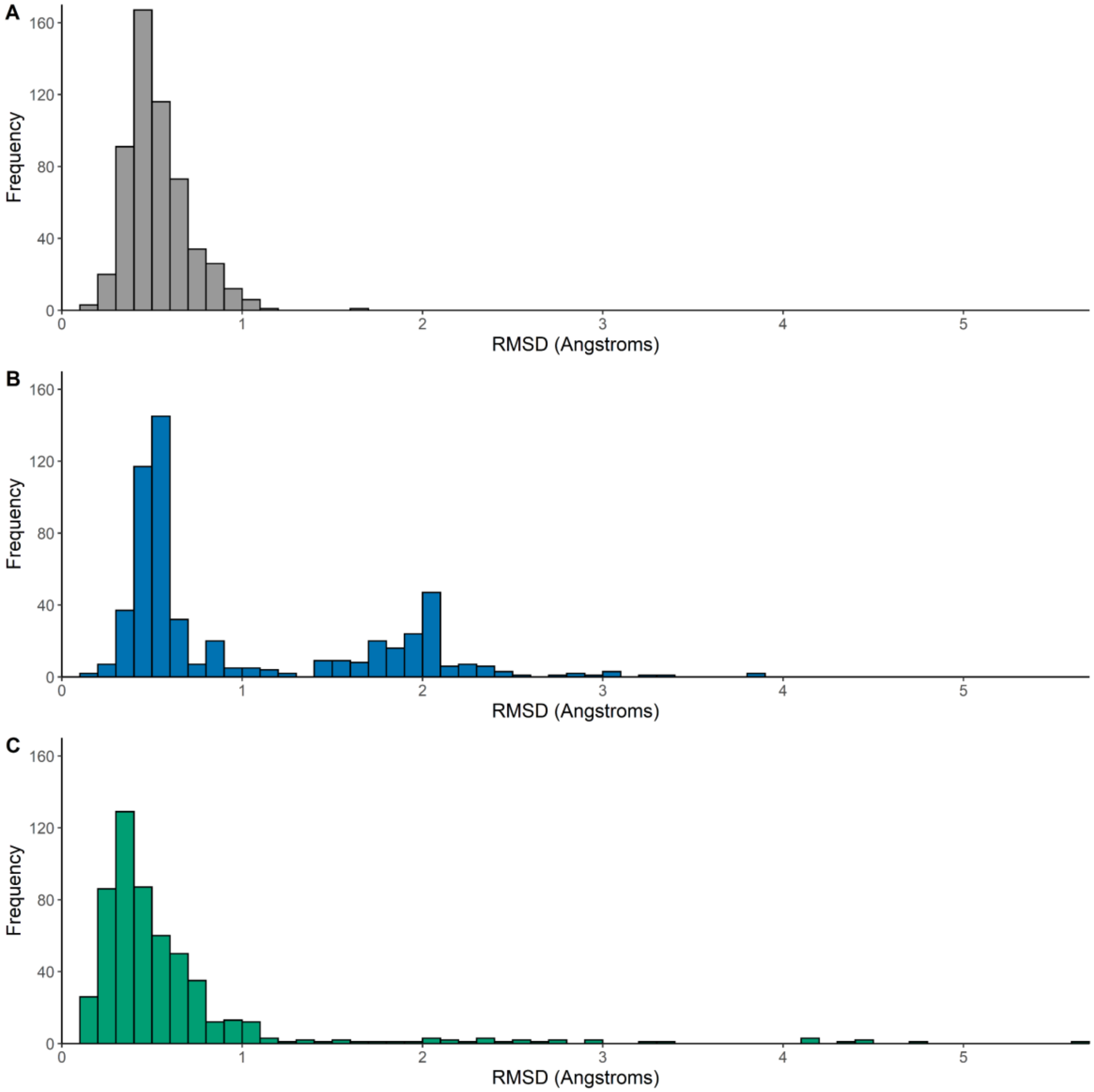


Figure 4. Distribution of the root mean square deviation (RMSD) of the non-hydrogen, unconstrained, optimized reactant atoms compared to the crystal structure coordinates. The distribution of RMSD for all the atoms in the model excluding SAM and CAT (A), for only the atoms of CAT (B), and for only the atoms of SAM (C) are shown.

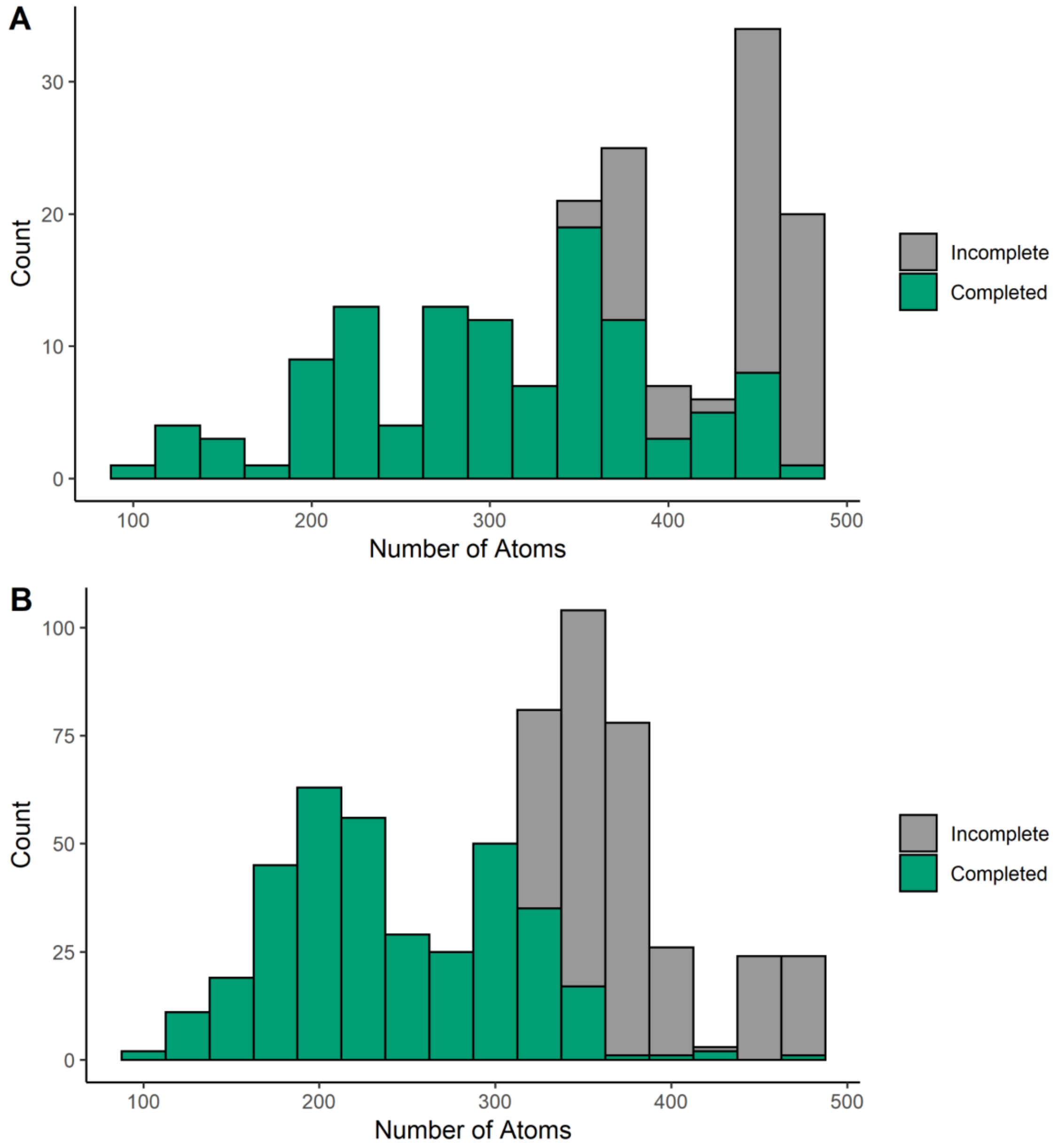


Figure 5. Histograms of the number of Combinatoric Scheme 1 models (A) and Combinatoric Scheme 2 models (B) completed for this work based upon the number of atoms present in the models.

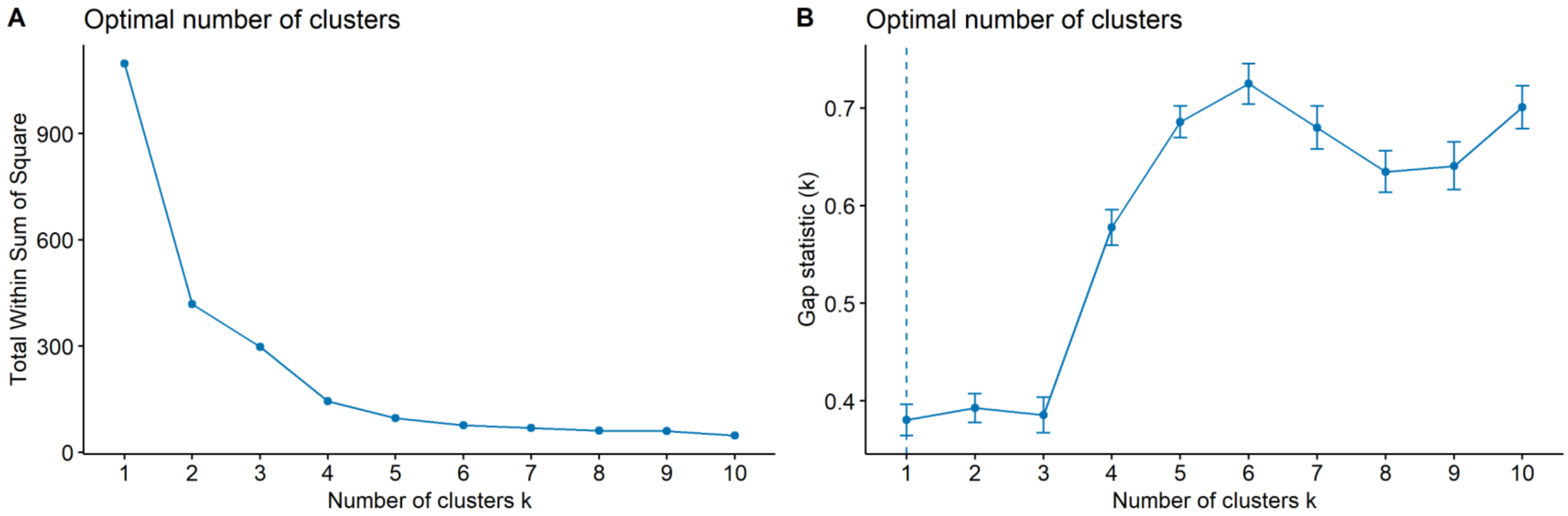


Figure 6. Elbow (A) and Gap (B) statistics for the computed *k*-means clustering.

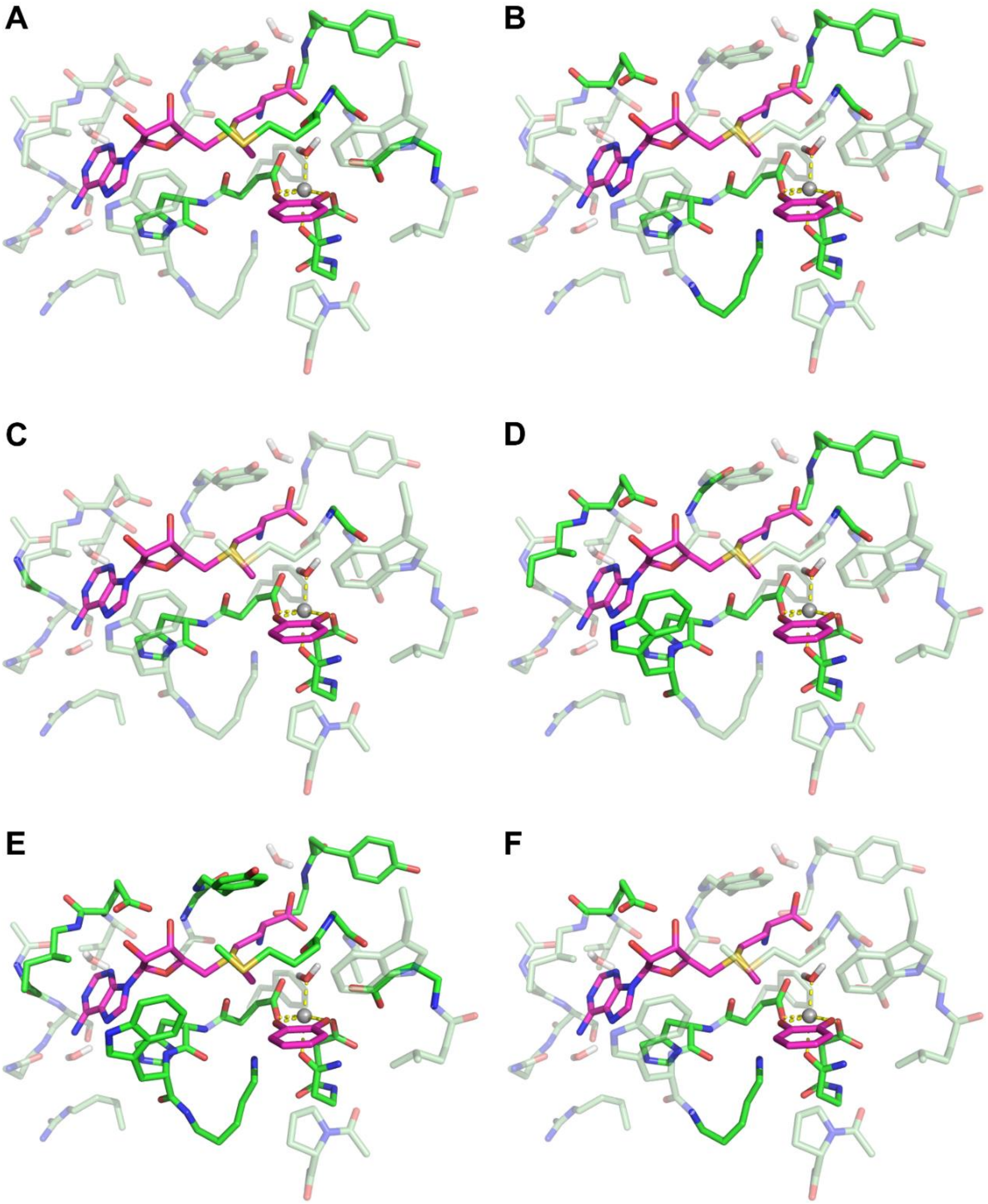
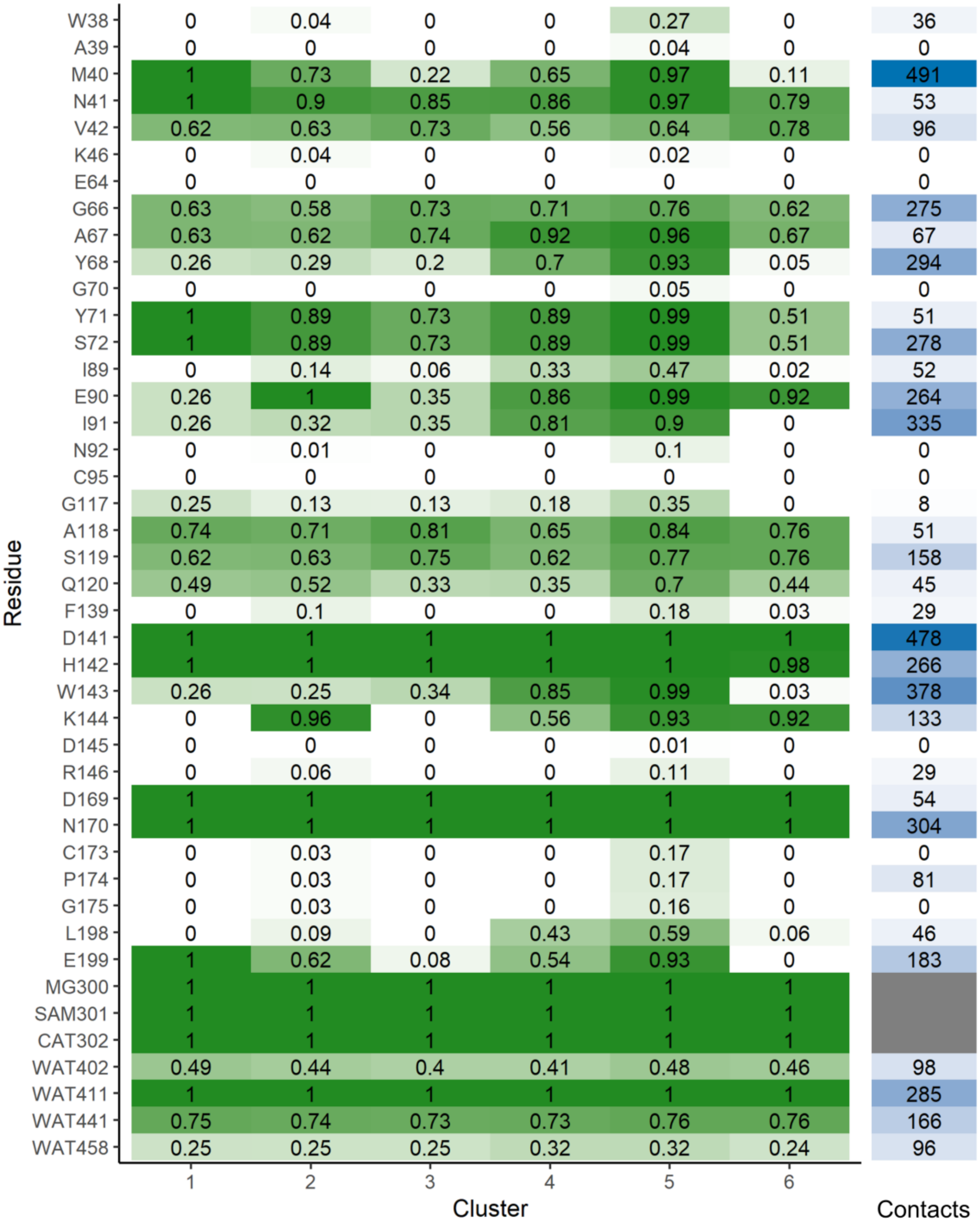


Figure 7. Visualization of the maximal 485-atom model highlighting the residues that occur in >80% of Clusters 1 (A), 2 (B), 3 (C), 4 (D), 5 (E) and 6 (F). The carbon atoms of the substrates are colored magenta. Residue frequency is tabulated in Appendix C: Table 2.

Table 1. Total contacts between the chemical active site and residue main chains (MC), side chains (SC) or waters (WAT).

Residue	Wide Contacts			Close Contacts			Hydrogen Bonding			Small Overlaps			Big Overlaps			Total Contacts
	MC	*SC*	*WAT*	*MC*	*SC*	*WAT*	*MC*	*SC*	*WAT*	*MC*	*SC*	*WAT*	*MC*	*SC*	*WAT*	
W38	0	27	0	0	9	0	0	0	0	0	0	0	0	0	0	36
M40	58	158	0	5	223	0	0	0	0	0	47	0	0	0	0	491
N41	38	0	0	15	0	0	0	0	0	0	0	0	0	0	0	53
V42	0	22	0	7	20	0	47	0	0	0	0	0	0	0	0	96
G66	142	0	0	113	0	0	6	0	0	14	0	0	0	0	0	275
A67	45	0	0	22	0	0	0	0	0	0	0	0	0	0	0	67
Y68	57	47	0	19	70	0	0	0	0	0	101	0	0	0	0	294
Y71	0	0	0	0	17	0	0	0	0	0	34	0	0	0	0	51
S72	10	53	0	33	120	0	17	1	0	0	44	0	0	0	0	278
I89	3	9	0	0	31	0	0	0	0	0	9	0	0	0	0	52
E90	47	54	0	10	37	0	0	116	0	0	0	0	0	0	0	264
I91	31	139	0	52	94	0	2	0	0	0	17	0	0	0	0	335
G117	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
A118	9	0	0	32	0	0	0	0	0	10	0	0	0	0	0	51
S119	7	71	0	7	10	0	63	0	0	0	0	0	0	0	0	158
Q120	0	4	0	0	32	0	0	9	0	0	0	0	0	0	0	45
F139	0	23	0	0	6	0	0	0	0	0	0	0	0	0	0	29
D141	23	88	0	92	126	0	0	0	0	55	65	0	0	29	0	478
H142	18	82	0	51	85	0	0	0	0	24	6	0	0	0	0	266
W143	28	96	0	18	138	0	0	0	0	0	96	0	0	2	0	378
K144	0	63	0	0	14	0	0	52	0	0	4	0	0	0	0	133
R146	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	29
D169	0	0	0	0	0	0	0	0	0	0	48	0	0	6	0	54
N170	0	22	0	0	95	0	0	32	0	0	123	0	0	32	0	304
P174	0	73	0	0	8	0	0	0	0	0	0	0	0	0	0	81
L198	0	0	0	0	23	0	0	0	0	0	23	0	0	0	0	46
E199	0	46	0	0	91	0	0	4	0	0	42	0	0	0	0	183
HOH402	0	0	0	0	0	17	0	0	72	0	0	9	0	0	0	98
HOH411	0	0	39	0	0	36	0	0	146	0	0	36	0	0	28	285
HOH441	0	0	2	0	0	16	0	0	136	0	0	12	0	0	0	166
HOH458	0	0	8	0	0	17	0	0	71	0	0	0	0	0	0	96
Total	524	1106	49	476	1249	86	135	214	425	103	659	57	0	69	28	5180

Table 2. Relative frequency (green) and number of *Probe* contacts (blue) for each residue being present in the models of a *k*-cluster. Values are proportionally shaded to emphasize differences in residue composition among *k*-clusters.



Residue	Cluster						Contacts
	1	2	3	4	5	6	
W38	0	0.04	0	0	0.27	0	36
A39	0	0	0	0	0.04	0	0
M40	1	0.73	0.22	0.65	0.97	0.11	491
N41	1	0.9	0.85	0.86	0.97	0.79	53
V42	0.62	0.63	0.73	0.56	0.64	0.78	96
K46	0	0.04	0	0	0.02	0	0
E64	0	0	0	0	0	0	0
G66	0.63	0.58	0.73	0.71	0.76	0.62	275
A67	0.63	0.62	0.74	0.92	0.96	0.67	67
Y68	0.26	0.29	0.2	0.7	0.93	0.05	294
G70	0	0	0	0	0.05	0	0
Y71	1	0.89	0.73	0.89	0.99	0.51	51
S72	1	0.89	0.73	0.89	0.99	0.51	278
I89	0	0.14	0.06	0.33	0.47	0.02	52
E90	0.26	1	0.35	0.86	0.99	0.92	264
I91	0.26	0.32	0.35	0.81	0.9	0	335
N92	0	0.01	0	0	0.1	0	0
C95	0	0	0	0	0	0	0
G117	0.25	0.13	0.13	0.18	0.35	0	8
A118	0.74	0.71	0.81	0.65	0.84	0.76	51
S119	0.62	0.63	0.75	0.62	0.77	0.76	158
Q120	0.49	0.52	0.33	0.35	0.7	0.44	45
F139	0	0.1	0	0	0.18	0.03	29
D141	1	1	1	1	1	1	478
H142	1	1	1	1	1	0.98	266
W143	0.26	0.25	0.34	0.85	0.99	0.03	378
K144	0	0.96	0	0.56	0.93	0.92	133
D145	0	0	0	0	0.01	0	0
R146	0	0.06	0	0	0.11	0	29
D169	1	1	1	1	1	1	54
N170	1	1	1	1	1	1	304
C173	0	0.03	0	0	0.17	0	0
P174	0	0.03	0	0	0.17	0	81
G175	0	0.03	0	0	0.16	0	0
L198	0	0.09	0	0.43	0.59	0.06	46
E199	1	0.62	0.08	0.54	0.93	0	183
MG300	1	1	1	1	1	1	
SAM301	1	1	1	1	1	1	
CAT302	1	1	1	1	1	1	
WAT402	0.49	0.44	0.4	0.41	0.48	0.46	98
WAT411	1	1	1	1	1	1	285
WAT441	0.75	0.74	0.73	0.73	0.76	0.76	166
WAT458	0.25	0.25	0.25	0.32	0.32	0.24	96

QM-model Construction

Residue Selection

Determining which residues to include/exclude in a model generally requires some form of residue ranking. Various selection criterion employed by others include distance from defined foci, partial charges, impact of excluding the residue from the model, and researcher's "chemical intuition." This work builds-up models from a 7-residue base composed of the chemically reactive and metal-coordinating residues (D141, D169, N170, Mg300, SAM301, CAT302, HOH411; Appendix C: Figure 2).

Models are built by either ranking the individual residues and adding them incrementally to the base model or by forming groups of residues by a common feature and adding group(s) of residues to the base model.

Expansion by Ranking Residues

- Distance from Reacting Species – Residues are incrementally added to models based upon shortest distance from any non-hydrogen atom of the reacting species (Mg^{2+}, SAM, CAT) to any non-hydrogen atom of a residue. The residues are added until all of the residues with contact dots (Appendix C: Table 1) have been included. Residues K46 and N92 are present in this list but do not have contact dots with the chemically reactive species. The order of residues added to the base model is as follows: M40, E90, HOH441, K144, V42, S119, L198, H142, Y68, F139, S72, I89, Y71, Q120, G66, E199, W143, R146, I91, K46, HOH402, A118, A67, W38, HOH458, N41, N92, G117, P174.

- Total Number of Contacts – Residues are incrementally added to models from most to fewest total contacts (see Appendix C: Table 1)
- Residue Frequency in Combinatoric Scheme 2 – Residues are cumulatively added to models based upon the frequency of a residue's occurrence in the unique models formed from the combinations of sets constructed in Combinatoric Scheme 2 (see below). While these models do not directly map back to a systematic cheminformatic method translatable to other studies, they do still provide additional model variants useful for this work's analysis on the impact of residue composition on model convergence. The order of residues added to the base model is as follows: S72, HOH441, H142, G66, S119, V42, A118, M40, E90, K144, E199, Q120, W143, Y68, I91, HOH402, N41, A67, L198, Y71, I89, G117, W38, HOH458, F139, P174, R146.

Expansion by Groupings of Interaction Features

- Combinatoric Scheme 1 – Models are formed from the sets of residues with particular combinations of contact types. Table 1 indicates 14 contact types (e.g. wide contact-main chain, wide contact-side chain, etc.) are present, leading to the total number of contact type combinations being $\sum_{i=1}^{14}\binom{14}{i}$ = 16,338. As residues often have more than two types of contact type, most combinations of contact types yield redundant models. This redundancy reduces the total number of unique residue sets to only 204 possible models.
- Combinatoric Scheme 2 – Similar to models formed from combinations of contact types, these models are formed from the combinations of the 15 sets listed below. The sets were derived using an older, no-longer-employed grouping method based

on similar types of contacts (contacts, hydrogen bonding, and overlaps). At this time, we are not able to directly map most of these models back to a systematic cheminformatic method appropriate for being templated to other works. Nevertheless, the variation in residue composition these sets of models possess do give significant insight into the residues that impact model convergence and are thus kept in the work. The total of 32,767 possible set combinations simplifies by redundancy to only 736 possible unique models. The sets and their general common feature are specified below. The sets titled "Contacts" are formed from residues that have either "Wide Contacts" or "Close Contacts" (as noted in Table 1); the sets titled "Hydrogen Bonding" are formed from residues that have "Hydrogen Bonding" contacts; and the sets titled "Overlaps" are formed from residues that have either "Small Overlaps" or "Big Overlaps".

Contacts: (HOH402, HOH441, HOH458) (M40, N41, G66, A67, Y68, S72, I91, G117, A118, H142, W143) (W38, M40, Y68, S72, E90, I91, S119, Q120, W143, K144, L198, E199) (N41, V42, G66, A67, Y68, S72, I89, E90, I91, A118, S119, H142, W143) (W38, M40, V42, Y68, Y71, S72, I89, E90, I91, S119, F139, H142, W143, K144, R146, P174, L198, E199)

Hydrogen Bonding: (G66) (Q120) (HOH441) (E90, K144) (V42, S72, S119)

Overlaps: (A118, H142) (HOH402, HOH441) (M40, S72, E199) (V42, G66, S119) (M40, Y68, Y71, S72, I91, H142, W143, K144, L198, E199)

Additional Discussion

Catechol Denticity

Of the 550 models examined in this work, 8 models (2% of all) optimized to monodentate catecholate structures (Figure 3) where the O1 oxygen of CAT is not bound to the Mg^{2+} (Mg-O1 distance > 3Å). Monodentate catecholates have been examined by Kulik et al.[127] and identify the difference in both activation and reaction free energies for mono- and bidentate arrangements within 1 kcal/mol of each other. Likewise, the activation and reaction thermodynamics for our monodentate models are not significantly different from bidentate models.

Sulfur and Magnesium Basis Set Benchmarking

A noncomprehensive benchmark of the impact of including polarization functions on sulfur (atom of SAM; directly involved with the methyl transfer) and magnesium (binds to CAT substrate; indirectly involved with the methyl transfer) was run on two different model types. Model 1 is a 254-atom, *RINRUS*-designed model composed of the residues present in >90% of the models within cluster 5; Model 2 is 306-atom, *RINRUS*-designed model composed of the residues present in >70% of the models within cluster 5 (see Table 2). The models were run using the same methodology mentioned previously (see QM-model Construction subsection Computational Methods) though with differing sulfur and magnesium basis sets. The results are shown below in Table 3 and illustrate that including polarization functions (present in the 6-31G(d') basis set but not in LANL2DZ) on sulfur are crucial for obtaining a result closer to experimental accuracy, but including them on magnesium has no significant effect.

Table 3. Free energies of activation and reaction for two models using differing basis sets for their Sulfur and Magnesium atoms

Model	Sulfur Basis Set	Magnesium Basis Set	$\Delta G^{\ddagger}$ (kcal/mol)	ΔG_{rxn} (kcal/mol)
Model 1	6-31G(d')	LANL2DZ	13.0	−8.1
	6-31G(d')	6-31G(d')	13.1	−8.1
	LANL2DZ	LANL2DZ	5.5	−17.6
Model 2	6-31G(d')	LANL2DZ	13.2	−3.8
	6-31G(d')	6-31G(d')	13.2	−4.0
	LANL2DZ	LANL2DZ	5.9	−13.1

Survey for Cartesian Coordinates within SI

Reporting the Cartesian coordinates for the starting and/or final model structures is one of the simplest and easiest ways to ensure others may be able to replicate, analyze, or utilize the models used in a study. Nevertheless, it is not necessarily a common practice for scientists to report their model structures, even if they are providing other data and supplementary information. To preview the frequency that protein or enzyme model Cartesian coordinates are being reported in supplementary materials, we conducted a survey of articles conducting QM-only, QM/MM, and ONIOM computations. The list of 148 entries was obtained using the Web of Science citation database search, filtering for the keywords “QM/MM” and “ONIOM” in manuscripts published between 1 January 2015 – 31 March 2015 and 1 January 2019 – 31 March 2019, along with grabbing articles published between the same range that cited either of two prominent QM-cluster works *Transition-Metal Systems in Biochemistry Studied by High-Accuracy Quantum Chemical Methods* by Siegbahn and Blomberg (doi: 10.1021/cr980390w)[49] or *Modeling Enzymatic Reactions Involving Transition Metals* by Siegbahn and Borowski (doi: 10.1021/ar050123u)[189]. A total of 90 entries from this list were excluded from this survey due to a variety of factors, the main one being that the actual publication dates were outside the desired ranges. Other factors that resulted entry exclusion include the entry not being an actual journal article (e.g. a journal supplement, erratum or review paper), the system of interest was not of proteins (e.g. studying inorganic metal clusters or lone molecules in solvent), or the study did not directly involve computation of QM-cluster or QM/MM models. Of the remaining 58 journal articles, 51 (88%) reported a supplementary information document of some kind, but only 20 (34%) reported Cartesian coordinates for any of their structures. This information is tabulated in data files that will become available at the publication of this work.

Appendix D: Chapter 5 Supplementary Information

Model cartesian coordinates and additional data are available at doi:10.1021/acs.jcim.9b00804

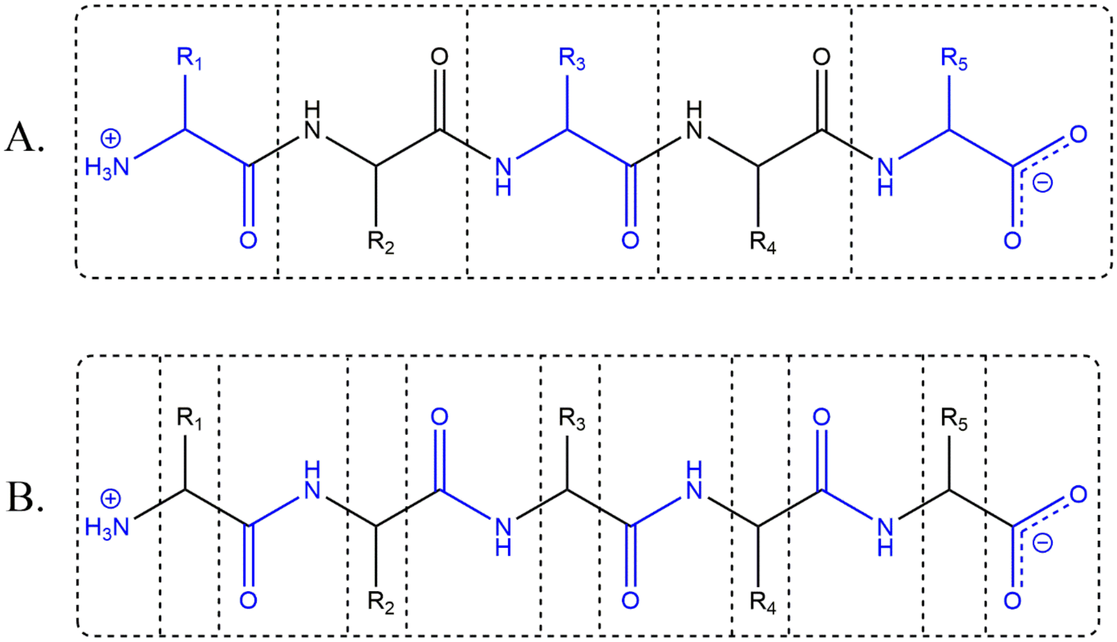


Figure 1. Atomic partitioning of a five-amino-acid peptide in terms of amino acids (A) and chemical functional groups (B).

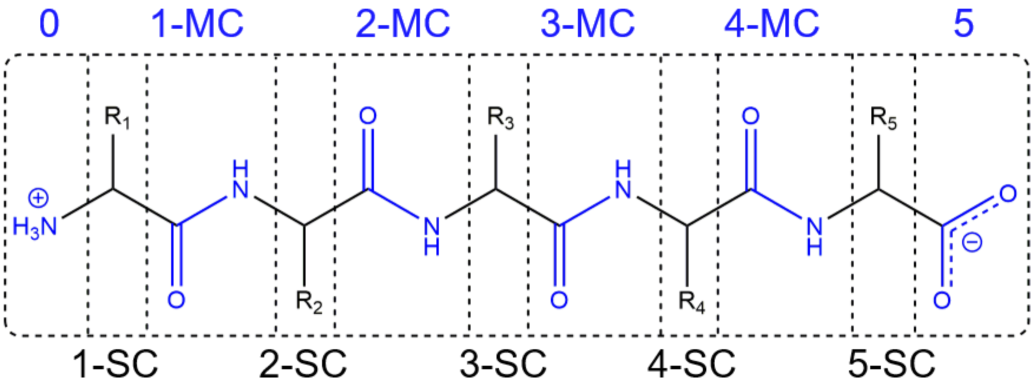


Figure 2. Example of the naming scheme used for this work on a five-residue peptide chain. The peptide is partitioned in terms of chemical main chain (MC) and side chain (SC) functional groups, and naming begins at the N-terminus.

Additional Information on Data Cleaning and F/ISAPT Model Construction

In generating the datasets for this work, there are several stages where additional steps were taken either to reduce the complexity of the modeling problem or to exclude interactions that would be or were infeasible to compute using the SAPT method detailed in this work.

Terminal Residues

As issues arise in how SAPT handles the charged amine and carboxylate functional groups of the N- and C- terminal residues, and as the PDBs 1UAI and 2EA3 are missing the N-terminus and C-terminus, respectively, interactions involving the N-terminus or SC of the first residue or the C-terminus or SC of the last residue resolved in the crystal structure were not included. In the example peptide in Figure 2, the Functional Groups 0, 1-SC, 5-SC, and 5 would be excluded from consideration.

Adjacent Functional Groups

Contact network graphs may be designed to interconnect residues (or functional groups in this work) that are adjacent (covalently bound) in accordance with the primary structure of the protein. As adjacent functional groups are covalently bound, their interaction energies were not computed in this work.

Proline

Given the tertiary amine of proline, there will be issues computing the ISAPT interactions involving MCs composed of the proline-N and involving the proline SC. Interactions that would involve computing ISAPT energies for proline-N MC or proline

SC are excluded from consideration. Interactions that would involve computing FSAPT energies for proline SC or MC are included and are trimmed according to the rules shown in Figure 3I and 3J, respectively.

Cystines

Cystines are present in the PDBs 1UAI, 2EA3, and 3WY8. Although the covalently bound side chains are effectively one unit, for this work we treated the two residues forming them in their reduced cysteine form as two separate cysteine side chains. As an example, PDB 1UAI has a cystine connecting residues 200-SC and 206-SC. The interaction between 200-SC and 203-MC was computed using a cysteine side chain for 200-SC.

Interaction Energies Unable to be Computed

Of the total dataset there were a total of 8 ISAPT SC-SC interaction energies unable to be computed due to software complications that were unable to be resolved over the course of this work. These were specifically the following:

Table 1. Network residue pairs whose interaction energies were unable to be computed.

PDB	Functional Group IDs	Functional Group Types
1UAI	120-SC_122-SC	ASP-ASP
	149-SC_151-SC	ASP-THR
1YW5	112-SC_114-SC	SER-GLU
	64-SC_65-SC	GLU-ASP
256L	20-SC_22-SC	LEU-ILE
	22-SC_24-SC	GLU-TYR
	88-SC_89-SC	TYR-ASP
3WY8	137-SC_139-SC	ASN-GLU

Table 2. General network information of the tested protein models.

PDB	Number of Residues	Number of Nodes	Number of Edges			
			MC-MC	MC-SC	SC-SC	Total
1UAI	223	430	233	369	408	1010
1YW5	177	336	234	273	287	794
256L	164	319	315	356	262	833
2EA3	183	347	200	281	299	780
3WY8	219	423	252	338	389	979

Table 3. Distribution of Interaction Data based on interaction charge and chemical type.

	1UAI	1YW5	256L	2EA3	3WY8
Interaction Charge					
MC-MC	233	234	315	200	252
MC-NEG	28	24	25	9	13
MC-NEU	316	219	189	257	310
MC-POS	25	30	42	14	15
NEG-NEG	3	1	1	0	0
NEG-NEU	43	33	30	19	23
NEG-POS	8	13	15	2	3
NEU-NEU	296	187	157	249	329
NEU-POS	48	49	55	26	30
POS-POS	6	2	1	1	1
Chemical Type					
ALI-ALI	60	44	56	65	52
ALI-ARO	67	41	41	48	74
ALI-NEG	12	16	13	6	8
ALI-POL	78	43	30	76	79
ALI-POS	14	25	28	9	11
ARO-ARO	20	12	1	7	15
ARO-NEG	17	4	7	4	8
ARO-POL	42	24	16	33	62
ARO-POS	15	8	12	4	7
MC-ALI	137	95	96	120	110
MC-ARO	75	42	29	38	80
MC-MC	233	234	315	200	252
MC-NEG	28	24	25	9	13
MC-POL	104	82	64	99	120
MC-POS	25	30	42	14	15
NEG-NEG	3	1	1	0	0
NEG-POL	14	13	10	9	7
NEG-POS	8	13	15	2	3
POL-POL	29	23	13	20	47
POL-POS	19	16	15	13	12
POS-POS	6	2	1	1	1
Total Interactions	1006	792	830	777	976

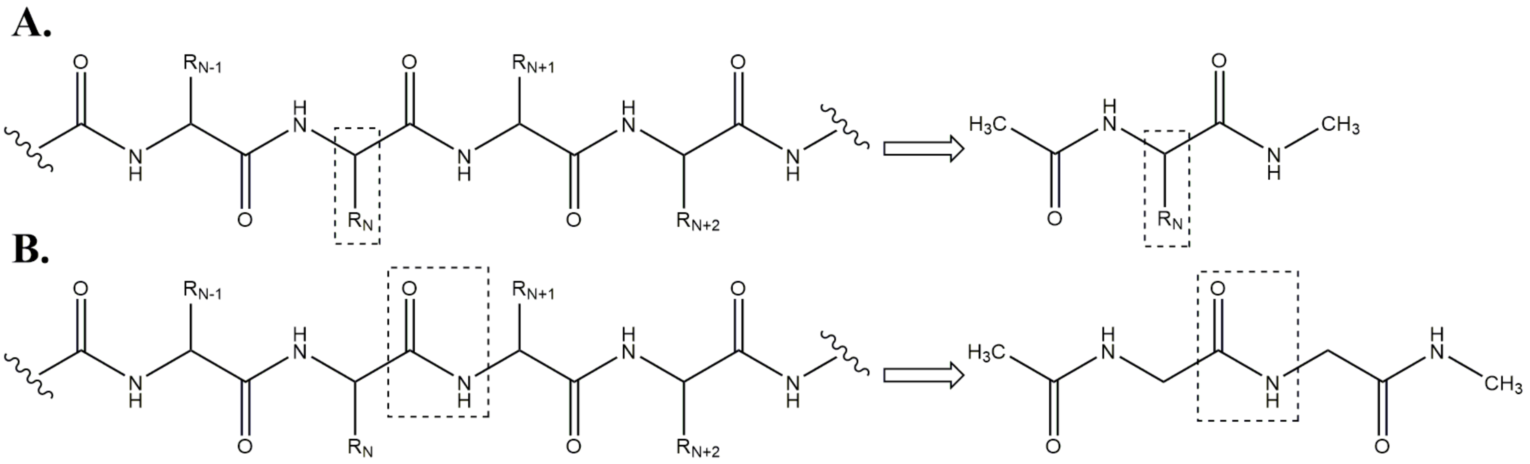


Figure 3. Model fragmentation schemes for computing the FSAPT interaction energy of a side chain (A) and main chain (B) with a second side/main chain

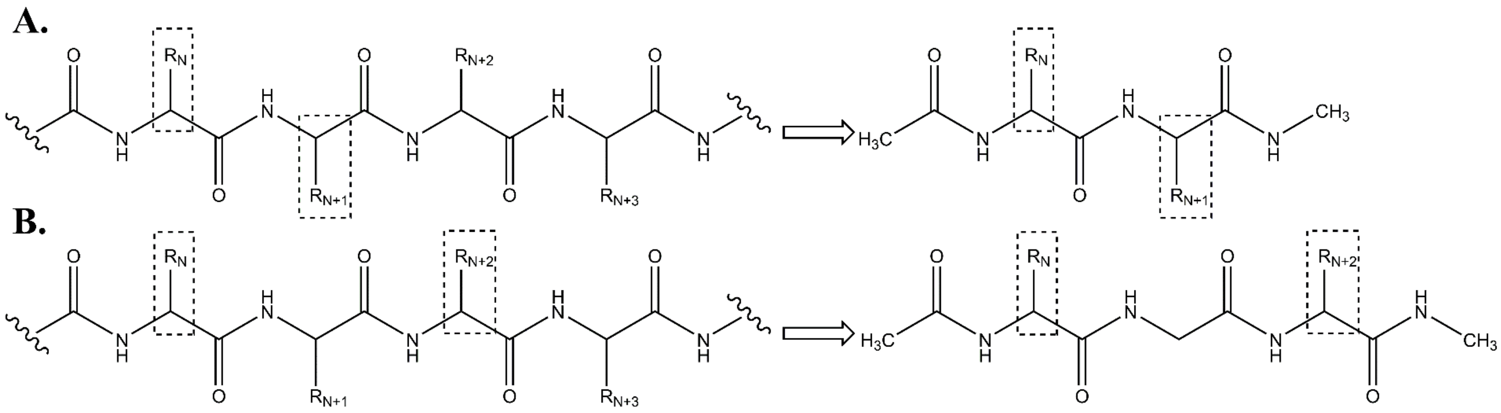


Figure 4. Model fragmentation schemes for computing the ISAPT interaction energy between the indicated side chains that are separated by exactly one (main chain) functional group (A) or exactly three (one side chain, two main chain) functional groups (B).

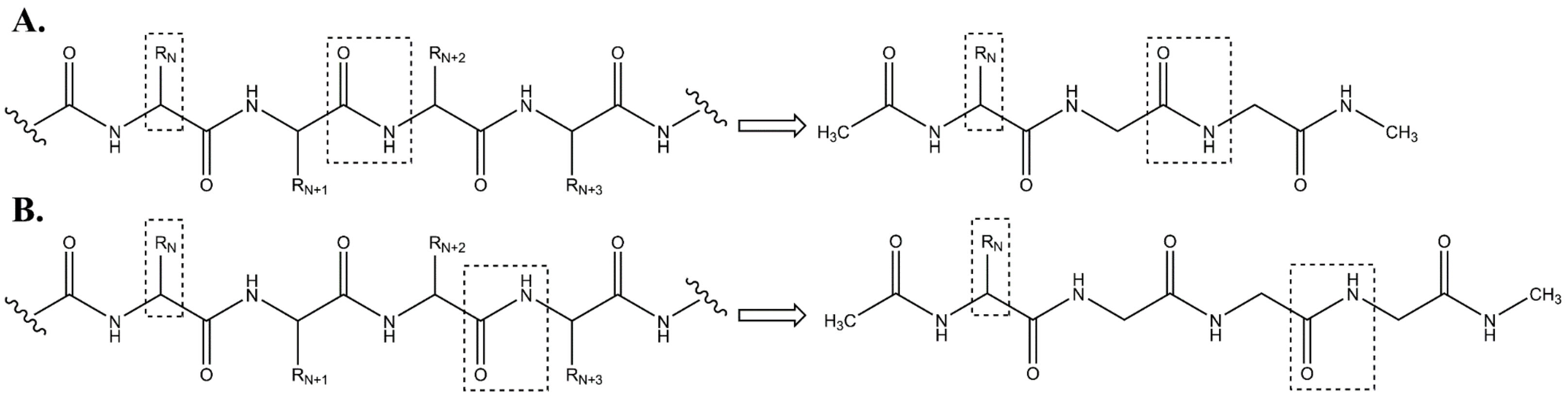


Figure 5. Model fragmentation schemes for computing the ISAPT interaction energy between the indicated side and main chains that are separated by exactly two (one main chain, one side chain) functional groups (A) or four (one main chain, one side chain) functional groups (B).

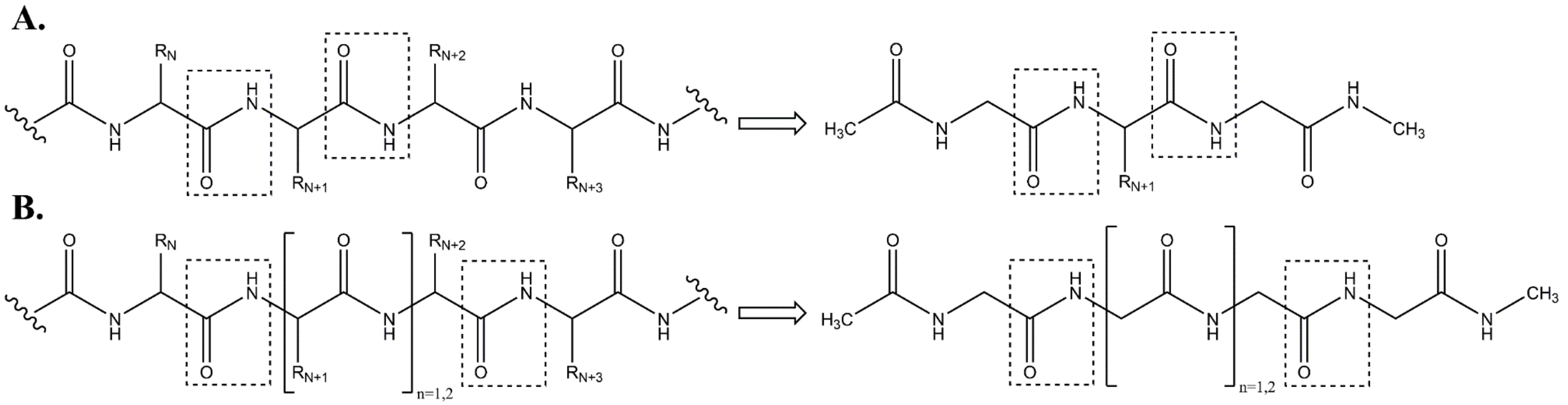


Figure 6. Model fragmentation schemes for computing the ISAPT interaction energy between the indicated main chains that are separated by exactly one (side chain) functional group (A) or either 3 (one main chain, two side chain) or five (two main chain, three side chain) functional groups (B).

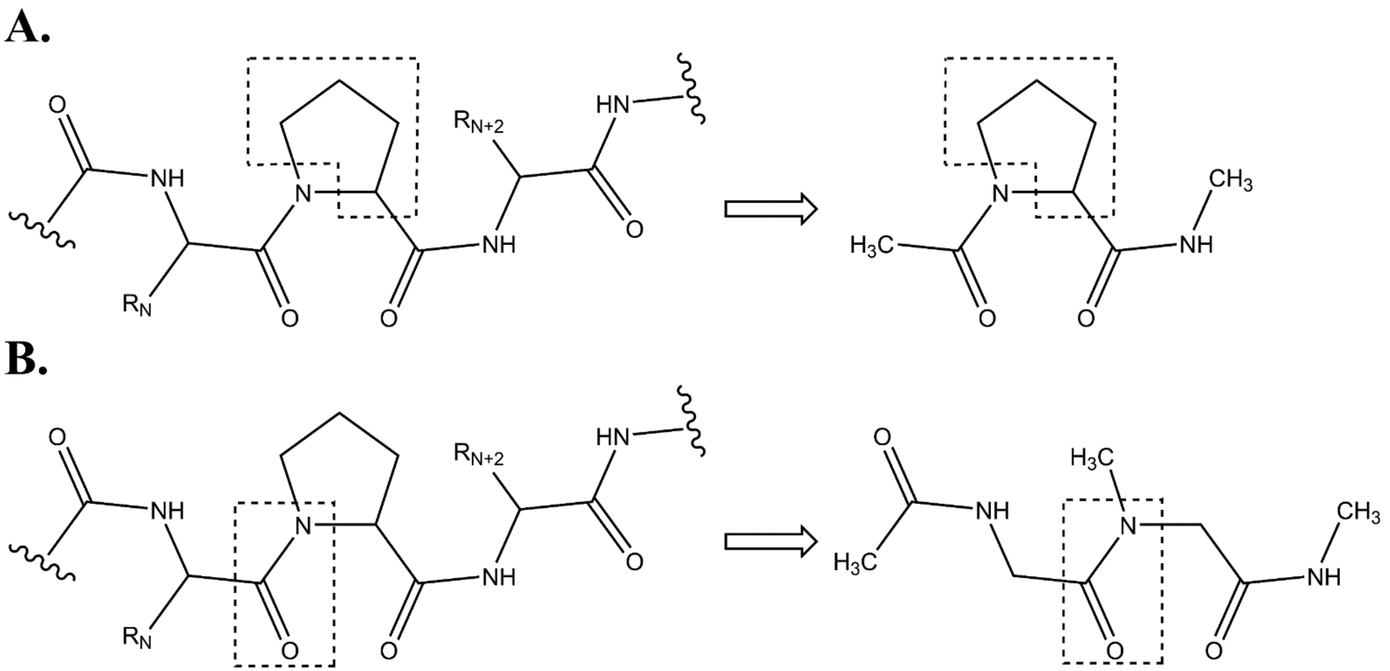


Figure 7. Model fragmentation scheme for computing the FSAPT interaction energy involving a proline side (A) or main (B) chain.

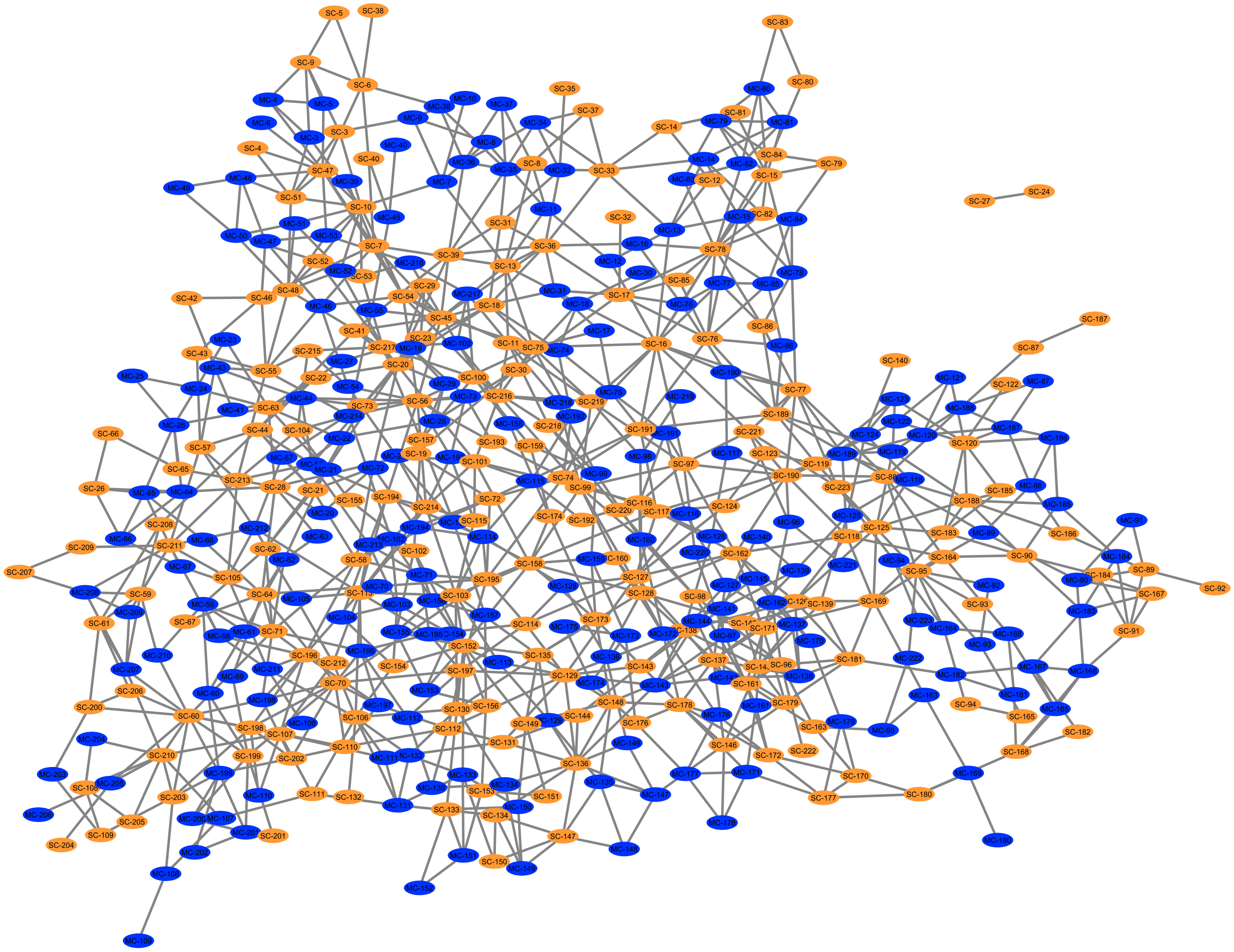


Figure 8. Graph of the functional group network of PDB 1UAI. Main chains are colored blue; side chains are colored orange.

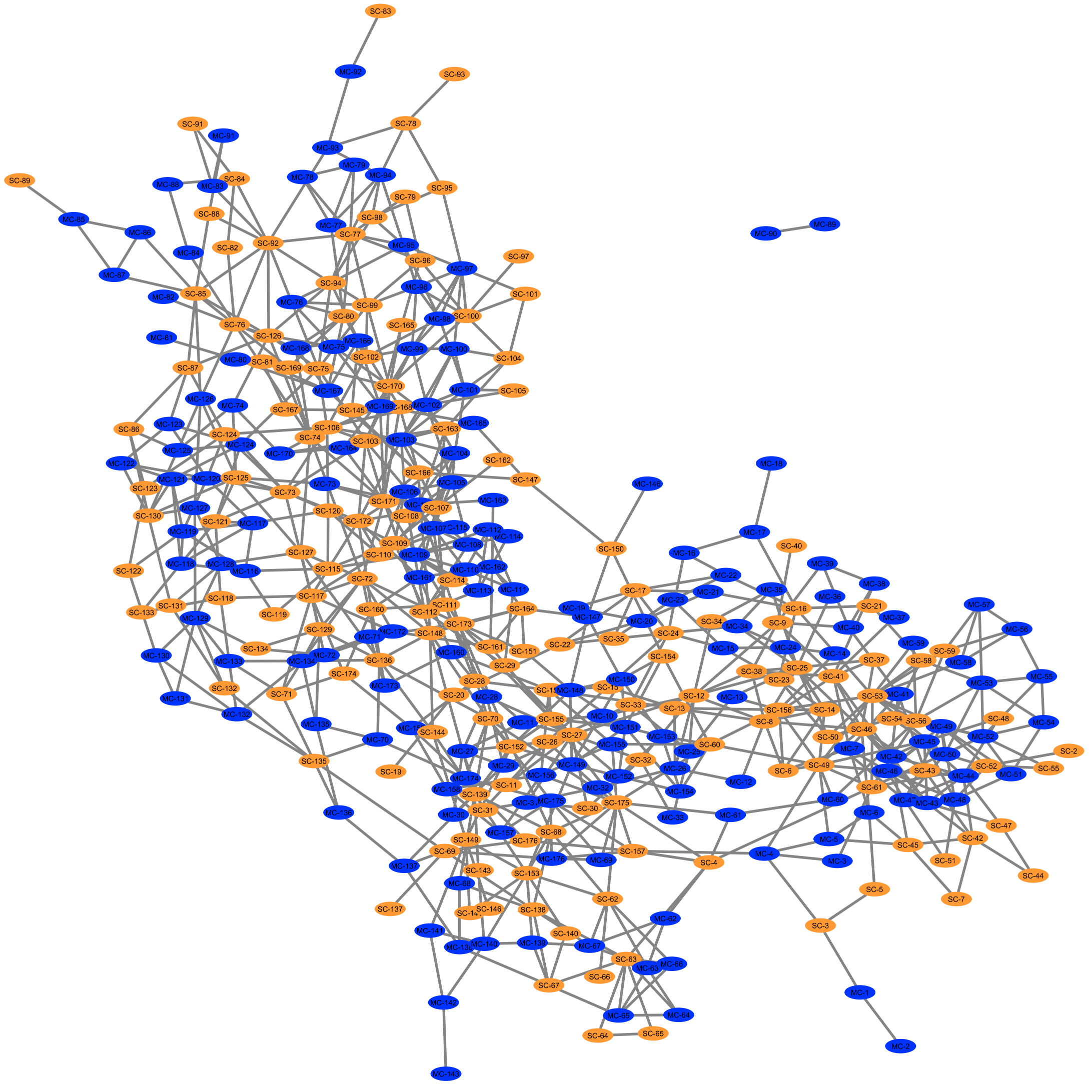


Figure 9. Graph of the functional group network of PDB 1YW5. Main chains are colored blue; side chains are colored orange.

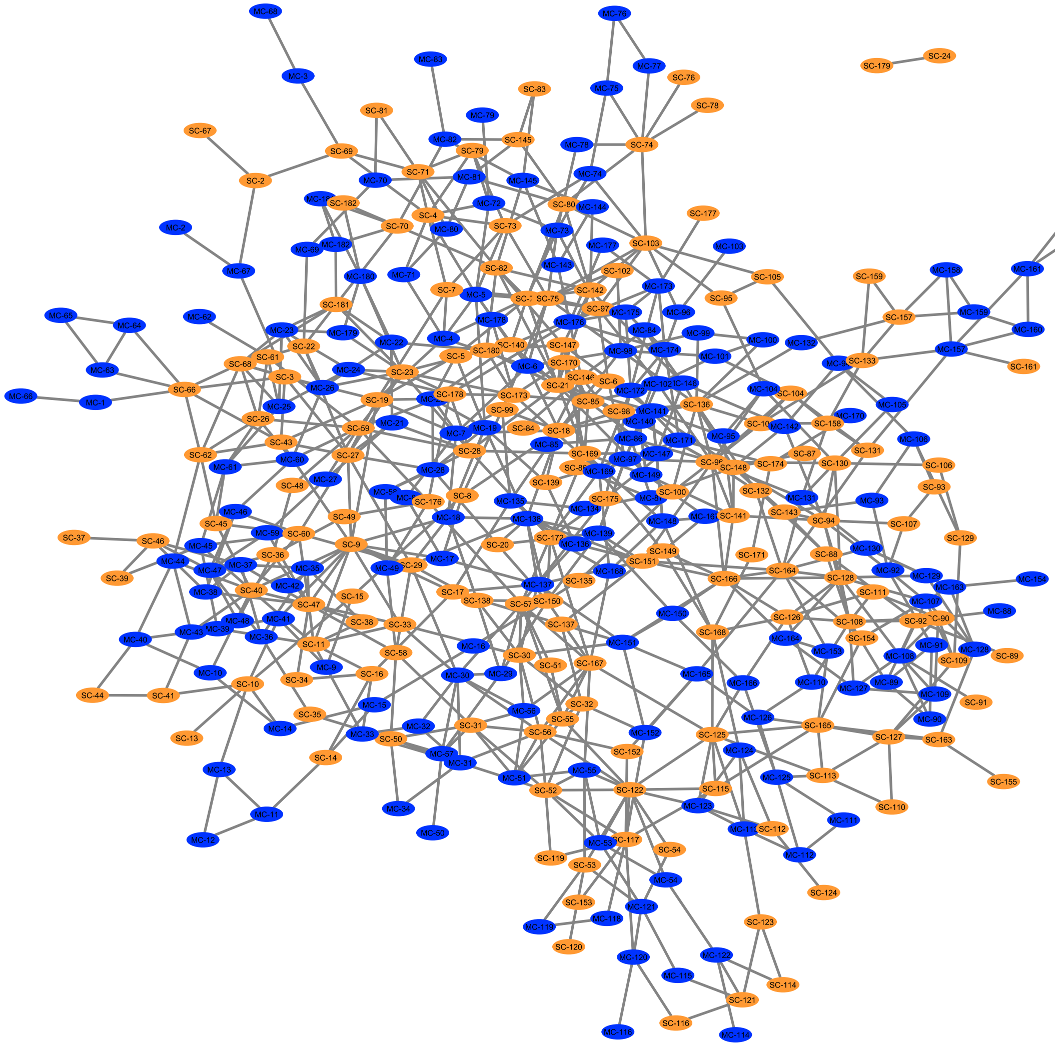


Figure 10. Graph of the functional group network of PDB 2EA3. Main chains are colored blue; side chains are colored orange.

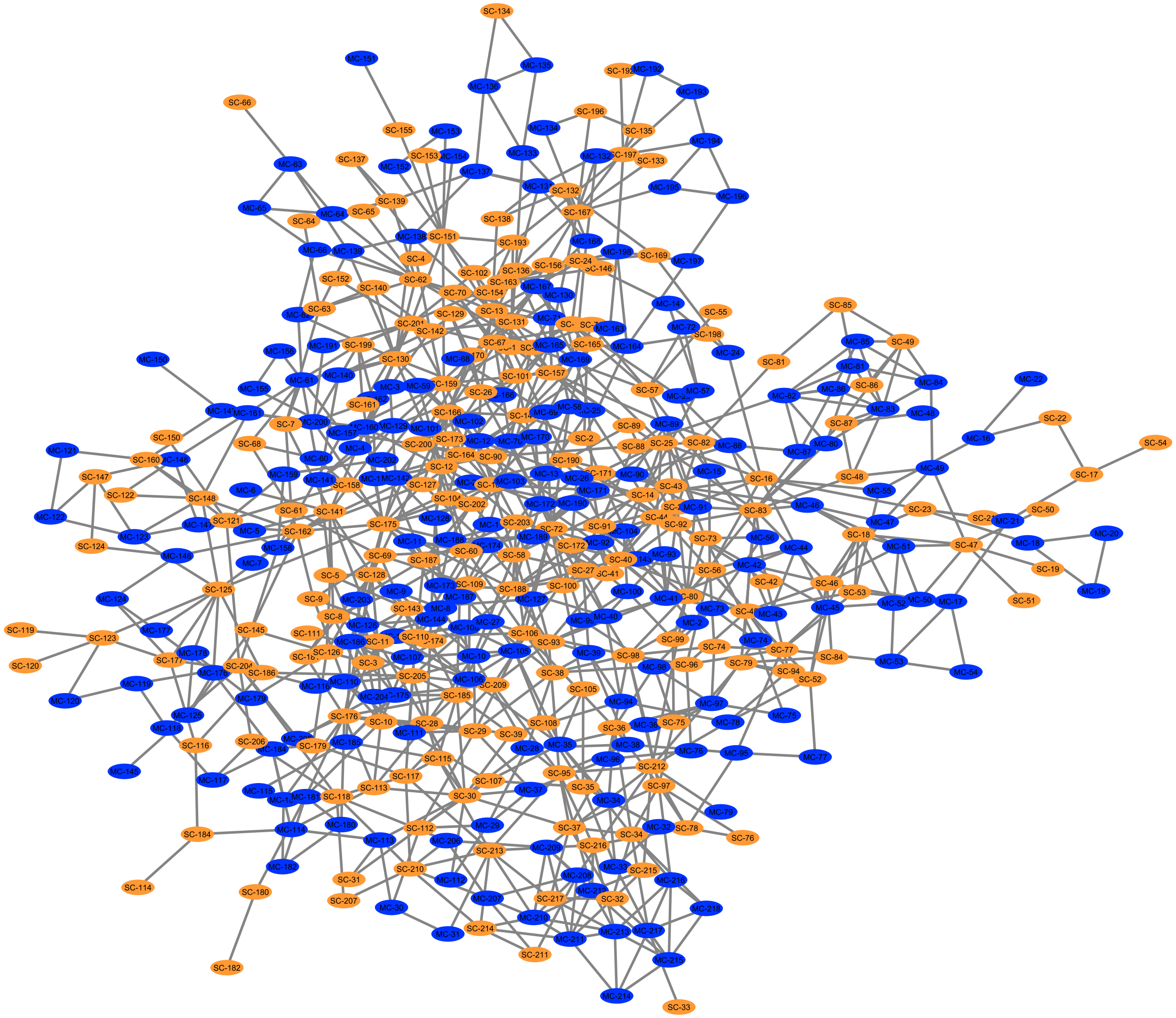


Figure 11. Graph of the functional group network of PDB 3WY8. Main chains are colored blue; side chains are colored orange.

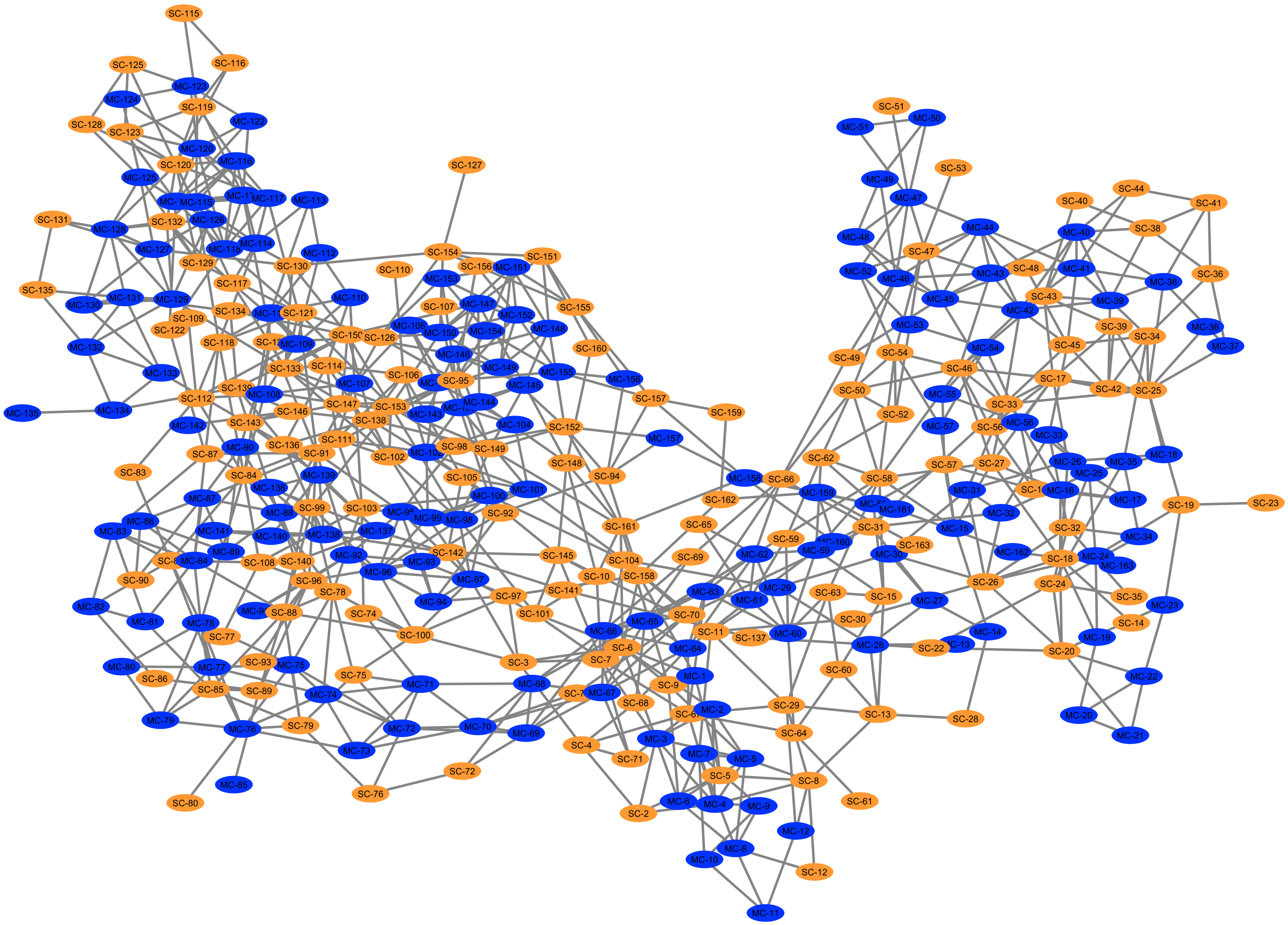


Figure 12. Graph of the functional group network of PDB 256L. Main chains are colored blue; side chains are colored orange.

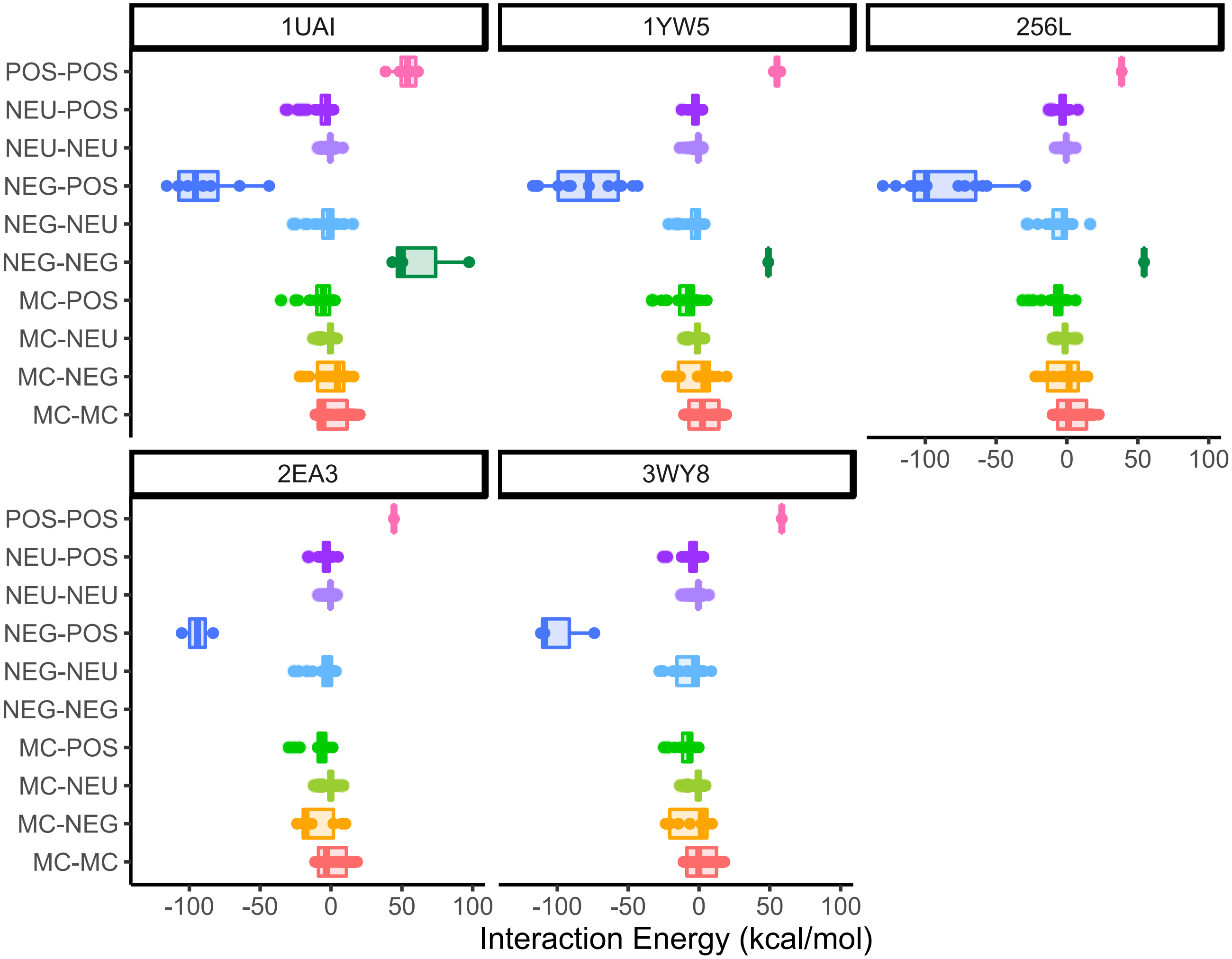


Figure 13. Distribution of interaction energy data among the test set. Color and partitioning is based upon the interaction charge of the two species. *MC* refers to main chains, and *POS*, *NEU*, and *NEG* refer to positive, neutral, and negative side chains, respectively.

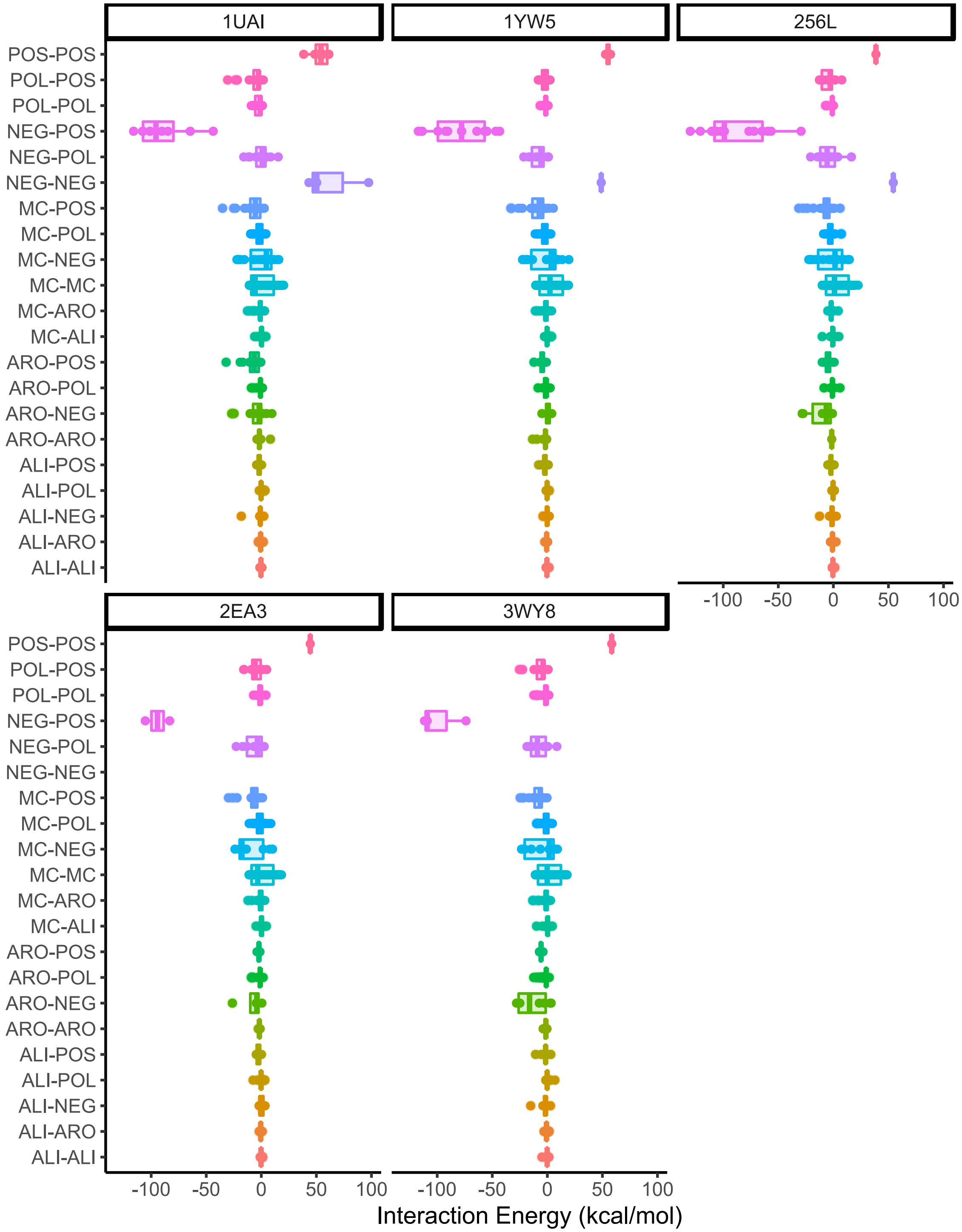


Figure 14. Distribution of interaction energy data among the test set. Color and partitioning is based upon the interaction type of the two species.

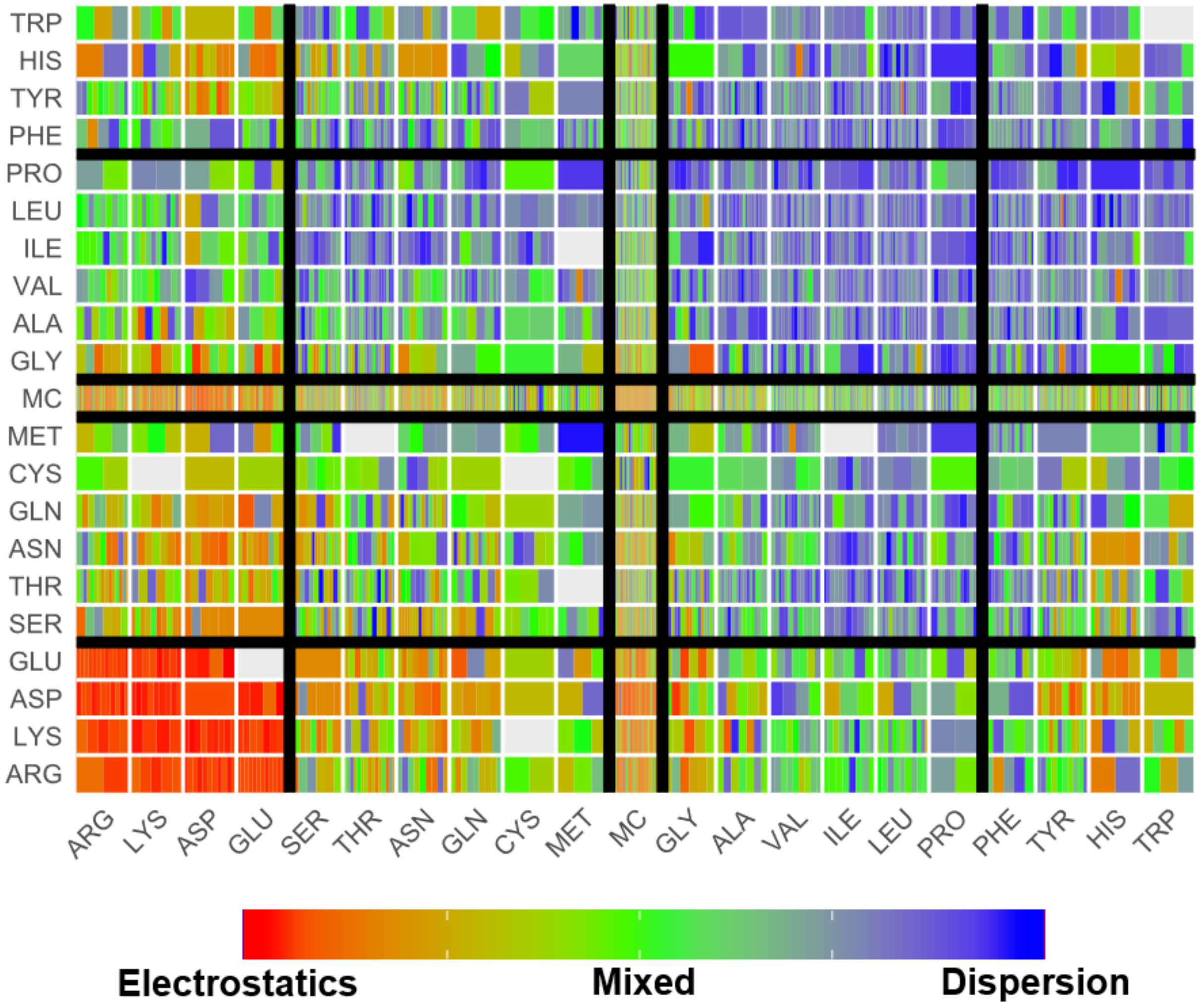


Figure 15. Grid of computed main and side chain pair interactions colored according to the proportion electrostatics and dispersion SAPT decomposition terms contribute to the interaction energy.

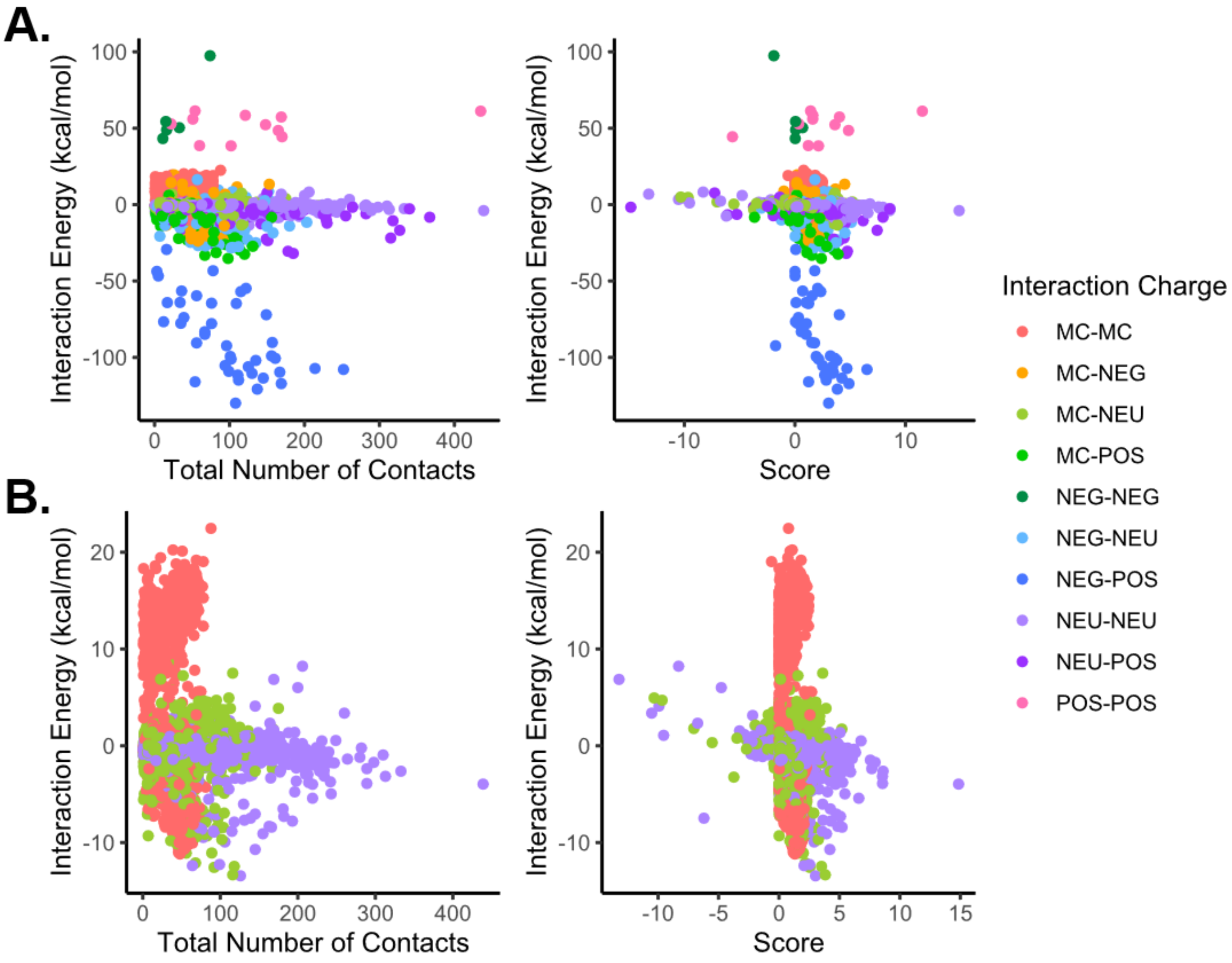


Figure 16. A) Correlation plots comparing the *Probe*-computed total number of contacts and score against SAPT-computed interaction energies. B) Correlation plots showing only the neutral-charged interactions.

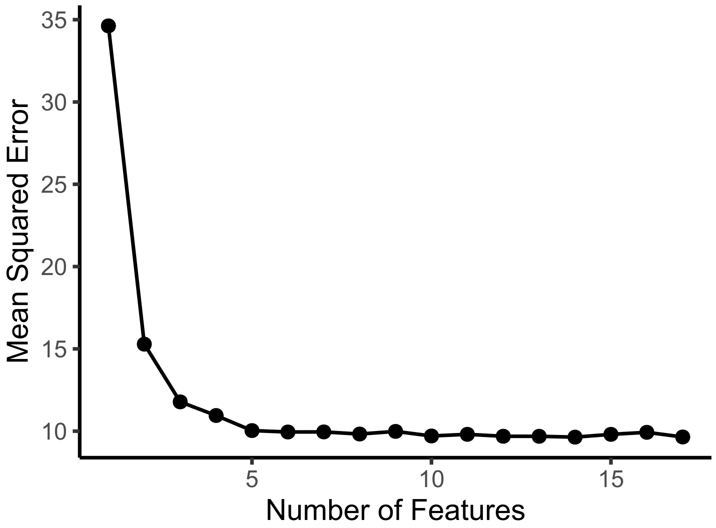


Figure 17. Convergence of mean squared error as the number of features tested at each node is increased (500 trees used in each forest test).

Table 4. The importance of the functional group (FG) descriptors in the training set random forest as determined by the increase in mean square error.

Descriptor Abbreviation	Descriptor Name	% Increase in Mean Square Error
IntType	Chemical Type	69.0
IntCharge	Interaction Charge	66.6
CoMDist	Center of Mass Distance	43.9
HBcount	Hydrogen Bonding Contacts	24.5
SeqDist	Sequence Distance	23.7
Type2	FG 2 Type	9.96
Func1	FG 1 Name	8.16
Func2	FG 2 Name	6.13
Score	Score	6.11
TotalCount	Total Contacts	3.43
Type1	FG 1 Type	3.41
WCcount	Wide Contacts	2.59
CCcount	Close Contacts	2.11
SOcount	Small Overlap Contacts	1.85
Pos1	Position of FG 1	0.16
Pos2	Position of FG 2	0.15
BOcount	Bad Overlap Contacts	0.02

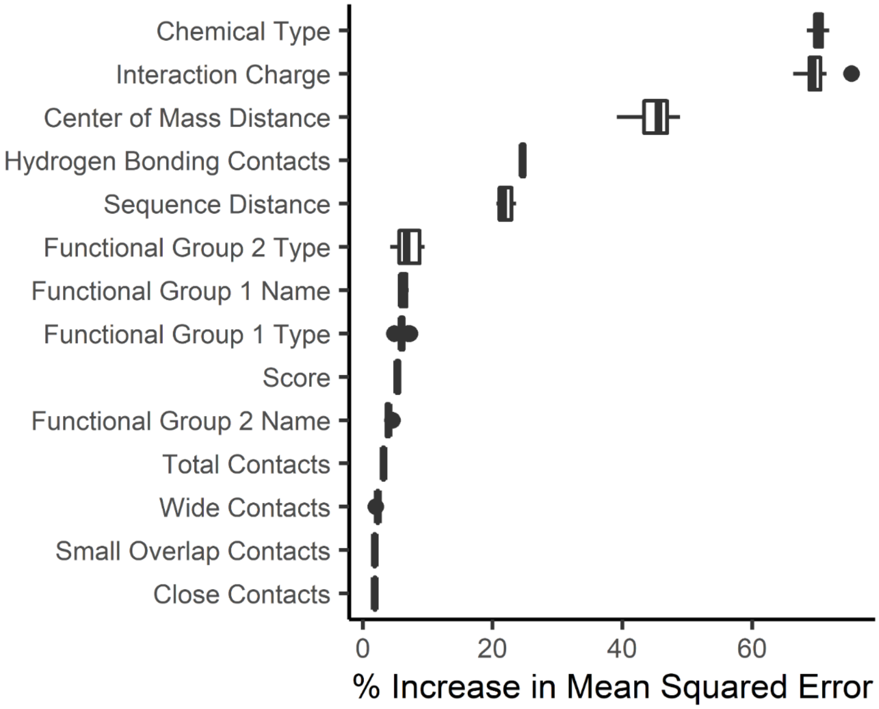


Figure 18. The importance of descriptors for the validation random forest as determined by the increase in mean square error.

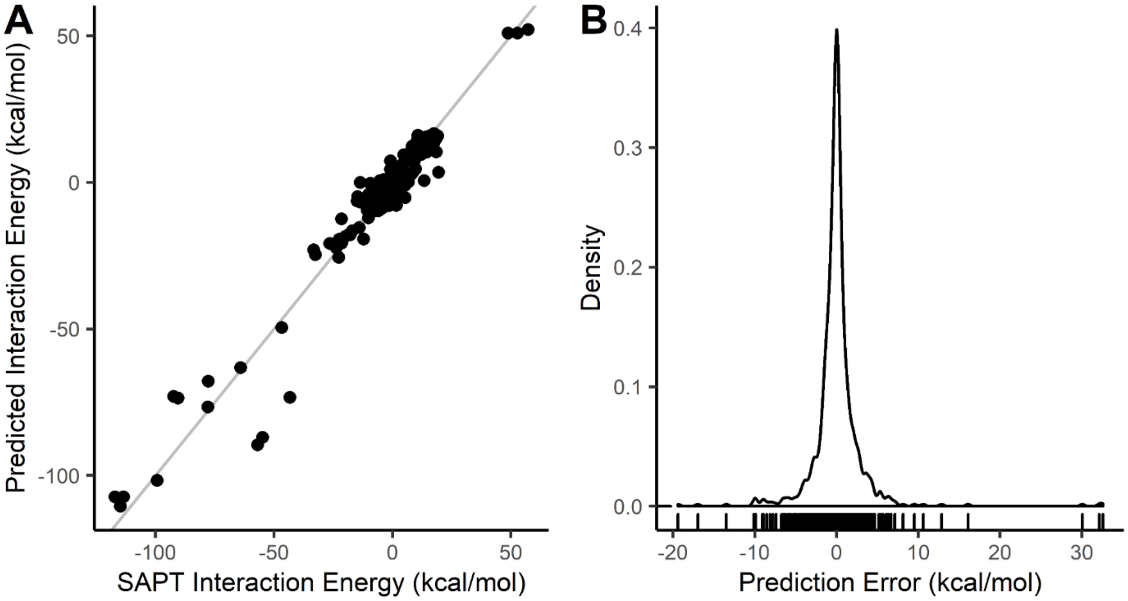


Figure 19. A) Plot of SAPT-computed vs random forest-predicted interaction energy. The grey line represents the line of equality where RF-predicted energies would equal SAPT-computed energies. B) Density plot of differences between SAPT-computed and RF-predicted energies.

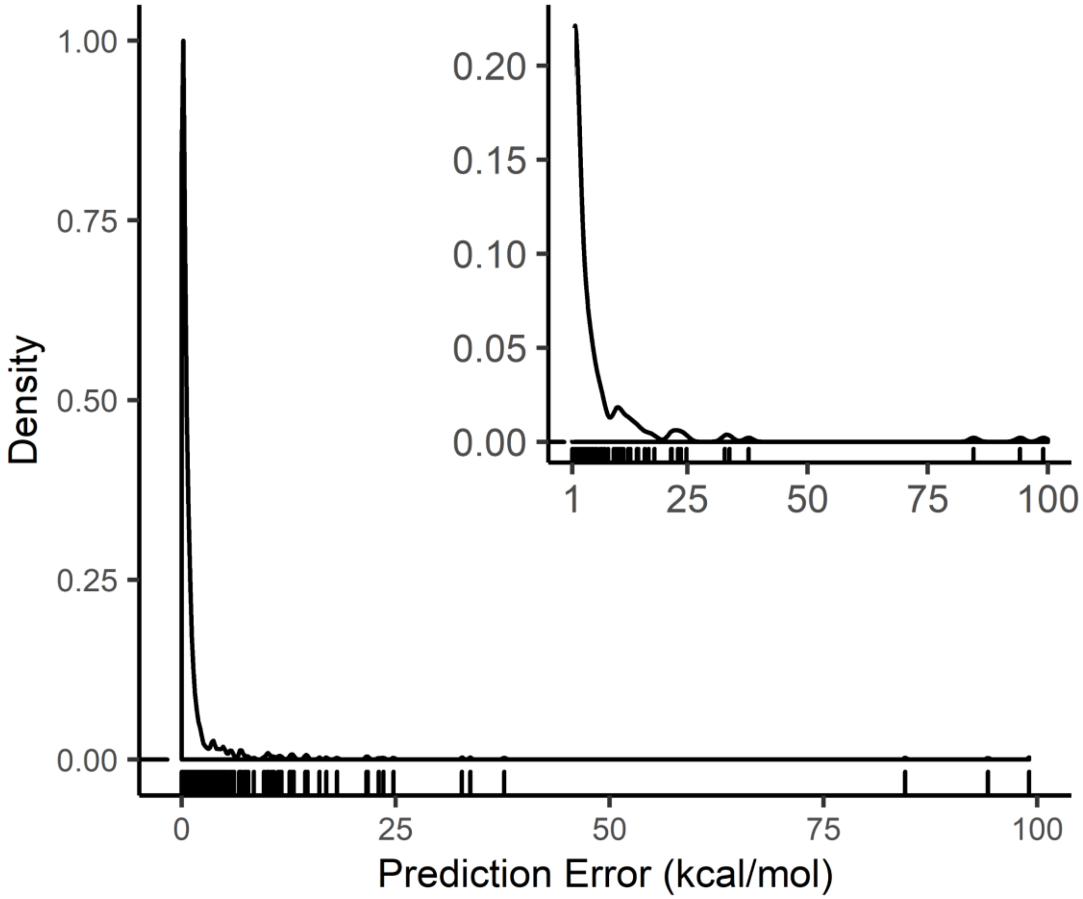


Figure 20. Density plot of relative error differences between SAPT-computed and RF-predicted energies

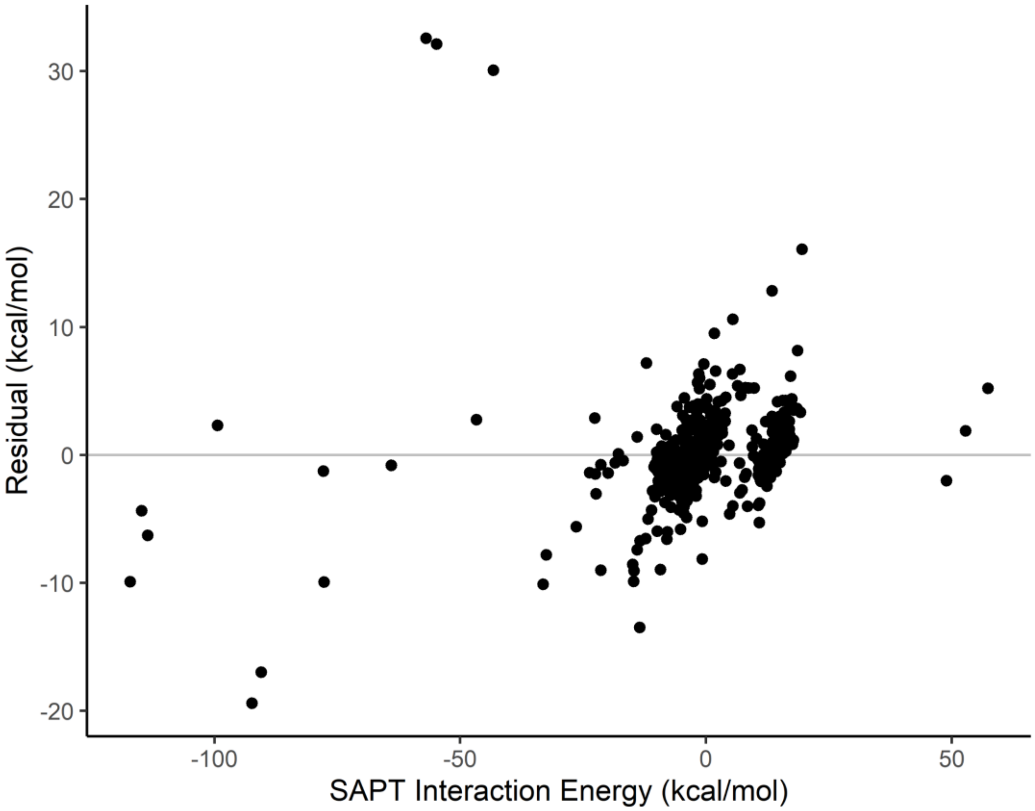


Figure 21. Plot of the residuals (difference between the SAPT-computed interaction energy and the RF-predicted interaction energy) for the 1YW5 test case.

Table 5. Distribution of Prediction Error for 1YW5 by Model Charge.

Model Charge Type	Mean Error (kcal/mol)	Standard Deviation (kcal.mol)	Mean Absolute Error (kcal/mol)	Number of Interaction Energies with Incorrect Sign
MC-MC	−0.18	1.9	1.3	11
MC-NEG	−1.0	6.2	4.4	2
MC-NEU	0.14	1.7	1.2	56
MC-POS	1.3	5.3	4.4	2
NEG-NEG	[1]2.0	[1]--	[1]--	0
NEG-NEU	0.37	3.2	2.3	12
NEG-POS	−2.4	17.9	13.0	0
NEU-NEU	0.05	1.1	0.62	19
NEU-POS	−0.54	2.8	2.1	10
POS-POS	[1]−3.6	[1]--	[1]--	0
Entire Data Set	−0.05	3.2	1.6	112

[1] NEG-NEG consists of one data point and POS-POS consists of only two data points so mean absolute error and standard deviation are inappropriate to report. Their respective mean errors are reported for reference, though it should not be held comparable to the other Charge Type values due to the sparsity of their datasets.