Fall 2021

Computational Enzymology on Sulfur-Containing Enzymes: From Method to Application

Paul Meister
University of Windsor

Computational Enzymology on Sulfur-Containing Enzymes: From Method to Application

By

Paul Meister

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Chemistry and Biochemistry
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Windsor

Windsor, Ontario, Canada

2021

Computational Enzymology on Sulfur-Containing Enzymes: From Method to Application

by

Paul Meister

APPROVED BY:

_____

G. Peslherbe, External Examiner

Concordia University

_____

S. Goodwin

School of Computer Science

_____

O. Vacratsis

Department of Chemistry and Biochemistry

_____

J. Trant

Department of Chemistry and Biochemistry

_____

J. Gauld, Advisor

Department of Chemistry and Biochemistry

May 21, 2021

Declaration of Co-authorship / previous publication

I. Co-Authorship

I hereby declare that this dissertation incorporates material that is result of joint research, as follows:

Chapter 3 of this dissertation was equally co-authored with Dr. Sahar Nikoo under the supervision of Prof. James W. Gauld. In all cases, the key ideas, primary contributions, experimental designs, data analysis, interpretation, and writing were performed by both co-authors, and the contribution of co-authors. The written manuscript was edited in collaboration with Dr. John J. Hayward under supervision of Prof. James W. Gauld.

Chapter 5 of this dissertation was equally co-authored with Dr. Bogdan F. Ion under the supervision of Prof. James W. Gauld. Key idea, primary contributions, experimental design, data analysis, and interpretation was performed by both co-authors. I was primarily responsible for the writing of the manuscript.

Chapter 7 of this dissertation was performed in collaboration with Aislyn Laurent including: primary contributions, experimental design, data analysis, and writing under the supervision of Prof. James W. Gauld.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my dissertation and have obtained written permission from each of the co-author(s) to include the above material(s) in my dissertation.

I certify that, with the above qualification, this dissertation, and the research to which it refers, is the product of my own work.

II. Previous Publication

This dissertation includes two original papers that have been previously published/submitted for publication in peer reviewed journals, as follows:

Dissertation Chapter	Publication title/full citation	Publication status*
Chapter 3	Nikoo, S., Meister, P. J., Hayward, J. J., Gauld, J. W. *Molecules* **2018**, 23, 3323.	Published

Chapter 5	Ion, B. F., Meister, P. J., Gauld, J. W. *J. Phys. Chem. A* **2019**, 123, 7710-19.	Published

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my dissertation. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.

III. General

I declare that, to the best of my knowledge, my dissertation does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my dissertation, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my dissertation.

I declare that this is a true copy of my dissertation, including any final revisions, as approved by my dissertation committee and the Graduate Studies office, and that this dissertation has not been submitted for a higher degree to any other University or Institution.

Abstract

Sulfur-containing biomolecules display incredible functional diversity. Indeed, in addition to thiols and thioethers, S-nitrosothiols, 3,4-coordinate, sulfoxides, persulfides and now even polysulfides are commonly observed intermediates. Unfortunately, however, their biological synthesis and roles remain poorly understood. In addition, sulfur-containing species can access a broad range of oxidation states and thus can act as either an electrophile or nucleophile giving rise to an even more diverse set of sulfur-derived functional groups. However, these unique properties can lead to difficulties in characterizing such compounds experimentally and reinforces the need for computational studies to reliably predict their structural and energetic properties. In this dissertation, we have applied a broad range of computational methodologies to important sulfur-containing biomolecules to assess their applicability and to elucidate the chemistry of such species.

The first two chapters highlight some of the roles of sulfur and challenges associated with studying sulfur systems. We detail current methods in computational enzymology, discussing examples of their application and outline the next generation of approaches.

In **Chapter 3**, we demonstrate the importance of selecting a proper computational method for application to biorelevant sulfur/selenium-containing molecules. We conduct a detailed benchmark study and find that sulfur-sulfur bonds in particular are sensitive to changes in basis set. The ωB97XD/6-311G(2d,p) level of theory is found to be the most accurate and reliable for obtaining geometries and energetics.

We apply the knowledge gained from these benchmark studies in **Chapter 4** to study biological persulfide/polysulfide formation pathways. Radical-containing charged intermediates are commonly proposed which can prove challenging for DFT methods to accurately describe. These species were found to be transient due to small contributions towards their stability from high-energy Rydberg states. Extending to an active site model

showed that sulfur-sulfur bond formation is favoured with radical intermediates, particularly when electron transfer cofactors such as FAD are involved.

Chapter 5 examines the catalytic mechanism of maleamate amidohydrolase (NicF) which converts an amide substrate to a carboxylic acid with assistance of a cysteine nucleophile. The reaction took place in two stages: a stepwise pathway for release of ammonia by concerted attack of a water molecule to form the acid product. Stabilization of the oxyanion hole by a critical threonyl is thought to facilitate the reaction instead of a metal ion. Then, in **Chapter 6**, several approaches for selecting a suitable structure from molecular dynamics simulations for use in QM/MM calculations were examined. Building upon our mechanistic studies on NicF in **Chapter 5**, we show that poor structure selection can result in problematic surfaces due to the erroneous energetic contributions from the MM layer.

To help alleviate some of the user burden, **Chapter 7** builds upon the conclusions of **Chapter 6** to investigate the ability of artificial intelligence to guide the structure selection process. The program "Pose Selector" was created which uses agglomerative machine learning to select a set of structures for analysis. Evaluation of this approach was done on the NicF system and revealed that this approach can be used to select a structure, and that different catalytic pathways may be identified.

Finally, in **Chapter 8**, we apply these combined learnings to the enzyme cysteine desulfurase (SufS) and examine how the active site prepares for sulfur transfer to an active site cysteinyl to produce an enzyme-bound persulfide. Protonation states of the key active site residues were clarified with dynamic pK_a calculations on these residues.

Throughout this dissertation we developed new protocols and demonstrated their application to sulfur-containing biomolecular systems. These insights can be extended beyond the realm of computational enzymology and may have impact in other fields such as the design of sulfa-drugs and calculating their properties using these approaches.

Acknowledgements

First and foremost, I would like to extend my gratitude to my supervisor Prof. James W. Gauld. Throughout the years, he has been incredibly supportive and available to provide guidance whenever it was required. With the onset of COVID-19, there were a lot of difficulties encountered, but he remained patient and we were able to persevere to the end of the process.

I would also like to thank my committee members, Prof. Vacratsis, Prof. Trant and Prof. Goodwin. Our discussions have always been valuable to help better the work in this thesis. Thank you all for taking your time to evaluate this dissertation. I would also like to thank Prof. Peslherbe for taking the time to be the external examiner for my committee.

A huge thanks to all past and present members of the Gauld group with whom I had the pleasure of meeting. Our group discussions have always been insightful. Thank you to Dr. John Hayward and Dr. Sahar Nikoo for their collaboration efforts in the publication of Chapter 3 and Aislyn Laurent for her expertise in all things related to computer science and really driving forward our machine learning project in Chapter 7.

Furthermore, I would also like to thank my parents for being incredibly supportive. Without them, producing this body of work would have been significantly more challenging.

Lastly, I extend my gratitude to the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ontario Graduate Scholarship (OGS) Program for financial support as well as Compute Canada and Sharcnet for providing not only valuable computational resources for these projects, but also for their wonderful technical support.

Table of Contents

List of Tables

List of Figures

List of Schemes

List of Appendices

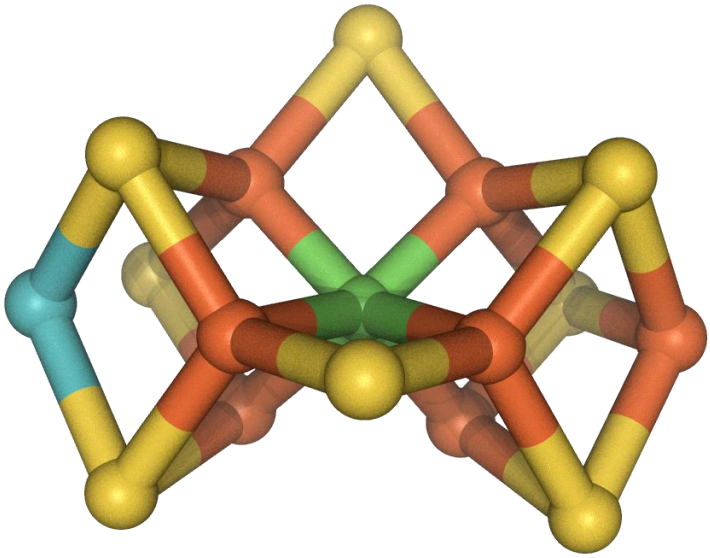
List of Abbreviations

RSS	Reactive sulfur species
ROS	Reactive oxygen species
PTP1B	Protein tyrosine phosphatase 1B
ADH	Liver alcohol dehydrogenase
NAD	Nicotinamide adenine dinucleotide
FAD	Flavin adenine dinucleotide
PLP	Pyridoxal 5'-phosphate
RC	Reactive complex
PRC	Pre-reactive complex
TS	Transition structure
IC	Intermediate complex
PC	Product complex
SufS	Cysteine desulfurase from *Synechocystsis* sp.
NicF	Maleamate amidohydrolase
SQR	Sulfide:quinone oxidoreductase
CysRS	Cysteinyl-tRNA synthetase
MD	Molecular Dynamics
DFT	Density functional theory
QM	Quantum mechanics
MM	Molecular mechanics
HF	Hartree-Fock
MP	Møller-Plesset perturbation theory
QCISD	Quadratic configuration interaction, singles and doubles
SCF	Self-consistent field
IEFPCM	Integral equation formalism of the polarizable continuum model
ZPVE	Zero-point vibrational energy
ONIOM	Our own *n*-layered integrated molecular orbital and molecular mechanics
AMBER	Assisted Model Building with Energy Refinement
CHARMM	Chemistry at Harvard Macromolecular Mechanics
GROMOS	Groningen Molecular Simulation

NAMD	Nanoscale molecular dynamics
CpHREMD	Constant pH replica exchange molecular dynamics
MOE	Molecular Operating Environment
RMSD	Root mean square deviation
NBO	Natural bond orbital
PA	Proton affinity
HA	Hydrogen affinity
GPB	Gas phase basicity
BDE	Bond dissociation enthalpy
PES	Potential energy surface
QTAIM	Quantum theory of atoms in molecules
PDB	Protein databank
AI	Artificial intelligence
PCA	Principal component analysis

Chapter 1:

Diversity and Challenges in the Study of (Bio)sulfur Chemistry



1.1 Sulfur Chemistry

Sulfur chemistry is unique amongst main group elements. For instance, it can possess and access a wide range of oxidation states, from -2 to +6, and even possibly in fractional oxidation states,[1] that as a result enables it to participate in a broad and fascinating range of bonding environments. An additional consequence is that sulfur-containing molecules can be both nucleophilic and electrophilic, enabling the sulfur and larger molecule to engage in a wide range of chemical reactions leading to a diverse range of sulfur functional groups (Figure 1.1). For example, a single sulfur can form formally single and double bonds with one or more oxygen atoms, S—C bonds in alkyl chains or rings, and interestingly tends to bond with itself to form S—S containing species.

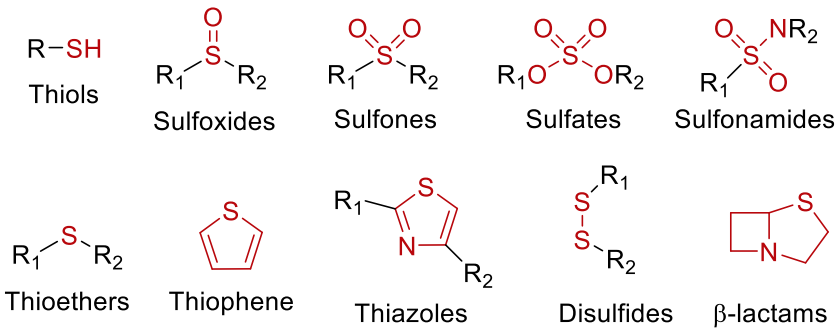


Figure 1.1. Examples showing the diverse range of sulfur functional groups commonly observed in commercial drugs and other important biological molecules.

1.2 Brief Overview of the Medicinal, Industrial and Biological Importance of Sulfur

1.2.1 Medicinal and industrial importance.

Due in large part to sulfur's diversity of chemistry and properties, it has long played key roles in drug design and development.[2] For example, one of the most important uses of sulfur in the 20th century was for development of novel β-lactams antibiotics (see Figure 1.1). In a recent review Scott and Njardarson highlighted the diversity of sulfur functional groups used in modern drug development;[2] sulfur-containing FDA-approved drugs

currently available on the market comprise over 249 unique structures that are divided into 14 drug categories. Indeed, sulfur, specifically sulfone (R'-SO2-R") functional groups are found in three of the top four small molecule therapeutic drugs of 2020 including the top choice, the HIV capsid inhibitor Lenacapvir (GS-6207) from Gilead Sciences for treatment of multidrug resistant HIV. Furthermore, sulfur-containing nucleobase analogues have now been developed as part of targeted gene therapy for spinal muscular atrophy.[3] Industrially, sulfur has also found broad applications from the first form of vulcanization, hardening of rubber (e.g., tires), to the widespread and common use of sulphites in the food and beverage industry as preservatives and antioxidants to help prevent spoilage[4] or stop fermentation.

1.2.2 Biological importance.

Within cells and organisms, sulfur is found in a broad variety of physiologically crucial or essential biomolecules. This is underscored by, for instance, that it is the only element other than carbon, hydrogen, nitrogen, and oxygen, found in the 20 common proteinogenic amino acids. In fact, it is present as part of the side chains of two amino acids: cysteine (Cys) as a thiol, and methionine (Met) as a thioether. Furthermore, these two amino acids serve as the source of sulfur for all sulfur-containing cofactors.

Several bio-critical roles of sulfur have been shown to be related to its redox properties. For example, methionine is known to often act as an antioxidant especially when within proteins,[5-6] while the most abundant cellular redox regulator in animals involves the reversible oxidative conversion of glutathione (GSH) to its dimeric disulfide GSSG.[7] However, there are other proposed forms of sulfur regulation.[7] Indeed, growing recognition of the diversity of possible oxidized-derived sulfur species deliberately formed in cells as part of metabolic pathways and their crucial physiological roles has led to a significant increased interest in the formation of reactive sulfur species (RSS).[1] For example, it is now known that in some proteins oxidation of the thiols of an active site cysteinyl is a deliberate

mechanism for protein regulation, in particular with sulfur adjacent nitrogen atoms and formation of hypervalent S—N bonds (e.g., formation of a sulfenyl amide in protein tyrosine phosphatase 1B (PTP1B) upon oxidation of an active site cysteine and its subsequent reaction with a nearby backbone nitrogen atom, or formation of a sulfurane in peroxiredoxin between an oxidized cysteinyl residue with a histidyl imidazole nitrogen, Figure 1.3).[8-9] Unfortunately, however, sulfur redox bio-signaling pathways are still poorly understood due to the complexity of the pathways (e.g., multiple oxidation states) and instability and reactivity of many of the proposed intermediates. This makes them difficult if not almost impossible to experimentally observe, especially in their radical forms.[10] Some of the known or proposed biopathways involving sulfur oxidation are shown in Figure 1.2.[7]

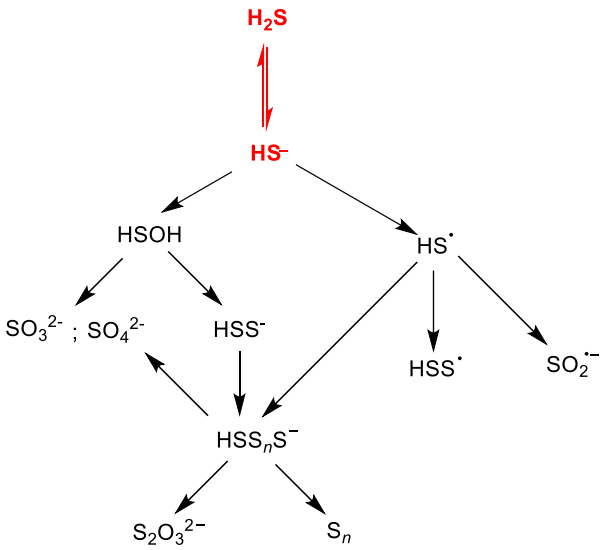


Figure 1.2. Examples of currently proposed hydrogen sulfide oxidation pathways leading to persulfide and O-containing derivatives. Pathways do not always represent direct transformations, i.e., some chemical species attributed to reactions omitted for simplicity.

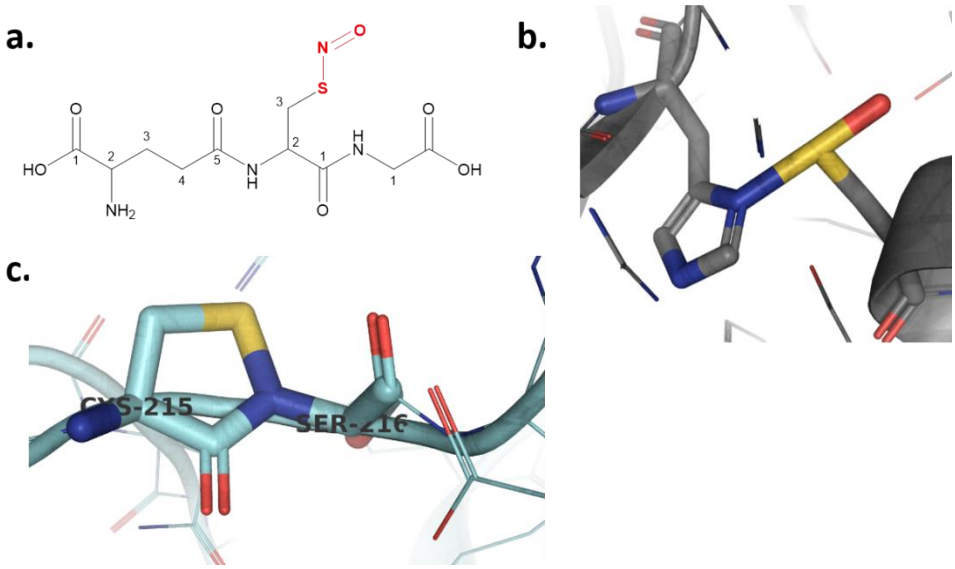


Figure 1.3. Examples of biological sulfur-nitrogen bonding. (A) S-nitrosoglutathione, part of NO signaling pathways, (B) sulfurane from archaeal peroxiredoxin (PDB: 2ZCT),[9] (C) sulfenyl amide from PTP1B (PDB: 1OEM).[11]

However, much of the biological reactivity attributed to sulfur is through the thiol group of cysteine.

1.3 Overview of potential roles of active site cysteinyl residues

Arguably the most well-known role(s) of cysteine is as a functionally important residue within the active sites of enzymes where they are known to perform a wide variety of activities. A common role of cysteinyl residues is as an acid (thiol) or base (thiolate). However, they are also often involved in coordination of functionally important (structurally or catalytically) metal ions including Zn^{2+}, $Fe^{2+/3+}$, Ni^{2+}, Cu^{+}, and Cd^{2+} due to the relative ease of deprotonation of their thiol.[12]

As metal ion ligands. A classic example of the potential roles of cysteinyl residues in binding metal ions in enzymes is arguably provided by the enzyme liver alcohol dehydrogenase (ADH).[12] This metalloenzyme catalyzes the oxidation of alcohols to ketones

with concomitant reduction of nicotinamide adenine dinucleotide (NAD^+). In particular, however, it contains two different Zn^{2+} ions; one being structural and the other being involved in the catalytic mechanism. While both ions possess tetrahedral coordination spheres, the structural Zn^{2+} is bound via the thiolates of 4 Cys residues. Meanwhile, the catalytically important Zn^{2+} is bound via the thiolates of 2 Cys residues and the side chain imidazole of a histidyl (His). Its coordination sphere is completed by ligation to a water molecule. Inactivation of the enzyme can be achieved via oxidation of the thiolates of one or more of these Cys residues causing the release of one or both Zn^{2+}. The interaction of the cysteine's thiol/thiolate with metal ions can also be exploited to control substrate recognition and binding. For instance, cysteinyl-tRNA synthetase is responsible for correctly attaching cysteine to its cognate tRNA ($tRNA^{Cys}$). It correctly distinguishes cysteine from the structurally and chemically similar serine with such high-fidelity, over 20000-fold.[13] As a result, unlike many other aminoacyl-tRNA synthetases, cysteinyl-tRNA synthetase does not have a second 'proof-reading' active site to ensure accuracy of its formed product. This high-fidelity is attributed to the proposed formation of a critical Zn-cysteine substrate thiolate interaction upon substrate binding.

As redox active residues. Cysteine, due to its sulfur, is also a redox-active residue and can often be reversibly oxidized/reduced. This ability is exploited in some enzymes in which a key cysteine undergoes post-translational modification (e.g., protein tyrosine phosphatase 1B (PTP1B)[8]) which have been shown in some to be important for protein regulation.[8, 14] However, cysteine's are also commonly involved in enzymatic redox mechanisms which can lead to a diverse variety of reactive sulfur species, and these can also be influenced by environment conditions (e.g., pH, concentration of reactive oxygen species etc.). While considerable experimental and computational efforts have been made to identify and characterize the chemistry and properties of redox active Cys-residues,[15-16] much unfortunately remains unclear or even unknown.

Two examples of enzymes that exploit the redox capabilities of one or more active site cysteinyl's are thioredoxin (Trx) and glutathione reductase (GSR). Thioredoxins are a family of small redox-active proteins containing an active site CXXC motif whose general physiological role is shown in Figure 1.4.[17] Their role is to reduce disulfide bonds to their thiol derivatives. As a result Trx plays an essential role in important cellular functions as a protein antioxidant, cell death regulator, inflammation response, protein folding, and viral life cycles.[18] The redox state of the two active site cysteinyl's of Trx are modulated by their substrate and Trx reductase (TrxR). The enzyme GSR also has redox active cysteinyl residues and performs a similar function to Trx except that its substrate is glutathione.[19] More specifically, GSR reduces oxidized dimeric glutathione (GSSG) back to its free constituent glutathiones (GSH) thus enabling the latter to successfully carry out its antioxidant scavenging role. The resulting oxidized GSR, i.e., its oxidized active cysteinyl residues now present as Cys-S–S-Cys, can be reduced back to its native state via the use of NADPH, thus enabling GSR to continue in its maintenance of redox homeostasis.

Cysteine desulfurases are a family of enzymes responsible for trafficking sulfur and furthermore are vital for the formation of metal ion-sulfur clusters (e.g., iron-sulfur clusters) and their homeostasis, nucleoside modifications, cofactor biosynthesis (biotin, molybdopterin, and thiamine), and branched chain amino acid biosynthesis.[20-21] Notably, these enzymes use a pyridoxal 5'-phosphate (PLP) cofactor as an electron sink to facilitate C—S bond cleavage in cysteine substrate and formation of an enzyme-bound persulfide. Unfortunately, despite their importance, the exact catalytic mechanism of cysteine desulfurases is still unknown. Indeed, even the identity of the required mechanistic base is unclear with several experimental studies seemingly proposing different identities for the base.[20, 22] Our computational study on the catalytic mechanism of the cysteine desulfurase SufS, as well as more detailed information on these enzymes, is provided in **Chapter 8**.

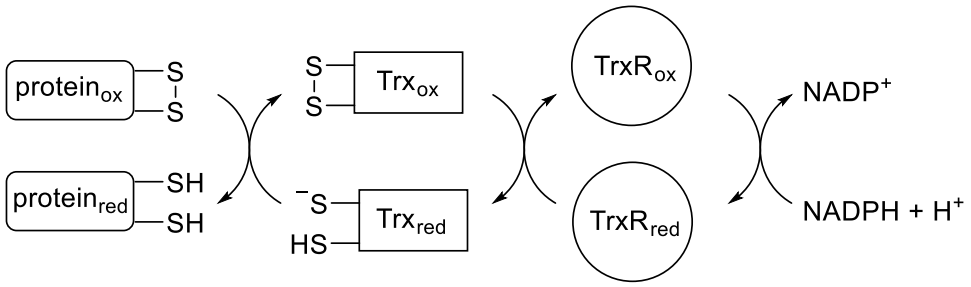


Figure 1.4. General depiction of the overall catalytic cycle involving Trx enzymes. The reduction of disulfide bonds in, for instance, proteins results in oxidation of the reduced native state of Trx (Trx_{red}) to Trx_{ox}. Restoration of the native state of Trx and its two free active site cysteinyl thiols involves thioredoxin reductase (TrxR) and NADPH.

Cysteine as a nucleophile. Cysteine proteases are enzymes in which an active site thiolate acts as a nucleophile in catalyzing the hydrolysis of peptide bonds. The general catalytic mechanism of cysteine proteases is shown in Figure 1.5. In such enzymes the cysteinyl residue is spatially adjacent to a basic residue, often a histidyl (see Figure 1.5; **RC**). The latter activates the cysteinyl by deprotonating the thiol and thus enabling the thiolate sulfur to nucleophilically attack the substrate at the carbonyl carbon of its peptide bond to form a tetrahedral intermediate (**I1**). Protonation of the peptide bond's nitrogen results in breaking of the peptide bond and formation of a covalent enzyme-intermediate species (**I2**) with release of peptide fragment **P2**. A water molecule can then enter the active site and nucleophilically attacks the carbonyl carbon of **I2** with concomitant loss of a proton to the basic (histidyl) residue to give the tetrahedral intermediate **I3**. The latter can then collapse to give the peptide fragment **P1**, which can then be released, and also regenerates the native state of the enzyme. Another example of a cysteine-catalyzed reaction is described in **Chapter 5** by the NicF enzyme.

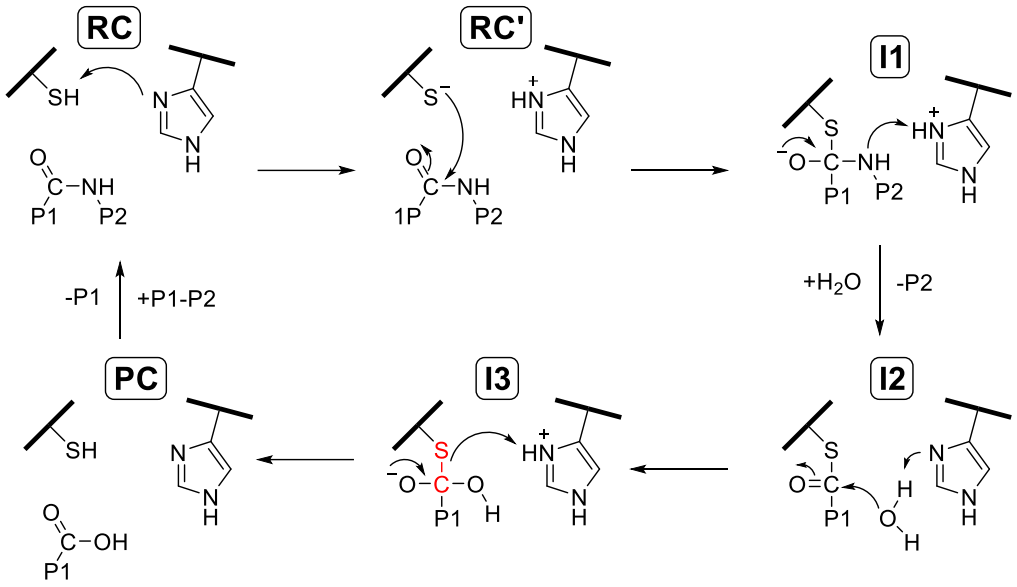


Figure 1.5. Schematic illustration of the general mechanism of cysteine proteases in which a histidyl acts as the catalytic base. The hydrolyzed bond is shown in red.

The above brief overview highlights the medicinal, industrial and, in particular, the physiological importance of sulfur as well as the diversity of the chemistry in which it can engage. Critical to a more complete understanding of its biochemical roles, and by extension how it can be used in other areas, e.g., therapeutic drug development, is elucidating how its chemistry may be exploited in physiologically important processes such as enzyme catalysis. However, in addition to the usual challenges with experimental characterization of enzyme mechanisms, the complexity of the multiple possible pathways as well as the high reactivity of many putative sulfur derivatives involved further adds to the difficulties that studies on biochemical sulfur species present.

1.4 Computational enzymology and sulfur-containing biomolecules

Computational enzymology presents an alternate approach that is now widely used successfully to elucidate the mechanisms and behaviors of a diverse array of enzymes. This

is due to the broad range of methods that can now be applied from docking and molecular dynamics (MD) to density functional theory (DFT), quantum mechanics (QM) and beyond that are able to individually or complementarily, reliably and accurately model the properties and chemistry of important biomolecular species such as enzymes. However, it presents its own challenges including but not limited to identifying a suitable methodology for the reaction or properties of interest or choosing a suitable structure from an MD simulation as a template structure for further and more detailed study.

These computational challenges are particularly true when it comes to the study of sulfur biochemistry as evidenced in the literature by the lack of consistent methodology applied to sulfur-containing biomolecules. Indeed, many different DFT methods have been used to study biosulfur systems, with range-separated functionals recently being shown to have the most success.[23] While high-level electron correlated *ab initio* methods have also shown some success,[24] their application with the necessarily larger biochemical models often required for modeling purposes is difficult. These challenges are discussed in more depth in **Chapter 4**. We recently demonstrated the performance of various functionals on polysulfur/selenium systems and have reached a similar conclusion.[25]

In addition to challenges with implementation of a reliable methodology, the literature is plagued with uncertainties, particularly with regard to per/polysulfide formation. While broad mechanisms for these pathways are outlined, specific mechanistic details are lacking with the chief example being sulfide:quinone oxidoreductase (SQR).[26] Formation of a trisulfur species is described in the proposed mechanism, however questions regarding chain extension arise. In particular, it remains unknown whether the addition of H_2S proceeds through the terminal or middle sulfur.

In this dissertation we assessed and established, in **Chapters 3** and **4**, computational approaches for the study of biomolecules containing sulfur atoms using small QM-cluster models. These studies are then built upon and expanded to the study of enzymes using QM/MM and MD in **Chapter 5**. A common computational approach when studying enzymes is to first perform a short MD simulation to obtain a solvated and thermally equilibrated structure(s). However, several approaches to choosing a structure have appeared in the literature. In **Chapter 6** we examine the challenges of selecting suitable chemical models from MD simulations and the possible implications of model choice. Subsequently, in **Chapter 7**, we demonstrate the potential use of artificial intelligence (AI) through machine learning to help inform model selection from MD simulations in these large systems. Finally, in **Chapter 8** the applications are tested on a cysteine desulfurase enzyme, SufS.

1.5 References

1. Giles, G. I.; Jacob, C., Reactive sulfur species: an emerging concept in oxidative stress. *Biol. Chem.* **2002,** *383*, 375-88.

2. Scott, K. A.; Njardarson, J. T., Analysis of US FDA-Approved Drugs Containing Sulfur Atoms. *Top. Curr. Chem. (Cham.)* **2018,** *376*, 5.

3. Wurster, C. D.; Ludolph, A. C., Nusinersen for spinal muscular atrophy. *Ther. Adv. Neurol. Disord.* **2018,** *11*, 1756285618754459.

4. D'Amore, T.; Di Taranto, A.; Berardi, G.; Vita, V.; Marchesani, G.; Chiaravalle, A. E.; Iammarino, M., Sulfites in meat: Occurrence, activity, toxicity, regulation, and detection. A comprehensive review. *Compr. Rev. Food. Sci. Food Saf.* **2020,** *19*, 2701-2720.

5. Levine, R. L.; Mosoni, L.; Berlett, B. S.; Stadtman, E. R., Methionine residues as endogenous antioxidants in proteins. *Proc. Natl. Acad. Sci. USA* **1996,** *93*, 15036-40.

6. Veredas, F. J.; Canton, F. R.; Aledo, J. C., Methionine residues around phosphorylation sites are preferentially oxidized in vivo under stress conditions. *Sci. Rep.* **2017,** *7*, 40403.

7. Benchoam, D.; Cuevasanta, E.; Moller, M. N.; Alvarez, B., Hydrogen Sulfide and Persulfides Oxidation by Biologically Relevant Oxidizing Species. *Antioxidants (Basel)* **2019,** *8*.

8. Dokainish, H. M.; Gauld, J. W., Formation of a Stable Iminol Intermediate in the Redox Regulation Mechanism of Protein Tyrosine Phosphatase 1B (PTP1B). *ACS Catalysis* **2015,** *5*, 2195-2202.

9. Nakamura, T.; Yamamoto, T.; Abe, M.; Matsumura, H.; Hagihara, Y.; Goto, T.; Yamaguchi, T.; Inoue, T., Oxidation of archaeal peroxiredoxin involves a hypervalent sulfur intermediate. *Proc. Natl. Acad. Sci. USA* **2008,** *105*, 6238-42.

10. Abedinzadeh, Z., Sulfur-centered reactive intermediates derived from the oxidation of sulfur compounds of biological interest. *Can. J. Physiol. Pharmacol.* **2001,** *79*, 166-70.

11. Salmeen, A.; Andersen, J. N.; Myers, M. P.; Meng, T. C.; Hinks, J. A.; Tonks, N. K.; Barford, D., Redox regulation of protein tyrosine phosphatase 1B involves a sulphenyl-amide intermediate. *Nature* **2003,** *423*, 769-73.

12. Giles, N. M.; Watts, A. B.; Giles, G. I.; Fry, F. H.; Littlechild, J. A.; Jacob, C., Metal and redox modulation of cysteine protein function. *Chem. Biol.* **2003,** *10*, 677-93.

13. Zhang, C. M.; Christian, T.; Newberry, K. J.; Perona, J. J.; Hou, Y. M., Zinc-mediated amino acid discrimination in cysteinyl-tRNA synthetase. *J. Mol. Biol.* **2003,** *327*, 911-7.

14. Giles, N. M.; Giles, G. I.; Jacob, C., Multiple roles of cysteine in biocatalysis. *Biochem. Biophys. Res. Commun.* **2003,** *300*, 1-4.

15. Fomenko, D. E.; Xing, W.; Adair, B. M.; Thomas, D. J.; Gladyshev, V. N., High-throughput identification of catalytic redox-active cysteine residues. *Science* **2007,** *315*, 387-9.

16. Weerapana, E.; Wang, C.; Simon, G. M.; Richter, F.; Khare, S.; Dillon, M. B.; Bachovchin, D. A.; Mowen, K.; Baker, D.; Cravatt, B. F., Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature* **2010,** *468*, 790-5.

17. Holmgren, A.; Johansson, C.; Berndt, C.; Lonn, M. E.; Hudemann, C.; Lillig, C. H., Thiol redox control via thioredoxin and glutaredoxin systems. *Biochem. Soc. Trans.* **2005,** *33*, 1375-7.

18. Collet, J. F.; Messens, J., Structure, function, and mechanism of thioredoxin proteins. *Antioxid. Redox. Signal.* **2010,** *13*, 1205-16.

19. Couto, N.; Wood, J.; Barber, J., The role of glutathione reductase and related enzymes on cellular redox homoeostasis network. *Free Radic. Bio.l Med.* **2016,** *95*, 27-42.

20. Hidese, R.; Mihara, H.; Esaki, N., Bacterial cysteine desulfurases: versatile key players in biosynthetic pathways of sulfur-containing biofactors. *Appl. Microbiol. Biotechnol.* **2011,** *91*, 47-61.

21. Mihara, H.; Esaki, N., Bacterial cysteine desulfurases: their function and mechanisms. *Appl. Microbiol. Biotechnol.* **2002,** *60*, 12-23.

22. Zheng, L.; White, R. H.; Cash, V. L.; Dean, D. R., Mechanism for the desulfurization of L-cysteine catalyzed by the nifS gene product. *Biochemistry* **1994,** *33*, 4714-20.

23. Smith, J. M.; Jami Alahmadi, Y.; Rowley, C. N., Range-Separated DFT Functionals are Necessary to Model Thio-Michael Additions. *J. Chem. Theory. Comput.* **2013,** *9*, 4860-5.

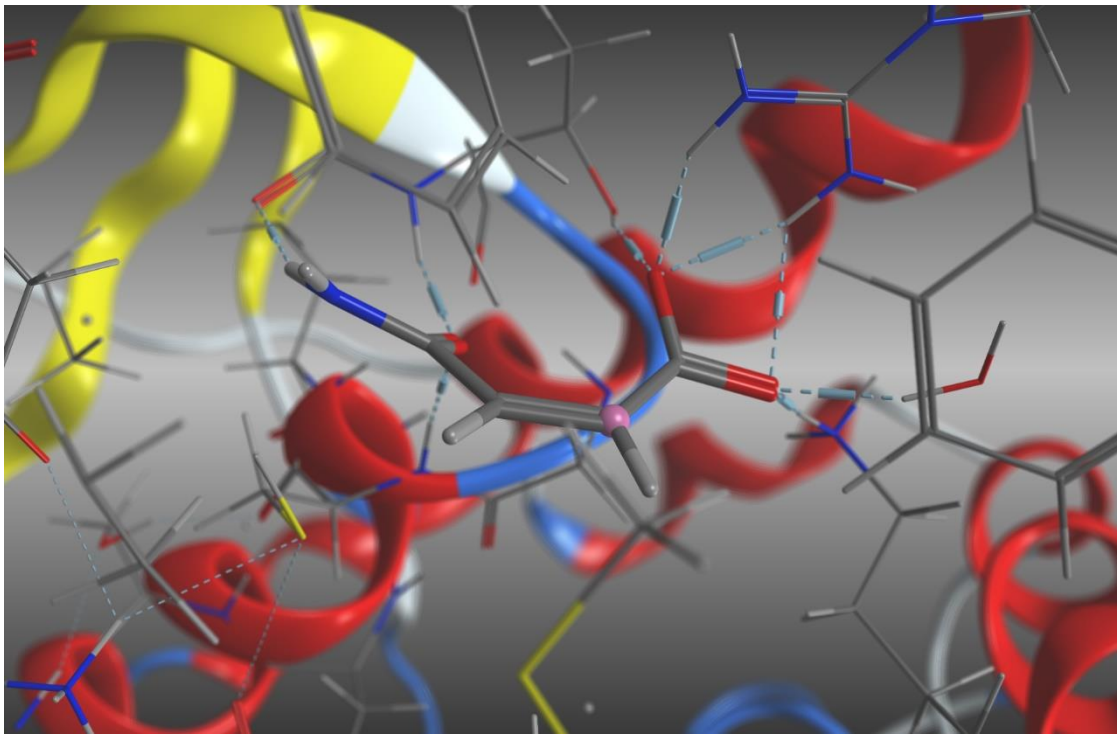
24. Orabi, E. A.; Peslherbe, G. H., Computational insight into hydrogen persulfide and a new additive model for chemical and biological simulations. *Phys. Chem. Chem. Phys.* **2019,** *21*, 15988-16004.

25. Nikoo, S.; Meister, P. J.; Hayward, J. J.; Gauld, J. W., An Assessment of Computational Methods for Calculating Accurate Structures and Energies of Bio-Relevant Polysulfur/Selenium-Containing Compounds. *Molecules (Basel)* **2018,** *23*.

26. Brito, J. A.; Sousa, F. L.; Stelter, M.; Bandeiras, T. M.; Vonrhein, C.; Teixeira, M.; Pereira, M. M.; Archer, M., Structural and functional insights into sulfide:quinone oxidoreductase. *Biochemistry* **2009,** *48,* 5613-22.

Chapter 2:

From *ab initio* to DFT: Applications of Modern Computational Enzymology



2.1 Introduction

Computational chemistry is a subset of chemistry wherein a chemist uses computer simulation software to investigate chemical phenomena. At the heart of these programs are the encoding of theoretical methods that aim to provide approximate solutions to the Schrödinger Equation which is itself at the core of quantum chemistry. Each of these methods is based on various approximations and/or assumptions; some may be mathematically rigorous while others may, for instance, be empirical. As a result of these various approaches to solving the Schrödinger Equation computational chemistry is now able to reliably and accurately investigate a diverse array of chemical problems. For example, it is now almost routine to predict structural properties of molecules,[1-5] investigate catalytic mechanisms,[6-8] examine charge distribution,[9-12] and predict spectroscopic quantities.[13-14] Furthermore, computational chemistry has evolved beyond studying small molecules and has expanded well into biochemistry, biology, and biophysics. Indeed, protein-drug interactions,[15-17] homology modelling for protein structure prediction,[18-20] modelling cellular membranes,[21-23] protein mutagenesis,[24-25] and enzymatic mechanisms[26-27] are all now possible. It is noted that excellent and detailed discussions and descriptions of the various theoretical concepts underlying the various computational methods can be found in textbooks and the journal literature.[28] Hence, in this chapter only a brief overview of those aspects most relevant to the studies discussed herein are provided (see below).

From an historical perspective, while the foundational theories of quantum chemistry have existed since the 1920s and 1930s, practical applications via computational chemistry only began with the development of digital computers. As computers have advanced so has the range of possible chemical problems that can be investigated. Conversely, new theoretical method developments have further expanded the capabilities of computational chemistry. For instance, in the early 1950s Roothaan[29] performed some of the first calculations on small atomic systems with calculations on diatomics following shortly thereafter in 1956.[30]

By the 1970s, molecules with as many as 18 atoms had become computationally tractable.[31] These calculations were all performed with Hartree-Fock using Slater-type orbitals and required highly specialized machines only available at a few research institutions. Ongoing computational and theoretical developments now mean that chemical models of hundreds and even thousands of atoms can be studied, though with a trade-off in accuracy and/or modelling power.

It can be said that there have been several major driving forces behind the development of new theoretical and computational methods, and their applications. Some of these include:

- the need and importance of including electron correlation for the study of, for example, chemical reactions;
- the need and desire for larger chemical models;
- not all parts of a chemical model, while possibly necessary, need to be modelled with equal accuracy.

This is further underscored by the fact that the 1998 Nobel Prize in Chemistry was awarded to Walter Kohn "for his development of the density-functional theory" and John Pople "for his development of computational methods in quantum chemistry", while the 2013 Nobel Prize in Chemistry was awarded to Martin Karplus, Arieh Warshel and Michael Levitt "for the development of multiscale models for complex chemical systems".

In this chapter, we briefly describe the Hartree-Fock (HF) theory, which is arguably the foundation for all other electronic structure methods, followed by wavefunction-based and Density Functional Theory (DFT) electron correlation methods. Subsequently, how these methods and others (e.g., molecular dynamics) can be applied to increasingly larger and more complex systems is then outlined.

2.2 Ab initio methods

2.2.1 The Hartree-Fock (HF) Method

The HF method is the foundational quantum chemical method used to approximate a wavefunction for a chemical system. It is important to note that while the Schrödinger Equation can be solved for a single electron system, there is no currently known solution for a many-electron system. To explore why this is the case, we can examine the full field-free non-relativistic molecular Hamiltonian which in atomic units is:

$$\hat{H} = \sum_{i=1}^{N_E} \frac{1}{2}\nabla_i^2 - \sum_{A=1}^{N_N} \frac{1}{2M_A}\nabla_A^2 - \sum_{i=1}^{N_E}\sum_{A=1}^{N_N} \frac{z_A}{r_{iA}} + \sum_{i=1}^{N_E}\sum_{j>i}^{N_E} \frac{1}{r_{ij}} + \sum_{A=1}^{N_A}\sum_{B>A}^{N_B} \frac{z_{AB}}{r_{AB}}$$

where N_E represents the total number of electrons and N_N represents the total number of nuclei in the system. The first term represents the kinetic energy of all electrons in the system, the second term represents the kinetic energy of all nuclei in the system, the third term represents all electron-nucleus interaction potential energies, the fourth term represents the sum of all electron-electron interactions, and the final term represents all nucleus-nucleus interaction potentials in the system. This equation is already too complex to solve and must be simplified. One of the main assumptions made is the Born-Oppenheimer Approximation. This can be stated as: since the nuclei are much heavier than electrons, from the perspective of the electrons the nuclei can be considered to be stationary.

As a result of the Born-Oppenheimer approximation, the sum of the nuclear kinetic energies is zero and nucleus-nucleus interacting potentials become a constant (V_{NN}) which can be calculated quickly and added in later. This leads to formation of the electronic

Schrödinger Equation whereby the electronic wavefunction is parametrically dependent on the nuclear coordinates:

$$\widehat{H}_{el} = \sum_{i=1}^{N_E} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N_E}\sum_{A=1}^{N_N} \frac{z_A}{r_{iA}} + \sum_{i=1}^{N_E}\sum_{j>i}^{N_E} \frac{1}{r_{ij}}$$

However, this equation is still too complicated to solve due to the final term accounting for the interaction of two electrons. In the orbital approximation the motions of electrons are assumed to be independent, allowing for separation of electrons into their own atomic or molecular orbitals. In effect, this approximation results in replacement of the intractable two-electron terms to essentially one-electron terms. Conceptually, each electron is now being modelled as moving through a field created by each of the other electrons. Mathematically, the central field approximation has been used to create an averaged electron smear that interacts with a single electron. It is noted that the wavefunction is expressed as a single Slater determinant to ensure that it is antisymmetric (i.e., obeys Pauli exclusion principle). We have arrived at the HF method.

Clearly, one of the major shortcomings of using the orbital approximation is that it does not properly account for electron correlation (Coulomb correlation; electron-electron repulsion). Combined with the fact that only the ground state of a system is calculated, HF fails in the study of complex chemical systems including for the calculation of reaction barriers, obtaining transition structures, and calculated distorted species. As a result, improvements to HF have given rise to post-HF methods to alleviate some of these issues, but they often come at a cost.

Importantly, HF calculations are still used in molecular dynamics parameterization (for more details, see the molecular mechanics section). For example, in the AMBER program package, ligands can be parameterized for inclusion in molecular dynamics simulations

following the same scheme used by Cornell et al.[32] HF/6-31G* calculations have been shown to be sufficient for obtaining a reasonable atomistic charge description as part of the restrained electrostatic potential (RESP) fitting model.

2.2.2 Wavefunction-based electron correlation methods

Methods that include electron correlation were developed in attempt to reach chemical accuracy; however, these are computationally much more expensive and only practically applicable to small systems. Calculations using these methods are incredibly rigorous and scale anywhere from $N^{5\text{-}10+}$ depending on the methodology used as shown in Table 2.1.

Table 2.1. Scaling costs of commonly used *ab initio* methods compared to DFT (in bold).

Scaling Behaviour	Method(s)
N^3	HF
N^4	**DFT**
N^5	MP2
N^6	MP3, CISD, MP4SDQ, CCSD
N^7	MP4, CCSD(T), QCISD(T)
N^8	MP5, CCSDT
N^9	MP6
N^{10}	MP7, CCSDTQ

Immediately, we can see a very intense cost-scaling effect for building in corrections to electron correlation. While results these methods produce are more reliable and accurate, the addition of electron correlation demands a large increase in resources making calculations for large models impossible. MP2 is the cheapest of these methods but still much more expensive and generally less reliable than DFT. Furthermore, *ab initio* methods tend to perform poorly for first-row transition metal chemistry due to the near degeneracy

of their low-lying excited states.[33] The large number of electrons and unusual bonding characteristics also necessitates higher resource usage through more complex basis sets. Notably, code used to perform these calculations has plenty of room for optimization and has been integrated as part of more complex methodology discussed below.

2.3 Density Functional Theory (DFT)

To alleviate the issues of the large cost increase, an alternative approach has been developed that does not require the calculation of a wavefunction. It is also noted that the wavefunction is not an observable property and has no physical interpretation, *i.e.*, it cannot be measured. Through both the Hohenberg-Kohn existence and variational theorems, the electron density was shown to be a reliable approach for calculation of chemical properties giving rise to DFT.

Compared to *ab initio* methods, DFT takes a significantly different approach to calculate energies of molecules. Rather than creating an approximate solution to the wavefunction, it ties the observable electron density to the ground state of a molecule. In other words, by knowing the electron density of a molecule, we can approximate its energy. The Kohn-Sham theorem defines the ground state energy as:

$$E[\rho] = T_s[\rho] + V_{NE}[\rho] + J[\rho] + E_{XC}[\rho]$$

Where T_s refers to the kinetic energy of the electrons, V_{NE} is the nuclear-electron potential energy, J is the electron-electron repulsion, and E_{XC} is the exchange-correlation energy. As can be seen, this is very similar to the electronic Hamiltonian but now these terms rely on the electron density rather than single electrons. The last term, E_{XC}, is the source of difficulty for all DFT methods. The exact functionals for exchange and correlation are not known and thus an empirical description is used. For example, B3LYP, arguably the most

successful hybrid functional, uses three different parameters to calculate the exchange correlation energy of a system as shown in the equation below.

$$E_{XC}^{B3LYP} = E_X^{LDA} + 0.20(E_X^{HF} - E_X^{LDA}) + 0.72(E_X^{GGA} - E_X^{LDA}) + 0.81(E_C^{GGA} - E_C^{LDA})$$

These variables are the amount of HF exchange energy, gradient-corrected exchange energy, and gradient-corrected correlation energy and are represented by the numbers shown, respectively. A functional can be tuned for different purposes by altering the contribution of each term to the total energy; however, this results in an entirely new functional. Notably, this has been done for B3LYP to improve results for chemical systems including first-row transition metal ions, specifically by lowering the amount of HF exchange included to 15% (B3LYP*).[34] Importantly, DFT and HF share important core principles making them conceptually and computationally related.[35]

The rise of hybrid DFT has proven to be a necessary boon for computational enzymology. Large model studies have become accessible as both code and computer hardware has developed, and hardware performance steadily increases every year. The search for an exact exchange-correlation functional is the cornerstone for further development of DFT methods. It is the approximation of the fourth term in the equation (the other terms are calculated exactly as in HF) that can be attributed to the success of DFT by lowering the cost of calculations significantly and making even the largest of systems approachable. Early developments of DFT looked at obtaining accurate geometries and dissociation energies and accurate prediction of barrier heights to benchmark their accuracy. Inclusion of van der Waals interactions drove further development in the field.[36-39] Through these advances, application of DFT to biological processes has become much more reliable and commonplace.

Although the operating cost of DFT is much lower compared to post-HF *ab initio* methods (Table 2.1), DFT has its own set of difficulties. Similarly to the HF method, DFT inherently lacks dispersion effects and has generally poor representation of long-range interactions at its most basic level. This created discrepancies between experimental data and calculated data which required refinement. Further research into dispersion correction terms to include in a variety of DFT functionals has increased the accuracy and reliability of results (ex. D3 and more recently, D4 corrections by Grimme). Excited state calculations are also impossible for DFT due to its single determinant nature. In its simplest form, it cannot be used to calculate excited states of molecules accurately. To address this issue, time-dependent DFT, or TDDFT, has become a method of choice.[40] TDDFT enables the use of DFT in systems where an external potential is included and has shown success in photochemical studies where excited state systems are commonly observed. Recently, this was applied to theoretical studies of photosystem II to gain insights into the pathways used to achieve photosynthesis.[41]

Unfortunately, however, DFT's largest practical drawback is the systematic improvement of functionals. Systematic improvement of DFT has been an often debated topic since these improvements are not rooted in fundamental first principles.[42] Recall, DFT relies upon the approximation of the exchange-correlation energy, an empirically derived quantity. A "Jacob's ladder" representation for the improvement of DFT[43] is often used to highlight key method improvements in the last 3 decades (Figure 2.1). Each improvement is in search of the exact functional, yet its components are not entirely understood. Because of this, performance of different functionals is not easily predictable for new chemical systems, and this often requires a different set of parameters for reliable applications. Throughout the years, we have seen the ever growing "functional zoo" as a result of continuous improvement upon previous functionals.[44] Unfortunately, this usually results in a functional tuned to work for a specific range of the benchmarks and performance outside this test set

is not known; it may only work on a narrow range of systems. Overall, this means there is no “black box” DFT method, so a careful selection of a DFT functional is recommended. It is not unusual to search the literature before choosing a method, or even to perform an assessment study to determine the best functional for a model of interest.

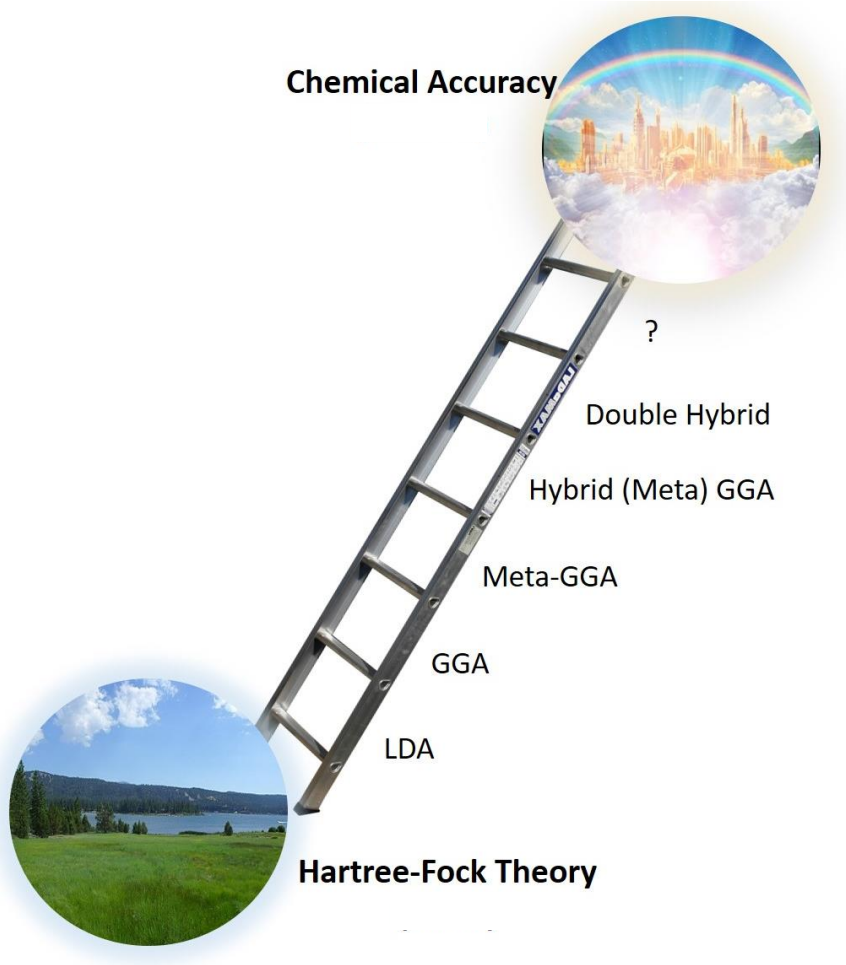


Figure 2.1. Jacob’s ladder representation for improvements in approximation of the exchange correlation functional over the past 30 years. The upper rungs on the ladder vary depending on the date of publication or the authors, and as new improvements are released. The bottom rungs do not change as they represent the early advancements of DFT.

Due to its modest cost, B3LYP/6-31G(d) has been the method and basis set of choice for the optimization of biomolecules and biological systems.[45] The basis set is of sufficient size to give reasonable geometry optimizations within a reasonable amount of time, and the method has been shown to give reliable structures for both organic and inorganic reactions with known shortcomings (for example B3LYP tends to underestimate barrier heights).[37] Fortunately, this level of theory has been successfully applied to a variety of biomolecular systems and since its shortcomings have been well studied, errors are generally predictable thus making it broadly applicable. For example, Shaik and coworkers have used B3LYP to model cytochrome P450 reactions.[46] While structures obtained using this level of theory are generally reliable, energetics are known to be inaccurate however and require long range corrections to be included and larger basis sets to ensure accuracy.[47] This can be easily remedied through the inclusion of previously discussed long range interactions and single point calculations at a higher level of theory.[48-52] With the steady advance in computational power, methods are being pushed to their limit by continuously increasing size of models.[53]

As mentioned previously, assessment studies are an incredibly useful tool to examine the shortcomings of a method and finding a level of theory for a chemical system of interest. The literature is filled with studies using B3LYP and its application to biological molecules. Yet the ever growing "functional zoo" is not nearly as well characterized. There are far too many new functionals in the literature for the same thorough analysis to occur, so users must perform this assessment themselves. These new functionals include many different parameters which are not always explicitly stated, making it impossible to predict where they will be useful. Thus, it is important to determine how these functionals perform on a system of interest by comparing to the highest-level *ab initio* method possible, the most rigorous approach. In especially sensitive systems, it is imperative that the core chemistry is well modeled so that structure properties and energetics reflect experimental results where available. For this purpose, an assessment study was done on chalcogen-

chalcogen bonding on biorelevant compounds in **Chapter 3**. Some previous studies have looked at the performance of DFT on chalcogen-containing compounds, but these were limited to thiols and selenols and did not examine sulfur or selenium bonding with itself or other chalcogens. Before continuing with more in-depth investigations into biological sulfur chemistry, it was important to ensure the current methods were able to describe these systems reliably and accurately.

2.4 Methodological Approaches

Proteins play important roles in all biological processes. Enzymes specifically are the workhorses of cells and are tasked with maintaining life. These biological macromolecules are incredibly complex and are thus challenging to study using experimental and theoretical methods. Mechanistic insights often require the combination of both techniques to understand what happens at the atomistic level in an enzyme active site. In particular, protein crystallography has enabled computational studies of enzymes by providing a starting point for the study of enzyme function.

Importantly, there are several challenges and choices that a user must make when selecting an appropriate model for study. There are many different approaches used by researchers around the world, each with its unique advantages and disadvantages and each method choice is carefully made for a specific purpose. Herein, we will give a brief overview of the different approaches used in computational enzymology to study various aspects of enzyme catalysis. Choosing a model is often a difficult task and several different approaches have been described in the literature.[54-59] There are numerous ways to approach this such as focusing on the active site in the QM-cluster approach, including the enzyme environment in the QM/MM approach, studying protein dynamics using MD, or perhaps combination studies of active site dynamics using QM/MM-MD. Each of these methods has its merits and drawbacks which is discussed below in its respective section. Examples will be shown

to highlight the successes of each approach. Importantly, a larger model is not necessarily indicative of a more accurate portrayal of the chemical problem; sometimes a smaller model that concisely captures the chemistry is a more ideal solution. Furthermore, selecting a model will necessarily inform the level of theory used. In general, the larger the model, the more sacrifices to accuracy are necessary for timely calculations.

2.4.1 QM-cluster

Enzymatic reactions can sometimes be reduced to small model systems to quickly gain insights into a problem of interest. The quantum mechanical-cluster (QM-cluster) approach allows a computational chemist to focus on a reaction of interest and use the highest level of theory possible to analyze it in detail. It is particularly useful when modeling transition metal-catalyzed enzymatic reactions without cofactors. These metal centers localize reactions to the vicinity of the metal ion and pose significant challenges in calculations arising from prediction of spin states. Accurate determination of energies for transition-metal reactions is an ongoing area of research and DFT methods are known to behave much more poorly for these systems.[60-61] In addition, they contain a large number of electrons and thus require significant investment of computational resources. The consideration of spin states adds to this problem generally mitigated through unrestricted spin calculations. Attempts to reduce these costs have tended towards using effective core potentials (ECPs, or pseudopotentials) to approximate the core electrons. This approach has been shown to be relatively accurate if a sufficient size basis set is used.[62] With consideration of all these shortcomings, a focused QM-cluster approach which includes the first shell and important second shell residues can be used to simplify the problem and gain more insight in a way that is not possible through the use of a large model.

One example system is from a collaboration project with Prof. Steven Bearne from Dalhousie. Here, we are examining the effect of different metal ions in the active site. While the enzyme shows some promiscuity in metal centers, we nonetheless chose to use the QM-cluster approach to perform calculations with models shown in Figure 2.2 for the Ni^{2+}-bound enzyme. Even though the method ignores the environment of the enzyme, modeling reactions in a polarizable medium with a dielectric constant of $\varepsilon = 4$ is used to overcome this drawback.[63-64] Another drawback is that effects of specific residues on the activity cannot be tested easily. Using a QM/MM approach, we can bring residues from the low layer into the high layer to calculate their exact contribution (more in section 2.4.3), but residues are deliberately selected for inclusion in QM-cluster calculations. If an additional residue is needed for the model, a new one must be constructed from the beginning.

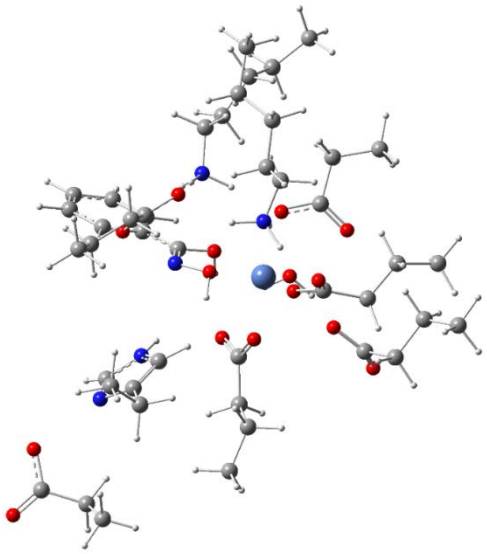


Figure 2.2. QM-cluster models for the active site of Ni^{2+}-bound mandelate racemase. Important first shell and second shell residues are included in the model with every other part of the enzyme omitted. Carbon atoms at the Cα are frozen to mimic constraints imposed by the protein backbone.

Another complication in QM-cluster modeling of enzymes has been in calculating the preferred substrate in asymmetric catalysis. Himo and coworkers have demonstrated the ability for this approach to show which propargyl alcohol and cyanohydrin enantiomer is preferred by *Mycobacterium smegmatis* acyl transferase.[65] A variety of conformations were tested in addition to the (*R)*- and (*S)*- enantiomers of a propargyl alcohol and a cyanohydrin. Potential energy surfaces were constructed for each substrate enantiomer to determine the origin of selectivity. The calculations correctly predicted the (*S)*-enantiomer of 1-isopropyl propargyl alcohol as the preferred substrate to match experiment, although the energy difference between the two enantiomers was ~3 kcal/mol off. Similarly, the (*R)*-enantiomer of 2-hydroxy propanenitrile was preferred and the energy difference replicated experimental observations of enantiomeric excess. These results show that with proper selection of a QM-cluster model, even the issue of enantioselectivity can be overcome.

One of the most difficult systems to study computationally are the carbon monoxide dehydrogenases. These enzymes catalyze the reversible reduction of CO_2 to CO and are a key enzyme in anaerobic carbon fixation.[66] Understanding its function is vital to understanding primordial life. Unfortunately, the active site binds a complicated 4Fe-4S cluster in addition to another metal ion such as Ni or Mo. Liao and Siegbahn have used QM-cluster with a modest size of 130 atoms to determine the energetics and transition structures of the initial oxidation/reduction step of the nickel-containing carbon monoxide dehydrogenase.[67] By relating studied structures to those found experimentally, they determined that a proton and a water molecule must be included in models to give accurate results. Based on this, they proposed a mechanism for this reaction whereby a protonated histidyl acts as a base to create an Fe-bound alcohol and CO. However, there remain many mechanistic details to be determined as this is one of the most complex examples of enzyme catalysis.

Recent studies tend to avoid QM-cluster models because it lacks a protein environment. Instead, residues are generally fixed at their Cα as a representation of the steric restraint imposed by the protein backbone. This also affects the polarity of the environment and necessitates the use of a solvation model. While it can be useful to use a reduced model, in cases of large active sites or where residues beyond the second shell influence the reaction, a different approach is required to ensure that these effects are captured. QM/MM and MD have grown to address this issue by modeling the entire enzyme.

2.4.2 Homology Modeling

Homology modeling is a technique used to build a model for tertiary and quaternary structure of proteins that do not have structures available to examine. This technique uses the sequence of a protein or a segment of the protein to match to a known structure. A sequence query is matched to a template structure and a model is built based on the template. Thus, the sequence query must have a significant match to the chosen template to ensure reliable and accurate modeling. In cases of less than 20-30% homology,[68-70] models are unpredictable, whereby useful *in silico* information it not likely to be gained (*i.e.*, garbage in, garbage out). Between 30 and 50% homology, a somewhat reliable model is produced where a large portion of uncertainty arises in loop regions. Ideally, a 50% homology match or better is obtained which allows us to make useful predictions about a protein and its interaction. This technique is incredibly useful in cases where proteins are difficult to crystallize, and the sequence is known. These sequences can be compared to available structures to create a model, however, assessing the model is tricky and can be done in a few ways: statistical methods which give probability scores,[71] energy calculations which examine the energy landscape,[72] and in rare cases matching to a known structure. Several online homology modeling tools are available to assist with model building[73-74] and they access structural data from online databases such as the Protein Databank (PDB) to obtain fold information. Naturally, the shorter the query, the more likely a reasonable match

will be identified. If we know a section of a protein that interacts with another protein, a homology model can be created for the interacting segment rather than attempt to model the entire protein. This is a crude approximation, but by taking this model and performing docking calculations for example, we can predict specific residues involved in the interaction of two proteins. Further experiments building upon these observations can drive mutagenesis studies or drug development to fit within predicted interacting regions. While this technique is not a focus for this dissertation, it is quite versatile and reliable when successful. Two examples include the ion selectivity of the potassium ion channel[75] and large scale identification of proteins in yeast.[76] Notably, homology modeling is usually a first step to model missing parts of a protein prior to application of any of the methods discussed below.

2.4.3 QM/MM

For many systems, the QM-cluster is not the best approach to model an enzymatic reaction. These models are not flexible so each time a model is deemed insufficient, a new one must be recreated. Thus, the effects of specific residues and their protonation states is more challenging to assess. Furthermore, most of the protein is removed so residues must be kept static in attempt to mimic steric effects imposed by the backbone of the enzyme. Quantum mechanics/molecular mechanics (QM/MM) is a much more flexible technique where the choice of which residues make up the active site remains with the user; however, the remainder of the protein is not excised from the model (Figure 2.3). This means the protein environment is now accounted for in addition to steric effects. While the initial setup is more complicated, the model is now modular with the ability to move residues in and out of the QM layer as required.

The QM layer is treated using DFT, often with a similar level of theory used for QM-cluster. The cost increase for this approach is minimal since molecular mechanics is

relatively cheap, and optimizations still converge relatively quickly. This technique is the most popular in the field of enzymology and has been the basis of driving further developments in computational protein modeling. Several reviews have been published recently highlighting the successes of the technique and giving tips on its usage.[54, 77-82]

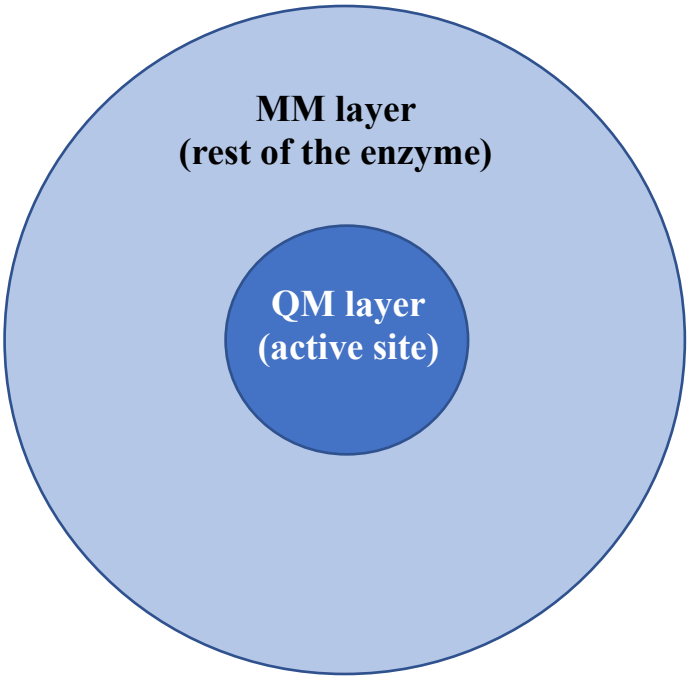


Figure 2.3. Schematic representation of the layering technique used by QM/MM to create a more inclusive multi-layer model. A small set of residues consisting of the active site is chosen to be the QM layer (dark blue) while the rest of the protein is treated using molecular mechanics force fields (light blue).

From Figure 2.3, there is the immediate question regarding the treatment of the coupling between the two layers. Two different schemes are available to address this issue: the additive scheme and the subtractive scheme (shown below).[83] In the additive scheme, the two different layers are addressed with their respective levels of theory and a coupling term is introduced between the two layers. Open valences are generally treated with the link atom approach where the cut bond is capped with a hydrogen atom on both sides.

Unfortunately, this requires special treatment of the coupling and parameters must be carefully included to ensure that there is no double-counting in the calculation. The subtractive scheme is slightly more complicated; MM energies are calculated for the entire protein as well as the QM layer. The MM energies for the QM layer are subtracted, thus giving rise to a subtractive scheme. Importantly, parameters for the link atoms must be well defined and exist within the program suite used for calculations. ONIOM is an example of a subtractive scheme and is the one used by Gaussian software *i.e.*, the approach used throughout this dissertation. The ONIOM method has been successfully applied in many chemistry subfields as outlined by Morokuma and coworkers.[77] Each of these schemes has challenges that appear at different stages of setup, however results are expected to be the same regardless of which is used.[83]

$$Additive\ Scheme: E_{ONIOM} = E_{QM} + E_{MM} + E_{\frac{QM}{MM},coupling}$$

$$Subtractive\ Scheme: E_{ONIOM} = E_{QM,model} + E_{MM,real} - E_{MM,model}$$

Misplaced cuts between the two layers can have severe consequences due to two different methodological approaches used to calculate energies. As a result, atoms may appear to have missing valence. To ensure that parameters are present for link atoms and charge distribution to the MM layer is minimal, six general rules are followed:

1. All bond formation or bond breaking must be present within the QM layer,
2. Any portions that result in hybridization change must be entirely in the QM layer,
3. Bond cuts should occur between a non-polar bond (ideally C-C) to represent the link atom approach most accurately,
4. Conjugated or aromatic regions must be kept entirely in one layer; no cutting between any of these types of bonds,

5. If a secondary or tertiary atom is expected to be important to the calculation, those should be included in the QM layer,
6. Partitioning should not divide areas of charge separation.

The last rule brings into question the treatment of charges between the two layers. There are three different approaches taken in QM/MM calculations: mechanical embedding, electronic embedding, and polarizable embedding. Mechanical embedding is the simplest approach where electrostatic effects are localized to the QM layer. The MM layer has no electrostatic effect on the QM layer, it only imposes steric restraints. In the electrostatic embedding scheme, the MM charges now polarize the QM layer. The structure in QM layer can adapt in response to the charges imposed by the MM layer. It is generally agreed that this gives more accurate energies,[81, 83] however accurate MM charges become vital to ensure this description is correct. Lastly, polarizable embedding schemes allow for polarization of both layers onto each other. This is by far the most difficult scheme to implement as a polarizable MM force field is required. More rigorous SCF calculations are also required, so these calculations are not routinely done. In this dissertation, energies obtained from electronic embedding and mechanical embedding were compared in **Chapter 5**. Importantly, where there are mostly hydrophobic interactions surrounding the active site, a mechanical embedding scheme is sufficient and electrostatic embedding is of similar quality or in some cases detrimental. The best example is **Chapter 5** where using electrostatic embedding increased barriers and transition structures were lost along the reaction surface.

The main advantage of QM/MM is that it enables detailed studies into enzymatic mechanisms. One example from our group is the post-transfer editing of ThrRS.[84] This enzyme is an important part of protein biosynthesis and is tasked with ensuring that threonine is added onto its cognate $tRNA^{Thr}$ before delivery to the ribosome for a growing

peptide chain. Due to structural similarities between serine and threonine, sometimes the enzyme can mischarge $tRNA^{Thr}$ with serine. To ensure fidelity of protein synthesis, these enzymes have developed editing mechanisms to remove mischarged amino acids. Mutagenesis experiments showed mutation of His73 and Cys182 to Ala decreased the rate of editing suggesting these residues are involved in the mechanism. However, the protonation states of these residues were unknown and further clarification on the catalytic base was required. Overall, 11 different mechanisms were tested in this study with varying the protonation states of each residue. It was found that a neutral His76 as well as a Cys182 thiolate resulted in the most kinetically and thermodynamically feasible pathway.

Another advantage is the ability to obtain a description of the potential energy surface in agreement with experiment as in, for example, the mechanism of aldehyde oxidase.[85] This enzyme detoxifies xenobiotic compounds and is involved in the drug metabolism pathway. It accepts a broad set of substrates including aldehydes and *N*-heterocycles. Notably, it is a molybdoflavin enzyme, containing a molybdenum cofactor (Moco), two Fe_2S_2 clusters, and flavin adenine dinucleotide (FAD) to aid in electron transfer. It is quite a challenging system to model due to the many possible spin states and large active site with many cofactors. Electron transport is also a challenge for optimizing transition structures as traditional DFT cannot model electron transport, but rather states before and after this occurs to suggest a possible reaction pathway. The pathway was divided into two different models for a manageable approach to track electrons: oxidation of Moco, followed by reduction of FAD. The catalytic reaction for the formation of phthalazine-1(2H)-one from phthalazine was elucidated. Electron transfer from the Moco to FAD had to be carefully calculated by comparing the redox half-reactions separately as FAD is not in the original model. FAD is very far from Moco in the active site and the electron/proton transfer steps between these two cofactors is unknown. However, it is possible through QM/MM to

calculate energy differences between the two intermediate steps to estimate thermodynamic properties.

QM/MM is an incredibly powerful technique that allows us to study many different aspects of catalysis beyond mechanistic studies. However, to achieve this, cost-effective strategies are required and often a less-than-ideal level of theory must be used. It is important to note that even today, further development is occurring for QM/MM. As always, the goal of accurate methods at a lower cost is one of the main driving forces of new developments. Protein dynamics is an essential part of catalysis that is ignored by current QM/MM approaches and is calculated by using a much simpler (in theory, not practice) technique.

2.4.4 Molecular Dynamics (MD)

Observing protein dynamics is useful in many scenarios such as in diseases whereby mutations influence interactions within a protein. Effects of these mutations can be tracked over time to see the impact they have on the reaction, whether through altering hydrogen bonding or solvent exposure pathways as examples. MD approximations consider the inverse of the Born-Oppenheimer approximation: electronic motions are ignored, and only nuclear motions are considered. As a result, this technique cannot be used to study reaction mechanisms; that remains the advantage of QM-cluster and QM/MM. However, as a result, this simplification enables the use of classical physics to calculate energies of each structure over a timescale. In this approach, atoms are treated as hard spheres connected by springs. As such, the energy equations for a typical MM system are shown below.

$$E_{MM} = E_{bonding} + E_{non-bonding}$$

$$E_{bonding} = \sum_{bonds} k_d(d - d_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi[1 + \cos(n\phi - \psi)]$$

$$E_{non-bonding} = \sum_{A,B} 4\varepsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right] + \sum_{A,B} \frac{q_A q_B}{4\pi\varepsilon_0 r_{AB}}$$

Energies are calculated by summing up bonding interactions and non-bonding interactions across the entire system. Bonding interactions are made up of bond lengths, bond angles, and dihedrals whereby two, three, and four atoms are connected respectively. Non-bonding interactions include van der Waals terms and electrostatic interactions between atoms that are not bonded. Cutoffs for non-bonding calculations are generally set to 10 Å as minimal effects are seen beyond that point.

To prevent bonds from floating apart and systems from exploding, penalty terms are set in place. Therefore, a set of descriptors for atom bonding is required which include equilibrium distances, usually generated from empirical data (or rudimentary HF calculations as stated previously). Together, this information is referred to as a force field, a set of parameters for bonding environments (including atom hybridization) containing information for bonding and non-bonding terms.

Force field developments specifically for biomolecules have created a surge in molecular dynamics usage. It provides a computationally cheap approach to study structural properties of biomolecules and has been particularly invaluable to the study of membrane-bound proteins. These systems contain interactions between a variety of different types of biomolecules and require the use of large, complex models to accurately simulate their properties. A variety of different force fields are available for proteins and enzymes including AMBER, CHARMM, GROMOS.[86-88] In this dissertation, the AMBER force field is used as it was developed for biomolecules and includes parametrization schemes for both organic and inorganic molecules.

One example use with respect to enzymology is through the study of differential solvent accessibility. Through this approach, dynamic aspects of residue mutations and the impact

they have on water access to the active site can be studied, an approach not viable using QM/MM. Wei et al. showed that ThrRS discriminates against serine in the active site by allowing access to a larger number of water molecules.[89]

Throughout this dissertation, molecular dynamics is used as a method to prepare structures for QM/MM. A minimum of 1 ns of simulation is required for proper equilibration,[56] although 10-20 ns is increasingly more common. Snapshot selection is important and can be difficult as there are many ways to reach a product state along the reaction surface. It was shown that for HIV-1 protease, the activation barrier can vary by over 40 kcal/mol depending on the structure chosen.[90] In **Chapter 6**, common snapshot selection techniques observed in the literature were reviewed. We showed that an x-ray structure is not always a good place to start QM/MM. Experimental mutations are often done on key residues to stabilize transition state analogues or various intermediates along a proposed pathway which are not conducive to catalysis. *In silico* structural adjustments require an equilibration period to allow the protein to relax and enter a lower energy state. This is particularly true when hydrogen bonding groups are removed or when docking calculations for a substrate are necessary.

2.4.4.1 Enhanced sampling techniques

Timescales required for relevant biological motions are usually in the millisecond range. This not readily accessible with atomistic MD and requires the use of enhanced sampling to allow access to these sorts of timescales.[91-92] One example technique is replica exchange.[93] The technique was developed as a way to overcome the issue of thermodynamic trapping in local energy minima at low temperatures by increasing the number of sampled states using replicas that sample into neighbouring states. Fortunately, the technique can be extended beyond temperature and has applications with other thermodynamic quantities as well. For example, calculating pK_as of active site residues is possible using constant pH replica

exchange molecular dynamics (CpHREMD). PROPKA, the traditional method for fast residue pK_a calculations, was shown to be problematic for cysteine residues since it considers an instantaneous snapshot.[94] CpHREMD provides an alternative mode of analysis that takes into consideration the transient nature of hydrogen bonding. Shen and coworkers have successfully used this approach to benchmark pK_a calculations of active site lysyl and cysteinyl residues.[95] While these are slightly underestimated, it is by a reproducible and predictable amount within 0.5 of the experimental value. This approach was used in **Chapter 8** to determine the protonation states of the active site residues in the presence/absence of the pyridoxal 5'-phosphate (PLP) cofactor and to determine the identity of a potential catalytic base.

2.4.5 QM/MM MD

While this technique has not been applied in this dissertation, its use has been rising significantly in the study of enzymatic reactions and thus warrants mention. QM/MM-MD is the combination of both QM/MM and MD methods to create a coupled technique. By doing so, quantum mechanical calculations are performed on a MD timescale, incorporating aspects of protein dynamics into QM/MM. While the active site is optimized, the protein remains mobile which in turn influences the direction of optimization and this continues back and forth until the calculation reaches convergence. Its greatest advantage is the ability to explore in detail an enzyme's energy landscape rather than looking at a single local minimum (such as an optimized complex from QM/MM). A reaction coordinate is defined to test a proposed reaction pathway, typically a bond breakage with a bond formation. The technique is remarkable for identifying potential transition states and intermediates as energetics are sampled along the pathway (Figure 2.4). It also works well for cases where protein dynamics have a large effect or when distinct protein conformations are observed (such as for transmembrane proteins). The major drawbacks are short timescales (< 200 ps) due to the incredibly intense calculations, the requirement for a

semiempirical method to minimize cost, a relatively simple reaction scheme that occurs in a minimal number of steps, and a heavy dependence on ensuring the force field parameters are correct. Umbrella sampling is a technique used to partially overcome the poor sampling from short timescales. It breaks up the reaction coordinate into a series of windows and applies a restraint to keep it to the center of the window. This increases the overall simulation time because each window is simulated for the same time, approximately 200 ps depending on the protocol. A relatively small QM model (< 60 atoms) is required for calculations to be completed in a reasonable amount of time, though increasing the number of atoms in the QM-layer at the cost of lower the sampling time is also a reasonable approach. Of note, this is simply a broad overview of QM/MM/MD and there have been many details omitted for the purpose of brevity. The following reviews are recommended for a more in-depth analysis.[54, 82, 96-97]

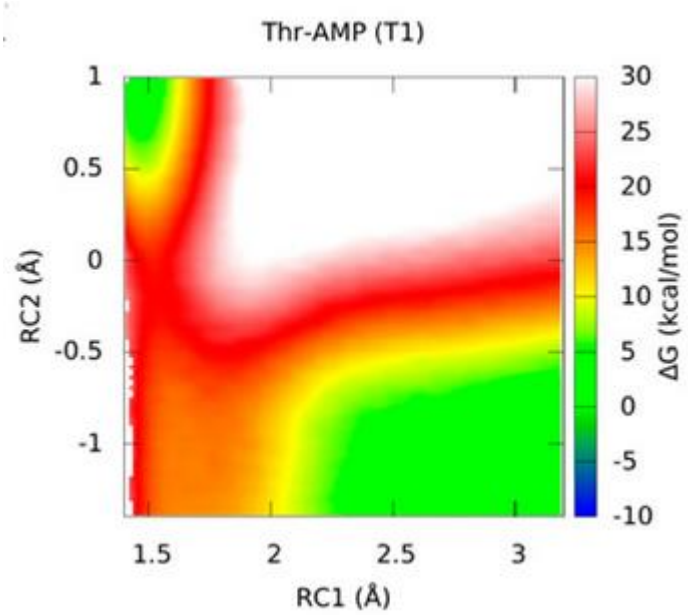


Figure 2.4. Example image of a free energy surface obtained from a defined reaction coordinate using umbrella sampling. High energy regions are shown in red and white, while lower energies are represented using green and blue. Paths can be traced from two regions of low energy to connect a reactant state to a product state. The example was taken from a study on pretransfer editing of ThrRS by Wei et al.[89]

2.5 Closing Remarks

The commonly used methods used in computational enzymology, QM-cluster, QM/MM, MD, and QM/MM-MD were briefly described along with their advantages and drawbacks. Understanding these is vital to the proper application of each technique. Neither is objectively superior to the other, and the choice of technique used is dependent on the chemistry of interest. Often, a study will make use of as many of these as possible to gain a fundamental understanding before applying to a larger model. Together, this multiscale approach gives more reliable results that build off each other and reinforce findings. DFT in particular has been an important driving force for the routine application of computational chemistry and has enabled scalability with reasonable accuracy at a much lower cost. Large active site models have been made possible which unlocked the potential to study enzymes considering as many catalytic factors as possible.

Computational power and program codes have developed to the point where any of today's consumer devices can perform calculations that were once considered intense and required specialized hardware. Protein dynamics and enzyme mechanistic pathways can be explored from a personal computer anywhere on the planet. Yet, creating a model remains a nontrivial task. Experimental studies give us insight into some of the key residues in the active site. Larger models are not necessarily more representative of the chemistry involved and is in practice more disadvantageous due to significantly increased costs. One solution may come from artificial intelligence. Machine learning approaches are quickly becoming prevalent in chemistry to study phenomena where large pools of data are available. Foundational studies have mainly demonstrated its use in drug discovery,[98] but now models have been extended to protein folding,[99] enzyme engineering,[100] and even catalysis.[101] Simply browsing through ChemRxiv many articles can be seen showcasing applications of machine learning within computational chemistry.

One other potential improvement may come from applications of quantum mechanics code to GPUs. This approach has been incredibly successful for MD,[102-104] yet application for quantum chemical calculations is still in its infancy. Initial studies have shown that use of GPU architecture can accelerate calculations by 20-30X,[105] but its implementation is not ideal. Post-SCF and smaller models are still more effectively optimized on CPUs using Gaussian leaving much room for improvement.

2.6 References

1. Lipkowitz, K. B., Applications of Computational Chemistry to the Study of Cyclodextrins. *Chem. Rev.* **1998,** *98,* 1829-1874.

2. Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S., Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. *J. Med. Chem.* **2008,** *51,* 769-79.

3. Ren, X. G.; Rinke, P.; Joas, C.; Scheffler, M., Random-phase approximation and its applications in computational chemistry and materials science. *J. Mater. Sci.* **2012,** *47,* 7447-7471.

4. Schlegel, H. B., Perspective on "Ab initio calculation of force constants and equilibrium geometries in polyatomic molecules. I. Theory" - Pulay P (1969) Mol Phys 17 : 197-204. *Theor. Chem. Acc.* **2000,** *103,* 294-296.

5. Zhang, D. W.; Xiang, Y.; Zhang, J. Z., New Advance in Computational Chemistry: Full Quantum Mechanical ab Initio Computation of Streptavidin-Biotin Interaction Energy. *J. Phys. Chem. B* **2003,** *107,* 12039-41.

6. Arakawa, H.; Aresta, M.; Armor, J. N.; Barteau, M. A.; Beckman, E. J.; Bell, A. T.; Bercaw, J. E.; Creutz, C.; Dinjus, E.; Dixon, D. A.; Domen, K.; DuBois, D. L.; Eckert, J.; Fujita, E.; Gibson, D. H.; Goddard, W. A.; Goodman, D. W.; Keller, J.; Kubas, G. J.; Kung,

H. H.; Lyons, J. E.; Manzer, L. E.; Marks, T. J.; Morokuma, K.; Nicholas, K. M.; Periana, R.; Que, L.; Rostrup-Nielson, J.; Sachtler, W. M.; Schmidt, L. D.; Sen, A.; Somorjai, G. A.; Stair, P. C.; Stults, B. R.; Tumas, W., Catalysis research of relevance to carbon management: progress, challenges, and opportunities. *Chem. Rev.* **2001,** *101*, 953-96.

7. Quesne, M. G.; Silveri, F.; de Leeuw, N. H.; Catlow, C. R. A., Advances in Sustainable Catalysis: A Computational Perspective. *Front. Chem.* **2019,** *7*, 182.

8. Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; St Clair, J. L.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L.; Houk, K. N.; Michael, F. E.; Baker, D., Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **2010,** *329*, 309-13.

9. Cherry, M.; Islam, M. S.; Gale, J. D.; Catlow, C. R. A., Computational Studies of Proton Migration in Perovskite Oxides. *Solid State Ion.* **1995,** *77*, 207-209.

10. Pascualahuir, J. L.; Silla, E.; Tomasi, J.; Bonaccorsi, R., Electrostatic Interaction of a Solute with a Continuum - Improved Description of the Cavity and of the Surface Cavity Bound Charge-Distribution. *J. Comput. Chem.* **1987,** *8*, 778-787.

11. Urban, J. J.; Tillman, B. G.; Cronin, W. A., Fluoroolefins as peptide mimetics: A computational study of structure, charge distribution, hydration, and hydrogen bonding. *J. Phys. Chem. A* **2006,** *110*, 11120-11129.

12. Verstraelen, T.; Sukhomlinov, S. V.; Van Speybroeck, V.; Waroquier, M.; Smirnov, K. S., Computation of Charge Distribution and Electrostatic Potential in Silicates with the Use of Chemical Potential Equalization Models. *J. Phys. Chem. C* **2012,** *116*, 490-504.

13. Izgorodina, E. I.; Seeger, Z. L.; Scarborough, D. L. A.; Tan, S. Y. S., Quantum Chemical Methods for the Prediction of Energetic, Physical, and Spectroscopic Properties of Ionic Liquids. *Chem. Rev.* **2017,** *117*, 6696-6754.

14. Neese, F., Prediction of molecular properties and molecular spectroscopy with density functional theory: From fundamental theory to exchange-coupling. *Coord. Chem. Rev.* **2009,** *253*, 526-563.

15. Ou-Yang, S. S.; Lu, J. Y.; Kong, X. Q.; Liang, Z. J.; Luo, C.; Jiang, H., Computational drug discovery. *Acta Pharmacol. Sin.* **2012,** *33*, 1131-40.

16. Wilson, G. L.; Lill, M. A., Integrating structure-based and ligand-based approaches for computational drug design. *Future Med. Chem.* **2011,** *3*, 735-50.

17. Wolber, G.; Seidel, T.; Bendix, F.; Langer, T., Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* **2008,** *13*, 23-9.

18. Evers, A.; Klabunde, T., Structure-based drug discovery using GPCR homology modeling: Successful virtual screening for antagonists of the Alpha1A adrenergic receptor. *J. Med. Chem.* **2005,** *48*, 1088-1097.

19. Krieger, E.; Nabuurs, S. B.; Vriend, G., Homology modeling. *Methods Biochem. Anal.* **2003,** *44*, 509-23.

20. Ogawa, H.; Toyoshima, C., Homology modeling of the cation binding sites of Na+K+-ATPase. *Proc. Natl. Acad. Sci. U. S. A.* **2002,** *99*, 15977-82.

21. Groot, R. D.; Rabone, K. L., Mesoscopic simulation of cell membrane damage, morphology change and rupture by nonionic surfactants. *Biophys. J.* **2001,** *81*, 725-736.

22. Shi, Q.; Izvekov, S.; Voth, G. A., Mixed atomistic and coarse-grained molecular dynamics: Simulation of a membrane-bound ion channel. *J. Phys. Chem. B* **2006,** *110*, 15045-15048.

23. Vanderploeg, P.; Berendsen, H. J. C., Molecular-Dynamics Simulation of a Bilayer-Membrane. *J. Chem. Phys.* **1982,** *76*, 3271-3276.

24. Bromberg, Y.; Rost, B., Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics* **2008,** *24*, i207-12.

25. You, L. C.; Yin, J., Dependence of epistasis on environment and mutation severity as revealed by in silico mutagenesis of phage T7. *Genetics* **2002,** *160*, 1273-1281.

26. Metz, S.; Thiel, W., QM/MM studies of xanthine oxidase: variations of cofactor, substrate, and active-site Glu802. *J Phys Chem B* **2010,** *114*, 1506-17.

27. Warshel, A.; Levitt, M., Theoretical Studies of Enzymic Reactions - Dielectric, Electrostatic and Steric Stabilization of Carbonium-Ion in Reaction of Lysozyme. *J. Mol. Biol.* **1976,** *103*, 227-249.

28. Cramer, C. J., *Essentials of computational chemistry: theories and models*. John Wiley & Sons: 2013.

29. Roothaan, C. C. J., New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951,** *23*, 69-89.

30. Boys, S. F.; Cook, G. B.; Reeves, C. M.; Shavitt, I., Automatic Fundamental Calculations of Molecular Structure. *Nature* **1956,** *178*, 1207-1209.

31. Buenker, R. J.; Peyerimhoff, S. D., Ab initio SCF calculations for azulene and naphthalene. *Chem. Phys. Lett.* **1969,** *3*, 37-42.

32. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993,** *97*, 10269-10280.

33. Veillard, A., Ab initio calculations of transition-metal organometallics: structure and molecular properties. *Chem. Rev.* **1991,** *91*, 743-766.

34. Salomon, O.; Reiher, M.; Hess, B. A., Assertion and validation of the performance of the B3LYP⋆ functional for the first transition metal row and the G2 test set. *J. Chem. Phys.* **2002,** *117*, 4729-4737.

35. Gill, P. M. W., Density Functional Theory (DFT), Hartree–Fock (HF), and the Self-consistent Field. In *Encyclopedia of Computational Chemistry*, 1998.

36. Bauschlicher, C. W., A comparison of the accuracy of different functionals. *Chem. Phys. Lett.* **1995,** *246*, 40-44.

37. Lynch, B. J.; Truhlar, D. G., How Well Can Hybrid Density Functional Methods Predict Transition State Geometries and Barrier Heights? *J. Phys. Chem. A* **2001,** *105*, 2936-2941.

38. Mourik, T. v.; Gdanitz, R. J., A critical note on density functional theory studies on rare-gas dimers. *J. Chem. Phys.* **2002,** *116*, 9620-9623.

39. Vondrášek, J.; Bendová, L.; Klusák, V.; Hobza, P., Unexpectedly Strong Energy Stabilization Inside the Hydrophobic Core of Small Protein Rubredoxin Mediated by Aromatic Residues: Correlated Ab Initio Quantum Chemical Calculations. *J. Am. Chem. Soc.* **2005,** *127*, 2615-2619.

40. Runge, E.; Gross, E. K. U., Density-Functional Theory for Time-Dependent Systems. *Phys. Rev. Lett.* **1984,** *52*, 997-1000.

41. Kavanagh, M. A.; Karlsson, J. K. G.; Colburn, J. D.; Barter, L. M. C.; Gould, I. R., A TDDFT investigation of the Photosystem II reaction center: Insights into the precursors to charge separation. *Proc. Natl. Acad. Sci. USA* **2020,** *117*, 19705-19712.

42. Medvedev, M. G.; Bushmarinov, I. S.; Sun, J.; Perdew, J. P.; Lyssenko, K. A., Density functional theory is straying from the path toward the exact functional. *Science* **2017,** *355*, 49-52.

43. Perdew, J. P.; Schmidt, K., Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conf. Proc.* **2001,** *577*, 1-20.

44. Goerigk, L.; Mehta, N., A Trip to the Density Functional Theory Zoo: Warnings and Recommendations for the User. *Aust. J. Chem.* **2019,** *72*, 563-573.

45. Friesner, R. A.; Guallar, V., Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.* **2005,** *56*, 389-427.

46. Meunier, B.; de Visser, S. P.; Shaik, S., Mechanism of oxidation reactions catalyzed by cytochrome p450 enzymes. *Chem. Rev.* **2004,** *104*, 3947-80.

47. Kruse, H.; Goerigk, L.; Grimme, S., Why the standard B3LYP/6-31G* model chemistry should not be used in DFT calculations of molecular thermochemistry: understanding and correcting the problem. *J. Org. Chem.* **2012,** *77*, 10824-34.

48. Austin, A.; Petersson, G. A.; Frisch, M. J.; Dobek, F. J.; Scalmani, G.; Throssell, K., A Density Functional with Spherical Atom Dispersion Terms. *J. Chem. Theory Comput.* **2012,** *8*, 4989-5007.

49. Ehrlich, S.; Moellmann, J.; Grimme, S., Dispersion-corrected density functional theory for aromatic interactions in complex systems. *Acc. Chem. Res.* **2013,** *46*, 916-26.

50. Grimme, S., Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006,** *27*, 1787-99.

51. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010,** *132*, 154104.

52. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011,** *32*, 1456-65.

53. Dokainish, H. M.; Kitao, A., Similarities and Differences between Thymine(6-4)Thymine/Cytosine DNA Lesion Repairs by Photolyases. *J. Phys. Chem. B* **2018,** *122*, 8537-8547.

54. 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. of Quantum Chem.* **2018,** *118*, e25558.

55. Harvey, J. N.; Himo, F.; Maseras, F.; Perrin, L., Scope and Challenge of Computational Methods for Studying Mechanism and Reactivity in Homogeneous Catalysis. *ACS Catalysis* **2019,** *9*, 6803-6813.

56. Quesne, M. G.; Borowski, T.; de Visser, S. P., Quantum Mechanics/Molecular Mechanics Modeling of Enzymatic Processes: Caveats and Breakthroughs. *Chemistry (Easton)* **2016,** *22*, 2562-81.

57. Quesne, M. G.; Borowski, T.; de Visser, S. P., Quantum Mechanics/Molecular Mechanics Modeling of Enzymatic Processes: Caveats and Breakthroughs. *Chem. Eur. J.* **2016,** *22*, 2562-2581.

58. Ryu, H.; Park, J.; Kim, H. K.; Park, J. Y.; Kim, S.-T.; Baik, M.-H., Pitfalls in Computational Modeling of Chemical Reactions and How To Avoid Them. *Organometallics* **2018,** *37*, 3228-3239.

59. Lonsdale, R.; Harvey, J. N.; Mulholland, A. J., A practical guide to modelling enzyme-catalysed reactions. *Chem. Soc. Rev.* **2012,** *41*, 3025-38.

60. Coskun, D.; Jerome, S. V.; Friesner, R. A., Evaluation of the Performance of the B3LYP, PBE0, and M06 DFT Functionals, and DBLOC-Corrected Versions, in the Calculation of Redox Potentials and Spin Splittings for Transition Metal Containing Systems. *J. Chem. Theory Comput.* **2016,** *12*, 1121-1128.

61. Moltved, K. A.; Kepp, K. P., Performance of Density Functional Theory for Transition Metal Oxygen Bonds. *Chemphyschem* **2019,** *20*, 3210-3220.

62. Yang, Y.; Weaver, M. N.; Merz, K. M., Jr., Assessment of the "6-31+G** + LANL2DZ" mixed basis set coupled with density functional theory methods and the effective core potential: prediction of heats of formation and ionization potentials for first-row-transition-metal complexes. *J. Phys. Chem. A* **2009,** *113*, 9843-9851.

63. Ramos, M. J.; Fernandes, P. A., Computational enzymatic catalysis. *Acc. Chem. Res.* **2008,** *41*, 689-98.

64. Siegbahn, P. E.; Himo, F., Recent developments of the quantum chemical cluster approach for modeling enzyme reactions. *J. Biol. Inorg. Chem.* **2009,** *14*, 643-51.

65. Kazemi, M.; Sheng, X.; Himo, F., Origins of Enantiopreference of Mycobacterium smegmatis Acyl Transferase: A Computational Analysis. *Chemistry (Easton)* **2019,** *25*, 11945-11954.

66. Adam, P. S.; Borrel, G.; Gribaldo, S., Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc. Natl. Acad. Sci. USA* **2018,** *115*, E1166-E1173.

67. Liao, R. Z.; Siegbahn, P. E. M., Energetics for the Mechanism of Nickel-Containing Carbon Monoxide Dehydrogenase. *Inorg. Chem.* **2019,** *58*, 7931-7938.

68. Baker, D.; Sali, A., Protein structure prediction and structural genomics. *Science* **2001,** *294*, 93-6.

69. Blake, J. D.; Cohen, F. E., Pairwise sequence alignment below the twilight zone. *J. Mol. Biol.* **2001,** *307*, 721-35.

70. Chothia, C.; Lesk, A. M., The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986,** *5*, 823-826.

71. Sippl, M. J., Recognition of errors in three-dimensional structures of proteins. *Proteins* **1993,** *17*, 355-62.

72. Lazaridis, T.; Karplus, M., Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **1999,** *288*, 477-87.

73. Buchan, D. W. A.; Jones, D. T., The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **2019,** *47*, W402-W407.

74. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T., SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018,** *46*, W296-W303.

75. Capener, C. E.; Shrivastava, I. H.; Ranatunga, K. M.; Forrest, L. R.; Smith, G. R.; Sansom, M. S., Homology modeling and molecular dynamics simulation studies of an inward rectifier potassium channel. *Biophys. J.* **2000,** *78*, 2929-42.

76. Sanchez, R.; Sali, A., Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc. Natl. Acad. Sci. USA* **1998,** *95*, 13597-602.

77. Chung, L. W.; Sameera, W. M.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z.; Liu, F.; Li, H. B.; Ding, L.; Morokuma, K., The ONIOM Method and Its Applications. *Chem. Rev.* **2015,** *115*, 5678-796.

78. Gherib, R.; Dokainish, H. M.; Gauld, J. W., Multi-scale computational enzymology: enhancing our understanding of enzymatic catalysis. *Int. J. Mol. Sci.* **2013,** *15*, 401-22.

79. Kellie, J. L.; Wetmore, S. D., Selecting DFT methods for use in optimizations of enzyme active sites: applications to ONIOM treatments of DNA glycosylases. *Can. J. Chem.* **2013,** *91*, 559-572.

80. Lin, H.; Truhlar, D. G., QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **2006,** *117*, 185.

81. Sousa, S. F.; Ribeiro, A. J. M.; Neves, R. P. P.; Brás, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J., Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. *WIREs Comput. Mol. Sci.* **2017,** *7*, e1281.

82. van der Kamp, M. W.; Mulholland, A. J., Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013,** *52*, 2708-2728.

83. Cao, L.; Ryde, U., On the Difference Between Additive and Subtractive QM/MM Calculations. *Front. Chem.* **2018,** *6*, 89-89.

84. Aboelnga, M. M.; Hayward, J. J.; Gauld, J. W., Enzymatic Post-Transfer Editing Mechanism of E. coli Threonyl-tRNA Synthetase (ThrRS): A Molecular Dynamics (MD) and Quantum Mechanics/Molecular Mechanics (QM/MM) Investigation. *ACS Catalysis* **2017,** *7*, 5180-5193.

85. Ferreira, P.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Romão, M. J.; Ramos, M. J., Catalytic Mechanism of Human Aldehyde Oxidase. *ACS Catalysis* **2020,** *10*, 9276-9286.

86. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983,** *4*, 187-217.

87. Scott, W. R. P.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F., The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* **1999,** *103* (19), 3596-3607.

88. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004,** *25*, 1157-1174.

89. Wei, W.; Gauld, J. W.; Monard, G., Pretransfer Editing in Threonyl-tRNA Synthetase: Roles of Differential Solvent Accessibility and Intermediate Stabilization. *ACS Catalysis* **2017,** *7*, 3102-3112.

90. Ribeiro, A. J. M.; Santos-Martins, D.; Russo, N.; Ramos, M. J.; Fernandes, P. A., Enzymatic Flexibility and Reaction Rate: A QM/MM Study of HIV-1 Protease. *ACS Catalysis* **2015,** *5*, 5617-5626.

91. Bernardi, R. C.; Melo, M. C. R.; Schulten, K., Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim Biophys Acta* **2015,** *1850*, 872-877.

92. Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q., Enhanced sampling in molecular dynamics. *J. Chem. Phys.* **2019,** *151*, 070902.

93. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999,** *314*, 141-151.

94. Awoonor-Williams, E.; Rowley, C. N., Evaluation of Methods for the Calculation of the pKa of Cysteine Residues in Proteins. *J. Chem. Theory Comput.* **2016,** *12*, 4662-73.

95. Harris, R. C.; Liu, R.; Shen, J., Predicting Reactive Cysteines with Implicit-Solvent-Based Continuous Constant pH Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **2020,** *16*, 3689-3698.

96. Kühne, T. D., Second generation Car–Parrinello molecular dynamics. *WIREs Comput. Mol. Sci.* **2014,** *4*, 391-406.

97. Brunk, E.; Rothlisberger, U., Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.* **2015,** *115*, 6217-6263.

98. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S., Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019,** *18*, 463-477.

99. Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D., Improved

protein structure prediction using potentials from deep learning. *Nature* **2020,** *577*, 706-710.

100. Mazurenko, S.; Prokop, Z.; Damborsky, J., Machine Learning in Enzyme Engineering. *ACS Catalysis* **2020,** *10*, 1210-1223.

101. Bonk, B. M.; Weis, J. W.; Tidor, B., Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* **2019,** *141*, 4108-4118.

102. Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012,** *8*, 1542-1555.

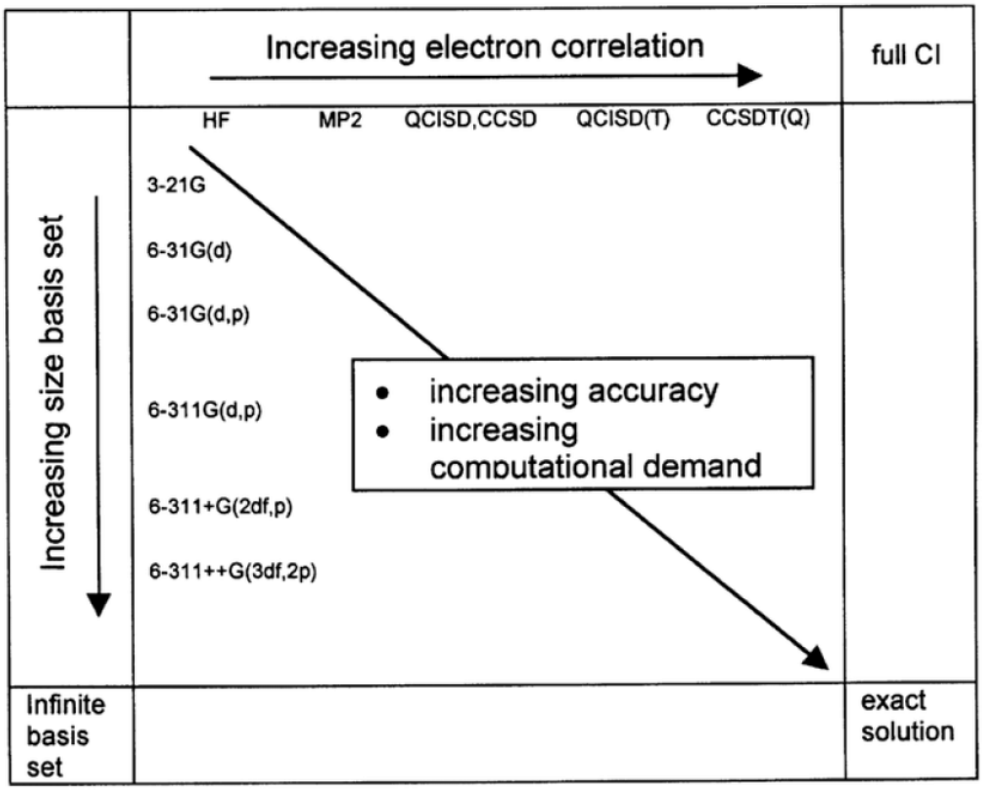
103. Le Grand, S.; Götz, A. W.; Walker, R. C., SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput. Phys. Commun.* **2013,** *184*, 374-380.

104. Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013,** *9*, 3878-3888.

105. Nitsche, M. A.; Ferreria, M.; Mocskos, E. E.; González Lebrero, M. C., GPU Accelerated Implementation of Density Functional Theory for Hybrid QM/MM Simulations. *J. Chem. Theory Comput.* **2014,** *10*, 959-967.

Chapter 3:

An Assessment on Computational Methods for Calculating Accurate Structures and Energies of Bio-Relevant Polysulfur/Selenium-containing Compounds



3.1 Introduction

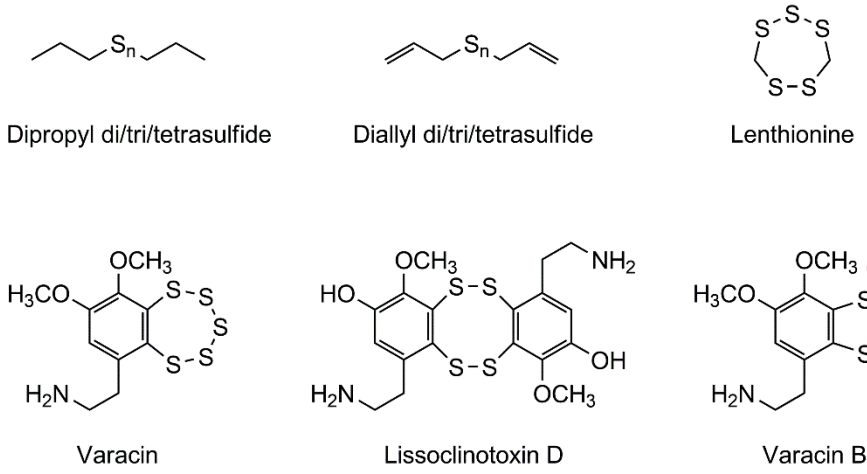
The chalcogens sulfur and selenium have long been known to play key roles in a diverse array of important physiological and biological processes including enzymatic mechanisms,[1-2] signalling,[3-5] and mediation and repair of oxidatively-damaged biomolecules.[6-8] Indeed, in addition to being found within 3 of the 22 proteinogenic amino acids; cysteine (Cys), methionine (Met), and selenocysteine (Sec), they are also found in many essential metabolites (*e.g.* thiazole). This is due to their ability to possess a broad range of oxidation states as well as bonding environments, and can often undergo reversible redox.[9] For instance, the most abundant antioxidant in animal cells is glutathione (GSH) which can mediate the cellular redox environment through interconversion with its oxidized disulfide form, GSSG.[10]

Recently, it has increasingly been recognized that Reactive Sulfur Species (RSS) comprise a rich and diverse range of physiologically important species (**Scheme 3.1**).[11] Indeed, hydrogen sulfide is now known to be a ubiquitous essential signaling molecule that plays key roles in many physiological and inflammatory processes including blood pressure regulation, cell proliferation and apoptosis, insulin signaling, and neurotransmission.[7, 12-16] However, sulfur can also form strong homonuclear single bonds and as a result can react with other sulfur species (for example, in proteins) to form a variety of polysulfur-containing RSS.[17]

Previously, the presence of per- (HSSH) and polysulfides (*e.g.* H_2S_n n=3-7) in biological systems were thought to be experimental artefacts or stores for H_2S.[18] Now, however, they are increasingly proposed or recognised as being biochemically important;[17, 19-22] for instance, some polysulfides have been shown to possess antibiotic or anticancer properties.[23-24] More recently, Cysteinyl-tRNA synthetase (CysRS), an ancient enzyme with a critical role in gene-encoded protein synthesis, has been shown to also catalyze the

formation of Cys-derived polysulfides.[21] This further underscores the potential biologically important activity of peptide-hydropersulfides.[25] Meanwhile, Se has been shown to form an Se-S intermediate in the selenoproteins thioredoxin reductase and formate dehydrogenase and plays a central role in the enzyme's activation.[26-27] Unfortunately, due to the high reactivity of RSS, specifically per/polysulfides within biological environments, many of their properties and much of their chemistry remains unclear or even unknown.

Scheme 3.1. Several examples of naturally occurring polysulfur species.[17, 28]



Computational chemistry has established itself as a key tool for the study of the properties and chemistry of biomolecular systems. Several such studies have been performed on S- and, to a notably lesser extent, Se- or mixed S/Se-containing per- and poly-seleno/sulfides. For example, Brzostowska *et al*. in part used the B3LYP method to examine the intramolecular reactions of the naturally occurring polysulfur-containing pentathiepins (such as varacin, Scheme 3.1) that generate S_3 and S_2 transfer units via a tetra- or trisulfide anion, respectively.[29] Recently, the high reactivity of several smaller hydropersulfides toward alkyl, alkoxy, peroxyl and thiyl radicals was investigated using both experimental and computational (CBS-QB3) methods.[30] It was concluded that such reactions are exothermic by 15-34 kcal mol^{-1} due to the low RSS–H bond dissociation enthalpy and

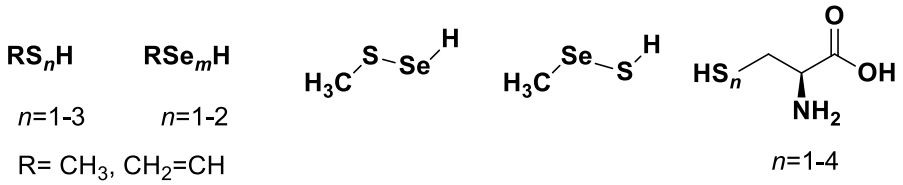
highy stability of perthiyl radicals.[30] Meanwhile, a computational study has used in part dispersion-corrected B3LYP (B3LYP-D3) to examine the role of the chalcogen atoms in the mechanism of glutathione peroxidase 4 which involves formation of a -Se–S- species.[31] Bachrach *et al*. used several computational methods including MP2, and B3LYP to examine possible mechanisms for nucleophilic attack at the Se in diselenides and selenosulfides and concluded that attack at Se is kinetically and thermodynamically preferred.[32] Using an ONIOM QM/MM approach wherein the DFT method M06-2X was used for the QM region, Huang *et al*. examined *S*-sulfhydration *via* a persulfide (RSS^-) intermediate as catalyzed by mercaptopyruvate sulfurtransferase (MST) and obtained reasonable agreement with experiment.[33] In all of these studies, smaller basis sets (*e.g.* 6-31G(d)) were used to obtain structures upon which they then based their calculations of thermochemical properties. The ability of a computational study to reliably and accurately provide insights into the structures and properties of any biomolecular system usually critically depends on the choice of QM method and basis set. Thus, an essential step towards computationally studying RSS is determining appropriate methodologies.

In this present study, the ability of a range of computational, in particular density functional theory (DFT) methods[34], to provide reliable and accurate structures and thermochemical properties of biologically relevant poly-sulfur/selenium containing compounds has been assessed. More specifically, the DFT methods B3LYP, B3PW91, ωB97XD, M06-2X, and M08-HX were applied to a systematic series of biologically relevant $RX_n(H)$ (X= S/Se, n=1-3, R=CH_3, CH_2=CH and cysteine, Scheme 3.2) species. As well as their structures, a variety of key thermochemical properties including their X–X (chalcogen-chalcogen) bond dissociation enthalpies, hydrogen affinities, and gas phase basicities were examined and benchmarked.

3.2 Methods

All calculations were performed using the Gaussian 09[35] and Gaussian 16[36] suites of programs. Optimized geometries for a systematic series of biologically relevant chemical models, shown in Scheme 3.2, were obtained using a variety of conventional wavefunction and density functional theory methods applied in conjunction with a range of Pople basis sets from 6-31G(d) to 6-311++G(3df,3pd).

Scheme 3.2. Schematic illustration of the species considered in this study.



Specifically, the hybrid DFT methods B3LYP and B3PW91, comprised of Becke's three parameter exchange functional[37] (B3) in combination with the Lee, Yang, and Parr correlation functional[38] (LYP) or the Perdew and Wang functional[39] (PW91), were assessed. B3LYP is commonly applied in the treatment of biomolecular systems while B3PW91 has been used previously for systems containing sulfur and selenium.[40-41] In addition, two *meta*-GGA functionals were assessed; M06-2X[42], a commonly employed functional in the study of enzymatic catalysis, and the more recently developed related functional M08-HX.[43] For this functional, optimized structures were obtained using Gaussian 16 whereas all other structures were obtained with Gaussian 09. Furthermore, the double-hybrid range-corrected functional ωB97XD[44] was also evaluated to further examine any effects of dispersion correction on geometric or energetic parameters. Geometries optimized at these levels of theory were compared to those obtained using the *ab initio* MP2 (see Appendix A) and QCISD methods. As DFT methods are the preferred choice for biochemical systems due to their computational cost and reliability, the MP2 results are

only shown in Appendix A3.1 for our baseline geometry assessment of $CH_3SSH/^-$ and $CH_3SeSeH/^-$. The Minnesota family of functionals (M06-2X and M08-HX) are more empirical and explicitly contain dispersion correction while B3LYP only contains 3 empirical parameters and has no explicit correction for dispersion.

All optimized structures were confirmed to be minima by harmonic vibrational frequency calculations performed at the same level of theory. These were also used to determine zero-point vibrational energy (ZPVE) and enthalpy corrections for subsequent calculations of a range of their thermochemical properties including proton affinities ($^{298.15K}PA_{A^-} = -\Delta H = -(H_{AH} - H_{A^-} - H_{H^+})$), gas-phase basicities ($^{298.15K}GPB_{A^-} = -\Delta G = -(-PA - T(S_{AH} - S_{A^-} - S_{H^+})$), and hydrogen affinities ($^{298.15K}HA_A = -\Delta H = -(H_{AH} - H_A - H_H)$).[45-46] In addition, homolytic bond dissociation enthalpies were calculated for production of the ˙SH or ˙SeH radicals ($^{298.15K}BDE_{AX-XH} = \Delta H = H_{AX}\bullet + H_{XH}\bullet - H_{AXXH}$).

3.3 Results and Discussion

3.3.1 Structural Assessment of CH_3XXH and CH_3XX^- (X=S,Se)

We began by using the broadest variety of DFT methods (B3LYP, B3PW91, ωB97XD, M06-2X, M08-HX) and range of basis sets (6-31G(d) to 6-311++G(3df,3pd)) used in this study to obtain optimized structures for CH_3XXH and CH_3XX^- (X=S,Se). These are the smallest homoatomic persulf/selenides considered in this present study. Due to a paucity of corresponding experimental data, benchmark optimized structures were obtained at the QCISD/6-311+G(2df,p) level of theory. For simplicity, only the key C–X, X–X, and X–H (X=S,Se) distances are discussed herein and are summarized in Table 3.1 and MP2 data is shown in Appendix A3.1.

Method Sensitivity to Basis Set Changes: All five DFT methods showed similar overall sensitivities to changes in the basis set from 6-31G(d) to 6-311++G(3df,3pd). For example, the smallest variations were observed for the C—X and X—H (X=S, Se) bond distances; for any given method and chemical system (be it neutral or anionic) they varied by ≤ 0.020 Å. The only exceptions occurred for the C—Se bond in CH_3SeSe^- when using M06-2X method which varied overall by 0.027 Å, and the Se—H bond using MP2 which varied by 0.025 Å (Appendix A3.1). Notably, for any DFT method and chemical system the C—S bond distances showed greater variation than that of C—Se bonds. The only exceptions to this trend occurred for the C–X bonds in the CH_3XX^- anions when using the B3LYP or M06-2X methods. However, the converse trend was observed for the X–H bonds; optimized Se–H bond lengths are more sensitive to basis set choice than S–H bonds (Table 3.1 and Appendix A3.1).

For all chemical systems (both neutral and anionic), regardless of the choice of DFT method, the largest variations upon changing basis set were observed in their X—X bonds. For CH_3SSH and CH_3SS^- it was found that the S—S bond distances varied by 0.032 – 0.043 and 0.044 – 0.051 Å, respectively. Meanwhile, in the corresponding CH_3SeSeH and CH_3SeSe^- species the Se—Se bond distances varied by 0.024 – 0.029 and 0.023 – 0.029 Å, respectively, except for using MP2 where the variation exceeded the trend for DFT data (Appendix A3.1). That is, except for Se—H bonds, those bonds involving Sulfur (*i.e.* C—X and X—X) in CH_3XXH and CH_3XX^- (X=S,Se) are most sensitive to the choice of basis set.

Effects of Increasing Basis Set Size: As can be seen in Table 3.1, for all DFT methods considered and for both neutral CH_3XXH and anionic CH_3XX^- (X=S,Se), similar trends were generally observed upon increasing the basis set from double- to triple-zeta, and then subsequently by inclusion of diffuse and polarization functions.

For instance, increasing the basis set from 6-31G(d) to 6-311++G(3df,3pd) generally caused a systematic shortening in the X—H and C—X bonds. Notably, one does not need to increase the basis set significantly in order to get reasonable agreement with the corresponding values obtained at the QCISD/6-311+G(2df,p) level of theory. In fact, upon changing the basis set from 6-31G(d) to 6-311G(d) (*i.e.* double- to triple-zeta) resulted in their lengths differing from the benchmark values by ≤ 0.015 Å. Further increases in basis set size by inclusion of diffuse (*i.e.* 6-311G(d,p) to 6-311+G(d,p)) or *f*- and/or *d*-polarization functions on heavy atoms or hydrogen resulted in only minor individual decreases.

In contrast, modifying the basis set showed quite different trends for the X—X (X = S, Se) bonds. For instance, for all methods assessed, increasing the basis set from 6-31G(d) to 6-311G(d) resulted in an increase in their optimized length for all chemical systems of up to 0.025 Å. The only exception occurred for the M06-2X method applied to CH_3SeSe^- for which the Se—Se bond shortened slightly by 0.008 Å. Notably, in CH_3SSH and CH_3SS^- the observed ranges of bond lengthening were 0.009 – 0.015 and 0.014 – 0.027 Å respectively, with the largest increases observed when using the B3LYP method. That is to say, the anionic persulfide is more sensitive to basis set changes than the neutral hydropersulfide. In contrast, for CH_3SeSeH and CH_3SeSe^- the observed increases were quite similar lying in the range of 0.017 – 0.025 Å (except where noted above). The further inclusion of *p*-functions on hydrogen (*i.e.* 6-311G(d) to 6-311G(d,p)) or diffuse functions on heavy atoms (*i.e.* 6-311G(d,p) to 6-311+G(d,p)) had negligible effect on the X—X bond lengths in both the neutral CH_3XXH and anionic CH_3XX^- (X=S,Se) systems.

Table 3.1. Selected optimized bond lengths in ångström (Å) for CH_3XXH and CH_3XX^- (X=S,Se).

		CH_3SSH			CH_3SS^-		CH_3SeSeH			CH_3SeSe^-	
Method	**Basis set**	C–S	S–S	S–H	C–S	S–S	C–Se	Se-Se	*Se*–H	C–Se	Se-Se
B3LYP	6-31G(d)	1.834	2.091	1.357	1.837	2.117	1.977	2.337	1.495	1.985	2.366
	6-311G(d)	1.830	2.103	1.357	1.831	2.134	1.979	2.362	1.488	1.987	2.391
	6-311G(d,p)	1.831	2.105	1.354	1.833	2.134	1.978	2.364	1.478	1.987	2.391
	6-311+G(d,p)	1.830	2.103	1.354	1.832	2.128	1.978	2.364	1.478	1.987	2.388
	6-311G(2d,p)	1.830	2.088	1.349	1.831	2.115	1.975	2.366	1.476	1.984	2.395
	6-311G(df,p)	1.829	2.093	1.354	1.831	2.120	1.973	2.351	1.479	1.981	2.376
	6-311+G(2df,p)	1.825	2.072	1.351	1.827	2.092	1.971	2.350	1.477	1.979	2.373
	6-311++G(3df,3pd)	1.820	2.064	1.349	1.822	2.083	1.971	2.351	1.476	1.978	2.373
B3PW91	6-31G(d)	1.821	2.072	1.354	1.824	2.096	1.961	2.315	1.491	1.960	2.336
	6-311G(d)	1.817	2.081	1.356	1.819	2.110	1.963	2.338	1.486	1.963	2.361
	6-311G(d,p)	1.818	2.083	1.353	1.820	2.110	1.962	2.340	1.476	1.962	2.360
	6-311+G(d,p)	1.818	2.083	1.353	1.821	2.105	1.962	2.339	1.476	1.962	2.358
	6-311G(2d,p)	1.817	2.067	1.349	1.819	2.091	1.959	2.342	1.474	1.959	2.363
	6-311G(df,p)	1.816	2.072	1.353	1.818	2.097	1.957	2.327	1.477	1.958	2.345
	6-311+G(2df,p)	1.812	2.053	1.351	1.814	2.071	1.955	2.327	1.475	1.955	2.342
	6-311++G(3df,3pd)	1.808	2.046	1.349	1.810	2.063	1.955	2.327	1.475	1.955	2.342
ωB97XD	6-31G(d)	1.819	2.070	1.349	1.821	2.095	1.954	2.309	1.486	1.968	2.342
	6-311G(d)	1.816	2.079	1.350	1.816	2.110	1.957	2.331	1.480	1.971	2.365
	6-311G(d,p)	1.816	2.081	1.348	1.817	2.110	1.956	2.333	1.470	1.970	2.365
	6-311+G(d,p)	1.816	2.080	1.348	1.817	2.105	1.956	2.333	1.470	1.970	2.362
	6-311G(2d,p)	1.816	2.067	1.344	1.816	2.092	1.953	2.334	1.469	1.967	2.369
	6-311G(df,p)	1.815	2.070	1.348	1.815	2.099	1.951	2.320	1.472	1.965	2.351
	6-311+G(2df,p)	1.810	2.053	1.346	1.811	2.074	1.950	2.319	1.470	1.963	2.348
	6-311++G(3df,3pd)	1.806	2.046	1.344	1.807	2.066	1.949	2.320	1.469	1.962	2.347
M06-2X	6-31G(d)	1.819	2.069	1.348	1.819	2.096	1.954	2.306	1.486	1.985	2.366
	6-311G(d)	1.816	2.078	1.348	1.816	2.110	1.960	2.330	1.480	1.964	2.358
	6-311G(d,p)	1.817	2.079	1.346	1.817	2.110	1.959	2.331	1.472	1.964	2.358
	6-311+G(d,p)	1.816	2.089	1.346	1.816	2.105	1.959	2.331	1.472	1.964	2.355
	6-311G(2d,p)	1.817	2.066	1.342	1.816	2.093	1.956	2.334	1.471	1.961	2.362
	6-311G(df,p)	1.814	2.068	1.347	1.816	2.097	1.954	2.320	1.473	1.960	2.345
	6-311+G(2df,p)	1.811	2.052	1.345	1.812	2.073	1.953	2.319	1.472	1.958	2.340
	6-311++G(3df,3pd)	1.808	2.046	1.342	1.808	2.065	1.953	2.319	1.472	1.958	2.341
M08-HX	6-31G(d)	1.817	2.068	1.350	1.817	2.094	1.955	2.305	1.488	1.958	2.330
	6-311G(d)	1.816	2.078	1.345	1.815	2.108	1.961	2.327	1.482	1.964	2.352
	6-311G(d,p)	1.816	2.079	1.348	1.815	2.108	1.960	2.328	1.474	1.963	2.352
	6-311+G(d,p)	1.816	2.079	1.349	1.815	2.103	1.960	2.328	1.474	1.964	2.348
	6-311G(2d,p)	1.814	2.067	1.344	1.814	2.093	1.957	2.331	1.473	1.960	2.357
	6-311G(df,p)	1.814	2.069	1.349	1.814	2.095	1.954	2.318	1.475	1.959	2.341
	6-311+G(2df,p)	1.809	2.053	1.347	1.810	2.072	1.953	2.317	1.474	1.957	2.335
	6-311++G(3df,3pd)	1.807	2.047	1.344	1.808	2.063	1.953	2.318	1.472	1.957	2.336
QCISD	6-311+G(2df,p)	1.815	2.065	1.347	1.816	2.088	1.965	2.333	1.474	1.961	2.359

In general, more significant changes in the X—X lengths were observed upon inclusion of either a second set of *d*- or a set of *f*-functions on heavy atoms (*i.e.* 6-311G(d,p) to 6-311G(2d,p) or 6-311G(df,p)). Specifically, for both these basis set changes the S—S bonds in CH_3SSH and CH_3SS^- shortened by 0.010 to 0.017 Å, except for MP2 where the distance increases for the 6-311G(2d,p) basis set. In contrast, for the corresponding selenium containing systems, the inclusion of a second set of *d*-functions on heavy atoms marginally lengthened Se—Se bonds by ≤ 0.004 Å, while the inclusion of a set of *f*-functions on heavy atoms shortened the Se—Se bonds by 0.010 – 0.017 Å. Combining or adding further diffuse and polarization functions by use of the 6-311+G(2df,p) or 6-311++G(3df,3pd) basis sets respectively, resulted in all DFT methods except B3LYP giving X—X (X= S, Se) bond lengths that were markedly shorter than their corresponding benchmark values. In the case of B3LYP (which overestimates the length of these bonds) increasing the basis set to 6-311+G(2df,p) or 6-311++G(3df,3pd) was in fact required in order to get good agreement with the benchmark values.

Importantly, for all DFT methods that were considered (with the exception of B3LYP), the 6-311G(2d,p) and 6-311G(df,p) basis sets gave optimized C—X, X—X, and X—H distances for CH_3XXH and CH_3XX^- (X = S) that were in closest general agreement with their corresponding benchmark values. Meanwhile, for the corresponding Se analogues, the best performing basis sets were generally 6-311G(d), 6-311G(d,p), 6-311+G(d,p), and 6-311G(2d,p). It is also noted that the M08-HX method does not offer much if any improvement over M06-2X. In fact, when Se is in the system, it slightly underestimates the bond lengths. As such, subsequent tables showing optimized parameters will only include results obtained using the B3PW91, ωB97XD and M06-2X methods in combination with the identified preferred basis sets. It should be noted that, for completeness, the corresponding values for the other methods are included in Appendix A. Since MP2 trends were mostly similar to DFT, we do not include results obtained with this method.

3.3.2 Effect of Conjugation: CH_2=CHXXH and CH_2=CHXX^- (X=S,Se)

These model systems were examined to gain insights into the influence of conjugation on the neutral and anionic persulfide and perselenide groups. Based on the results obtained for the CH_3XX^-/H systems, optimized structures were obtained using only the B3PW91, ωB97XD, and M06-2X methods in combination with the 6-311+G(d,p), 6-311G(2d,p), and 6-311G(df,p) basis sets. The most significant changes were observed in their C—X and X—X bonds, hence only these optimized values are shown in Table 3.2. Appendix A3.2 and Appendix A3.3 include the full set of optimized parameters with all DFT functionals and the X—H bonds.

Table 3.2. Selected optimized bond lengths (ångström, Å) for CH_2=CHXXH and CH_2=CHXX^- (X=S,Se).

		CH_2=CHSSH		CH_2=CHSS^-		CH_2=CHSeSeH		CH_2=CHSeSe^-	
Method	**Basis set**	C–S	S–S	C–S	S–S	C–Se	Se-Se	C–Se	Se-Se
B3PW91	6-311+G(d,p)	1.768	2.088	1.745	2.098	1.913	2.347	1.899	2.359
	6-311G(2d,p)	1.767	2.074	1.747	2.084	1.910	2.350	1.898	2.364
	6-311G(df,p)	1.768	2.077	1.748	2.090	1.908	2.334	1.894	2.346
ωB97XD	6-311+G(d,p)	1.771	2.084	1.748	2.101	1.913	2.338	1.898	2.357
	6-311G(2d,p)	1.770	2.072	1.750[a]	2.088[a]	1.910	2.339	1.897	2.361
	6-311G(df,p)	1.770	2.074	1.751	2.096	1.908	2.325	1.895	2.343
M06-2X	6-311+G(d,p)	1.770	2.084	1.747	2.102	1.914	2.337	1.901	2.354
	6-311G(2d,p)	1.770	2.071	1.751	2.090	1.911	2.340	1.900	2.359
	6-311G(df,p)	1.770	2.072	1.750	2.095	1.909	2.326	1.897	2.343
QCISD	6-311+G(2df,p)	1.771	2.069	1.750	2.086	1.914	2.329	1.902	2.359

[a] Had one negligible imaginary frequency.

In general, upon changing basis sets and methods similar trends were observed as for the CH_3XX^-/H systems, though with some key differences. For instance, for X=S, the closest agreement of the optimized bond lengths with the benchmark values was obtained for the 6-311G(df,p) and 6-311G(2d,p) basis sets. In contrast, for X=Se the X—X bond is more sensitive to the choice of basis set. In particular, for $CH_2CHSeSe^-$ the best agreement with

the benchmark values are obtained using the 6-311+G(d,p) or 6-311G(2d,p) basis sets; inclusion of *f*-functions results in too short an Se—Se bond. Meanwhile, for neutral $CH_2CHSeSeH$ it is essential to include *f*-functions in order to obtain good agreement with the corresponding benchmark values.

Comparison of the optimized bond lengths in CH_3XX^-/H and CH_2CHXX^-/H (X = S,Se) shows that for all species, at all levels of theory, the optimized C—X distance in CH_2CHXX^-/H is 0.04 – 0.05 Å shorter than in the corresponding CH_3XX^-/H analogue (*cf.* Table 3.1). This is also observed when structures are optimized at the M06-2X/aug-cc-pVTZ level of theory (data not shown). In contrast, the X—X bond lengths in CH_2CHXX^-/H (X = S,Se) are all within ±0.01 Å of their optimized values in the corresponding CH_3XX^-/H analogue. When X=S the X—X bond in the anion is 0.017 Å longer in the conjugated species compared to 0.023 Å in the alkyl terminated species. It is noted that this again indicates that deprotonation of RXXH (X=S,Se) causes the X—X bond to lengthen, although conjugation lessens the bond lengthening upon going to the anion.

3.3.3 Effects of Mixed Sulfur/Selenium: CH_3XYH (X = S,Se; Y = Se,S)

We then considered mixed sulfur/selenide species for which selected optimized parameters are given in Table 3.3. Full optimized parameters for all DFT methods and basis sets are shown in Table Appendix A3.4.

For all four mixed species the optimized lengths of the C—X bond (X=S,Se) are all within 0.01 Å of those obtained for their analogous persulfide or perselenide (*i.e.* CH_3XXH (X = S,Se)) at the same level of theory (*cf.* Table 3.1). Thus, similar method and basis set trends were also observed and do not require further detailed discussion. However, it should be noted that while the X—Y bond lengths in the neutral species CH_3SSeH and CH_3SeSH are close to each other, the bond is consistently predicted to be marginally shorter (< 0.01 Å) in

CH_3SSeH. The changes observed in the CH_3X—Y bond length upon deprotonation (*i.e.* CH_3XYH to CH_3XY^-), depends on whether S or Se is the terminal atom. When S is the terminal atom, upon deprotonation of the thiol group the Se—S bond lengthens marginally by <0.01 Å. In contrast, when Se is the terminal atom deprotonation causes the S—Se bond to lengthen by ≥0.04 Å. The larger impact of deprotonating a selenol *versus* thiol group is also seen in the species shown in Tables 3.1 and 3.2, though to a smaller extent. We found that when the sulfur atom is in the middle of the chain, it has a larger Mulliken charge compared to selenium (−0.23 *vs*. −0.13) in the anionic species. By comparison, the charge of the terminal chalcogen is the same in both species. Thus, repulsive interactions cause the bond length to increase when sulfur is in the center of the chain.

Table 3.3. Selected optimized bond lengths (ångström, Å) for CH_3XYH and CH_3XY^- (X=S,Se; Y=Se,S).

		CH_3SSeH		**CH_3SSe^-**		**CH_3SeSH**		**CH_3SeS^-**	
Method	**Basis set**	C–S	S–Se	C–S	S–Se	C–Se	Se-S	C–Se	Se-S
B3PW91	6-311+G(d,p)	1.821	2.214	1.820	2.256	1.959	2.218	1.971	2.221
	6-311G(2d,p)	1.819	2.204	1.819	2.244	1.957	2.208	1.968	2.220
	6-311G(df,p)	1.819	2.199	1.819	2.241	1.953	2.204	1.964	2.212
ωB97XD	6-311+G(d,p)	1.820	2.208	1.817	2.252	1.953	2.213	1.962	2.218
	6-311G(2d,p)	1.818	2.199	1.817	2.242	1.951	2.204	1.959	2.216
	6-311G(df,p)	1.818	2.195	1.817	2.239	1.948	2.200	1.957	2.209
M06-2X	6-311+G(d,p)	1.819	2.208	1.817	2.250	1.956	2.212	1.964	2.220
	6-311G(2d,p)	1.817	2.200	1.816	2.239	1.954	2.203	1.960	2.221
	6-311G(df,p)	1.817	2.194	1.816	2.237	1.951	2.198	1.959	2.212
QCISD	6-311+G(2df,p)	1.819	2.198	1.816	2.239	1.952	2.202	1.961	2.207

3.3.4 Extending a Sulfide: CH_3SSSH and $CH_2CHSSSH$.

Polysulfides, but to-date not polyselenides, have been suggested to be potentially biochemically important. Thus for completeness we considered the effect of extending the persulfide chain. Specifically, the simplest alkyl and conjugated trisulfides CH_3SSSH and $CH_2CHSSSH$ were examined. The C—S_1 bond in all species had very similar optimized lengths (see Appendix A3.5), for all methods and basis sets used, to those obtained for the

analogous alkyl persulfides (*cf.* Table 3.1); the observed shortened C—X bond in the conjugated neutral and anionic persulfides (Table 3.2) was not observed in $CH_2CHSSSH$, further evidence that it is due to delocalization across the persulf/selenide and conjugated R-group (see Appendix A). Thus, in Table 3.4 only the optimized lengths of the S—S bonds are given.

As can be seen in Table 3.4, the B3PW91 method gives the worst agreement with the benchmark values, especially for the deprotonated (anionic) species with errors of up to 0.07 Å. Furthermore, for any DFT method, the largest errors in the optimized S_i—S_j bond lengths are observed upon use of the 6-311+G(d,p) basis set. The best agreement is instead obtained using the M06-2X and ωB97XD methods, the former performing slightly better in conjunction with the 6-311G(2d,p) basis set. Indeed, their errors lie in the ranges of 0.000 – 0.011 Å and −0.003 – 0.009 Å, respectively.

Table 3.4. Selected optimized bond lengths (Å) for RSSSH (R=CH_3, CH_2=CH).

		CH_3SSSH		**CH_3SSS^-**		**$CH_2CHSSSH$**		**CH_2CHSSS^-**	
Method	**Basis set**	S_1–S_2	S_2–S_3	S_1–S_2	S_2–S_3	S_1–S_2	S_2–S_3	S_1–S_2	S_2–S_3
B3PW91	6-311+G(d,p)	2.069	2.102	2.147	2.067	2.082	2.094	2.170	2.052
	6-311G(2d,p)	2.057	2.086	2.120	2.058	2.071	2.079	2.136	2.047
	6-311G(df,p)	2.060	2.091	2.130	2.063	2.073	2.084	2.149	2.050
ωB97XD	6-311+G(d,p)	2.068	2.093	2.124	2.073	2.078	2.086	2.134	2.062
	6-311G(2d,p)	2.057	2.080	2.105	2.063	2.070	2.074	2.115	2.054
	6-311G(df,p)	2.059	2.084	2.111	2.070	2.071	2.077	2.123	2.060
M06-2X	6-311+G(d,p)	2.066	2.090	2.121	2.074	2.079	2.083	2.143	2.063
	6-311G(2d,p)	2.056	2.078	2.103	2.065	2.070	2.071	2.118	2.055
	6-311G(df,p)	2.057	2.080	2.107	2.069	2.070	2.073	2.123	2.060
QCISD	6-311+G(2df,p)	2.056	2.078	2.096	2.065	2.066	2.073	2.107	2.057

Notably, in CH_3SSSH and $CH_2CHSSSH$, the CS_1—S_2 bond is predicted to be shorter than the S_2—S_3H bond by >0.02 and <0.01 Å respectively. However, upon deprotonation of the terminal thiol group in each, *i.e.* formation of CH_3SSS^- and CH_2CHSSS^-, the CS_1—S_2 bond

lengthens significantly by ~0.05 Å, from ~2.06 and 2.07 Å in CH_3SSSH and $CH_2CHSSSH$ respectively, to approximately 2.1 and 2.12 Å. In contrast, the S_2—S_3 bond in CH_3SSS^- and CH_2CHSSS^- has shortened by 0.01-0.02 Å. Mulliken charges on S_1 were found to decrease more than they do on S_2 upon deprotonation in the benchmark calculation, for both alkyl and conjugated polysulfur species. However, the difference between the two was small compared to the large charge on S_3; it becomes much more negative upon deprotonation. This indicates there is some degree of charge delocalization along the sulfur chain (Appendix A3.6).

3.3.5 Obtaining Reliable and Accurate Thermochemistry for CH_3SSH and CH_2CHSSH

Two of the most common goals when applying computational methods to the study of chemical problems are obtaining reliable and accurate optimized structures and thermochemical data. For biochemical or related problems, common reactions often require knowledge of proton affinities (PAs), gas phase basicities (GPBs), hydrogen affinities (HAs) and bond dissociation enthalpies (BDEs). Furthermore, given the size of the systems often encountered there is simultaneously considerable interest in identifying a DFT-based approach for calculating such properties. Thus, having assessed the use of DFT methods for the accurate optimization of structures, we also assessed the ability of the DFT methods B3LYP, B3PW91, ωB97XD, M06-2X, M08-HX, in combination with a range of basis sets to provide reliable and accurate biochemically-relevant thermochemical data.

Given the poor performance of B3LYP and minimal improvement of M08-HX over M06-2X in obtaining reliable structures, vide supra, here we only report the performance of B3PW91, ωB97XD, and M06-2X, unless otherwise noted. Furthermore, we have focused on reporting basis sets that for such systems have been previously used (*e.g.* 6-31G(d)), shown herein to be most consistently reliable (*i.e.* 6-311G(2d,p)), or often used for calculating accurate thermochemistry (*i.e.* 6-311+G(2df,p) and 6-311++G(3df,3pd)). The

results obtained are summarized in Table 3.5, although the data for all functionals and basis sets that were studied are shown in Appendix A.

Table 3.5. Homolytic S−S bond dissociation enthalpy (BDE) of RSSH, proton affinity (PA) and gas-phase basicity (GPB) of RSS^-, and hydrogen affinity (HA) of $RSS^\bullet$ (R=CH_3, CH_2CH). All energies calculated at 298.15K and in kJ mol^{-1} (see Computational Methods).

		CH_3SSH	CH_3SS^-		$CH_3SS^\bullet$
Method	**Basis Set**	BDE(S—S)	PA	GPB	HA
B3PW91	6-31G(d)	235.4	1455.2	1424.8	274.6
	6-311G(2d,p)	250.3	1447.7	1417.3	283.8
	6-311+G(2df,p)	254.7	1445.8	1415.4	279.5
	6-311++G(3df,3pd)	255.8	1448.7	1418.4	281.6
ωB97XD	6-31G(d)	244.1	1455.0	1425.1	284.0
	6-311G(2d,p)	258.5	1447.3	1417.5	292.0
	6-311+G(2df,p)	264.6	1445.5	1415.8	288.0
	6-311++G(3df,3pd)	265.6	1449.2	1419.6	289.8
M06-2X	6-31G(d)	252.0	1440.0	1411.4	285.4
	6-311G(2d,p)	264.9	1430.0	1401.6	294.4
	6-311+G(2df,p)	270.3	1428.5	1400.1	291.0
	6-311++G(3df,3pd)	272.3	1432.3	1404.2	293.9
QCISD	6-311+G(2df,p)	**236.3**	**1445.4**	**1415.1**	**282.7**
		CH_2CHSSH	**CH_2CHSS^-**		**$CH_2CHSS^\bullet$**
		BDE(S—S)	PA	GPB	HA
B3PW91	6-31G(d)	238.0	1427.6	1396.5	272.8
	6-311G(2d,p)	253.0	1423.1	1391.5	282.1
	6-311+G(2df,p)	255.7	1417.7	1386.8	278.4
	6-311++G(3df,3pd)	256.8	1420.5	1389.7	280.1
ωB97XD	6-31G(d)	244.1	1429.8	1397.3	283.7
	6-311G(2d,p)	258.5	1422.4[a]	1395.9[a]	291.5
	6-311+G(2df,p)	262.7	1420.2	1382.0	288.3
	6-311++G(3df,3pd)	263.5	1423.6	1388.7	289.7
M06-2X	6-31G(d)	249.8	1414.5	1381.1	283.7
	6-311G(2d,p)	263.0	1408.2	1374.4	292.8
	6-311+G(2df,p)	267.4	1403.4	1366.4	290.2
	6-311++G(3df,3pd)	269.1	1406.7	1368.1	292.6
QCISD	6-311+G(2df,p)	**242.0**	**1421.8**	**1390.7**	**283.5**

[a] One of the required species for determining this value had one negligible imaginary frequency.

The benchmark values were again obtained using the QCISD/6-311+G(2df,p) level of theory. The thermochemical values obtained for the conjugated persulfides compared to the alkyl persulfides reflects the trends observed structurally. For instance, the BDE(S–S) for CH_2CHSSH is slightly higher by 5.7 kJ mol^{-1} than that of CH_3SSH, while the PA and GPB of CH_2CHSS^- are 23.6 and 24.4 kJ mol^{-1} lower, respectively, than those of CH_3SS^-. This is due to delocalization across the persulfide and CH_2CH group in $CH_2CHSSH/^-$. Meanwhile, the hydrogen affinity of $CH_2CHSS^{\bullet}$ is predicted to be only marginally higher than that of $CH_3SS^{\bullet}$ by 0.8 kJ mol^{-1}. Notably, we did see spin contamination in the QCISD calculations with a spin of 0.88. This may be the reason for lower energies calculated by QCISD compared to DFT methods where spin contamination was much lower (0.77).

A DFT-based model was determined to be accurate if it gave values within 10 kJ mol^{-1}, generally held to be experimental accuracy, of the benchmark values. From Table 3.5 it can be seen that not all methods or basis set choices were reliable, nor did all thermochemical properties exhibit the same method/basis set requirement. For instance, for accurate determination of the BDE(S—S) of both CH_3SSH and CH_2CHSSH, the best performance was observed for B3PW91 and ωB97XD in combination with the 6-31G(d) basis set. The former method slightly underestimating compared to the benchmark value while the latter overestimated slightly. The M06-2X method only gave an accurate BDE(S—S) for CH_2CHSSH and again when using the 6-31G(d) basis set. In general, the use of basis sets larger than 6-31G(d) gave BDE(S—S) values that are markedly higher (14 – 36 kJ mol^{-1}) than those obtained at the QCISD/6-311+G(2df,p) level of theory.

For the PA and GPB of the CH_3SS^- and CH_2CHSS^- anions the M06-2X method again only gives accurate values when used in combination with the 6-31G(d) basis set. In contrast, the B3PW91 and ωB97XD give accurate values for all the basis sets considered. However, the triple-zeta basis sets (6-311G(2d,p), 6-311+G(2df,p), and 6-311++G(3df,3pd)) gave best

agreement with calculated values within 5 kJ mol^{-1} of their corresponding benchmark value (see Table 3.5).

The calculated values of the hydrogen affinity of $CH_3SS^{\bullet}$ and $CH_2CHSS^{\bullet}$ follow almost the same method and basis set trends and accuracy as that observed for the PA and GPB of CH_3SS^- and CH_2CHSS^-. Namely, the B3PW91 and ωB97XD methods in conjunction with any of the basis sets considered give calculated values within ±10 kJ mol^{-1}. The only exception occurs when at the B3PW91/6-31G(d) level of theory for $CH_2CHSS^{\bullet}$, which gives a HA value 10.7 kJ mol^{-1} lower than the corresponding QCISD/6-311+G(2df,p) benchmark value (see Table 3.5). Meanwhile, the M06-2X method is inconsistent; for $CH_3SS^{\bullet}$ only the 6-31G(d) and 6-311+G(2df,p) basis sets give values within 10 kJ mol^{-1} of the benchmark values while for $CH_2CHSS^{\bullet}$ it gives good agreement for all basis sets considered herein.

Several overall trends are suggested in this examination of the performance of the DFT methods B3PW91, ωB97XD, and M06-2X, in conjunction with a range of basis sets, for the noted important thermochemical properties. In particular, the M06-2X functional is the least consistent and usually gives values that differ from the benchmark values by more than 10 kJ mol^{-1}. Furthermore, for all DFT functionals the values obtained using the 6-311+G(2df,p) basis set are within 4 kJ mol^{-1} of the corresponding values obtained using the much larger and more expensive 6-311++G(3df,3pd) basis set. Hence, for the remainder of this report, for simplicity, only thermochemical values obtained using the B3PW91 and ωB97XD functionals in combination with basis sets no larger than 6-311+G(2df,p) are discussed, unless otherwise noted.

3.3.6 Thermochemistry of Selenium-Containing Species

As noted above in the structural assessment, for selenium containing species considered herein; that is $CH_3SeSe^{\bullet/-}/H$, $CH_2CHSeSe^{\bullet/-}/H$, $CH_3SSe^{\bullet/-}/H$, and $CH_3SeS^{\bullet/-}/H$, the smallest consistently reliable basis set was 6-311+G(d,p), though with exceptions as noted. Hence, for these species we have limited our discussion herein to results obtained using B3PW91 and ωB97XD in conjunction with the 6-311+G(d,p) and 6-311+G(2df,p) basis sets. The results are shown in Table 3.6.

As for the analogous purely sulfur-containing species (*cf.* Table 3.5), the calculated PA's and GPB's of all species obtained using B3PW91 or ωB97XD with either basis set choice gives values within 10 kJ mol^{-1} of the corresponding benchmark values. A similar consistency is observed for the calculated HA's, though with some exceptions. In particular, at the ωB97XD/6-311+G(d,p) level of theory the calculated HA's of $CH_2CHSeSe^{\bullet}$ and $CH_3SSe^{\bullet}$ are 11.2 and 12.4 kJ mol^{-1}, respectively, higher than their corresponding QCISD/6-311+G(2df,p) values. Again, as seen in Table 3.5, the calculated RX—YH BDE's of all species are generally markedly overestimated by 12.7 – 26.3 kJ mol^{-1} using either DFT method and basis set. Only three values fall within the desired 10 kJ mol^{-1} error margin and all occur for the mixed chalcogen species: all using the 6-311+G(d,p) basis set.

Table 3.6. Calculated homolytic Se−Se bond dissociation enthalpies (BDE) of RSeSeH, proton affinity (PA) and gas-phase basicities (GPB) of $RSeSe^-$, and Se−H homolytic bond dissociation enthalpies (HA) of $RSeSe^\bullet$ (R=CH_3, CH_2CH). All energies in kJ mol^{-1}.

		CH_3SeSeH	**CH_3SeSe^-**		**$CH_3SeSe^\bullet$**
Method	**Basis Set**	BDE(Se—Se)	PA	GPB	HA
B3PW91	6-311+G(d,p)	219.2	1409.0	1378.7	270.5
	6-311+G(2df,p)	222.9	1411.4	1381.1	266.0
ωB97XD	6-311+G(d,p)	218.0	1410.8	1380.7	276.5
	6-311+G(2df,p)	222.6	1413.3	1383.4	270.7
QCISD	6-311+G(2df,p)	**204.1**	**1404.3**	**1374.0**	**267.3**
		$CH_2CHSeSeH$	**$CH_2CHSeSe^-$**		**$CH_2CHSeSe^\bullet$**
		BDE(Se—Se)	PA	GPB	HA
B3PW91	6-311+G(d,p)	222.1	1388.6	1357.4	270.2
	6-311+G(2df,p)	225.3	1390.7	1359.7	265.4
ωB97XD	6-311+G(d,p)	221.3	1391.4	1359.6	277.1
	6-311+G(2df,p)	225.3	1393.8	1361.1	271.0
QCISD	6-311+G(2df,p)	**208.3**	**1386.7**	**1355.7**	**265.9**
		CH_3SSeH	**CH_3SSe^-**		**$CH_3SSe^\bullet$**
		BDE(S—Se)	PA	GPB	HA
B3PW91	6-311+G(d,p)	221.7	1413.2	1382.8	273.8
	6-311+G(2df,p)	233.3	1415.7	1385.5	265.3
ωB97XD	6-311+G(d,p)	219.6	1414.0	1384.2	279.9
	6-311+G(2df,p)	232.5	1417.2	1387.5	270.9
QCISD	6-311+G(2df,p)	**218.7**	**1411.3**	**1381.1**	**267.5**
		CH_3SeSH	**CH_3SeS^-**		**$CH_3SeS^\bullet$**
		BDE(Se—S)	PA	GPB	HA
B3PW91	6-311+G(d,p)	229.4	1435.0	1404.8	287.1
	6-311+G(2df,p)	241.6	1441.7	1411.4	283.6
ωB97XD	6-311+G(d,p)	231.4	1435.6	1405.8	295.6
	6-311+G(2df,p)	245.0	1442.2	1412.4	290.8
QCISD	6-311+G(2df,p)	**224.5**	**1438.9**	**1408.6**	**286.1**

As for the analogous purely sulfur-containing species (*cf.* Table 3.5), the calculated PA's and GPB's of all species obtained using B3PW91 or ωB97XD with either basis set choice gives values within 10 kJ mol^{-1} of the corresponding benchmark values. A similar

consistency is observed for the calculated HA's, though with some exceptions. In particular, at the ωB97XD/6-311+G(d,p) level of theory the calculated HA's of $CH_2CHSeSe^{\bullet}$ and $CH_3SSe^{\bullet}$ are 11.2 and 12.4 kJ mol^{-1}, respectively, higher than their corresponding QCISD/6-311+G(2df,p) values. Again, as seen in Table 3.5, the calculated RX—YH BDE's of all species are generally markedly overestimated by 12.7 – 26.3 kJ mol^{-1} using either DFT method and basis set. Only three values fall within the desired 10 kJ mol^{-1} error margin and all occur for the mixed chalcogen species: all using the 6-311+G(d,p) basis set.

The thermochemical values provided in Tables 3.5 and 3.6 show several key differences between persulfides and perselenides and the mixed chalcogens. These are most clearly and simply illustrated by examination of the calculated benchmark values. In particular, increasing the number of Se atoms in an RX—YH bond reduces its BDE as shown by comparing those of CH_3SSH (236.3 kJ mol^{-1}), CH_3SSeH (218.7 kJ mol^{-1}), CH_3SeSH (223.2 kJ mol^{-1}), and CH_3SeSeH (202.7 kJ mol^{-1}). In addition, a conjugated group adjacent to the RX—XH group increases its BDE slightly by 3-6 kJ mol^{-1}. Meanwhile, the HA's are reasonably consistent and depend primarily on whether the formal radical terminal is a Sulfur or Selenium. For the former, all values lie within the range 282.7 kJ mol^{-1} ($CH_3SS^{\bullet}$) to 286.1 kJ mol^{-1} ($CH_3SeS^{\bullet}$), while the latter are notably lower, between 265.9 kJ mol^{-1} ($CH_2CHSeSe^{\bullet}$) and 267.3 kJ mol^{-1} ($CH_3SeSe^{\bullet}$ and $CH_3SSe^{\bullet}$).

Comparison of the PA's and GPB's of these perselenides and mixed per-sulf/selenides with their analogous persulfides (*cf.* Table 3.5), shows several interesting trends. It is noted that for simplicity, as the observed trends were the same for all the DFT methods, the values discussed here refer to those obtained at the benchmark level of theory. Firstly, systematically increasing the occurrence of Se in a per-chalcogenide group decreases their PA's and GPB values. For instance, the PA's of CH_3SS^{-}, CH_3SeS^{-}, CH_3SSe^{-}, and CH_3SeSe^{-} are 1445.4 kJ mol^{-1}, 1438.9 kJ mol^{-1}, 1411.3 kJ mol^{-1}, and 1404.3 kJ mol^{-1}. Simultaneously, their GPB values decrease from 1415.1 kJ mol^{-1}, to 1408.6 kJ mol^{-1} and

1381.1 kJ mol^{-1}, to 1374.0 kJ mol^{-1}. In addition, replacing CH_3- with CH_2CH- decreases the PA and GPB values of the pure persulfides and perselenides by 24-25 and 17-18 kJ mol^{-1}, respectively. This perhaps reflects a larger delocalization when a conjugated group is adjacent *and* that the effect is less for the selenides.

Extending the persulfides to trisulfides. We also examined the effect of extending the persulfide group by an additional sulfur to a trisulfide; specifically, we considered key bio-relevant thermochemistry of $CH_3SSS^{\bullet/-}$/H and $CH_2CHSSS^{\bullet/-}$/H. Based on the trends observed for persulfide systems, thermochemical values were obtained using the B3PW91 and ωB97XD DFT methods in combination with the 6-311+G(2df,p) basis set, and again at the QCISD/6-311+G(2df,p) benchmark level of theory (all other data shown in Appendix A3.10). As can be seen in Table 3.7, both DFT methods give thermochemical results in generally good agreement with the benchmark values with B3PW91 slightly preferred, although the differences in average errors are relatively small.

Comparison of the calculated benchmark values of the trisulfides with those of the corresponding persulfides (*cf.* Table 3.6) shows that the BDE of RS_1S_2—S_3H (*i.e.* the BDE of the terminal S_2—S_3 bond), where R = CH_3- and CH_2CH-, decreases significantly by 53.3 and 56.3 kJ mol^{-1}, respectively. Similarly, their calculated PAs and GPBs decrease markedly by 31-33 and 18-19 kJ mol^{-1} for R = CH_3- and CH_2CH-, respectively. As a result of these changes the RSS—SH BDE, and PAs and GPBs of $RSSS^-$ all lie within a narrower range (≤13 kJ mol^{-1}) of each other. This perhaps reflects in part a decrease in the influence of the R group on the increasingly removed S—SH bond. In contrast, the hydrogen affinities of $CH_3SSS^{\bullet}$ and $CH_2CHSSS^{\bullet}$ are 6.7 and 11.5 kJ mol^{-1} higher than that of their corresponding persulfide analogues with values of 289.4 and 295.0 kJ mol^{-1}, respectively.

Table 3.7. Homolytic RS−SH bond dissociation enthalpy (BDE) of RSSSH, proton affinity (PA) and gas-phase basicity (GPB) of $RSSS^-$, and hydrogen affinity (HA) of $RSSS^\bullet$ ($R=CH_3$, CH_2CH). All energies calculated at 298.15K and in kJ mol^{-1}.

		CH_3SSSH	**CH_3SSS^-**		**$CH_3SSS^\bullet$**
Method	**Basis Set**	BDE(RSS—SH)	PA	GPB	HA
B3PW91	6-311+G(2df,p)	193.8	1413.9	1382.4	287.9
ωB97XD	6-311+G(2df,p)	200.4	1415.0	1385.1	297.7
QCISD	6-311+G(2df,p)	**183.0**	**1413.9**	**1384.1**	**289.4**
		$CH_2CHSSSH$	**CH_2CHSSS^-**		**$CH_2CHSSS^\bullet$**
		BDE(RSS—SH)	PA	GPB	HA
B3PW91	6-311+G(2df,p)	192.8	1399.5	1367.0	290.5
ωB97XD	6-311+G(2df,p)	201.0	1403.0	1370.9	302.0
QCISD	6-311+G(2df,p)	**185.7**	**1403.7**	**1371.2**	**295.0**

3.3.7 Cysteine-derived polysulfides: $_{Cys}SS_nH$ (n=1-3)

Within biological systems, as noted above, per- and polysulfide derivatives of cysteine play important roles. Hence, we also considered the structures and thermochemistry of such species using the ωB97XD/6-311G(2d,p) level of theory. This level was selected as it emerged in our earlier study as the most able to provide reliable structures and thermochemistry for related model systems (see above). Selected parameters of optimized structures obtained for $_{Cys}SS_nH$ (n=0-3) and $_{Cys}SS_n^-$ (n=0-3) are provided in Table 3.8 and representative optimized structures are shown in Figure 3.1, at this level of theory.

Table 3.8. Selected optimized bond lengths (ångström, Å) obtained at the ωB97XD/6-311G(2d,p) level of theory for $_{Cys}SS_nH$ (*n*=0-3) and $_{Cys}SS_n^-$ (*n*=0-3).

Molecule	C–S_1	S_1–S_2	S_2–S_3	S_3–S_4
$_{Cys}$**SH**	1.827			
$_{Cys}$**S–SH**	1.822	2.073		
$_{Cys}$**S–S–SH**	1.825	2.056	2.084	
$_{Cys}$**S–S–S–SH**	1.827	2.060	2.080	2.084
$_{Cys}$**S$^-$**	1.832			
$_{Cys}$**S–S$^-$**	1.820	2.098		
$_{Cys}$**S–S–S$^-$**	1.829	2.091	2.064	
$_{Cys}$**S–S–S–S$^-$**	1.833	2.088	2.080	2.061

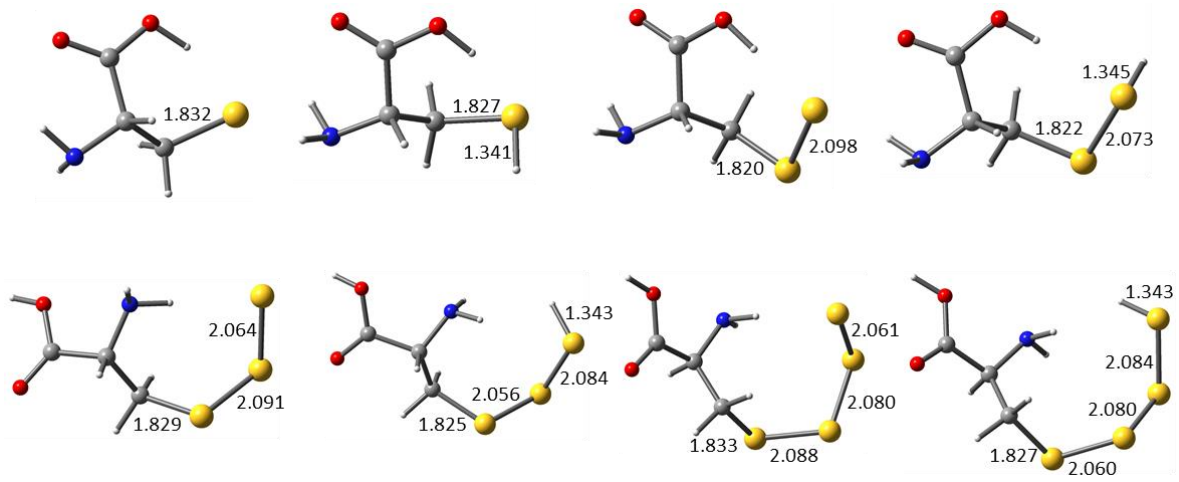


Figure 3.1. Optimized structures of cysteine, cysteine persulfide and cysteine polysulfides with the selected bond lengths in ångström (Å) using the ωB97XD/6-311G(2d,p) level of theory.

As can be seen, for both the neutral and deprotonated derivatives, the *r*(C−S) bond length decreases slightly upon forming a perthiol/sulfide group, but then gradually lengthens, essentially returning to its length in Cysteine, as one goes to the corresponding trithiol/sulfide and tetrathiol/sulfide derivatives. It is also noted that for both $_{Cys}$SSH and

$_{Cys}SS^-$, the C—S and S—S bond lengths are in close agreement with the corresponding values obtained at the same level of theory for CH_3SSH (1.816 and 2.067 Å) and CH_3SS^- (1.816 and 2.092 Å), respectively (*cf.* Table 3.1). In addition, for the neutral polysulfides the S—S bond in the chain that is farthest removed from the alkyl group has the longest length, with the S—S bonds being sequentially shorter the closer they are to the alkyl group. In contrast, for the corresponding anionic deprotonated series ($_{Cys}SS_n^-$, n=0-3), the opposite trend is observed; the S—S bond in the chain farthest from the alkyl group is shortest, and they get longer the closer they are to the alkyl group. In the latter species this trend may reflect a diminishing effect of the negative charge on the terminal sulfur atom the further removed the bond.

Thermochemistry of Cysteine-derived polysulfides. As for the other species considered above, we calculated the RS_n—SH homolytic BDE's for $_{Cys}SS_nH$ (n=1-3), the PA's and GPB's of the $_{Cys}SS_n^-$ (n=0-3) series of derivatives, and the hydrogen affinities (HA's) of the $_{Cys}SS_n^{\bullet}$ (n=0-3) series of species. However, all values were again only obtained using the chosen ωB97XD/6-311G(2d,p) level of theory. The results obtained are shown in Figure 3.2A, B, and C respectively.

The calculated RS_n—SH (n=1-3) homolytic bond dissociation enthalpies, that is, the BDE of the terminal S—SH bond in the perthiol chain decreases by 33.7 kJ mol^{-1} from 248.1 to 214.4, upon going from the perthiol (n=1) to trithiol (n=2). It then decreases a further 15.4 kJ mol^{-1} to 199 kJ mol^{-1} upon increasing to the chain further to the tetrathiol (n=3); Figure 3.2A. This trend suggests that while the strength of the terminal S—SH bond does weaken as the chain is lengthened, it approaches a limiting value below 199.0 kJ mol^{-1}, although possibly not too much lower than that. In contrast, as seen in Figure 3.2C, the hydrogen affinities of the $_{Cys}SS_n^{\bullet}$ (n=0-3) species decrease significantly by 67.8 kJ mol^{-1} upon going from a cysteinyl thiyl radical (n=0; 356.5 kJ mol^{-1}) to the perthiyl radical (n=1; 288.7 kJ

mol^{-1}). However, extending the chain further to *n*=2 or *n*=3 results in a slight increase in the HA's to 298.7 kJ mol^{-1} and 295.6 kJ mol^{-1}, respectively. This suggests that at least for HA's of the polysulfur radicals, they are reasonably constant for the perthiyl and beyond.

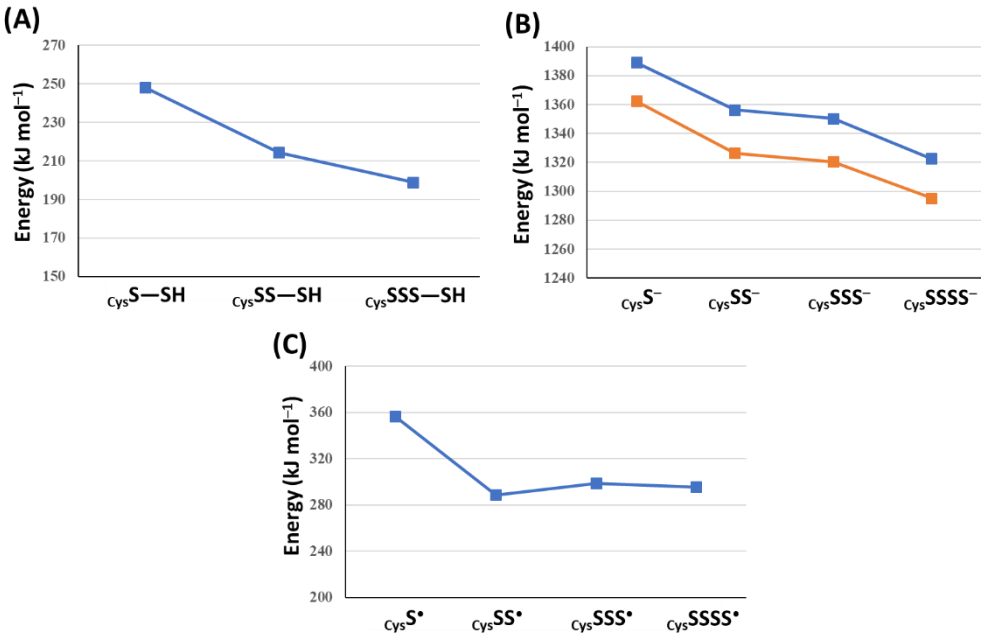


Figure 3.2. Calculated (at the ωB97XD/6-311G(2d,p) level of theory) values of (A) RS$_n$—SH BDE's for $_{Cys}SS_nH$ (*n*=1-3), (B) PA's (blue) and GPB's (orange) of $_{Cys}SS_n^-$ (*n*=0-3), and (C) HA's of $_{Cys}SS_n^{\bullet}$ (*n*=0-3). All energies are in kJ mol^{-1}.

Figure 3.2B shows that as the polysulfide chain is extended from the cysteine thiolate to $_{Cys}SSSS^-$ both the calculated PA's and GPB's decrease. However, while there are significant decreases of 33.0 (PA) and 35.9 (GPB) kJ mol^{-1} going from the thiolate ($_{Cys}S^-$) to perthiolate ($_{Cys}SS^-$), this does not appear to continue upon extending the chain further. Indeed, extending the chain to $_{Cys}SSS^-$ results in only comparatively small further decreases of 5.8 and 5.9 kJ mol^{-1} to 1350.5 and 1320.6 kJ mol^{-1}. But, upon extending the chain to $_{Cys}SSSS^-$, larger decreases in both the PA and GPB are again observed; they decrease by 27.9 and 25.4 kJ mol^{-1} to 1322.6 and 1295.2 kJ mol^{-1}. This may in part reflect that in the

cysteinyl polysulfide derivatives, weak intramolecular hydrogen bonding was observed in some optimized structures between the terminal S^- centre and the cysteine's amino group. This interaction would also help decrease PA and GPB values. It should also be noted that comparison of the calculated PA's and GPB's of the cysteine-derived sulfides $_{Cys}SS^-$ and $_{Cys}SSS^-$ with that of their smaller corresponding analogs CH_3SS^- (*cf.* Table 3.5) and CH_3SSS^- (*cf.* Table 3.7) shows that that those of the former two species are markedly lower by 60-90 kJ mol^{-1}. Thus, overall, these results suggest that in biological systems, extending the polysulfide chain will increase its likelihood of being deprotonated and that hydrogen bonding may help to stabilize such anions.

3.4 Conclusions

The reliability and accuracy of several commonly used DFT functionals (*e.g.* B3LYP, B3PW91, ωB97XD, M06-2X, M08-HX) as well as MP2 was assessed for a systematic series of bio-relevant polysulfur/selenium-containing systems. In particular, optimized structures and thermodynamic properties of a range of $RX_n(H)$ (X=S, Se, R=CH_3, CH_2=CH, and cysteine, *n*=1-4) were examined with a variety of Pople basis sets of increasing size. We offer the following conclusions from this detailed study:

1. Evaluation of the bond lengths in $CH_3XXH/^-$ (X=S,Se, *cf.* Table 3.1) showed that the S—S bond is the most sensitive to changes in basis set. The smallest basis set used (6-31G(d)) frequently resulted in optimized geometries that were very similar to the benchmark given by QCISD/6-311+G(2df,p), although the preferred basis sets were 6-311G(2d,p) and 6-311G(df,p) due to the higher sensitivity of sulfur to basis set choice.
2. We also saw that M08-HX did not offer much improvement over M06-2X and in some cases was detrimental. The best functionals for geometry optimization were found to be B3PW91, ωB97XD and M06-2X.

3. In the conjugated system CH_2CHXXH, the C—X bond length decreases due to more delocalized electron density.
4. In mixed chalcogen species, the location of the chalcogen atom has a significant effect on the X—Y bond length. The RSe—S^- bond is slightly longer than the RS—Se^- bond since sulfur takes on more negative charge. Deprotonation of polysulfide species increases the C—S bond as well, for the same reason: charge delocalization along a sulfur chain.
5. In the polysulfide species PA, GPB, BDE, and HA all decrease with the increasing sulfur chain. The lower PA and GPB values indicate that in a biological system, polysulfide chains are more likely to exist as deprotonated rather than neutral species compared to the parent thiol or selenol.
6. To perform reliable thermochemical calculations, the ωB97XD/6-311G(2d,p) level of theory was found to give the most accurate results relative to the benchmark and this was then used to evaluate the geometry and thermochemistry of the more complex cysteine and cysteine per/polysulfide species.

3.5 References

1. Zhong, L. W.; Holmgren, A., Essential role of selenium in the catalytic activities of mammalian thioredoxin reductase revealed by characterization of recombinant enzymes with selenocysteine mutations. *J. Biol. Chem.* **2000,** *275*, 18121-18128.

2. Fernandez, F. J.; Arda, A.; Lopez-Estepa, M.; Aranda, J.; Pena-Soler, E.; Garces, F.; Round, A.; Campos-Olivas, R.; Bruix, M.; Coll, M.; Tunon, I.; Jimenez-Barbero, J.; Vega, M. C., Mechanism of Sulfur Transfer Across Protein-Protein Interfaces: The Cysteine Desulfurase Model System. *ACS Catal.* **2016,** *6*, 3975-3984.

3. Kahya, M. C.; Naziroglu, M.; Ovey, I. S., Modulation of Diabetes-Induced Oxidative Stress, Apoptosis, and Ca(2+) Entry Through TRPM2 and TRPV1 Channels in

Dorsal Root Ganglion and Hippocampus of Diabetic Rats by Melatonin and Selenium. *Mol. Neurobiol.* **2017,** *54*, 2345-2360.

4. Paulsen, C. E.; Carroll, K. S., Cysteine-mediated redox signaling: chemistry, biology, and tools for discovery. *Chem. Rev.* **2013,** *113*, 4633-79.

5. Garcia-Santamarina, S.; Boronat, S.; Hidalgo, E., Reversible Cysteine Oxidation in Hydrogen Peroxide Sensing and Signal Transduction. *Biochemistry* **2014,** *53*, 2560-2580.

6. Huang, Z.; Rose, A. H.; Hoffmann, P. R., The role of selenium in inflammation and immunity: from molecular mechanisms to therapeutic opportunities. *Antioxid. Redox. Signal.* **2012,** *16*, 705-43.

7. Ida, T.; Sawa, T.; Ihara, H.; Tsuchiya, Y.; Watanabe, Y.; Kumagai, Y.; Suematsu, M.; Motohashi, H.; Fujii, S.; Matsunaga, T.; Yamamoto, M.; Ono, K.; Devarie-Baez, N. O.; Xian, M.; Fukuto, J. M.; Akaike, T., Reactive cysteine persulfides and S-polythiolation regulate oxidative stress and redox signaling. *Proc. Natl. Acad. of Sci. USA* **2014,** *111*, 7606-7611.

8. Kim, H. J.; Ha, S.; Lee, H. Y.; Lee, K. J., Rosics: Chemistry and proteomics of cysteine modifications in redox biology. *Mass Spec. Rev.* **2015,** *34*, 184-208.

9. Jacobs, E. T.; Jiang, R.; Alberts, D. S.; Greenberg, E. R.; Gunter, E. W.; Karagas, M. R.; Lanza, E.; Ratnasinghe, L.; Reid, M. E.; Schatzkin, A.; Smith-Warner, S. A.; Wallace, K.; Martinez, M. E., Selenium and colorectal adenoma: results of a pooled analysis. *J. Natl. Cancer Inst.* **2004,** *96*, 1669-75.

10. Wu, G.; Fang, Y. Z.; Yang, S.; Lupton, J. R.; Turner, N. D., Glutathione metabolism and its implications for health. *J. Nutr.* **2004,** *134*, 489-92.

11. Giles, G. I.; Tasker, K. M.; Jacob, C., Hypothesis: The role of reactive sulfur species in oxidative stress. *Free Radical Biology and Medicine* **2001,** *31* (10), 1279-1283.

12. Toohey, J. I., Sulfur signaling: Is the agent sulfide or sulfane? *Anal. Biochem.* **2011,** *415*, 221-221.

13. Kimura, H., Physiological role of hydrogen sulfide and polysulfide in the central nervous system. *Neurochem. Int.* **2013,** *63*, 492-497.

14. Kaneko, Y.; Kimura, Y.; Kimura, H.; Niki, I., L-cysteine inhibits insulin release from the pancreatic beta-cell - Possible involvement of metabolic production of hydrogen sulfide, a novel gasotransmitter. *Diabetes* **2006,** *55*, 1391-1397.

15. Lo Faro, M. L.; Fox, B.; Whatmore, J. L.; Winyard, P. G.; Whiteman, M., Hydrogen sulfide and nitric oxide interactions in inflammation. *Nitric Oxide* **2014,** *41*, 38-47.

16. Yu, B.; Zheng, Y.; Yuan, Z.; Li, S.; Zhu, H.; De La Cruz, L. K.; Zhang, J.; Ji, K.; Wang, S.; Wang, B., Toward Direct Protein S-Persulfidation: A Prodrug Approach That Directly Delivers Hydrogen Persulfide. *J. Am. Chem. Soc.* **2018,** *140*, 30-33.

17. Giles, G. I.; Nasim, M. J.; Ali, W.; Jacob, C., The Reactive Sulfur Species Concept: 15 Years On. *Antioxidants* **2017,** *6*.

18. Shen, X. G.; Peter, E. A.; Bir, S.; Wang, R.; Kevil, C. G., Analytical measurement of discrete hydrogen sulfide pools in biological specimens. *Free Radic. Biol. Med.* **2012,** *52*, 2276-2283.

19. Iciek, M.; Wlodek, L., Biosynthesis and biological properties of compounds containing highly reactive, reduced sulfane sulfur. *Pol. J. Pharmacol.* **2001,** *53*, 215-225.

20. Mishanina, A. V.; Libiad, M.; Banerjee, R., Biogenesis of reactive sulfur species for signaling by hydrogen sulfide oxidation pathways. *Nat. Chem. Biol.* **2015,** *11*, 457-464.

21. Akaike, T.; Ida, T.; Wei, F. Y.; Nishida, M.; Kumagai, Y.; Alam, M. M.; Ihara, H.; Sawa, T.; Matsunaga, T.; Kasamatsu, S.; Nishimura, A.; Morita, M.; Tomizawa, K.; Nishimura, A.; Watanabe, S.; Inaba, K.; Shima, H.; Tanuma, N.; Jung, M.; Fujii, S.; Watanabe, Y.; Ohmuraya, M.; Nagy, P.; Feelisch, M.; Fukuto, J. M.; Motohashi, H., Cysteinyl-tRNA synthetase governs cysteine polysulfidation and mitochondrial bioenergetics. *Nat. Commun.* **2017,** *8*, 1177.

22. Koike, S.; Nishimoto, S.; Ogasawara, Y., Cysteine persulfides and polysulfides produced by exchange reactions with H2S protect SH-SY5Y cells from methylglyoxal-induced toxicity through Nrf2 activation. *Redox Biol.* **2017,** *12*, 530-539.

23. Munchberg, U.; Anwar, A.; Mecklenburg, S.; Jacob, C., Polysulfides as biologically active ingredients of garlic. *Org. Biomol. Chem.* **2007,** *5*, 1505-1518.

24. Gruhlke, M. C. H.; Nicco, C.; Batteux, F.; Slusarenko, A. J., The Effects of Allicin, a Reactive Sulfur Species from Garlic, on a Selection of Mammalian Cell Lines. *Antioxidants* **2017,** *6*.

25. Millikin, R.; Bianco, C. L.; White, C.; Saund, S. S.; Henriquez, S.; Sosa, V.; Akaike, T.; Kumagai, Y.; Soeda, S.; Toscano, J. P.; Lin, J.; Fukuto, J. M., The chemical biology of protein hydropersulfides: Studies of a possible protective function of biological hydropersulfide generation. *Free Radic. Biol. Med.* **2016,** *97*, 136-147.

26. Cerqueira, N. M.; Fernandes, P. A.; Gonzalez, P. J.; Moura, J. J.; Ramos, M. J., The sulfur shift: an activation mechanism for periplasmic nitrate reductase and formate dehydrogenase. *Inorg. Chem.* **2013,** *52*, 10766-72.

27. Cheng, Q.; Sandalova, T.; Lindqvist, Y.; Arner, E. S., Crystal structure and catalysis of the selenoprotein thioredoxin reductase 1. *J. Biol. Chem.* **2009,** *284*, 3998-4008.

28. Searle, P. A.; Molinski, T. F., 5 New Alkaloids from the Tropical Ascidian, Lissoclinum Sp - Lissoclinotoxin-a Is Chiral. *J. Org. Chem.* **1994,** *59*, 6600-6605.

29. Brzostowska, E. M.; Greer, A., The role of amine in the mechanism of pentathiepin (polysulfur) antitumor agents. *J. Am. Chem. Soc.* **2003,** *125*, 396-404.

30. Chauvin, J. P. R.; Griesser, M.; Pratt, D. A., Hydropersulfides: H-Atom Transfer Agents Par Excellence. *J. Am. Chem. Soc.* **2017,** *139*, 6484-6493.

31. Bortoli, M.; Torsello, M.; Bickelhaupt, F. M.; Orian, L., Role of the Chalcogen (S, Se, Te) in the Oxidation Mechanism of the Glutathione Peroxidase Active Site. *Chemphyschem* **2017,** *18*, 2990-2998.

32. Bachrach, S. M.; Demoin, D. W.; Luk, M.; Miller, J. V., Nucleophilic attack at selenium in diselenides and selenosulfides. A computational study. *J. Phys. Chem. A* **2004,** *108*, 4040-4046.

33. Huang, G. T.; Yu, J. S. K., Enzyme Catalysis that Paves the Way for S-Sulfhydration via Sulfur Atom Transfer. *J. Phys. Chem. B* **2016,** *120*, 4608-4615.

34. Koch, W.; Holthausen, M. C., *A chemist's guide to density functional theory*. 2nd ed.; Wiley-VCH: Weinheim ; New York, 2001; p xiii, 300 p.

35. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, ; M. A. Robb, J. R. C., G. Scalmani, V. Barone, B. Mennucci, ; G. A. Petersson, H. N., M. Caricato, X. Li, H. P. Hratchian, ; A. F. Izmaylov, J. B., G. Zheng, J. L. Sonnenberg, M. Hada, ; M. Ehara, K. T., R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, ; Y. Honda, O. K., H. Nakai, T. Vreven, J. A. Montgomery, Jr., ; J. E. Peralta, F. O., M. Bearpark, J. J. Heyd, E. Brothers, ; K. N. Kudin, V. N. S., T. Keith, R. Kobayashi, J. Normand, ; K. Raghavachari, A. R., J. C. Burant, S. S. Iyengar, J. Tomasi, ; M. Cossi, N. R., J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, ; V. Bakken, C. A., J. Jaramillo, R. Gomperts, R. E. Stratmann, ; O. Yazyev, A. J. A., R. Cammi, C. Pomelli, J. W. Ochterski, ; R. L. Martin, K. M., V. G. Zakrzewski, G. A. Voth, ; P. Salvador, J. J. D., S. Dapprich, A. D. Daniels, ; O. Farkas, J. B. F., J. V. Ortiz, J. Cioslowski, ; Fox, a. D. J. *Gaussian 09*, E.01; Gaussian, Inc.: Wallingford CT, 2013.

36. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, ; M. A. Robb, J. R. C., G. Scalmani, V. Barone, ; G. A. Petersson, H. N., X. Li, M. Caricato, A. V. Marenich, ; J. Bloino, B. G. J., R. Gomperts, B. Mennucci, H. P. Hratchian, ; J. V. Ortiz, A. F. I., J. L. Sonnenberg, D. Williams-Young, ; F. Ding, F. L., F. Egidi, J. Goings, B. Peng, A. Petrone, ; T. Henderson, D. R., V. G. Zakrzewski, J. Gao, N. Rega, ; G. Zheng, W. L., M. Hada, M. Ehara, K. Toyota, R. Fukuda, ; J. Hasegawa, M. I., T. Nakajima, Y. Honda, O. Kitao, H. Nakai, ; T. Vreven, K. T., J. A. Montgomery, Jr., J. E. Peralta, ; F. Ogliaro, M. J. B., J. J. Heyd, E. N. Brothers, K. N. Kudin, ; V. N. Staroverov, T. A. K., R. Kobayashi, J. Normand, ; K. Raghavachari, A. P. R., J. C. Burant, S. S. Iyengar, ; J. Tomasi, M. C., J. M.

Millam, M. Klene, C. Adamo, R. Cammi, ; J. W. Ochterski, R. L. M., K. Morokuma, O. Farkas, ; J. B. Foresman, a. D. J. F. *Gaussian 16*, Gaussian Inc.: Wallingford, CT, 2016.

37. Becke, A. D., A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J. of Chem. Phys.* **1993,** *98* (2), 1372-1377.

38. Lee, C. T.; Yang, W. T.; Parr, R. G., Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. *Physical Review B* **1988,** *37* (2), 785-789.

39. Perdew, J. P.; Wang, Y., Accurate and Simple Analytic Representation of the Electron-Gas Correlation-Energy. *Physical Review B* **1992,** *45* (23), 13244-13249.

40. Jursic, B. S., Density functional theory and complete basis set ab initio evaluation of proton affinity for some selected chemical systems. *J. Mol. Struct.* **1999,** *487*, 193-203.

41. Bras, N. F.; Perez, M. A. S.; Fernandes, P. A.; Silva, P. J.; Ramos, M. J., Accuracy of Density Functionals in the Prediction of Electronic Proton Affinities of Amino Acid Side Chains. *J. Chem. Theory Comput.* **2011,** *7*, 3898-3908.

42. Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008,** *120*, 215-241.

43. Zhao, Y.; Truhlar, D. G., Exploring the Limit of Accuracy of the Global Hybrid Meta Density Functional for Main-Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2008,** *4*, 1849-1868.

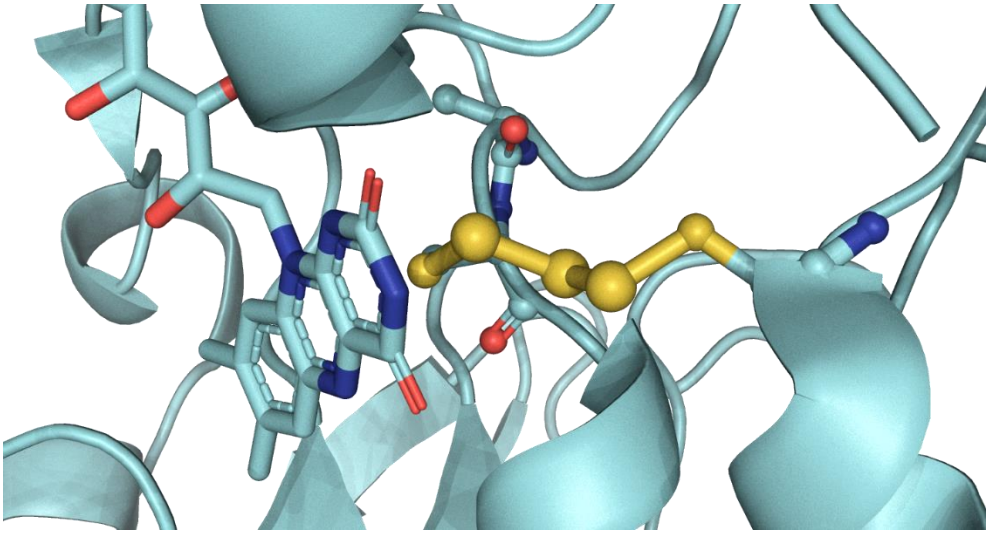
44. Chai, J. D.; Head-Gordon, M., Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008,** *10*, 6615-6620.

45. Hunter, E. P. L.; Lias, S. G., Evaluated gas phase basicities and proton affinities of molecules: An update. *J. Phys. Chem. Ref. Data* **1998,** *27*, 413-656.

46. Moser, A.; Range, K.; York, D. M., Accurate Proton Affinity and Gas-Phase Basicity Values for Molecules Important in Biocatalysis. *J. Phys. Chem. B* **2010,** *114*, 13911-13921.

Chapter 4:

A Theoretical Investigation on Biological Per/Polysulfide Formation: From Small Model to Enzyme



4.1 Introduction

Within biological systems, sulfur-containing species have increasingly been recognized as important as, for example, mediators of oxidative stress and regulators of cellular homeostasis.[1-4] For instance, the antioxidant glutathione (GSH) is readily oxidized to its disulfide form (GSSG) as reactive oxygen species (ROS) and free radicals are concomitantly quenched. Meanwhile, the transsulfuration metabolic pathway is central to the biosynthesis of essential sulfur-containing metabolites such as cysteine and hydrogen sulfide (H_2S).[5-6] Disruptions to this pathway have been shown to lead to a variety of pathological conditions including neurodegeneration,[7-8] vascular abnormalities,[9] and skeletal defects.[10-11]

With the introduction of the 'reactive sulfur species'[12] concept in the early 2000's, there has been a rapidly growing realization of the structural and physicochemical diversity of sulfur-containing biomolecules, and that these are critical for physiological functions. In particular, oxidized sulfur species were regularly observed but they were not thought to be experimentally relevant intermediates of signaling pathways. However, it has now been shown that these species do in fact have important roles in redox signaling and oxidative stress responses. As a result, the biochemistry of disulfide,[13-19] persulfide (RSSH), and polysulfide (RSS_nH, $n>1$) containing biomolecules is a rapidly rising topic of interest.[20-26] For instance, such species are now known to act as electron acceptors in anaerobic bacteria,[27] or involved in sulfur biostorage[28] or as intermediates in sulfur transport,[29-30] as H-atom donors/acceptors,[31] and even in gene expression regulation.[32-33]

The bioformation of such sulfur-sulfur containing species, however, presently remains a much-debated topic.[34] This is due at least in part to the fact that polysulfides are inherently less stable than thiols and decompose in a pH-dependent manner. They are more acidic than their parent thiols by 1-2 pK_a units and more likely exist in their deprotonated forms (RSS^-)

at biological pH where they are more reactive. In addition, they can be challenging to detect experimentally and often require specialized detection methods such as fluorescent sensors.[35-38] This has led to many hypotheses for their formation, including the use of radical and oxidized sulfur intermediates; which of these, if any, occur remains unclear.

Computational chemistry presents a possible alternative approach for the study of such reactive species. However, results can vary significantly depending on the methodology applied. For example, recent benchmarking studies on thiols and polysulfides have shown that DFT can experience difficulties, particularly with predicting proper reaction pathways. Indeed, Hartree-Fock methods have been historically employed to investigate reaction pathways, to varying degrees of success.[39] Fortunately, informed method selection can remedy this to a certain degree with reasonable results.[40-41] In particular, range-separated functionals, notably ωB97X-D, have been shown to be required for correct observation of carbanion intermediates in thiol Michael additions to alkenes by Rowley and coworkers.[42] These important benchmark studies have made insightful strides in computational modeling and provide an improved computational framework for correctly modeling the chemistry of such reactions. Recently, Orabi and Peslherbe studied hydrogen persulfides using high-level *ab initio* methods.[43] They found that these methods were effective in describing not only geometries of hydrogen persulfides, but also the energetics of their isomers and bond dissociation energies. Unfortunately, the application of computational methods to per/polysulfide formation remains unclear, especially with regard to energetics of such proposed reactions.

In this study, we looked at the formation of persulfides and polysulfides using a quantum mechanical-cluster (QM-cluster) approach. A variety of proposed reactions regarding their formation were tested.[34] Small models were built to study differences in reactivity between sulfenic acid, sulfide, and sulfur radicals as they are commonly predicted intermediate

species. Orbital interactions of oxidized sulfur species with a radical suggests a low barrier for S—S bond formation. We applied these findings to the mechanism of sulfur chain extension in *A. ambivalens* sulfide:quinone oxidoreductase (SQR). Our findings similarly suggest that radical sulfur intermediates are more conducive to sulfur-sulfur bond formation rather than nucleophilic attack of a thiolate on a thiol as is typically proposed. In particular, SQR relies on flavin adenine dinucleotide (FAD) to abstract a hydrogen atom followed by an electron transfer to form a neutral polysulfide, thus enabling S–S bond formation between the enzyme's sulfur radical and HS^- as the sulfur source for the growing chain.

4.2 Computational Methods

All *ab initio* and DFT calculations were performed using the Gaussian 09 package.[44] Optimized structures were obtained at the dispersion corrected ωB97X-D[45]/6-311+G(2d,p) level of theory as we have previously shown it to be able to provide reliable structures for related and bio-relevant polysulfide species.[41-42] It is noted that for comparisons, optimized structures were also obtained using the same basis set in conjunction with the QCISD[46] and MP2 *ab initio* methods. Frequency calculations were performed at the same level of theory to confirm each structure as a minimum or a first-order transition structure. Single point calculations were performed using the same DFT method as for the optimized structure, but now using the more extensive 6-311+G(2df,p) basis set, i.e., Method1/6-311+G(2df,p)//Method1/6-311G(2d,p). The environments polarity was modelled using the IEFPCM solvation method with dielectric constants (ε) of 4, 10, or 80. Calculations were performed to test proton, hydrogen, and hydride affinities of FAD, cysteine, cysteine persulfide, and cysteine polysulfide at the ωB97X-D/6-311+G(2df,p)//ωB97X-D/6-311G(2d,p) level of theory. Results from single point calculations on the smallest models were also submitted for NBO analysis.[47] Orbitals were visualized using the NBOPro@Jmol interface. For the computational investigations on the polysulfide chain extension mechanism of sulfide:quinone oxidoreductase (SQR) from the archaea *A. ambivalens*, the

X-ray crystal structure PDB ID 3H8L was used as a template.[48] The structure was submitted to the PDB2PQR webserver for assignment of protonation states of side chains.[49] A QM-model consisting of 251 atoms was built from this template including: the Cys350-bound trisulfide, Cys178, FAD, Ala177, Glu179, Gly180, Pro181, Asp215, Pro314, Lys315, Leu316, Val349, Val351, Ala352, Asp353, and a water molecule which was replaced with HS^-. Atoms at the edges of the active site were fixed along the protein backbone to represent restrictions to movement imposed by the protein. Calculations on this model were done at the ωB97X-D/6-311+G(2df,p)//ωB97X-D/6-311G(2d,p) level of theory.

4.3 Results and Discussion

4.3.1 Small-model benchmark calculations

Literature results have shown ωB97X-D is a suitable DFT method for calculating chemical properties of thiol compounds, though data concerning radical systems is somewhat lacking. Herein, we performed benchmark calculations to evaluate the performance of neutral and radical anion sulfur-containing molecules. Such radical-containing systems can be particularly challenging to obtain accurate results using DFT methods due to complex electronic behaviour. However, the importance of these radical intermediates cannot be underscored as they are commonly theorized in biological sulfur pathways. Thus, we investigated the possibility for DFT, in particular ωB97X-D, to accurately describe their geometry and energetics. We compared the stability of $CH_3S_nH^{\bullet-}$ (n=1-3) in gas phase and in water using QCISD and MP2. In solvent, these radicals are highly unstable and dissociate to a methyl sulfide (n=0–2) and $HS^{\bullet}$ (shown below, Figure 4.4). In gas phase, this is also observed for the per- and polysulfide, however, methyl thiyl radical anion does not dissociate. Optimized geometries were similar across all methods, indicating ωB97X-D can be used to obtain structural data on sulfur compounds with a reliable accuracy. In addition, the sulfur environment must have a low polarity to accommodate the unstable nature of

these radical anions, if they indeed are present. Due to their instability, such transient states must be stabilized for a sufficient length of time to avoid undesirable products.

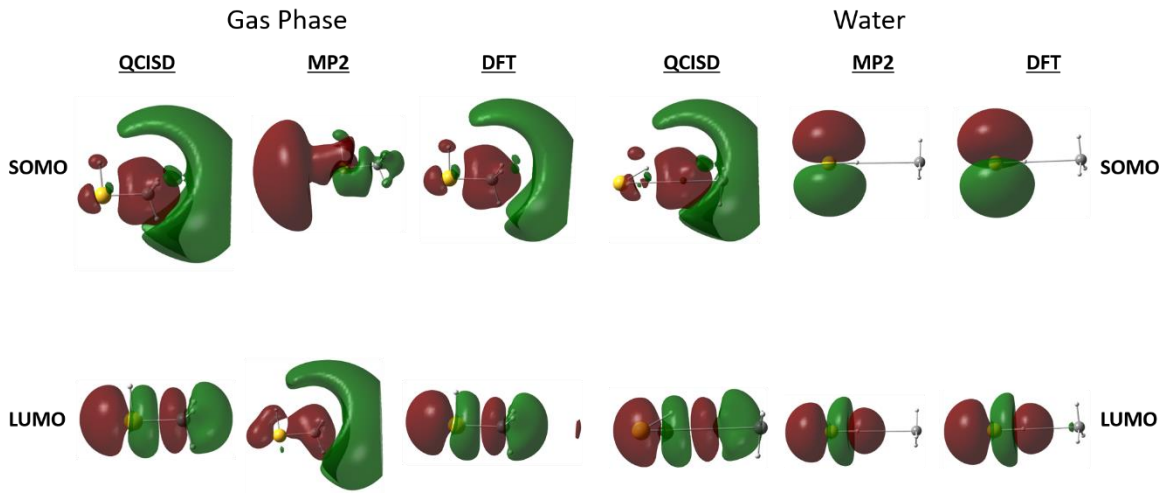


Figure 4.1. Visualized SOMO and LUMO orbitals of $CH_3SH^{\bullet-}$ in gas phase (left) and water (right).

We further investigated $CH_3SH^{\bullet-}$ by visualizing the orbitals from gas phase and solvent calculations (Figure 4.1) and using NBO to visualize and quantify the most stabilizing interactions (Figure 4.2, Table 4.1). NBO analysis calculates the energetic contribution of each stabilizing interaction and the orbitals involved, giving an idea of the relative stability of a compound. In solvent MP2/DFT calculations, we observed the dissociation to give $HS^{\bullet}$ and methanide. Notably, MP2 orbitals show that the SOMO is present on sulfur rather than methyl as predicted with the other two methods. Results between the two methods agreed that the radical electron is present on the methyl group (specifically for the non-dissociated systems in gas phase calculations). For both QCISD and MP2 calculation, four stabilizing interactions were observed. The highest in magnitude were the sulfur lone pair interaction with a C—H antibonding orbital (two degenerate interactions). The next stabilizing interaction was between the S—H bond and a Rydberg orbital on C. Since DFT is not an electron correlation method, it cannot compute interaction energies from higher energy

states. This manifests in 3 stabilizing interactions rather than 4 as observed with QCISD and MP2. The most stabilizing interactions are by far the most significant with a 70% contribution. Interestingly, interaction of the radical with any orbital was not a contributing factor and is not observed in any case. These results not only indicate that this species is likely transient and quite unstable and demonstrate once again that while DFT can give accurate results stabilizing contributions from high energy states are not accounted for in these calculations. For well behaved systems, these are not significant since these are low in energy, but it is particularly impactful for smaller systems as errors are much more exaggerated and become more apparent.

Table 4.1. Most stabilizing interactions for $CH_3SH^{\bullet-}$ calculated by NBO analysis. LP represents a lone pair orbital, BD represents a bonding orbital, BD* represents an antibonding orbital, and RY* represents a Rydberg orbital.

Method	NBO Donor	NBO Acceptor	Energy (kcal/mol)
QCISD	LP(S)	BD* (C—H)	2.16
	LP(S)	BD* (C—H)	2.16
	BD(S—H)	RY* (C)	1.48
	BD(S—H)	BD* (C—H)	1.27
MP2	LP(S)	BD* (C—H)	2.23
	LP(S)	BD* (C—H)	2.22
	BD(S—H)	RY* (C)	1.49
	BD(S—H)	BD* (C—H)	1.30
ωB97X-D	LP(S)	BD* (C—H)	1.97
	LP(S)	BD* (C—H)	1.97
	BD(S—H)	BD* (C—H)	1.26

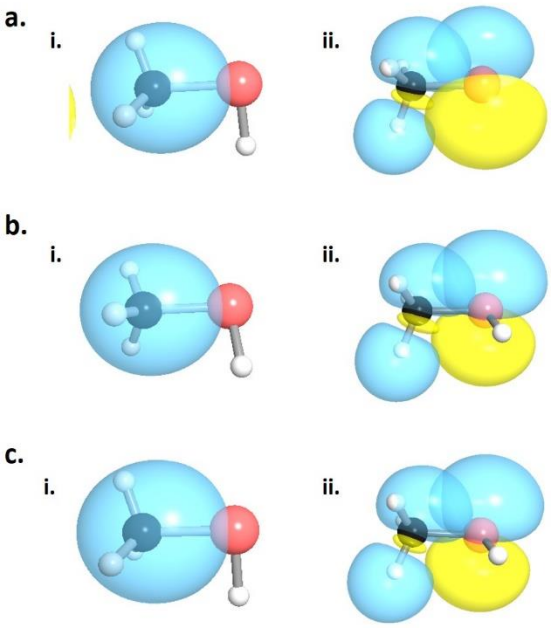


Figure 4.2. NBO orbitals of the (i) singly-occupied molecular orbital and (ii) the most stabilizing orbital interaction calculated using (a) QCISD/6-311+G(2df,p), (b) MP2/6-311+G(2df,p), and (c) ωB97X-D/6-311+G(2df,p) gas phase calculations.

We also studied methyl thiyl using the same approach as a comparison to $CH_3SH^{\bullet-}$. We expected to observe fewer differences in *ab initio* compared to DFT calculations due to lack of charge and thus any meaningful contributions from charged excited states. This was indeed reflected in the calculations as the SOMO is now present on the sulfur atom instead. Furthermore, the largest contributors to the stability of methyl thiyl did not contain any Rydberg states as seen in $CH_3SH^{\bullet-}$ (Table 4.2). Rather, these were from a C—H orbital interaction with the radical orbital, and an interaction from the sulfur lone pair with a pair of near-degenerate C—H antibonding orbitals (Figure 4.3, Table 4.2 QCISD and MP2). Using DFT however, there are four observed interactions: the same three and an additional interaction from a C—H orbital with the SOMO. The magnitude of these interactions is quite different across these four contributions, however, and one of the interactions of a

sulfur lone pair with a C—H orbital becomes degenerate with the most stabilizing interaction predicted with QCISD and MP2. The other orbital is slightly less energetic, although also present in *ab initio* calculations. Notably, the magnitude of stabilizing interactions for methyl thiyl is significantly higher than for $CH_3SH^{\bullet-}$ (Table 4.2, magnitude of stabilizing interactions). In addition, DFT predicts methyl thiyl to be more stable than the other *ab initio* methods. These results once again suggest that stability and energetics calculations using DFT methods can become problematic even in the simplest chemical systems from the lack of consideration of higher energy contributions. Although it would be ideal to consider contributions of higher energy states to the structure as done in QCISD (or other high-level electron correlation methods), this is unfeasible in larger systems, particularly enzymes, due to the high associated computational cost. Indeed, ωB97X-D struggles to accurately describe some of these systems but can give reasonable results at a much lower cost.

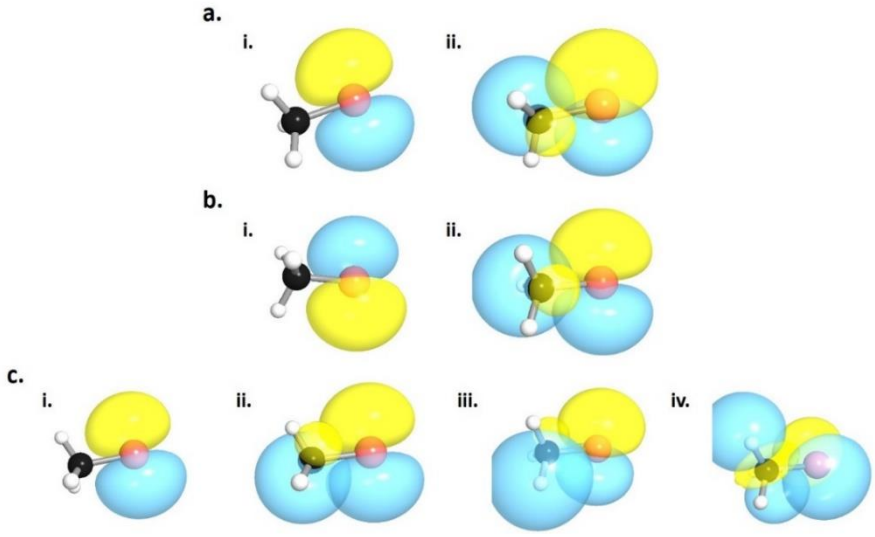


Figure 4.3. NBO calculations for $CH_3S^{\bullet}$ in gas phase using (a) QCISD/6-311+G(2df,p), (b) MP2/6-311+G(2df,p), and (c) ωB97X-D/6-311+G(2df,p). The two sets of orbitals shown in (c) are from degenerate contributions.

Table 4.2. Most stabilizing interactions for $CH_3S^{\bullet}$ obtained from NBO calculations. LP represents a lone pair orbital, LP* represents and antibonding lone pair orbital, BD represents a bonding orbital, BD* represents an antibonding orbital, and RY* represents a Rydberg orbital.

Method	NBO Donor	NBO Acceptor	Energy (kcal/mol)
QCISD	BD(C—H)	LP*(S)	4.54
	LP(S)	BD* (C—H)	3.46
	LP(S)	BD* (C—H)	3.43
MP2	BD(C—H)	LP*(S)	4.75
	LP(S)	BD* (C—H)	3.55
	LP(S)	BD* (C—H)	3.52
ωB97X-D	BD(C—H)	LP*(S)	4.57
	LP(S)	BD* (C—H)	4.52
	BD(C—H)	LP*(S)	4.45
	LP(S)	BD* (C—H)	1.04

We then investigated the ability for ωB97X-D to describe electronic behaviour in more complex sulfur systems, in particular per/polysulfide radical anions. SOMO and LUMO orbitals of each compound were identified (Figure 4.4). For polysulfide calculations, the LUMO orbitals were very similar across the three methods tested. As mentioned previously, these compounds dissociate due to their inherent instability. Notably, there were some differences observed in the SOMOs originating from the terminal S—S bond. QCISD calculations in solvent indicated the radical electron lies in an orthogonal *p*-orbital (there are 2 with degenerate energy levels). LUMOs are very similar regardless of method or solvent used, but QCISD predicts the SOMO to be on the terminal sulfur atom rather than the inner sulfur (Figure 4.4a). For these larger systems, DFT results are reliable, specifically in the polysulfide. These results also highlight the importance of choosing a proper method for benchmarking accuracy.

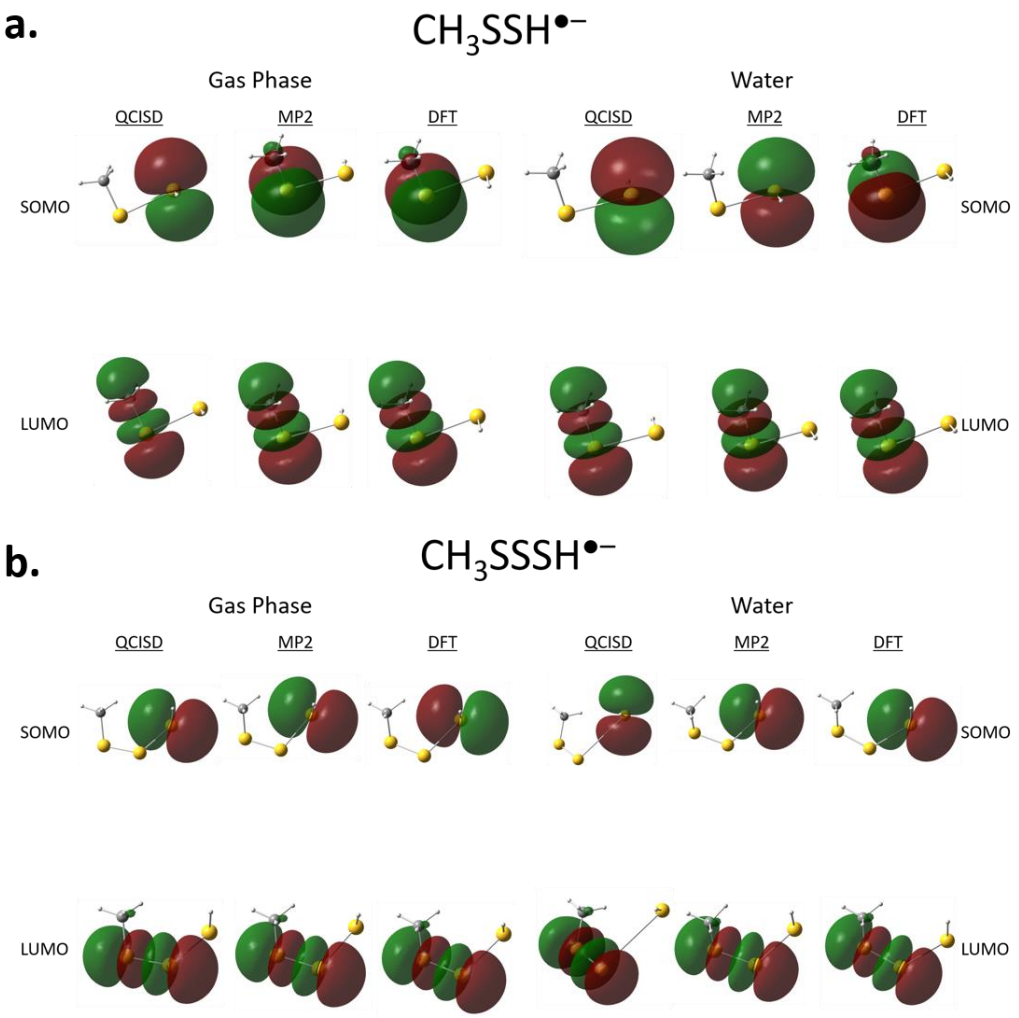


Figure 4.4. Calculated SOMO and LUMO orbitals of (a) $CH_3SSH^{\bullet-}$ and (b) $CH_3SSSH^{\bullet-}$ in gas phase (left) and water (right). Orbitals are obtained from Gaussian checkpoint files.

4.3.2 Small model reactivity

With the knowledge in hand that DFT struggles with charged radical species, we investigated model reactions of sulfur addition to determine if the DFT experiences similar problems. It is at this point QCISD starts to become prohibitively expensive, so we must assess the applicability of ωB97X-D to biological models. Here, we tested the addition of $HS^{\bullet}$ onto CH_3SH to represent a highly simplified system and extended to a model of a simple persulfide (CH_3SS^-) with isobutene and dehydroalanine using gas phase NBO calculations at the same level of theory to model a biological sulfur transfer process (Figure

4.5). The small model study examined the addition of a radical sulfur intermediate onto a thiol, a simplified model of a commonly proposed biological sulfur transfer pathway. Reaction modeling indicated that addition of HS^- onto thiols and RS^- did not result in formation of a sulfur bond or indicate any orbital overlap in calculations (data not shown). Instead, these species tend to repel each other regardless of solvent polarity (ε = 1 (gas-phase), 4, 10, 80) or inclusion of explicit water molecules for added hydrogen bonding stabilization. With the optimization of $SH^\bullet$ + CH_3SH, we observed a bonding orbital formed between the two sulfur atoms (Figure 4.5a). This bonding interaction is rather weak (3.79 kcal mol^{-1}) and is stabilized by C—H antibonding.

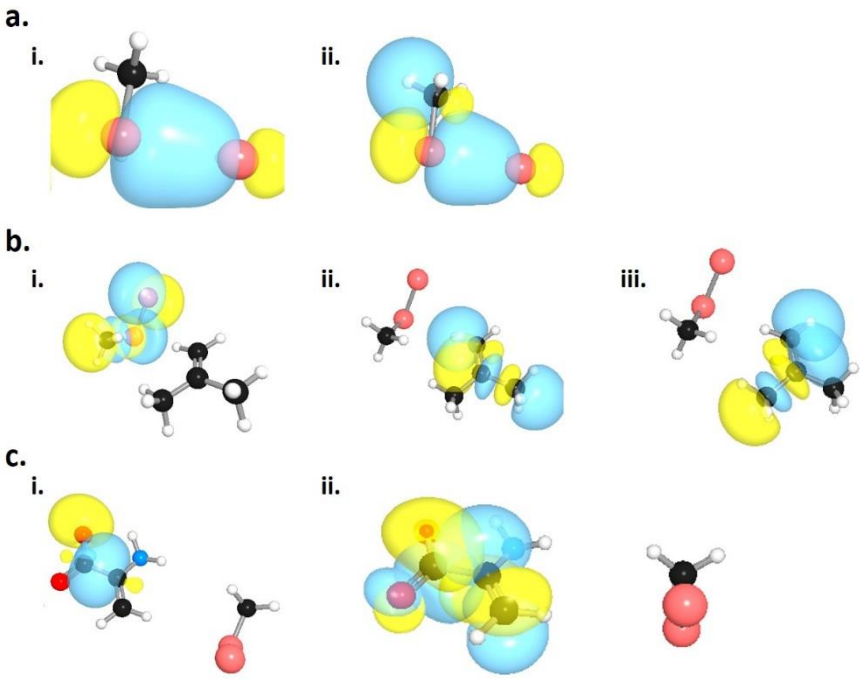


Figure 4.5. Top two stabilizing NBO interactions of (a) $HS^\bullet$ + CH_3SH, (b) CH_3SS^- + isobutene, (c) CH_3SS^- + dehydroalanine as shown in NBO. In (c, ii), degenerate interactions are shown.

Expanding the model to the larger system of CH_3SS^- with isobutene and dehydroalanine, the reactants were oriented to encourage orbital overlap between a deprotonated carbon and

a sulfur *p*-orbital as a model for the proposed sulfur transfer step in SufS (see **Chapter 8**). No interactions between the two molecules were observed at any point (Figure 4.6c). We attempted to perform a scan on the potential energy surface to obtain the reactants CH_3S^- and S-isobutene/dehydrocysteine from the product, however, we were met with no results. Instead, the two negative charges on sulfur repel each other as in the small models and negative charge on the persulfide is repelled from the isobutene/dehydroalanine intermediate. This leads us to believe that sulfur transfer likely does not occur between two sulfide species and, if possible, one of the sulfur atoms likely exists as a radical. However, protonation states of cysteinyl residues or the substrate are unknown (for more detail, see **Chapter 8**). In enzymes with redox-active cofactors such as FAD, a hydrogen atom from cysteine may be abstracted to produce a sulfur-centered radical which can interact with hydrogen sulfide from solution (discussed below).

Since reactants could not be obtained from scans on the reverse of the potential energy surface, we instead started from a set of reactants to investigate the possibility of different conditions (i.e., protonation states of sulfur species) to form a S—S bond. The reactions tested were as follows: CH_3SH + dehydroCys-SH, CH_3S^- + dehydroCys-S^-, CH_3SH + dehydroCys-S^-, and CH_3S^- + dehydroCys-SH (Figure 4.6). In these enzymes, knowledge of protonation states and pK_as of the active site residues is vital to understand the driving factors of sulfur transfer. This is because they lack electron transfer cofactors and may rely on transient, unstable intermediates to catalyze S—S bond formation. Thus, no radical-containing models were built to model persulfide formation. For more details, please refer to **Chapter 8**. Our model was built based on the SufS active site with dehydrocysteine representing the pyridoxal 5'-phosphate (PLP)-bound first reaction intermediate and methyl thiol as the active site Cys367 identified as the sulfur acceptor. The carboxylate group is protonated since in the enzyme, it is stabilized by hydrogen bonding. In Figure 4.7a, two neutral thiol reactants result in no transfer and are likely repelled by lone pair electrons. The

resulting S—S distance is 4.02 Å suggesting there is no interaction between the two sulfur atoms. By stark contrast, in Figure 4.6b, two thiolates optimized to a sulfur transfer from dehydrocysteine to methyl thiolate. The S—S bond length was 2.06 Å in the optimized complex, and the S—C distance (from the transferred sulfur atom to the dehydroalanine fragment) was 1.99 Å. This suggests that there is perhaps an enzyme-substrate complex formed to aid in the transfer of sulfur and there may be a subsequent proton transfer to facilitate S—C bond clevage. In Figure 4.6c, we consider the possibility for the thiolate to exist on the substrate rather than the active site cysteinyl. In this model, the cysteinyl thiol proton readjusts to hydrogen bond with the substrate thiolate. Figure 4.6d represents the proposed mechanism with attack of the cysteinyl thiolate on the substrate sulfur atom. The product of this reaction is proton abstraction from the amine to restore the neutral thiol.

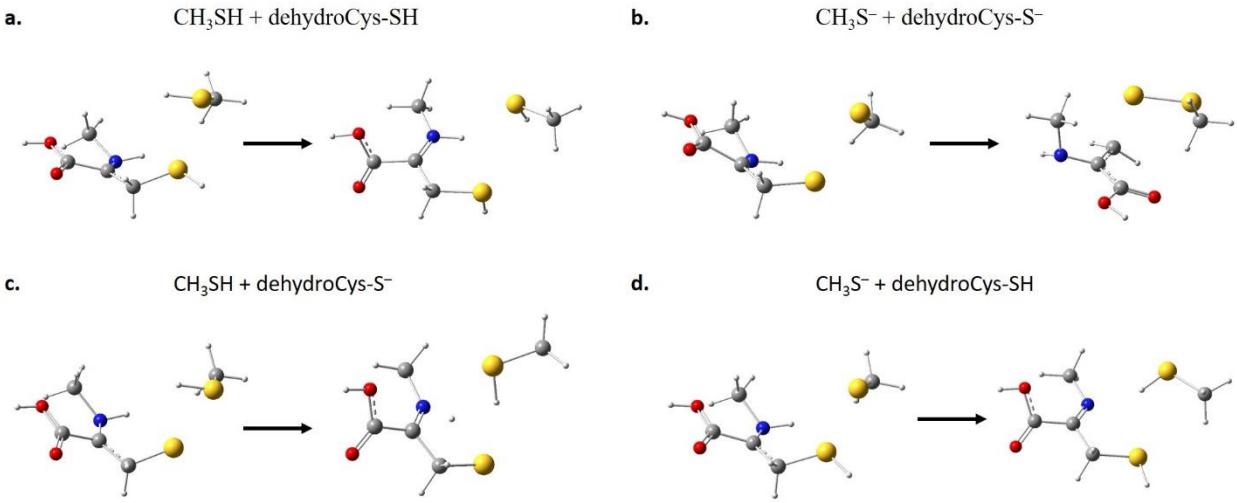


Figure 4.6. Model reactions tested for S—S bond formation. Each model has been optimized. No scans were performed to obtain products.

We further examined the potential for persulfide formation via radical sulfur addition onto an oxidized sulfur intermediate. Oxidized sulfur is part of a larger group of reactive sulfur species (RSS) involved in many signaling pathways. In particular, cysteine oxidation to the sulfenic acid has been suggested as a route to formation of a sulfenyl amide or other S—N

derivatives such as in PTP1B.[50] It has also been shown that Cys—SOH is part of the pathway for disulfide formation and spontaneous protein folding.[51] That is, it acts as a regulator of protein activity and protects against further oxidative damage (ex. to cysteine sulfinic/sulfonic acid). Reaction pathway intermediates are difficult to study experimentally since they are transient; oxidation of sulfur is proposed to render it more electrophilic and thus more susceptible to nucleophilic attack. To investigate this possibility, we tested a model reaction of $CH_3SOH + HS^-$ and $CH_3SOH + HS^{\bullet}$ using the same DFT method (Figure 4.7). In the addition of HS^-, a S—S bond could not be formed. Attempted scans along this reaction coordinate were not successful, so we attempted calculations from an intermediate state with both HO^- and HS^- bound to methyl thiolate. While the oxidized sulfur has been rendered more electrophilic, HO^- must act as a leaving group in competition with HS^-, which is not favourable. Indeed, HS^- is a better leaving group and this was reflected in the calculations whereby the reactant state is restored.

HOMO and LUMO orbitals post optimization were examined to determine if there was possibility for attack, Figure 4.7c. In the HOMO, we see there was some interaction, but to a much lesser degree than in the SOMO (Figure 4.7d). The LUMO however, indicates that HS^- can replace HO^-, although the energy cost associated with this is quite large and requires specific environment conditions. In the case of $HS^{\bullet}$, bond formation seems to be much more plausible. The S—S is significantly shorter at 2.51 Å compared to 3.19 Å when sulfide is a reactant. The SOMO closely resembles the LUMO of the HS^- addition and the LUMO shows electrons present on HO^-. Together, these results suggest that addition using a radical mechanism is much more favourable rather than competition between two nucleophiles where HO^- is a better nucleophile than HS^-, unless the protein environment plays a role. We also tested the possibility for hydrogen sulfide to form an S—S bond through including solvent stabilization from 2 and 3 explicit solvent molecules, though results indicated that a proton transfer pathway formed instead (results not shown).

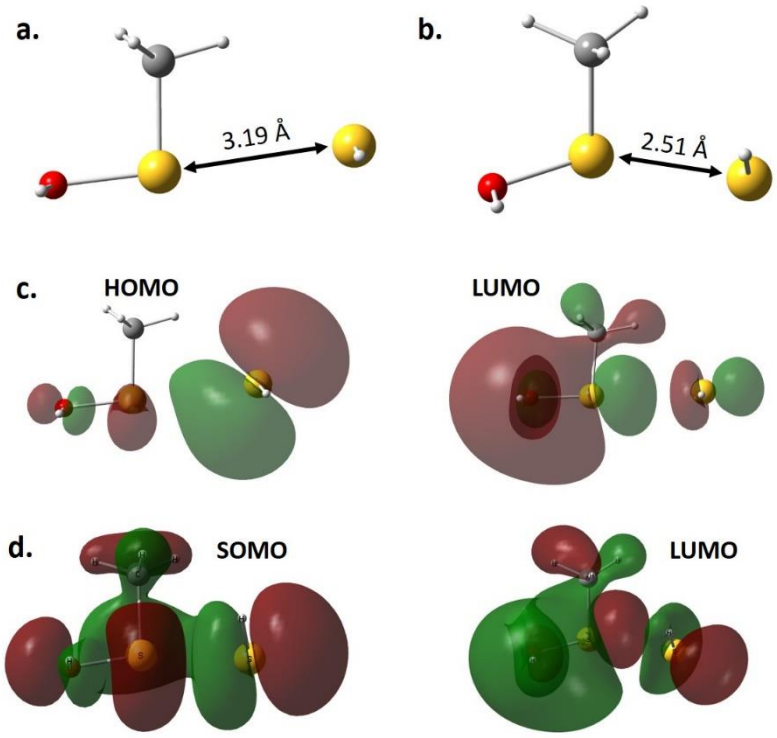


Figure 4.7. Tested reactions for S—S bond formation from an oxidized sulfur intermediate. (a) Optimized $CH_3SOH + HS^-$, (b) optimized $CH_3SOH + HS^{\bullet}$, (c) HOMO and LUMO for $CH_3SOH + HS^-$, and (c) SOMO and LUMO for $CH_3SOH + HS^{\bullet}$.

4.3.3 Sulfide:Quinone Oxidoreductase (SQR)

Sulfur oxidation pathways to form protein-bound persulfides have also been characterized from the mitochondrial enzyme sulfide:quinone oxidoreductase (SQR).[52-54] This enzyme oxidizes cysteine to form cysteine-bound polysulfides from H_2S. In particular, we examined SQR from the thermoacidophilic archaea *Acidianus ambivalens* which uses sulfur as an energy source.[48] In this organism, SQR forms polysulfide using H_2S as a sulfur source and extending a sulfur chain bridged by two active site cysteinyl residues. Once these sulfur chains grow to a sufficient length, they are released from the enzyme as polysulfide. There are several reported literature mechanisms for this enzyme,[48, 55-56] and the presence of a bound FAD complicates studies due to its redox nature. Its presence

suggests the possibility for a hydrogen-based (proton-coupled electron transfer) or hydride-based sulfur addition mechanism.

Thus, we investigated the involvement of FAD in the enzyme's mechanism by examining its potential to act as a proton, hydrogen atom, and hydride acceptor from cysteinyl and its polysulfur derivatives (Figure 4.8). While the nitrogen atom is typically represented as the proton acceptor, we also investigated the possibility for the oxygen atom to act as an acceptor based on the spatial orientation of FAD in the active site. Hydrogen bonding between the oxygen atom (Figure 4.8a) and the cysteinyl polysulfide is observed. It has also been previously shown that this oxygen atom can act as a catalytic base to initiate a reaction in other systems.[57] Calculations were done in both gas phase and a dielectric constant of 10 to observe the differences imposed by the enzyme environment on the affinities. Proton affinities are much higher in a dielectric than in gas phase to avoid charged intermediates. Hydride affinities are curiously much higher in gas phase and a solvent environment decreased the affinities by up to 300 kJ mol^{-1}. This may reflect the role imposed by the enzyme active site to better match the ability for the substrate to give up hydride, thus lowering the reaction barrier. Hydrogen atom affinities are very similar regardless of solvent polarity with the oxygen atom having a lower affinity by 52.6 kJ mol^{-1}. While the nitrogen atom is a better hydrogen atom acceptor, FAD(O) is better positioned to receive the hydrogen atom and may act as an intermediate for transfer to FAD(N). We compare these affinities to those of cysteinyl and its per/polysulfide derivatives to understand which of these processes is most likely to occur in the enzyme (Figure 4.8b). Proton affinities and hydride affinities are very large barriers to overcome (>300 kJ mol^{-1}) and we observe that hydrogen atom affinities are within 100 kJ mol^{-1}. As the sulfur chain extends, the terminal sulfur group becomes a better hydrogen atom donor which has also been previously shown by Pratt and coworkers.[31]

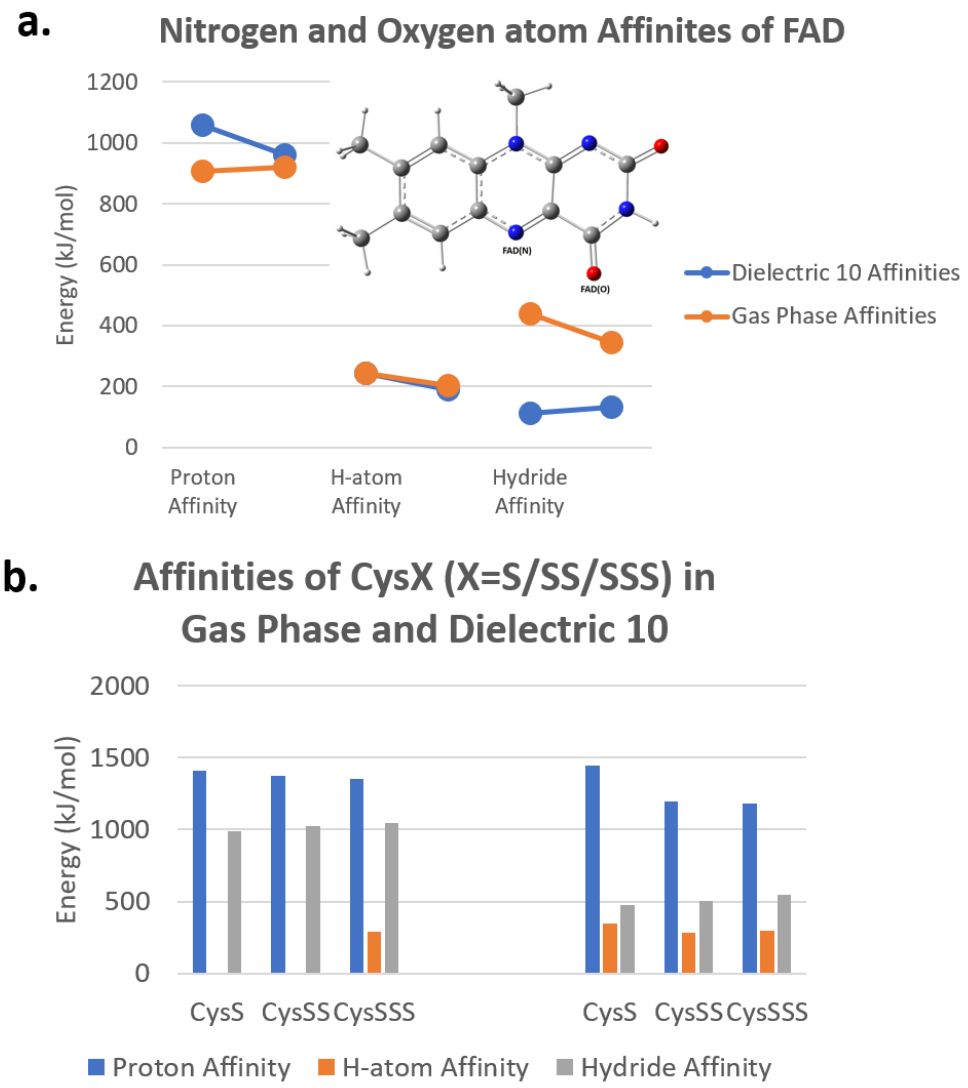


Figure 4.8. (a) Proton, hydrogen atom, and hydride affinities for the oxygen and nitrogen atoms of FAD in gas phase and a dielectric of 10. An inset is shown in (a) highlighting the atoms of interest on FAD (optimized at ωB97X-D/6-311G(2d,p)). (b) Proton, hydrogen atom, and hydride affinities of Cys and Cys-per/polysulfide derivatives in gas phase (left) and a dielectric of 10 (right).

While the affinities gave us a rough estimate of the enzyme mechanism, the surrounding protein environment is not accounted for in these calculations. We extended our model to contain the enzyme active site and FAD (Figure 4.9). The model consisted of 251 atoms

taken from the crystal structure of SQR (PDB ID: 3H8L[48]), including two important proposed aspartyl residues, the bridging cysteinyl residues, FAD, and a significant portion of the active site environment. Importantly, the protonation state of the growing sulfur chain is unknown due to the nature of crystal structure prediction. The oxygen atom from FAD is in hydrogen bond proximity to the terminal sulfur, so we propose that this begins in a protonated state. Furthermore, this supports the work of Pratt and coworkers since the growing polysulfide chain can act as a hydrogen atom donor to FAD. Asp353 is proposed to abstract a proton from H_2S to give HS^-, however the distance in the model was too large to consider the possibility of deprotonation in our modeling. Since the active site is partially exposed to allow for substrate to enter, HS^- replaced a crystal water to act as the sulfur source. We speculate that since the pK_a of H_2S is 7.04, it can enter as hydrogen sulfide. The optimized reactive complex is shown in Figure 4.9. The distance to the oxygen atom of FAD is only 2.92 Å compared to 3.82 Å to the nitrogen atom (not shown).

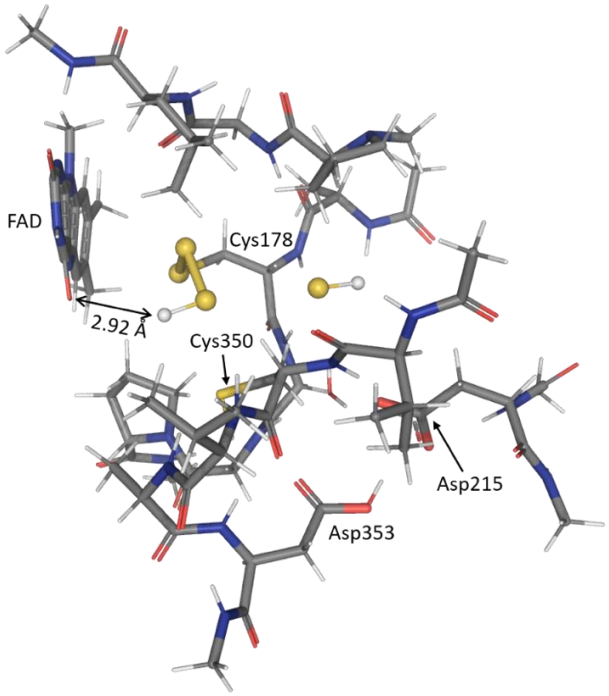


Figure 4.9. Optimized reactive complex (**RC**) for the model of sulfur chain extension from three to four sulfur atoms. The crystal structure was taken from the PDB ID: 3H8L.

The presence of the FAD cofactor indicates the necessity of a redox reaction; thus, we proceed through a proton-coupled electron transfer mechanism (Figure 4.10). In **IC1**, a proton from the terminal sulfur atom is transferred to FAD(O) to give **IC2**. This step is thermodynamically favoured and occurs without a barrier. We also attempted backwards scans from this step to confirm and indeed obtained **RC**. From **IC1**, the proton is transferred from FAD(O) to FAD(N). This results in a shortening of the ring C atom to the second sulfur atom in the chain (2.14 Å; Figure 4.10 **IC2**) to resemble the intermediate proposed in the literature whereby a sulfur atom binds to this carbon atom.[58] Following **IC2**, an electron was moved from the trisulfur group to the FAD system to model a complete H-atom transfer. Since this step is thermodynamically unfavourable and comes with an associated barrier of 206.5 kJ/mol, there may be an additional step associated with electron transfer. Once the electron transfer occurred, formation of a S—S bond occurred with relative ease to give **IC3**. The energy listed for the formation of **PC** is not listed as there are many associated steps with the exit of FAD from the active site and any required rearrangements. To model this step, steered MD is recommended, although a template structure of the active site lacking FAD is required.

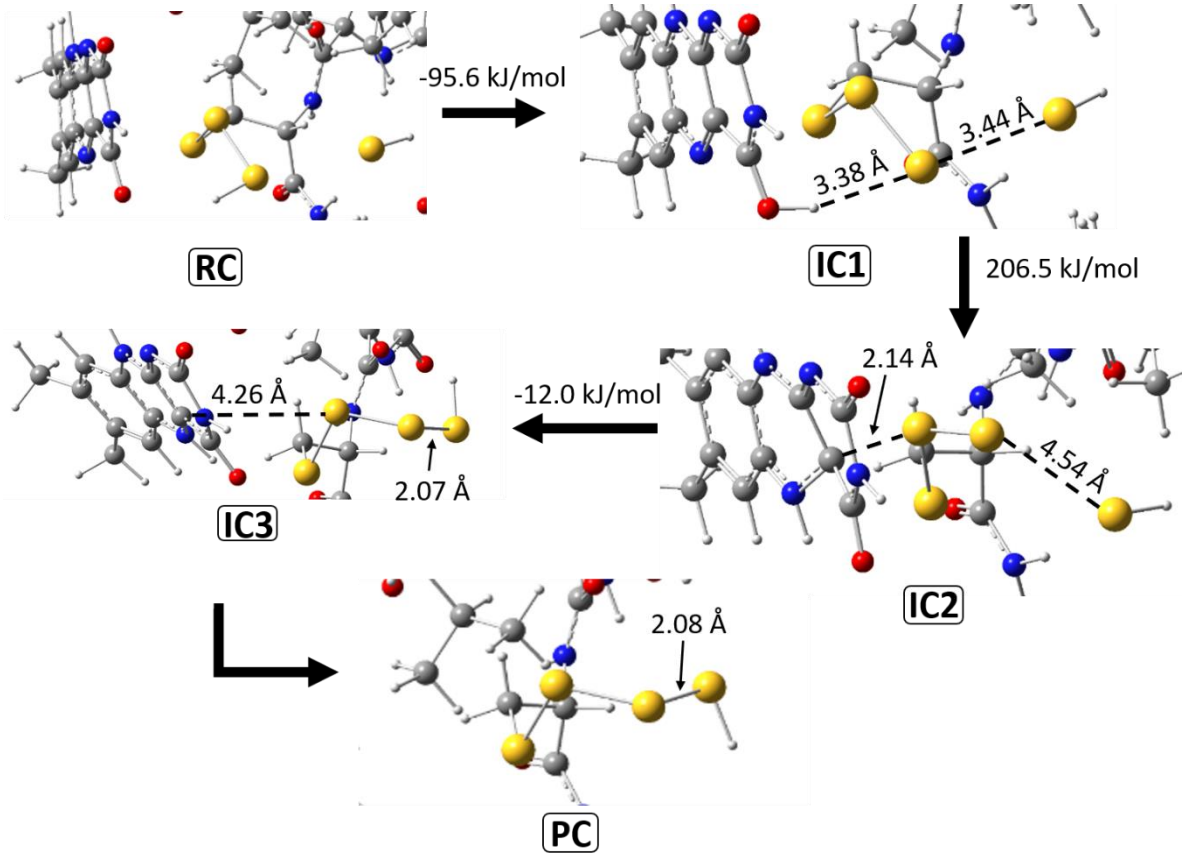


Figure 4.10. Selected optimized structures (see Computational Methods) and relative energies along the proposed mechanism of SQR (arrows may represent multiple transformations). For clarity, only the cysteinyl's, sulfurs and FAD and shown (Sulfur = yellow, Carbon = grey, Nitrogen = blue, Oxygen = red, Hydrogen = white).

4.4 Conclusions

In this study, we examined the possibility for the DFT method ωB97X-D to accurately model proposed biological per/polysulfide pathway intermediates. These are often represented as radical species or oxidized cysteine derivatives that are difficult to detect experimentally. Benchmarking studies on small models that ωB97X-D is a reliable method for larger sulfur chains ($n \geq 3$), and that radical anion species are not well described due to stabilizing contributions from higher energy states. MP2 also struggles in niche cases, indicating it is not a suitable method to use for benchmark calculations. Thiolates tend not

to interact with thiols, a commonly proposed persulfide formation pathway, but may proceed through two thiolate species followed by protonation. Reactions involving sulfur atoms are thermodynamically favoured to occur else their barrier is very high, and systems tend towards favourable hydrogen bonding orientations. Our model SQR results suggest FAD(O) can act as a proton acceptor to mediate transfer to FAD(N) due to sulfur proximity. Following proton transfer, electron transfer in SQR is likely the largest barrier to reactivity and as soon as an electron is lost, S—S bond formation is promoted. Overall, our results inform that sulfur reactions are most favourable to occur through radical intermediates. Further improvements to DFT methodology are required to improve accuracy in the study of sulfur oxidation. Although high-level *ab initio* methods can accurately model these intermediates, they are unfeasible for large enzyme models. However, even though DFT performance is not ideal, we can still use ωB97X-D to gain valuable insights into biological redox sulfur mechanisms.

4.5 References

1. Giles, G. I.; Nasim, M. J.; Ali, W.; Jacob, C., The Reactive Sulfur Species Concept: 15 Years On. *Antioxidants (Basel)* **2017,** *6*.

2. Lau, N.; Pluth, M. D., Reactive sulfur species (RSS): persulfides, polysulfides, potential, and problems. *Curr. Opin. Chem. Biol.* **2019,** *49*, 1-8.

3. Olson, K. R.; Straub, K. D., The Role of Hydrogen Sulfide in Evolution and the Evolution of Hydrogen Sulfide in Metabolism and Signaling. *Physiology (Bethesda)* **2016,** *31*, 60-72.

4. Kimura, H., Signaling molecules: hydrogen sulfide and polysulfide. *Antioxid. Redox Signal.* **2015,** *22*, 362-76.

5. Kabil, O.; Banerjee, R., Enzymology of H2S biogenesis, decay and signaling. *Antioxid. Redox Signal.* **2014,** *20*, 770-82.

6. Sbodio, J. I.; Snyder, S. H.; Paul, B. D., Regulators of the transsulfuration pathway. *Br. J. Pharmacol.* **2019,** *176*, 583-593.

7. McBean, G. J.; Aslan, M.; Griffiths, H. R.; Torrao, R. C., Thiol redox homeostasis in neurodegenerative disease. *Redox Biol.* **2015,** *5*, 186-194.

8. Paul, B. D.; Snyder, S. H., Neurodegeneration in Huntington's disease involves loss of cystathionine gamma-lyase. *Cell Cycle* **2014,** *13*, 2491-3.

9. Watanabe, M.; Osada, J.; Aratani, Y.; Kluckman, K.; Reddick, R.; Malinow, M. R.; Maeda, N., Mice deficient in cystathionine beta-synthase: animal models for mild and severe homocyst(e)inemia. *Proc. Natl. Acad. Sci. USA* **1995,** *92*, 1585-9.

10. Kruger, W. D., Cystathionine beta-synthase deficiency: Of mice and men. *Mol. Genet. Metab.* **2017,** *121*, 199-205.

11. Parrot, F.; Redonnet-Vernhet, I.; Lacombe, D.; Gin, H., Osteoporosis in late-diagnosed adult homocystinuric patients. *J. Inherit. Metab. Dis.* **2000,** *23*, 338-40.

12. Giles, G. I.; Jacob, C., Reactive sulfur species: an emerging concept in oxidative stress. *Biol. Chem.* **2002,** *383*, 375-88.

13. Deponte, M., Glutathione catalysis and the reaction mechanisms of glutathione-dependent enzymes. *Biochim. Biophys. Acta* **2013,** *1830*, 3217-66.

14. Go, Y. M.; Jones, D. P., Thiol/disulfide redox states in signaling and sensing. *Crit. Rev. Biochem. Mol. Biol.* **2013,** *48*, 173-81.

15. Jones, D. P.; Sies, H., The Redox Code. *Antioxid. Redox Signal.* **2015,** *23*, 734-46.

16. Lu, J.; Holmgren, A., The thioredoxin antioxidant system. *Free Radic Biol Med* **2014,** *66*, 75-87.

17. Naumov, S.; Schoneich, C., Intramolecular addition of cysteine thiyl radical to phenylalanine and tyrosine in model peptides, Phe (CysS*) and Tyr(CysS*): a computational study. *J. Phys. Chem. A* **2009,** *113*, 3560-5.

18. Northrop, B. H.; Coffey, R. N., Thiol-ene click chemistry: computational and kinetic analysis of the influence of alkene functionality. *J. Am. Chem. Soc.* **2012,** *134*, 13804-17.

19. Turell, L.; Radi, R.; Alvarez, B., The thiol pool in human plasma: the central contribution of albumin to redox processes. *Free Radic. Biol. Med.* **2013,** *65*, 244-253.

20. Bailey, T. S.; Pluth, M. D., Reactions of isolated persulfides provide insights into the interplay between H2S and persulfide reactivity. *Free Radic. Biol. Med.* **2015,** *89*, 662-7.

21. Fukuto, J. M.; Ignarro, L. J.; Nagy, P.; Wink, D. A.; Kevil, C. G.; Feelisch, M.; Cortese-Krott, M. M.; Bianco, C. L.; Kumagai, Y.; Hobbs, A. J.; Lin, J.; Ida, T.; Akaike, T., Biological hydropersulfides and related polysulfides - a new concept and perspective in redox biology. *FEBS Lett.* **2018,** *592*, 2140-2152.

22. Ida, T.; Sawa, T.; Ihara, H.; Tsuchiya, Y.; Watanabe, Y.; Kumagai, Y.; Suematsu, M.; Motohashi, H.; Fujii, S.; Matsunaga, T.; Yamamoto, M.; Ono, K.; Devarie-Baez, N. O.; Xian, M.; Fukuto, J. M.; Akaike, T., Reactive cysteine persulfides and S-polythiolation regulate oxidative stress and redox signaling. *Proc. Natl. Acad. Sci. USA* **2014,** *111*, 7606-11.

23. Liu, R.; Yue, Z.; Tsai, C. C.; Shen, J., Assessing Lysine and Cysteine Reactivities for Designing Targeted Covalent Kinase Inhibitors. *J Am Chem Soc* **2019,** *141* (16), 6553-6560.

24. Mueller, E. G., Trafficking in persulfides: delivering sulfur in biosynthetic pathways. *Nat. Chem. Biol.* **2006,** *2*, 185-94.

25. Olson, K. R.; Gao, Y.; Arif, F.; Arora, K.; Patel, S.; DeLeon, E. R.; Sutton, T. R.; Feelisch, M.; Cortese-Krott, M. M.; Straub, K. D., Metabolism of hydrogen sulfide (H2S) and Production of Reactive Sulfur Species (RSS) by superoxide dismutase. *Redox Biol.* **2018,** *15*, 74-85.

26. Ono, K.; Akaike, T.; Sawa, T.; Kumagai, Y.; Wink, D. A.; Tantillo, D. J.; Hobbs, A. J.; Nagy, P.; Xian, M.; Lin, J.; Fukuto, J. M., Redox chemistry and chemical biology of H2S, hydropersulfides, and derived species: implications of their possible biological activity and utility. *Free Radic. Biol. Med.* **2014,** *77*, 82-94.

27. Ghosh, W.; Dam, B., Biochemistry and molecular biology of lithotrophic sulfur oxidation by taxonomically and ecologically diverse bacteria and archaea. *FEMS Microbiol. Rev.* **2009,** *33*, 999-1043.

28. Berg, J. S.; Schwedt, A.; Kreutzmann, A. C.; Kuypers, M. M.; Milucka, J., Polysulfides as intermediates in the oxidation of sulfide to sulfate by Beggiatoa spp. *Appl. Environ. Microbiol.* **2014,** *80*, 629-36.

29. Akaike, T.; Ida, T.; Wei, F. Y.; Nishida, M.; Kumagai, Y.; Alam, M. M.; Ihara, H.; Sawa, T.; Matsunaga, T.; Kasamatsu, S.; Nishimura, A.; Morita, M.; Tomizawa, K.; Nishimura, A.; Watanabe, S.; Inaba, K.; Shima, H.; Tanuma, N.; Jung, M.; Fujii, S.; Watanabe, Y.; Ohmuraya, M.; Nagy, P.; Feelisch, M.; Fukuto, J. M.; Motohashi, H., Cysteinyl-tRNA synthetase governs cysteine polysulfidation and mitochondrial bioenergetics. *Nat. Commun.* **2017,** *8*, 1177.

30. Kimura, H., Signalling by hydrogen sulfide and polysulfides via protein S-sulfuration. *Br. J. Pharmacol.* **2020,** *177*, 720-733.

31. Chauvin, J. R.; Griesser, M.; Pratt, D. A., Hydropersulfides: H-Atom Transfer Agents Par Excellence. *J. Am. Chem. Soc.* **2017,** *139*, 6484-6493.

32. Arakaki, K.; Uehara, A.; Higa-Nakamine, S.; Kakinohana, M.; Yamamoto, H., Increased expression of EGR1 and KLF4 by polysulfide via activation of the ERK1/2 and ERK5 pathways in cultured intestinal epithelial cells. *Biomed. Res.* **2020,** *41*, 119-129.

33. Hou, N.; Yan, Z.; Fan, K.; Li, H.; Zhao, R.; Xia, Y.; Xun, L.; Liu, H., OxyR senses sulfane sulfur and activates the genes for its removal in Escherichia coli. *Redox Biol.* **2019,** *26*, 101293.

34. Benchoam, D.; Cuevasanta, E.; Moller, M. N.; Alvarez, B., Hydrogen Sulfide and Persulfides Oxidation by Biologically Relevant Oxidizing Species. *Antioxidants (Basel)* **2019,** *8* (2).

35. Chen, W.; Pacheco, A.; Takano, Y.; Day, J. J.; Hanaoka, K.; Xian, M., A Single Fluorescent Probe to Visualize Hydrogen Sulfide and Hydrogen Polysulfides with Different Fluorescence Signals. *Angew. Chem. Int. Ed. Engl.* **2016,** *55*, 9993-6.

36. Doka, E.; Pader, I.; Biro, A.; Johansson, K.; Cheng, Q.; Ballago, K.; Prigge, J. R.; Pastor-Flores, D.; Dick, T. P.; Schmidt, E. E.; Arner, E. S.; Nagy, P., A novel persulfide detection method reveals protein persulfide- and polysulfide-reducing functions of thioredoxin and glutathione systems. *Sci. Adv.* **2016,** *2*, e1500968.

37. Hamid, H. A.; Tanaka, A.; Ida, T.; Nishimura, A.; Matsunaga, T.; Fujii, S.; Morita, M.; Sawa, T.; Fukuto, J. M.; Nagy, P.; Tsutsumi, R.; Motohashi, H.; Ihara, H.; Akaike, T., Polysulfide stabilization by tyrosine and hydroxyphenyl-containing derivatives that is important for a reactive sulfur metabolomics analysis. *Redox Biol.* **2019,** *21*, 101096.

38. Yin, G.; Niu, T.; Yu, T.; Gan, Y.; Sun, X.; Yin, P.; Chen, H.; Zhang, Y.; Li, H.; Yao, S., Simultaneous Visualization of Endogenous Homocysteine, Cysteine, Glutathione, and their Transformation through Different Fluorescence Channels. *Angew. Chem. Int. Ed. Engl.* **2019,** *58*, 4557-4561.

39. Thomas, B. E.; Kollman, P. A., An ab initio molecular orbital study of the first step of the catalytic mechanism of thymidylate synthase: the Michael addition of sulfur and oxygen nucleophiles.*J. Org. Chem.* **1995,** *60*, 8375-8381.

40. Awoonor-Williams, E.; Isley, W. C., 3rd; Dale, S. G.; Johnson, E. R.; Yu, H.; Becke, A. D.; Roux, B.; Rowley, C. N., Quantum Chemical Methods for Modeling Covalent Modification of Biological Thiols. *J. Comput. Chem.* **2020,** *41*, 427-438.

41. Nikoo, S.; Meister, P. J.; Hayward, J. J.; Gauld, J. W., An Assessment of Computational Methods for Calculating Accurate Structures and Energies of Bio-Relevant Polysulfur/Selenium-Containing Compounds. *Molecules* **2018,** *23*.

42. Smith, J. M.; Jami Alahmadi, Y.; Rowley, C. N., Range-Separated DFT Functionals are Necessary to Model Thio-Michael Additions. *J Chem. Theory Comput.* **2013,** *9*, 4860-5.

43. Orabi, E. A.; Peslherbe, G. H., Computational insight into hydrogen persulfide and a new additive model for chemical and biological simulations. *Phys. Chem. Chem. Phys.* **2019,** *21*, 15988-16004.

44. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox *Gaussian 09, Revision E.01*, Gaussian, Inc: Wallingford, CT, 2016.

45. Chai, J. D.; Head-Gordon, M., Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008,** *10* (44), 6615-20.

46. Pople, J. A.; Head-Gordon, M.; Raghavachari, K., Quadratic configuration interaction. A general technique for determining electron correlation energies. *J. Chem. Phys.* **1987,** *87*, 5968-5975.

47. E. D. Glendening, J., K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, P. Karafiloglou, C. R. Landis, and F. Weinhold *NBO 7.0*, Theoretical Chemistry Institute: University of Wisconsin, Madison, 2018.

48. Brito, J. A.; Sousa, F. L.; Stelter, M.; Bandeiras, T. M.; Vonrhein, C.; Teixeira, M.; Pereira, M. M.; Archer, M., Structural and functional insights into sulfide:quinone oxidoreductase. *Biochemistry* **2009,** *48*, 5613-22.

49. Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A., PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004,** *32*, W665-7.

50. Forman, H. J.; Davies, M. J.; Kramer, A. C.; Miotto, G.; Zaccarin, M.; Zhang, H.; Ursini, F., Protein cysteine oxidation in redox signaling: Caveats on sulfenic acid detection and quantification. *Arch. Biochem. Biophys.* **2017,** *617*, 26-37.

51. Rehder, D. S.; Borges, C. R., Cysteine sulfenic acid as an intermediate in disulfide bond formation and nonenzymatic protein folding. *Biochemistry* **2010,** *49*, 7748-55.

52. Jackson, M. R.; Melideo, S. L.; Jorns, M. S., Human sulfide:quinone oxidoreductase catalyzes the first step in hydrogen sulfide metabolism and produces a sulfane sulfur metabolite. *Biochemistry* **2012,** *51*, 6804-15.

53. Landry, A. P.; Ballou, D. P.; Banerjee, R., Modulation of Catalytic Promiscuity during Hydrogen Sulfide Oxidation. *ACS Chem. Biol.* **2018,** *13*, 1651-1658.

54. Libiad, M.; Yadav, P. K.; Vitvitsky, V.; Martinov, M.; Banerjee, R., Organization of the human mitochondrial hydrogen sulfide oxidation pathway. *J. Biol. Chem.* **2014,** *289*, 30901-10.

55. Griesbeck, C.; Schutz, M.; Schodl, T.; Bathe, S.; Nausch, L.; Mederer, N.; Vielreicher, M.; Hauska, G., Mechanism of sulfide-quinone reductase investigated using site-directed mutagenesis and sulfur analysis. *Biochemistry* **2002,** *41*, 11552-65.

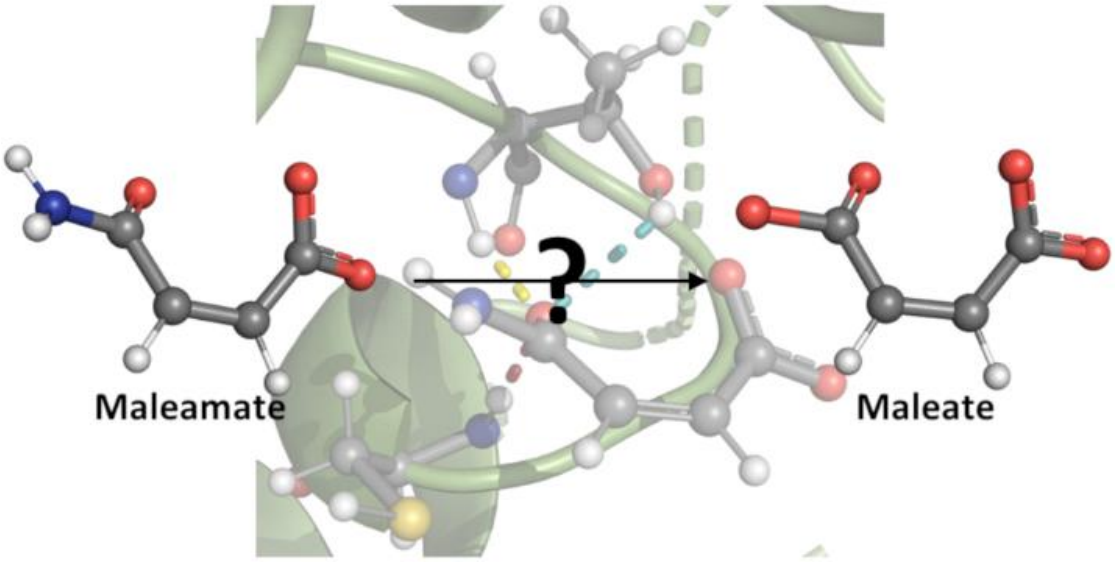
56. Marcia, M.; Ermler, U.; Peng, G.; Michel, H., The structure of Aquifex aeolicus sulfide:quinone oxidoreductase, a basis to understand sulfide detoxification and respiration. *Proc Natl Acad Sci USA* **2009,** *106*, 9625-30.

57. Huang, W.; Gauld, J. W., Tautomerization in the UDP-galactopyranose mutase mechanism: a DFT-cluster and QM/MM investigation. *J. Phys. Chem. B* **2012,** *116*, 14040-50.

58. Jackson, M. R.; Loll, P. J.; Jorns, M. S., X-Ray Structure of Human Sulfide:Quinone Oxidoreductase: Insights into the Mechanism of Mitochondrial Hydrogen Sulfide Oxidation. *Structure* **2019,** *27*, 794-805 e4.

Chapter 5:

A Multi-scale Computational Investigation on the Catalytic Mechanism of the Non-metallo Amidase Maleamate Amidohydrolase (NicF)



5.1 Introduction

Vitamins play several important metabolic roles including in catalysis and receptor binding.[1] For example, vitamin B_6 (pyridoxal-5'-phosphate) is important in Schiff base enzymatic catalysis,[2] while vitamin D_3 binds to receptors involved in bone and calcium metabolism.[3] Due to their insufficient production within the cell, vitamins must be obtained via our diets.[4] However, they must often be modified after ingestion in order to become biologically active. For instance, vitamin B_3 (nicotinate), critical for NAD^+ metabolism,[5] must be obtained from nicotinamide by the catalytic action of amide hydrolases such as nicotinamidases.[6] This catabolism is also essential for many physiologically important bacterial enzymes that require N-heteroaromatic derivatives as substrates.[7-8] Many of these bacteria are present in soil and/or are human pathogens,[9-10] such as *Bordetella bronchiseptica*,[11] *Streptococcus pneumoniae*,[12] and *Mycobacterium tuberculosis*.[13] As a result, there is tremendous interest from both agricultural and human health perspectives in understanding the catalytic chemistry of amide hydrolases.[11, 14-15]

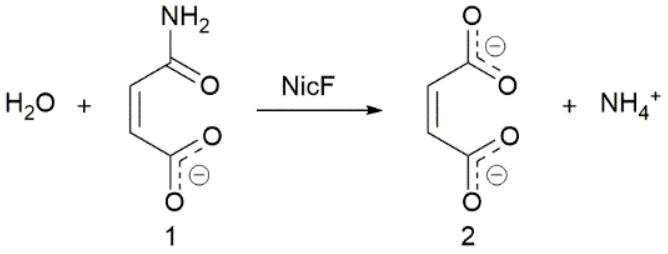
The majority of amide hydrolases contain a metal ion within their active site.[6, 16-17] These ions are important for both substrate recognition and binding[6] and furthermore, often have catalytic roles that exploit their Lewis acid nature.[18] Some well-known examples are the Zn(II)-dependent nicotinamidase[19] and thermolysin.[20] In both cases the metal helps bind and orient the amide substrate, as well as facilitate cleavage and hydrolysis of its amide bond.[18-19, 21] Notably, the substrates bind quite differently to the Zn(II) in these enzymes. In thermolysin, the substrate's amide oxygen binds directly to the Zn(II) whereas in nicotinamidase it is the substrate's pyridyl ring nitrogen.[20-21] Nevertheless, in both systems the Zn(II) acts as a Lewis acid.

Maleamate amidohydrolase (NicF) is a related member of the same enzyme family, with a central role in the nicotinate catabolic pathway of aerobic bacteria.[9] More specifically, it

catalyzes the conversion of maleamate to the fumarate-precursor, maleate (Scheme 5.1).[10] Based on gene studies in *Pseudomonas putida*, Jimenez et al.[9] proposed that, as in homologous enzymes such as nicotinamidases,[17] NicF may possess a catalytic triad composed of an aspartyl (Asp31), lysyl (Lys121), and cysteinyl (Cys154).

More recently, Kincaid et al.[10] obtained an X-ray crystal structure of an unbound NicF enzyme from *Bordetella bronchiseptica* (PDB ID: 3UAO) in which Cys150 had been mutated to a cysteine sulfenate (Cso). Based on the structure obtained, they also suggested that Asp29, Lys117, Cys/Cso150 likely form a catalytic triad. However, they further proposed that as no Zn(II) was observed to be present, it did not play a role in catalysis. Indeed, they performed kinetic studies and showed that the enzyme functioned efficiently without a metal ion. Thus, unlike the majority of amide hydrolases,[19] NicF appears to be a metal-independent (non-metallo) amidase.

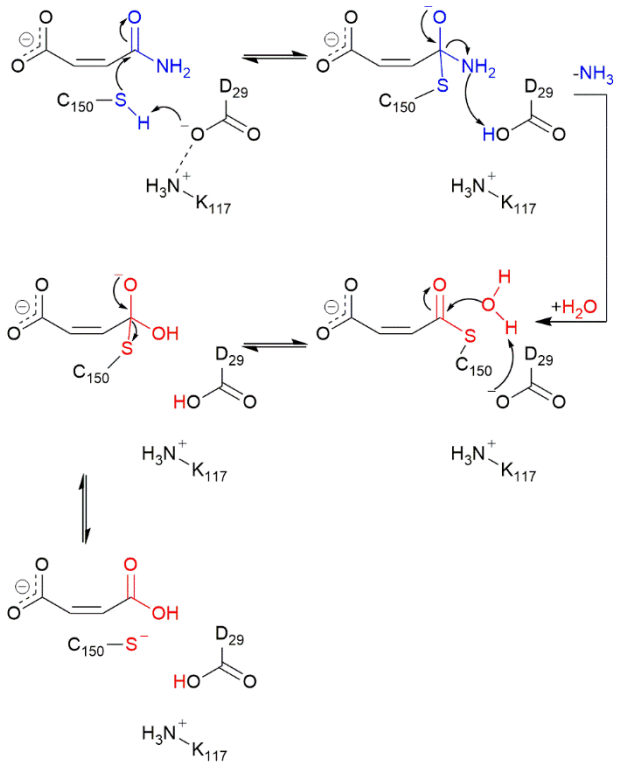
Scheme 5.1. The overall deamination reaction of maleamate (**1**) to maleate (**2**) catalyzed by the amidohydrolase NicF.[10]



As a result, they proposed the catalytic mechanism shown in Scheme 5.2. More specifically, the thiol of the active site cysteinyl (Cys150) is deprotonated, activated, by the R-group carboxylate (R–COO^-) of the nearby aspartyl (Asp29) that was observed to be within hydrogen bonding distance.[10] The resulting cysteinyl thiolate then attacks the carbonyl carbon (C_{carb}) of the maleamate substrate's amide bond. This gives rise to a tetrahedral oxyanion intermediate with a covalent C_{carb}–S_{Cys150} substrate-enzyme cross-link.

The now neutral Asp29 R-group carboxylic acid then protonates the substrate's amine group. This induces cleavage of the C_{carb}–NH_2 bond to give a thioester cross-linked enzyme-intermediate complex along with free NH_3. The latter is then replaced by a H_2O molecule. It has been suggested that the Asp29 is able to activate the water for hydrolysis of the thioester bond in part due to the presence of the nearby R-group of Lys117.[10] The incoming H_2O nucleophilically attacks the C_{carb} centre while transferring a proton to Asp29. The resulting tetrahedral oxyanion intermediate ultimately collapses to maleate with cleavage of the C_{carb}–S_{Cys150} bond. Meanwhile, Arg175 and Lys190' (not shown for simplicity) are thought to help stabilize the substrate's anionic charge.[10] However, a number of key questions still remain about the roles of active site residues and the catalytic mechanism of NicF. Furthermore, insights into its non-metallo mechanism can provide an opportunity to compare with the more common metalloamidases.

Scheme 5.2. Proposed mechanism for conversion of maleamate to monoanionic maleate catalyzed by NicF.[10]



In this study we have used a multi-scale computational approach to elucidate the catalytic mechanism of maleamate amidohydrolase from *Bordetella bronchiseptica* (NicF). In particular, molecular dynamics (MD) simulations and QM/MM methods have been used to investigate its substrate-bound active site, catalytic mechanism, and role(s) of the protein environment and key active site residues. Specific non-bonding and bonding interactions are examined with the use of the quantum theory of atoms in molecules (QTAIM). This analysis provides a detailed topological analysis of the active site giving insight into the key interactions required for stabilization, binding[22-23] and catalysis.[24] These interactions are examined throughout the catalytic mechanism *via* the intermediates.

5.2 Methods

Molecular Dynamics (MD) Simulations

MD simulations and model preparation were done using the Molecular Operating Environment (MOE)[25] program. The X-ray crystal structure of maleamate amidohydrolase (NicF) from *Bordetella bronchiseptica* complexed with an acetate ion (PDB ID: 3UAO) was used as a template from which to obtain the chemical model.[10] Specifically, each active site in the enzyme complex contained an acetate ion that was modified to the maleamate substrate. The cysteine–sulfenate (Cso150) was mutated to cysteine while hydrogen atoms were added using the default MOE method. The enzyme-substrate complex was solvated by a 4 Å spherical layer of water molecules (i.e., to a depth of at least two water molecules retaining all peptide chains). The resulting complex was restrained within the volume established by an ellipsoidal potential wall. The electrostatic and van der Waals potentials decayed smoothly by a damping functional factor with a final cutoff of 10 Å. The structure was then minimized using the AMBER99 force field until the root mean square gradient of the total energy fell below 0.42 $kJ \cdot mol^{-1} \cdot Å^{-2}$. The complex then underwent thermal relaxation at constant pressure and temperature. The Nosé-Poincaré thermostat[26] was coupled with the equations of motion, where a 2 fs time step was set for numerical

integration. The system was equilibrated at 150 K for 100 ps, after which the system was set to 300 K for an interval of 10 ns. Based on root mean square deviations (RMSD) and cluster analyses, a representative structure from the trajectory was then optimized using the AMBER99 force field. This MD protocol has been successfully applied in other enzymatic studies.[27-28]

QM/MM calculations and QTAIM analysis

All QM/MM calculations were performed using the ONIOM[29-37] formalism in the Gaussian 09[38] program. Optimized geometries and harmonic vibrational frequencies were obtained at the ONIOM(DFT/6-31G(d):AMBER96) level of theory where DFT=B3LYP or M06-2X, within a mechanical embedding (ME) formalism. That is, the reactive region (QM-layer) was described at the DFT/6-31G(d)[39-41] level of theory while the surrounding protein environment (MM-layer) was described using the AMBER[42] force field. To obtain more reliable relative energies, single-point calculations were done at the ONIOM(ωB97X-D/6-311G(2df,p):AMBER96) level of theory within an ME formalism for all structures with the electronic embedding (EE) formalism also being used for structures in the first half-reaction to examine the effects of its inclusion. Single-point calculations at the DFT/6-311+G(2d,2p) (DFT=B3LYP, M06-2X, ωB97XD) were also benchmarked and are shown in Appendix B (Appendix B5.5). This protocol has been successfully applied in other enzymatic studies.[27-28]

A suitable QM/MM chemical model was derived from the final optimized AMBER99 structure (see MD section above). All residues and waters up to 26 Å from the amide carbon of the maleamate substrate were included as the steric and electrostatic effects of the environment surrounding the active site can have important effects on the mechanism.[43] The reactive region included the maleamate substrate, R-groups of the proposed catalytic triad residues (Lys117, Asp29, and Cys150), partial R-groups of Lys190' and Arg175

which may play a role in stabilizing the substrate's $–COO^-$ group, and peptide backbones between Ala145–Thr146 (α-C_{145}–CO–NH–α-C_{146} including β-OH moiety of Thr146) and Gly149–Cys150 (α-C_{149}–CO–NH–α-C_{150}). All surrounding residues and waters were included in the MM-layer. The α-carbon of each residue in the MM-layer was kept fixed at its final MD optimized position. This computational approach has been successfully applied to related enzymes.[19] Optimized structures of all pathway intermediates and transition states were further subjected to QTAIM analysis to provide a detailed topology of the active site and key interactions therein at the single point level.[44]

5.3 Results and Discussion

5.3.1 The maleamate-bound active site

As noted above it has been proposed that in related amidohydrolases the oxyanion hole is composed of the backbone –NH– groups of Thr146 and Cys150,[10, 45] while in, for example, a nicotinamidase from *S. pneumoniae* (SpNic), other groups are also observed to be involved in substrate binding.[19] An MD simulation was performed on the substrate-bound enzyme complex in which Cys150 is in its neutral thiol state. As the thiol must be deprotonated to act as a nucleophile this complex can be considered as the pre-reactive complex, **PRC** (Appendix B5.1). Over the course of the 10.1 ns MD simulation, the backbone -NH- groups of Thr146 and Cys150, as well as the R-group hydroxyl of Thr146 were all observed to hydrogen bond with the substrate's O_{carb} center, even if only transiently. Hence, root mean square deviation (RMSD) analyses were performed on their interactions (i.e., the -NH-'s of Thr146, Cys150, and Thr146β-OH) with the substrate's key O_{carb} centre and are shown in Figure 5.1.

As can be seen in Figure 5.1A, the total RMSD for the O_{carb} centre and above groups involved in the putative oxyanion hole generally fluctuate over a reasonably narrow range of 0.26 ± 0.06 Å. The individual $O_{carb}\cdots X$ (X = HO-β_{Thr146}, HN_{Thr146}, and HN_{Cys150})

distances observed over the course of the MD are shown in Figure 5.1B-D, respectively. The strongest and most consistent interactions are observed for the $O_{carb}\cdots HN_{Thr146}$/$HN_{Cys150}$ interactions which both have an average length of approximately 1.88 Å over the course of the MD; Figure 5.1C and D, respectively. In contrast, the $O_{carb}\cdots$ HO-β_{Thr146} distance fluctuates from 1.96 to 3.43 Å. This reflects the observation that the $_{Thr146}\beta$-OH was observed during the simulation to intermittently shift and bind instead with the substrate's carboxyl group (Figure 5.1B).

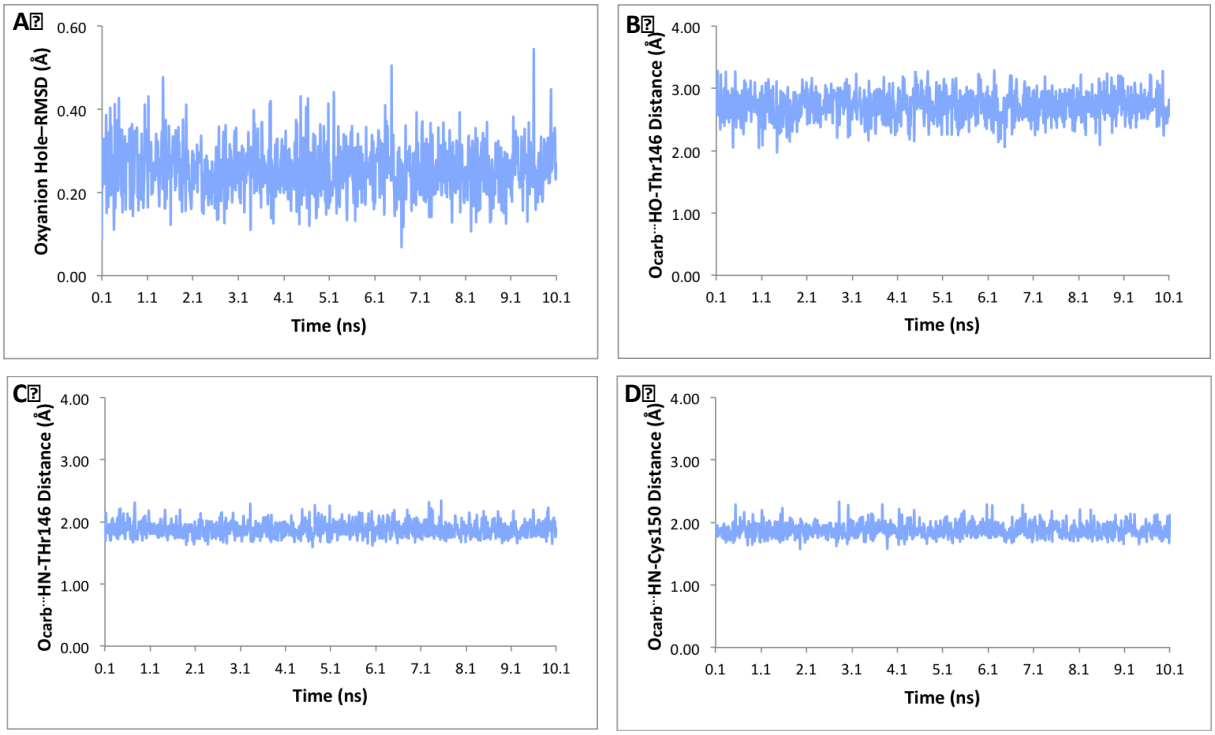


Figure 5.1. RMSD plots of the (A) oxyanion hole groups and the (B) $O_{carb}\cdots$HO-βThr146, (**C**) $O_{carb}\cdots$HN-Thr146, and (**D**) $O_{carb}\cdots$HN-Cys150 distances (Angstroms), obtained over a 10.1 ns MD simulation (See Computational Methods).

In the QM/MM optimized **PRC** structure (Appendix B5.1) the $O_{carb}\cdots HN_{Cys150}$ and $O_{carb}\cdots HN_{Thr146}$ distances are 3.04 and 2.15 Å respectively, with angles of 119.4º and 128.9º. This suggests that there is substantially less oxyanion hole stabilization in the **PRC** from hydrogen bonding. Indeed, concomitantly, the substrates C_{carb}-O_{carb} bond length is

1.25 Å; indicating that it has changed little upon binding (i.e., has not developed oxyanion character).

In the mechanism of NicF, the Cys150 sulfur acts as a nucleophile. Hence, the enzymatic mechanism can be considered to begin upon formation of the reactive complex, **RC**, wherein Cys150 is in its thiolate state. It is noted that the barrier for deprotonation of Cys150 thiol is low relative to the rest of the reaction and thus is not the rate-limiting step (Appendix B5.2). As described in the Introduction, in the experimentally proposed mechanism, the maleamate carboxylate is suggested to bind with Arg175 and Lys190'.[10] As seen in Figure 5.2 in **RC**, the substrate's carboxylate does indeed form strong hydrogen bonds with the R-groups of Arg175 and Lys190' with $_{Arg175}NH\cdots O$ and $_{Lys190'}N\cdots HO$ distances of 1.86 and 1.62 Å, respectively. It is noted that the amine of Lys190' has transferred a proton to the substrate's carboxyl thus forming maleamic acid, neutralizing both Lys190' and the substrate. Furthermore, the R-group of Arg175 also forms a moderately strong hydrogen bond to the $_{Thr146}$β-OH oxygen with $r(_{Arg175}NH_{\eta}\cdots O_{Thr146})$ = 1.83 Å. Meanwhile, the $_{Thr146}$β-OH group itself forms a strong hydrogen bond to the substrate's carbonyl oxygen with $r(_{Thr146}\beta\text{-}OH\cdots O_{carb})$ = 1.74 Å.

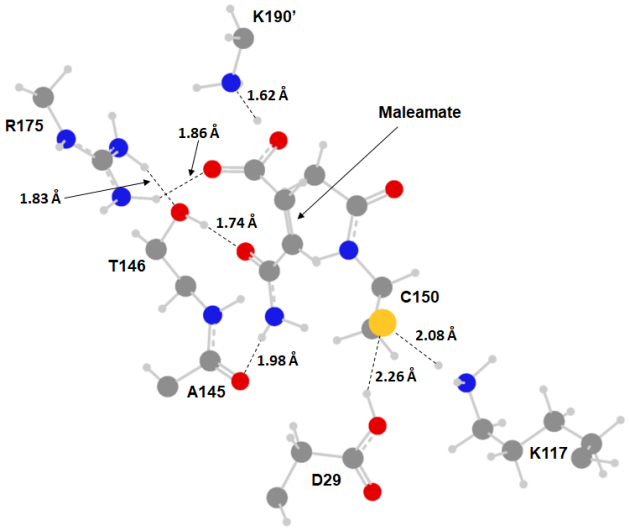


Figure 5.2. Optimized structure of the maleamic acid-bound active site (**RC**). Hydrogen bonding is shown with a dashed line.

5.3.2 Catalytic Mechanism of NicF

As noted above, it has been proposed[10] that the NicF mechanism proceeds via a nucleophilic addition-elimination sequence. More specifically, the first phase involves formation of a thioester cross-linked enzyme-intermediate complex with concomitant loss of the substrate's amine as ammonia. The second phase is hydrolysis of the enzyme-intermediate's thioester bond and formation of the maleamate product.

Stage 1: Formation of a thioester-enzyme intermediate. The NicF potential energy surface (PES) for Stage 1 of the overall mechanism is shown in Figure 5.3 while the corresponding optimized structures are shown in Figure 5.4.

The first step of the mechanism is nucleophilic attack of the Cys150 thiolate sulfur at the carbonyl carbon (C_{carb}) of the maleamic acid substrate. This occurs via **TS1** at a cost of 88.8 kJ mol^{-1} to give the tetrahedral oxyanion intermediate **IC1**. The latter lies somewhat high in

energy relative to **RC** at 80.0 kJ mol^{-1}. Using the present computational model, this is in fact the rate-limiting step of the first-half reaction and overall mechanism.

In **IC1**, as shown in Figure 5.4, a C_{carb}–S_{Cys150} bond has now been formed but with a long bond length of 2.08 Å. For comparison, using the same DFT methodology the C–S bond length in CH_3SH is 1.84 Å (data not shown). This bond formation occurs with significant concomitant lengthening of the C_{carb}–N bond by 0.17 Å to 1.51 Å. In addition, the C_{carb}=O_{carb} bond has lengthened slightly by 0.05 Å to 1.30 Å. That is, there is an increase in its alkoxy anion character as also evidenced in the significant increase in negative charge from –0.52 (in **RC**) to –0.76 on the O_{carb} centre (see also Appendix B5.4). As a result, the Thr146/Cys150–NH···O_{carb} hydrogen bonds have also shortened markedly to 1.97 and 2.05 Å, respectively (Figure 5.4). It is noted that while the Thr146β–OH···O_{carb} distance has lengthened considerably by 0.37 Å to 2.11 Å, it is still a moderately strong hydrogen bond. It is noted that if the substrate's carboxylate group is kept anionic (i.e., no proton transfer allowed from the R-groups of either Arg175 and Lys190') the negative charge on O_{carb} increases slightly to –0.79 (Appendix B5.4). This suggests that the R-groups of Lys190' and Arg175 may also play a part in stabilizing the oxyanionic intermediate albeit indirect from the oxyanion centre.

In the second and final step of Stage 1, the neutral R-carboxylic acid group of Asp29 protonates the amine leaving group of the intermediate **IC1**. This transfer occurs concurrently with proton transfer from the protonated R-group amine of Lys117 to the R-group carboxyl of Asp29. More importantly, protonation of the amine leaving group results in cleavage of the C_{carb}···N bond in **IC1** and formation of the thioester intermediate, **IC2**. This reaction step proceeds via **TS2** with a barrier of just 5.5 kJ·mol^{-1} relative to **IC1** (85.5 kJ·mol^{-1} with respect to **RC**) as shown in Figure 5.3. The thioester intermediate complex **IC2** lies higher in energy than **RC** by 45.5 kJ·mol^{-1} and thus, the first stage is predicted to

be endothermic. The $C_{carb}\cdots NH_3$ distance in **IC2** has lengthened even further to 1.74 Å while the C_{carb}–S_{Cys150} bond has shortened markedly by 0.18 Å to 1.90 Å (Figure 5.4).

The electrostatic potential density maps for smaller models of **IC1** and **IC2**, no enzyme residues or solvent included, are shown in Appendix B5.4 for when the substrates carboxyl group is neutralized (A) or anionic (B). This helps approximate the effects of hydrogen bonding by active site residues, specifically Arg175 and Lys190'. As can be seen, in both electrostatic potential maps for the neutral and anionic **IC2**-models there is less electron density around the C_{carb} center, especially in the neutral carboxyl group case. Likewise, calculated ESP charges indicate that the positive charge on C_{carb} has decreased in **IC2** while that of O_{carb} is now less negative. That is, its oxyanion character has decreased relative to that observed in **IC1**, as expected.

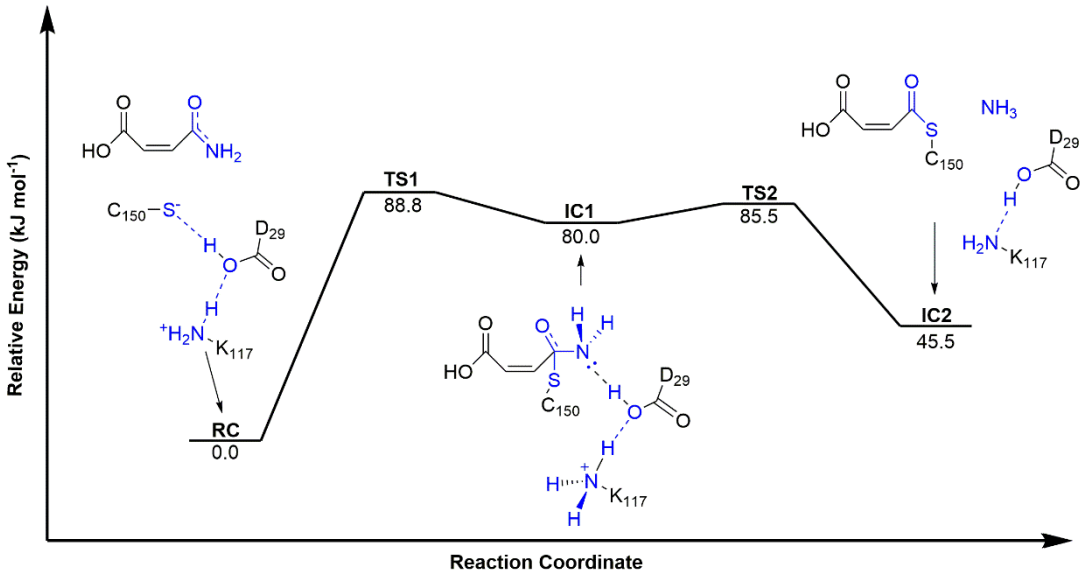


Figure 5.3. PES (kJ mol^{-1}) for the first stage of the overall mechanism of NicF obtained at the ONIOM(ωB97X-D/6-311+G(2df,p):AMBER96)//ONIOM(B3LYP/6-31G(d):AMBER96) level of theory within the ME formalism.

Using a similar computational methodology we have previously studied the related Zn(II)-dependent cysteine amidase SpNic.[19] Comparing the results for Stage 1 (formation of a thioester crosslinked intermediate with release of ammonia) for both SpNic and the present non-metallo NicF enzyme shows that both proceed *via* similar reaction steps. Notably, the relative energies of all stationary points along the Stage 1 PES are predicted to be higher in NicF, by 26-80 $kJ \cdot mol^{-1}$, but still enzymatically feasible. Additionally, the rate-limiting barrier in NicF is observed to be formation of the oxyanion intermediate (**IC1**), i.e., nucleophilic attack of the cysteinyl thiolate, rather than deamination as in SpNic.[19] This comparison suggests that while the zinc-dependent SpNic has lower reaction barriers, in agreement with experiment, NicF is catalytically viable in the absence of Zn(II).[10]

Stage 2: Product (maleic acid) formation via hydrolysis of the thioester-enzyme intermediate. For the second stage, as proposed for related enzymes such as SpNic, the now free NH_3 exits the active site and is replaced by a solvent H_2O. Thus, in the present computational investigation the NH_3 was removed and replaced by a H_2O. This provided a suitable reactive complex for the second stage of the NicF mechanism. The resulting **IC2** with NH_3 replaced by a H_2O molecule is denoted as **IC2'** herein. The optimized structures obtained for Stage 2 are shown in Figure 5.4, while electrostatic potential density maps are shown in Appendix B5.4 and the Stage 2 PES is shown in Figure 5.5.

In **IC2'** (Figure 5.4) the oxygen (O_w) of the incoming (co-substrate) water is positioned 2.68 Å from the intermediates C_{carb} centre. In addition, it acts as a hydrogen bond donor to both the R-group carboxylate of Asp29 and the backbone carbonyl oxygen of Ala145 with bond lengths of 2.17 and 1.95 Å, respectively. Notably, these interactions also help orient the water so that a lone pair on O_w is in the direction of the substrates C_{carb} centre as required for the first step of Stage 2.

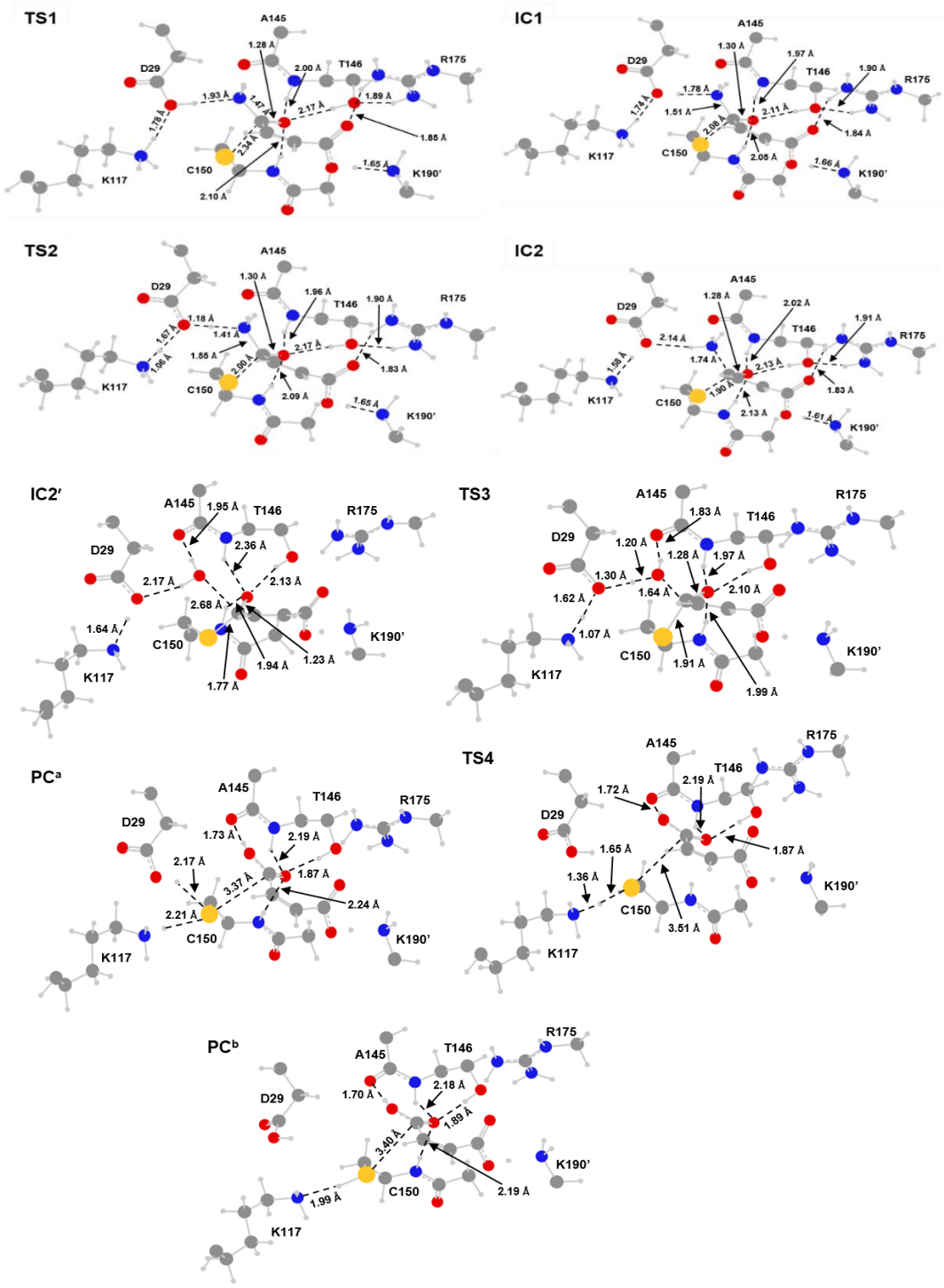


Figure 5.4. Optimized geometries of all transition states and intermediates for Stages 1 and 2 of the NicF mechanism, obtained at the ONIOM(ωB97X-D/6-311+G(2df,p):AMBER96)// ONIOM(B3LYP/6-31G(d):AMBER96) level of theory within the ME formalism.

The first step is nucleophilic attack of the incoming water's oxygen (O_w) at the intermediates C_{carb} centre. This proceeds via **TS3** with a barrier of 84.7 kJ mol^{-1} to give the energetically low-lying product complex **PC^a** (Figure 5.5). The latter complex lies 7.1 kJ mol^{-1} lower in energy than **IC2'**; that is, the reaction is exothermic. In **TS3** the hydrogen bonds the co-substrate water makes with Asp29 and Ala145 have both shortened (Figure 5.4). In particular, the $O_wH\cdots OOC_{Asp29}$ distance has decreased markedly to 1.30 Å while the O_w–H bond has lengthened to 1.20 Å. It is noted, that in **IC2'** the Asp29 carboxylate was neutral, however, this proton is transferred early to the amine group of Lys117 thus allowing Asp29 to accept a proton from the water (H_2O_w). This suggests that Asp29 plays a key role in helping to activate the water, by accepting a proton from the water, facilitating the nucleophilic attack of O_w on C_{carb}. In addition, the role of $_{Thr146}$OH as a member of the oxyanion hole appears to be further supported by its role in this step. In particular, the $O_{carb}\cdots$HO–Thr146 hydrogen bond distance shortens in **TS3** with respect to that observed in **IC2'** by 0.03 Å to 2.10 Å (Figure 5.4). That is, it again helps to stabilize the negative charge build up on O_{carb} during the reaction step.

In **PC^a** the enzyme is restored to the hypothesized cysteine thiolate state. However, for completeness, a proton transfer from the now neutral Asp29 R-group carboxylic acid to the thiolate of Cys150 to generate a neutral thiol. This step proceeds via **TS4** with a very low barrier of only 17.4 kJ·mol^{-1} to give the product complex **PC^b**. The latter lies 16.2 kJ mol^{-1} lower in energy than **IC2'** indicating that this reaction step is also exothermic.

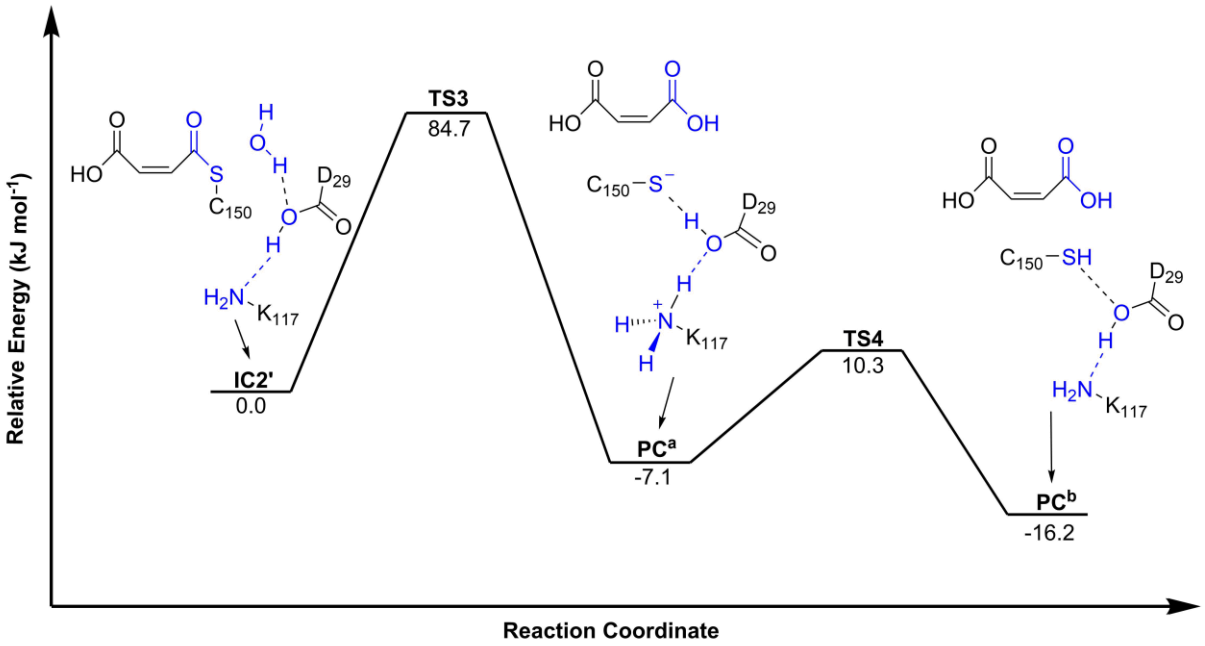


Figure 5.5. PES (kJ·mol^{-1}) for the second stage of the overall mechanism of NicF, obtained at the ONIOM(ωB97X-D/6-311+G(2df,p):AMBER96)//ONIOM(B3LYP/6-31G(d):AMBER96) level of theory within the ME formalism.

5.3.3 QTAIM Analysis

The active site and oxyanion hole of NicF were further characterized using the Quantum Theory of Atoms In Molecules (QTAIM) as described by Bader et al.[46] Molecular graphs of the intermediates identified along the pathway are shown in Figure 5.6 and show key interactions within each complex. It should be noted that the density (ρ_{BCP}) and Laplacian ($\nabla^2\rho_{BCP}$) at the bond critical point are used to characterize the interaction type. In particular, a closed-shell interaction is characterized by a $\rho_{BCP} < 0.10$ and positive $\nabla^2\rho_{BCP}$.[22] Meanwhile, for hydrogen bonds ρ_{BCP} is between 0.002 and 0.034, and $\nabla^2\rho_{BCP}$ lies between 0.024 and 0.139.[47]

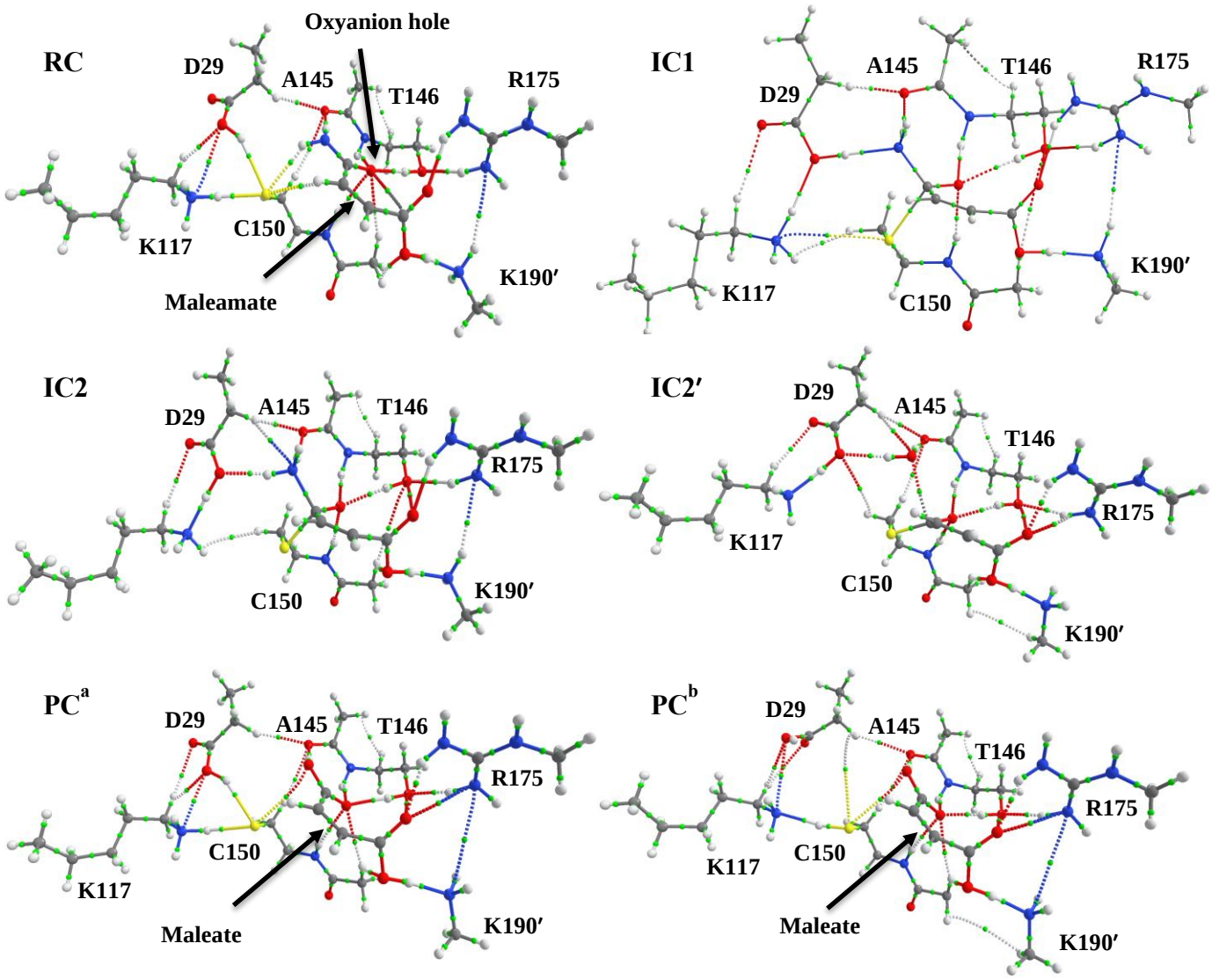


Figure 5.6. QTAIM molecular graphs calculated for all non-TS structures along both stages of the NicF mechanism. Bond critical points are shown as green dots.

In **RC**, a bond path exists between $_{Cys150}$S and the substrate's C_{amide} centre. As indicated by ρ_{BCP} and $\nabla^2\rho_{BCP}$, this is a weak closed-shell interaction suggesting that the two centres are suitably close to each other for further reaction. The thiolate group of Cys150 is stabilized via three different strength hydrogen bonds: specifically, with the protonated R-group amine of Lys117, neutral R-group carboxylic acid of Asp29, and the maleamate substrate's

own amide NH_2 group. In addition, there is further interaction with a C-H group of the substrate. Meanwhile the key amide oxygen of the maleamate substrate is involved in several hydrogen bonding interactions, specifically the Thr146 and Cys150 backbone –NH– groups and R-group hydroxyl of Thr146. These presumably aid formation of the oxyanion.

In intermediate **IC1** a weak covalent bond between $_{Cys150}S$ and C_{amide} has now been formed as indicated by the larger ρ_{BCP} and negative $\nabla^2\rho_{BCP}$. The bond path between the leaving ammonia and C_{amide} has now also changed in character from partial double to a single bond. As a consequence, the former amide nitrogen is now *sp*3 hybridized. Importantly, it is primed to accept a proton from the neutral R-group carboxylic acid of Asp29. Indeed, a bond path is now observed between Asp29-COOH and the substrate's nitrogen centre. It is noted that a weak hydrogen bond also forms between the substrate's leaving $–NH_2$ group and the backbone carbonyl oxygen of Ala145. Such an interaction would help to position the $-NH_2$ group and possibly also help enhance the basicity of its nitrogen centre. Concomitantly, the protonated R-group amine of Lys117 forms a strong hydrogen bond with the R-group carboxylic acid of Asp29. This suggests that a proton relay mechanism occurs to produce NH_3. Due to a slight structural rearrangement to accommodate the formation of the oxyanion hole, the hydrogen bonding distances between the substrate's amide oxygen (O_{amide}) and the putative oxyanion hole groups have decreased, thus helping to stabilize the oxyanion O_{amide}.

Upon formation of **IC2**, that is, protonation of the nitrogen followed by C—N bond cleavage to form ammonia, the same set of stabilizing interactions is observed for the oxyanion hole (Figure 5.6). It is noted that an interaction between the nitrogen of the leaving ammonia and C_{Carb} centre is observed. The ρ_{BCP} and $\nabla^2\rho_{BCP}$ values are indicative of a closed-shell, non-covalent, interaction. In **IC2'** the leaving NH_3 is replaced by a H_2O

molecule. As seen in Figure 5.6 this incoming H_2O forms hydrogen bonds to both the R-group carboxylic acid of Asp29 and the backbone carbonyl oxygen of Ala145. In addition, a weak interaction is observed with a CH of Asp29. These interactions help position the water such that a lone-pair of its oxygen is directed towards the C_{amide} centre of **IC2'**. That is, the water is now well positioned for reaction with the intermediate's thioester bond. Meanwhile, the oxyanion hole retains the same interactions observed in **IC2**.

As can be seen upon comparison of the two product complexes **PC^{a}** and **PC^{b}**, Cys150 regardless of whether it is in its thiolate or thiol state forms similar interactions. Notably, in both complexes no bond path is observed between $_{Cys150}$S and the products C_{acid} centre.

5.4 Conclusions

In the present study, we used multi-scale computational methods to gain insights into the catalytic mechanism of *Bordetella bronchiseptica* maleamate amidohydrolase (NicF) including substrate binding, oxyanion hole formation, and elucidation of the role of key active site residues. In particular, DFT, molecular dynamics (MD), ONIOM quantum mechanics/molecular mechanics (QM/MM), and QTAIM analyses have been complementarily applied.

The identity and role of active site residues involved in the putative oxyanion hole were examined using both MD and ONIOM(QM/MM). It has been suggested that this region helps facilitate the stabilization of the anionic character of O_{carb} substrate via $O_{carb}\cdots$HN–Thr146 and $O_{carb}\cdots$HN–Cys150 interactions. However, the present results suggest that Thr146β–OH may also play an important role in the oxyanion hole by hydrogen bonding with the substrate's O_{carb} centre along the reaction pathway.

The overall catalytic mechanism of NicF proceeds in two stages. In Stage 1, Cys150 begins in a thiolate state and performs a nucleophilic addition onto C_{amide} of the substrate maleamate. This is the rate-limiting step for this stage and the overall mechanism with a barrier of 88.8 kJ mol^{-1}. Subsequent formation of NH_3 by protonation of the substrate's amide nitrogen by the R-group carboxylic acid of Asp29, with concomitant cleavage of the C—NH_2 bond, proceeds with a low barrier of just 5.5 kJ mol^{-1} with respect to **IC1**. That is, loss of ammonia is step-wise within the current computational model. In Stage 2, hydrolysis of the C_{carb}–S_{Cys150} thioester bond in **IC2'**, occurs in one step within the present models with a barrier of 84.7 kJ·mol^{-1} to form the maleic acid product complex **PC^a**. That is, stage 2 proceeds *via* a concerted process.

Consequently, even though NicF lacks a metal for substrate placement and Lewis acid catalysis, it is still able to effectively catalyse amide bond hydrolysis and control positioning of reactants and active side residues and groups.

5.5 References

1. Bridwell-Rabb, J.; Drennan, C. L., Vitamin B12 in the spotlight again. *Curr. Opin. Chem. Biol.* **2017,** *37*, 63-70.

2. Lin, B. B.; Tian, G.; Liu, Y. J., Mechanistic insights into the gamma-elimination reaction of L-methionine catalyzed by methionine gamma-lyase (MGL). *Theor. Chem. Acc.* **2017,** *136*.

3. Hayashi, K.; Yasuda, K.; Sugimoto, H.; Ikushiro, S.; Kamakura, M.; Kittaka, A.; Horst, R. L.; Chen, T. C.; Ohta, M.; Shiro, Y.; Sakaki, T., Three-Step Hydroxylation of Vitamin D3 by a Genetically Engineered CYP105A1: Enzymes and Catalysis. *FEBS J.* **2010,** *277*, 3999-4009.

4. Falcone, T. D.; Kim, S. S.; Cortazzo, M. H., Vitamin K: Fracture Prevention and Beyond. *PM&R* **2011,** *3*, S82-7.

5. Sauve, A. A., NAD(+) and Vitamin B3: From Metabolism to Therapies. *J. Pharmacol. Exp. Ther.* **2008,** *324*, 883-893.

6. French, J. B.; Cen, Y.; Sauve, A. A.; Ealick, S. E., High-Resolution Crystal Structures of Streptococcus pneumoniae Nicotinamidase with Trapped Intermediates Provide Insights into the Catalytic Mechanism and Inhibition by Aldehydes. *Biochemistry-Us* **2010,** *49*, 8803-8812.

7. Fetzner, S.; Tshisuaka, B.; Lingens, F.; Kappl, R.; Hüttermann, J., Bacterial Degradation of Quinoline and Derivatives—Pathways and Their Biocatalysts. *Angew. Chem. Int. Ed. 37*, 576-597.

8. Taniguchi, H.; Sungwallek, S.; Chotchuang, P.; Okano, K.; Honda, K., A Key Enzyme of the NAD(+) Salvage Pathway in Thermus thermophilus: Characterization of Nicotinamidase and the Impact of Its Gene Deletion at High Temperatures. *J. Bacteriol.* **2017,** *199*.

9. Jimenez, J. I.; Canales, A.; Jimenez-Barbero, J.; Ginalski, K.; Rychlewski, L.; Garcia, J. L.; Diaz, E., Deciphering the Genetic Determinants for Aerobic Nicotinic acid Degradation: The nic Cluster from Pseudomonas putida KT2440. *Proc. Natl. Acad. Sci. U SA* **2008,** *105*, 11329-11334.

10. Kincaid, V. A.; Sullivan, E. D.; Klein, R. D.; Noel, J. W.; Rowlett, R. S.; Snidert, M. J., Structure and Catalytic Mechanism of Nicotinate (Vitamin B3) Degradative Enzyme Maleamate Amidohydrolase from Bordetella bronchiseptica RB50. *Biochemistry* **2012,** *51*, 545-554.

11. Yevsa, T.; Ebensen, T.; Fuchs, B.; Zygmunt, B.; Libanova, R.; Gross, R.; Schulze, K.; Guzman, C. A., Development and characterization of attenuated metabolic mutants of Bordetella bronchiseptica for applications in vaccinology. *Environ. Microbiol.* **2013,** *15*, 64-76.

12. Galeazzi, L.; Bocci, P.; Amici, A.; Brunetti, L.; Ruggieri, S.; Romine, M.; Reed, S.; Osterman, A. L.; Rodionov, D. A.; Sorci, L.; Raffaelli, N., Identification of Nicotinamide

Mononucleotide Deamidase of the Bacterial Pyridine Nucleotide Cycle Reveals a Novel Broadly Conserved Amidohydrolase Family. *J. Biol. Chem.* **2011,** *286*, 40365-40375.

13. Boshoff, H. I.; Xu, X.; Tahlan, K.; Dowd, C. S.; Pethe, K.; Camacho, L. R.; Park, T. H.; Yun, C. S.; Schnappinger, D.; Ehrt, S.; Williams, K. J.; Barry, C. E., 3rd, Biosynthesis and recycling of nicotinamide cofactors in mycobacterium tuberculosis. An essential role for NAD in nonreplicating bacilli. *J. Biol. Chem.* **2008,** *283*, 19329-41.

14. Zhang, Y.; Mitchison, D., The curious characteristics of pyrazinamide: a review. *Int. J. Tuber. Lung Dis.* **2003,** *7*, 6-21.

15. Rodionova, I. A.; Zuccola, H. J.; Sorci, L.; Aleshin, A. E.; Kazanov, M. D.; Ma, C.-T.; Sergienko, E.; Rubin, E. J.; Locher, C. P.; Osterman, A. L., Mycobacterial Nicotinate Mononucleotide Adenylyltransferase Structure, Mechanism and Implications for Drug Discovery. *J. Biol. Chem.* **2015,** *290*, 7693-7706.

16. Luo, H. B.; Zheng, H. P.; Zimmerman, M. D.; Chruszcz, M.; Skarina, T.; Egorova, O.; Savchenko, A.; Edwards, A. M.; Minor, W., Crystal structure and molecular modeling study of N-carbamoylsarcosine amidase Ta0454 from Thermoplasma acidophilum. *J. Struct. Biol.* **2010,** *169*, 304-311.

17. Du, X. L.; Wang, W. R.; Kim, R.; Yakota, H.; Nguyen, H.; Kim, S. H., Crystal Structure and Mechanism of Catalysis of a Pyrazinamidase from Pyrococcus horikoshii. *Biochemistry* **2001,** *40*, 14166-14172.

18. Smith, B. C.; Anderson, M. A.; Hoadley, K. A.; Keck, J. L.; Cleland, W. W.; Denu, J. M., Structural and kinetic isotope effect studies of nicotinamidase (Pnc1) from Saccharomyces cerevisiae. *Biochemistry* **2012,** *51*, 243-256.

19. Ion, B. F.; Kazim, E.; Gauld, J. W., A Multi-Scale Computational Study on the Mechanism of Streptococcus pneumoniae Nicotinamidase (SpNic). *Molecules* **2014,** *19*, 15735-53.

20. Wu, R. B.; Hu, P.; Wang, S. L.; Cao, Z. X.; Zhang, Y. K., Flexibility of Catalytic Zinc Coordination in Thermolysin and HDAC8: A Born-Oppenheimer ab Initio QM/MM Molecular Dynamics Study. *J. Chem. Theory Comput.* **2010,** *6*, 337-343.

21. Fyfe, P. K.; Rao, V. A.; Zemla, A.; Cameron, S.; Hunter, W. N., Specificity and Mechanism of Acinetobacter baumanii Nicotinamidase: Implications for Activation of the Front-Line Tuberculosis Drug Pyrazinamide. *Angew. Chem. Int. Edit.* **2009,** *48*, 9176-9179.

22. Astani, E. K.; Heshmati, E.; Chen, C. J.; Hadipour, N. L., A Theoretical Study on the Characteristics of the Intermolecular Interactions in the Active Site of Human Androsterone Sulphotransferase: DFT Calculations of NQR and NMR Parameters and QTAIM Analysis. *J Mol Graph Model* **2016,** *68*, 14-22.

23. Vega-Hissi, E. G.; Tosso, R.; Enriz, R. D.; Gutierrez, L. J., Molecular Insight into the Interaction Mechanisms of Amino-2H-Imidazole Derivatives With BACE1 Protease: A QM/MM and QTAIM Study. *Int. J. Quantum Chem.* **2015,** *115*, 389-397.

24. Dokainish, H. M.; Simard, D. J.; Gauld, J. W., A Pseudohypervalent Sulfur Intermediate as an Oxidative Protective Mechanism in the Archaea Peroxiredoxin Enzyme ApTPx. *J Phys. Chem. B* **2017,** *121*, 6570-6579.

25. *Molecular Operating Environment (MOE)*, 2013.08; Chemical Computing Group Inc.: Montréal, QC, Canada, 2015.

26. Bond, S. D.; Leimkuhler, B. J.; Laird, B. B., The Nosé-Poincaré Method for Constant Temperature Molecular Dynamics. *J. Comput. Phys.* **1999,** *151* (1), 114-134.

27. Ion, B. F.; Bushnell, E. A. C.; De Luna, P.; Gauld, J. W., A Molecular Dynamics (MD) and Quantum Mechanics/Molecular Mechanics (QM/MM) Study on Ornithine Cyclodeaminase (OCD): A Tale of Two Iminiums. *Int. J. Mol. Sci.* **2012,** *13*, 12994-13011.

28. Bushnell, E. A. C.; Erdtman, E.; Llano, J.; Eriksson, L. A.; Gauld, J. W., The First Branching Point in Porphyrin Biosynthesis: A Systematic Docking, Molecular Dynamics and Quantum Mechanical/Molecular Mechanical Study of Substrate Binding and Mechanism of Uroporphyrinogen-III Decarboxylase. *J. Comput. Chem.* **2011,** *32*, 822-834.

29. Bearpark, M. J.; Ogliaro, F.; Vreven, T.; Boggio-Pasqua, M.; Frisch, M. J.; Larkin, S. M.; Morrison, M.; Robb, M. A., CASSCF Calculations for Photoinduced Processes in Large Molecules: Choosing When to Use the RASSCF, ONIOM and MMVB Approximations. *J. Photochem. Photobiol., A* **2007,** *190*, 207-227.

30. Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J., A New ONIOM Implementation in Gaussian98. Part I. The Calculation of Energies, Gradients, Vibrational Frequencies and Electric Field Derivatives. *J. Mol. Struct.* **1999,** *461*, 1-21.

31. Humbel, S.; Sieber, S.; Morokuma, K., The IMOMO Method: Integration of Different Levels of Molecular Orbital Approximations for Geometry Optimization of Large Systems: Test for n-butane Conformation and S(N)2 Reaction: RCl + Cl-. *J. Chem. Phys.* **1996,** *105*, 1959-1967.

32. Maseras, F.; Morokuma, K., IMOMM – A New Integrated Ab Initio Plus Molecular Mechanics Geometry Optimization Scheme of Equilibrium Structures and Transition States. *J. Comput. Chem.* **1995,** *16*, 1170-1179.

33. Morokuma, K.; Musaev, D. G.; Vreven, T.; Basch, H.; Torrent, M.; Khoroshun, D. V., Model Studies of the Structures, Reactivities, and Reaction Mechanisms of Metalloenzymes. *IBM J. Res. Dev.* **2001,** *45*, 367-395.

34. Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K., ONIOM: A Multilayered Integrated MO+MM Method for Geometry Optimizations and Single Point Energy Predictions. A Test for Diels-Alder Reactions and Pt(P(t-Bu)(3))(2)+H2 Oxidative Addition. *J. Phys. Chem.* **1996,** *100*, 19357-19363.

35. Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J., Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM. *J. Chem. Theory Comput.* **2006,** *2*, 815-826.

36. Vreven, T.; Morokuma, K., On the Application of the IMOMO (Integrated Molecular Orbital Plus Molecular Orbital) Method. *J. Comput. Chem.* **2000,** *21*, 1419-1432.

37. Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J., Geometry Optimization with QM/MM, ONIOM, and Other Combined Methods. I. Microiterations and Constraints. *J. Comput. Chem.* **2003,** *24*, 760-769.

38. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, M. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09, Revision D.01*, Gaussian, Inc.: Wallingford CT, 2009.

39. Becke, A. D., Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993,** *98*, 5648-5652.

40. Becke, A. D., A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993,** *98*, 1372-1377.

41. Lee, C. T.; Yang, W. T.; Parr, R. G., Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. *Phys. Rev. B* **1988,** *37*, 785-789.

42. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005,** *26*, 1668-1688.

43. Hu, L. H.; Soderhjelm, P.; Ryde, U., On the Convergence of QM/MM Energies. *J. Chem. Theory Comput.* **2011,** *7*, 761-777.

44. Keith, T. A. *AIMAll (Version 17.11.14)*, TK Gristmill Software: Overland Park KS, USA, 2017.

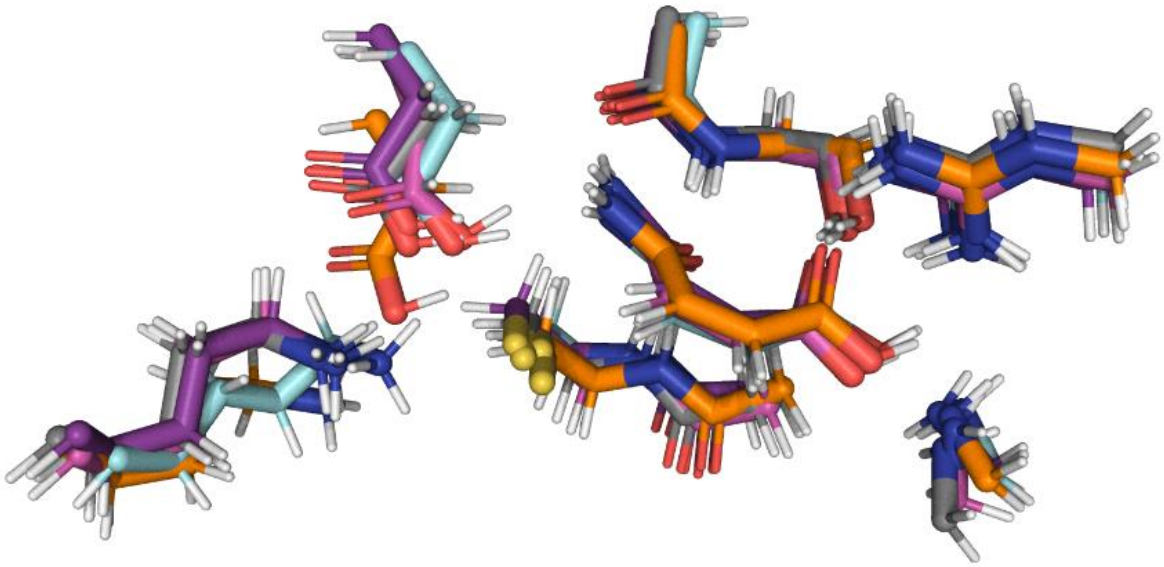
45. French, J. B.; Cen, Y.; Vrablik, T. L.; Xu, P.; Allen, E.; Hanna-Rose, W.; Sauve, A. A., Characterization of nicotinamidases: steady state kinetic parameters, classwide inhibition by nicotinaldehydes, and catalytic mechanism. *Biochemistry* **2010,** *49*, 10421-39.

46. Bader, R. F. W., *Atoms in molecules : a quantum theory*. Clarendon Press: Oxford ; New York, 1990; p xviii, 438 p.

47. Koch, U.; Popelier, P. L. A., Characterization of C-H-O Hydrogen-Bonds on the Basis of the Charge-Density. *J. Phys. Chem.* **1995,** *99*, 9747-9754.

Chapter 6:

The Importance of Informed MD Pose Selection for QM/MM Calculations: the Maleamate Amidohydrolase (NicF) Case Study



6.1 Introduction

Computational chemistry has been proven to be an extremely powerful tool for the study of catalysis[1-3] by, for instance, providing atomistic-level insights into chemical species along reaction pathways. In particular, the quantum mechanics/molecular mechanics (QM/MM) method has been shown to be particularly useful and is now arguably the preferred approach.[4-7] Recent advances in computing power have driven advancements in computational enzymology allowing for more complex analyses. In particular, a combined QM/MM and molecular dynamics (MD) multiscale approach is frequently used and is a powerful tool to study many features of enzyme mechanisms.[8-10] Molecular dynamics has also made significant advances due to implementation of GPU-processing capabilities. It is now not uncommon to see microsecond scale simulations.[11-12]

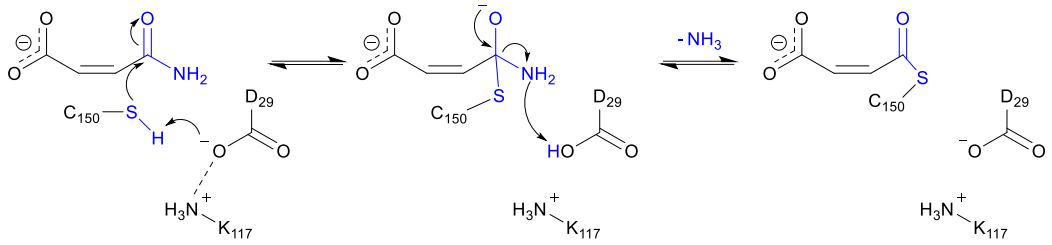
In general, a similar protocol is employed to study enzymes in the field of computational enzymology.[13-15] Initially, a protein structure of interest is either built from homology modeling software or is obtained from an online database such as the Protein Data Bank (PDB). The correct substrate is then added by docking if not present in the crystal structure. A proper docked pose is selected for further relaxation by MD. From the thousands of structures sampled, a limited selection must be made to conduct QM/MM calculations to study a mechanism of interest. It is at this stage that a computational chemist encounters the largest hurdle. Selecting any number of poses that can be considered catalytically active is a difficult task lacking a black-box approach for initial structure selection. Some groups have looked at iterative selection of poses (eg. after every nanosecond),[16] others select the last structure,[17] and others will select a structure from the most populated cluster.[18] Recent efforts have even turned to machine learning algorithms to pick catalytically relevant poses generated from molecular dynamics rather than relying on visual examination.[19]

The timescales accessible to molecular dynamics allow for what has been previously termed “instantaneous disorder”.[18] Throughout several picoseconds, the global structure of the enzyme does not change much, however even small structural changes in the active site have significant effects on the calculated barriers. For example, in the case of HIV-1 protease, the reaction barrier changes by as much as 40 kcal mol^{-1} depending on the selected pose.[18] Thus, selecting structures that are catalytically relevant is a significant challenge.

Each of the methods above have one commonality: the use of chemical intuition is avoided. Several types of cluster analyses may be employed along the proposed reaction mechanisms to study the chemistry relevant to the system of interest. Using this approach, key catalytic features can be studied with focus. We use a model system to compare these pose selection strategies and the effect on the calculated barrier. The deamination reaction catalyzed by maleamate amidohydrolase (NicF) was chosen as the model system.[20] We have previously elucidated the mechanism for this enzyme and it contains several bond-breaking, bond-forming, and proton transfer processes, thus possessing sufficient complexity and diversity of chemical processes.

NicF is an amide hydrolase with a central role in the nicotinate catabolic pathway of aerobic bacteria. More specifically, it catalyzes the conversion of maleamate to the fumarate precursor, maleate. Based on previous studies,[21] NicF contains a catalytic triad composed of aspartyl, lysyl, and cysteinyl. Experimental evidence suggests stabilization of the oxyanion drives the reaction. Many pathogenic and soil bacteria such as *Streptococcus pneumoniae*,[22] and *Mycobacterium tuberculosis*[23] utilize this pathway for carbon generation. Therefore, gaining catalytic insight into the function of NicF has significant impact on human health.

Scheme 6.1. Formation of the acyl-enzyme intermediate. The catalytic cysteinyl-150 is deprotonated and attacks the amide oxygen to form a tetracoordinate intermediate. Ammonia is then lost to generate the sulfoxide.



In this study, we calculated the reaction barriers from 6 different snapshots from MD simulations and compared the barriers to those observed in our previous study. The system studied is shown in Scheme 6.1. The X-ray structure was selected, as well as a structure every 2 ns, including the final structure. Furthermore, we also compared the reaction pathways and barriers in these structures for a stationary MM layer. Our results indicate that a calculated mechanism depends on the chosen structure and barriers have large variations, even when most of the protein is kept stationary.

6.2 Methods

Molecular Dynamics Simulations

MD simulations were conducted as we have described previously (see **Chapter 5**).[20] In summary, the Molecular Operating Environment (MOE)[24] program was used to prepare structures and conduct MD simulations. The X-ray crystal structure of maleamate amidohydrolase complexed with active site-bound acetate (PDB ID: 3UAO) was used as a template.[21] Each acetate ion was modified to the maleamate substrate. The resulting enzyme-substrate complex was solvated by a 4 Å spherical layer of water molecules within a restraining volume established by an ellipsoidal potential wall. The structure was then minimized using the AMBER99 force field until the root mean square gradient of the total

energy fell below 0.42 $kJ \cdot mol^{-1} \cdot Å^{-2}$. The complex then underwent thermal relaxation at constant pressure and temperature. The Nosé-Poincaré thermostat[25] was coupled with the equations of motion, where a 2 fs time step was set for numerical integration. The system was equilibrated at 150 K for 100 ps, after which the system was set to 300 K for an interval of 10 ns. Based on root mean square deviations (RMSD) and cluster analyses, a representative structure from the trajectory was then optimized using the AMBER14 force field.

QM/MM calculations

All QM/MM calculations were performed using the ONIOM[26-34] formalism in the Gaussian 09[35] program. Optimized geometries and harmonic vibrational frequencies were obtained at the ONIOM(B3LYP/6-31G(d):AMBER96) level of theory within a mechanical embedding (ME) formalism. This level of theory was chosen as it has been previously shown that it is successful for this reaction.[20] To obtain more reliable relative energies, single-point calculations were done at the ONIOM(ωB97X-D/6-311G(2df,p):AMBER96) level of theory within an ME formalism.

Structures from the MD simulation were selected every 2 ns as a method of random pose selection. In total, 6 snapshots were compared to a structure selected by cluster analysis. The reactive region included the maleamate substrate, R-groups of the proposed catalytic triad residues (Lys117, Asp29, and Cys150), partial R-groups of Lys190' and Arg175 which may play a role in stabilizing the substrate's $–COO^-$ group, and peptide backbones between Ala145–Thr146 (α-C_{145}–CO–NH–α-C_{146} including β-OH moiety of Thr146) and Gly149–Cys150 (α-C_{149}–CO–NH–α-C_{150}). All surrounding residues and waters were included in the MM-layer. The α-carbon of each residue in the MM-layer was kept fixed at its final MD optimized position. This computational approach has been successfully applied to related enzymes.[36]

6.3 Results and Discussion

Predicting enzymatic reactions is particularly challenging since experimentally, it can be difficult to track unstable intermediates. However, obtaining barriers from computational studies that are reflective of those from experiment is a difficult task. Usually, a structure is selected for computational study based on a previously proposed mechanism. Ideally, selected structures position the substrate in such a way to enable the proposed pathway. For example, in acid-base catalysis, there must be an acid or base positioned relatively close to the substrate and be in proper hydrogen bonding orientation for the reaction to favourably proceed. Some of these key interactions may be transient and thus only occur sporadically throughout simulations. Differentiating between structures where these interactions exist requires in-depth analyses (for example, clustering). The goal of this study was to identify the importance of making an informed selection from MD simulations for mechanistic studies using a QM/MM approach.

The first half of the NicF mechanism was an ideal system to study due to the significant complexity of chemical processes it possesses: proton transfer pathways, bond formation, and bond breaking. An overview of this mechanism is shown in Figure 6.1. In the first step, hybridization of the amide nitrogen changes to facilitate proton transfer from Asp29. This enables nucleophilic attack of the cysteinyl thiolate to form **IC1** (Figure 6.1A). Ammonia then acts as a leaving group to form the enzyme-substrate complex in **IC2**. Hydrolysis occurs in the second half of the mechanism whereby an activated water molecule is used to regenerate the native enzyme (see **Chapter 5** for more detail).

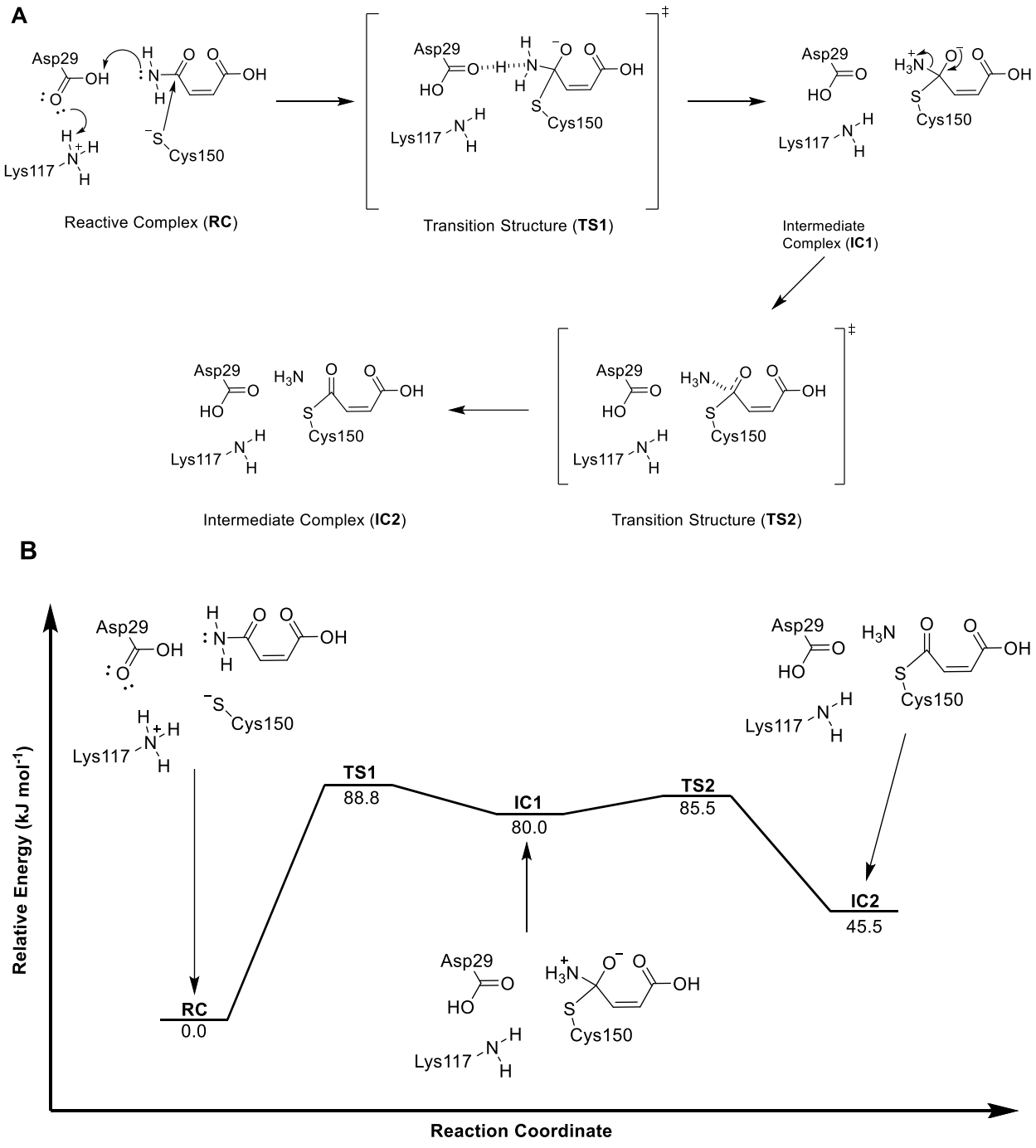


Figure 6.1. (A) Proposed mechanism for the elimination of ammonia and formation of the acyl-enzyme complex in NicF. (B) Potential energy surface of the proposed reaction pathway for the cluster-selected pose. Adapted from Ref. 20.

Structures for QM/MM calculations were selected iteratively every 2 ns and potential energy surfaces were constructed for each of these, in addition to the x-ray structure. The motivation behind including the x-ray structure was to demonstrate whether MD was required to obtain a starting structure. All structures were compared to the cluster-selected pose from **Chapter 5** (total of 7 starting structures and surfaces). Each was optimized to obtain the reactive complex (**RC**). Figure 6.2 shows an overlay of each structure for comparison. The 4 ns, 6 ns, 8 ns, and 10 ns structures have very similar active site orientations and are thus predicted to give similar potential energy surfaces. However, this may not be the case due to contributions from the MM environment. Notably, there are some differences in the orientation of Lys29 and Lys190′ in the 8 ns **RC**. Furthermore, in the 2 ns **RC**, we observe a very different active site orientation compared to the other structures. The active site is considerably shifted and the catalytic Cys150 is oriented away from the substrate and is positioned in a hydrogen bonding orientation to Lys117, suggesting an unproductive complex will be formed (previously termed **PRC**, see Appendix B).

Distances between the Cys150 thiolate and maleamate amide carbon as well as $_{Asp29}$COOH and maleamate nitrogen are shown in Table 1. In the final MD structure, $_{Asp29}$COOH···N_{amide} (Å) is too long after optimization (Table 1) for the proton to influence a hybridization change of the nitrogen and facilitate the attack of the Cys150 thiolate on the amide carbon. Attempts to obtain the first intermediate complex (**IC1**) always resulted in reformation of the **RC** indicating the structure is thermodynamically trapped. As a result, further calculations were not done on this complex. Importantly, examination of these key distances along the proposed mechanism have outlined a poor structure, but it cannot be used to determine the best structures especially when these distances are similar (*i.e.*, are similar to the cluster pose). This acts as a good initial filter, but further analysis is required.

Thus, the proposed mechanism was followed along the same potential energy surface for the 2, 4, 6, 8 ns poses to determine other differences that arise due to poor structure choice.

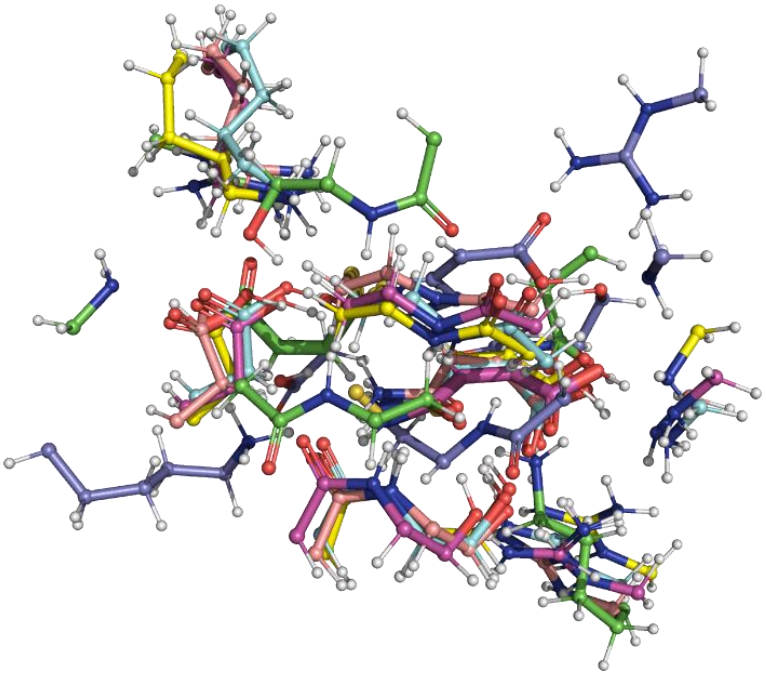


Figure 6.2. Overlay of the optimized reactant complexes for Cluster analysis pose (dark blue), after 2 ns MD (green), after 4 ns MD (light blue), after 6 ns MD (dark pink), after 8 ns MD (yellow), and the final MD structure (light pink).

Table 6.1. $_{Cys150}S^{-}\cdots C_{amide}$ (Å) and $_{Asp29}H\cdots N_{amide}$ (Å) distances along the first step of the proposed reaction pathway for acyl-enzyme formation in NicF. Distances are shown for each structure analyzed in this study.

Structure	**$_{Cys150}S\cdots C_{amide}$ (Å)**	**$_{Asp29}H\cdots N_{amide}$ (Å)**
Cluster pose	3.64	3.59
X-ray	5.60	Transferred
2 ns	3.55	3.90
4 ns	3.22	3.40
6 ns	3.19	3.48
8 ns	3.58	3.87
Final	3.77	5.08

In addition, the x-ray structure did not optimize to a **RC** that is comparable to other structures (Figure 6.3). The catalytic cysteinyl residue (Cys150) is not well positioned to attack the amide carbon of the maleamate substrate and is instead hydrogen bonding to $_{Lys117}NH_3^+$. This resulted in a proton transfer to $_{Cys150}S^-$ to form a thiol and fall into an unreactive complex ($_{Cys150}S\cdots C_{amide}$ (Å) 5.6 Å (Table 1)). These results suggest that in this structure, the active site was positioned such that the proposed mechanistic pathway is a large barrier to overcome compared to neutralization of the thiolate. Instead, during optimization, the complex fell into a previously characterized pre-reactive complex (**PRC**), as predicted from the initial structure.[20] Further calculations on this resulting complex were not done since there is no pathway leading to **IC2** due to the large barrier associated with this conversion. Attempted scans to promote **IC1** formation were unsuccessful and always fell back to **PRC**. This may have been due to the absence of a minimization protocol on this structure. The enzyme was crystallized with a mutated cysteine residue, S-hydroxycysteine (CSO), and an acetate ion in the active site rather than the maleamate substrate. To restore the native state, acetate was modified to maleamate and CSO was changed back to the thiolate *in silico* to model the reaction surface more accurately. By simply taking the x-ray structure without applying a minimization protocol, there was no opportunity given for the structure to reach an equilibrated state and optimize hydrogen bonding. This is amplified by the fact there were structural alterations made that introduced the potential for more hydrogen bonds to exist. The substrate was not given a chance to re-orient and these inconsistencies were carried into QM/MM calculations. This indicates that an MD simulation, or at least a minimization protocol, is beneficial in cases where the crystal structure is experimentally mutated to allow the modified structure to relax and allow for proper hydrogen bonding networks to form.

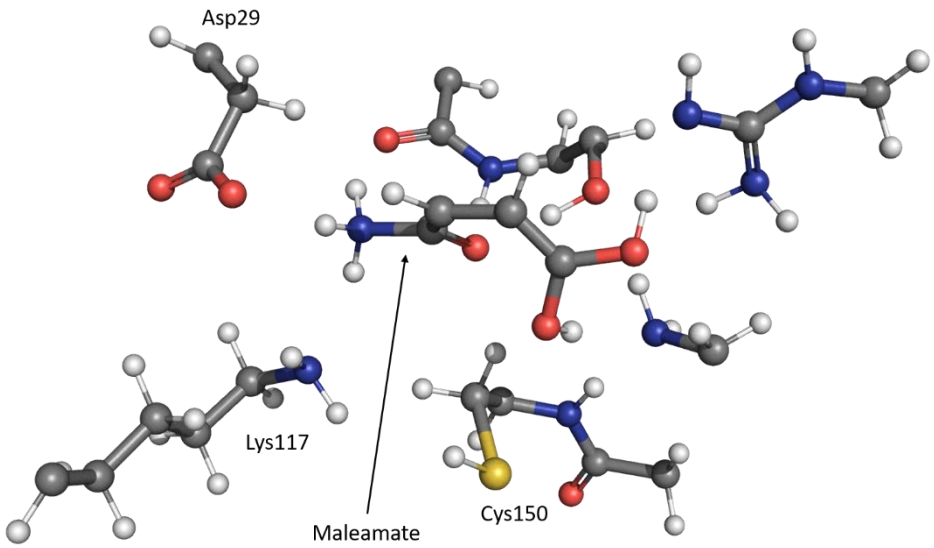


Figure 6.3. Active site of the x-ray **RC** after 120 optimization cycles.

The potential energy surfaces for each starting structure studied are shown in Figure 6.4 following the proposed mechanism. There are several notable points of difference between the surface explored using a selected pose compared to random, iterative selection. Firstly, all the structures have lower energy barriers relative to the cluster-selected pose except for **IC2** after 4 ns. To investigate the reasons for these observations, we decomposed QM and MM energy contributions to the ONIOM energy for each structure along the proposed reaction coordinate (see **Chapter 2** for more details).

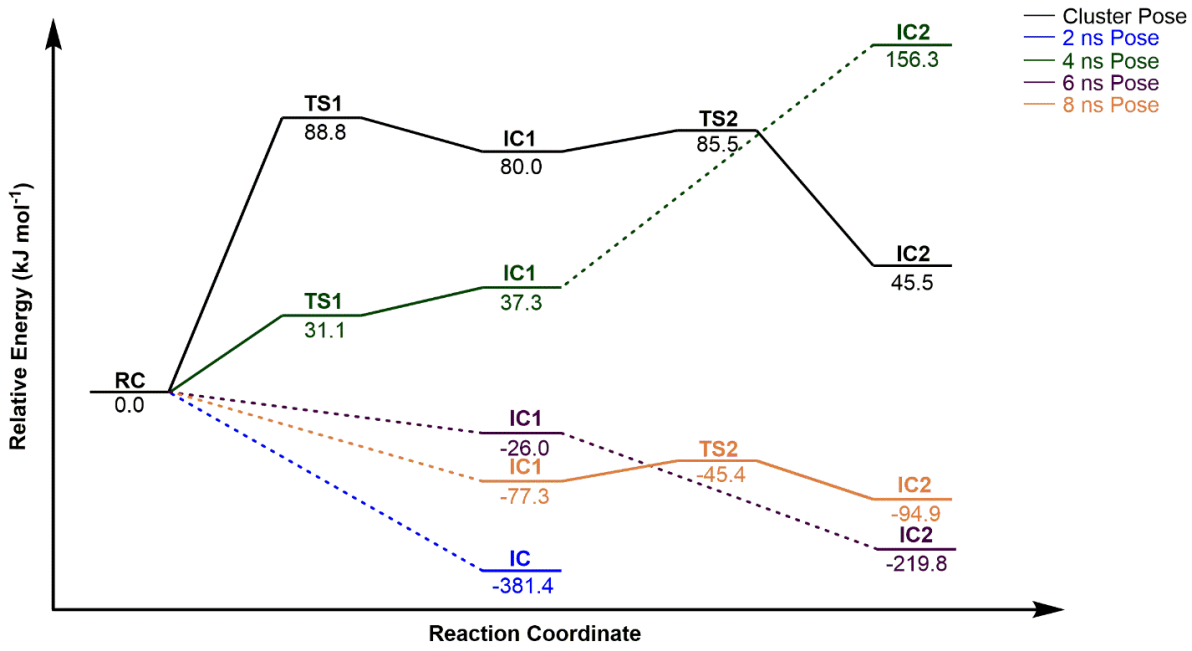


Figure 6.4. Combined potential energy surface for the proposed reaction from each structure as a starting point. Some transition structures could not be obtained and disconnects in the surface are shown by a dotted line connection.

In a typical calculation, it is expected the main contribution to the ONIOM energy will be from the QM-layer where the chemistry occurs. Contributions from the MM-layer are still expected from hydrogen bonding and sterics, but these are expected to be minimal relative to electronic contributions. Figure 6.5 shows the decomposed ONIOM energy for the selected cluster pose. The QM energy is the main contributor to the ONIOM energy and the trend in the ONIOM energy is qualitatively similar to the QM energy. The contribution of the MM energy to the ONIOM energy is flat by comparison and results in a minor impact (relative to the QM-layer). Notably, this decomposition is the only one in which contributions from the MM layer increase the energy for the reaction to occur (positive rather than negative MM energies as seen in Figures 6.6-6.9). Furthermore, **TS2** is

predicted to exist because of the low layer contributions where they further destabilize the optimized structure compared to the QM contribution which is relatively flat.

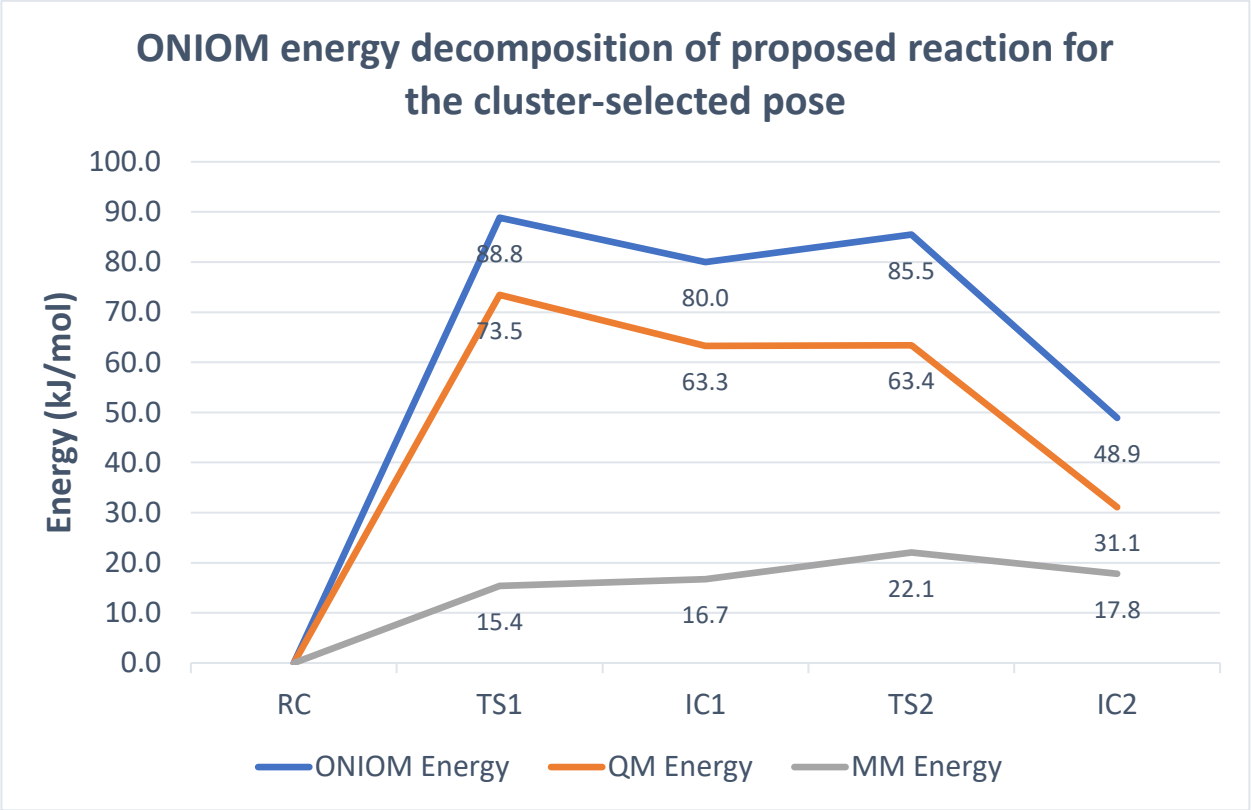


Figure 6.5. ONIOM energy decomposition of the cluster-selected pose along the potential energy surface for the proposed mechanism.

The same ONIOM decomposition from the 2, 4, 6, and 8 ns starting structures were analyzed and it was found that each of these is affected by significant MM contributions to the ONIOM energy. In particular, the 2 ns structure falls into a thermodynamically stable state as a tetrahedral intermediate. In Figure 6.6, we can see that the major contribution to the ONIOM energy indeed results from relaxation in the MM layer rather than contributions from the QM layer. This may be an indicator that MD was not performed for enough time to allow the enzyme to equilibrate, causing major reorganizations in the MM layer leading to overall stabilization. Further attempts to obtain **IC2** were unproductive due

to this. At this point, this structure is thermodynamically trapped, and no reaction can be predicted.

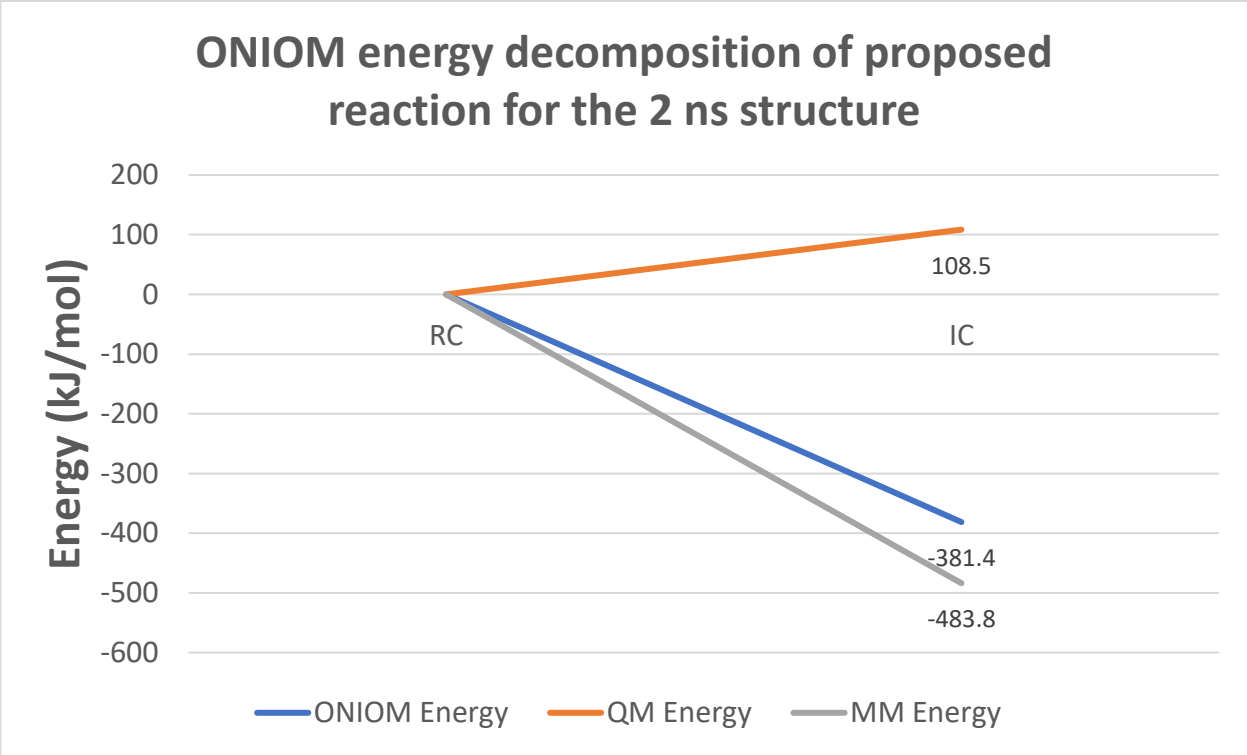


Figure 6.6. ONIOM-decomposed potential energy surface for the 2 ns starting structure.

By contrast, the potential energy surface followed by the 4 ns starting structure shows a different result (Figure 6.7). The contribution from the MM layer is much smaller than for the 2 ns structure, however the contribution is still significant, except for **IC2.** The overall shape of the surface follows the QM layer, but points to other problematic issues that may lead to a misrepresentation of the reaction; for example, **TS1** is found to be lower in energy than **IC1**. We observed a similar trend when applying electrostatic embedding to the cluster analyzed pose in the study performed in **Chapter 5**. This raised the question of whether **TS1** was a transition structure, though this was confirmed with frequency analysis through a single imaginary frequency. On the other hand, the reaction proceeds through **IC2** with a very high barrier of 156.2 kJ mol^{-1}. Reasons for this are not known, especially since **IC1** is

by no means thermodynamically preferred relative to **RC**. As a result, the rate-determing step for this reaction may be interpreted to be skewed towards the substrate-bound state.

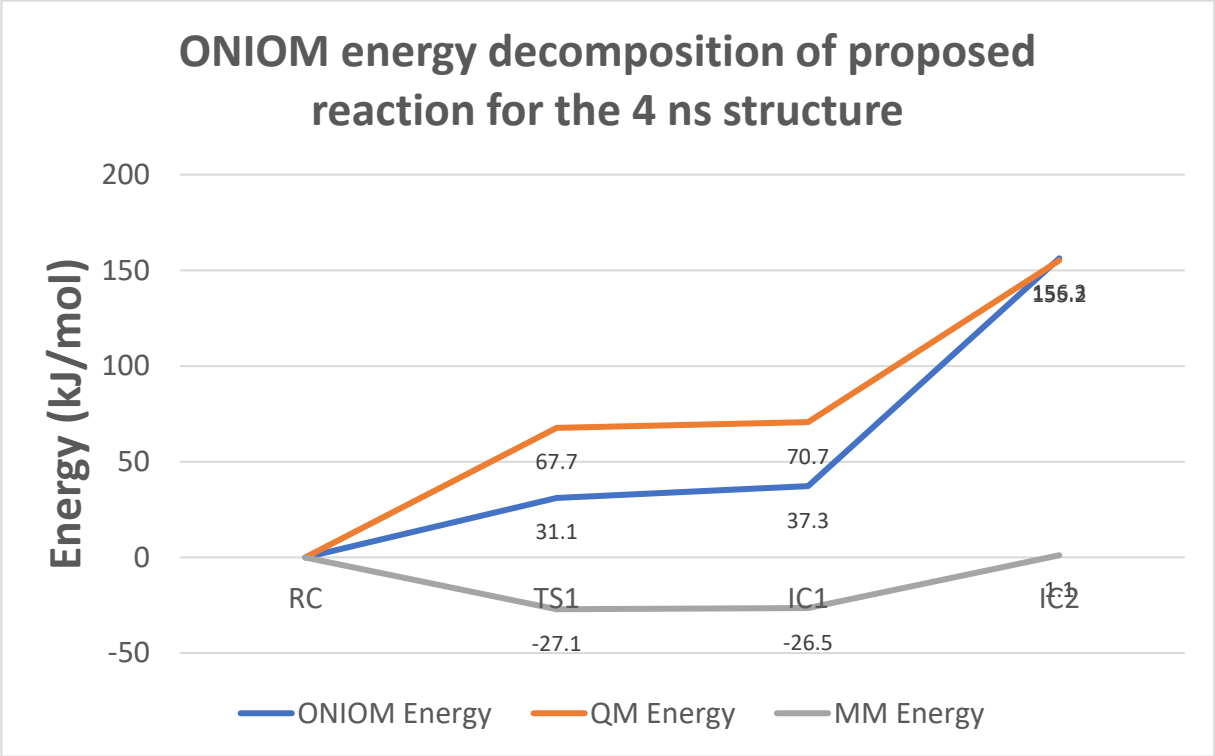


Figure 6.7. ONIOM-decomposed potential energy surface for the 4 ns starting structure.

For the 6 ns starting structure, we observe a similar pattern to the 2 ns structure (Figure 6.8). However, unlike the 2 ns structure, the QM energy suggests that this is a low barrier reaction resulting in a stable intermediate (*i.e.* both energies point towards the same story). Transition structures could not be obtained along this surface and are omitted from analysis. This energy decomposition is difficult to interpret due to the stabilization of **IC2** relative to **RC**. It suggests an entirely different mode of catalysis, at least in the first half of the reaction. If the second half reaction were to proceed similar to cluster-analyzed results, the barrier could be too large to overcome and may leave the enzyme in a thermodynamically trapped enzyme-substrate complex. This is a particularly problematic structure due to the

instability observed in the MM layer and may represent a transient high-energy state MD simulation (see Figure 5.1A, a spike is observed in the oxyanion hole RMSD at 6 ns).

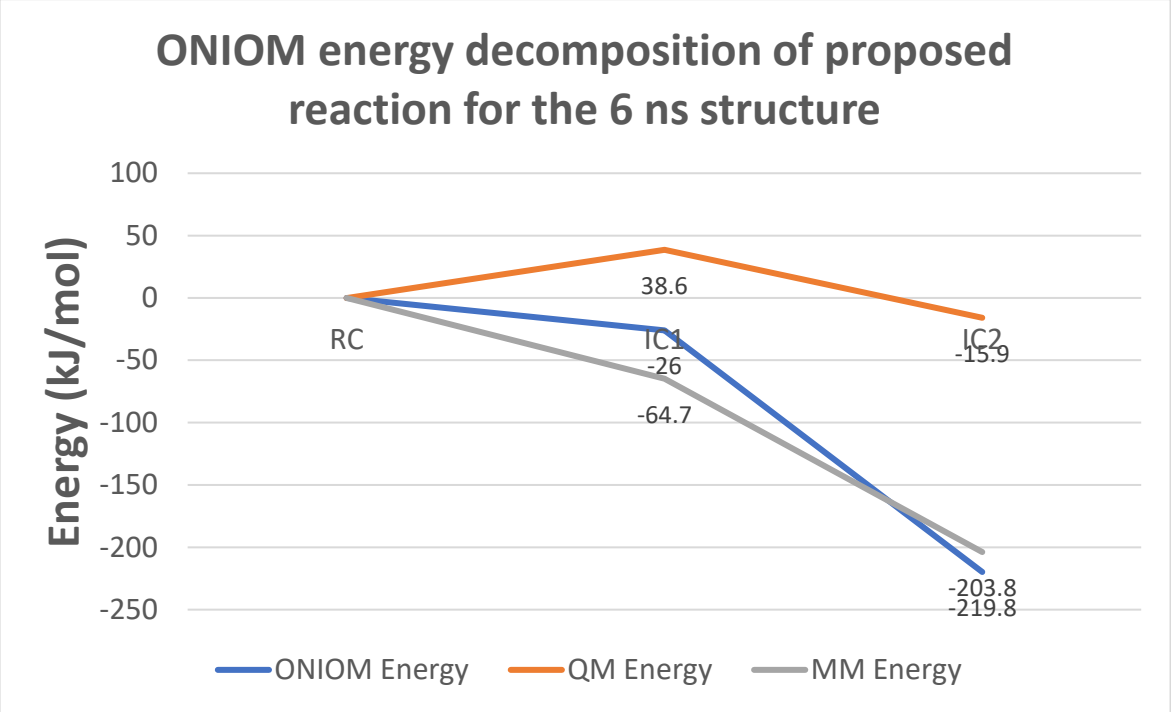


Figure 6.8. ONIOM-decomposed surface for the 6 ns starting structure. Transition structures could not be successfully characterized along this surface.

Results so far may be interpreted such that perhaps more than 6 ns of simulation is required to obtain an equilibrated starting structure. In the decomposed surface from the structure obtained after 8 ns, we see a dichotomy between QM and MM contributions to the ONIOM energy (Figure 6.9). In contrast to MM energies, the QM energies follow a similar trend to that observed in the cluster pose, except ~30 kJ mol^{-1} higher. These barriers are too large for a typical enzymatic reaction; thus, the proposed mechanism is not predicted to be feasible for this structure. Attempts to characterize **TS1** yielded no results and is omitted. These results also highlight the importance of decomposing ONIOM energies into its constituents. Had the MM contribution been more reasonable, the ONIOM energy would have suggested that the mechanism is feasible and very similar energetically to the cluster-

selected pose. The main difference would be from relaxation in the secondary shell residues and a larger barrier for C—N_{amide} bond cleavage. Visual inspection of the structures would be required to analyze structural differences observed at each point along the surface. Overall, this suggests there is potential for many structures to be a good starting point for QM/MM calculations following MD. However, once again, we observe an a large amount of added stability attributed to the MM layer.

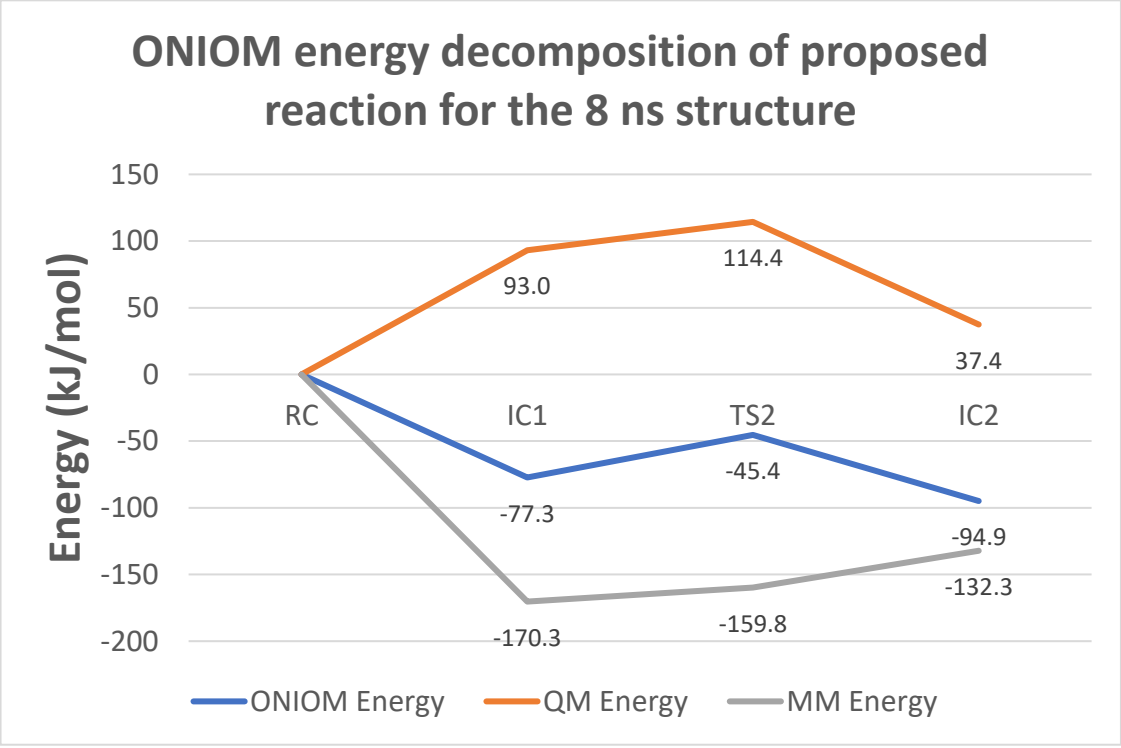


Figure 6.9. ONIOM-decomposed surface for the 8 ns starting structure.

As part of a typical protocol, residues in the low layer beyond a certain distance from the active site are fixed in place to prevent large rearrangement of the protein and significant energetic contribution from the MM layer to the ONIOM energy (typically at least 10 Å). We attempted to take this one step further and fix in place all residues that did not participate in the reaction (*i.e.*, the active site) to impose a fixed protein environment which would only bring about energetic contributions from the QM layer. Unfortunately, none of

the structures could be optimized. We faced many challenges in our attempts to obtain complexes which was a direct result of the imposed constraints. Second-shell residues require flexibility to allow the enzyme to adjust and accommodate such steric changes for the reaction to proceed.

6.4 Conclusion

In this study, we examined several post-MD pose selection strategies for their ability to give reliable potential energy surfaces in QM/MM investigations. A total of 7 starting structures were examined including 5 chosen iteratively from a MD simulation (after 2, 4, 6, 8, and 10 ns), the x-ray structure, and a structure selected from cluster analyses. Examination of key distances along the proposed mechanism can be used as an initial filtering method, but there are more considerations that do not necessarily indicate the quality of the surface they will produce. For example, the final structure from MD remained thermodynamically trapped in the optimized reactive complex due to the distance of Asp29 to the amide nitrogen atom. Furthermore, the x-ray structure optimized to a pre-reactive complex. Apart from the cluster-selected pose, none of the remaining structures was able to produce a reliable surface. Decomposing the ONIOM energies revealed that this was due to large fluctuations in the portion of the protein modeled using molecular mechanics. These results demonstrate the importance of the MM layer in QM/MM calculations; although it is expected to have a minor effect, it is of noteworthy consequence especially when energies fluctuate dramatically. Since the cluster pose exhibited no such issues, it is highly recommended to perform cluster analyses on MD simulations to obtain a suitable starting structure, though caution is still encouraged. ONIOM energy decomposition along the proposed reaction coordinate serves as a secondary check and can ensure each layer is well-behaved throughout optimization. An unstable MM environment has large implications on the QM layer and affects positioning of important residues and the substrate. In addition,

the lack of any surface obtained for x-ray structure demonstrates the importance of performing MD studies, particularly in cases where docking is required, or experimental mutations are made and must be restored to their native state *in silico*. Further studies can examine the use of newly developing polarizable embedding schemes to correct for these issues.

6.5 References

1. Bornscheuer, U. T.; Huisman, G. W.; Kazlauskas, R. J.; Lutz, S.; Moore, J. C.; Robins, K., Engineering the third wave of biocatalysis. *Nature* **2012,** *485*, 185-94.

2. Norskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H., Towards the computational design of solid catalysts. *Nat. Chem.* **2009,** *1*, 37-46.

3. Sperger, T.; Sanhueza, I. A.; Kalvet, I.; Schoenebeck, F., Computational Studies of Synthetically Relevant Homogeneous Organometallic Catalysis Involving Ni, Pd, Ir, and Rh: An Overview of Commonly Employed DFT Methods and Mechanistic Insights. *Chem. Rev.* **2015,** *115*, 9532-86.

4. van der Kamp, M. W.; Mulholland, A. J., Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry* **2013,** *52*, 2708-28.

5. Chung, L. W.; Sameera, W. M. C.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z. F.; Liu, F. Y.; Li, H. B.; Ding, L. N.; Morokuma, K., The ONIOM Method and Its Applications. *Chem. Rev.* **2015,** *115*, 5678-5796.

6. Senn, H. M.; Thiel, W., QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed. Engl.* **2009,** *48*, 1198-229.

7. Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W., P450 Enzymes: Their Structure, Reactivity, and Selectivity-Modeled by QM/MM Calculations. *Chem. Rev.* **2010,** *110*, 949-1017.

8. Romero-Tellez, S.; Lluch, J. M.; Gonzalez-Lafont, A.; Masgrau, L., Comparing Hydrolysis and Transglycosylation Reactions Catalyzed by Thermus thermophilus beta-Glycosidase. A Combined MD and QM/MM Study. *Front. Chem.* **2019,** *7*, 200.

9. Timmins, A.; Saint-Andre, M.; de Visser, S. P., Understanding How Prolyl-4-hydroxylase Structure Steers a Ferryl Oxidant toward Scission of a Strong C-H Bond. *J. of Am. Chem. Soc.* **2017,** *139*, 9855-9866.

10. Wilson, K. A.; Fernandes, P. A.; Ramos, M. J.; Wetmore, S. D., Exploring the Identity of the General Base for a DNA Polymerase Catalyzed Reaction Using QM/MM: The Case Study of Human Translesion Synthesis Polymerase eta. *ACS Catal.* **2019,** *9*, 2543-2551.

11. Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem Theory Comput.* **2012,** *8*, 1542-1555.

12. Hong, C.; Tieleman, D. P.; Wang, Y., Microsecond Molecular Dynamics Simulations of Lipid Mixing. *Langmuir* **2014,** *30*, 11993-12001.

13. Ahmadi, S.; Herrera, L. B.; Chehelamirani, M.; Hostas, 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*.

14. Lonsdale, R.; Harvey, J. N.; Mulholland, A. J., A practical guide to modelling enzyme-catalysed reactions. *Chem. Soc. Rev.* **2012,** *41*, 3025-3038.

15. Quesne, M. G.; Borowski, T.; de Visser, S. P., Quantum Mechanics/Molecular Mechanics Modeling of Enzymatic Processes: Caveats and Breakthroughs. *Chem. Eur. J.* **2016,** *22*, 2562-2581.

16. Lence, E.; van der Kamp, M. W.; Gonzalez-Bello, C.; Mulholland, A. J., QM/MM simulations identify the determinants of catalytic activity differences between type II dehydroquinase enzymes. *Org. Biomol. Chem.* **2018,** *16*, 4443-4455.

17. Molakarimi, M.; Mohseni, A.; Taghdir, M.; Pashandi, Z.; Gorman, M. A.; Parker, M. W.; Naderi-Manesh, H.; Sajedi, R. H., QM/MM simulations provide insight into the mechanism of bioluminescence triggering in ctenophore photoproteins. *PLoS One* **2017,** *12*, e0182317.

18. Ribeiro, A. J. M.; Santos-Martins, D.; Russo, N.; Rarnos, M. J.; Fernandes, P. A., Enzymatic Flexibility and Reaction Rate: A QM/MM Study of HIV-1 Protease. *ACS Catal.* **2015,** *5*, 5617-5626.

19. Bonk, B. M.; Weis, J. W.; Tidor, B., Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* **2019,** *141*, 4108-4118.

20. Ion, B. F.; Meister, P. J.; Gauld, J. W., Multiscale Computational Study on the Catalytic Mechanism of the Nonmetallo Amidase Maleamate Amidohydrolase (NicF). *J. Phys. Chem. A* **2019,** *123*, 7710-7719.

21. Kincaid, V. A.; Sullivan, E. D.; Klein, R. D.; Noel, J. W.; Rowlett, R. S.; Snidert, M. J., Structure and Catalytic Mechanism of Nicotinate (Vitamin B3) Degradative Enzyme Maleamate Amidohydrolase from Bordetella bronchiseptica RB50. *Biochemistry* **2012,** *51*, 545-554.

22. Galeazzi, L.; Bocci, P.; Amici, A.; Brunetti, L.; Ruggieri, S.; Romine, M.; Reed, S.; Osterman, A. L.; Rodionov, D. A.; Sorci, L.; Raffaelli, N., Identification of nicotinamide mononucleotide deamidase of the bacterial pyridine nucleotide cycle reveals a novel broadly conserved amidohydrolase family. *J. Biol. Chem.* **2011,** *286*, 40365-75.

23. Boshoff, H. I. M.; Xu, X.; Tahlan, K.; Dowd, C. S.; Pethe, K.; Camacho, L. R.; Park, T. H.; Yun, C. S.; Schnappinger, D.; Ehrt, S.; Williams, K. J.; Barry, C. E., Biosynthesis and recycling of nicotinamide cofactors in Mycobacterium tuberculosis - An essential role for NAD in nonreplicating bacilli. *J. Biol. Chem.* **2008,** *283*, 19329-19341.

24. *Molecular Operating Environment (MOE)*, 2013.08; Chemical Computing Group Inc.: Montréal, QC, Canada, 2015.

25. Bond, S. D.; Leimkuhler, B. J.; Laird, B. B., The Nosé-Poincaré Method for Constant Temperature Molecular Dynamics. *J. Comput. Phys.* **1999,** *151*, 114-134.

26. Bearpark, M. J.; Ogliaro, F.; Vreven, T.; Boggio-Pasqua, M.; Frisch, M. J.; Larkin, S. M.; Morrison, M.; Robb, M. A., CASSCF Calculations for Photoinduced Processes in Large Molecules: Choosing When to Use the RASSCF, ONIOM and MMVB Approximations. *J. Photochem. Photobiol., A* **2007,** *190*, 207-227.

27. Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J., A New ONIOM Implementation in Gaussian98. Part I. The Calculation of Energies, Gradients, Vibrational Frequencies and Electric Field Derivatives. *J. Mol. Struct.* **1999,** *461*, 1-21.

28. Humbel, S.; Sieber, S.; Morokuma, K., The IMOMO Method: Integration of Different Levels of Molecular Orbital Approximations for Geometry Optimization of Large Systems: Test for n-butane Conformation and S(N)2 Reaction: RCl + Cl-. *J. Chem. Phys.* **1996,** *105*, 1959-1967.

29. Maseras, F.; Morokuma, K., IMOMM – A New Integrated Ab Initio Plus Molecular Mechanics Geometry Optimization Scheme of Equilibrium Structures and Transition States. *J. Comput. Chem.* **1995,** *16*, 1170-1179.

30. Morokuma, K.; Musaev, D. G.; Vreven, T.; Basch, H.; Torrent, M.; Khoroshun, D. V., Model Studies of the Structures, Reactivities, and Reaction Mechanisms of Metalloenzymes. *IBM J. Res. Dev.* **2001,** *45*, 367-395.

31. Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K., ONIOM: A Multilayered Integrated MO+MM Method for Geometry Optimizations and Single Point Energy Predictions. A Test for Diels-Alder Reactions and Pt(P(t-Bu)(3))(2)+H2 Oxidative Addition. *J. Phys. Chem.* **1996,** *100*, 19357-19363.

32. Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J., Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM. *J. Chem. Theory Comput.* **2006,** *2*, 815-826.

33. Vreven, T.; Morokuma, K., On the Application of the IMOMO (Integrated Molecular Orbital Plus Molecular Orbital) Method. *J. Comput. Chem.* **2000,** *21*, 1419-1432.

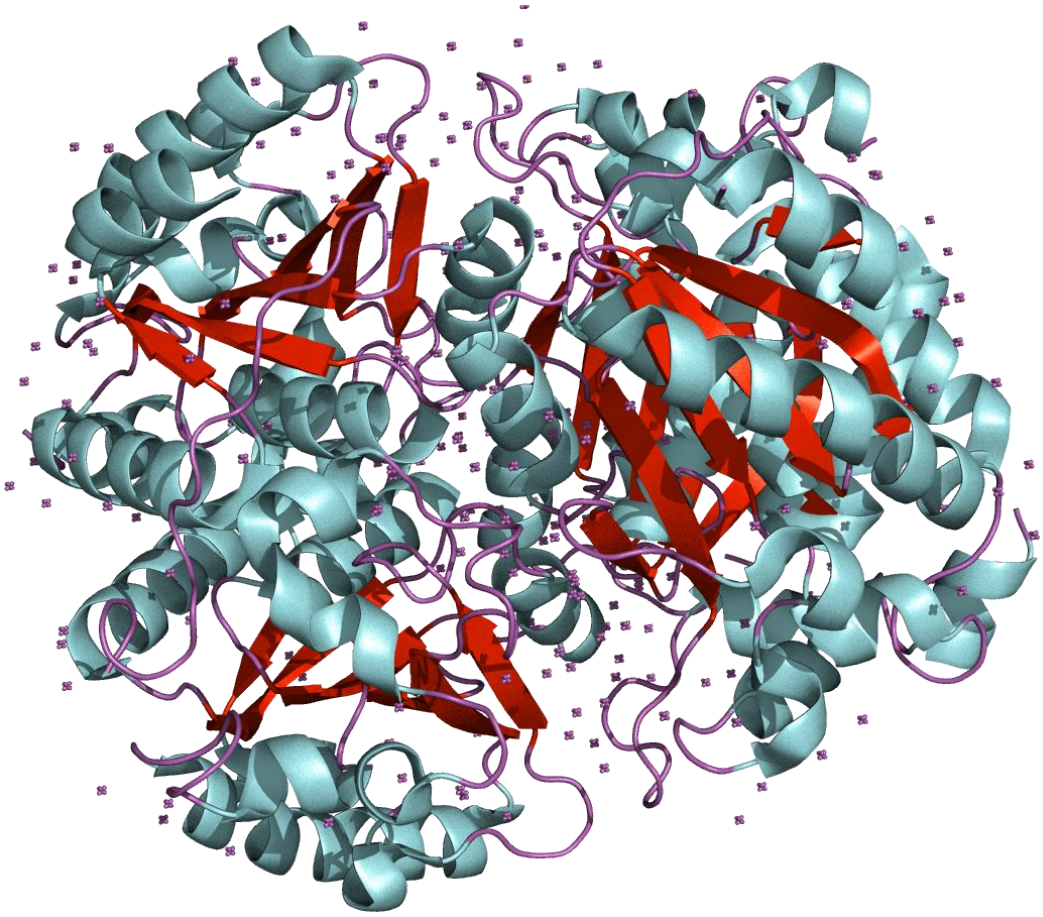
34. Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J., Geometry Optimization with QM/MM, ONIOM, and Other Combined Methods. I. Microiterations and Constraints. *J. Comput. Chem.* **2003,** *24*, 760-769.

35. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, M. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09, Revision D.01*, Gaussian, Inc.: Wallingford CT, 2009.

36. Ion, B. F.; Kazim, E.; Gauld, J. W., A Multi-Scale Computational Study on the Mechanism of Streptococcus pneumoniae Nicotinamidase (SpNic). *Molecules* **2014,** *19* (10), 15735-53.

Chapter 7

Simplifying QM/MM for Enzymes: Molecular Dynamics Pose Selection using Agglomerative Machine Learning



7.1 Introduction

Computational enzymology has become a robust field for the study of enzymatic mechanisms.[1-5] It gives us an atomistic picture of the active site and enables the capture of unstable transition structures. Several protocols have been established for obtaining structures for quantum calculations.[5-7] Typically, a crystal structure found on the Protein Databank (PDB) will be used as a starting point. It is common to find mutations in the protein for the purpose of crystallography; however, these must be restored to their native state using various computational tools. Missing substrates will be docked into the active site and the entire enzyme is solvated for molecular dynamics (MD) simulations. At least 1 ns of production simulation is required for a reasonable relaxation time, though 10-20 ns is increasingly more common.

It is at this stage that computational chemists face a potentially critical challenge. MD produces several thousands of structures (time steps of 2 fs), so choosing a representative structure or group of structures is nontrivial. To aid this process, several different protocols have been used and described in the literature, the most common of which is to analyze structural fluctuations of interest (for example, along a proposed reaction coordinate).[8] Meanwhile, others have iteratively selected structures,[9] while yet others may select the final structure in the MD simulation.[10] However, enzymatic catalysis is often complex and there are many different factors that contribute to their rate-enhancing capabilities.[11] For example, enzymes may employ any number of strategies from: transition state stabilization, compensation for enzyme-substrate binding entropy, desolvation effects, steric effects, electrostatic pre-organization, and/or donor-acceptor matching pK_as. The impacts of some of these strategies can be assessed using MD methods (such as steric effects or pK_a matching), but others require in-depth quantum chemical calculations. For these reasons, visual inspection of results is often not reliable; chemical intuition can be a useful guidepost. Oftentimes, in-depth knowledge on enzymatic catalysis and existing

experimental data is relied upon to make a well-informed decision. This process is further complicated by lengthy simulations (> 100 ns) since large domain shifts can occur with increasing time scales, thus increasing the complexity of structural analyses.[12]

Current methods of pose selection are labour intensive and have the potential to introduce user error. Researchers must make predictions about what chemical features will be relevant based solely on data available in the literature and visual inspection of the system. If difficulties in the modeling process are encountered, it is common to reassess data and create new models. For those with relatively little experience (for example new graduate students or undergraduates in training), or for instances where a proposed mechanism paints an incomplete picture, MD simulation results can be difficult to interpret and this may lead to further complications in subsequent analyses such as QM/MM. Valuable data can be overlooked, albeit inadvertently, while analyzing MD trajectories due to the sheer volume of information. Results must be narrowed down to a manageable number of poses so a large portion of data cannot be inspected. With all of these considerations, pose selection can be a daunting task with significant consequences.

Herein, we aim to develop a system which is both automated for efficiency and helps to remove user bias to help alleviate these issues. Previous studies have looked at using machine learning for identifying key features of enzymatic catalysis,[13] but our aim is to help improve the user experience. A machine learning algorithm is applied to a molecular dynamics trajectory to produce a set of "catalytically viable" starting structures for QM/MM. In this way, the efficiency of the researcher's workflow can be considerably increased as less time would be dedicated to identifying parameters and selection poses manually while still allowing the researcher to visibly sort through a small set of selected data and apply their chemical intuition. Additionally, the program sorts through all the data and may select a structure that would otherwise have been ignored.

An agglomerative clustering algorithm is applied to the structures produced from a MD simulation of NicF. We have previously studied the NicF mechanism and found that it proceeds in two stages.[8] In the first stage, the enzyme-substrate complex is formed through Cys150, and ammonia is released from the maleamate substrate in a stepwise manner. In the second stage, a water molecule is activated and used to form the acid product and restore the enzyme to its reactant state. As was done in **Chapter 6**, the first stage of the mechanism will be studied to determine the impact of changing the initial structure. In that chapter, we demonstrated the importance of making an informed decision which manifested as an instability in the MM layer, producing inaccurate potential energy surfaces. Likewise, structures selected by the program are then submitted for QM/MM calculation to examine the reaction pathway and compare results.

7.2 Computational Methods and Machine Learning Workflow

7.2.1 Molecular Dynamics Simulations

Molecular dynamics simulations were performed in our previously published study.[8] To summarize, the Molecular Operating Environment (MOE) program suite[14] was used to prepare the starting structure (PDB ID: 3UAO).[15] The x-ray structure was complexed with an acetate molecule and Cys150 was mutated to the sulfenic acid. Acetate was changed to maleamate and the sulfenic acid was reverted to a thiol. The resulting complex was solvated by a 4 Å spherical layer of water molecules restrained within an ellipsoidal potential wall. The entire structure was then minimized using the AMBER99 force field until the root mean square gradient of the total energy fell below 0.42 $kJ \cdot mol^{-1 \cdot Å^{-2}}$. This is then heated in 100 ps at constant pressure. Production simulations were run using the Nosé-Poincaré thermostat with a time step of 2 fs for a total of 10 ns.[16] Structures identified using the machine learning workflow below were then minimized in the AMBER14 force field before continuing with QM/MM calculations.

7.2.2 Agglomerative Clustering and Workflow

In this study, we chose to use a hierarchical clustering algorithm to identify MD poses suitable for QM/MM calculations. Unsupervised clustering groups together samples based on their features. An algorithmic model is trained to associate features with a particular group.[17] New samples can be sorted into groups, or clusters, based on these identified features. An overview is shown in Figure 7.1 with an example. For our purpose, each frame is considered a sample. The agglomerative clustering approach was used to create two clusters of poses.

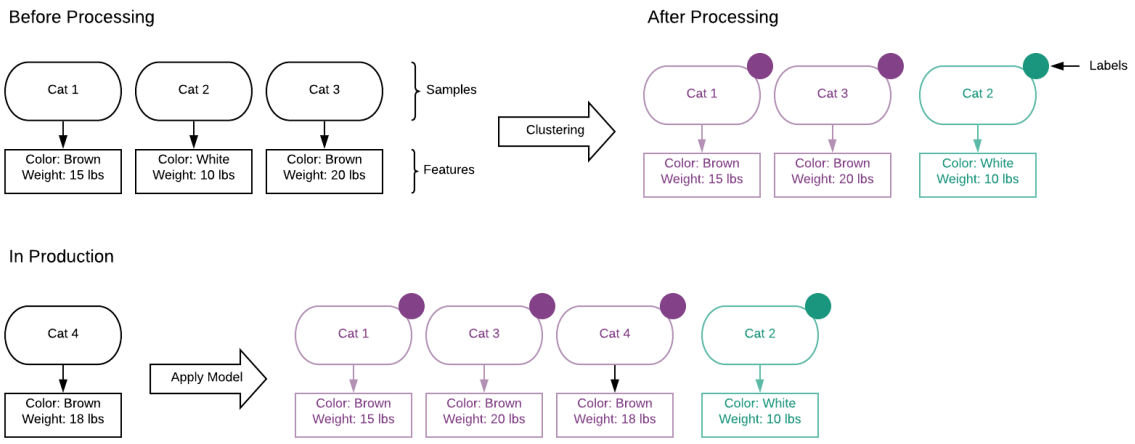


Figure 7.1. An overview of unsupervised clustering. Samples are labeled based on their features and grouped. The algorithm is then applied to a new sample for placing in the appropriate group.

Agglomerative Clustering takes a bottom-up approach to clustering. Samples are merged into successively larger groups until clusters are too different to be merged (or when a specified number of clusters is reached). The features, in this case, are chemical in nature. The movement of atoms in the active site of the protein and the ligand is the activity of interest. Therefore, atoms of interest were first identified, afterward information about their

distance from frame to frame was passed to the clustering algorithm for training. Two clusters of poses were created: a smaller "catalytically viable" group, and a larger "non-viable" group. This workflow is summarized in Figure 7.2.

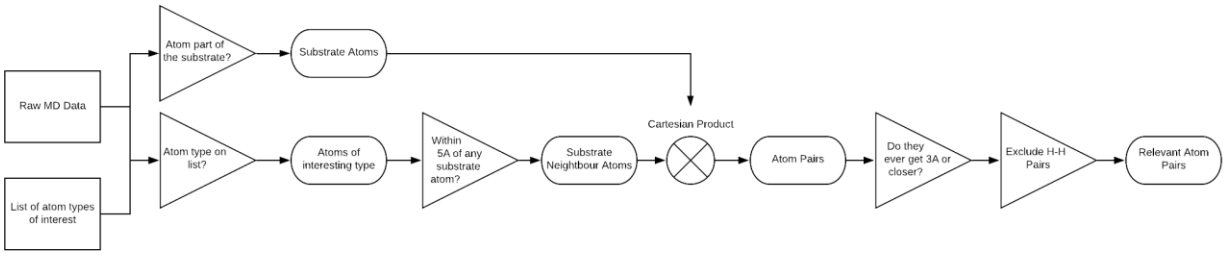


Figure 7.2. Workflow leading into model training. Atoms are selected based on their type, residue, and distance from the substrate. They are then paired up. Pairs are filtered based on the proximity of each atom to one another; a pair with atoms on opposite ends of the substrate is rejected. This final list of atom pairs is what is finally processed by the algorithm.

All code was written in Python since several scientific libraries are available. In particular, MDTraj[18] was used to process MD simulation output. Binary files generated by MD are converted into dynamic trajectory objects, making each atom in the system easily searchable by their location or identity. It also allows more detailed information, like the displacement of atoms frame by frame, to be obtained. NumPy[19] was used in conjunction with these functions to add enhanced array (or matrix) functionality.

Hierarchical clustering functions were imported from SciKit-Learn[20]. Scikit-learn is a robust machine learning library including modules for pre-processing data, training and testing algorithms, and model evaluation. In addition to Agglomerative Clustering functions, two modules were employed. The Principal Component Analysis (PCA) module is used here for dimensionality reduction, compressing three-dimensional displacement

vectors into a single number and the silhouette score[21] was used to evaluate the accuracy of the model. The silhouette coefficient, part of the clustering performance evaluation module, is an indication of the quality of the clusters produced by a model. It does this by measuring the cohesion between members in a group, and the distance between groups. Mathematically, the silhouette coefficient is calculated as:

$$S = \frac{b - a}{max(a, b)}$$

Where *a* is the mean distance from a given sample to all other members of the same group, and *b* is the mean distance between a sample and all members of the next closest group.

7.2.3 QM/MM Calculations

All QM/MM calculations were performed using the ONIOM[22-28] formalism in Gaussian 09 program suite.[29-30] Geometries were optimized at the ONIOM(B3LYP/6-31G(d):AMBER96) level of theory using mechanical embedding.[31-34] Harmonic vibrational frequencies were also obtained at this level to confirm intermediates as minima and transition structures as first-order saddle points. Single point energy calculations were performed at a higher level of theory, namely ONIOM(ωB97X-D/6-311G(2df,p):AMBER96) as we found it to be more accurate for this reaction.[8] To save computational time, residues within a 24 Å sphere around the maleamate substrate and Cys150 are considered in the model. Active sites were represented as they were in our previous studies, containing the R-groups of the catalytic triad (Lys117, Asp29, and Cys150), partial side chains of Arg175 and Lys190' that stabilize the substrate carboxylate, and the peptide backbones of Ala145-Thr146 (including the β-OH moiety) and Gly149-Cys150. Surrounding residues and water molecules were included in the MM layer. Atoms at the edge of the system are frozen at their MM coordinates to represent steric restrictions imposed by the enzyme.

7.3 Results and Discussion

7.3.1 The "Pose Selector" Program

We created a program that helps the user select a set of "catalytically viable" poses for application in QM/MM by implementing an agglomerative machine learning algorithm to data produced in MD simulations. The program operates in three steps.

Step 1: A user can input MD trajectory data in a variety of file formats. MDTraj takes this data and converts it to dynamic trajectory and topology objects where each atom is searchable. Several filters are applied to the data to remove interactions observed outside of the active site.

- Water molecules are ignored; they make up a bulk of the atoms in the system,
- Substrate atoms are ignored; these are added back in later,
- Atoms not included on list of atom types provided by the user are ignored ,
- Atoms outside a 5 Å sphere around the active site are ignored.

In the NicF system, this reduces the number of atoms considered from 87,155 to only 35.

Step 2: The list of atoms generated in *Step 1* is paired up with an atom from the substrate, and duplicates are removed from the list. H-H atom pairs were found to reduce the quality of the model produced, so these are removed. Furthermore, pairs of atoms that do not come closer than 3 Å throughout the simulation are ignored. Applying this approach to NicF helped improve data significantly. However, this is not recommended for systems where long-range interactions are expected to be observed. This may come at a cost of lowering the quality of results, so we recommend testing to see which approach works best for the data set in question. A total of 28 pairs were selected after processing for NicF.

Step 3: Displacement vectors for each pair of atoms in are computed across each frame. This produces a 3D array, organized by frame, then pair, then a three-component vector for

each pair. In the clustering algorithm, each frame is considered a sample and the atom distance is a feature. Machine learning requires a 2D array, organized by sample then feature. Principal component analysis projects our three-component vectors onto a 1D space, reducing them to single value scalars. At this stage, agglomerative machine learning is applied, and frames are clustered into one of two groups. The results are evaluated *via* silhouette coefficient to determine the quality of the model.

7.3.2 Application to Maleamate Amidohydrolase (NicF)

In our previous study on NicF in **Chapter 5**, we identified a representative frame from the MD simulation by examining the stabilization provided in the active site through a proposed oxyanion hole around the substrate amide oxygen. Hydrogen bonding interactions from $_{Thr146}$OH and $_{Thr146}$NH were crucial to enable hybridization change of the amide nitrogen to encourage release as ammonia (Figure 7.3). In particular, positioning of the former was vital as it could also hydrogen bond to the substrate carboxylic acid, as reflected in the increased RMSD. As such, we analyzed these interactions to select a representative frame that was ideal for modeling the proposed reaction and obtaining realistic barriers. Although in this case we were able to develop a clearer picture of the enzymatic mechanism based on the observations in our MD simulations and experimental results, this may not be the case for other enzymes. We chose to examine the viability of machine learning to help reduce the complexity of structure selection for QM/MM, especially in difficult cases. With data on-hand and knowledge about the detailed mechanism of NicF, we used this as an initial benchmark.

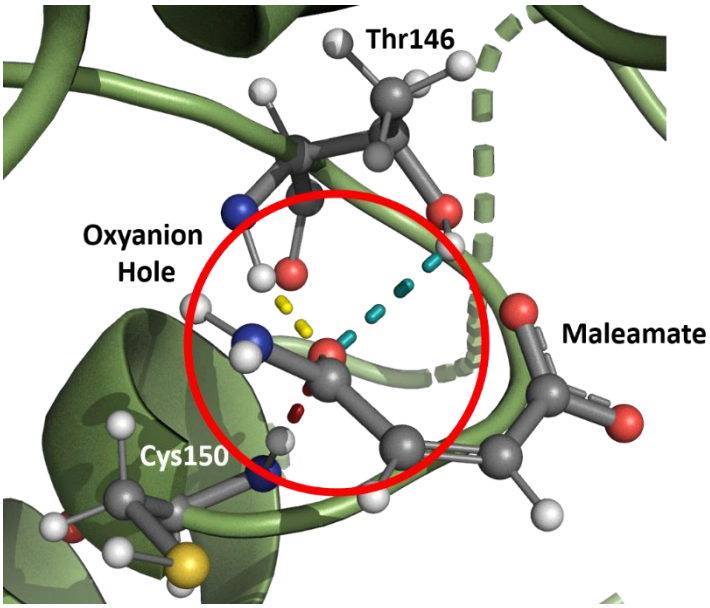


Figure 7.3. Stabilizing interactions around the proposed oxyanion hole of NicF. Two of these are to Thr146 and one is to Cys150.

Examination of only the key interactions observed in Figure 7.3 ignores the contribution of the remainder of the active site. Thus, we chose to examine all interactions to remain as general as possible and consider the effect of interactions that were not originally described. The MD model was built as an octamer with a total of 8 active sites present over the entire 10 ns simulation. Each active site was identified from the unique substrate PDB identifier (ACT) and numbered according to the order of appearance in the PDB file (ACT 1610 is active site 1, ACT 1611 is active site 2, etc…). Atom types that have the potential to be involved in stabilizing interactions (*e.g.*, hydrogen bonding) were broadly identified based on their AMBER atom type and were exported as a CSV file to read into the algorithm. Atom pairs from the substrate to any active site atom (within 5 Å of maleamate) that had an atom type listed in the CSV file were generated. Agglomerative clustering is applied to this data, and dendrograms are created to show the ideal number of clusters for each active site (Appendix C), suggesting 2 clusters are ideal for each active site. The algorithm is then evaluated *via* the silhouette coefficient to examine the quality of clusters produced.

Interestingly, each active site produced a different number of frames with differing silhouette scores (Table 7.1). Notably, there were no results obtained for active site 7. Results were checked to ensure that the workflow was followed, yet there were no distinct clusters produced. This may be due to smaller fluctuations in this active site and the algorithm could not discern between the small changes from frame to frame. Silhouette scores for each active site indicate that the clusters produced were reliable as they were mostly around 0.9, except for active site 3 with a score 0.297. While the clusters produced from this active site are more random relative to the others, these scores suggest this approach can be generally applied.

Table 7.1. Number of "Catalytically viable" frames and the silhouette score for each active site.

Active Site	Catalytically Viable Frames	Silhouette Score
1	462	0.917
2	37	0.924
3	158	0.297
4	128	0.903
5	232	0.925
6	285	0.893
7	0	N/A
8	482	0.912

Although there is the potential from a frame from active sites 1, 2, 4-6, or 8 to be a good starting structure, we chose to select a structure from active site 2 for further QM/MM calculations. This active site had the fewest viable frames to consider and a high silhouette score, making it an ideal candidate. This allowed us to visually inspect each frame to determine if some poses were not oriented favourably. Out of the 37 frames identified; 17 frames were discarded after visual analysis due to unfavourable hydrogen bonding orientation from Asp29 to the substrate amide nitrogen atom. From the remaining 20 frames, five were selected for further analysis using QM/MM. These were from frames: 78,

225, 446, 581, and 776 and represent points from the beginning, middle, and end of the simulation time.

The first half of the mechanism was studied for each frame with the mechanism from the selected structure pose as the point of comparison. Results were obtained for the structure from Frame 78 and the surface is shown in Figure 7.4a. Notably, the reaction is still found to be stepwise, however there are a few different features to this surface. Importantly, **TS1** could not be identified from the Frame 78 **RC**. This may indicate a low transition barrier that is difficult to characterize and shifts the rate-limiting step from C—S bond formation to C—N bond cleavage. Furthermore, the energy barriers are much lower than those obtained from the cluster pose. As in **Chapter 6**, the ONIOM energy was decomposed to determine if the lower energies were a result of a poorly optimized MM environment (Figure 7.4b). The MM layer did not show any large fluctuations as observed previously and did not skew results from the QM layer. Another important observation is that the energy of the resulting enzyme-substrate complex is lower than **RC**. This may lead to a larger barrier in the second half of the reaction since the enzyme is in a thermodynamically favoured state. Overall, these results suggest the possibility for NicF to undergo different modes of catalysis which depend upon the initial configuration of **RC**.

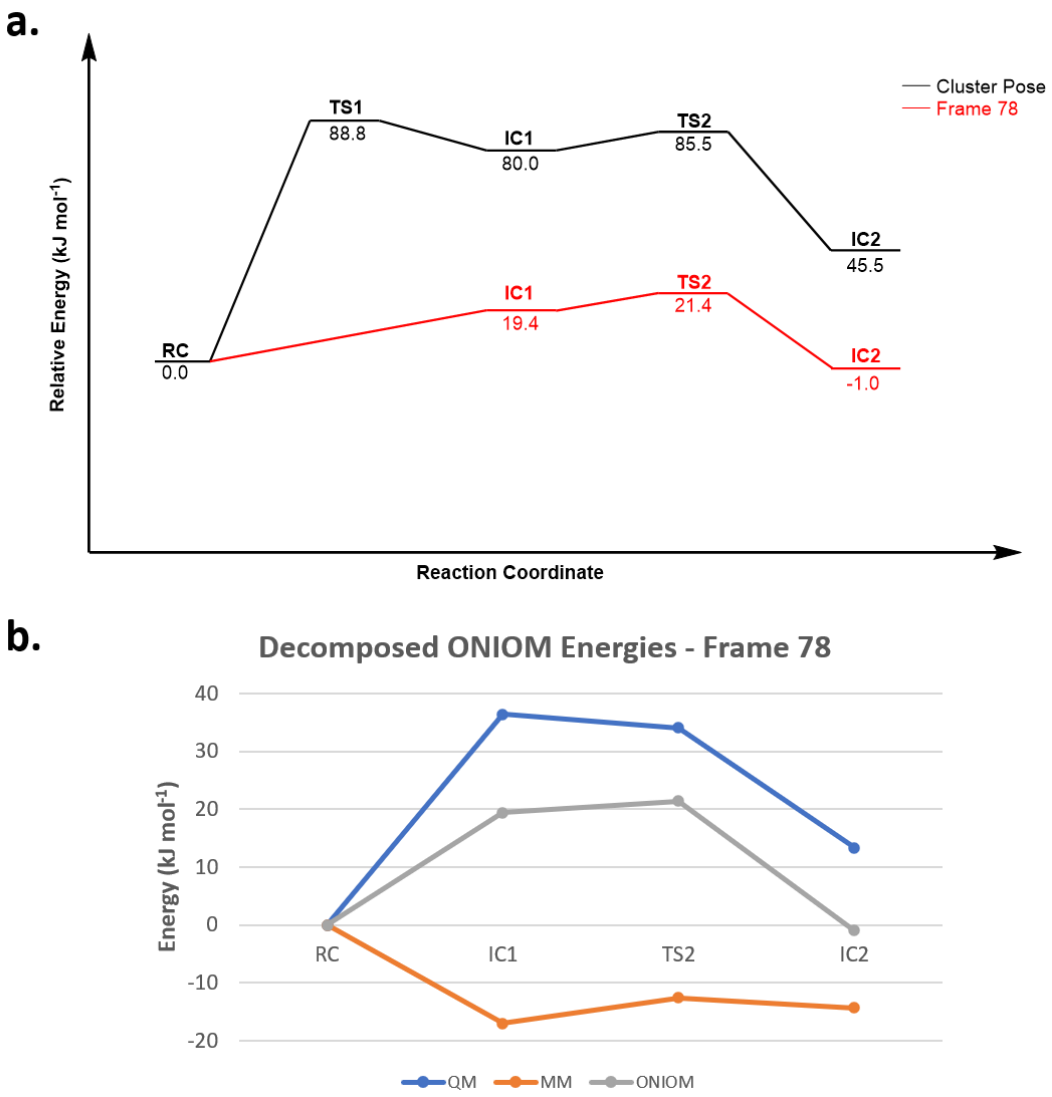


Figure 7.4. (A) Potential energy surface for Frame 78 (red) compared to the cluster selected pose (black). (B) Decomposed energies for the reaction pathway in Frame 78.

7.4 Conclusions

We examined the applicability of machine learning to help guide the selection of a molecular dynamics structure for elucidating an enzymatic mechanism using QM/MM. The program "Pose Selector" was created to read in a variety of different molecular dynamics outputs, identify key interactions within the active site, and use agglomerative clustering to output a set of "catalytically viable" poses based on these interactions. The NicF enzyme

was used as a test model for the program. It consists of a total of 8 active site present over a 10 ns MD simulation with 1010 frames saved as output. Each active site was submitted to "Pose Selector" and the second active site produced the fewest number of viable frames that also allowed for visual inspection with a silhouette score of 0.924. Two clusters were found to be ideal, and these can be divided into "catalytically viable" and "non-viable" poses. From a total of 37 "catalytically viable" frames, 17 were discarded as poor selections after visual inspection, and 5 of the remaining 20 were taken for QM/MM calculations from different time points *i.e.*, frames 78, 225, 446, 581, and 776. The potential energy surface from frame 78 indicated a stepwise reaction takes place with lower barriers. Interestingly, TS1 was not observed which may be due to a transition structure lower in energy than the intermediate. Overall, these results suggest that there are different catalytic modes available for NicF dependent on the initial structure used in calculations. Future calculations will look at the potential energy surfaces produced by the other four selected frames, as well as examine frames output from other active sites in their ability to reproduce the surface of a pose selected from cluster analyses. Furthermore, the applicability of "Pose Selector" to output reliable starting structures will be examined in different enzyme systems, consisting of, for example, more than one substrate or containing additional cofactors.

7.5 References

1. 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*, e25558.

2. Chung, L. W.; Sameera, W. M.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z.; Liu, F.; Li, H. B.; Ding, L.; Morokuma, K., The ONIOM Method and Its Applications. *Chem. Rev.* **2015,** *115* (12), 5678-796.

3. Gherib, R.; Dokainish, H. M.; Gauld, J. W., Multi-scale computational enzymology: enhancing our understanding of enzymatic catalysis. *Int. J. Mol. Sci.* **2013,** *15*, 401-22.

4. Lonsdale, R.; Ranaghan, K. E.; Mulholland, A. J., Computational enzymology. *Chem. Commun.* **2010,** *46*, 2354-72.

5. Quesne, M. G.; Borowski, T.; de Visser, S. P., Quantum Mechanics/Molecular Mechanics Modeling of Enzymatic Processes: Caveats and Breakthroughs. *Chem. Eur. J.* **2016,** *22*, 2562-2581.

6. Cao, L.; Ryde, U., On the Difference Between Additive and Subtractive QM/MM Calculations. *Front. Chem.* **2018,** *6*, 89-89.

7. Sousa, S. F.; Ribeiro, A. J. M.; Neves, R. P. P.; Brás, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J., Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. *WIREs Comput. Mol. Sci.* **2017,** *7* (2), e1281.

8. Ion, B. F.; Meister, P. J.; Gauld, J. W., Multiscale Computational Study on the Catalytic Mechanism of the Nonmetallo Amidase Maleamate Amidohydrolase (NicF). *J. Phys. Chem. A* **2019,** *123*, 7710-7719.

9. Lence, E.; van der Kamp, M. W.; Gonzalez-Bello, C.; Mulholland, A. J., QM/MM simulations identify the determinants of catalytic activity differences between type II dehydroquinase enzymes. *Org. Biomol. Chem.* **2018,** *16*, 4443-4455.

10. Molakarimi, M.; Mohseni, A.; Taghdir, M.; Pashandi, Z.; Gorman, M. A.; Parker, M. W.; Naderi-Manesh, H.; Sajedi, R. H., QM/MM simulations provide insight into the mechanism of bioluminescence triggering in ctenophore photoproteins. *PLoS One* **2017,** *12*, e0182317.

11. Quantum Biochemistry. Wiley-VCH, Weinheim, Germany, 2010, pp 643-666.

12. Ribeiro, A. J. M.; Santos-Martins, D.; Russo, N.; Rarnos, M. J.; Fernandes, P. A., Enzymatic Flexibility and Reaction Rate: A QM/MM Study of HIV-1 Protease. *ACS Catal.* **2015,** *5*, 5617-5626.

13. Bonk, B. M.; Weis, J. W.; Tidor, B., Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* **2019,** *141*, 4108-4118.

14. *Molecular Operating Environment (MOE)*, 2019.01; Chemical Computing Group ULC: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021.

15. Kincaid, V. A.; Sullivan, E. D.; Klein, R. D.; Noel, J. W.; Rowlett, R. S.; Snidert, M. J., Structure and Catalytic Mechanism of Nicotinate (Vitamin B3) Degradative Enzyme Maleamate Amidohydrolase from Bordetella bronchiseptica RB50. *Biochemistry* **2012,** *51*, 545-554.

16. Bond, S. D.; Leimkuhler, B. J.; Laird, B. B., The Nosé–Poincaré Method for Constant Temperature Molecular Dynamics. *J. Comput. Phys.* **1999,** *151*, 114-134.

17. Rokach, L.; Maimon, O., Clustering Methods. In *Data Mining and Knowledge Discovery Handbook*, Maimon, O.; Rokach, L., Eds. Springer US: Boston, MA, 2005; pp 321-352.

18. McGibbon, Robert T.; Beauchamp, Kyle A.; Harrigan, Matthew P.; Klein, C.; Swails, Jason M.; Hernández, Carlos X.; Schwantes, Christian R.; Wang, L.-P.; Lane, Thomas J.; Pande, Vijay S., MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015,** *109*, 1528-1532.

19. Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E., Array programming with NumPy. *Nature* **2020,** *585* (7825), 357-362.

20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É., Scikit-learn: Machine Learning in Python. 2018.

21. Rousseeuw, P. J., Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987,** *20*, 53-65.

22. Bearpark, M. J.; Ogliaro, F.; Vreven, T.; Boggio-Pasqua, M.; Frisch, M. J.; Larkin, S. M.; Morrison, M.; Robb, M. A., CASSCF calculations for photoinduced processes in large molecules: Choosing when to use the RASSCF, ONIOM and MMVB approximations. *J. Photochem. Photobiol. A: Chem.* **2007,** *190*, 207-227.

23. Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J., A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives1Dedicated to Professor Keiji Morokuma in celebration of his 65th birthday.1. *J. Mol. Struct.* **1999,** *461-462*, 1-21.

24. Humbel, S.; Sieber, S.; Morokuma, K., The IMOMO method: Integration of different levels of molecular orbital approximations for geometry optimization of large systems: Test for n-butane conformation and SN2 reaction: RCl+Cl−. *J. Chem. Phys.* **1996,** *105*, 1959-1967.

25. Maseras, F.; Morokuma, K., IMOMM: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *J. Comput. Chem.* **1995,** *16*.

26. Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K., ONIOM: A Multilayered Integrated MO + MM Method for Geometry Optimizations and Single Point Energy Predictions. A Test for Diels−Alder Reactions and Pt(P(t-Bu)3)2 + H2 Oxidative Addition. *J. Phys. Chem.* **1996,** *100*, 19357-19363.

27. Vreven, T.; Byun, K. S.; Komáromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J., Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM. *J. Chem. Theory Comput.* **2006,** *2*, 815-826.

28. Vreven, T.; Morokuma, K.; Farkas, O.; Schlegel, H. B.; Frisch, M. J., Geometry optimization with QM/MM, ONIOM, and other combined methods. I. Microiterations and constraints. *J. Comput. Chem.* **2003,** *24*, 760-9.

29. Dennington, R. K., Todd A.; Millam, John M. *GaussView, Version 5.0.9*, Semichem Inc.: Shawnee Mission, KS, 2016.

30. M. J. Frisch, G. W. Trucks., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox *Gaussian 09, Revision E.01*, Gaussian, Inc: Wallingford, CT, 2016.

31. Becke, A. D., Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993,** *98*, 5648-5652.

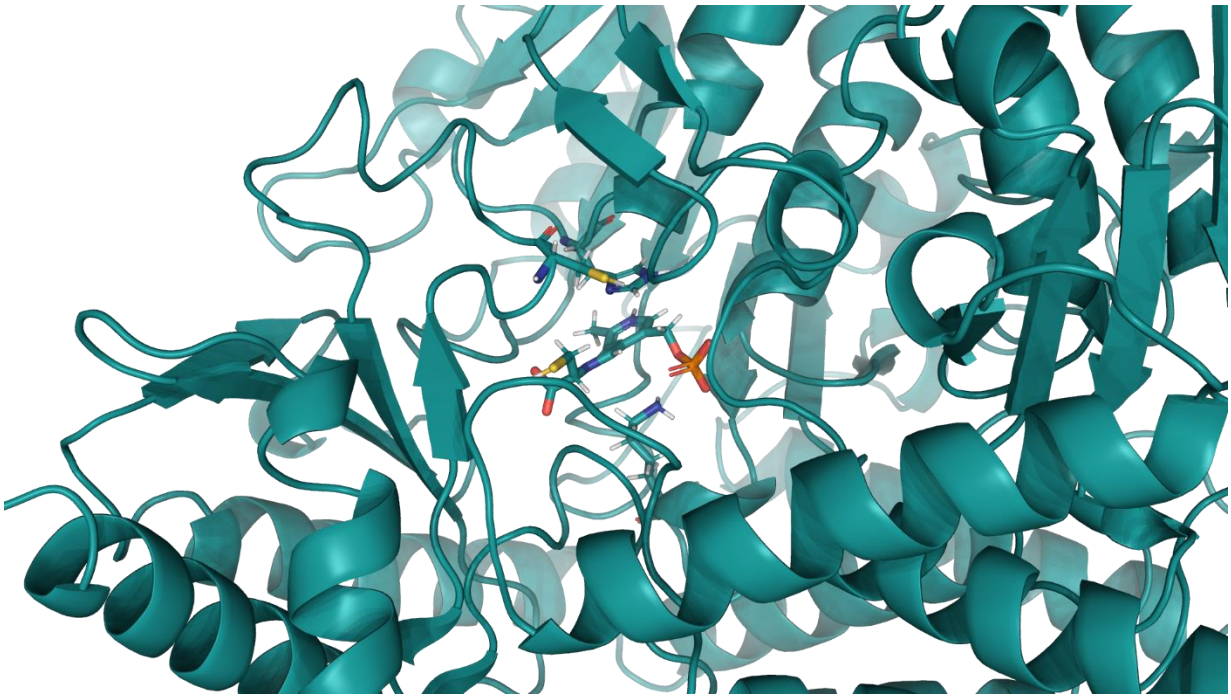
32. Becke, A. D., A new mixing of Hartree–Fock and local density-functional theories. *J. Chem. Phys.* **1993,** *98*, 1372-1377.

33. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005,** *26* (16), 1668-88.

34. 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*, 785-789.

Chapter 8

Insights into the Active Site of a *Synechocystis* sp. Cysteine Desulfurase (SufS): A Computational Investigation



8.1 Introduction

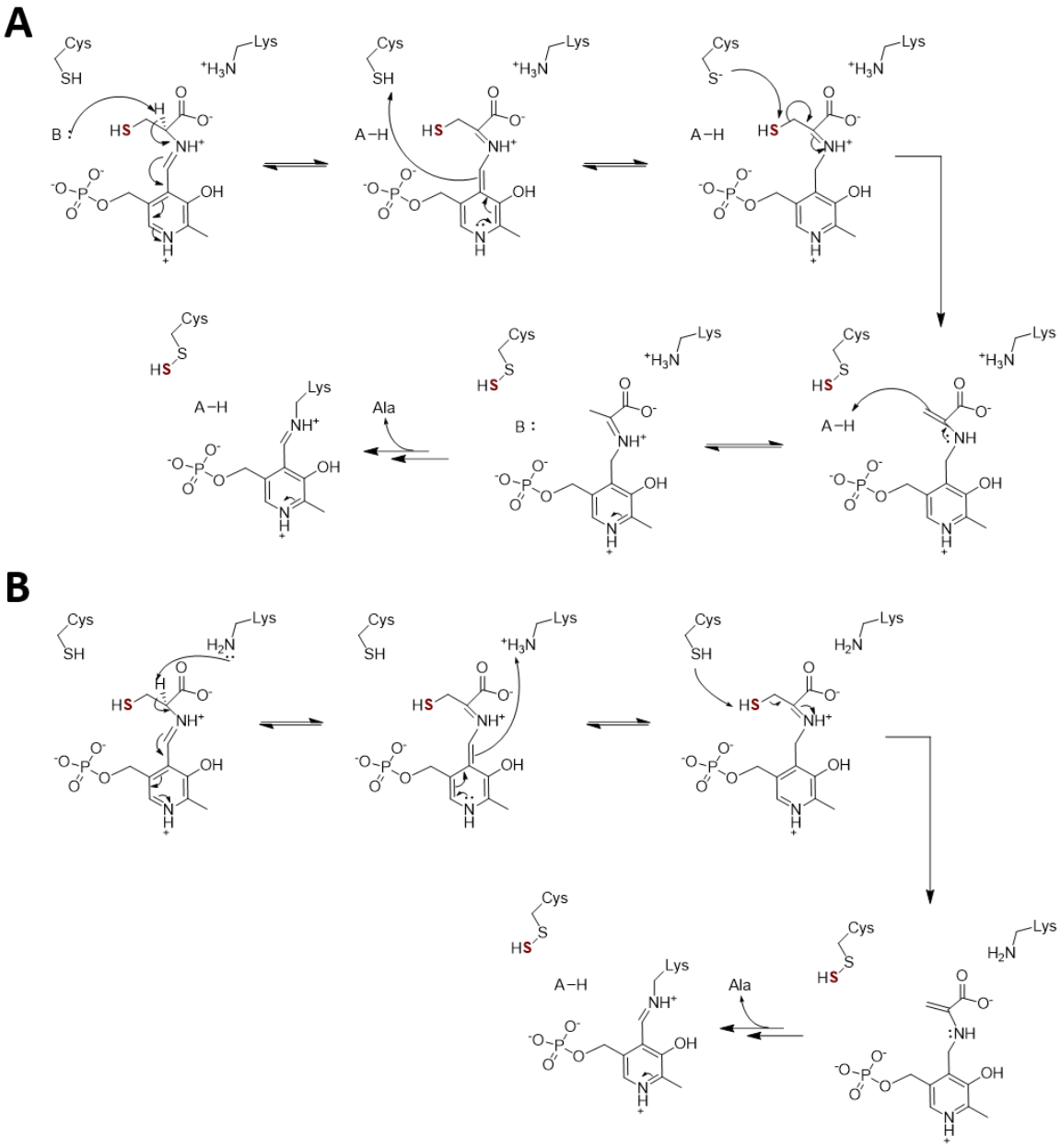
Cysteine is an essential amino acid as the source of sulfur for the biosynthesis of thiocofactors.[1-4] These biosynthesis pathways are incredibly complex and remain poorly characterized, yet persulfide intermediates are commonly observed. Compared to their thiol counterparts, persulfides have unique redox properties as a result of their additional sulfur atom.[5] For instance, their pK_a is lower and are thus more reactive at a biological pH. Most notably, they play important roles in redox signaling and oxidative stress, and, in particular, cysteine persulfide (CysSSH) is demonstrably involved in biological sulfur trafficking.[6-8]

Cysteine desulfurases are a family of enzymes often involved in iron-sulfur cluster formation which act as cofactors in the catalysis of a wide range of biologically important reactions including transamination, racemization, and β-elimination.[9] In bacteria, iron-sulfur cluster formation is governed by three main gene clusters: CSD, ISC, SUF. Each of these functions under different biological conditions and are essential for electron transfer reactions, radical mechanisms, iron sensing, and/or operate at different oxygen levels. Most importantly, the first step in each pathway utilizes a cysteine desulfurase to mobilize sulfur from the free amino acid cysteine and transfer it through a series of enzyme-bound persulfides to its intended target. Notably, the SUF pathway operates under iron starvation or oxidative stress.[10] The SUF gene cluster is an operon consisting of 6 proteins: *sufABCDSE*.[3] Of these proteins, SufS is a bacterial type II cysteine desulfurase responsible for extracting the sulfur atom from a free cysteine and transfering it to its accessory protein SufE as the first step in the SUF pathway.[11] In this reaction, an enzyme-bound cysteine-persulfide intermediate is formed, and alanine is released as product.

A key feature of SufS is the use of a pyridoxal phosphate (PLP) cofactor to help catalyze sulfur transfer. It acts by stabilizing the negative charge development at Cα position in the transition state after formation of the Schiff base with the amino acid adduct.[9] In return the

pK_a of the Cα proton is reduced drastically, facilitating its deprotonation by a nearby active site base. This affords the enzyme an extra layer of control over product formation as persulfides are known to be more reactive than their thiol counterparts. Thus, by ensuring interaction with their target and limiting access to solution, side reactions are avoided. Interestingly, bacterial SufS also shows activity towards selenocysteine, a known selective function of mammalian selenocysteine lyase.[12]

Studies on NifS, a cysteine desulfurase in the nitrogen fixation system also utilizing a PLP cofactor, have suggested the active site sulfur from a conserved cysteine residue attacks the substrate thiol.[13-15] Several mechanisms for the formation of the persulfide have been suggested, yet the identity of the catalytic base responsible for deprotonating the Cα remains elusive (Scheme 8.1). The pK_a of each proposed important active site residue is unclear and there are several interactions that influence this phenomenon. Therefore, there is potential for multiple residues to act as the catalytic base to initiate the reaction and this uncertainty is reflected in the literature. One study even suggested the role of an important water molecule coordinated by an active site His.[16] The earliest work by Zheng et al. suggested that an active site cysteine thiolate is required to form a cysteine persulfide.[17] Subsequently, they proposed mechanism for NifS, a similar sulfur transfer enzyme where the identity of the base is unknown and is summarized in Scheme 8.1a.[13] Later on, Mihara and Esaki suggested that the lysyl residue responsible for coordinating PLP can act as the base as shown in Scheme 8.1b.[18] In the latter mechanism, the protonation state of the lysyl is also in constant flux, and by contrast, the cysteinyl residue is suggested to exist as the thiol.



Scheme 8.1. Proposed mechanisms for the formation of the SufS-bound persulfide and release of the alanine product available in the literature. (A) The mechanism proposed by Black et al.[14] whereby an unknown base abstracts the Cα proton. (B) The mechanism proposed by Mihara et al.[18] whereby the lysyl residue that binds PLP to the enzyme is responsible for Cα proton abstraction.

In this study, we aim to clarify the protonation states of potentially mechanistically important active site residues as well as characterize key active site interactions that lead to the first step of the reaction catalyzed by SufS. Since the identity of the active site base residue was still up for debate, we performed molecular dynamics simulations to calculate the pK_a of each residue in question and performed QM/MM studies on the optimized reactant complex (RC) to gain further insight into the identity of the catalytic base and protonation states of these key residues. Our results indicate that the conserved cysteinyl residue has an increased pK_a and that a nearby histidyl coordinates with the PLP-cysteine adduct, potentially acting as a base to deprotonate and form a thiolate. Proton abstraction may occur through an active site water molecule.

8.2 Computational Methods

Molecular Dynamics Simulations

The 1.8 Å crystal structure of SufS (PDB ID: 1T3I)[15] was used as a template structure for MD simulations. The Molecular Operating Environment (MOE) was used to prepare structures for simulations.[19] More specifically, the octanoyl-sucrose molecule found in the active site and glycerol molecules were removed to restore the native complex. The cysteine-aldimine intermediate (the ligand) was built from the PLP since it was not present in the crystal structure. The ligand was parametrized using AmberTools16[20] and GAFF2. The resulting complex is then run through a minimization protocol. A short equilibration is then done at 10 K, followed by heating to 300 K. Production MD simulations were then run for 10 ns using AMBER16 and the ff14SB force field[20] in the NVT ensemble. SHAKE was used to constrain bonds containing hydrogen.[21] A time step of 2 fs was used in the simulations of this system. The Andersen thermostat was used with temp0 set to 300 K and no pressure scaling.[22] Complexes of both the thiol/thiolate form of Cys367 and neutral and protonated forms of Lys367 and His123 were built for simulations.

Constant pH replica exchange MD (CpHREMD) simulations were also run on the enzyme both in the presence and lack of the ligand. Setup for these simulations followed the protocols recommended by Shen and coworkers.[23-26] System preparation was done with an initial minimization period and a 50 kcal mol^{-1} $Å^{-2}$ restraint on heavy atoms, heating to 300 K with a 5 kcal mol^{-1} $Å^{-2}$ restraint on heavy atoms, and a constant pH equilibration with no exchange of protonation states for 2 ns before submitting for replica exchange. For each simulation, the GB-Neck2 (igb=8) implicit solvent model was used. Modifications were made to the mbondi3 implicit Born radii as recommended for both cysteinyl and histidyl residues.[27] Bonds containing hydrogen were constrained using SHAKE,[21] the salt concentration was set to 0.15 M, and a 2 fs timestep was used. A total of 12 replicas were used to sample pH 4-9.5 in increments of 0.5 units. Simulations were run at an effectively infinite cutoff (1000 Å) at a temperature of 300 K. Exchanges between adjacent replicas were attempted every 250 steps and replicas were run for a total of 4 ns for a total sample time of 48 ns. Cys367, Lys226, and His123 were allowed to titrate in these simulations.

QM/MM Calculations

All QM/MM calculations were performed using the ONIOM formalism in Gaussian 09 revision E.01.[28] Optimized geometries and harmonic vibrational frequencies were obtained at the ONIOM(B3LYP/6-31G(d,p):AMBER16) level of theory using mechanical embedding. QM/MM input files were prepared using the molUP[29] extension of VMD.[30] Parameters for the enzyme complex are generated for the chosen MD frame using AmberTools16.[20] These are defined explicitly within the input file including proper bond stretch, angle bending, and torsion parameters. The entire enzyme is modeled with a total system size of 12646 atoms and the QM layer containing 139 atoms. This includes the side chains of Lys226, Cys367, His123, Thr95, Cys202, Ser223, Arg362, Arg382, Thr688, 4 coordinating water molecules, and the ligand. The phosphate group on PLP was placed in the MM layer. All atoms 12 Å beyond the substrate are kept frozen. Energies for each

structure along the reaction pathway were obtained at the ONIOM(ωB97XD/6-311+G(2df,p):AMBER16) as this functional was shown to be more accurate for sulfur-containing systems.[31-32]

8.3 Results and Discussion

8.3.1 pK_a prediction of the key active site residues

The differences in the literature mechanisms raised questions about the identity of the catalytic base and the protonation state of the active site cysteinyl, lysyl, and histidyl. To address this issue, we investigated these questions using CpHREMD simulations, calculated the proton affinities of each residue in question, and obtained an optimized reactive complex (RC). Furthermore, we performed molecular dynamics simulations on the two protonation states of Cys367, Lys226, and His123 for a total of 6 simulations.

Our initial step was to visually examine the active site and determine if a catalytic base can be inferred based on proximity (Figure 8.1). Since the crystal structure did not include the substrate, a minimized structure from the prepared complex was used as the basis for comparison. As a result, each residue in question is shown in the neutral state. In both cases, the distance to the nitrogen atom is too far for catalysis (4.3 Å to Lys226 and 5.5 Å to His123). While there is certainly the potential for hydrogen bonding networks to transfer a proton, we note that there is a lack of water molecules nearby the Cα proton. Furthermore, His123 assists in the positioning of PLP through stacking interactions. From the active site configuration, we determine that His123 cannot act as a base to abstract the Cα proton, but rather Lys226 can perform this function due to its much closer proximity. Hydrogen bonding from the Schiff base keeps Lys226 near the Cα proton, although there is the potential for fluctuation from hydrogen bonding to the imine nitrogen.

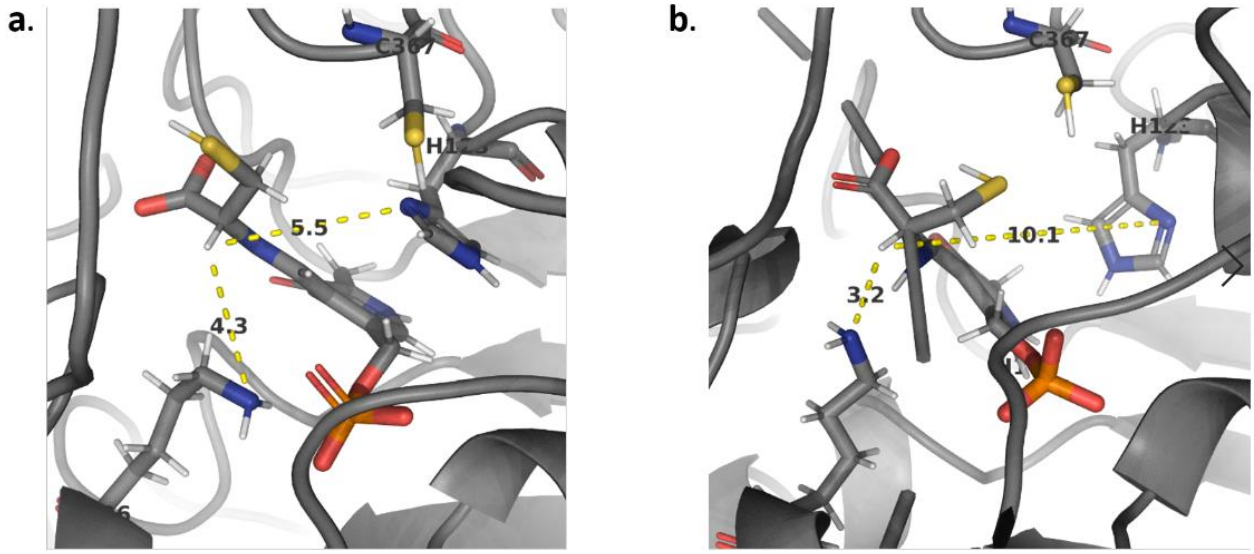


Figure 8.1 Active site configuration of (A) the prepared structure for MD simulation and (B) a selected structure from the simulation data. The distances of $_{His123}N\delta$ and $_{Lys226}H_2N$ to the Cα proton of the substrate are highlighted in yellow and shown in Å.

CpHREMD simulations were conducted on the active site to gain more insight into potential protonation states of the catalytic residues. As a broad overview, these simulations combine sampling of both conformational and protonation states, a thermodynamic property. While the system has conformational freedom, attempts to exchange between different protonation states are made with a differing pH environment. The equilibrium probability of exchange between two adjacent replicas is calculated and taken to calculate an overall exchange probability. These probabilities are fitted to the Hill equation to give a predicted pK_a and a titration curve as well as the Hill constant.[33] In our simulations, we only considered the protonation states of the key catalytic residues His123, Cys367, and Lys226. Since the substrate is a non-standard residue with no available experimental pK_a data, we chose to omit it from this analysis. Furthermore, there are several titratable groups on the substrate: the thiol group, aspartic acid group, PLP nitrogen atom, and PLP oxygen atom. To avoid misinterpretation of results, we chose to utilize a similar approach as

common pK_a prediction software whereby the instantaneous environment of the substrate thiol's environment is used to infer these protonation state.

In Figure 8.2, the titration curves for each residue are shown in the presence and absence of the cysteinyl-PLP substrate. Interestingly, the calculated pK_a of Cys367 and Lys226 are very similar regardless of the presence of substrate, and only slightly decrease from 10.15 to 9.94 and 10.02 to 9.88, respectively. By contrast, in the presence of the substrate, the pK_a of His123 increases by 1 unit, more closely matching the solvent-accessible environment and and may suggest that a significant amount of time is spent as $_{His123}H^+$. From Figure 8.1, $_{His123}N\delta$ is hydrogen bonded to the substrate thiol group. As such, His123 has the potential to act as a catalytic base, but rather than deprotonation of the susbstrate Cα, it can deprotonate the substrate thiol group to generate a reactive thiolate. This would result in the thiol/thiolate interaction proposed by both mechanisms; however, the substrate is the thiolate rather than the active site Cys367. The titration curves for Cys367 are also indicative of this feature as its pK_a remains relatively unaffected regardless of the presence or absence of the substrate and is elevated compared to a free cysteine thiol.

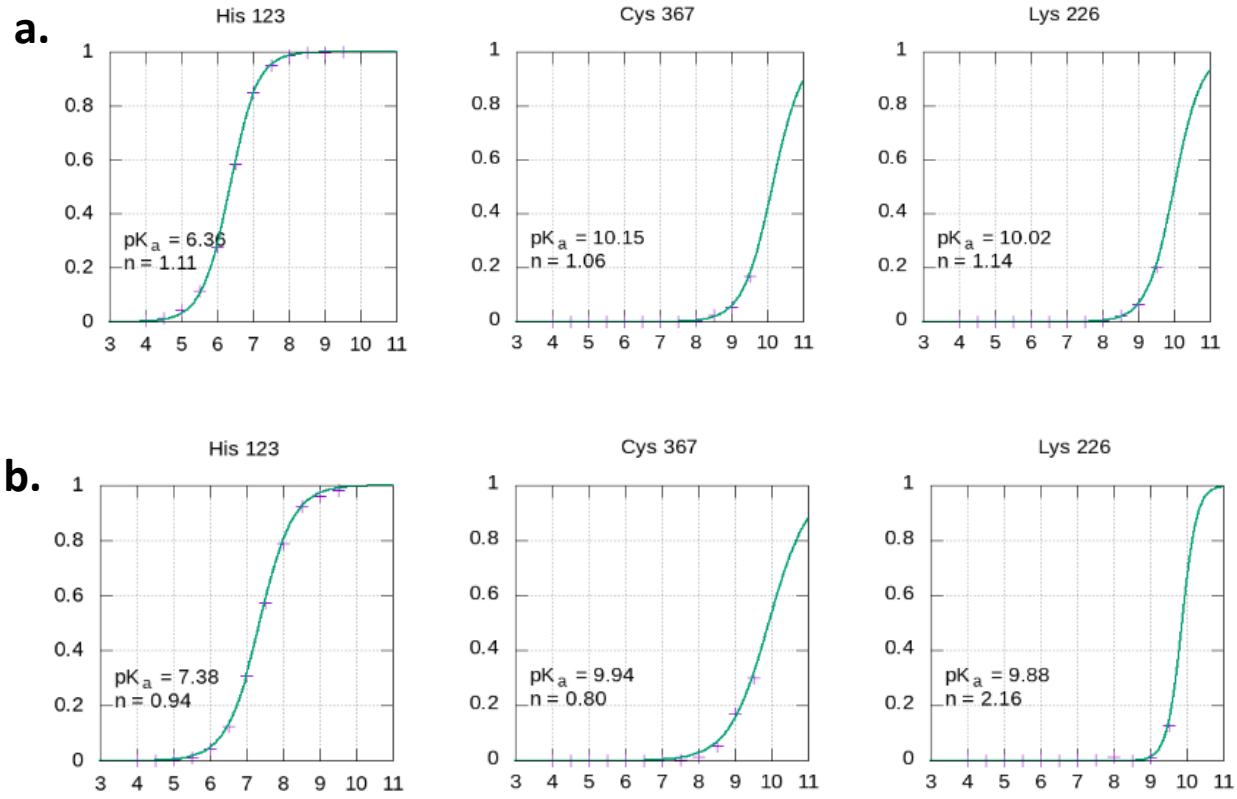


Figure 8.2. Titration curves for His123, Cys367, and Lys226 (A) without cysteine or PLP in the active site and (B) cysteinyl-PLP adduct.

8.3.2 Active site interactions of the catalytic residues

To gain further understanding into the influence of the Cys367 protonation state on the active site, we performed molecular dynamics simulations on the substrate-bound enzyme with Cys367 as both a thiol and as a thiolate. Since the pK_a of Cys367 is relatively unaffected by the presence of substrate, we sought to examine the active site environment to determine the reasons for this observation. We examined the stability of the active site and the ligand in both protonation states throughout the simulations to measure the localized effect of Cys367's protonation state (Figure 8.3). Notably, the protein is equilibrated in both states and few fluctuations are observed as indicated by the backbone RMSD. Focusing on the active site environment, the major fluctuations are observed within

the ligand. In the thiol state (Figure 8.3a), there is some flux observed in ligand orientation which is mostly attributed to the hydrogen bonding environment. The presence of a phosphate group and several nearby hydrogen bonding partners can cause the substrate to shift with larger degrees of freedom, especially since at this point PLP is bound to the enzyme through the Schiff base linkage to Lys226. However, when Cys367 is deprotonated, the ligand takes on a few distinct conformations with less flux and can shift through them quickly.

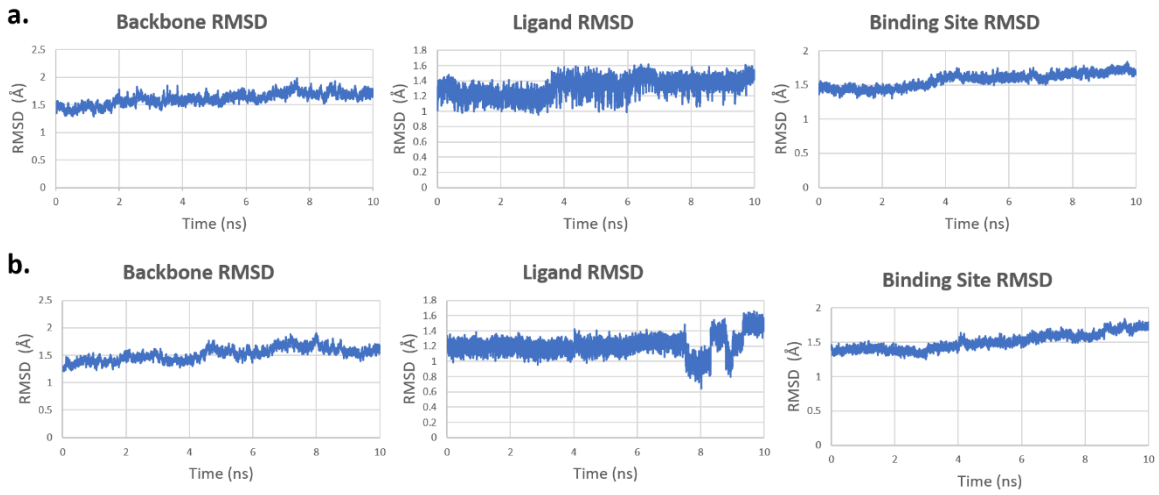


Figure 8.3. Backbone, ligand, and binding site RMSDs for MD simulations performed in the (A) thiol and (B) thiolate states of Cys367. Binding site RMSDs exclude the ligand and are evaluated in a 7 Å radius around the ligand.

Due to the significant flux of the ligand, we also examined the effect of this fluxionality on two key bond distances along the proposed reaction coordinate in Scheme 8.1b, namely, the sulfur-sulfur distance and the distance of $_{Lys226}$N to the Cα proton (Figure 8.4). Comparison of the S—S distance in both cases reveals that the conformational change of the ligand in the thiolate state results in a longer distance which would significantly increase the energy barrier for sulfur transfer. We also notice there is significantly less fluctuation of the distance in the thiolate state due to the added hydrogen bonding stability afforded by

His123 and nearby water molecules. In both cases, however, the distance does not decrease below 3 Å. Notably, this effect also extends into the first proton abstraction coordinate whereby fluctuation in ligand positioning caused an increase in the $_{Lys226}$N-Hα_{Ligand} distance.

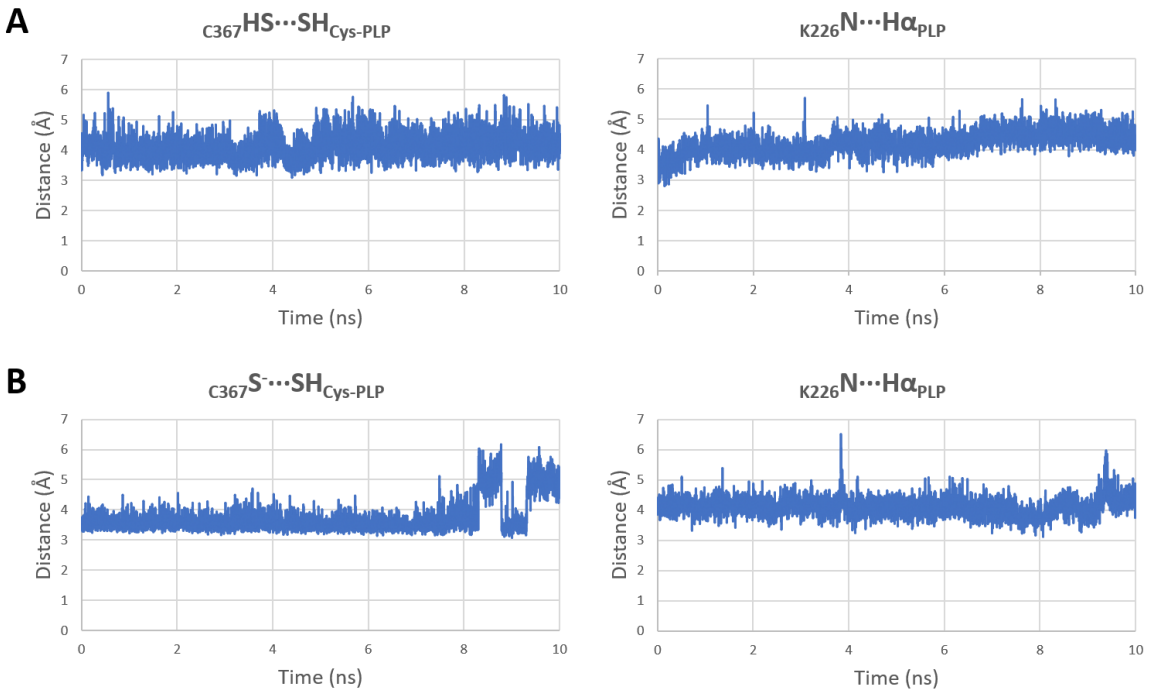


Figure 8.4. Key bond distances highlighted along the proposed reaction coordinate (Scheme 8.1b) in the (A) thiol and (B) thiolate state of Cys367.

Furthermore, we also performed MD simulations on the first proposed intermediate (deprotonated substrate Cα, neutral His123) in both the thiol and thiolate states of Cys367 with the assumption that Lys226 acts as the catalytic base in the first proposed reaction step. After the first step, a proton transfer is proposed from an active site acidic residue to the C4 atom on the linked PLP, facilitating C—S bond cleavage in the following stages. To gain further insight into the protonation state of Cys367 at this stage, we analyzed the distance between the two sulfur atoms throughout the course of both simulations to

determine if either protonation state encourages their interaction (Figure 8.5). When Cys367 is modeled as a thiol, the sulfur-sulfur distance can be as low as 3 Å compared to the thiolate state where the two sulfur atoms are an average of 5 Å apart. Therefore, we determined that in the early stages of the mechanism, Cys367 is likely a thiol as it encourages sulfur-sulfur interactions. However, deprotonation of Cys367 may occur beyond the second step of the mechanism since it is solvent accessible, and this can occur through a water molecule.

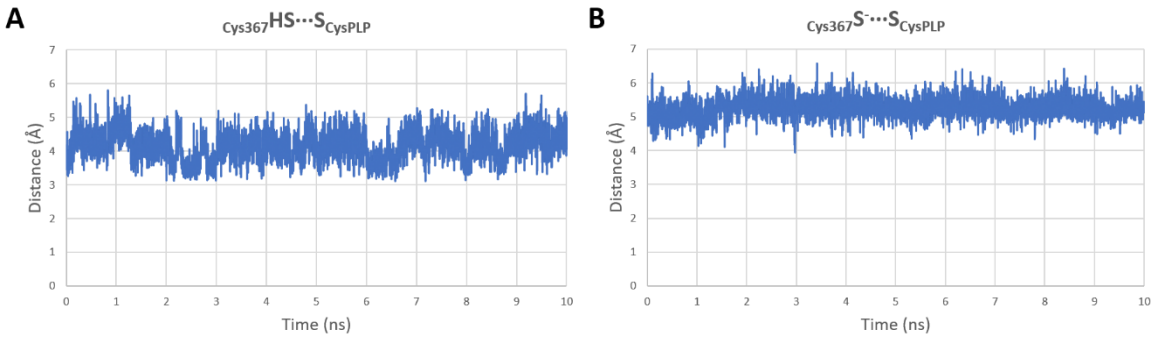


Figure 8.5. Sulfur-sulfur distance of the catalytic Cys367 to the PLP-bound substrate as simulated in the first intermediate state where Cys367 is modeled as the (A) thiol, and (B) thiolate. In each case, the ligand is also in the thiol state.

8.3.3 A model Reactant Complex (RC)

Based on our molecular dynamics studies, we selected a structure from the fully neutral active site (His123, $_{Lys226}NH_2$, Cys367) for optimization as the reactive complex (RC) and the result is shown in Figure 8.6. The phosphate group was omitted in these and further calculations since its inclusion resulted in its neutralization by abstracting protons from its environment. During the optimization, the S—S distance increased to 5.03 Å from the hydrogen bonding stabilization of both thiol groups; the substrate thiol is held firmly in place by the $_{His123}N\delta$ at 1.95 Å and Cys367 is hydrogen bonded to a network of water molecules. Furthermore, the distance from the Cα proton to Lys226 increased to 3.98 Å,

suggesting it is unlikely deprotonation occurs without assistance. No additional residues are found to be close to the Cα. From this stage, it is unclear how the mechanism might proceed.

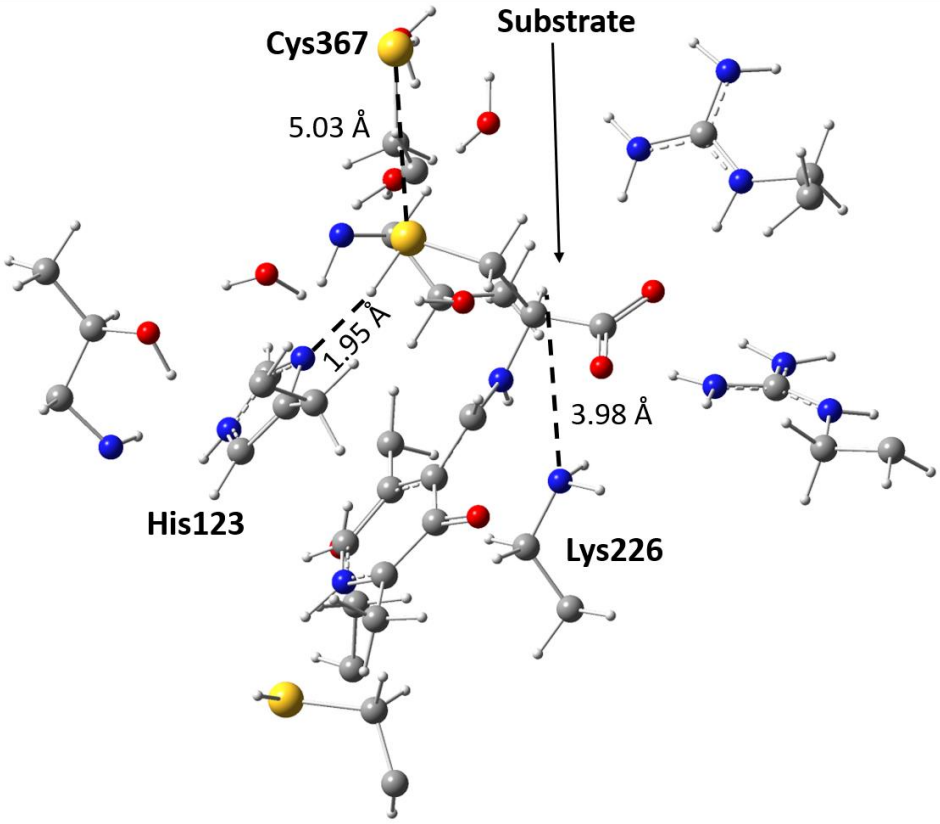


Figure 8.6. Optimized RC of SufS. Key interactions along the proposed reaction coordinate are highlighted with dashed bonds.

We attempted to further understand the protonation states of His123, Cys367, and the substrate in the optimized RC and gain further insight into a first reaction step by performing proton affinity calculations on these residues (Table 8.1). His123 has a lower proton affinity than the ligand's thiol group by approximately 448 kJ mol^{-1}, indicating deprotonation at this RC is unfavourable. In fact, the proton affinity is even higher than Cys367 by approximately 49 kJ mol^{-1} suggesting that it prefers to be protonated in the RC.

Notably, the proton affinity of the ligand's Cα is lower than the other more basic residues and is relatively more acidic in this environment. While Lys226 is only calculated to have a proton affinity of 832.2 kJ mol^{-1}, this is due to hydrogen bond stabilization from the imine hydrogen. This may indicate that PLP is not the sole factor that dictates this phenomenon, but instead assists through positioning for proper hydrogen bonding to a nearby base and likely requires assistance from a water molecule. In another study by our group member on the PLP enzyme alliin lyase, it was found that a similar deprotonation occurs through the assistance of a coordinating water.

Table 8.1. Proton affinities (in kJ mol^{-1}) of the three active site residues His123, Cys367, and Lys226, as well as the ligand's thiol group and Cα. Calculations were done in the QM/MM model.

	Proton Affinity (kJ mol^{-1})
$_{His123}N\delta$	1376.0
$_{Cys367}SH$	1775.0
$_{Lys226}NH_2$	832.2
$_{Ligand}SH$	1824.5
$_{Ligand}C\alpha$	1687.3

We further conducted QTAIM analysis on the optimized RC to examine existing stabilizing interactions, particularly around the Cα proton. Importantly, there is no bond path observed to any active site residue. Instead, the proposed Lys226 is hydrogen bonded to the imine hydrogen. This may occur for several reasons. Firstly, a water molecule may be required to facilitate deprotonation. MD simulations did not show a water network to Lys226; however, this is not unreasonable since it is solvent accessible. Secondly, SufS activity has been shown to be dependent on its interaction with SufE through dimer interface mutations whereby its activity increased 6-fold when in contact with SufE.[11, 34] Unfortunately, the studied complex was a homodimer and the PDB does not currently contain any structures

of the SufS-SufE complex. Further studies into the catalytic mechanism of SufS would benefit from this crystal structure. Additionally, the ligand's sulfur atom has many stabilizing interactions suggesting there is potential for it to become deprotonated at some point in the catalytic mechanism, similarly to the stabilization provided by the oxyanion hole observed for NicF in **Chapter 5**.

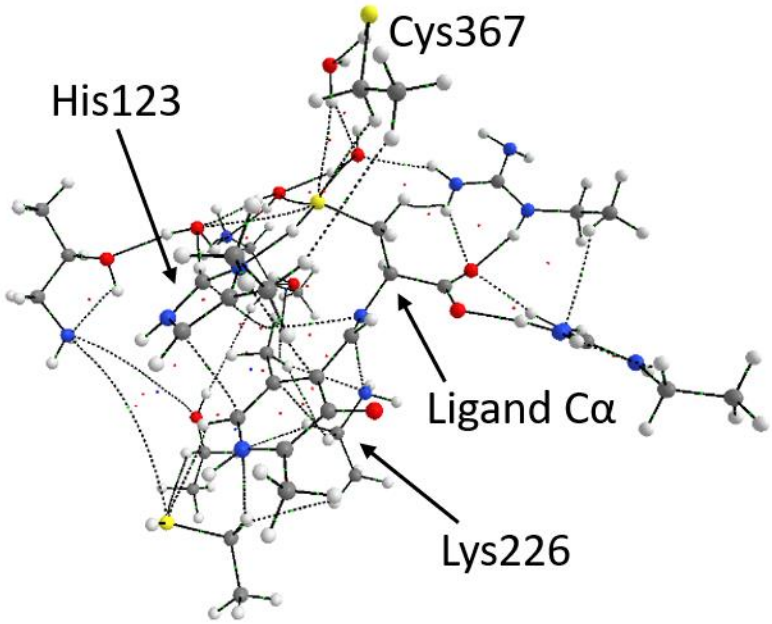


Figure 8.7. Molecular graph of the SufS RC. Solid and dashed lines represent bond paths, reflective of the relative strength of the interaction between two atoms. Green dots represent bond critical points along each bond path. Red dots represent ring critical points.

8.4 Conclusions

The active site of SufS was studied using molecular dynamics and QM/MM to determine the protonation states of the key active site residues His123, Lys226, and Cys367 and clarify the identity of the catalytic base. Using CpHREMD simulations, we found that the pK_a of His123 increases from 6.36 to 7.38 in the presence of the substrate. By contrast, the pK_a of Cys367 and Lys226 are relatively unaffected and rather decrease from 10.15 to 9.94 and 10.02 to 9.88, respectively. From visual inspection of the active site, we determined

His123 could not act as the catalytic base because of its distance to the ligand's Cα proton. Instead, it participates in hydrogen bonding with the substrate thiol, indicated this thiol group may become deprotonated at some point in the reaction. Rather, it is more likely that Lys226 acts as the catalytic base due to its proximity to this proton. We further conducted MD simulations to examine the effect of each protonation state on key active site distances along the proposed reaction mechanism, namely the sulfur-sulfur distance and $_{Lys226}H_2N\cdots H\alpha_{Ligand}$. Throughout the course of the simulations, the neutral states of His123, Lys226, and Cys367 are found to be preferred since they minimize the sulfur-sulfur interaction distance from 5 Å to 3 Å. Thus, a model for the reactive complex was chosen and optimized to reflect these observations. This RC revealed that Lys226 is not well positioned to abstract the Cα proton from the ligand at 3.98 Å. Studies in our group on a related PLP system showed that this deprotonation is water-assisted. A water molecule was found between a similar proton and the corresponding lysyl which may suggest that SufS operates in the same way. These studies set the foundation for future mechanistic studies by highlighting key features of the active site that influence catalysis.

8.5 References

1. Begley, T. P.; Xi, J.; Kinsland, C.; Taylor, S.; McLafferty, F., The enzymology of sulfur activation during thiamin and biotin biosynthesis. *Curr. Opin. Chem. Biol.* **1999,** *3* (5), 623-9.

2. Marquet, A., Enzymology of carbon-sulfur bond formation. *Curr. Opin. Chem. Biol.* **2001,** *5*, 541-9.

3. Hidese, R.; Mihara, H.; Esaki, N., Bacterial cysteine desulfurases: versatile key players in biosynthetic pathways of sulfur-containing biofactors. *Appl. Microbiol. Biotechnol.* **2011,** *91*, 47-61.

4. Mueller, E. G., Trafficking in persulfides: delivering sulfur in biosynthetic pathways. *Nat. Chem. Biol.* **2006,** *2*, 185-194.

5. Fujii, S.; Sawa, T.; Motohashi, H.; Akaike, T., Persulfide synthases that are functionally coupled with translation mediate sulfur respiration in mammalian cells. *Br. J. Pharmacol.* **2019,** *176*, 607-615.

6. Nishida, M.; Sawa, T.; Kitajima, N.; Ono, K.; Inoue, H.; Ihara, H.; Motohashi, H.; Yamamoto, M.; Suematsu, M.; Kurose, H.; van der Vliet, A.; Freeman, B. A.; Shibata, T.; Uchida, K.; Kumagai, Y.; Akaike, T., Hydrogen sulfide anion regulates redox signaling via electrophile sulfhydration. *Nat. Chem. Biol.* **2012,** *8*, 714-24.

7. Ida, T.; Sawa, T.; Ihara, H.; Tsuchiya, Y.; Watanabe, Y.; Kumagai, Y.; Suematsu, M.; Motohashi, H.; Fujii, S.; Matsunaga, T.; Yamamoto, M.; Ono, K.; Devarie-Baez, N. O.; Xian, M.; Fukuto, J. M.; Akaike, T., Reactive cysteine persulfides and S-polythiolation regulate oxidative stress and redox signaling. *Proc. Natl. Acad. Sci. USA* **2014,** *111*, 7606-11.

8. Ihara, H.; Kasamatsu, S.; Kitamura, A.; Nishimura, A.; Tsutsuki, H.; Ida, T.; Ishizaki, K.; Toyama, T.; Yoshida, E.; Abdul Hamid, H.; Jung, M.; Matsunaga, T.; Fujii, S.; Sawa, T.; Nishida, M.; Kumagai, Y.; Akaike, T., Exposure to Electrophiles Impairs Reactive Persulfide-Dependent Redox Signaling in Neuronal Cells. *Chem. Res. Toxicol.* **2017,** *30*, 1673-1684.

9. Eliot, A. C.; Kirsch, J. F., Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations. *Annu. Rev. Biochem.* **2004,** *73*, 383-415.

10. Boyd, E. S.; Thomas, K. M.; Dai, Y.; Boyd, J. M.; Outten, F. W., Interplay between Oxygen and Fe–S Cluster Biogenesis: Insights from the Suf Pathway. *Biochemistry* **2014,** *53*, 5834-5847.

11. Kim, D.; Singh, H.; Dai, Y.; Dong, G.; Busenlehner, L. S.; Outten, F. W.; Frantom, P. A., Changes in Protein Dynamics in Escherichia coli SufS Reveal a Possible Conserved Regulatory Mechanism in Type II Cysteine Desulfurase Systems. *Biochemistry* **2018,** *57* (35), 5210-5217.

12. Mihara, H.; Maeda, M.; Fujii, T.; Kurihara, T.; Hata, Y.; Esaki, N., A nifS-like gene, csdB, encodes an Escherichia coli counterpart of mammalian selenocysteine lyase. Gene cloning, purification, characterization and preliminary x-ray crystallographic studies. *J. Biol. Chem.* **1999,** *274*, 14768-72.

13. Zheng, L.; White, R. H.; Cash, V. L.; Dean, D. R., Mechanism for the Desulfurization of L-Cysteine Catalyzed by the nifS Gene Product. *Biochemistry* **1994,** *33* (15), 4714-4720.

14. Black, K. A.; Dos Santos, P. C., Shared-intermediates in the biosynthesis of thio-cofactors: Mechanism and functions of cysteine desulfurases and sulfur acceptors. *Biochim. Biophys. Acta* **2015,** *1853*, 1470-80.

15. Tirupati, B.; Vey, J. L.; Drennan, C. L.; Bollinger, J. M., Kinetic and Structural Characterization of Slr0077/SufS, the Essential Cysteine Desulfurase from Synechocystis sp. PCC 6803. *Biochemistry* **2004,** *43*, 12210-12219.

16. Lima, C. D., Analysis of the E. coli NifS CsdB protein at 2.0 A reveals the structural basis for perselenide and persulfide intermediate formation. *J. Mol. Biol.* **2002,** *315*, 1199-208.

17. Zheng, L.; White, R. H.; Cash, V. L.; Jack, R. F.; Dean, D. R., Cysteine desulfurase activity indicates a role for NIFS in metallocluster biosynthesis. *Proc. Natl. Acad. Sci. USA* **1993,** *90*, 2754-2758.

18. Mihara, H.; Esaki, N., Bacterial cysteine desulfurases: their function and mechanisms. *Appl. Microbiol. Biotechnol.* **2002,** *60*, 12-23.

19. *Molecular Operating Environment (MOE)*, Chemical Computing Group ULC, 1010: Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 237, 2020.

20. D.A. Case, K. B., I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M.

Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York and P.A. Kollman *AMBER 2016,* University of California, San Francisco, 2016.

21. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977,** *23*, 327-341.

22. Andrea, T. A.; Swope, W. C.; Andersen, H. C., The role of long ranged forces in determining the structure and properties of liquid water. *J. Chem. Phys.* **1983,** *79*, 4576-4584.

23. Harris, R. C.; Liu, R.; Shen, J., Predicting Reactive Cysteines with Implicit-Solvent-Based Continuous Constant pH Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **2020,** *16*, 3689-3698.

24. Harris, R. C.; Shen, J., GPU-Accelerated Implementation of Continuous Constant pH Molecular Dynamics in Amber: pKa Predictions with Single-pH Simulations. *J. Chem. Inf. Model.* **2019,** *59*, 4821-4832.

25. Huang, Y.; Harris, R. C.; Shen, J., Generalized Born Based Continuous Constant pH Molecular Dynamics in Amber: Implementation, Benchmarking and Analysis. *J. Chem. Inf. Model.* **2018,** *58*, 1372-1383.

26. Liu, R.; Yue, Z.; Tsai, C.-C.; Shen, J., Assessing Lysine and Cysteine Reactivities for Designing Targeted Covalent Kinase Inhibitors. *J. Am. Chem. Soc.* **2019,** *141*, 6553-6560.

27. Liu, R.; Yue, Z.; Tsai, C. C.; Shen, J., Assessing Lysine and Cysteine Reactivities for Designing Targeted Covalent Kinase Inhibitors. *J. Am. Chem. Soc.* **2019,** *141*, 6553-6560.

28. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox *Gaussian 09*, Gaussian Inc: Wallingford CT, 2009.

29. S. Fernandes, H.; Ramos, M. J.; M. F. S. A. Cerqueira, N., molUP: A VMD plugin to handle QM and ONIOM calculations using the gaussian software. *J. Comput. Chem.* **2018,** *39*, 1344-1353.

30. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J. Mol. Graphics* **1996,** *14*, 33-38.

31. Nikoo, S.; Meister, P. J.; Hayward, J. J.; Gauld, J. W., An Assessment of Computational Methods for Calculating Accurate Structures and Energies of Bio-Relevant Polysulfur/Selenium-Containing Compounds. *Molecules* **2018,** *23*.

32. Ion, B. F.; Meister, P. J.; Gauld, J. W., Multiscale Computational Study on the Catalytic Mechanism of the Nonmetallo Amidase Maleamate Amidohydrolase (NicF). *J. of Phys. Chem. A* **2019,** *123*, 7710-7719.

33. Itoh, S. G.; Damjanovic, A.; Brooks, B. R., pH replica-exchange method based on discrete protonation states. *Proteins* **2011,** *79*, 3420-36.

34. Charan, M.; Singh, N.; Kumar, B.; Srivastava, K.; Siddiqi, M. I.; Habib, S., Sulfur mobilization for Fe-S cluster assembly by the essential SUF pathway in the Plasmodium

falciparum apicoplast and its inhibition. *Antimicrob. Agents Chemother.* **2014,** *58*, 3389-3398

Chapter 9:

Conclusions and Future Work

9.1 Conclusions

In this dissertation, we developed, validated, and applied new approaches for reliable computational studies into sulfur chemistry. Each chapter built upon the list of available techniques by introducing different tools, starting from the most basic QM-cluster to QM/MM, to the application of many types of MD simulations, and even incorporating machine learning. Combining some or all these techniques is shown to be a potent approach to gain fundamental insights not only into the chemistry of sulfur species but can be applied to any chemical system.

Beginning in **Chapter 3**, we evaluated DFT's ability to accurately describe structure and energetics of biorelevant sulfur/selenium molecules. Examination of the literature revealed a lack of a consistent methodology applied to these compounds and the context behind method selection was not clearly described. This was especially apparent for per/polysulfide species. Inspired by the detection of these chemical species in enzymes vital to biological function, we were motivated to explore commonly applied DFT methods and their applicability to these types of chemical systems at a fundamental level. Our findings showed that sulfur in particular is sensitive to the choice of method and basis set; changing these had significant effects on the geometries and thermochemistry of these compounds. A proper balance between basis set size and choice of functional proved to be critical for accurate geometric and energetic calculations. Even the most recent functional tested in this study, M08HX, showed relatively little improvement over existing functionals, demonstrating the importance of benchmark studies to understand where and how to apply a computational method before applying it.

These challenges were further accentuated in **Chapter 4** in the study of biological per/polysulfide formation pathways. These are incredibly complex and are often predicted to contain sulfur in many different oxidation states, including as radical anions.

Stabilization provided to these charged radical species is often transient, making experimental detection difficult. Computational chemistry has become an established approach for the characterization and identification of such unstable species. Unfortunately, however, traditional DFT is designed to study ground state molecules and not very applicable to these species where contributions from high energy states provide this stabilization. Using an *ab initio* method is possible, albeit restricted to small systems (< 20 atoms). As was shown in **Chapter 3**, calculating energetics for these compounds is a difficult task and relies on selection of a proper methodology. Future development of DFT functionals should consider per/polysulfides as part of the benchmark sets to address these issues.

In **Chapter 5**, we apply these learnings to the enzyme NicF. Unlike a typical amidohydrolase, it lacks a metal center which facilitates the reaction by positioning the substrate for attack and lowering the pK_a of a ligated water molecule. Instead, NicF relies on active site residues to perform these functions. We used a multiscale approach (MD, QM/MM, and QTAIM) to study this active site and determine the key interactions that drive the reaction. In our MD simulations, we observed the formation of a putative oxyanion through a critical threonyl. Using QM/MM, we found that hydrogen bonding from the threonyl to the amide oxygen atom facilitates a hybridization change of the nitrogen atom through protonation by the catalytic Asp29. QTAIM revealed that there was additional stabilization provided to the oxyanion hole in addition to hydrogen bonding and that atomic interactions can be observed along a proposed reaction coordinate as a preparatory step. Together, the combination of these techniques contributed to a greater understanding of this active site which would not be possible through their individual application. This approach can be applied to gain fundamental insights into enzymatic catalysis.

Selecting a structure from MD for analysis using QM/MM methods was done in a deliberate and informed way in **Chapter 5**. However, this was not the only approach found in the literature. Indeed, such stabilization may not always be detected in MD simulations and structure selection must proceed through alternative means. **Chapter 6** analyses the application of these approaches to the NicF system. Poor structure selections were problematic in ways that were not immediately obvious from visual inspection. Instead, this manifested as erroneous potential energy surfaces arising from a destabilized MM environment. From these studies, we found that decomposing the ONIOM energies is a good way to check that the system is performing well. In addition to MD-derived structures, we also used the x-ray structure as a starting point to determine if MD is necessary at all. We showed that MD is required to produce a relaxed structure that has optimal hydrogen bonding orientation rather than take the crystal structure from the PDB. Crystal structures can have multiple mutations present, may be missing a substrate, have regions missing from the model, and are generally obtained from extreme conditions. These challenges can all be addressed using *in silico* methods and is vital to ensure proper modeling of an enzymatic reaction. In particular, these findings have significant impact on the methodology used for the rational drug design targeting enzymes.

Given the importance of selecting a proper model, we took that a step further in **Chapter 7** to examine the ability for machine learning to identify structures. Recent literature is tending towards the use of longer simulations, further complicating analysis. Thousands of poses are generated with complicated patterns in simulation arising from different types of protein motions. Unfortunately, only a small set of structures can be selected for QM/MM analysis due to the large time constraint associated with system preparation. As a result, significant portions of simulation data may be overlooked in attempt to find the most interesting regions. Through machine learning, an algorithm can attempt to find a pattern regardless of the amount of data. Thus, we created the program “Pose Selector” which can

read in simulation data and output a set of "catalytically viable" and "non-viable" poses. We tested this approach on the NicF simulation data from **Chapter 5** where we compared machine learning-selected structures to the one selected by cluster analyses. A total of 8 active sites were present, and these all produced a different number of "catalytically viable" poses with high reliability. Constructing potential energy surfaces from these structures revealed that there were no issues in the MM layer of the protein and that there are potentially many catalytic pathways accessible to the enzyme. Future work will entail testing the applicability of Pose Selector to other enzymes, particularly those containing more than one substrate and/or a cofactor. Furthermore, as databases grow, and more simulation data is made available, more complex algorithms such as neural networks can potentially be applied.

Lastly, in **Chapter 8**, we applied the sum of our knowledge and used it to investigate the enzymatic case of persulfide formation in SufS. In **Chapter 4**, QM-cluster studies showed how an electron transfer cofactor assists in the formation of polysulfides through facilitating radical intermediates, but this is not possible for SufS. Instead, PLP is used to increase the acidity of a cysteine's Cα proton whereby a nearby catalytic base can abstract it. Several proposed mechanisms are available in the literature; however, they were unclear on the identity of the catalytic base. His123 and Lys226 were suggested, but the pK_a of each residue was unknown. To address this problem, we applied our multiscale approach, including the new method of CpHREMD simulations. These simulations allow us to create titration curves for each residue to determine their protonation state. The active site environment was found to be basic, the pK_as of each residue were modified to align more closely in presence of substrate. For instance, the pK_a of His123 increases and the pK_as of Lys226 and Cys367 were relatively close. Using QM/MM, proton affinities of these residues were calculated and the substrate's Cα proton was found to be more acidic by comparison. With a reactant complex identified and an understanding of how the enzyme

prepares for sulfur transfer, this study facilitates future mechanistic studies. In addition, small but significant chemical changes to PLP can be modeled to show their influence on the reaction.

9.2 Future Work

In the previous section, we discussed future work as it pertains to each chapter specifically. However, given the increasing usage of artificial intelligence within computational chemistry, there is the potential for many exciting applications across a broad range of enzymology research. For example, we can apply machine learning to known published reactions for the same family of enzymes to highlight small but notable mechanistic differences between them, such as for cysteine proteases. Using such approaches, we can identify the impact of the protein environment or examine the roles of active site residues that differ between each enzyme. These insights can be extended to the study of enzyme evolution where these differences help guide the design enzymes for specific catalytic functions. Furthermore, a molecular dynamics database similar to the PDB can be created to help curate data for enzymes to create accessible computational research which can be applied to the growing problem of structure prediction. Beyond gaining key insights into enzymatic catalysis, these approaches can be leveraged to improve the computational enzymology workflow through automating difficult aspects. All these improvements can have a significant impact in the drug development industry through gaining a broad range of specific insights, an essential requirement for targeted drug design.

Appendices

Appendix A

Appendix A3.1. Selected optimized bond lengths in ångström (Å) for CH_3XXH and CH_3XX^- (X=S,Se) using MP2.

		CH_3SSH			**CH_3SS^-**		**CH_3SeSeH**			**CH_3SeSe^-**	
Method	**Basis set**	C–S	S–S	S–H	C–S	S–S	C–Se	Se-Se	*Se*–H	C–Se	Se-Se
MP2	6-31G(d)	1.812	2.062	1.346	1.812	2.089	1.962	2.328	1.489	1.965	2.353
	6-311G(d)	1.808	2.073	1.346	1.807	2.104	1.961	2.348	1.483	1.965	2.374
	6-311G(d,p)	1.805	2.073	1.338	1.804	2.103	1.957	2.348	1.464	1.961	2.374
	6-311+G(d,p)	1.805	2.074	1.339	1.805	2.098	1.957	2.348	1.464	1.962	2.371
	6-311G(2d,p)	1.818	2.082	1.337	1.819	2.107	1.952	2.350	1.464	1.958	2.380
	6-31G11(df,p)	1.799	2.044	1.341	1.798	2.067	1.947	2.321	1.470	1.952	2.343
	6-311+G(2df,p)	1.809	2.053	1.345	1.810	2.069	1.948	2.319	1.469	1.954	2.340
	6-311++G(3df,3pd)	1.802	2.042	1.338	1.802	2.059	1.946	2.322	1.465	1.952	2.341

Appendix A3.2. C–S and S–S bond lengths (ångström, Å) for CH_2CHSSH and CH_2CHSS^-.

		CH_2CHSSH		CH_2CHSS^-	
Method	**Basis set**	C–S	S–S	C–S	S–S$^-$
B3LYP	6-31G(d)	1.781	2.099	1.764	2.111
	6-311G(d)	1.779	2.111	1.762	2.131
	6-311+G(d)	1.778	2.109	1.756	2.124
	6-311+G(d,p)	1.778	2.110	1.755	2.123
	6-311G(2d,p)	1.777	2.096	1.759	2.109
	6-311+G(df,p)	1.776	2.096	1.753	2.109
	6-311+G(2df,p)	1.771	2.078	1.747	2.081
	6-311++G(3df,3pd)	1.767	2.069	1.742	2.076
B3PW91	6-31G(d)	1.772	2.078	1.753	2.088
	6-311G(d)	1.769	2.088	1.751	2.105
	6-311+G(d)	1.768	2.086	1.745	2.098
	6-311+G(d,p)	1.768	2.088	1.745	2.098
	6-311G(2d,p)	1.767	2.074	1.747	2.084
	6-311+G(df,p)	1.766	2.075	1.743	2.084
	6-311+G(2df,p)	1.762	2.058	1.737	2.063
	6-311++G(3df,3pd)	1.758	2.050	1.733	2.054
ωB97XD	6-31G(d)	1.766	2.075	1.755	2.090
	6-311G(d)	1.762	2.058	1.753	2.108
	6-311+G(d)	1.758	2.050	1.748	2.102
	6-311+G(d,p)	1.766	2.075	1.748	2.101
	6-311G(2d,p)	1.762	2.058	1.750	2.088
	6-311+G(df,p)	1.769	2.072	1.746	2.088
	6-311+G(2df,p)	1.764	2.057	1.740	2.069
	6-311++G(3df,3pd)	1.760	2.049	1.736	2.061
M06-2X	6-31G(d)	1.773	2.075	1.754	2.092
	6-311G(d)	1.772	2.083	1.752	2.108
	6-311+G(d)	1.770	2.083	1.747	2.102
	6-311+G(d,p)	1.770	2.084	1.747	2.102
	6-311G(2d,p)	1.770	2.071	1.751	2.090
	6-311+G(df,p)	1.768	2.071	1.744	2.086
	6-311+G(2df,p)	1.764	2.056	1.739	2.069
	6-311++G(3df,3pd)	1.761	2.050	1.735	2.061
M08-HX	6-31G(d)	1.773	2.073	1.755	2.091
	6-311G(d)	1.771	2.082	1.753	2.106
	6-311+G(d)	1.770	2.080	1.748	2.100
	6-311+G(d,p)	1.770	2.081	1.747	2.100
	6-311G(2d,p)	1.769	2.071	1.750	2.090
	6-311+G(df,p)	1.768	2.070	1.745	2.085
	6-311+G(2df,p)	1.763	2.055	1.739	2.068
	6-311++G(3df,3pd)	1.805	2.041	1.736	2.060
QCISD	6-311+G(2df,p)	1.771	2.069	1.750	2.086

Appendix A3.3. C–Se and Se–Se bond lengths (ångström, Å) for $CH_2CHSeSeH$ and $CH_2CHSeSe^-$.

		$CH_2CHSeSeH$		$CH_2CHSeSe^-$	
Method	**Basis set**	C–Se	Se–Se	C–Se	Se–Se$^-$
B3LYP	6-31G(d)	1.921	2.347	1.908	2.362
	6-311G(d)	1.927	2.372	1.915	2.388
	6-311G(d,p)	1.926	2.373	1.914	2.388
	6-311+G(d,p)	1.926	2.373	1.912	2.386
	6-311G(2d,p)	1.922	2.375	1.911	2.391
	6-311G(df,p)	1.921	2.360	1.908	2.373
	6-311+G(2df,p)	1.919	2.358	1.905	2.371
	6-311++G(3df,3pd)	1.918	2.359	1.905	2.371
B3PW91	6-31G(d)	1.908	2.324	1.894	2.337
	6-311G(d)	1.914	2.347	1.902	2.360
	6-311G(d,p)	1.913	2.347	1.901	2.360
	6-311+G(d,p)	1.913	2.347	1.899	2.359
	6-311G(2d,p)	1.910	2.350	1.898	2.364
	6-311G(df,p)	1.908	2.334	1.894	2.346
	6-311+G(2df,p)	1.906	2.333	1.892	2.344
	6-311++G(3df,3pd)	1.906	2.331	1.892	2.344
ωB97XD	6-31G(d)	1.908	2.310	1.892	2.333
	6-311G(d)	1.914	2.337	1.901	2.358
	6-311G(d,p)	1.913	2.338	1.900	2.358
	6-311+G(d,p)	1.913	2.338	1.898	2.357
	6-311G(2d,p)	1.910	2.339	1.897	2.361
	6-311G(df,p)	1.908	2.325	1.895	2.343
	6-311+G(2df,p)	1.906	2.324	1.893	2.341
	6-311++G(3df,3pd)	1.906	2.325	1.893	2.341
M06-2X	6-31G(d)	1.906	2.313	1.893	2.331
	6-311G(d)	1.914	2.336	1.903	2.356
	6-311G(d,p)	1.914	2.337	1.902	2.356
	6-311+G(d,p)	1.914	2.337	1.901	2.354
	6-311G(2d,p)	1.911	2.340	1.900	2.359
	6-311G(df,p)	1.909	2.326	1.897	2.343
	6-311+G(2df,p)	1.907	2.325	1.895	2.340
	6-311++G(3df,3pd)	1.907	2.326	1.895	2.341
M08-HX	6-31G(d)	1.910	2.306	1.892	2.326
	6-311G(d)	1.915	2.332	1.901	2.349
	6-311G(d,p)	1.915	2.333	1.901	2.349
	6-311+G(d,p)	1.914	2.333	1.904	2.372
	6-311G(2d,p)	1.911	2.335	1.897	2.353
	6-311G(df,p)	1.910	2.320	1.896	2.338
	6-311+G(2df,p)	1.908	2.321	1.893	2.334
	6-311++G(3df,3pd)	1.907	2.321	1.893	2.335
QCISD	6-311+G(2df,p)	1.914	2.329	1.902	2.359

Appendix A3.4. C–Se, C–S and S–Se bond lengths (ångström, Å) for CH_3XYH and CH_3XY^- (X=S,Se; Y=Se,S).

		CH_3SSeH		CH_3SSe^-		CH_3SeSH		CH_3SeS^-	
Method	**Basis set**	C–S	S–Se	C–S	S–Se	C–Se	Se-S	C–Se	Se-S
B3LYP	6-31G(d)	1.837	2.219	1.836	2.266	1.973	2.226	1.985	2.234
	6-311G(d)	1.833	2.235	1.832	2.286	1.976	2.240	1.986	2.253
	6-311G(d,p)	1.834	2.237	1.833	2.286	1.975	2.242	1.986	2.253
	6-311+G(d,p)	1.834	2.237	1.833	2.281	1.975	2.241	1.988	2.244
	6-311G(2d,p)	1.832	2.227	1.832	2.270	1.973	2.231	1.983	2.244
	6-311G(df,p)	1.832	2.221	1.832	2.266	1.969	2.226	1.980	2.236
	6-311+G(2df,p)	1.826	2.213	1.826	2.250	1.968	2.216	1.981	2.221
	6-311++G(3df,3pd)	1.822	2.212	1.822	2.248	1.968	2.215	1.980	2.219
B3PW91	6-31G(d)	1.824	2.199	1.823	2.242	1.957	2.205	1.969	2.212
	6-311G(d)	1.820	2.212	1.820	2.260	1.960	2.217	1.970	2.228
	6-311G(d,p)	1.821	2.214	1.820	2.260	1.959	2.219	1.970	2.228
	6-311+G(d,p)	1.821	2.214	1.820	2.256	1.959	2.218	1.971	2.221
	6-311G(2d,p)	1.819	2.204	1.819	2.244	1.957	2.208	1.968	2.220
	6-311G(df,p)	1.819	2.199	1.819	2.241	1.953	2.204	1.964	2.212
	6-311+G(2df,p)	1.814	2.191	1.813	2.226	1.953	2.195	1.965	2.198
	6-311++G(3df,3pd)	1.810	2.190	1.809	2.224	1.953	2.194	1.964	2.196
ωB97XD	6-31G(d)	1.823	2.194	1.821	2.238	1.951	2.201	1.960	2.209
	6-311G(d)	1.820	2.207	1.818	2.256	1.947	2.190	1.962	2.224
	6-311G(d,p)	1.820	2.209	1.818	2.256	1.953	2.213	1.961	2.224
	6-311+G(d,p)	1.820	2.208	1.817	2.252	1.953	2.213	1.962	2.218
	6-311G(2d,p)	1.818	2.199	1.817	2.242	1.951	2.204	1.959	2.216
	6-311G(df,p)	1.818	2.195	1.817	2.239	1.948	2.200	1.957	2.209
	6-311+G(2df,p)	1.813	2.187	1.811	2.224	1.947	2.191	1.956	2.196
	6-311++G(3df,3pd)	1.809	2.186	1.807	2.223	1.947	2.190	1.956	2.195
M06-2X	6-31G(d)	1.822	2.193	1.820	2.236	1.950	2.199	1.958	2.210
	6-311G(d)	1.819	2.207	1.817	2.254	1.957	2.211	1.963	2.228
	6-311G(d,p)	1.819	2.208	1.818	2.254	1.956	2.212	1.963	2.228
	6-311+G(d,p)	1.819	2.208	1.817	2.250	1.956	2.212	1.964	2.220
	6-311G(2d,p)	1.817	2.200	1.816	2.239	1.954	2.203	1.960	2.221
	6-311G(df,p)	1.817	2.194	1.816	2.237	1.951	2.198	1.959	2.212
	6-311+G(2df,p)	1.812	2.188	1.811	2.221	1.950	2.190	1.959	2.198
	6-311++G(3df,3pd)	1.809	2.187	1.808	2.221	1.950	2.190	1.958	2.197
M08-HX	6-31G(d)	1.820	2.194	1.818	2.227	1.951	2.198	1.957	2.218
	6-311G(d)	1.818	2.207	1.816	2.245	1.957	2.209	1.962	2.232
	6-311G(d,p)	1.819	2.208	1.816	2.245	1.956	2.210	1.962	2.232
	6-311+G(d,p)	1.818	2.208	1.816	2.240	1.956	2.210	1.964	2.224
	6-311G(2d,p)	1.816	2.200	1.814	2.229	1.954	2.201	1.959	2.225
	6-311G(df,p)	1.817	2.195	1.815	2.228	1.951	2.197	1.958	2.217
	6-311+G(2df,p)	1.811	2.188	1.809	2.213	1.950	2.190	1.958	2.203
	6-311++G(3df,3pd)	1.809	2.187	1.807	2.212	1.951	2.189	1.958	2.201
QCISD	6-311+G(2df,p)	1.819	2.198	1.816	2.239	1.952	2.202	1.961	2.207

Appendix A3.5. Optimized S–S bond lengths (ångström, Å) for RSSSH/$^{-}$ (R=CH_3, CH_2=CH).

Method	Basis set	CH_3SSSH		CH_3SSS^-		$CH_2CHSSSH$		CH_2CHSSS^-	
		S_1–S_2	S_2–S_3	S_1–S_2	S_2–S_3	S_1–S_2	S_2–S_3	S_1–S_2	S_2–S_3
B3LYP	6-31G(d)	2.079	2.113	2.176	2.073	2.095	2.105	2.206	2.059
	6-311G(d)	2.089	2.123	2.183	2.088	2.108	2.114	2.213	2.071
	6-311+G(d)	2.088	2.122	2.176	2.085	2.104	2.113	2.206	2.069
	6-311+G(d,p)	2.088	2.124	2.176	2.085	2.103	2.115	2.206	2.068
	6-311G(2d,p)	2.075	2.108	2.147	2.077	2.091	2.101	2.169	2.064
	6-311+G(df,p)	2.076	2.112	2.152	2.077	2.089	2.104	2.179	2.062
	6-311+G(2df,p)	2.060	2.094	2.122	2.061	2.075	2.086	2.147	2.047
	6-311++G(3df,3pd)	2.052	2.087	2.110	2.053	2.066	2.080	2.135	2.040
B3PW91	6-31G(d)	2.062	2.092	2.145	2.057	2.076	2.085	2.169	2.044
	6-311G(d)	2.070	2.100	2.151	2.070	2.086	2.091	2.174	2.055
	6-311+G(d)	2.070	2.099	2.147	2.067	2.083	2.091	2.171	2.052
	6-311+G(d,p)	2.069	2.102	2.147	2.067	2.082	2.094	2.170	2.052
	6-311G(2d,p)	2.057	2.086	2.120	2.058	2.071	2.079	2.136	2.047
	6-311+G(df,p)	2.058	2.090	2.126	2.059	2.070	2.084	2.145	2.046
	6-311+G(2df,p)	2.044	2.073	2.097	2.043	2.056	2.066	2.117	2.032
	6-311++G(3df,3pd)	2.036	2.067	2.087	2.036	2.048	2.061	2.105	2.025
ωB97XD	6-31G(d)	2.061	2.084	2.122	2.063	2.073	2.077	2.136	2.053
	6-311G(d)	2.069	2.091	2.127	2.077	2.084	2.084	2.139	2.067
	6-311+G(d)	2.069	2.091	2.124	2.074	2.080	2.084	2.137	2.063
	6-311+G(d,p)	2.068	2.093	2.124	2.073	2.078	2.086	2.134	2.062
	6-311G(2d,p)	2.057	2.080	2.105	2.063	2.070	2.074	2.115	2.054
	6-311+G(df,p)	2.058	2.083	2.108	2.065	2.067	2.077	2.118	2.055
	6-311+G(2df,p)	2.044	2.068	2.087	2.048	2.054	2.062	2.098	2.039
	6-311++G(3df,3pd)	2.037	2.062	2.078	2.042	2.047	2.056	2.089	2.032
M06-2X	6-31G(d)	2.060	2.082	2.121	2.063	2.072	2.076	2.144	2.053
	6-311G(d)	2.067	2.088	2.124	2.078	2.083	2.081	2.144	2.067
	6-311+G(d)	2.067	2.088	2.121	2.074	2.080	2.082	2.143	2.063
	6-311+G(d,p)	2.066	2.090	2.121	2.074	2.079	2.083	2.143	2.063
	6-311G(2d,p)	2.056	2.078	2.103	2.065	2.070	2.071	2.118	2.055
	6-311+G(df,p)	2.055	2.079	2.104	2.065	2.067	2.074	2.119	2.055
	6-311+G(2df,p)	2.043	2.065	2.085	2.049	2.056	2.059	2.100	2.039
	6-311++G(3df,3pd)	2.036	2.060	2.077	2.042	2.048	2.055	2.092	2.032
M08-HX	6-31G(d)	2.063	2.080	2.124	2.063	2.081	2.071	2.144	2.053
	6-311G(d)	2.071	2.088	2.128	2.076	2.089	2.078	2.148	2.065
	6-311+G(d)	2.070	2.087	2.126	2.073	2.085	2.079	2.148	2.061
	6-311+G(d,p)	2.070	2.089	2.126	2.073	2.084	2.081	2.149	2.061
	6-311G(2d,p)	2.060	2.077	2.110	2.064	2.075	2.069	2.125	2.055
	6-311+G(df,p)	2.059	2.079	2.110	2.064	2.072	2.071	2.130	2.053
	6-311+G(2df,p)	2.046	2.064	2.090	2.049	2.060	2.057	2.108	2.039
	6-311++G(3df,3pd)	2.041	2.059	2.083	2.041	2.053	2.053	2.101	2.032
QCISD	6-311+G(2df,p)	2.056	2.078	2.096	2.065	2.107	2.057	2.066	2.073

Appendix A3.6. Mulliken charges on every S of RSSSH/$^{-}$ (R=CH_3, CH_2=CH).

	CH3SSSH			**CH_3SSS^-**		
Method/Basis Set	**S_1**	**S_2**	**S_3**	**S_1**	**S_2**	**S_3**
QCISD/6-311+G(2df,p)	0.15	0.05	-0.25	-0.05	-0.10	-0.77
	$CH_2CHSSSH$			**CH_2CHSSS^-**		
QCISD/6-311+G(2df,p)	-0.01	0.04	-0.20	-0.23	-0.08	-0.74
	CYS-SSSH			**CYS-SSS^-**		
ωB9XD/6-311G(2d,p)	0.00	-0.03	-0.12	-0.12	-0.13	-0.66

Appendix A3.7. Homolytic S−S bond dissociation enthalpy (BDE) of RSSH, proton affinity (PA) and gas-phase basicity (GPB) of RSS^-, and hydrogen affinity (HA) of $RSS^\bullet$ (R=CH_3, CH_2CH). All energies calculated at 298.15K and in kJ mol^{-1}.

		CH_3SSH	**CH_3SS^-**		**$CH_3SS^\bullet$**	**CH_2CHSSH**	**CH_2CHSS^-**		**$CH_2CHSS^\bullet$**
Methods	**Basis set**	BDE(S—S)	PA	GPB	HA	BDE(S—S)	PA	GPB	HA
B3LYP	6-31G(d)	224.8	1456.4	1426.1	276.4	228.1	1430.0	1398.6	274.7
	6-311G(2d,p)	235.6	1440.5	1410.1	285.6	239.2	1418.0	1386.0	259.6
	6-311G+(2df,p)	240.3	1437.7	1407.3	281.1	241.9	1411.5	1380.2	280.2
	6-311G++(3df,3pd)	241.6	1440.9	1410.6	283.5	243.2	1414.7	1383.5	282.1
B3PW91	6-31G(d)	235.4	1455.2	1424.8	274.6	238.0	1427.6	1396.5	272.8
	6-311G(2d,p)	250.3	1447.7	1417.3	283.8	253.0	1423.1	1391.5	254.5
	6-311G+(2df,p)	254.7	1445.8	1415.4	279.5	255.7	1417.7	1386.8	278.4
	6-311G++(3df,3pd)	255.8	1448.7	1418.4	281.6	256.8	1420.5	1389.7	280.1
WB97XD	6-31G(d)	244.1	1455.0	1425.1	284.0	244.1	1429.8	1397.3	283.7
	6-311G(2d,p)	258.5	1447.3	1417.5	292.0	258.5	1422.4	1395.5	291.5
	6-311G+(2df,p)	264.6	1445.5	1415.8	288.0	262.7	1420.2	1382.0	288.3
	6-311G++(3df,3pd)	265.6	1449.2	1419.6	289.8	263.5	1423.6	1388.6	289.7
M062X	6-31G(d)	252.0	1440.0	1411.4	285.4	249.8	1414.5	1381.1	283.7
	6-311G(2d,p)	264.9	1430.0	1401.6	294.4	263.0	1408.2	1374.4	292.8
	6-311G+(2df,p)	270.3	1428.5	1400.1	291.0	267.4	1403.4	1366.4	290.2
	6-311G++(3df,3pd)	272.3	1432.3	1404.2	293.9	269.1	1406.7	1368.1	292.6
M08HX	6-31G(d)	258.0	1446.1	1414.2	295.4	255.5	1419.6	1387.9	291.6
	6-311G(2d,p)	268.3	1435.5	1405.6	298.2	268.7	1414.3	1382.0	297.1
	6-311G+(2df,p)	273.6	1433.5	1404.3	294.0	274.0	1411.4	1376.4	296.2
	6-311G++(3df,3pd)	275.2	1437.7	1408.6	296.4	272.7	1412.9	1383.4	295.7
QCISD	6-311G+(2df,p)	236.3	1445.4	1415.1	282.7	242.0	1421.8	1390.7	283.5

Appendix A3.8. Homolytic S−S bond dissociation enthalpy (BDE) of RSeSeH, proton affinity (PA) and gas-phase basicity (GPB) of $RSeSe^-$, and hydrogen affinity (HA) of $RSeSe^\bullet$ (R=CH_3, CH_2CH). All energies calculated at 298.15K and in kJ mol^{-1}.

		CH_3SeSeH	**CH_3SeSe^-**		**$CH_3SeSe^\bullet$**	**$CH_2CHSeSeH$**	**$CH_2CHSeSe^-$**		**$CH_2CHSeSe^\bullet$**
Methods	**Basis set**	BDE(Se—Se)	PA	GPB	HA	BDE(Se—Se)	PA	GPB	HA
B3LYP	6-311+G(d,p)	208.0	1401.6	1371.3	272.2	211.5	1382.6	1350.9	272.0
	6-311G(df,p)	212.9	1421.0	1390.7	267.0	193.2	1400.0	1369.1	266.2
	6-311G+(2df,p)	212.0	1404.0	1373.7	267.7	214.9	1384.8	1353.2	267.1
	6-311G++(3df,3pd)	212.7	1404.3	1374.1	268.5	215.4	1385.5	1353.9	268.0
B3PW91	6-311+G(d,p)	219.2	1409.0	1378.7	270.5	222.1	1388.6	1357.4	270.2
	6-311G(df,p)	223.7	1424.2	1393.8	265.3	204.7	1401.9	1371.0	264.5
	6-311G+(2df,p)	222.9	1411.4	1381.1	266.0	225.3	1390.7	1359.7	265.4
	6-311G++(3df,3pd)	223.6	1411.8	1381.5	266.7	226.6	1392.1	1360.7	266.7
WB97XD	6-311+G(d,p)	218.0	1410.8	1380.7	276.5	221.3	1391.4	1359.6	277.1
	6-311G(df,p)	223.7	1426.5	1396.4	270.3	226.2	1405.4	1373.9	270.5
	6-311G+(2df,p)	222.6	1413.3	1383.4	270.7	225.3	1393.8	1361.1	271.0
	6-311G++(3df,3pd)	223.7	1413.6	1383.7	271.4	226.3	1391.9	1366.1	271.8
M062X	6-311+G(d,p)	203.5	1384.2	1353.6	275.3	206.2	1367.6	1335.5	277.3
	6-311G(df,p)	209.0	1397.8	1367.0	271.6	212.0	1377.0	1352.2	273.0
	6-311G+(2df,p)	207.7	1386.0	1355.3	272.2	211.0	1369.0	1336.1	273.6
	6-311G++(3df,3pd)	223.4	1387.4	1356.8	273.4	212.4	1370.5	1337.8	272.3
M08HX	6-311+G(d,p)	233.7	1380.0	1349.7	291.4	235.9	*	*	291.8
	6-311G(df,p)	237.6	1393.8	1363.4	288.1	240.7	1373.8	1339.7	288.5
	6-311G+(2df,p)	223.4	1381.9	1351.6	288.8	224.9	1362.6	1329.0	288.5
	6-311G++(3df,3pd)	223.4	1383.4	1353.2	289.5	225.7	1362.8	1337.3	290.0
QCISD	6-311G+(2df,p)	204.1	1404.3	1374.0	267.3	208.3	1386.7	1355.7	265.9

* Poor optimization of anion led to high energy conformation.

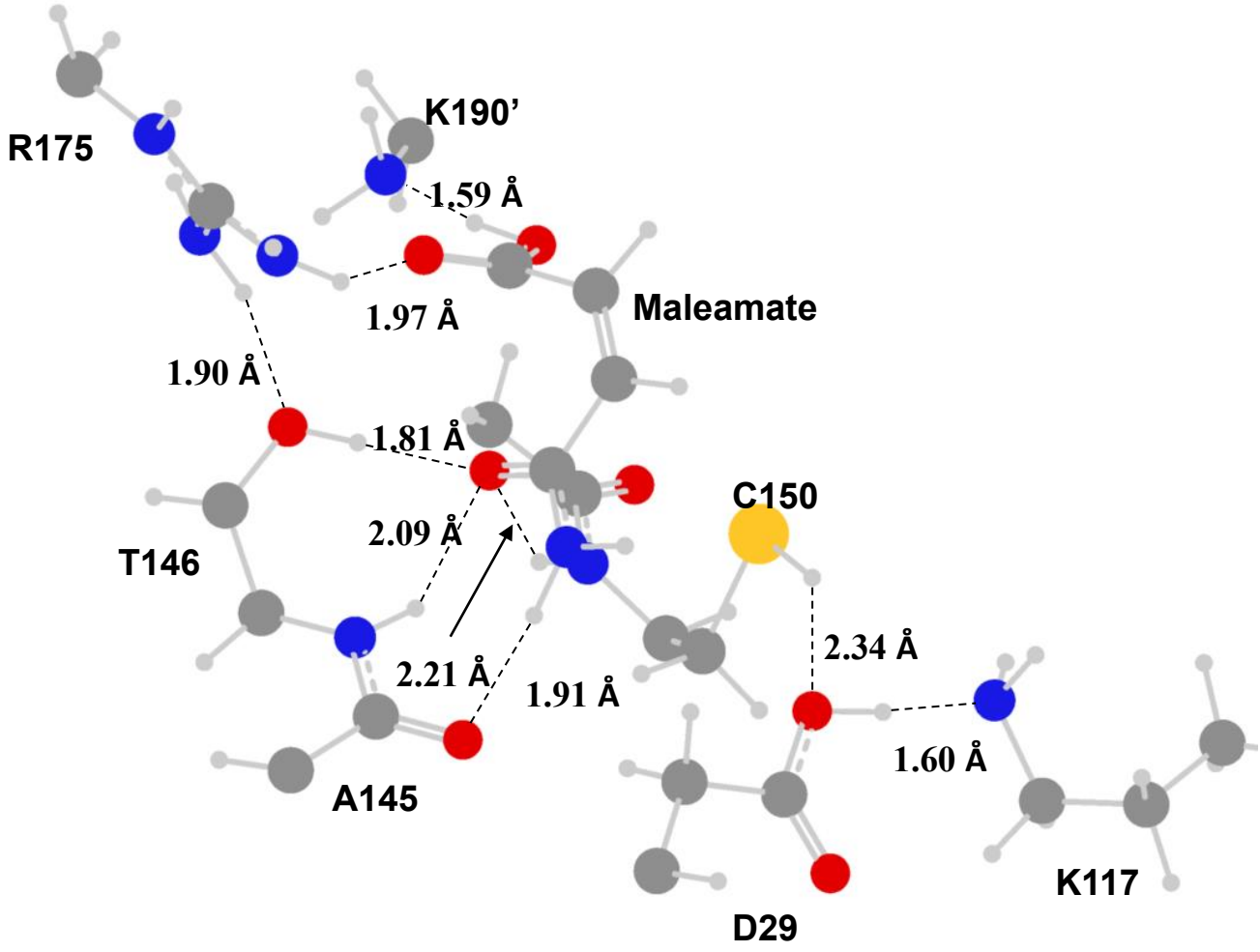
Appendix A3.9. Homolytic S−S bond dissociation enthalpy (BDE) of RXYH, proton affinity (PA) and gas-phase basicity (GPB) of RXY^-, and hydrogen affinity (HA) of $RXY^\bullet$ ($R=CH_3$, CH_2CH; X=S,Se, Y=Se,S). All energies calculated at 298.15K and in kJ mol^{-1}.

		CH_3SSeH	CH_3SSe^-		$CH_3SSe^\bullet$	CH_3SeSH	CH_3SeS^-		$CH_3SeS^\bullet$
Methods	**Basis set**	BDE(S—Se)	PA	GPB	HA	BDE(Se—S)	PA	GPB	HA
B3LYP	6-311+G(d,p)	209.8	1405.0	1374.8	275.2	217.3	1427.8	1397.6	289.2
	6-311G(df,p)	215.2	1421.2	1391.0	269.8	223.4	1435.1	1404.8	283.5
	6-311G+(2df,p)	220.7	1407.8	1377.7	267.0	229.0	1434.1	1403.9	285.3
	6-311G++(3df,3pd)	220.7	1408.0	1377.9	267.6	229.1	1437.2	1407.0	288.6
B3PW91	6-311+G(d,p)	221.7	1413.2	1382.8	273.8	229.4	1435.0	1404.8	287.1
	6-311G(df,p)	227.1	1425.8	1395.7	268.2	235.2	1440.5	1410.1	281.7
	6-311G+(2df,p)	233.3	1415.7	1385.5	265.3	241.6	1441.7	1411.4	283.6
	6-311G++(3df,3pd)	233.2	1415.9	1385.8	265.9	241.5	1444.5	1414.3	286.5
WB97XD	6-311+G(d,p)	219.6	1414.0	1384.2	279.9	231.4	1435.6	1405.8	295.6
	6-311G(df,p)	226.1	1427.3	1397.5	273.9	238.5	1440.7	1410.9	289.3
	6-311G+(2df,p)	232.5	1417.2	1387.5	270.9	245.0	1442.2	1412.4	290.8
	6-311G++(3df,3pd)	232.3	1417.6	1388.0	271.6	244.9	1445.2	1415.7	293.6
M062X	6-311+G(d,p)	222.3	1390.1	1359.8	277.7	226.8	1415.5	1384.6	297.8
	6-311G(df,p)	229.2	1400.9	1370.7	268.1	234.5	1420.3	1389.4	294.9
	6-311G+(2df,p)	234.3	1392.1	1362.0	270.5	240.0	1422.3	1391.7	296.3
	6-311G++(3df,3pd)	235.0	1393.2	1363.2	271.9	254.9	1425.7	1395.1	299.5
M08HX	6-311+G(d,p)	231.7	1386.7	1357.9	293.8	242.0	1420.6	1390.1	309.5
	6-311G(df,p)	236.8	1397.4	1368.2	288.3	246.5	1426.9	1396.1	306.6
	6-311G+(2df,p)	228.4	1387.9	1359.1	285.2	203.1	1382.7	1260.1	263.3
	6-311G++(3df,3pd)	227.9	1388.9	1359.7	285.8	259.9	1431.8	1401.2	311.8
QCISD	6-311G+(2df,p)	218.7	1411.3	1381.1	267.5	224.5	1438.9	1408.6	286.1

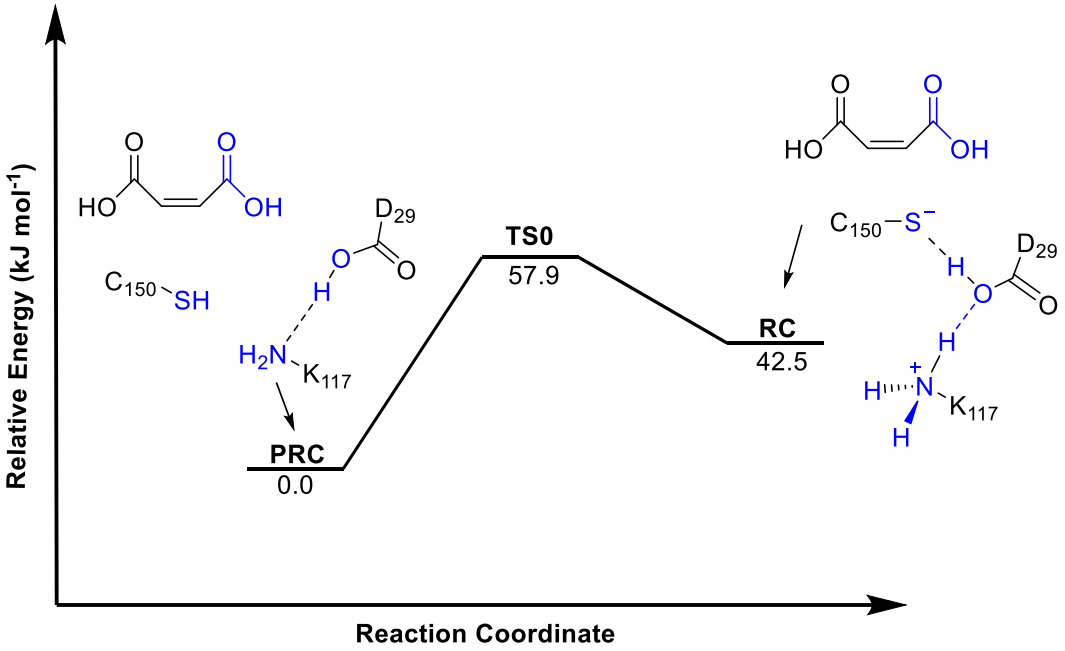
Appendix A3.10. Homolytic S−S bond dissociation enthalpy (BDE) of RSSSH, proton affinity (PA) and gas-phase basicity (GPB) of $RSSS^-$, and hydrogen affinity (HA) of $RSSS^\bullet$ (R=CH_3, CH_2CH). All energies calculated at 298.15K and in kJ mol^{-1}.

		CH_3SSSH	CH_3SSS^-		$CH_3SSS^\bullet$	$CH_2CHSSSH$	CH_2CHSSS^-		$CH_2CHSSS^\bullet$
Methods	**Basis set**	BDE(SS—S)	PA	GPB	HA	BDE(SS—S)	PA	GPB	HA
B3LYP	6-31G(d)	167.6	1415.2	1383.7	283.9	166.9	1402.9	1369.7	288.1
	6-311G(2d,p)	178.4	1407.1	1375.6	292.8	177.7	1384.9	1351.4	297.2
	6-311G+(2df,p)	180.4	1406.7	1375.2	289.5	179.7	1392.7	1360.0	292.9
	6-311G++(3df,3pd)	180.1	1410.1	1378.7	291.9	179.0	1396.1	1363.4	294.9
B3PW91	6-31G(d)	177.9	1413.9	1382.6	282.6	176.8	1427.5	1394.2	287.1
	6-311G(2d,p)	191.5	1413.9	1382.3	290.9	190.5	1401.8	1369.3	293.9
	6-311G+(2df,p)	193.8	1413.9	1382.4	287.9	192.8	1399.5	1367.0	290.5
	6-311G++(3df,3pd)	195.9	1419.4	1387.9	292.3	192.0	1402.4	1370.1	292.2
WB97XD	6-31G(d)	184.2	1415.6	1385.2	293.1	185.0	1406.0	1376.3	297.8
	6-311G(2d,p)	197.3	1415.1	1385.0	300.8	198.0	1405.6	1375.1	305.5
	6-311G+(2df,p)	200.4	1415.0	1385.1	297.7	201.0	1403.0	1370.9	302.0
	6-311G++(3df,3pd)	199.6	1418.7	1388.8	299.7	199.4	1406.3	1375.3	303.8
M062X	6-31G(d)	192.6	1398.7	1369.0	292.5	192.7	1389.5	1357.5	297.3
	6-311G(2d,p)	204.7	1395.6	1365.4	300.9	205.2	1387.3	1356.3	306.5
	6-311G+(2df,p)	208.6	1395.7	1365.3	298.4	208.7	1385.4	1356.1	303.2
	6-311G++(3df,3pd)	209.1	1399.2	1368.7	300.9	208.6	1388.6	1358.7	305.7
M08HX	6-31G(d)	196.0	1402.8	1376.5	300.8	195.7	1393.8	1362.2	305.8
	6-311G(2d,p)	206.8	1400.7	1371.2	304.5	206.9	1392.9	1361.8	310.1
	6-311G+(2df,p)	209.6	1400.6	1371.4	301.3	209.4	1389.8	1359.3	306.1
	6-311G++(3df,3pd)	209.8	1404.8	1376.2	303.7	208.8	1393.8	1363.8	308.3
QCISD	6-311G+(2df,p)	183.0	1413.9	1384.1	289.4	185.7	1403.7	1371.2	295.0

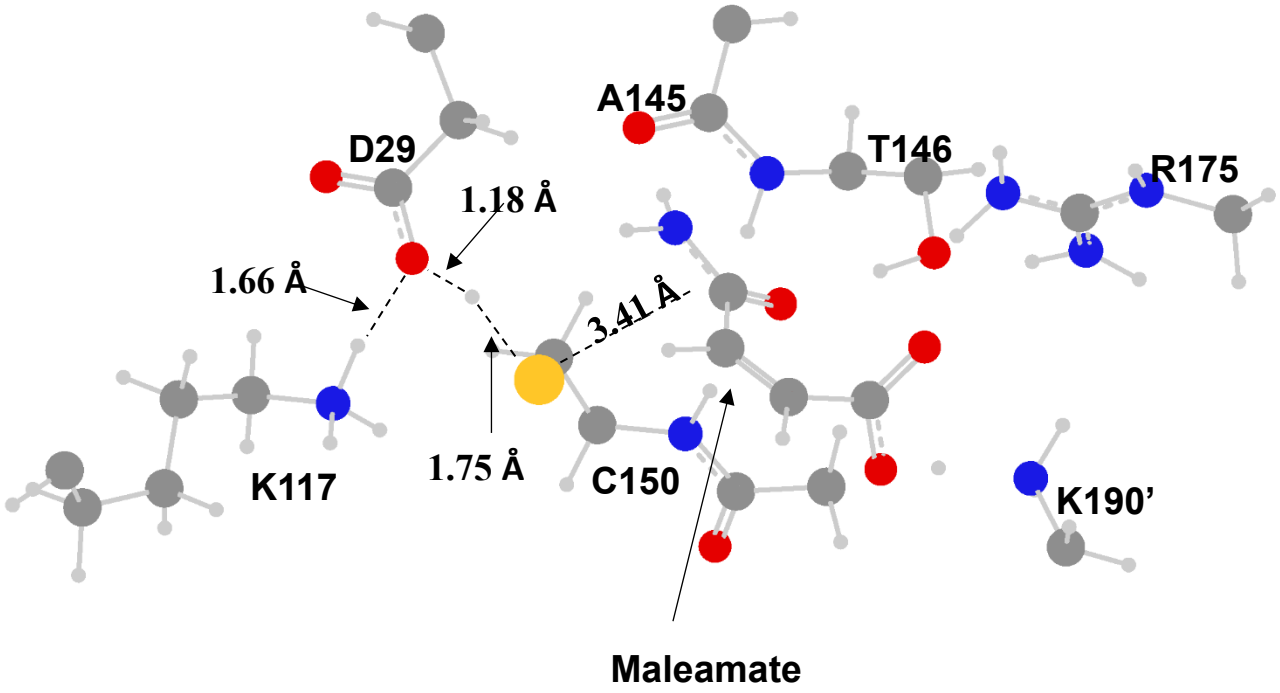
Appendix B



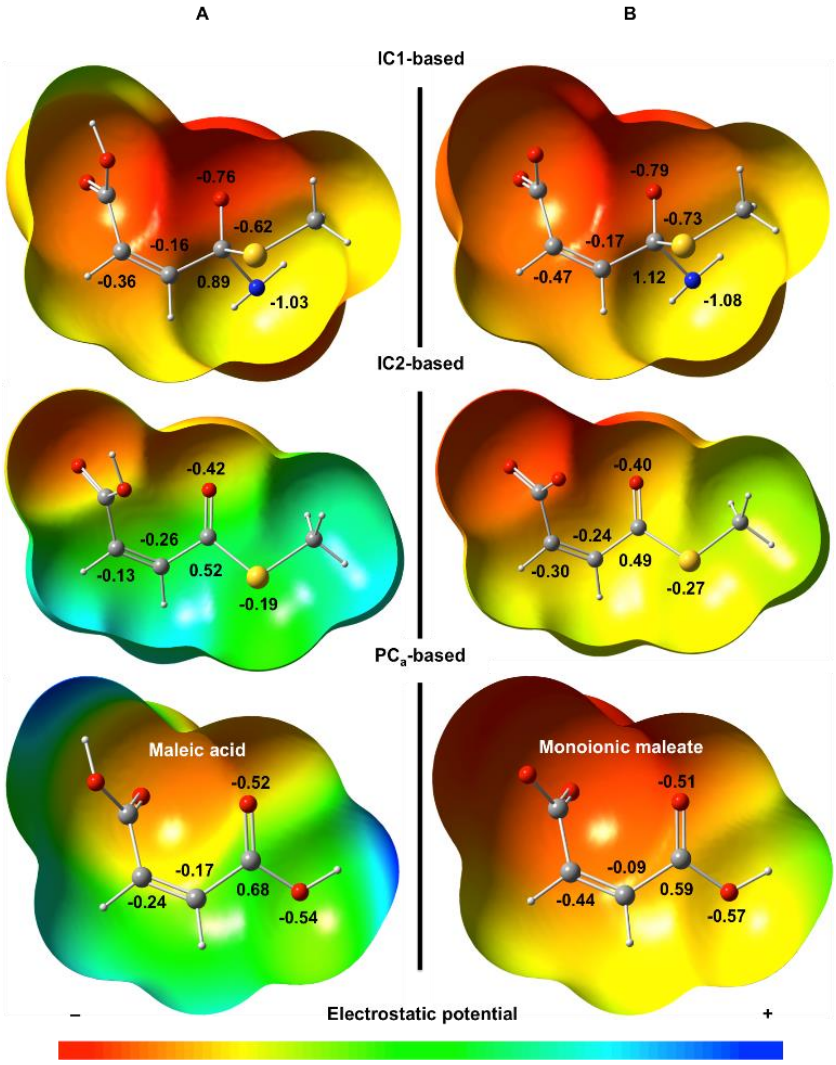
Appendix B5.1. Optimized structure of the pre-reactive complex (**PRC**). The nucleophile Cys150 is in its neutral thiol state.



Appendix B5.2. Potential Energy Surface for the deprotonation of Cys150; that is, conversion of the **PRC** (neutral Cys150 thiol) to generate **RC** (deprotonated Cys150 thiolate).



Appendix B5.3. Optimized transition structure (**TS0**) for the interconversion of **PRC** and **RC**.

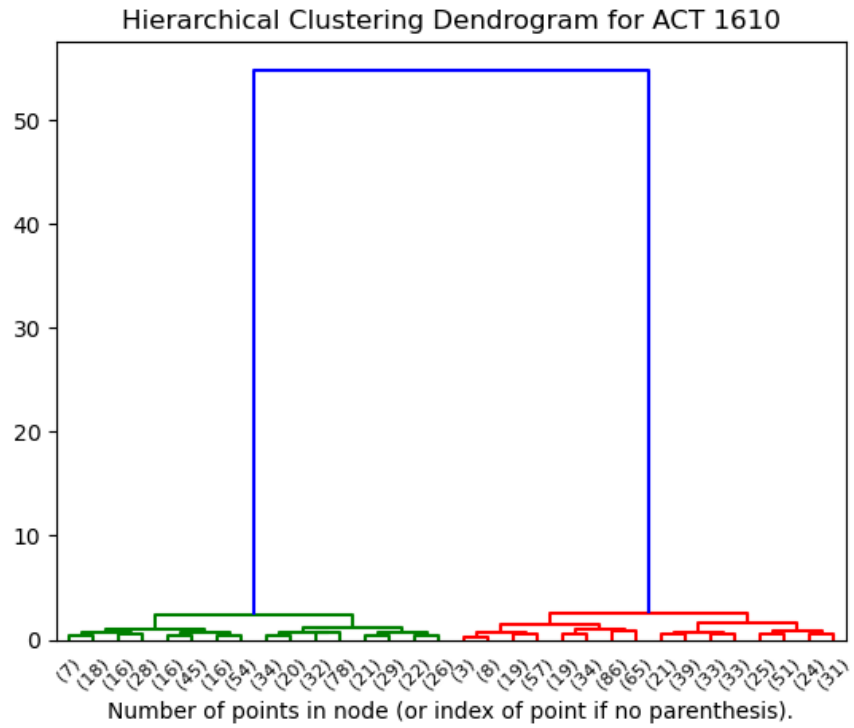


Appendix B5.4. Electrostatic potential surfaces mapped onto the electron density of select key species in the NicF mechanism for (A) neutral and (B) ionic carboxyl groups, obtained at B3LYP/6-311+G(2df,p) level of theory.

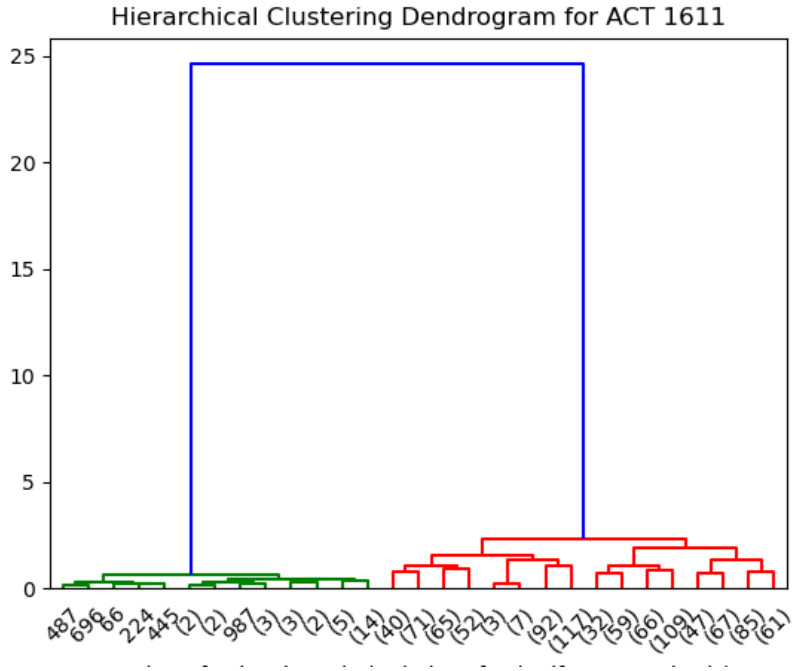
Appendix B5.5. Benchmark data for the second stage of the NicF reaction. Three different DFT methods are compared; B3LYP, M06-2X, and ωB97X-D.

Method	**Structure**	**Energy (kJ mol^{-1})**
B3LYP/6-311+G(2d,2p)	IC2′	0
	TS3	99.1
	PC[a]	0.4
	TS4	14.7
	PC[b]	-15.4
M06-2X/6-311+G(2d,2p)	IC2′	0
	TS3	81.5
	PC[a]	-7.9
	TS4	12.3
	PC[b]	-6.5
ωB97X-D/6-311+G(2d,2p)	IC2′	0
	TS3	85.0
	PC[a]	-8.3
	TS4	10.6
	PC[b]	-15.3
B3LYP/6-311+G(2df,p)	IC2′	0
	TS3	98.6
	PC[a]	1.6
	TS4	14.4
	PC[b]	-16.1
M06-2X /6-311+G(2df,p)	IC2′	0
	TS3	82.6
	PC[a]	-6.2
	TS4	12.6
	PC[b]	-6.8
ωB97X-D/6-311+G(2df,p)	IC2′	0
	TS3	84.7
	PC[a]	-7.1
	TS4	10.3
	PC[b]	-16.2

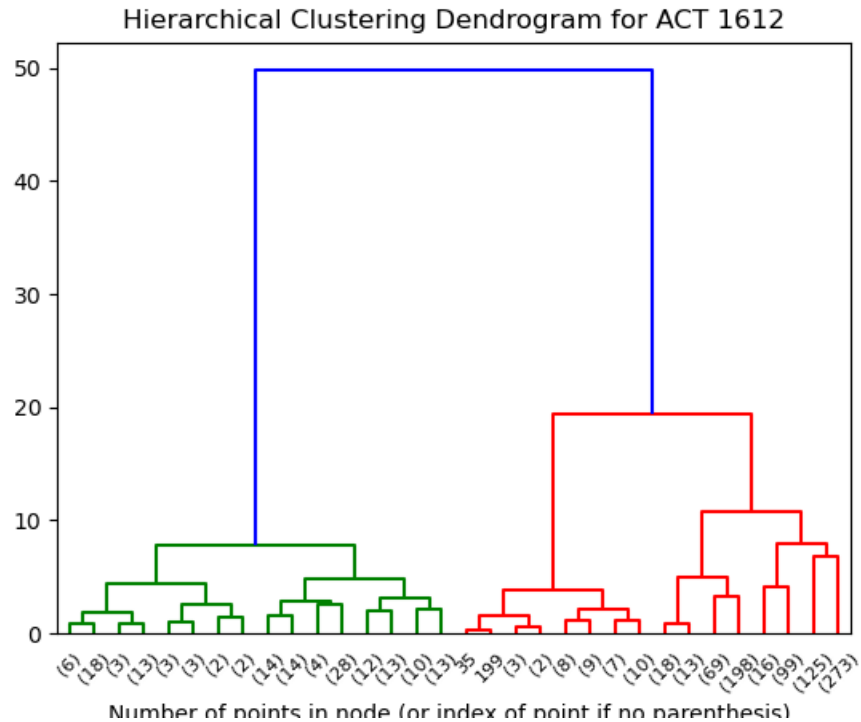
Appendix C



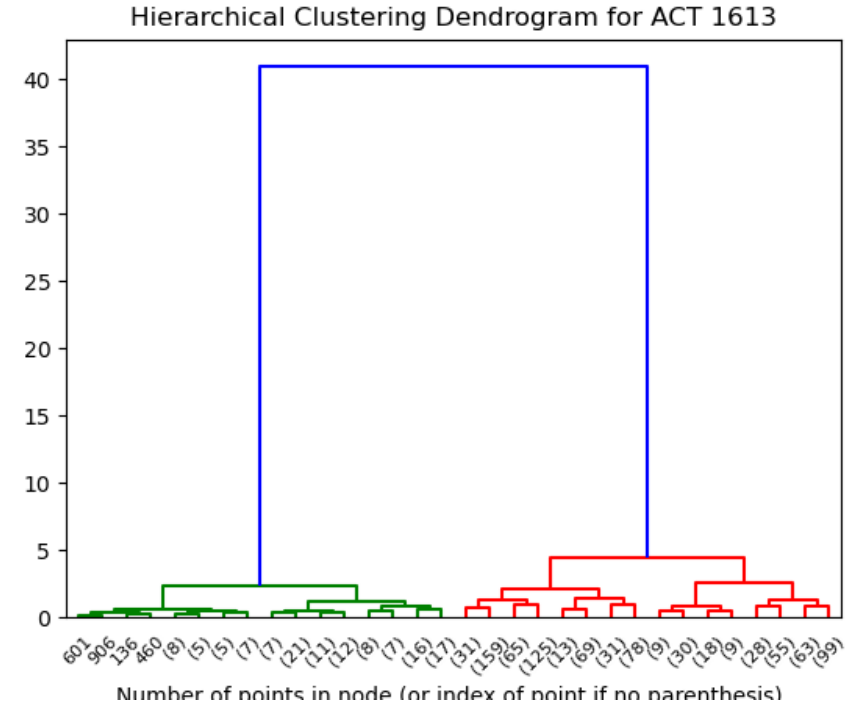
Appendix C7.1. Hierarchical Clustering Dendrogram for active site 1.



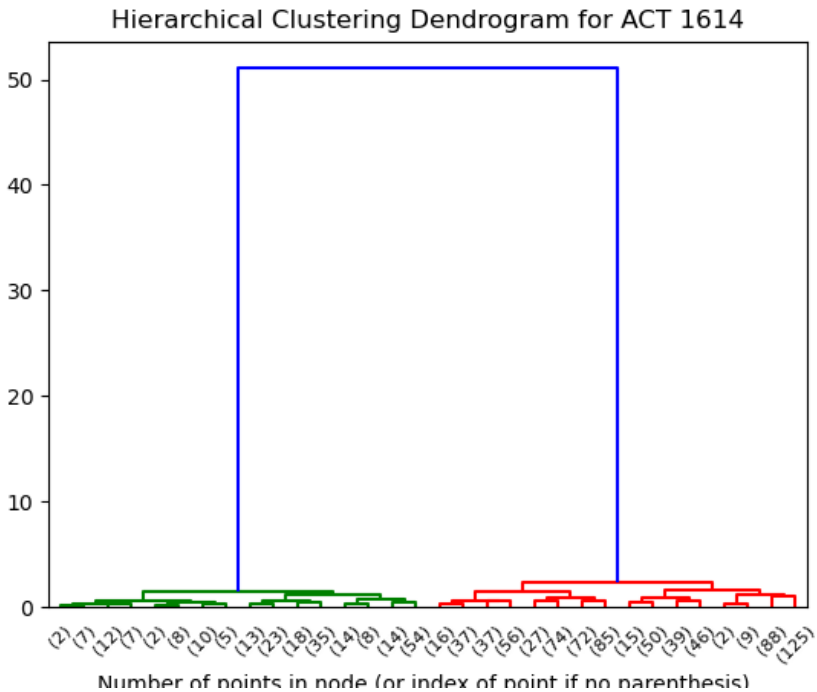
Appendix C7.2. Hierarchical Clustering Dendrogram for active site 2.



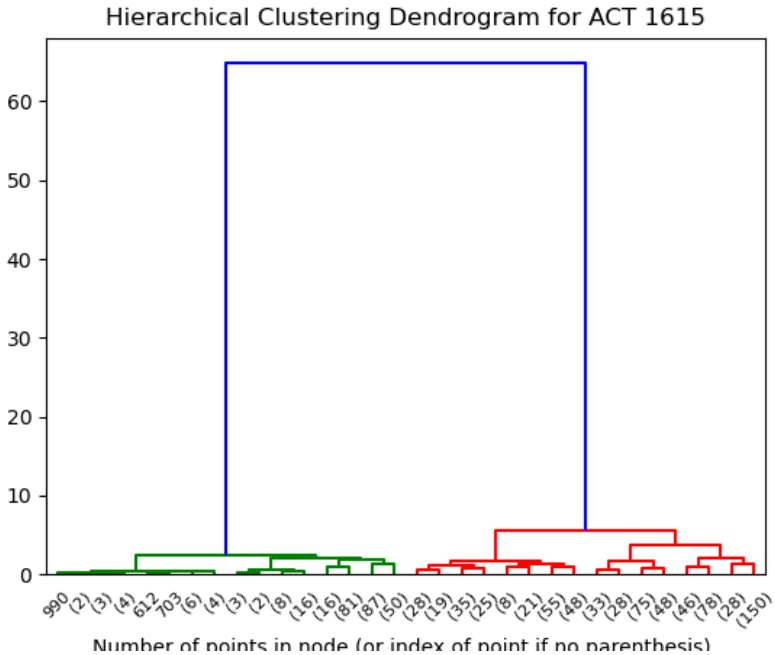
Appendix C7.3. Hierarchical Clustering Dendrogram for active site 3.



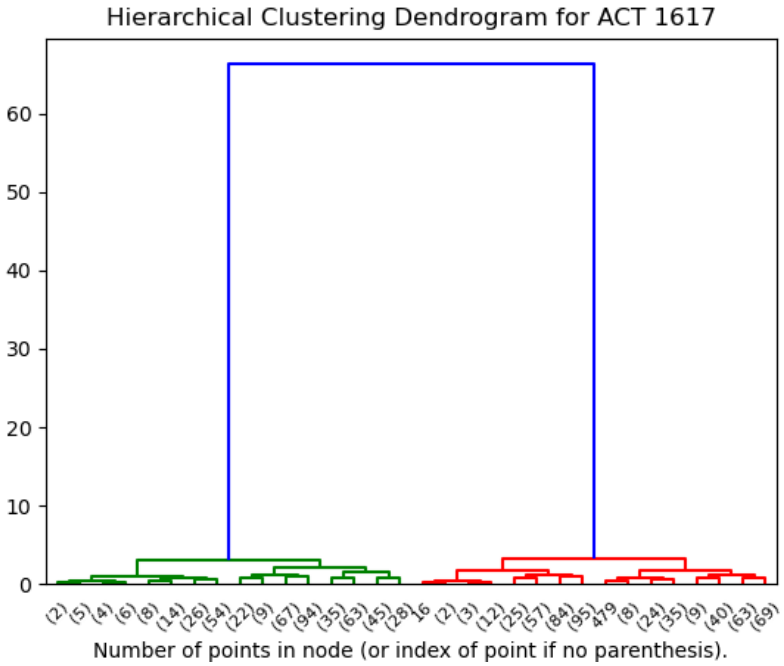
Appendix C7.4. Hierarchical Clustering Dendrogram for active site 4.



Appendix C7.5. Hierarchical Clustering Dendrogram for active site 5.



Appendix C7.6. Hierarchical Clustering Dendrogram for active site 6.



Appendix C7.7. Hierarchical Clustering Dendrogram for active site 8.

Appendix C7.8. Program output for the Pose Selector program running on each active site of NicF showing the number of frames identified and the resulting silhouette score.

ACT 1610

Labeling frames... Success!

Group labeled 0: 548

-> Group labeled 1: 462

Selected frames (numbering starts at 0):

[1, 4, 6, 7, 9, 11, 12, 16, 17, 18, 19, 22, 24, 25, 26, 27, 29, 31, 33, 37, 39, 40, 43, 45, 48, 49, 61, 65, 66, 67, 68, 70, 71, 76, 80, 82, 84, 85, 86, 89, 90, 91, 93, 95, 99, 100, 103, 107, 109, 110, 113, 116, 117, 118, 120, 127, 132, 133, 134, 137, 140, 141, 142, 144, 146, 148, 149, 151, 155, 159, 160, 161, 162, 163, 165, 166, 168, 169, 171, 173, 176, 177, 178, 181, 185, 188, 189, 192, 195, 199, 205, 208, 209, 211, 213, 218, 219, 220, 223, 229, 236, 238, 240, 241, 244, 246, 247, 248, 249, 251, 252, 254, 258, 259, 262, 263, 264, 266, 267, 268, 269, 272, 273, 276, 279, 285, 286, 287, 288, 292, 294, 301, 307, 311, 312, 315, 318, 319, 320, 324, 325, 326, 327, 330, 332, 334, 336, 337, 338, 339, 343, 344, 345, 347, 348, 349, 353, 354, 356, 361, 362, 366, 367, 369, 371, 373, 374, 376, 377, 378, 379, 382, 383, 384, 385, 388, 391, 392, 393, 395, 396, 397, 399, 401, 403, 407, 408, 410, 412, 414, 415, 417, 419, 421, 423, 425, 429, 430, 433, 435, 437, 440, 445, 450, 453, 458, 460, 463, 467, 469, 470, 471, 473, 475, 477, 479, 480, 483, 484, 486, 487, 491, 492, 495, 497, 498, 499, 500, 505, 508, 509, 510, 512, 513, 515, 516, 517, 519, 520, 521, 522, 523, 529, 533, 537,

539, 540, 551, 554, 555, 556, 559, 560, 562, 563, 566, 567, 569, 570, 575, 585, 590, 592, 594, 602, 603, 607, 608, 609, 612, 613, 614, 615, 618, 622, 623, 624, 626, 628, 630, 631, 632, 633, 635, 636, 640, 641, 642, 645, 646, 648, 650, 652, 653, 654, 656, 659, 661, 663, 664, 667, 669, 672, 675, 677, 678, 679, 680, 682, 684, 689, 690, 691, 693, 694, 696, 697, 701, 702, 705, 707, 710, 711, 714, 719, 720, 721, 725, 727, 728, 733, 734, 737, 740, 742, 743, 744, 745, 748, 749, 755, 759, 760, 762, 763, 765, 766, 768, 769, 773, 783, 785, 791, 793, 794, 797, 798, 800, 803, 804, 806, 809, 811, 813, 814, 815, 819, 820, 821, 822, 824, 825, 829, 831, 832, 833, 834, 835, 837, 838, 839, 842, 848, 849, 850, 851, 852, 853, 855, 857, 858, 862, 863, 864, 865, 867, 868, 869, 871, 872, 875, 879, 883, 884, 885, 886, 890, 892, 894, 897, 900, 901, 903, 904, 905, 906, 907, 912, 914, 917, 918, 919, 922, 923, 924, 926, 927, 930, 932, 934, 936, 938, 940, 946, 947, 949, 951, 956, 959, 960, 962, 964, 967, 968, 971, 973, 975, 979, 980, 982, 990, 991, 992, 993, 994, 995, 998, 999, 1000, 1006, 1008, 1009]

Assessing the quality of the model...
Silhouette score (1 is best, 0 is random): 0.91712695

ACT 1611
Labeling frames... Success!
Group labeled 0: 973
-> Group labeled 1: 37

Selected frames (numbering starts at 0):
[34, 66, 77, 79, 84, 91, 133, 163, 165, 166, 189, 224, 295, 319, 348, 431, 445, 487, 502, 580, 584, 597, 627, 655, 696, 707, 775, 808, 856, 870, 890, 898, 950, 982, 987, 988, 993]

Assessing the quality of the model...
Silhouette score (1 is best, 0 is random): 0.9244331

ACT 1612
Labeling frames... Success!
Group labeled 0: 513
-> Group labeled 1: 497

Selected frames (numbering starts at 0):
[28, 31, 36, 37, 43, 47, 52, 55, 56, 65, 72, 73, 76, 77, 81, 83, 84, 86, 87, 88, 89, 92, 93, 94, 96, 98, 99, 100, 101, 102, 103, 106, 111, 113, 114, 115, 116, 119, 122, 124, 126, 128, 131, 133, 134, 136, 139, 141, 142, 145, 148, 149, 151, 152, 153, 154, 156, 157, 158, 160, 163, 167, 171, 172, 173, 175, 180, 184, 185, 186, 187, 193, 194, 195, 196, 201, 205, 207, 208, 209, 210, 212, 213, 214, 215, 217, 218, 219, 220, 221, 225, 227, 230, 232, 233, 234, 237, 239, 240, 243, 246, 247, 250, 253, 254, 256, 260, 271, 279, 280, 285, 296, 306, 310, 312, 317, 351, 360, 379, 387, 391, 395, 403, 405, 421, 425, 432, 450, 464, 465, 484,

485, 516, 517, 524, 542, 561, 562, 568, 571, 580, 606, 678, 680, 682, 722, 729, 752, 783, 854, 867, 887, 913, 924, 934, 953, 958, 995]

Assessing the quality of the model...
Silhouette score (1 is best, 0 is random): 0.29697767

ACT 1613
Labeling frames... Success!
Group labeled 0: 882
-> Group labeled 1: 128

Selected frames (numbering starts at 0):
[1, 6, 7, 8, 9, 22, 33, 36, 46, 51, 76, 78, 84, 89, 100, 104, 108, 113, 118, 133, 136, 139, 140, 142, 153, 164, 173, 188, 217, 219, 251, 253, 256, 271, 272, 298, 310, 311, 313, 322, 328, 332, 340, 342, 343, 349, 350, 372, 374, 381, 384, 396, 402, 403, 405, 411, 412, 414, 418, 422, 447, 460, 474, 475, 480, 484, 490, 505, 531, 542, 543, 547, 558, 570, 571, 572, 576, 595, 597, 601, 603, 608, 620, 628, 629, 630, 639, 641, 644, 657, 665, 666, 669, 670, 673, 674, 675, 680, 695, 709, 711, 715, 734, 739, 756, 757, 764, 765, 783, 793, 798, 817, 851, 857, 860, 872, 879, 884, 888, 893, 900, 905, 906, 939, 948, 970, 972, 973]

Assessing the quality of the model...
Silhouette score (1 is best, 0 is random): 0.9025177

ACT 1614
Labeling frames... Success!
Group labeled 0: 778
-> Group labeled 1: 232

Selected frames (numbering starts at 0):
[12, 22, 27, 28, 35, 41, 43, 44, 46, 59, 67, 68, 74, 75, 76, 80, 98, 106, 107, 115, 124, 128, 129, 130, 134, 137, 138, 140, 146, 147, 148, 153, 155, 156, 160, 162, 163, 167, 169, 172, 181, 183, 185, 189, 192, 194, 195, 202, 208, 211, 215, 216, 219, 223, 224, 227, 231, 232, 234, 242, 243, 260, 261, 269, 277, 290, 294, 298, 305, 322, 324, 326, 333, 336, 340, 344, 347, 348, 353, 363, 372, 376, 378, 383, 385, 388, 390, 397, 406, 411, 412, 415, 417, 425, 427, 442, 443, 444, 448, 456, 457, 459, 462, 466, 477, 482, 490, 494, 499, 500, 501, 503, 505, 506, 510, 515, 516, 522, 524, 534, 535, 537, 538, 544, 547, 558, 559, 564, 565, 566, 573, 576, 584, 589, 593, 599, 601, 607, 608, 617, 621, 626, 635, 637, 639, 640, 643, 645, 646, 651, 657, 661, 662, 667, 688, 697, 699, 700, 701, 703, 706, 712, 718, 722, 725, 739, 744, 747, 749, 760, 761, 762, 765, 768, 782, 788, 792, 793, 800, 807, 809, 813, 815, 817, 819, 822, 823, 824, 825, 830, 831, 838, 842, 847, 850, 854, 862, 865, 872, 873, 876, 879, 882, 890, 891, 895, 897, 899, 900, 902, 905, 907, 913, 918, 919, 921, 924, 925, 927, 938, 943, 944, 955, 961, 983, 985, 992, 995, 996, 1004, 1005, 1006]

Assessing the quality of the model...
Silhouette score (1 is best, 0 is random): 0.924787

ACT 1615
Labeling frames... Success!
Group labeled 0: 725
-> Group labeled 1: 285

Selected frames (numbering starts at 0):
[129, 130, 131, 132, 133, 134, 135, 139, 140, 143, 144, 147, 153, 154, 161, 162, 164, 165, 166, 167, 168, 169, 170, 171, 172, 178, 181, 183, 184, 185, 186, 188, 195, 197, 198, 199, 202, 203, 205, 208, 213, 222, 225, 231, 232, 233, 238, 239, 240, 243, 247, 250, 253, 254, 256, 261, 263, 265, 268, 269, 270, 271, 272, 273, 274, 275, 276, 278, 279, 280, 281, 282, 288, 291, 292, 294, 295, 296, 298, 300, 302, 303, 305, 307, 309, 310, 311, 313, 315, 316, 328, 330, 331, 332, 333, 334, 346, 349, 350, 355, 359, 362, 364, 365, 366, 367, 368, 369, 371, 375, 380, 382, 384, 385, 386, 388, 390, 394, 396, 397, 399, 400, 402, 405, 410, 411, 415, 417, 418, 420, 423, 425, 427, 428, 430, 432, 433, 435, 436, 437, 438, 439, 440, 444, 445, 446, 450, 452, 453, 455, 457, 458, 460, 464, 472, 475, 478, 479, 483, 487, 488, 489, 490, 491, 492, 493, 498, 500, 502, 504, 505, 506, 507, 511, 513, 515, 521, 522, 523, 524, 528, 529, 530, 532, 534, 536, 538, 539, 540, 543, 544, 545, 550, 551, 552, 553, 612, 627, 645, 646, 659, 660, 663, 664, 674, 695, 699, 703, 709, 713, 718, 730, 734, 741, 748, 750, 755, 760, 761, 765, 785, 788, 794, 798, 800, 803, 804, 806, 807, 808, 822, 823, 831, 835, 837, 851, 854, 862, 870, 877, 878, 879, 880, 881, 882, 883, 885, 887, 888, 889, 890, 892, 893, 894, 895, 896, 897, 898, 899, 900, 901, 902, 903, 904, 906, 907, 912, 913, 914, 915, 916, 923, 925, 936, 937, 945, 949, 958, 972, 977, 982, 987, 990, 993, 1005]

Assessing the quality of the model...
Silhouette score (1 is best, 0 is random): 0.8931388

ACT 1616
Labeling frames... Success!
Group labeled 0: 1010
-> Group labeled 1: 0

Selected frames (numbering starts at 0):
[]

ACT 1617
Labeling frames... Success!
Group labeled 0: 528
-> Group labeled 1: 482

Selected frames (numbering starts at 0):
[12, 13, 17, 18, 19, 22, 23, 25, 26, 28, 30, 33, 37, 38, 43, 44, 47, 48, 49, 50, 51, 52, 53, 54, 55, 57, 60, 61, 63, 64, 66, 68, 69, 70, 71, 72, 73, 74, 76, 78, 80, 81, 84, 85, 88, 90,

94, 95, 99, 101, 105, 106, 107, 108, 109, 110, 111, 112, 114, 116, 117, 118, 121, 122, 123, 126, 127, 128, 131, 133, 136, 138, 139, 140, 144, 145, 147, 148, 149, 150, 151, 153, 154, 156, 157, 159, 160, 162, 163, 171, 173, 174, 175, 178, 179, 180, 181, 183, 186, 187, 188, 189, 190, 192, 193, 195, 196, 201, 202, 203, 204, 205, 207, 208, 209, 211, 213, 214, 215, 216, 218, 219, 220, 221, 225, 226, 227, 230, 231, 232, 234, 236, 237, 241, 242, 244, 245, 246, 247, 248, 250, 253, 256, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 274, 275, 276, 278, 279, 280, 281, 284, 285, 286, 287, 288, 289, 290, 292, 293, 298, 300, 301, 303, 304, 305, 306, 307, 308, 309, 310, 311, 314, 315, 316, 317, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 341, 343, 344, 345, 346, 347, 351, 352, 355, 357, 358, 361, 362, 364, 365, 366, 367, 369, 370, 371, 372, 376, 380, 381, 382, 384, 388, 391, 392, 395, 397, 399, 400, 401, 404, 405, 406, 407, 408, 411, 413, 414, 415, 416, 417, 418, 421, 423, 424, 426, 428, 430, 432, 433, 434, 435, 436, 437, 438, 440, 441, 442, 443, 444, 445, 446, 448, 450, 451, 452, 458, 461, 464, 465, 466, 468, 470, 472, 473, 475, 476, 480, 481, 482, 483, 484, 485, 486, 487, 488, 491, 492, 494, 497, 499, 501, 502, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 522, 523, 524, 525, 527, 530, 531, 534, 538, 540, 541, 542, 543, 548, 549, 550, 551, 554, 556, 557, 559, 566, 567, 568, 569, 571, 576, 577, 578, 579, 580, 581, 582, 584, 586, 590, 592, 593, 594, 596, 599, 601, 603, 604, 605, 607, 610, 611, 615, 617, 620, 622, 625, 626, 627, 628, 630, 631, 633, 634, 636, 637, 638, 639, 640, 642, 644, 647, 648, 649, 650, 651, 653, 655, 657, 659, 661, 664, 667, 674, 675, 678, 682, 689, 699, 701, 706, 720, 726, 727, 729, 737, 739, 741, 747, 751, 752, 758, 759, 760, 764, 766, 769, 774, 775, 779, 780, 784, 788, 790, 792, 793, 795, 796, 797, 798, 804, 807, 810, 813, 815, 817, 818, 819, 822, 831, 837, 841, 845, 846, 847, 854, 865, 868, 875, 876, 887, 888, 891, 892, 895, 897, 900, 907, 913, 916, 919, 931, 933, 943, 947, 950, 953, 956, 959, 960, 964, 966, 967, 968, 969, 972, 974, 976, 984, 988, 995, 1006, 1008]

Assessing the quality of the model...
Silhouette score (1 is best, 0 is random): 0.91219497

Vita Auctoris

NAME:	Paul Meister
YEAR OF BIRTH:	1994
EDUCATION:	University of Windsor, B.Sc., Windsor, ON, 2016
	University of Windsor, Ph.D., Windsor, ON, 2021