Title

Computational Studies on the [FeFe] Hydrogenase Maturation and the Metamorphism of the Circadian Protein KaiB

Author

Chen, Nanhao

Publication Date

2023

Peer reviewed|Thesis/dissertation

Computational Studies on the [FeFe] Hydrogenase Maturation and the Metamorphism of the Circadian Protein KaiB

By

Nanhao Chen
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____
Lee-Ping Wang, Chair

_____
R. David Britt

_____
Davide Donadio

Committee in Charge

2023

Table of Contents

1 Introduction **1**

1.1 Computations in Chemistry[1] 1

1.1.1 QM/MM Simulations[2] 1

1.1.2 Molecular Dynamics (MD)[9,10] 5

1.2 Brief Introduction of Hydrogenases and Circadian Protein KaiB. . . . 8

1.2.1 Hydrogenases 8

1.2.2 Circadian Protein KaiB 10

1.3 Research in my Ph.D. Period 12

2 Quantum chemical study of a radical relay mechanism for the HydG-catalyzed synthesis of a Fe(II)$(CO)_2$(CN)cysteine precursor to the H-cluster of [FeFe] hydrogenase[1] **20**

2.1 Introduction 20

2.2 Computational Methods 24

2.3 Results and Discussions 25

2.3.1 Tyrosine Radical Formation. 25

2.3.2 Tyrosine Decomposition. 26

2.3.3 DHG Decomposition and CN and COOH Generation. 29

2.3.4 Spin crossover and first CN^- substitution. 32

2.3.5 First $COOH^\bullet$ substitution, decomposition, and reduction. . . . 34

2.3.6 Reduction and decomposition of a second $COOH^\bullet$. 36

2.3.7 Completion of the catalytic cycle. 38

2.4 Conclusions 40

3 How does HydE work? A Comparison Between A Radical Mechanism and A Proton-Transfer Mechanism 52

3.1 Introduction . . . 52

3.2 Methods . . . 56

3.3 Results and Discussions . . . 58

3.3.1 Formation of the 10-s intermediate. . . . 58

3.3.2 Conversion of the 10-s intermediate . . . 61

3.3.2.1 Hypothesis A: Cleavage of the cysteinyl Cβ-S bond via a unconventional β-elimination. . . . 62

3.3.2.2 Hypothesis B: Cysteine decomposition pathway via radical-relay mechanism . . . 63

3.3.2.3 Conversion of 2-iminopropanoate cation into pyruvate. 66

3.3.3 Dimerization in HydE . . . 68

3.3.3.1 Decomposition of the 10-min intermediate . . . 68

3.3.3.2 Dimerization mechanism of Fe(I) complexes within HydE . . . 71

3.4 Conclusions . . . 72

4 Sequence-based Prediction of Metamorphic Behavior in Proteins[2] 78

4.1 Introduction . . . 78

4.2 Theory . . . 81

4.2.1 Secondary Structure Prediction (SSP) . . . 81

4.2.2 Metamorphic Proteins and Diversity Index . . . 82

4.3 Dataset Setup . . . 84

4.3.1 Construction of the metamorphic reference dataset . . . 84

4.3.2 Construction of the monomorphic reference dataset . . . 85

4.4 Results and Discussion 87
4.4.1 Behavior of the diversity index (DI) 87
4.4.2 Diversity index-based classifier performance 89
4.4.3 Comparison with other methods 93
4.4.4 Classification using multiple diversity indices 93
4.4.5 Analysis of outliers in diversity index-based classification . . . 96
4.4.6 Dependence of results on sequence database 99
4.5 Conclusions 100

A Supporting Information for Chapter 2: Quantum chemical study of a radical relay mechanism for the HydG-catalyzed synthesis of a Fe(II)$(CO)_2$(CN)cysteine precursor to the H-cluster of [FeFe] hydrogenase 109

A.1 Computational methods 109

B Supporting Information for Chapter 3: How does HydE work? A Comparison Between A Radical Mechanism and A Proton-Transfer Mechanism 139

C Supporting Information for Chapter 4: Sequence-based Prediction of Metamorphic Behavior in Proteins 147

Abstract

Hydrogenases are a family of enzymes that catalyze the reversible redox reaction of molecule hydrogen (H_2). There are several kinds of hydrogenases, including [Fe] hydrogenases, [NiFe] hydrogenases, and [FeFe] hydrogenases, found in a variety of organisms. Hydrogenases have attracted much attention from chemists, physicists, and biologists due to their special roles in energy metabolism, as they can produce the cleanest carbon-neutral fuel, H_2. Among these hydrogenases, [FeFe] hydrogenases are characterized by their special di-iron ([FeFe]) cluster called the "H-cluster". The maturation process of the [FeFe] hydrogenases is the biosynthesis of the H-cluster by the enzymes HydE, HydF, and HydG, and the delivery of the H-cluster into HydA. In this thesis we describe hybrid quantum mechanics (QM)/ molecular mechanics (MM) simulations of the maturation process, including the catalytic processes in HydG and HydE. The results were published in *Biochemistry* journal as an article titled "Quantum chemical study of a radical relay mechanism for the HydG-catalyzed synthesis of a Fe(II)$(CO)_2$(CN)cysteine precursor to the H-cluster of [FeFe] hydrogenase". We proposed a radical-relay mechanism for how HydG catalyzes the decomposition of the tyrosine substrate into $COO^{\bullet -}$ and HCN. These species are converted into CO and CN at the [5Fe-5S] auxiliary cluster in HydG, which bind to the fifth "dangler" Fe and result in the [Fe(II)$(CO)_2$(CN)cysteine] "synthon" product. HydE, as the downstream protein of HydG, modifies this Fe(II) complex into a 5-coordinate Fe(I) cluster via a radical mechanism. Using QM/MM simulations we proposed a feasible radical mechanism for this conversion, as well as a dimerization pathway from the 5-coordinated Fe(I) cluster to a diamagnetic di-iron cluster that is proposed to be the product of HydE.

In addition, studies have been done to understand the behavior of a circadian clock protein, KaiB. KaiB is a key component of the KaiABC circadian clock system in cyanobacteria. KaiB changes its folding to bind to KaiA and KaiC respectively to adjust the expression of different kinase proteins. The folding change in KaiB, also named fold-switching, is classified as metamorphic behavior because it involves changes between the secondary structures and three-dimensional structures in contrast to conventional protein conformational changes. Classical, all-atom molecular dynamics simulations exceeding (>100 μs) in length were carried out to study the fold-switching process of the KaiB in explicit solvent. We also developed a new feature, named diversity index (DI), that can distinguish the metamorphic protein sequences from other monomorphic sequences (i.e. one native fold sequences), and this work was published in *Biophysical Journal* in 2021 titled "Sequence-based Prediction of Metamorphic Behavior in Proteins".

Acknowledgements

I would like to appreciate all the support from my family, especially my parents and my wife. Without their support, I cannot make this happen.

I also would like to express my gratitude to my academic advisor, Dr. Lee-Ping Wang, for his patient guidance and advice on my Ph.D. projects. It is my great fortune to have Lee-Ping as my advisor in my Ph.D. period. Besides, I would like to thank Dr. Andy LiWang at UC Merced for the funding support in the KaiB and metamorphic project. I also acknowledge financial support from U.S. Army Research Office, award W911NF-17-1-0434.

In addition, I would like to thank the current and former members of Wang's group, particularly Yudong Qiu for his mentorship and support when I first came to Davis, and Hyesu Jang, Lisa Oh, Marshall Hutchings, Zhecheng He, Nathan Yoshino, Jesi Lee, and Heejune Park for their kind help when I was in need.

Chapter 2 is a reprint of the published work: "Quantum chemical study of a radical relay mechanism for the HydG-catalyzed synthesis of a Fe (II)(CO) 2 (CN) cysteine precursor to the H-cluster of [FeFe] hydrogenase." *Biochemistry* 60 (40), 3016-3026. Chapter 3 is a preprint of the HydE manuscript. I gratefully acknowledge my coauthors, Dr. R. David Britt, Guodong Rao, and Lizhi Tao for their kind help and useful discussion in these projects, as well as discussions with Yudong Qiu, Tom Rauchfuss and Dan Suess.

Chapter 4 is also a reprint of the published work: Chen, N., Das, M., LiWang, A., & Wang, L. P. (2020). "Sequence-based prediction of metamorphic behavior in proteins." Biophysical journal, 119(7), 1380-1390. I would like to gratefully acknowledge Yudong Qiu, Archana Chavan, and Xuejun Yao for their valuable discussions.

1 Introduction

1.1 Computations in Chemistry[1]

Computational chemistry or theoretical chemistry is a relatively 'young' field in chemistry that applies the fundamental principles in physics and mathematics to study chemistry problems. The first theoretical computation in chemistry was done in 1927 by Walter Heitler and Fritz London. The development of computational chemistry has revolutionized the study of chemistry by providing tools for researchers to explore the behavior of molecules in a virtual environment and predicting many molecular properties without conducting expensive and time-consuming 'wet' experiments. In recent decades, the advancement of computing power and algorithms has enabled the application of modeling and simulations to study complicated systems accurately. In addition, experimental techniques such as nuclear magnetic resonance, X-ray crystallography, and cryogenic electron microscopy not only provide the molecular simulations with often-needed initial conditions but also validate the results of simulations. Nowadays, many complex systems and research questions can be studied by simulations, including optimizing catalysts, designing drugs to bind a given protein, predicting the properties of given materials, and investigating complicated protein structures, functions, and catalysis. Among all the methods in molecular simulations, the hybrid quantum mechanics/molecular mechanics (QM/MM) method and the molecular dynamics (MD) method are two widely-used tools for bio-molecular studies.

1.1.1 QM/MM Simulations[2]

As the name says, the QM/MM method combines accurate quantum mechanics (QM) calculations with efficient classical molecular mechanics (MM) calculations. This feature enables the QM/MM method to describe the behavior of the electrons in the

central region of a protein while considering the behavior of the whole protein simultaneously. The QM/MM method was first introduced in the 1970s by Arieh Warshel and Michael Levitt, and they were awarded the 2013 Nobel Prize in Chemistry with Martin Karplus for "the development of multiscale models for complex chemical systems".[3] In their publications, the behavior of the atoms of the carbonium ion intermediate in the lysozyme was described in the MM method while the behavior of the electrons of the intermediate was described using a semi-empirical QM approach (QCFF/ALL). In recent years, QM/MM simulations have been used to study a wide range of chemical and biological systems, including enzyme catalytic mechanisms[4], photochemistry[5], photophysics[6], and others.

A general QM/MM simulation requires dividing the system into QM and MM parts and the energy of each part is calculated by the corresponding methods. As for the QM parts, the Hamiltonian can be written in the following way, assuming fixed nuclei and setting physical constants to one:

$$H_{\mathrm{QM}} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{A,i}^{M} \frac{Z_A}{|r_i - R_A|} + \sum_{i<j} \frac{1}{|r_i - r_j|} \tag{1}$$

where Z_A is atomic number of nucleus A, $|r_i - R_A|$ is the distance between electron i and nucleus A, and $|r_i - r_j|$ is the distance between electrons i and j. Various *ab initio* QM methods attempt to solve for the eigenstates of the Hamiltonian with approximate wavefunctions; for example, Hartree-Fock (HF) uses a Slater determinant ansatz of the wavefunction. By contrast, the commonly used Kohn-Sham density functional theory (KS-DFT) replaces parts of the *ab initio* Hamiltonian with a functional of the density that captures higher order post-Hartree-Fock effects in an approximate way. Considering the interaction between the QM part and the MM part, there are two schemes, namely subtractive schemes and additive schemes. Subtractive schemes

require the QM calculation on the QM part ($\mathbb{QM}$) and two MM calculations, one for the QM part and the other for the whole system ($\mathbb{S}$). The QM/MM energy of the whole system is obtained as the following equation.

$$E^{sub}_{\text{QM/MM}}(\mathbb{S}) = E_{\text{QM}}(\mathbb{QM}) + E_{\text{MM}}(\mathbb{S}) - E_{\text{MM}}(\mathbb{QM})$$

The subtractive schemes are simple since there are no explicit QM-MM coupling terms. The coupling between QM and MM is handled at the MM level, which can introduce inaccuracies for the electrostatic interaction, and the MM parameters for the QM part sometimes are difficult to obtain.

Instead of implicitly describing the QM-MM coupling terms, the additive schemes have an explicit QM-MM coupling term ($E_{\text{QM-MM}}(\mathbb{QM}, \mathbb{MM})$)for the QM and MM interaction. Following is the equation of the additive schemes.

$$E^{add}_{\text{QM/MM}}(\mathbb{S}) = E_{\text{QM}}(\mathbb{QM}) + E_{\text{MM}}(\mathbb{MM}) + E_{\text{QM-MM}}(\mathbb{QM}, \mathbb{MM})$$

This QM-MM coupling term $E_{\text{QM-MM}}$ includes bonded and non-bonded interactions between the QM and MM regions.

$$E_{\text{QM-MM}}(\mathbb{QM}, \mathbb{MM}) = E^{bonded}_{\text{QM-MM}} + E^{vdW}_{\text{QM-MM}} + E^{elec}_{\text{QM-MM}}$$

Among the three terms in $E_{\text{QM-MM}}$, the electrostatic coupling term is generally the most difficult term to calculate. There are several schemes to describe the electrostatic QM-MM interaction, and here only the electrostatic embedding is discussed. The electrostatic embedding scheme incorporates the MM point charges as one-electron

terms in the QM Hamiltonian as follows:

$$\hat{H}^{elec}_{\text{QM-MM}} = -\sum_{i}^{N}\sum_{j\in\mathbb{MM}}^{L}\frac{q_j}{|\mathbf{r}_i - \mathbf{R}_j|} + \sum_{a\in\mathbb{QM}}^{M}\sum_{j\in\mathbb{MM}}^{L}\frac{q_j Q_a}{|\mathbf{R_a} - \mathbf{R_j}|}$$

where the q_j are the MM point partial charges at $\mathbf{R}_j$; Q_a are the nuclear charges of the QM atoms at $\mathbf{R}_a$; $\mathbf{r}_i$ are the electron positions. The subscripts, i, j, and a represent the indices of the electrons, MM point charges, and QM nuclei respectively.

The van der Waals (vdW) interactions between QM and MM regions are treated at the MM level, and generally described by the Lennard-Jones potential as following:

$$E^{vdW} = \sum_{\text{pairs of atoms AB}} 4\epsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right]$$

Generally speaking, all atoms in the QM region are involved in the vdW interactions with all atoms in MM regions. However, only the atoms close to the QM-MM boundary contribute significantly.

Regarding the bonded term in the QM-MM interactions, mainly three kinds of schemes have been proposed: link-atom schemes, boundary-atom schemes, and localized-orbital schemes. Among these three schemes, boundary-atom schemes replace the MM boundary atom with a special atom that contributes electrons and basis functions to saturate the free valency of the connected QM atom. Most of the protocols in boundary-atom schemes are parameterized to reproduce certain properties, such as bond length and bond energy, by incorporating a boundary atom-centered pseudopotential. Here, the pseudoatom and pseudobond approach for *ab initio* DFT methods is the method of choice for fitting the parameters.[7,8] In the pseudoatom and pseudobond approach, the Cα atoms in proteins are generally chosen as the boundary atoms, which are defined to have seven valence electrons and nuclear charges. These

boundary carbon atoms are described by an angular-momentum-independent effective core potential (ECP) and an STO-2G basis set. The parameters of the ECP and basis set of the boundary carbon atoms are fitted to mimic the Cα-Cβ bond length, Cα-Hα bond length, and the Cα-Cβ-Hβ angle. In addition, these boundary carbon atoms still act as regular Cα atoms in the MM interactions.

1.1.2 Molecular Dynamics (MD)[9,10]

Molecular dynamics simulation is another pervasive computer simulation method for studying physical motions and interactions of atoms and molecules in a system over a period of time. In classical MD, the accelerations are computed by applying Newton's Second Law to each atom in a molecular system, $\mathbf{F}_i = m_i\mathbf{a}_i$, where F_i is the force on atom i and is equal to the nuclear gradient of the potential energy times -1, i.e. $F_i = -\nabla_{\mathbf{r}_i} E(\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_N)$. MD simulations can be run under either MM methods, or QM methods, or even QM/MM methods, depending on the choice of how to compute the potential energy. The first development and application of MD was reported in the 1960s by Alder and Wainwright, who simulated the motion of argon atoms in a box. Over the years, the performance of MD has improved significantly due to advancements in both the software algorithms and the computer hardware. In particular, graphical processing units (GPUs) are a recent technology that significantly speeds up MD simulations.

MD simulations use numerical integration, such as the Verlet algorithm, to propagate the system in time. During the simulations, the equations of motion are solved at discrete time steps, which are normally in femtoseconds due to the limitation imposed by the short periods of the bond vibrations. The positions and velocities of particles are updated according to the forces acting on them, which can be done by either QM, MM, or QM/MM methods as mentioned above. General speaking, when studying the

protein motions, pure MM MD simulations will be carried out due to the long-time scale of the protein motions. While studying the catalysis of the enzymes, QM/MM MD simulations or pure QM MD simulations can be used to describe the reactions.

Here we focus on introducing the MM methods. The classical potential energy function (or force field) contains bonded terms and nonbonded terms. The bonded terms typically include bond stretching, angle bending, proper torsion, and improper torsion while the nonbonded terms are described using the van der Waals terms and Coulomb interaction terms. As mentioned above, the van der Waals interaction is generally described by the Lennard-Jones equation. Therefore, the energy function can be written as follows.

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

$$+ \sum_{\text{vdW AB}} 4\epsilon_{AB} \left[\left(\frac{\sigma_{AB}}{r_{AB}} \right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}} \right)^{6} \right] + \sum_{\text{elec AB}} \frac{1}{4\pi\epsilon_0} \frac{q_A q_B}{r_{AB}}$$

where $d, \theta,$ and ϕ represent bond lengths, angles, and torsions correspondingly; d_0 and θ_0 are the bond and angle equilibrium values, respectively; n and δ are torsional multiplicity and phase; r_{AB} is the nonbonded distance between atoms A and B, and ϵ_{AB} and σ_{AB} are the Lennard-Jones parameters; q_A, q_B are atomic partial charges for atoms A and B, and ϵ_0 is the permittivity of free space.

Although the MD simulations nowadays can propagate for multi-microsecond of a moderately sized system in an explicit solvent model ($\approx 10^5$ atoms), it is still shorter than many of the processes in biomolecular systems including the majority of protein folding processes. Besides, sometimes it is not efficient to carry out a long-time MD simulation to capture a simple process. Therefore, enhanced sampling is necessary. Enhanced sampling is a range of techniques in MD simulations to overcome

the limitations of conventional MD and help to sample the rare events that cannot be easily captured. There are several methods of enhanced sampling[11] in MD simulations, including replica exchange MD (REMD)[12], umbrella sampling[13], metadynamics[14], and accelerated MD (aMD).[15] Here I briefly introduce umbrella sampling since it has been used a lot in my research.

Umbrella sampling has been used to study a wide range of applications, including protein folding, chemical reactions, ligand binding and transfer, and others in biomolecules or material systems. It has been proven to be powerful and efficient for investigating complicated situations at the atomic level that are hardly studied by experiments. These simulations are carried out by applying a biasing harmonic potential along the defined reaction coordinate (RC). Multiple simulations are run with different biasing potentials spaced out along the reaction coordinate. The *unbiased* free energy profile along the RC can then be calculated from the sampled distribution of the system along the defined RC. Several methods can combine the sample data from different windows and calculate the free energy difference along the RC, such as weighted histogram analysis method (WHAM)[16] or multistate Bennett acceptance ratio (MBAR).[17] WHAM is one of frequently used methods to analyze the umbrella sampling data, and calculates the relative probability of observing the states from the histogram with discrete sampling windows and further converts the probability to a potential of mean force (PMF), also called the free energy profile.

1.2 Brief Introduction of Hydrogenases and Circadian Protein KaiB.

1.2.1 Hydrogenases

Hydrogenases are enzymes catalyzing the interconversion between hydrogen gas (H_2) and protons (H^+).[18] They are found in many organisms, including bacteria, and some eukaryotes. Hydrogenases are attractive since they require a relatively low over potential to reduce proton to hydrogen gas. According to the structures, hydrogenases can be classified into three classes, namely [Fe]-hydrogenases, [NiFe] hydrogenases, and [FeFe] hydrogenases.[19] [Fe]-hydrogenases contain only one Fe at the catalytic centers (in Figure 1.1) and no iron-sulfur clusters. Due to this special structure, the [Fe]-hydrogenases have been found to catalyze the reversible heterolytic cleavage of H_2 by $H_2 \rightleftharpoons H^+ + H^-$ instead of a true redox reaction. [NiFe] and [FeFe] hydrogenases both have a di-ionic cluster and iron-sulfur cluster in the active site and catalyze a true redox reaction to generate H_2 as $H_2 \rightleftharpoons 2H^+ + 2e^-$. Among these three kinds of hydrogenases, [FeFe] hydrogenases are generally the most active in the production of hydrogen gas. The turnover frequency of [FeFe]-hydrogenases have been reported in the order of 10,000 s^{-1}. This has led to intense studies on [FeFe]-hydrogenases for sustainable production of hydrogen gas.[20]

As shown in Figure 1.1, the catalytic center of [FeFe] hydrogenases, called the H-cluster, is constructed with a traditional [Fe_4S_4] cluster connected to a di-iron cluster with a bridging dithiolate cofactor. The maturation of [FeFe] hydrogenases is the process to build the H-cluster and deliver it to the Hydrogenases A (HydA).[21] Experiments have been done to show the essential roles of HydE, HydF, and HydG in H-cluster biosynthesis. Isotopic labels in the terminal CN^- ligands of HydA have been found when labeling the tyrosine as [^{15}N]-Tyr or [^{13}C]-Tyr. Besides, ^{57}Fe and ^{57}Fe-

CO/CN vibrational modes have been detected in HydA when only the HydG lysate is ^{57}Fe-enriched. These results led to the hypothesis that HydG generates its product by decomposing tyrosine. Recent experiments proposed that the HydG product, [Fe(II)$(CO)_2$(CN)Cys] synthon (Fe-synthon), acts as the substrate of HydE, which converts it into an Fe(I) intermediate. Furthermore, the di-iron cluster [Fe(II)$(CO)_2$(CN)Cys] combined with HydF and lysate, has been proven to be able to maturate HydA, which led to the hypothesis that the bridging dithiolate cofactor is assembled in HydF and later the whole cluster is delivered to HydA to build up the H-cluster.

Figure 1.2 shows the current hypothetical mechanism of [FeFe] hydrogenase maturation. In the beginning, HydG uses tyrosine and cysteine as the substrates to synthesize the Fe-synthon ([Fe(II)$(CO)_2$(CN)Cys]) where the diatomic ligands CN and CO are sourced from the tyrosine substrate. According to the stochiometric number of the CO ligand in the Fe-synthon, it takes two tyrosines in HydG to complete the synthesis.[22] Later, the Fe-synthon is delivered to HydE and turns into a 5-coordinated Fe(I) cluster. It is proposed that the 5-coordinated Fe(I) cluster would further dimerize in HydE into a diamagnetic [Fe(I)$_2$$(SH)_2$$(CO)_4$$(CN)_2$] cluster.[23] The di-iron cluster is finally decorated by HydF with the dithiolate cofactor and transferred to *apo* HydA to finish the maturation process. The possible catalytic mechanism of the H-cluster in HydA is also displayed in Figure 2. The distal Fe (Fe_d) has 5 connected ligands most of the time, which enables the Fe_d to accommodate another small ligand. The traditional [Fe_4S_4] cluster acts as an electron reservoir by providing electrons to reduce the Fe_d, while bridging dithiomethylamine (DTMA) plays a role as a proton repository by transferring protons from the environment to the Fe_d.

Although many experiments have been done to study the functions of HydG and HydE, many mechanistic questions remain to be answered, such as: (1) How does the tyrosine in HydG become two diatomic ligands? (2) What is the function of the

auxiliary cluster in HydG? (3) How does HydE cleave the cysteine and deoxyadenosine from the Fe(I) intermediate? (4) Where does the Fe(I) cluster dimerize? These questions are highly difficult to answer using experiments alone, and computations can play a key role in understanding these mechanisms in detail.

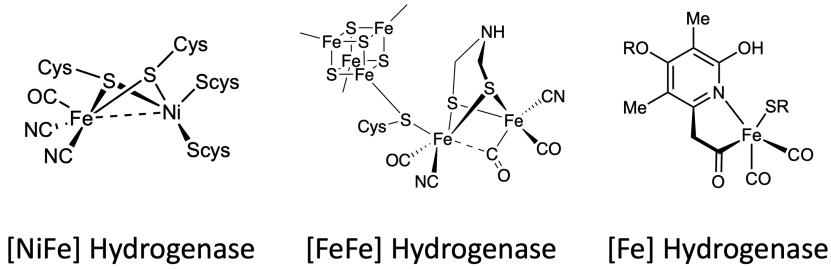


Figure 1.1: The catalytic centers of three kinds of hydrogenases.

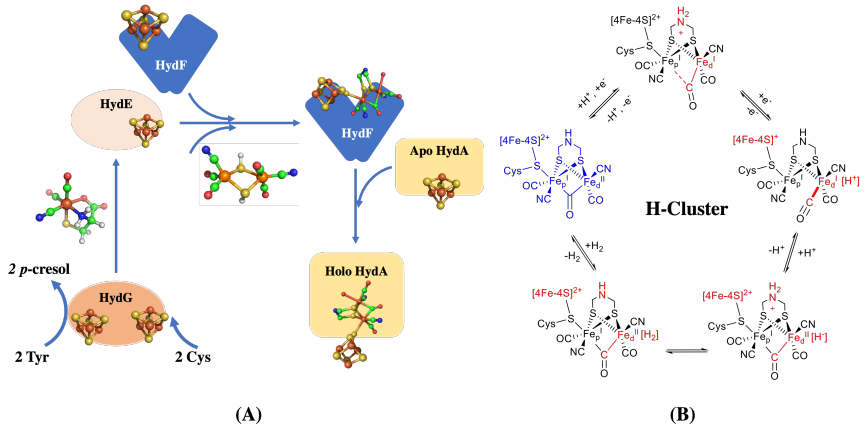


Figure 1.2: The hypothetical mechanism of [FeFe]-hydrogenase maturation (A) and the possible catalytic cycle of the H-cluster (B).

1.2.2 Circadian Protein KaiB

Circadian clocks are biochemical systems with predictable oscillations that track the rotation of the Earth, and are ubiquitous in living organisms. In the past decade, scientists, including Dr. Andy LiWang at UC Merced, found that cyanobacteria have a robust circadian clock that can continue working when reconstituted in vitro.[24–26] The

circadian oscillator of cyanobacteria comprises three proteins, namely KaiA, KaiB, and KaiC.[25,27] Among these three circadian proteins, the phosphorylation of KaiC at residues Ser431 and Thr432 oscillates on a 24-hour time scale, and the autokinase and autophosphatase functions of KaiC are essential to keep the stable 24-hour period.[28–30] As shown in Figure 1.3, KaiC is the ring-shaped, bi-lobal homo-hexameric protein whose N and C terminals form CI and CII rings respectively. The peptide in the C terminal is called the A-loop which can be bound by KaiA.[31] KaiA is a dimer in nature that can bind to the A-loop of KaiC to stimulate the KaiC auto-kinase activity.[32] In contrast to KaiA, KaiB can promote the KaiC intrinsic autophosphatase activity by sequestering the alternative structure of KaiA.[33] As illustrated in Figure 1, KaiC starts from an unphosphorylated state at dawn. As the morning progresses, KaiA binds to the A-loop of KaiC, stimulating the phosphorylation of KaiC.[34,35] The phosphorylation happens in a highly ordered manner involving Thr432 first, followed by Ser431.[35] At dusk, both Thr432 and Ser431 are phosphorylated, and KaiC changes the structure to make CI and CII rings in a stacking structure.[33,35] At the same time, the A-loop recedes inside the CII ring while CI is exposed to bind to the KaiB.[36] KaiB undergoes a metamorphic change from a ground state (GS) dimer to a fold-switched (FS) monomer.[37–39] Due to the release of KaiA and KaiC, and the binding of KaiB and KaiC, KaiC begins to dephosphorylate starting with Thr432 followed by Ser431.[40,41] Finally, at dawn, both Ser431 and Thr432 are dephosphorylated, and CI and CII return to their unstacked structure, completing the cycle. According to cyanobacterial circadian processes, KaiB plays a key role in regulating the KaiC phosphorylation oscillation by a rare change in structure from dimer GS to monomer FS.

Understanding the fold-switching mechanism of KaiB can help people adjust the circadian rhythm of cyanobacteria. Long-time MD simulations can provide a sight

of the fold-switching pathway and predict possible intermediates. Besides, the fold-switching property of KaiB protein also provides a good example to explore the other possible metamorphic proteins. By extracting comment features on these metamorphic proteins, one can have a model to predict the metamorphic behavior of the existing proteins and their structures as well.

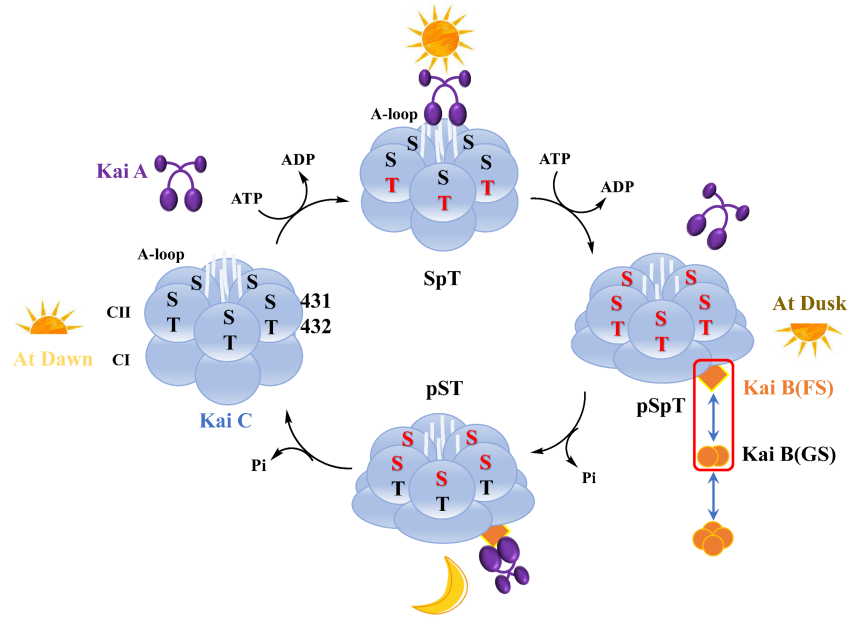


Figure 1.3: Mechanism for the cyanobacterial circadian oscillator. The red letters "S" and "T" represent the phosphorylation of serine 431 and threonine 432 on KaiC respectively, while the black letters represent their unphosphorylated state. KaiB is a metamorphic protein that undergoes a dramatic structural change from the ground state (GS, orange circle) to the fold-switched state (FS, orange diamond). The FS KaiB binds to KaiC, activating the KaiC autophosphatase activities and dephosphorylating S431 and T432. Finally, KaiC returns to its original unphosphorylated state.

1.3 Research in my Ph.D. Period

In my Ph.D. period, a lot of my attention was paid to the maturation of [FeFe]-hydrogenases. I first conducted theoretical studies on the catalytic process of HydG. The HydG crystal structures were reported by Dinis et al. in 2015, providing a good

initial structure for theoretical studies. The structure reveals that, unlike the traditional SAM enzyme, HydG has two iron-sulfur clusters. One is the traditional [Fe_4S_4] cluster, named canonical iron-sulfur cluster, coordinated with the SAM molecule, and the other is a [Fe_5S_5], named auxiliary iron-sulfur cluster, constructed by a [Fe_4S_4] cluster and a dangler iron with a cysteine. By carrying out a series of QM/MM simulations, I proposed a radical-relay mechanism for HydG to decompose the tyrosine substrate into $COO^{\bullet-}$ and CN^-. Furthermore, the auxiliary iron-sulfur cluster is postulated to help reduce the $COO^{\bullet-}$ after excluding all the other possibilities around the canonical iron-sulfur cluster.

After solving the HydG catalytic mechanism, I start to study the HydE catalytic mechanism. As mentioned above, HydE is the downstream protein of HydG in [FeFe]-hydrogenase maturation. Similar to the HydG simulation process, the QM/MM simulations were set up based on the HydE crystal structure (PDB ID: 7O1P).[42] Unlike the traditional radical SAM enzyme that operates by hydrogen atom transfer, HydG carries out a C-S radical addition and transfers the radical to the Fe(II) in the Fe-synthon, reducing it to Fe(I). This first Fe(I) is named the 10s intermediate (IM) due to its generation time. This intermediate then transforms into a new 5-coordinate Fe(I) cluster within 10 minutes while releasing pyruvate as the side-product. The existence of this 5-coordinated Fe(I) intermediate has been confirmed by electronic paramagnetic resonance (EPR) experiments, and is named the 10-min IM. Finally, the dimerization of the 10-min IM has been discussed, and a possible pathway has been proposed to show the feasibility of the dimerization.

Besides the [FeFe]-hydrogenases mechanism, I also spent my Ph.D. period studying the fold-switching pathway of the KaiB protein, and the possible feature for metamorphic proteins. Since the KaiB fold-switching is a structural change at the time scale of 24 hours, which is unaffordable for traditional MD simulations, the enhanced

sampling technique combined with the Markov state model (MSM) has been utilized. In addition, I also searched for the common features of metamorphic proteins by collecting a dataset of other existing metamorphic proteins from the literature. In my publication, a called the diversity index (DI) was proposed and I proved that it is capable of distinguishing the metamorphic proteins from the others. The DI is calculated based on the secondary structure prediction results. By checking how much the protein sequence "confuses" the secondary structure prediction programs, we can have a probability that whether the given sequence is a metamorphic protein sequence or not.

References

[1]F. Jensen, Introduction to computational chemistry, 2nd ed. (John Wiley and Sons, 2007).

[2]H. M. Senn and W. Thiel, "Qm/mm methods for biomolecular systems", Angewandte Chemie International Edition **48**, 1198–1229 (2009).

[3]A. Warshel and M. Levitt, "Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme", Journal of molecular biology **103**, 227–249 (1976).

[4]H. M. Senn and W. Thiel, "Qm/mm studies of enzymes", Current opinion in chemical biology **11**, 182–187 (2007).

[5]E. Boulanger and J. N. Harvey, "Qm/mm methods for free energies and photochemistry", Current Opinion in Structural Biology **49**, 72–76 (2018).

[6]M. Rossano-Tapia and A. Brown, "Quantum mechanical/molecular mechanical studies of photophysical properties of fluorescent proteins", Wiley Interdisciplinary Reviews: Computational Molecular Science **12**, e1557 (2022).

[7]Y. Zhang, "Pseudobond ab initio qm/mm approach and its applications to enzyme reactions", Theoretical Chemistry Accounts **116**, 43–50 (2006).

[8]Y. Zhou, S. Wang, Y. Li, and Y. Zhang, "Born–oppenheimer ab initio qm/mm molecular dynamics simulations of enzyme reactions", in Methods in enzymology, Vol. 577 (Elsevier, 2016), pp. 105–118.

[9]M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules", Nature structural biology **9**, 646–652 (2002).

[10]T. Hansson, C. Oostenbrink, and W. van Gunsteren, "Molecular dynamics simulations", Current opinion in structural biology **12**, 190–196 (2002).

[11]R. C. Bernardi, M. C. Melo, and K. Schulten, "Enhanced sampling techniques in molecular dynamics simulations of biological systems", Biochimica et Biophysica Acta (BBA)-General Subjects **1850**, 872–877 (2015).

[12]A. E. Garcia, H. Herce, and D. Paschek, "Simulations of temperature and pressure unfolding of peptides and proteins with replica exchange molecular dynamics", Annual Reports in Computational Chemistry **2**, 83–95 (2006).

[13]J. Kästner, "Umbrella sampling", Wiley Interdisciplinary Reviews: Computational Molecular Science **1**, 932–942 (2011).

[14]A. Barducci, M. Bonomi, and M. Parrinello, "Metadynamics", Wiley Interdisciplinary Reviews: Computational Molecular Science **1**, 826–843 (2011).

[15]D. Perez, B. P. Uberuaga, Y. Shim, J. G. Amar, and A. F. Voter, "Accelerated molecular dynamics methods: introduction and recent developments", Annual Reports in computational chemistry **5**, 79–98 (2009).

[16]B. Roux, "The calculation of the potential of mean force using computer simulations", Computer physics communications **91**, 275–282 (1995).

[17]Z. Tan, E. Gallicchio, M. Lapelosa, and R. M. Levy, "Theory of binless multi-state free energy estimation with applications to protein-ligand binding", The Journal of chemical physics **136**, 04B608 (2012).

[18]W. Lubitz, H. Ogata, O. Rüdiger, and E. Reijerse, "Hydrogenases", Chemical reviews **114**, 4081–4148 (2014).

[19]S. Shima and R. K. Thauer, "A third type of hydrogenase catalyzing h2 activation", The chemical record **7**, 37–46 (2007).

[20]F. Wittkamp, M. Senger, S. Stripp, and U.-P. Apfel, "[fefe]-hydrogenases: recent developments and future perspectives", Chemical Communications **54**, 5934–5942 (2018).

[21]A. Pagnier, B. Balci, E. M. Shepard, W. E. Broderick, and J. B. Broderick, "[fefe]-hydrogenase in vitro maturation", Angewandte Chemie International Edition **61**, e202212074 (2022).

[22]P. Dinis, D. L. Suess, S. J. Fox, J. E. Harmer, R. C. Driesener, L. De La Paz, J. R. Swartz, J. W. Essex, R. D. Britt, and P. L. Roach, "X-ray crystallographic and epr spectroscopic analysis of hydg, a maturase in [fefe]-hydrogenase h-cluster assembly", Proceedings of the National Academy of Sciences **112**, 1362–1367 (2015).

[23]Y. Zhang, L. Tao, T. J. Woods, R. D. Britt, and T. B. Rauchfuss, "Organometallic fe2(μ-sh)2(co)4(cn)2 cluster allows the biosynthesis of the [fefe]-hydrogenase with only the hydf maturase", Journal of the American Chemical Society **144**, 1534–1538 (2022).

[24]T.-H. Chen, T.-C. Huang, and T.-J. Chow, "Calcium requirement in nitrogen fixation in the cyanobacterium synechococcus rf-1", Planta **173**, 253–256 (1988).

[25]M. Ishiura, S. Kutsuna, S. Aoki, H. Iwasaki, C. R. Andersson, A. Tanabe, S. S. Golden, C. H. Johnson, and T. Kondo, "Expression of a gene cluster kaiabc as a circadian feedback process in cyanobacteria", Science **281**, 1519–1523 (1998).

[26]T. Kondo, C. A. Strayer, R. D. Kulkarni, W. Taylor, M. Ishiura, S. S. Golden, and C. H. Johnson, "Circadian rhythms in prokaryotes: luciferase as a reporter of circadian gene expression in cyanobacteria.", Proceedings of the National Academy of Sciences **90**, 5672–5676 (1993).

[27]T. Kondo, N. F. Tsinoremas, S. S. Golden, C. H. Johnson, S. Kutsuna, and M. Ishiura, "Circadian clock mutants of cyanobacteria", Science **266**, 1233–1236 (1994).

[28]T. Nishiwaki, Y. Satomi, M. Nakajima, C. Lee, R. Kiyohara, H. Kageyama, Y. Kitayama, M. Temamoto, A. Yamaguchi, A. Hijikata, et al., "Role of kaic phosphorylation in the circadian clock system of synechococcus elongatus pcc 7942", Proceedings of the National Academy of Sciences **101**, 13927–13932 (2004).

[29]Y. Xu, T. Mori, R. Pattanayek, S. Pattanayek, M. Egli, and C. H. Johnson, "Identification of key phosphorylation sites in the circadian clock protein kaic by crystallographic and mutagenetic analyses", Proceedings of the National Academy of Sciences **101**, 13933–13938 (2004).

[30]T. Nishiwaki, H. Iwasaki, M. Ishiura, and T. Kondo, "Nucleotide binding and autophosphorylation of the clock protein kaic as a circadian timing process of cyanobacteria", Proceedings of the National Academy of Sciences **97**, 495–499 (2000).

[31]R. Pattanayek, J. Wang, T. Mori, Y. Xu, C. H. Johnson, and M. Egli, "Visualizing a circadian clock protein: crystal structure of kaic and functional insights", Molecular cell **15**, 375–388 (2004).

[32]Y.-I. Kim, G. Dong, C. W. Carruthers Jr, S. S. Golden, and A. LiWang, "The day/night switch in kaic, a central oscillator component of the circadian clock of cyanobacteria", Proceedings of the National Academy of Sciences **105**, 12825–12830 (2008).

[33]Y.-G. Chang, R. Tseng, N.-W. Kuo, and A. LiWang, "Rhythmic ring–ring stacking drives the circadian oscillator clockwise", Proceedings of the National Academy of Sciences **109**, 16847–16851 (2012).

[34]T. Nishiwaki, Y. Satomi, Y. Kitayama, K. Terauchi, R. Kiyohara, T. Takao, and T. Kondo, "A sequential program of dual phosphorylation of kaic as a basis for circadian rhythm in cyanobacteria", The EMBO journal **26**, 4029–4037 (2007).

[35]Y.-G. Chang, N.-W. Kuo, R. Tseng, and A. LiWang, "Flexibility of the c-terminal, or cii, ring of kaic governs the rhythm of the circadian clock of cyanobacteria", Proceedings of the National Academy of Sciences **108**, 14431–14436 (2011).

[36]R. Tseng, Y.-G. Chang, I. Bravo, R. Latham, A. Chaudhary, N.-W. Kuo, and A. LiWang, "Cooperative kaia–kaib–kaic interactions affect kaib/sasa competition in the circadian clock of cyanobacteria", Journal of molecular biology **426**, 389–402 (2014).

[37]K. Hitomi, T. Oyama, S. Han, A. S. Arvai, and E. D. Getzoff, "Tetrameric architecture of the circadian clock protein kaib: a novel interface for intermolecular interactions and its impact on the circadian rhythm", Journal of Biological Chemistry **280**, 19127–19135 (2005).

[38]I. Vakonakis, D. A. Klewer, S. B. Williams, S. S. Golden, and A. C. LiWang, "Structure of the n-terminal domain of the circadian clock-associated histidine kinase sasa", Journal of molecular biology **342**, 9–17 (2004).

[39]R. Tseng, N. F. Goularte, A. Chavan, J. Luu, S. E. Cohen, Y.-G. Chang, J. Heisler, S. Li, A. K. Michael, S. Tripathi, et al., "Structural basis of the day-night transition in a bacterial circadian clock", Science **355**, 1174–1180 (2017).

[40]M. Egli, T. Mori, R. Pattanayek, Y. Xu, X. Qin, and C. H. Johnson, "Dephosphorylation of the core clock protein kaic in the cyanobacterial kaiabc circadian oscillator proceeds via an atp synthase mechanism", Biochemistry **51**, 1547–1558 (2012).

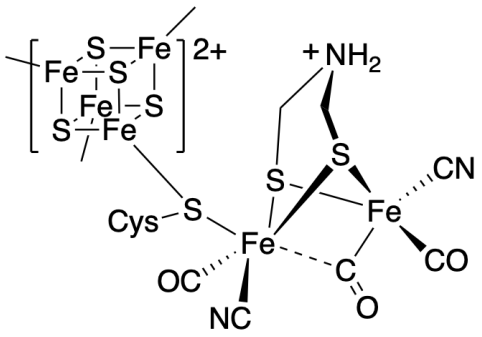
[41]T. Nishiwaki and T. Kondo, "Circadian autodephosphorylation of cyanobacterial clock protein kaic occurs via formation of atp as intermediate", Journal of Biological Chemistry **287**, 18030–18035 (2012).

[42]R. Rohac, L. Martin, L. Liu, D. Basu, L. Tao, R. D. Britt, T. B. Rauchfuss, and Y. Nicolet, "Crystal structure of the [fefe]-hydrogenase maturase hyde bound to complex-b", Journal of the American Chemical Society **143**, 8499–8508 (2021).

2 Quantum chemical study of a radical relay mechanism for the HydG-catalyzed synthesis of a Fe(II)$(CO)_2$(CN)cysteine precursor to the H-cluster of [FeFe] hydrogenase[1]

2.1 Introduction

Hydrogenases are fascinating metalloenzymes that catalyze the reversible interconversion of H_2 and H^+/e^-. They are categorized into [FeFe], [NiFe], and [Fe] subtypes according to the metal composition of their active site cofactors.[1,2] The [FeFe] hydrogenase is highly active in H_2 production, with rates up to 10^4/s, making it of great interest to the renewable energy community.[3] This high activity is rendered by the unique catalytic center of [FeFe] hydrogenases, a six-Fe "H-cluster" consisting of a $[4\,Fe{-}4\,S]_H$ cluster linked through a cysteine S to a $[2\,Fe]_H$ cluster in which the two iron centers are coordinated by diatomic CO and CN^- ligands, as well as an unusual azadithiolate (adt, $NH(CH_2S^-)_2$) bridging ligand (Scheme 1). The unique structure and activity of the H-cluster thus raise the intriguing question as to its biosynthesis; a multi-component, step-by-step assembly made challenging by toxic ligands, oxygen sensitivity, and the inherent chemical instability of the adt moiety.



Scheme 1: The H-cluster in the active site of [FeFe] hydrogenases.

Genetic and biochemical studies have shown that three Fe-S cluster proteins, HydE, HydF and HydG, are essential to the biosynthesis of the H-cluster.[4,5] In particular, it is demonstrated that the radical S-adenosyl-L-methionine (rSAM) enzyme HydG is responsible for the biogenesis of the toxic CO and CN^- ligands and their passivation by the formation of a $[Fe(CO)_2(CN)(cysteinate)]$ organometallic synthon.[6–9] Recent work shows that this HydG product synthon serves as the substrate for another rSAM enzyme, HydE, which in turn generates $[Fe^{I}(CO)_2(CN)S]$-containing intermediates that are proposed to undergo dimerization to form the core of the $[2\,Fe]_H$ subcluster.[10] It was also recently shown that the $NH(CH_2)_2$ component of the adt ligand is sourced from serine.[11] Once assembled, presumably on the HydF protein,[12] the $[2Fe]_H$ cluster is delivered to an apo-hydrogenase that harbors only the $[4Fe\text{-}4S]_H$ subcluster, allowing the completion of the fully active H-cluster.[13]

Despite these recent advances, questions remain regarding the detailed steps of the bio-assembly pathway, even those involving the best understood maturase, HydG. Specifically, the molecular mechanism of CO and CN^- formation is only partially defined by experiments. HydG contains two Fe-S clusters at either end of a 24 Å hydrophobic TIM barrel. These two clusters fulfill two distinct functions.[14] As shown in Figure 2.1, the N-terminal cluster, a rSAM $4\,Fe-4\,S$ cluster such as found in all

radical SAM enzymes, initiates the rSAM chemistry to generate the 5′-dAdo$^{\bullet}$ radical which abstracts an amino hydrogen atom from tyrosine and induces Cα-Cβ homolysis to form a 4-hydroxylbenzyl radical (4-OB$^{\bullet}$) as detected by EPR spectroscopy, along with a proposed dehydroglycine (DHG) molecule.[15] It is thought that the DHG is then converted into a CO and CN^- pair inside HydG; these diatomic ligands are delivered to the C-terminal [4Fe-4S]-[Fe(cysteinate)] auxiliary cluster, where they bind to the unique fifth "dangler Fe" to form a [4Fe-4S]-[Fe(CO)(CN)(cysteinate)] intermediate revealed by stop-flow FTIR and EPR spectroscopy.[16–18] A second pair of CO/CN^- generated from Tyr further convert this intermediate to the [$Fe(CO)_2$(CN)(cysteinate)] synthon product and a [4Fe-4S]-CN cluster that is subsequently reconverted to the resting state configuration with a fresh cysteine ligand replacing the CN^- and then binding a new Fe^{2+}.[16]

It remains elusive how DHG undergoes C-C bond cleavage to form CO and CN^-; the simple hydrolysis reaction would yield ammonium and glyoxylate instead. It is also unclear whether and how the auxiliary cluster is involved in the decomposition of DHG, although this relevance is inferred from the observation that the auxiliary cluster knock-out mutant of HydG generates only CN^- at a much slower rate without any detectable CO,[9] and that the H265N mutant, which abolishes the dangler Fe in the auxiliary cluster, generates cyanide and formate only, again without CO.[19] These results seem to imply, as alleged in the latter study, that the formation of CN^- may occur at the rSAM site, whereas the formation of CO requires the dangler Fe site in the auxiliary cluster. This proposal is yet incomplete however, in that the detailed reaction mechanism for both CO and CN^- formation needs to be clarified, and also given that the ligand environment and the electronic structure of the dangler Fe are dramatically altered upon binding of the first CO/CN^- pair (high-spin, S = 2 to low-spin, S = 0),[18] so the chemistry leading to the second CO ligand must differ from

that of the first CO and CN^- addition. Despite these previous efforts, it is challenging to address these remaining issues with purely experimental approaches. Tracking the fate of DHG in the active site pocket and the protein channel is difficult to achieve since relevant spectroscopic markers are currently lacking.

Here we investigate the reactions in HydG using computational quantum chemistry in order to gain insights into the experimentally inaccessible portions of the catalytic mechanism and guide further experimental design. The study involves two main parts: the first part is a series of hybrid QM/MM molecular dynamics calculations of the reactions at the canonical Fe-S cluster. Here, the 5'-deoxyadenosyl radical (5'dAdo$^\bullet$) initiator cleaves the tyrosine substrate, the products of which proceed through a relay of radical intermediates ending in HCN and a $COO^{\bullet-}$ radical anion. The second part is a broken symmetry DFT study of the reactions at the auxiliary Fe-S cluster where two equivalents of CN^- and $COOH^\bullet$ coordinate to the dangler Fe in a series of substitution and redox reactions that yield the synthon as the final product. The presented mechanistic hypothesis is supported by computational data and consistent with experimental results, and reveals important and previously hidden features of the catalytic mechanism of HydG.

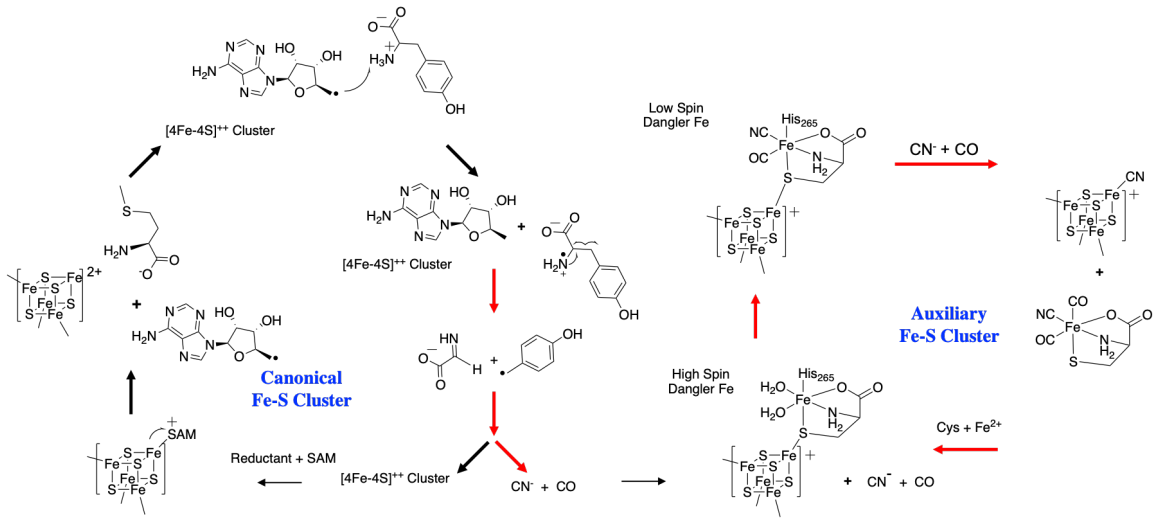


Figure 2.1: Summary of reactivity at the HydG rSAM Fe-S cluster (left) and auxiliary cluster (right) leading to the production of the $Fe(CO)_2CN$ synthon. The cycles on the left and right represent the reactions that occur at the canonical and auxiliary iron-sulfur clusters, respectively. Herein, the red arrows indicate the controversial steps in HydG that we focus on in the study.

2.2 Computational Methods

Herein, the crystal structure of HydG (pdb ID: 4WCX) was used as a starting point to study the catalytic mechanism.[14,16,20] The reactions that occur around the canonical [4Fe-4S] cluster were simulated using a hybrid density functional theory QM/MM umbrella sampling approach with the help of the Q-Chem/AMBER interface.[21–23] The reactions occurring at the auxiliary [4Fe-4S] cluster were studied using a cluster model of the active site due to elevated computational cost, with calculations performed by the TeraChem software package.[24–26] EPR properties for selected structures were computed using the ORCA software package.[27,28] Further details of the calculations including structural modeling, the level of QM theory,[29,30] basis set,[31–34] and the choice of QM region and force fields[35–38] used in the QM/MM calculations are provided in the Supporting Information.

2.3 Results and Discussions

2.3.1 Tyrosine Radical Formation.

It has been reported by Britt, et al. that the HydG catalytic reaction is initiated by electron transfer from the reduced rSAM Fe-S cluster to its bound SAM cofactor, leading to the formation of a 5'-deoxyadenosyl radical (5'dAdo$^{\bullet}$) that initiates tyrosine homolytic cleavage.[39,40] Because the Fe-S cluster induced decomposition of SAM is a well-known and well-studied radical reaction in biological systems, it is not a focal point of this study, and here we only aim to demonstrate consistency with experiments. By driving the defined reaction coordinate RC1 = d(C. . .S) – d(Fe. . .S) (defined in Figure S4), we found that the activation energy of this reaction is approximately 26.3 kcal/mol (Figure S4) which is comparable with the HydG experimental kinetics studies (about 23 kcal/mol).[41] Following this step, Dinis et al. proposed that the 5'-dAdo$^{\bullet}$ radical abstracts a hydrogen atom from the tyrosine amino group[14], by analogy to the tryptophan lyase NosL, where X-ray structural analysis and computations indicated that H-atom abstraction occurs at a tryptophan amino group.[42] As shown in Figure 2.2, the key structures of tyrosine radical generation (Figure 2(A)) and the corresponding diagrams (Figure 2(B) and 2(C)) were depicted. The activation free energy of H$^{\bullet}$ abstraction from the tyrosine amino group was calculated as 15 kcal/mol, demonstrating that this H-atom abstraction is feasible at room temperature and is exothermic, releasing more than 6 kcal/mol of energy. This exothermic nature of the reaction is attributed to the increased electronegativity of the tyrosyl N relative to the 5'-dAdo$^{\bullet}$ C1 atom and by the rapid delocalization of the radical over the aromatic tyrosyl side chain, which we observed by monitoring the spin density along the reaction coordinate. We also considered an alternative pathway involving H-atom abstraction from the tyrosine α carbon in Figure S5, and ruled it out due to

the activation barrier being higher by ≈5 kcal/mol.

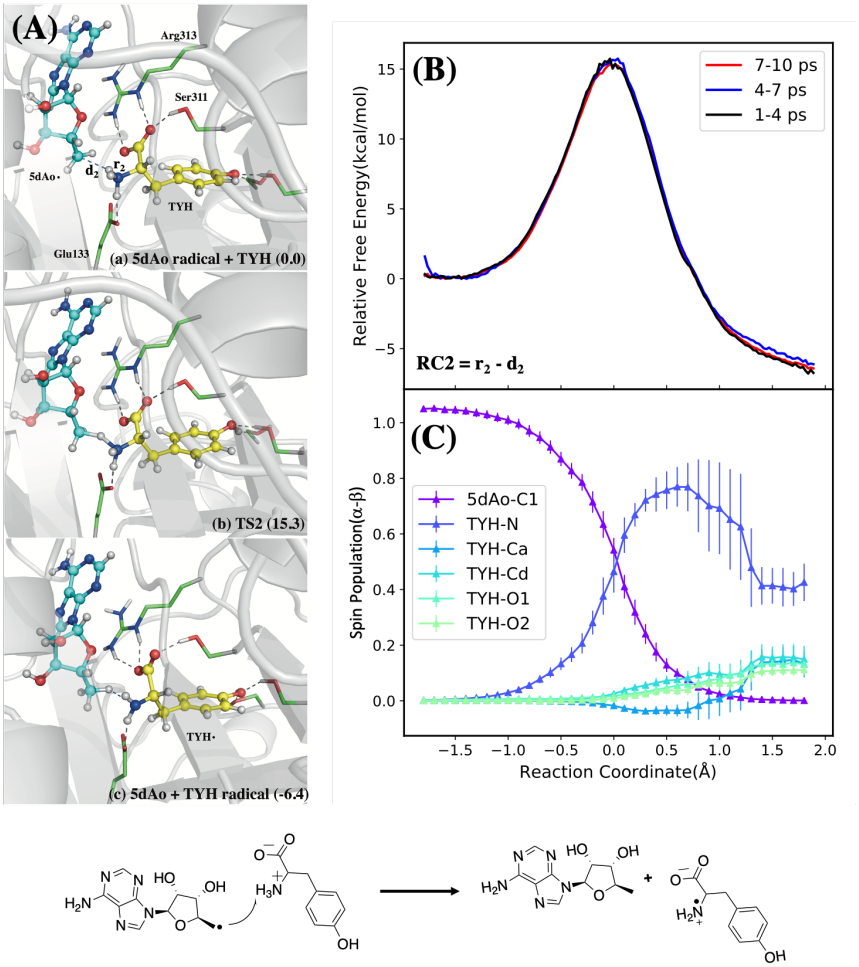


Figure 2.2: Radical transfer from 5-Ado radical to tyrosine substrate. (A): Structures of reactant, transition state, and product. (B): QM/MM free energy profile calculated using different QM/MM trajectory segments with the defined reaction coordinate (RC). (C): Mulliken spin populations along the reaction coordinate with error bars taken from the standard deviations of the QM/MM MD simulations. Bottom: Reaction scheme.

2.3.2 Tyrosine Decomposition.

As mentioned above, HydG and NosL share many similarities both in terms of protein sequence and functionality.[14,20] However, there is one significant difference in the

catalytic mechanism between NosL and HydG: it is reported that NosL catalyzes the cleavage of the tryptophan Cα—C(=O) bond which generates a $COO^{\bullet -}$ radical instead of a radical localized on the indole side chain.[42] In HydG, it is believed that the DHG is a key intermediate, which indicates that the cleavage of Cα—Cβ is required instead. We calculated the activation energies starting from different protonation states of the Tyr carboxyl group and found that it plays an important role in directing which Cα—C bond undergoes homogenous cleavage. As shown in Figure S6, the relative activation energies of the two different C—C cleavage mechanisms depends strongly on the protonation state of the model, and the zwitterionic TYH model tends to undergo Cα—Cβ cleavage while the neutral TYY model tends to undergo Cα—C(=O) cleavage. The analysis of frontier orbitals in Figure S6 supports this result; the β LUMO, which is singly occupied, possesses σ-bonding character between Cα—Cβ in TYH that is absent in TYY, and this is consistent with TYH favoring Cα—Cβ cleavage. All of these results suggest that TYH is the appropriate protonation state in HydG, and differences in protonation state may contribute to the distinct mechanisms of NosL and HydG, along with the intrinsic difference between Tyr and Trp side chains.

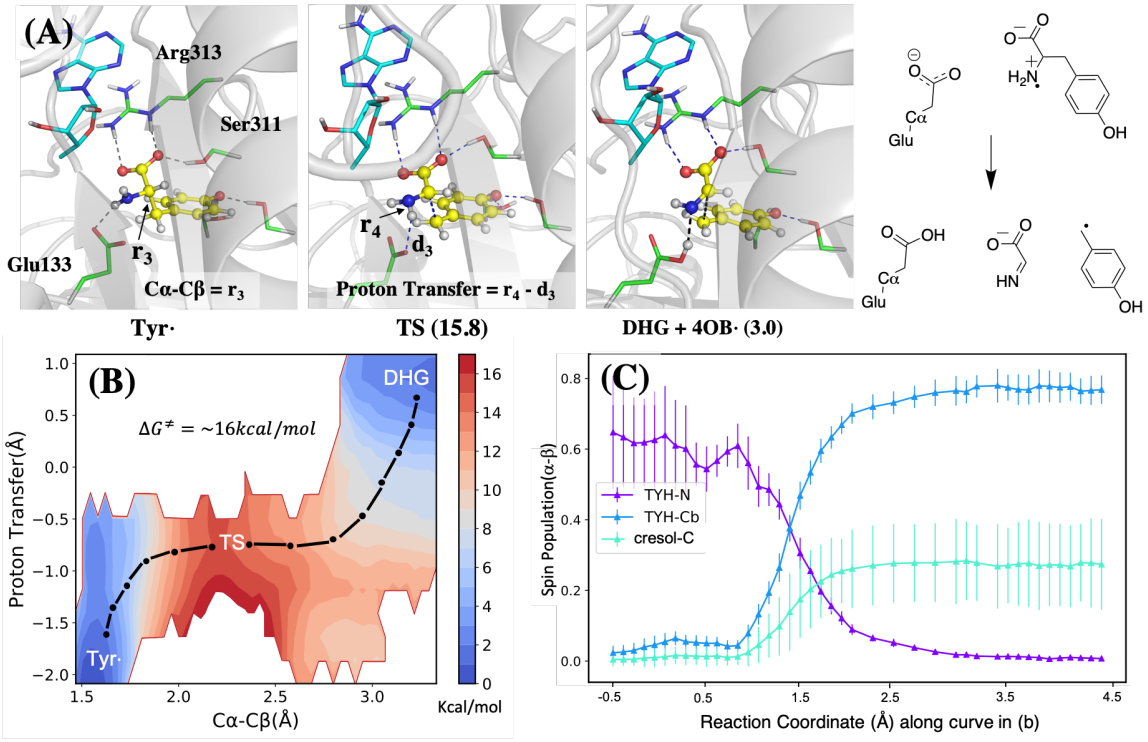


Figure 2.3: The calculated mechanism of tyrosine radical decomposition and DHG formation. The structures of three key states are shown on top (A), and the free energy map and spin density of different atoms are shown at the bottom left (B) and right (C) respectively. The color bar in (B) indicates the scale of the energy in kcal/mol. The spin population in (C) is plotted along the one-dimensional path indicated using the black dot-dash in the free energy map (B). The 2D scheme of the reaction is shown at the top right.

The results of tyrosine decomposition to DHG and 4-OB$^\bullet$ are shown in Figure 2.3. The mechanism involves the transfer of a proton from the tyrosyl radical H2N$^\bullet$ moiety to Glu133, thus a two-dimensional umbrella sampling calculation was carried out to calculate the reaction free energy along the Cα—Cβ bond length and the proton transfer coordinate. According to the free energy map and the key structures in Figure 2.3, the highest point on the barrier mainly involves C—C bond dissociation and the proton is transferred afterward. After the reaction, the intermediates DHG and 4-OB$^\bullet$ are formed, and the radical on the latter is stabilized by the aromatic

system.

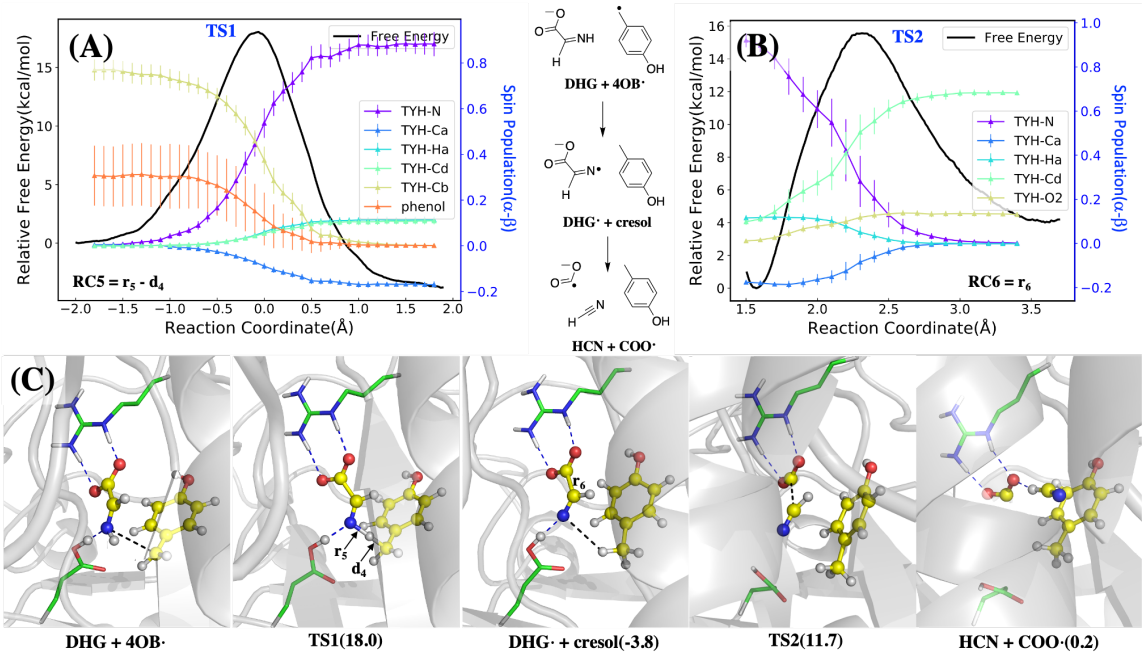


Figure 2.4: The calculated mechanism of DHG radical formation and decomposition, starting with H$^\bullet$ abstraction from the DHG nitrogen, followed by decomposition of DHG radical to COO$^{\bullet-}$ and HCN. Top: Free energy profile (A) and spin populations profiles (B) along the umbrella sampling reaction coordinates. The reaction coordinate of the DHG radical formation (top left) is defined as $r_5 - d_4$ where distances are indicated in the bottom left panel, and for the DHG radical decomposition it is defined as r_6 shown in the bottom panel. The energy curves are both colored in black with the y-axis on the left, while the spin density profiles shown for selected atoms in color with the y-axis are shown on the right. Bottom (C): Key structures during these two reactions with hydrogen bonds are shown in blue dotted lines and the distances of the reaction coordinates are shown in black dotted lines. As the structures shown in the bottom, the radical transfers from the 4-OB$^\bullet$ to the nitrogen in the DHG, which helps the decomposition of the DHG into COO$^\bullet$ and HCN. The barriers of these two steps are about 18 and 15.5 kcal/mol respectively. The radical transfer is exothermic while the DHG decomposition is endothermic and makes these two reactions overall barely absorb energy.

2.3.3 DHG Decomposition and CN and COOH Generation.

Isotope labeling studies have confirmed that the CO and CN ligands originate from the atoms of tyrosine that belong to the DHG homolysis product. However, our cal-

culations show that directly breaking the C-C bond in DHG to form HCN and CO_2 is not feasible at room temperature ($E_a > 35$ kcal/mol). In addition, CO_2 would then need to be reduced to CO, and the highly negative reduction potential of CO_2 makes this process even less feasible. Herein, we propose a "radical relay" mechanism where the 4-hydroxybenzyl radical (4-OB$^\bullet$) produced by tyrosine lysis now abstracts an H atom from the DHG nitrogen (Figure 2.4 left). We calculate that the radical transfer process from the 4-OB$^\bullet$ to the DHG is exothermic with ΔG = –4 kcal/mol and $\Delta G^\ddagger$ = 18 kcal/mol, and more favorable than an alternative pathway involving H-atom abstraction from carbon (Figure S5). Next, the homolytic cleavage of the C—C bond in this newly formed DHG radical occurs with ΔG = +4 kcal/mol and $\Delta G^\ddagger$ = 15 kcal/mol (Figure 2.4 right), making the sum of these two DHG-centered radical reaction steps thermodynamically neutral. The products of this new fission are HCN and another radical, $COO^{\bullet-}$. The well-studied radical anion of carbon dioxide has most of the unpaired spin localized on carbon.[43] Importantly, this proposed mechanism provides a crucial role for the 4-OB$^\bullet$; rather than being simply quenched to form p-cresol, it plays an active role in fragmenting the co-formed DHG along the proper reaction pathway to produce CO and CN^- rather than ammonia and glyoxylate.

The HCN and $COO^{\bullet-}$ species are structurally similar but not identical to the CN^- and CO that ultimately binds to the dangler iron located near the auxiliary cluster. Reduction of $COO^{\bullet-}$ to CO requires a hydrogen atom donor, which we could not find in the vicinity of the rSAM Fe-S cluster; attempts to transfer an H atom from cresol or nearby ionizable residues all resulted in activation energies of 30 kcal/mol or higher (Figure S7). Although it has been reported that $COO^{\bullet-}$ can react with tryptophan side chains[42] and disulfide bonds[44], the TIM barrel does not contain any of these structures.[14] Due to the lack of other plausible reaction pathways for $COO^{\bullet-}$, we propose that these species diffuse through the TIM barrel to the dangler

iron of the auxiliary cluster. Changes in the protonation state of $COO^{\bullet -}$ and HCN to $COOH^{\bullet}$ and CN^{-} are also thermodynamically accessible, as HCN is weakly acidic, and $COO^{\bullet -}$ is able to accept a proton from Glh133 (which was itself protonated during the prior Tyr decomposition step) with a slightly positive ΔE of 5.0 kcal/mol. Following the diffusion and changes in the protonation state, the remaining reaction steps are proposed to take place at the auxiliary cluster, summarized in Figure 2.5.

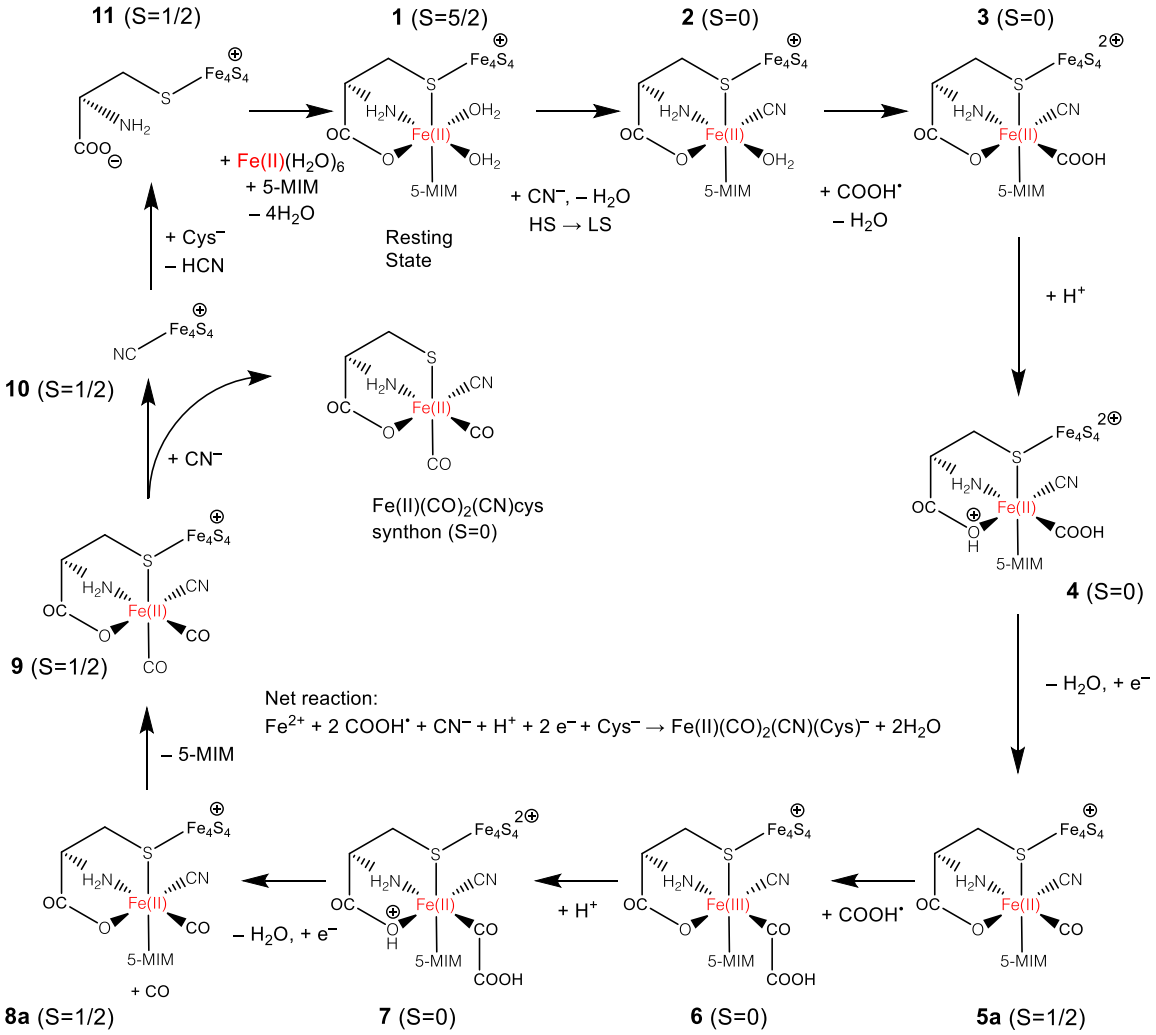


Figure 2.5: Catalytic cycle for the formation of Fe(II) synthon at the auxiliary cluster. The total spin is indicated for each species; additional properties are listed in Table S1. The resting state of the catalyst is shown at the top (**1**). The catalytic cycle involves three ligand substitutions in the coordination sphere of the dangler Fe and the reduction of two $COOH^{\bullet}$ species to CO. The 5-methylimidazole (5-MIm) species models the histidine residue in the protein.

2.3.4 Spin crossover and first CN^- substitution.

The assembly of the synthon at the dangler Fe is proposed to begin with the coordination of cyanide by displacement of a labile aqua ligand on the ferrous center (**1** + CN^- → **2** + H_2O in Figure 2.5). Prior to the reaction, the Fe_4S_4 cluster is in the reduced

state with a total charge of +1 and 1 unpaired electron (α), the dangler Fe(II) is high-spin, having 4 unpaired electrons (α) and the $COOH^{\bullet}$ has one unpaired electron (β). The coordination of CN^- stabilizes the low spin electronic state, a ubiquitous feature of octahedral ferrous cyanides.[45,46] To support the result that the coordination of a single CN^- ligand is sufficient to stabilize the low-spin state, we carried out multireference density matrix renormalization group (DMRG) calculations performed on Fe(cys)(5-MIm) in the gas phase, with the results summarized in Table S2 and Figure S8. Consistent with our DFT results, these calculations predict the ground state is high-spin with two H_2O ligands, and changes to low spin when one of the ligands is replaced by CO or CN^-.

Prior to any ligand substitution, the structure of **1** was optimized in both high-spin and low-spin electronic states as well as the minimum energy crossing point (MECP) between the two. All three optimized structures are highly similar with the largest differences coming from the Fe—S bond lengths, which are 0.16 Å shorter in the LS state. The MECP is energetically slightly uphill from the HS minimum by $\Delta E = 8.2$ kcal/mol, and the LS minimum is only 0.4 kcal/mol lower than the MECP. With the addition of free energy corrections, we computed ΔG(LS-HS) = 11.9 kcal/mol for spin crossover prior to ligand substitution. Starting from the LS energy minimum of **1**, the substitution reaction of the H_2O ligand by CN^- is found to proceed with a modest energy barrier, involving a number of structural intermediates as described in Figure S9. The overall energy parameters of the **1**(HS) $\rightarrow$ **2**(LS) reaction are given as $\Delta E = -7.2; E_a = 17.2; \Delta G = -3.9, \Delta G^{\ddagger} = 20.1$ (all values in kcal/mol).

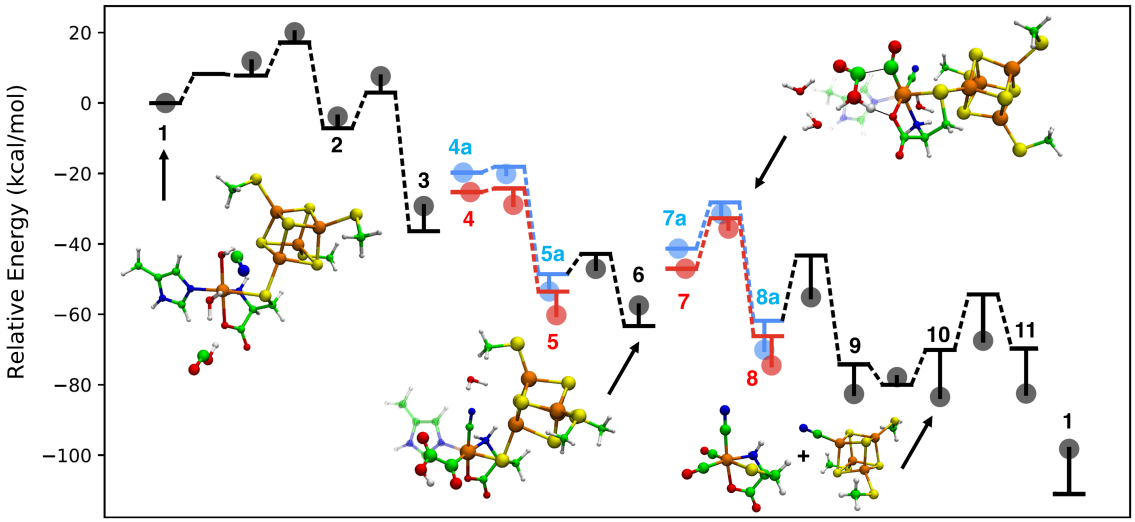


Figure 2.6: Energy diagram of the catalytic cycle expressed in Fig. 2.5 for formation of Fe(II) synthon at auxiliary cluster. Bolded numbers correspond to labeled species in Figure 2.5. Horizontal lines connected by dotted lines represent relative electronic energies that are connected via a minimum energy path. Circles represent relative Gibbs free energies computed using the harmonic oscillator/rigid rotor approximation. In the blue paths, proton-coupled reduction occurs first followed by decomposition. In the red paths, protonation is followed by deprotonation, then followed by reduction.

2.3.5 First COOH• substitution, decomposition, and reduction.

Starting from the low-spin Fe(II)-CN complex **2**, the substitution reaction of the second H_2O ligand by COOH• (**2** + COOH• → **3** + H_2O) was also found to be easily accessible and exothermic, with the energy parameters computed as ΔE = –29.1; E_a = 10.2; $\Delta G = -25.4$; $\Delta G^{\ddagger} = 11.4$ (all values in kcal/mol). The spin densities on COOH and Fe, respectively, change from (–0.92, –0.02) → (0.04, –1.09) during this reaction step, clearly indicating that the spin density on COOH• moves to Fe, and resulting in an electronic state that we interpret as (+, III). We also found that a second ET step from the auxiliary cluster to the dangler Fe is energetically favorable with $\Delta E = -8.9$ kcal/mol, resulting in a ground state of (2+, II). Based on this finding, we think one possible role for the auxiliary cluster is to act as an electron

source for the addition of $COOH^{\bullet}$, which would otherwise oxidize the dangler Fe.

In order for the COOH ligand to decompose into CO + H_2O, an electron, and proton must be transferred to the OH group. The electrochemical reduction of M-COOH to CO is well-known in the broad literature on CO_2 reduction electrocatalysis.[47–52] We investigated two possible pathways of this reaction step, summarized as $\mathbf{3}+H^+ + e^- \rightarrow \mathbf{5a}+H_2O$ overall. In the first possible pathway, protonation occurs first, followed by ligand decomposition, then reduction. We modeled this step by adding a proton to one of the displaced H_2O molecules and placing both molecules close to the OH group in a hydrogen bonding conformation. After optimization of the minimum energy path, the reactant structure was protonated at the cysteine carboxyl oxygen coordinating to the dangler Fe(II). The free energy of protonation is found to be positive, as $\Delta G(\mathbf{3} + H^+ \rightarrow \mathbf{4}) = 12.9$ kcal/mol. The proton is transferred via a H_2O relay to the OH group on COOH with a very small energy barrier of < 1.0 kcal/mol, and the resulting H_2O dissociates leaving a CO coordinated to Fe. The free energy diagram of the steps $\mathbf{3} \rightarrow \mathbf{4} \rightarrow \mathbf{5}$ are shown as the first red pathway in Figure 2.6, and the free energy parameters are calculated as $\Delta G = -22.1; \Delta G^{\ddagger} = 12.9$ kcal/-mol relative to **3** (electronic energy differences are not reported in steps involving electron and proton transfer). Finally, an electron is added to the system to reduce the cluster from (2+) to (+), represented as $\mathbf{5} \rightarrow \mathbf{5a}$ in Figure 2.6, and the redox potential is computed as -0.95 V, which corresponds to $\Delta E = 6.7$ kcal/mol when using dithionite as the reducing agent, (-0.66 V), or ≈ 12.7 kcal/mol using ferredoxin ($\approx -0.40 \pm 0.03$V) under physiological conditions.[53] The g-tensor eigenvalues of **5a** are computed as [2.030, 1.916, 1.862] and found to be in reasonable agreement with the experimental measurement of g =[2.058, 1.922, 1.882] (RMS error = 0.020).[18] The hyperfine couplings are computed as A(dangler Fe)=[-1.18, -3.04, -1.58] MHz and A(cysteine Cβ)=[-6.32, -0.80, 0.82] MHz. While these values differ significantly

from the experimental measurements of A(dangler Fe)=[0.45, 0.30, 0.50] MHz and A(cysteine Cβ)=[1.00, 0.20, 1.00] MHz, they are consistent in their order of magnitude, indicating only small amounts of spin density on the dangler Fe and cysteine Cβ centers.

In the second possible pathway shown as the first blue pathway of Figure 2.6, reduction of the auxiliary cluster and protonation of the cysteine carbonyl oxygen occur simultaneously and can be summarized as $\mathbf{3} + \mathrm{H}^+ + \mathrm{e}^- \rightarrow \mathbf{4a}$. The redox potential is computed as -1.46 V, which corresponds to ΔG = 18.4 kcal/mol if the reducing agent is dithionite. The proton transfer to the OH group of COOH and the resulting dissociation of H_2O ($\mathbf{4a} \rightarrow \mathbf{5a}$) proceeds in a similar fashion to the first pathway, but due to the higher relative free energy of the reduced state, the overall activation free energy is found to be slightly higher ($\Delta G^{\ddagger}$ = 18.4 kcal/mol relative to **3**) compared to the first pathway.

2.3.6 Reduction and decomposition of a second COOH•.

In order to produce the Fe(II) synthon, the 5-methylimidazole ligand must be replaced by CO sourced from a second COOH• ligand produced as a result of a second tyrosine lysis, which we presume follows the same mechanism as the first tyrosine lysis described above. The main question for the second CO formation is how the mechanism is affected by the altered state of the auxiliary cluster, given that the first CO and CN are already bound and the Fe(II) shifted from high spin to low spin. Our attempt to compute this substitution step led to an unexpected result: a transition state for 5-MIm substitution by COOH• was found, but the energy minimization started from the TS led to a reactant structure where COOH• forms a covalent C-C bond with the first CO coordinated to the dangler Fe (**6**, lower middle structure in Figure. 2.6). We found that this new intermediate, a glyoxylyl (OC-COOH) ligand to Fe, forms

easily from **5** in a pathway where $COOH^{\bullet}$ directly forms a C-C bond with the CO ligand and the energy parameters are $\Delta E = -14.7, E_a = 5.8, \Delta G = -4.0, \Delta G^{\ddagger} = 6.4$ kcal/mol; this unusual C—C coupling is reminiscent of CO_2 electrolysis experiments at inert electrodes where oxalate is one of the products formed.[54] The glyoxylyl intermediate is characterized by a (+, III) oxidation state because the spin density is transferred to the dangler Fe, analogous to the coordination of the first $COOH^{\bullet}$. The alternate (2+, II) electronic state, in which an electron is transferred from the cluster, is found to be slightly higher in energy $\Delta G = 1.5$ kcal/mol, indicating that the calculations do not significantly favor one state over the other. From this new intermediate, we investigated whether this ligand could be decomposed to CO by an additional reduction step.

We modeled the decomposition of the glyoxylyl ligand by placing a H_3O^+ cation close to the OH group and searching for a TS involving proton transfer and dissociation of the C-C bond to form H_2O and CO, similar to before. In the optimized path, we again found that the reactant structure was protonated at the cysteine carboxyl oxygen coordinating to the Fe atom (**7a** in figure 2.6), with $\Delta G = 10.4$ kcal/mol relative to **6**. The ground state of the protonated form has nearly zero spin on the dangler Fe indicating a (2+, II) electronic state. Proton transfer results in the decomposition of the glyoxylyl ligand to yield $CO + H_2O +$ **8**, shown as the second red pathway in Figure 2.6. The free energy barrier of the overall reaction **6** $\rightarrow$ **8** including the protonation step the highest in the overall cycle ($\Delta G = -17.0, \Delta G^{\ddagger} = 21.8$ kcal/mol), but are somewhat smaller compared to the highest barrier of 26.3 kcal/mol found for the reactions at the canonical cluster. The Fe_4S_4 cluster is reduced back to the (+) electronic state (i.e. **8** $\rightarrow$ **8a** with a potential of -0.85 V or $\Delta G = 4.4$ kcal/mol when dithionite is used as the reducing agent, corresponding to $\Delta G = -12.6$ kcal/mol relative to **6**.

A second possible pathway was considered in which simultaneous protonation and reduction of **6** occurs first, followed by PT and decomposition of the COOH group. The proton-coupled redox potential was computed as -1.36 V at pH 7, corresponding to an overpotential of 0.70 V, or $\Delta G = 16.1$ kcal/mol. The decomposition reaction then proceeds in an analogous fashion, in which the proton is transferred to the glyoxylyl OH group, followed by ligand decomposition to yield $CO + H_2O$ + **8a**. In the reduced electronic state, the free energy parameters of the reaction $\mathbf{6} + H^+ + e^- \rightarrow$ **7a** $\rightarrow$ **8a** were computed as $\Delta G = -12.6, \Delta G^{\ddagger} = 26.0$ kcal/mol, shown as the second blue pathway in Figure 2.6.

Because the overall free energy barrier is 4.2 kcal/mol higher when proceeding from the reduced state **7a**, we think the most likely sequence of reaction steps is protonation of **6** coupled to cluster $\rightarrow$ dangler electron transfer, decomposition, then reduction; however, the alternate ordering of proton-coupled reduction followed by decomposition is a close alternative possibility. The product includes a free CO ligand which is able to displace the 5-MIm ligand, leading to **9**, with energy parameters $\Delta E = -12.3, E_a = 18.6, \Delta G = -12.5, \Delta G^{\ddagger} = 14.9$ kcal/mol.

2.3.7 Completion of the catalytic cycle.

At this point in the cycle, the coordination sphere of the Fe(II) synthon is completed, and this Fe(II)$(CO)_2$(CN)cysteine complex needs to be released from HydG so it can serve as the substrate for the other rSAM enzyme, HydE.[55] We placed the second CN^- species close to the auxiliary cluster (about 6Å away from the Fe_4S_4 cluster) and found that it spontaneously coordinates to the Fe_4S_4 iron closest to the dangler Fe without a barrier. This coordination mode is highly similar to how the first CN^- species coordinates to **1**, also seen in titration experiments.[18] We note that although we computed the CN^- association step after the completion of the coordination sphere

of the Fe(II) synthon, it is possible that this coordination may take place at an earlier point in the catalytic cycle, perhaps as early as the delivery of the second $COOH^{\bullet}$ equivalent (**6**).

The dissociation of the synthon is energetically slightly uphill by $\Delta E = 9.9$ kcal/-mol, and the overall reaction of replacing the synthon by CN^- has energy parameters given by $\Delta E, \Delta G(\mathbf{9} \rightarrow \mathbf{10}) = 4.11, -0.85$ kcal/mol. The g-tensor eigenvalues of **10** are computed as [2.023, 1.961, 1.931], which can be compared with g =[2.09, 1.94, 1.93] from the experiment (RMS error = 0.041).[16] The hyperfine tensor on the CN carbon is computed as A =[5.11, 27.39, 27.54] MHz, which is much larger than the experimental measurement of A =[-5.0, -4.0, 0.9] MHz; the overestimation is likely due to the tendency to over-delocalize spin density in DFT.[56] To close the cycle, the final substitution of this CN^- by CH_3SH, the side chain analogue of cysteine, is found to be nearly isoenergetic with a moderate barrier (ΔE = +0.4, E_a = 15.7, ΔG = +1.0, $\Delta G^{\ddagger}$ = 16.0, all values in kcal/mol); a proton is transferred from the thiol group to CN^- during this step. We computed g =[2.016, 1.967, 1.938], which when compared to the experimental measurement of g =[2.06, 1.90, 1.87], has a relatively large RMS error of 0.061 compared to the other species; this could be due to the use of the Cys side chain analogue in place of the complete amino acid in our calculations. The hyperfine tensor on the CH_3SH carbon is computed as A =[5.08, 9.05, 9.64] MHz, which is significantly larger than the experimental A =[0.83, 0.83, 1.09] MHz, a consistent trend with all of the EPR property calculations. The ultimate fate of the HCN species is still experimentally undetermined. Subsequent coordination of Fe(II) to cysteine and 5-MIm completes the catalytic cycle and returns the system to its resting state ($\Delta E = -40.9, \Delta G = -14.8$, all values in kcal/mol).

This computed mechanism produces an isomer of the synthon where the CN ligand is opposite to the cysteinyl carboxylate oxygen, whereas a recent experimental study

on the crystal structure of HydE with the synthon as its substrate[57] assigned the CN ligand as opposite to sulfur based on the proximity of hydrogen bond donating residues. We computed the relative electronic energies of the synthon isomers and found them to be very close in energy (ΔE (opposite S) = 0.00; ΔE (opp. N) = 0.65; ΔE (opp. O) = 1.65, values in kcal/mol). Therefore, we think it is thermodynamically possible for the synthon to undergo isomerization prior to binding to HydE though the isomerization mechanism has yet to be determined.

2.4 Conclusions

In this theoretical study, we have been able to elucidate additional mechanistic details concerning the HydG catalytic cycle, building on previous experimental results. We propose that after the initial HydG tyrosine lysis produces a 4-hydroxybenzyl radical (4-OB$^\bullet$) and dehydroglycine (DHG), the nascent 4-OB$^\bullet$ radical in turn abstracts an H-atom from the nitrogen of DHG, resulting in a DHG radical. This DHG radical in turn undergoes a spontaneous C-C bond cleavage to form HCN and a new COOH$^{\bullet-}$ anion radical. In the overall formation of the HydG Fe(II)$(CO)_2$(CN)cysteine product this radical cascade occurs twice. However, the specific reactions at the five-Fe auxiliary Fe-S cluster differs between the first and second subcycles. Following the first HydG induced tyrosine lysis and subsequent radical cascade, the fifth "dangler Fe" of the auxiliary cluster is high spin Fe(II) with two aqua ligands. The CN^- and COOH$^{\bullet-}$ anion radical can bind in these two positions, substituting for the aqua ligands, and causing a spin crossover of the dangler Fe(II) from high spin to low spin. The COOH is further protonated and decomposes to the CO ligand and an H_2O. The auxiliary cluster reaction must differ on the second subcycle since the dangler Fe(II) is now low spin with CO and CN ligands. We model the second cycle with the second COOH$^\bullet$

forming a covalent C-C bond at the existing CO Fe ligand, forming a transient glyoxylyl ligand that decomposes upon further reduction. The second CO resulting from this decomposition displaces the histidine ligand to the dangler Fe, while the second CN^- attacks the Fe_4S_4 cluster, releasing the entire HydG Fe(II)$(CO)_2$(CN)cysteine product.

This Fe(II) "synthon" is transferred to HydE where it serves as a substrate for another set of reactions that lead to a highly reactive Fe(I)$(CO)_2$CNS species poised to form the Fe_2S_2 core of the $[2Fe]_H$ subcluster.[55,58] The HydG cycle is completed by another cysteine replacing the CN^- bound to the Fe_4S_4 cluster, with this new cysteine and the HydG histidine binding a new Fe(II) to regenerate the HydG resting state. According to the energy barriers of all the reactions above, the rate-limiting step of the HydG catalytic process is either the SAM decomposition, whose energy barrier is 26.3 kcal/mol, or the redox-coupled decomposition of the second $COOH^\bullet$ equivalent with a free energy barrier of 21.8 kcal/mol. Both numbers are comparable to experimental kinetics studies (23 kcal/mol). All of the other computed barrier heights for all the steps in this complex reaction are found to be thermodynamically accessible and consistent with the observed (rather slow) timescale of the HydG reaction chemistry, and thus we consider this a plausible overall model for the HydG chemistry, consistent with experimental observables. Looking forward, we think that theoretical studies can also provide insights into the catalytic mechanism of HydE and HydF to complement the experimental data and furnish a complete mechanism for the biosynthesis of the unique $[2Fe]_H$ subcluster of the [FeFe] hydrogenase active site.

References

[1]W. Lubitz, H. Ogata, O. Rüdiger, and E. Reijerse, "Hydrogenases", Chemical reviews **114**, 4081–4148 (2014).

[2]J. C. Fontecilla-Camps, A. Volbeda, C. Cavazza, and Y. Nicolet, "Structure/function relationships of [nife]-and [fefe]-hydrogenases", Chemical reviews **107**, 4273–4303 (2007).

[3]K. Pandey, S. T. A. Islam, T. Happe, and F. A. Armstrong, "Frequency and potential dependence of reversible electrocatalytic hydrogen interconversion by fefe -hydrogenases", Proceedings of the National Academy of Sciences of the United States of America **114**, 3843–3848 (2017).

[4]M. C. Posewitz, P. W. King, S. L. Smolinski, L. P. Zhang, M. Seibert, and M. L. Ghirardi, "Discovery of two novel radical s-adenosylmethionine proteins required for the assembly of an active fe hydrogenase", Journal of Biological Chemistry **279**, 25711–25720 (2004).

[5]J. M. Kuchenreuther, R. D. Britt, and J. R. Swartz, "New insights into fefe hydrogenase activation and maturase function", PLOS ONE **9**, 1–9 (2012).

[6]R. C. Driesener, M. R. Challand, S. E. McGlynn, E. M. Shepard, E. S. Boyd, J. B. Broderick, J. W. Peters, and P. L. Roach, "Fefe -hydrogenase cyanide ligands derived from s-adenosylmethionine-dependent cleavage of tyrosine", Angewandte Chemie-International Edition **49**, 1687–1690 (2010).

[7]E. M. Shepard, S. E. McGlynn, A. L. Bueling, C. S. Grady-Smith, S. J. George, M. A. Winslow, S. P. Cramer, J. W. Peters, and J. B. Broderick, "Synthesis of the 2fe subcluster of the fefe -hydrogenase h cluster on the hydf scaffold", Proceedings

of the National Academy of Sciences of the United States of America **107**, 10448–10453 (2010).

[8]J. M. Kuchenreuther, W. K. Myers, D. L. M. Suess, T. A. Stich, V. Pelmenschikov, S. A. Shiigi, S. P. Cramer, J. R. Swartz, R. D. Britt, and S. J. George, "The hydg enzyme generates an fe(co)(2)(cn) synthon in assembly of the fefe hydrogenase h-cluster", Science **343**, 424–427 (2014).

[9]J. M. Kuchenreuther, W. K. Myers, T. A. Stich, S. J. George, Y. NejatyJahromy, J. R. Swartz, and R. D. Britt, "A radical intermediate in tyrosine scission to the co and cn- ligands of fefe hydrogenase", Science **342**, 472–475 (2013).

[10]L. Tao, S. A. Pattenaude, S. Joshi, T. P. Begley, T. B. Rauchfuss, and R. D. Britt, "Radical sam enzyme hyde generates adenosylated fe(i) intermediates en route to the [fefe]-hydrogenase catalytic h-cluster", Journal of the American Chemical Society **142**, 10841–10848 (2020).

[11]G. D. Rao, L. Z. Tao, and R. D. Britt, "Serine is the molecular source of the nh(ch2)(2) bridgehead moiety of the in vitro assembled fefe hydrogenase h-cluster", Chemical Science **11**, 1241–1247 (2020).

[12]G. Berggren, A. Adamska, C. Lambertz, T. R. Simmons, J. Esselborn, M. Atta, S. Gambarelli, J. M. Mouesca, E. Reijerse, W. Lubitz, T. Happe, V. Artero, and M. Fontecave, "Biomimetic assembly and activation of fefe -hydrogenases", Nature **499**, 66–70 (2013).

[13]R. D. Britt, G. Rao, and L. Tao, "Biosynthesis of the catalytic h-cluster of [fefe] hydrogenase: the roles of the fe–s maturase proteins hyde, hydf, and hydg", Chemical Science (2020).

[14]P. Dinis, D. L. Suess, S. J. Fox, J. E. Harmer, R. C. Driesener, L. De La Paz, J. R. Swartz, J. W. Essex, R. D. Britt, and P. L. Roach, "X-ray crystallographic and epr spectroscopic analysis of hydg, a maturase in [fefe]-hydrogenase h-cluster assembly", Proceedings of the National Academy of Sciences **112**, 1362–1367 (2015).

[15]J. M. Kuchenreuther, W. K. Myers, T. A. Stich, S. J. George, Y. NejatyJahromy, J. R. Swartz, and R. D. Britt, "A radical intermediate in tyrosine scission to the co and cn- ligands of fefe hydrogenase", Science **342**, 472–475 (2013).

[16]D. L. Suess, I. Bürstel, L. De La Paz, J. M. Kuchenreuther, C. C. Pham, S. P. Cramer, J. R. Swartz, and R. D. Britt, "Cysteine as a ligand platform in the biosynthesis of the fefe hydrogenase h cluster", Proceedings of the National Academy of Sciences **112**, 11455–11460 (2015).

[17]D. L. Suess, C. C. Pham, I. Bürstel, J. R. Swartz, S. P. Cramer, and R. D. Britt, "The radical sam enzyme hydg requires cysteine and a dangler iron for generating an organometallic precursor to the [fefe]-hydrogenase h-cluster", Journal of the American Chemical Society **138**, 1146–1149 (2016).

[18]G. Rao, L. Tao, D. L. Suess, and R. D. Britt, "A [4fe–4s]-fe (co)(cn)-l-cysteine intermediate is the first organometallic precursor in [fefe] hydrogenase h-cluster bioassembly", Nature chemistry **10**, 555–560 (2018).

[19]A. Pagnier, L. Martin, L. Zeppieri, Y. Nicolet, and J. C. Fontecilla-Camps, "Co and cn- syntheses by fefe -hydrogenase maturase hydg are catalytically differentiated events", Proceedings of the National Academy of Sciences of the United States of America **113**, 104–109 (2016).

[20]Y. Nicolet, L. Zeppieri, P. Amara, and J. C. Fontecilla-Camps, "Crystal structure of tryptophan lyase (nosl): evidence for radical formation at the amino group of tryptophan", Angewandte Chemie International Edition **53**, 11840–11844 (2014).

[21]Y. Shao, Z. Gan, E. Epifanovsky, A. T. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kuś, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. W. III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. D. Jr., H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. Hanson-Heine, P. H. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. D. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. Thom, T. Tsuchimochi, V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. Chan, D. M. Chipman, C. J. Cramer, W. A. G. III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. S. III, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. V. Voorhis, J. M. Herbert, A. I. Krylov, P. M.

Gill, and M. Head-Gordon, "Advances in molecular quantum chemistry contained in the q-chem 4 program package", Molecular Physics **113**, 184–215 (2015).

[22]D. Case, K. Belfon, I. Ben-Shalom, S. Brozell, D. Cerutti, I. T.E. Cheatham, V. Cruzeiro, T. Darden, R. Duke, G. Giambasu, M. Gilson, H. Gohlke, A. Goetz, R. Harris, S. Izadi, S. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R. Walker, J. Wang, L. Wilson, R. Wolf, X. Wu, Y. Xiong, Y. Xue, D. York, and P. Kollman, Amber 2020, ucsf.

[23]Y. Zhou, S. Wang, Y. Li, and Y. Zhang, "Born–oppenheimer ab initio qm/mm molecular dynamics simulations of enzyme reactions", in Methods in enzymology, Vol. 577 (Elsevier, 2016), pp. 105–118.

[24]S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, S. I. L. Kokkila-Schumacher, N. Luehr, J. W. Snyder, C. Song, A. V. Titov, I. S. Ufimtsev, and T. J. Martínez, "Terachem: accelerating electronic structure and ab initio molecular dynamics with graphical processing units", The Journal of Chemical Physics **152**, 224110 (2020).

[25]S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, C. M. Isborn, S. I. L. Kokkila-Schumacher, X. Li, F. Liu, N. Luehr, J. W. Snyder Jr., C. Song, A. V. Titov, I. S. Ufimtsev, L.-P. Wang, and T. J. Martínez, "Terachem: a graphical processing unit-accelerated electronic structure package for large-scale ab initio molecular dynamics", WIREs Computational Molecular Science **n/a**, e1494.

[26]L.-P. Wang and C. Song, "Geometry optimization made simple with translation and rotation coordinates", The Journal of Chemical Physics **144**, 214108 (2016).

[27]F. Neese, "Prediction of electron paramagnetic resonance g values using coupled perturbed hartree–fock and kohn–sham theory", The Journal of Chemical Physics **115**, 11080–11096 (2001).

[28]F. Neese, F. Wennmohs, U. Becker, and C. Riplinger, "The orca quantum chemistry program package", The Journal of Chemical Physics **152**, 224108 (2020).

[29]O. Salomon, M. Reiher, and B. A. Hess, "Assertion and validation of the performance of the b3lyp* functional for the first transition metal row and the g2 test set", The Journal of Chemical Physics **117**, 4729–4737 (2002).

[30]R. K. Szilagyi and M. A. Winslow, "On the accuracy of density functional theory for iron—sulfur clusters", Journal of Computational Chemistry **27**, 1385–1397 (2006).

[31]F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: design and assessment of accuracy", Phys. Chem. Chem. Phys. **7**, 3297–3305 (2005).

[32]L. E. Roy, P. J. Hay, and R. L. Martin, "Revised basis sets for the lanl effective core potentials", Journal of Chemical Theory and Computation **4**, 1029–1031 (2008).

[33]J. Zheng, X. Xu, and D. G. Truhlar, "Minimally augmented karlsruhe basis sets", Theoretical Chemistry Accounts **128**, 295–305 (2011).

[34]H. Jang, Y. Qiu, M. E. Hutchings, M. Nguyen, L. A. Berben, and L.-P. Wang, "Quantum chemical studies of redox properties and conformational changes of a four-center iron co2 reduction electrocatalyst", Chem. Sci. **9**, 2645–2654 (2018).

[35]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field", Journal of Computational Chemistry **25**, 1157–1174 (2004).

[36]C. H. Chang and K. Kim, "Density functional theory calculation of bonding and charge parameters for molecular dynamics studies on [fefe] hydrogenases", Journal of chemical theory and computation **5**, 1137–1145 (2009).

[37]L.-P. Wang, T. J. Martinez, and V. S. Pande, "Building force fields: an automatic, systematic, and reproducible approach", The Journal of Physical Chemistry Letters **5**, 1885–1891 (2014).

[38]L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martınez, and V. S. Pande, "Building a more predictive protein force field: a systematic and reproducible route to amber-fb15", The Journal of Physical Chemistry B **121**, 4023–4039 (2017).

[39]R. I. Sayler, T. A. Stich, S. Joshi, N. Cooper, J. T. Shaw, T. P. Begley, D. J. Tantillo, and R. D. Britt, "Trapping and electron paramagnetic resonance characterization of the 5-dado radical in a radical s-adenosyl methionine enzyme reaction with a non-native substrate", ACS central science **5**, 1777–1785 (2019).

[40]R. I. Sayler, T. A. Stich, S. Joshi, N. Cooper, J. T. Shaw, T. P. Begley, D. J. Tantillo, and R. D. Britt, "Trapping and electron paramagnetic resonance characterization of the 5-dado radical in a radical s-adenosyl methionine enzyme reaction with a non-native substrate", ACS Central Science **5**, 1777–1785 (2019).

[41]R. C. Driesener, B. R. Duffus, E. M. Shepard, I. R. Bruzas, K. S. Duschene, N. J.-R. Coleman, A. P. Marrison, E. Salvadori, C. W. Kay, J. W. Peters, et al., "Biochemical and kinetic characterization of radical s-adenosyl-l-methionine enzyme hydg", Biochemistry **52**, 8696–8707 (2013).

[42]P. Amara, J.-M. Mouesca, M. Bella, L. Martin, C. Saragaglia, S. Gambarelli, and Y. Nicolet, "Radical s-adenosyl-l-methionine tryptophan lyase (nosl): how the protein

controls the carboxyl radical• co2–migration", Journal of the American Chemical Society **140**, 16661–16668 (2018).

[43]F. A. Villamena, E. J. Locigno, A. Rockenbauer, C. M. Hadad, and J. L. Zweier, "Theoretical and experimental studies of the spin trapping of inorganic radicals by 5, 5-dimethyl-1-pyrroline n-oxide (dmpo). 1. carbon dioxide radical anion", The Journal of Physical Chemistry A **110**, 13253–13258 (2006).

[44]V. Favaudon, H. Tourbez, C. Houee-Levin, and J. M. Lhoste, "Carboxyl radical induced cleavage of disulfide bonds in proteins. a .gamma.-ray and pulse radiolysis mechanistic investigation", Biochemistry **29**, 10978–10989 (1990).

[45]R. L. Lord and M.-H. Baik, "Why does cyanide pretend to be a weak field ligand in [cr(cn)5]3-?", Inorganic Chemistry **47**, 4413–4420 (2008).

[46]M. Nakamura, "Is cyanide really a strong-field ligand?", Angewandte Chemie International Edition **48**, 2638–2640 (2009).

[47]E. Simón-Manso and C. P. Kubiak, "Dinuclear Nickel Complexes as Catalysts for Electrochemical Reduction of Carbon Dioxide", Organometallics **24**, Publisher: American Chemical Society, 96–102 (2005).

[48]A. M. Appel, J. E. Bercaw, A. B. Bocarsly, H. Dobbek, D. L. DuBois, M. Dupuis, J. G. Ferry, E. Fujita, R. Hille, P. J. A. Kenis, C. A. Kerfeld, R. H. Morris, C. H. F. Peden, A. R. Portis, S. W. Ragsdale, T. B. Rauchfuss, J. N. H. Reek, L. C. Seefeldt, R. K. Thauer, and G. L. Waldrop, "Frontiers, Opportunities, and Challenges in Biochemical and Chemical Catalysis of CO2 Fixation", Chemical Reviews **113**, Publisher: American Chemical Society, 6621–6658 (2013).

[49]C. Costentin, M. Robert, and J.-M. Savéant, "Catalysis of the electrochemical reduction of carbon dioxide", en, Chemical Society Reviews **42**, 2423–2436 (2013).

[50]J. Shen, R. Kortlever, R. Kas, Y. Y. Birdja, O. Diaz-Morales, Y. Kwon, I. Ledezma-Yanez, K. J. P. Schouten, G. Mul, and M. T. M. Koper, "Electrocatalytic reduction of carbon dioxide to carbon monoxide and methane at an immobilized cobalt protoporphyrin", en, Nature Communications **6**, 8177 (2015).

[51]S. Lin, C. S. Diercks, Y.-B. Zhang, N. Kornienko, E. M. Nichols, Y. Zhao, A. R. Paris, D. Kim, P. Yang, O. M. Yaghi, and C. J. Chang, "Covalent organic frameworks comprising cobalt porphyrins for catalytic CO2 reduction in water", Science **349**, 1208–1213 (2015).

[52]S. Gao, Z. Sun, W. Liu, X. Jiao, X. Zu, Q. Hu, Y. Sun, T. Yao, W. Zhang, S. Wei, and Y. Xie, "Atomic layer confined vacancies for atomic-level insights into carbon dioxide electroreduction", en, Nature Communications **8**, 10.1038/ncomms14503 (2017).

[53]S. J. Maiocco, A. J. Arcinas, S. J. Booker, and S. J. Elliott, "Parsing redox potentials of five ferredoxins found within thermotoga maritima", Protein Science **28**, 257–266 (2019).

[54]A. Gennaro, A. A. Isse, M.-G. Severin, E. Vianello, I. Bhugun, and J.-M. Savéant, "Mechanism of the electrochemical reduction of carbon dioxide at inert electrodes in media of low proton availability", J. Chem. Soc., Faraday Trans. **92**, 3963–3968 (1996).

[55]L. Tao, S. A. Pattenaude, S. Joshi, T. P. Begley, T. B. Rauchfuss, and R. D. Britt, "The radical sam enzyme hyde generates adenosylated fe (i) intermediates en route to the [fefe]-hydrogenase catalytic h-cluster", Journal of the American Chemical Society (2020).

[56]D. G. Artiukhin and J. Neugebauer, "Frozen-density embedding as a quasi-diabatization tool: charge-localized states for spin-density calculations", The Journal of Chemical Physics **148**, 214104 (2018).

[57]R. Rohac, L. Martin, L. Liu, D. Basu, L. Tao, R. D. Britt, T. B. Rauchfuss, and Y. Nicolet, "Crystal structure of the [fefe]-hydrogenase maturase hyde bound to complex-b", Journal of the American Chemical Society **143**, 8499–8508 (2021).

[58]R. D. Britt, G. Rao, and L. Tao, "Biosynthesis of the catalytic h-cluster of [fefe] hydrogenase: the roles of the fe–s maturase proteins hyde, hydf, and hydg", Chemical Science **11**, 10313–10323 (2020).

3 How does HydE work? A Comparison Between A Radical Mechanism and A Proton-Transfer Mechanism

3.1 Introduction

Hydrogenases are enzymes catalyzing the reversible redox conversion of H^+ into molecular H_2. This fascinating reaction also has important practical applications since H_2 is a clean and renewable fuel that can be utilized in many industries. The catalytic center of the hydrogenase contains a canonical [Fe_4S_4] subcluster bonded via a cysteine residue to an organometallic subcluster known as the "H-cluster", a diiron complex with CO and CN^- ligands and an azadithiolate ($NH(CH_2S)_2$, adt) bridge. The biosynthesis of this key structure involves three enzymes called HydG, HydE, and HydF, the first two of which are radical S-adenosyl-L-methionine (rSAM). The function and reaction mechanism of these maturases is a highly interesting biochemical question that has been under active study for nearly two decades. Some of the key questions about the mechanism include how the CO and CN^- ligands are managed, given that they are highly toxic as free species, and how the adt moiety is incorporated given its instability in the free state.

In our model of H-cluster assembly, the radical SAM enzyme HydG performs the initial reaction, lysing tyrosine and forming an [Fe(cysteine)$(CO)_2$(CN)] organometallic product starting from a resting state auxiliary [Fe_4S_4] cluster that is linked to a Fe(II)-cysteine complex via a cysteine S bridge.[1–8] The use of a synthetic [Fe(cysteine)$(CO)_2$(CN)] analog (SynB) of this proposed HydG product showed that HydG itself can then be eliminated in the enzymatic synthesis of active H-cluster.[9] Our recent computational study[7] shows that HydG decomposes the tyrosine (Tyr) residue into a diatomic lig-

and CN^- and a $COOH^\bullet$ radical via a radical relay process initialized by the $[Fe_4S_4]$ cluster and the rSAM. These two ligands are delivered to another Fe-S cluster in HydG, namely the auxiliary cluster, and are transformed into two diatomic ligands, CO and CN^- on the dangler Fe. The presence of 4-hydroxybenzyl radical, an intermediate along the computed reaction pathway, was detected in HydG mutants that had the dangler Fe knocked out.[10] With another tyrosine decomposition, the $[Fe^{II}(Cys)(CO)_2(CN)]$ synthon (syn-B) is generated and released as the final product of HydG.

EXAFS comparing the structure of the H-cluster formed with the maturation using the selenocysteine version of SynB compared to the normal cysteine version showed that the bridging S/Se of the binuclear cluster are sourced from this cysteine S/Se. Also, HYSCORE reveals no $^{13}C/^{15}N$ coupling using $^{13}C3$,^{15}N-cysteine in the maturation, although $^{13}C3$ pyruvate accumulates in the maturation media, demonstrating that the S-Cβ bond of cysteine must be cleaved in the maturation reaction. Rao et al.[11] showed that the CH_2NHCH_2 fragment of the ADT bridge is actually sourced from serine.

Tao et al. (2020)[12] showed that this S-Cβ bond cleavage of cysteine is catalyzed by the second radical SAM maturase, HydE. The rSAM generated 5′-dAdo$^\bullet$ attacks the cysteine sulfur to form a crosslinked Fe(I) adenosyl complex, giving an $S = \frac{1}{2}$ EPR signal that peaks at about 10 s in the reaction. The S-Cβ bond cleavage follows the formation of this species, resulting in a different Fe(I) EPR signal arising from a different complex, which we now believe is a 5-coordinate $Fe(I)S(CO)_2(CN)$ species (Tao et al. indicate an interaction with the ribose O, but we do not favor this assignment now). This new EPR signal peaks at about 10 min and then decays. We discussed this in light of the tendency for such 5 coordinate Fe(I) organometallic species to dimerize, which would give rise to an antiferromagnetically coupled, EPR

silent reaction product $Fe_2S_2(CO)_4(CN)_2$.

In 2021, Rohac et al. published a crystal structure showing the binding conformation of syn-B within HydE bound to the methionine residue M224.[13] By triggering the radical chemistry of HydE using a chemical reducing agent prior to crystallization, it was found that HydE converts syn-B into a new 5-coordinate $Fe(I)S(CO)_2(CN)$ intermediate. This intermediate is further hypothesized to undergo dimerization in HydE, in which residue M291 transiently binds one Fe(I) complex while a second complex is produced. Recently, Zhang et al. showed that the hypothesized dimerization product $[Fe_2(\mu\text{-}SH)_2(CN)_2(CO)_4]^{-2}$ allows the maturation of HydA with only HydF,[14] which is consistent with the hypothesis that the dimer is the product of HydE.

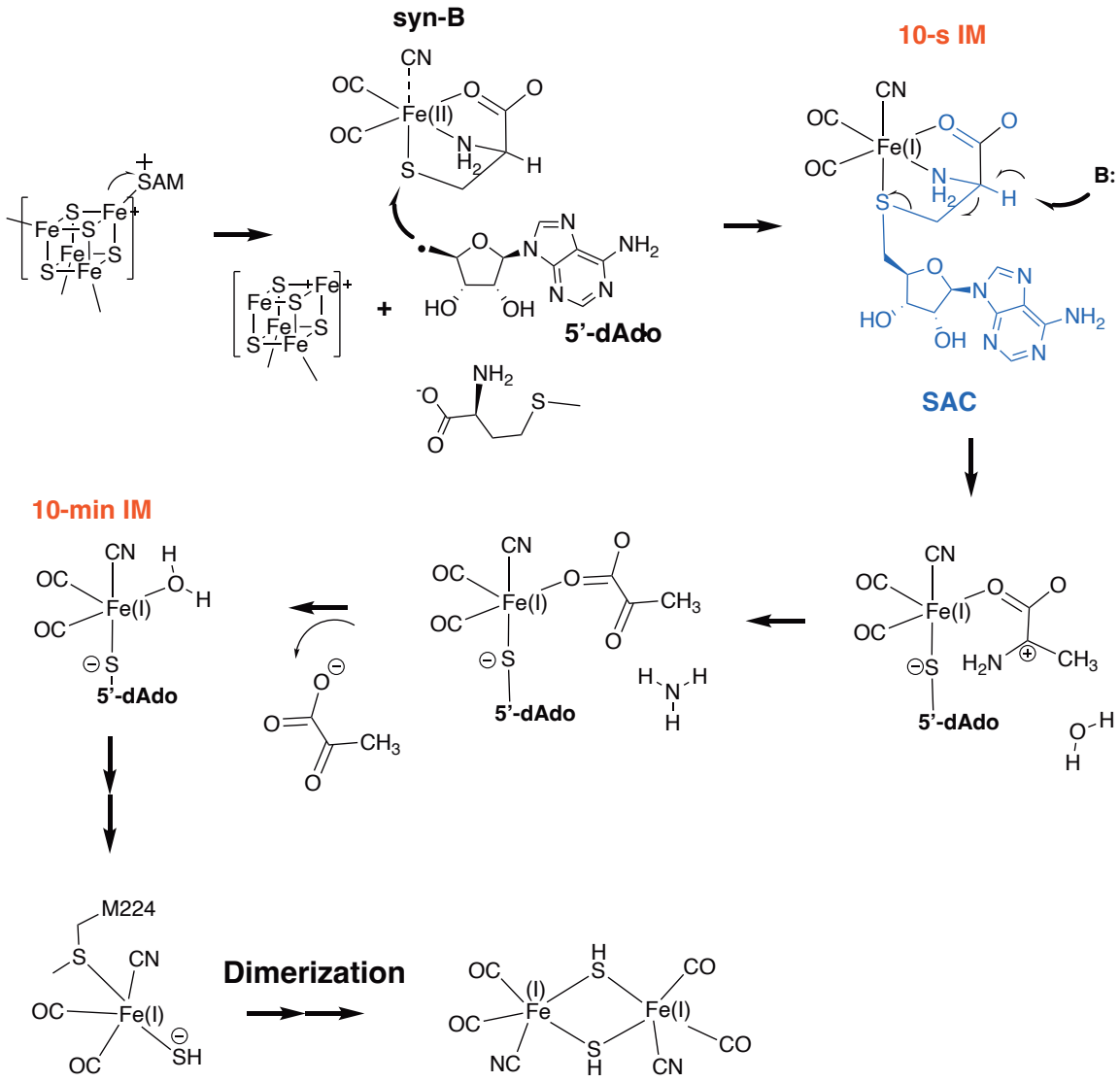


Figure 3.1: Proposed reaction scheme of HydE catalysis in the published literature. This work provides a detailed computational description of the following mechanistic steps: the conversion of the 10-s intermediate to the 10-min intermediate, the decomposition of the intermediate to form the 5-coordinate Fe(I) species, and the dimerization of two Fe(I) complexes to form a $Fe_2(SH)_2(CO)_4(CN)_2$ product.

Herein, the current catalytic mechanism of HydE is summarized in Figure 3.1. As a regular SAM enzyme, HydE initializes the reaction by decomposing the SAM and generating the 5′-dAdo$^{\bullet}$. A 10-second intermediate (10-s IM) with dangler Fe(I) has been detected in experiments after the SAM decomposition. Later, the basic residue

is purposed to deprotonate the Hα from the 10-s IM yielding dehydroalanine. After a hydrolysis reaction, the pyruvate is formed and released as a side-product, and the dangler Fe cluster becomes a 5-coordinate structure named the 10-minute intermediate (10-min IM). In addition, the dimerization of the 10-min IM is also proposed to occur in HydE. Although experiments have provided much evidence (mentioned above) about the catalytic mechanism of HydE, many of the key mechanistic steps are still hypothetical, such as the mechanism for cleaving the S-Cβ bond. Herein, we employ quantum mechanics/molecular mechanics (QM/MM) molecular dynamics (MD) simulation to study the catalytic mechanism in HydE. We first show the feasibility of the C-S bond radical addition to generate the 10-s intermediate. In addition, although one possible basic residue, deprotonated Lys 309, has been discovered around the catalytic pocket, an alternative radical mechanism is shown to be more energetically feasible than the Hα deprotonation pathway. Finally, a possible dimerization mechanism in HydE has been proposed and computationally verified.

3.2 Methods

A structure of a HydE mutant (PDB ID: 7O1O) from Rohac et al.[13] with bound syn-B and S-adenosyl-L-homocysteine (SAH) ligands was chosen as the base structure for modeling as it closely matched our system of interest. The mutation is missing a [Fe_2S_2] cluster in the C-terminal region of HydE but it is not expected to affect the catalysis. The inactive SAH was substituted by the active SAM, which is the exact substrate in HydE. The protonation states of all residues were decided according to their standard side chain pKa values and the experimental pH value (7.0). The Lys 309 was deprotonated to study its potential to be a proton acceptor. The force field parameters of the [Fe_4S_4] cluster from ref[15] are chosen to describe the behavior of

the cluster in the MD simulations. Meanwhile, the parameters of the syn-B from our previous HydG simulations[7] are adopted in this system. As for the protein and solvent, the AMBER-FB15 protein force field and the TIP3P-FB water model were used to describe their behaviors respectively.[16] The GAFF small molecule force field was used to parameterize the SAM substrate.[17] The structure preparations were all done with the help of *tleap* in Amber16.[18]

The simulation process started with the classical MD simulations followed by the QM/MM simulations. The following preparatory steps were used in equilibrating the classical MD simulations to relax any close contacts and other structural defects in the initial structure: First, energy minimization was carried out with the protein atoms frozen, followed by minimization of the whole system. Next, a sequence 200 ps heating simulations followed by a 200 ps equilibrium simulation was taken to heat up the system to 300 K and equilibrate the density under 1 atm pressure. Finally, a 50 ns MD simulation was carried out in the NVT ensemble, and the last frame of the simulation was chosen as the starting point in the QM/MM simulations. During the simulations, a Langevin thermostat algorithm with a collision frequency of 1.0 ps^{-1} was set to control the temperature. The cutoff values for both short-range electrostatics and van der Waals interactions were set to 12 Å, while the particle-mesh Ewald method was chosen to describe the long-range summation of the electrostatic interactions.

The QM/MM simulations were carried out using the Qchem 4.0[19] / AMBER12[18] software packages for the QM and MM regions respectively and joined together using a pseudobond Q-Chem/AMBER interface.[20] Different QM regions were used during the catalytic process depending on which region of the protein was involved in reactivity for each elementary step. Based on the other QM/MM simulations on iron-sulfur proteins,[21] the QM regions were described by the unrestricted B3LYP density functional approximation, and the basis set used was def2-SVP for most atoms and def2-SV(P)

for Fe. In instances where the QM and MM regions are covalently bonded, we adopted a pseudo-bond approach in which the boundary atom on the MM side of the covalent bond is modeled using seven electrons, a spherically symmetric pseudopotential and an STO-2G basis set.[22] The QM atoms directly bonded to the boundary atom used the 6-31G* basis set.

The catalytic reactions were studied by the following procedure: The initial structure was obtained by energy minimization of the entire QM/MM system, followed by a reaction coordinate (RC) driving procedure consisting of a series of energy minimizations in which the chosen RC consisting of linear combinations of one or more interatomic distances was constrained to a range of values. After verifying that scanning along the RC leads to the expected reactivity, the MM region of each constrained optimized structure was relaxed by carrying out a 500 ps MM/MD simulation with the QM region frozen. Following this, QM/MM MD simulations in the NVT ensemble with umbrella sampling along the defined RC were carried out using a 1.0 fs time step, totaling about 15 ps for each window. The energies and structures for the last 10 ps of simulation data were retained for structural analysis and generating the free energy profile with the weighted histogram analysis method (WHAM).[23,24]

3.3 Results and Discussions

3.3.1 Formation of the 10-s intermediate.

Similar to other SAM enzymes, the HydE catalytic process is initialized by the decomposition of SAM and the generation of 5′-dAdo$^{\bullet}$, which will trigger downstream radical reactions. Since the SAM decomposition is a well-known and well-studied biochemical reaction not only in experiments but also in computations, herein, we only use reaction coordinate driving without further free energy simulations (Figure S1)

to compute the reaction pathway resulting in 5′-dAdo$^{\bullet}$. Following this, we studied the radical addition of SAM to the cysteinyl sulfur, one of the key reactions in the HydE catalytic process (Figure 3.2). Although SAM enzymes generally follow a reactivity pattern of using the uses the 5′-dAdo$^{\bullet}$ to abstract a hydrogen atom from the substrate, HydE instead uses the 5′-dAdo$^{\bullet}$ to carry out radical addition. According to the free energy profile in Figure 3.2(A), this C-S radical addition is a very feasible reaction with a 5 kcal/mol barrier and releasing more than 11 kcal/mol energy. As shown in Figure 3.2(B), Figure 3.2(C), and Figure S2, the spin density in the 5′-dAdo$^{\bullet}$ is directed toward the sulfur atom of syn-B, providing a good orientation for the C-S addition. During the addition process, the spin density was observed to transfer from the C1 in 5′-dAdo$^{\bullet}$ to the dangler Fe as shown in Figure 3.2(C), corresponding to a change in the formal oxidation state from Fe(II) to Fe(I). This Fe(I) structure has been previously detected in EPR experiments and named the 10-second intermediate.[12] According to the geometries, the Fe—S distance increases from 2.4 Å to 3.2 Å during the C-S addition, which weakens the Fe-S coordination bond, and the resulting ligand is referred to as an S-adenosyl cysteine (SAC) molecule, equivalent to a demethylated SAM.

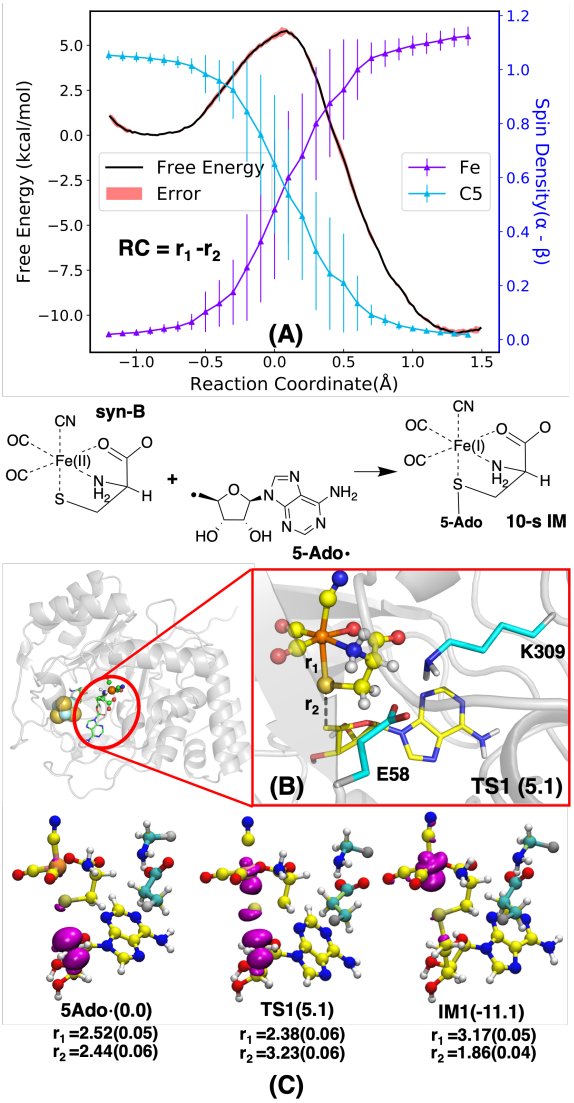


Figure 3.2: Formation of the 10-s intermediate by radical addition of 5′-dAdo• to the cysteinyl sulfur of syn-B. The free energy profile of the reaction is displayed in the top panel (A) with the spin density changes of the 5′-dAdo• C5 and dangler Fe. The transition state structure of this addition reaction is shown in ball and stick representation in the middle panel (B) highlighting the key distances along the reaction. The spin densities of key states are drawn in the bottom panel (C) to show the radical transfer during the reaction. The spin density isosurfaces are colored in magenta. Energies and key distances are given in kcal/mol and Å respectively.

3.3.2 Conversion of the 10-s intermediate

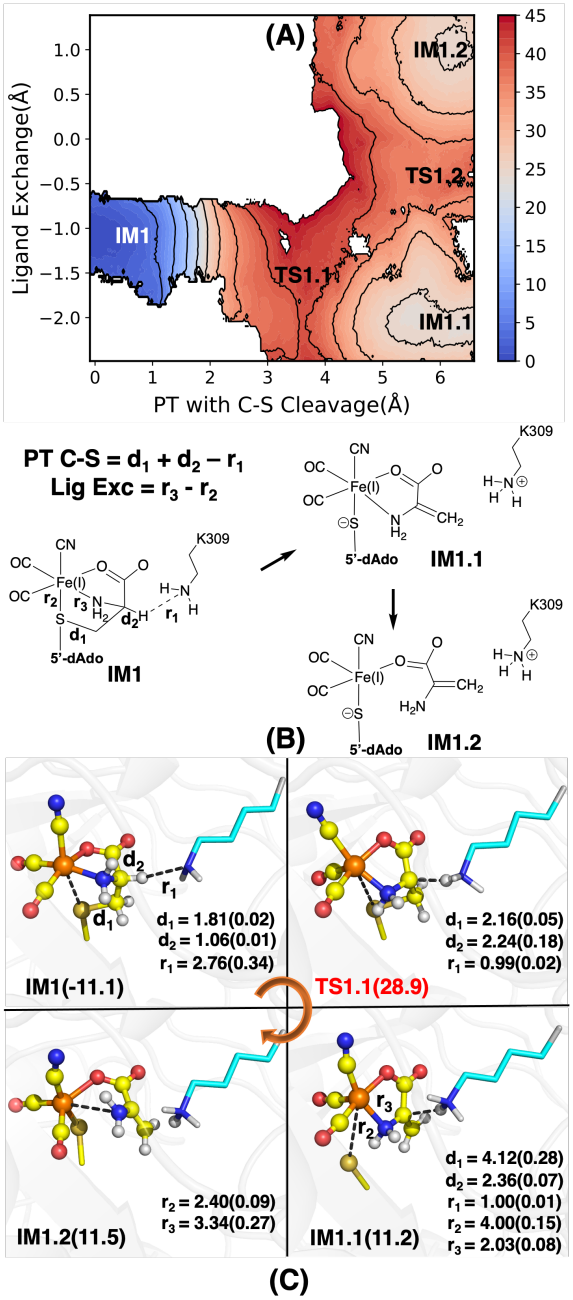


Figure 3.3: The hypothetical mechanism of cysteine decomposition via the Cβ-S bond cleavage with an unconventional proton transfer to a lysine side chain followed by ligand exchange. The 2-D free energy map of two reaction coordinates is displayed in the top panel (A). The key structures are shown in 2D in panel (B) along with the RC definitions, and panel (C) shows the 3D structures with key distances labeled. The free energies of the corresponding structures relative to the starting species are written in given in kcal/mol and distances are given in Å.

3.3.2.1 Hypothesis A: Cleavage of the cysteinyl Cβ-S bond via a unconventional β-elimination. Experiments show that the 10-second intermediate, formed in the previous step, is converted to a second intermediate on a 10-minute timescale that is coupled to the release of pyruvate. According to the current experimental data, this is a rate-limiting step taking roughly 10 minutes to accomplish in the HydE catalytic reaction. The cysteine (or SAC) decomposition was proposed as a β-elimination involving a proton transfer from the Hα of cysteine to a basic residue, which further triggers the cleavage of the Cβ-S bond, forming a 5′-S-dAdo species and dehydroalanine (dHA) coordinated to Fe(I).[12] Possible candidates for the proton acceptor include the side chains of E58 and K309, which form a salt bridge in the crystal structure.[13] We computed a free energy profile for a concerted proton transfer (PT) and C-S bond cleavage step in Figure 3.3, in which the proton is transferred from Cα to the neutral K309 side chain. The activation free energy of this step is as high as $\approx$ 40 kcal/mol relative to the 10-s intermediate; the size of the barrier could explain why this intermediate persists for 10 minutes in the experiment. The proton transfer is endothermic relative to the 10-s intermediate with $\Delta G \approx 22.3$ kcal/mol and forms a temporary intermediate, labeled IM1.1, that easily can interconvert with another structure labeled IM1.2. These two structures differ in whether the 5′-S-dAdo or the dHA amine group is coordinated to the Fe(I). According to the 2D free energy map and Figure S3(a), this mechanism occurs in a stepwise manner with Hα proton transferring first to the neutral K309, followed by the C-S bond cleavage, then the coordination change. In a separate calculation, we showed that leading with the coordination change increased the free energy by ¿25 kcal/mol with no local minimum found (in Figure S3(b)). The hybridization of the nitrogen atom in dHA changes during the conversion of IM1.1 to IM1.2. In IM1.1, the nitrogen is sp^3 hybridized and making four bonds to two hydrogens, carbon, and Fe, and a double bond is formed

between Cα and Cβ, whereas in IM1.2 the N atom is sp^2 hybridized and forms a conjugated π-system with Cα and Cβ.

Dehydroalanine is very easy to convert to 2-iminopropanoate (IM2) with the K309 and E58 side chains playing an important role as proton repositories. Figure S4 shows how this conversion occurs with the existence of a hydrogen bond network. According to the mechanism of this reaction, the dHA and a water molecule play a role as a proton shuttle to transfer the hydrogen from K309 to E58 and alkylate the dHA. The reaction is quite feasible with a barrier lower than 16 kcal/mol and releases about 10 kcal/mol of free energy. Among the reactions, one carbocation metastable state is generated, labeled MS1. The MS1 state is essential for the following hydrolytic reaction to further generate the pyruvate.

3.3.2.2 Hypothesis B: Cysteine decomposition pathway via radical-relay mechanism Although the generation of the 10-min intermediate is slow, the free energy barrier from the unconventional proton transfer (Cα-H to N) is still much higher than what is traditionally considered to be kinetically feasible in enzymes. Here we propose an alternative pathway with a lower barrier via a radical-relay mechanism. The first step in this pathway is the homogeneous cleavage of Cβ-S resulting in the radical transfer from the dangler iron to the Cβ, oxidizing the dangler iron to Fe(II) and leaving a primary radical at Cβ. Next, the K309 side chain transfers one hydrogen to the Cβ and then accepts the Hα from the Cα, converting the primary radical to a more stable secondary radical. The radical transfers back to the dangler Fe via a Fe-N coordination bond cleavage, reducing the dangler iron to Fe(I) again.

The energy profiles and the key structures with their spin densities are displayed in Figure 3.4 and Figures S5, S6, and S7. The cysteine decomposition via C—S bond homogeneous cleavage is the rate-limiting step of this mechanism, consistent with the

experimental observation that no other radical intermediate was detected in between the 10-second intermediate and 10-min intermediate. This Cβ-S bond homogeneous cleavage is endothermic with a barrier as high as 26 kcal/mol and absorbing 11 kcal/-mol of free energy, yielding the RIM1.1 state. The primary radical on Cβ (RIM1.1) is less stable, and the hydrogen atom transfer from the K309 side chain (RIM1.2 state) occurs with a barrier of 16.3 kcal/mol and is energetically neutral. Next, the hydrogen atom transfer from Cα radical transfer occurs with a similar energy barrier (17.8 kcal/mol) and releases 9.2 kcal/mol of free energy, which is consistent with the spin density plots of the RIM1.2 and RIM1.3 states in Figure 3.4(B). We also considered an alternative pathway in which the H atom was transferred directly from the Cα to Cβ, and the barrier was found to be as high as 38 kcal/mol (Figure S8), indicating that invoking K309 as a radical intermediate can significantly lower the barrier; similarly, we investigated whether the tyrosyl side chain of Y306 could substitute for K309 as a radical repository but found that the barrier was $\approx$ 42 kcal/mol from reaction coordinate driving, and did not pursue it further (Figure S8). The radical mechanism ends with the dissociation between the dangler Fe and the coordinating amine group, leading to the radical transfer from Cα back to the dangler Fe and the generation of the RIM2 state, which is chemically equivalent to the IM2 state in the proton-transfer pathway; the two structures can be made identical by reorienting the 2-iminopropanoate ligand and further elongating the Fe—N distance. As the energy diagram in Figure 3.5 shows, the alternative radical mechanism yields the 2-iminopropanoate cation with a barrier of about 27 kcal/mol, which is much lower than the 40 kcal/mol of the closed shell mechanism (Hypothesis A). These advantages reveal that the radical pathway is more favorable than the proton-transfer pathway for cysteine decomposition.

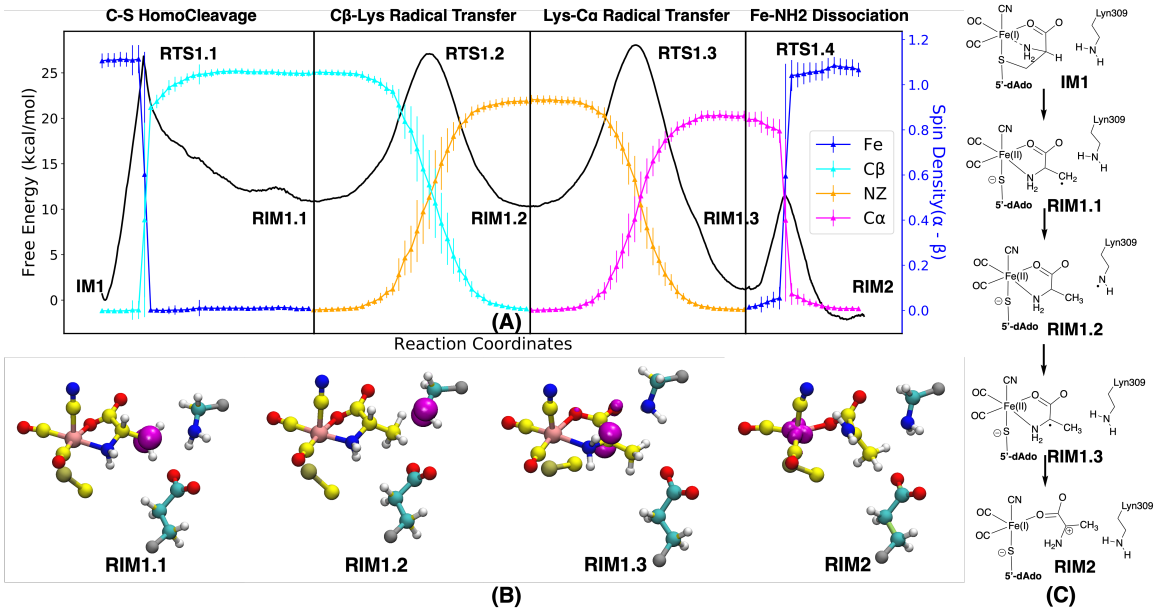


Figure 3.4: The radical mechanism of cysteine decomposition in HydE. The free energy profile (black curve) and the key spin densities (cyan, blue, orange, and magenta curves) are shown in panel (A). The spin densities of the stable states are drawn in panel (B) with the same color scheme as Figure 2. The 2D structures of the pathway are highlighted in panel (C) with unpaired electron density shown as magenta isosurfaces.

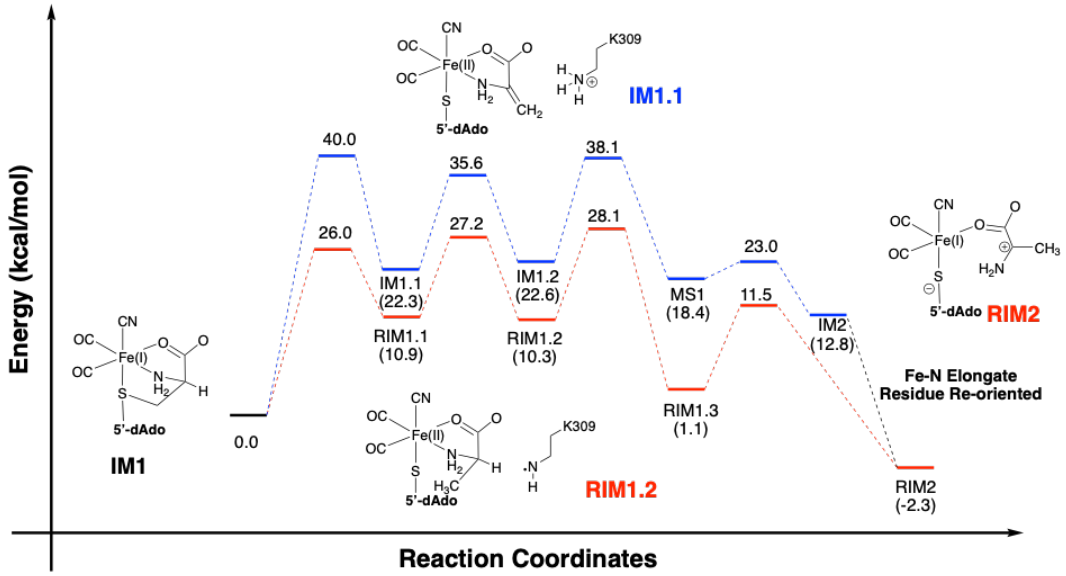


Figure 3.5: The comparison between the unconventional proton transfer pathway and the radical transfer pathway. Here, the proton transfer pathway is colored in blue while the radical transfer pathway is colored in red. Besides, the structures of the IM1 state, the IM1.1 state, the RIM1.2 state, and the RIM2 state are highlighted.

3.3.2.3 Conversion of 2-iminopropanoate cation into pyruvate. According to the experiments, the cysteine in Syn-B is converted into pyruvate on a 10-minute time scale and is released as a side product in the HydE catalytic reaction. We found that pyruvate generation is energetically accessible and involves hydrolysis of the Cα—N bond and several proton transfer steps that take advantage of the side chains of E58 and K309 as proton repositories. The energy diagram with several critical structures is depicted in Figure 3.6. The 2-iminopropanoate cation contains an electrophilic carbocation which is suitable for the nucleophilic attack of a water molecule, while the neutral K309 side chain accepts a proton from the water, leaving a hydroxyl group bonded to Cα. This step is energetically facile and the temporary product 2-amino-2-hydroxypropanoate, labeled as PIM1.1, is energy favorable compared to the RIM2 state ($\Delta G^{\ddagger} = 8.2, \Delta G = -1.8$ kcal/mol). A hydrogen bonding network between the 2-amino-2-hydroxypropanoate E58, and K309, allows protons to

be easily transferred back and forth between the amino group of the former and the side chains of the latter. The amino group can be protonated as shown in PIM1.3, which makes it into a good leaving group; the dissociation of NH_3 has the highest barrier of the pyruvate forming sequence with $\Delta G^{\ddagger} = 14.2, \Delta G = 8.7$ kcal/mol leading to PIM1.4 Following this, the protonated lysine K309 is able to protonate the ammonia with $\Delta G = -9.3$ kcal/mol. The entire reaction is exergonic with respect to the RIM2 starting structure with the 2-iminopropanoate ligand. The pyruvate has been detected in experiments as a side-product of the catalytic reaction and is presumably displaced by a water molecule, leading to the "10-min intermediate" observed in EPR experiments (Fig. 3.1).

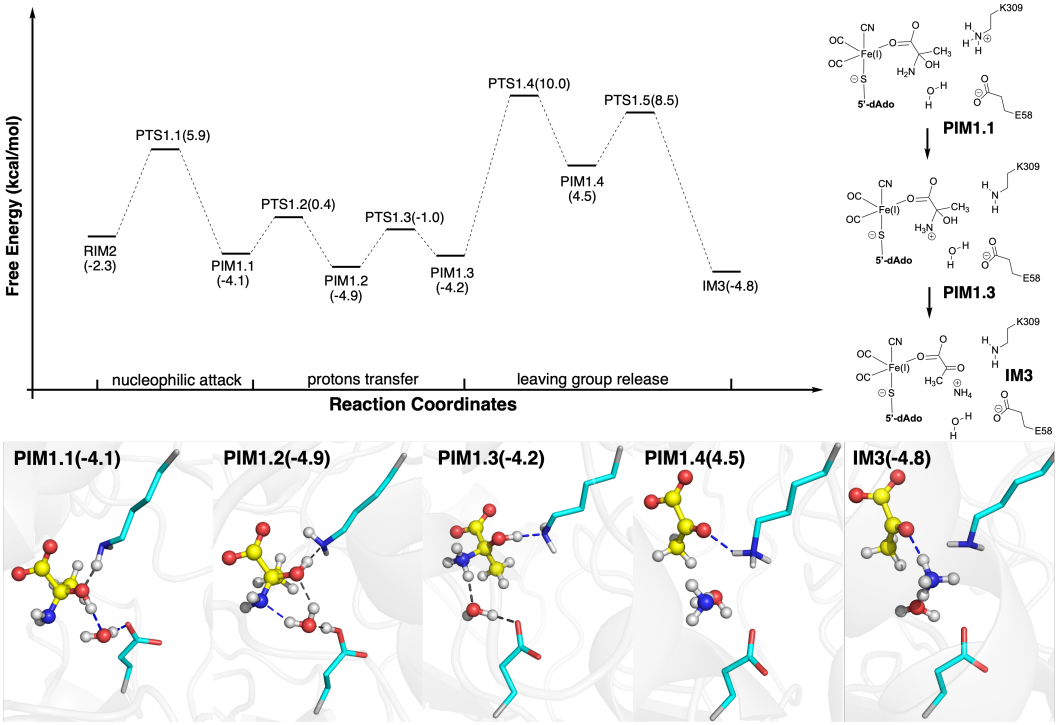


Figure 3.6: The energy diagram with key structures of the conversion of 2-iminopropanoate cation into pyruvate, consisting mainly of nucleophilic attack and proton transfer (PIM 1.1 and PIM 1.2), the protonation of the amine group (PIM 1.3), and the release and protonation of ammonia (PIM 1.4 and IM3). The bottom panel displays five key stable structures during the pyruvate generation process. The color scheme is identical to the previous figure.

3.3.3 Dimerization in HydE

3.3.3.1 Decomposition of the 10-min intermediate In Ref [13], a 5-coordinate $Fe(I)(CO)_2(CN)(Cl)(M-224)$ complex was observed in the HydE crystal structure by providing a syn-B substrate and triggering the radical reaction prior to crystallization. Considering the possibility of a diamagnetic Fe_2S_2 dimer product of HydE, the chloride ligand observed in the crystal structure is thought to be substituted by the cysteinyl sulfur in the HydE catalytic reaction. Therefore, we searched for a pathway in which the S-5′-dAdois cleaved from the 10-min intermediate. The mechanism with the lowest barrier was found to involve the homolytic dissociation of the C5—S bond coupled to temporary oxidation of Fe(I) to Fe(II), resulting in $Fe(II)(CO)_2(CN)(SH)(OH)$ and a 5′-dAdo$^\bullet$ radical, labeled as RIM4.1. The reaction proceeds by radical addition between 5'-dAdo$^\bullet$ and the coordinating OH group, resulting in the radical transfer back to the dangler Fe and yielding $Fe(I)(CO)_2(CN)(SH)$ weakly coordinated to an adenosine molecule (IM5 state). As shown in Figure 3.7(A), the overall barrier of this 2-step reaction is about 28.6 kcal/mol, which is comparable to the Cβ—S bond dissociation in the 10-s intermediate discussed above, and the analysis of spin density clearly shows the transfer of unpaired electron density from Fe to the C5′ of 5′-dAdo$^\bullet$, then back to Fe. The adenosine has been detected by experiment, which indirectly supports our mechanism. (*ref) Moreover, this mechanism is chemically analogous to our radical mechanism for the decomposition of the 10-s intermediate, as well as the initial radical addition of 5′-dAdo$^\bullet$ to syn-B in reverse. In all three cases, the Fe(I) atom acts as a repository for the unpaired electron and is oxidized to Fe(II) when radical intermediates are present.

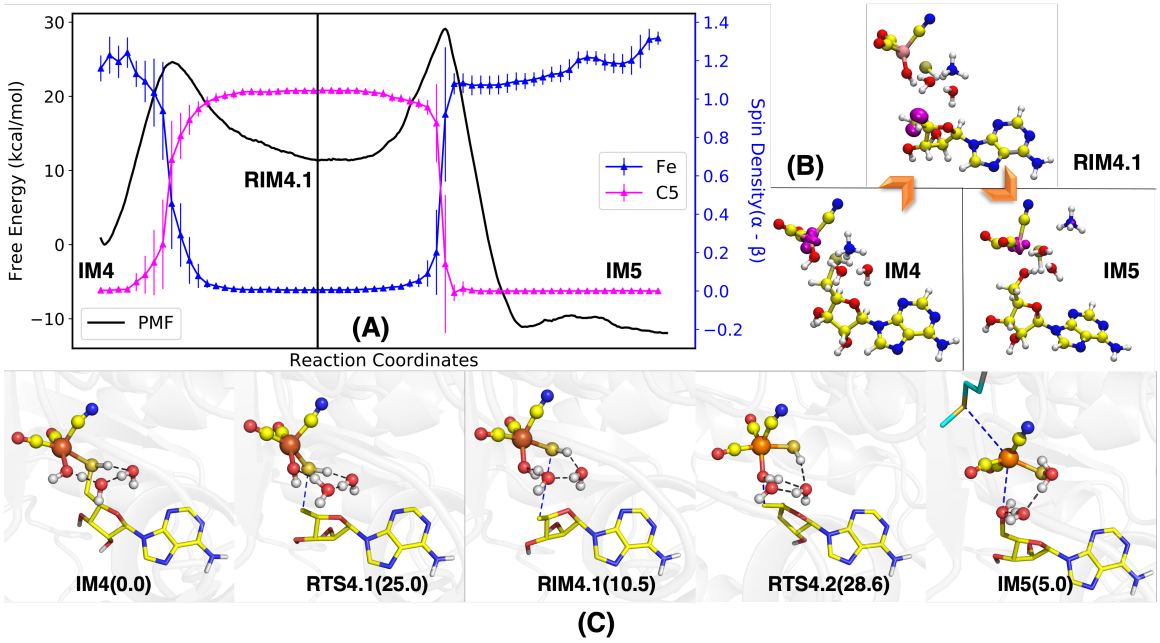


Figure 3.7: The decomposition mechanism of S-5′-dAdoin the 10-min intermediate. Starting with IM4, homolytic cleavage of the C—S bond occurs concomitantly with the oxidation of Fe(I) to Fe(II) and results in a 5′-dAdo$^{\bullet}$ radical. The transfer of OH$^{\bullet}$ from the Fe(II) complex quenches the radical and generates a Fe(I)$(CO)_2(CN)$ species weakly coordinated to an adenosine molecule; the release of adenosine was detected in experiments (IM5). The side chain of M291 is shown to be nearby to the Fe(I) complex in IM5.

Thus far, the decomposition of the 10-min intermediate has led to IM5, which contains Fe(I) weakly coordinated to adenosine, a good leaving group. The experimental structure has Fe(I) coordinated to the M224 side chain, suggesting that it displaces adenosine, but it is more than 8 Åaway from the generated Fe(I) cluster in our calculations. On the other hand, the M291 residue is much closer to the Fe-cluster in our structure of IM5 as shown in Fig. 3.7 Herein, we hypothesize that M291 will be the residue to harbor the Fe-cluster first instead of M224, and later the cluster can be transferred to the M224. The replacement of adenosine by the M291 was calculated and shown in Figure S9, and is found to be energetically very feasible with an energy barrier of $\approx$ 1.0 kcal/mol and releasing about 0.6 kcal/mol of energy. The formation

of $Fe(I)(CO)_2(CN)(SH)(M-291)$ shortens the distance between the Met224 and Fe(I) to 6 Å, making Fe-cluster transfer more possible from the Met291 to Met224.

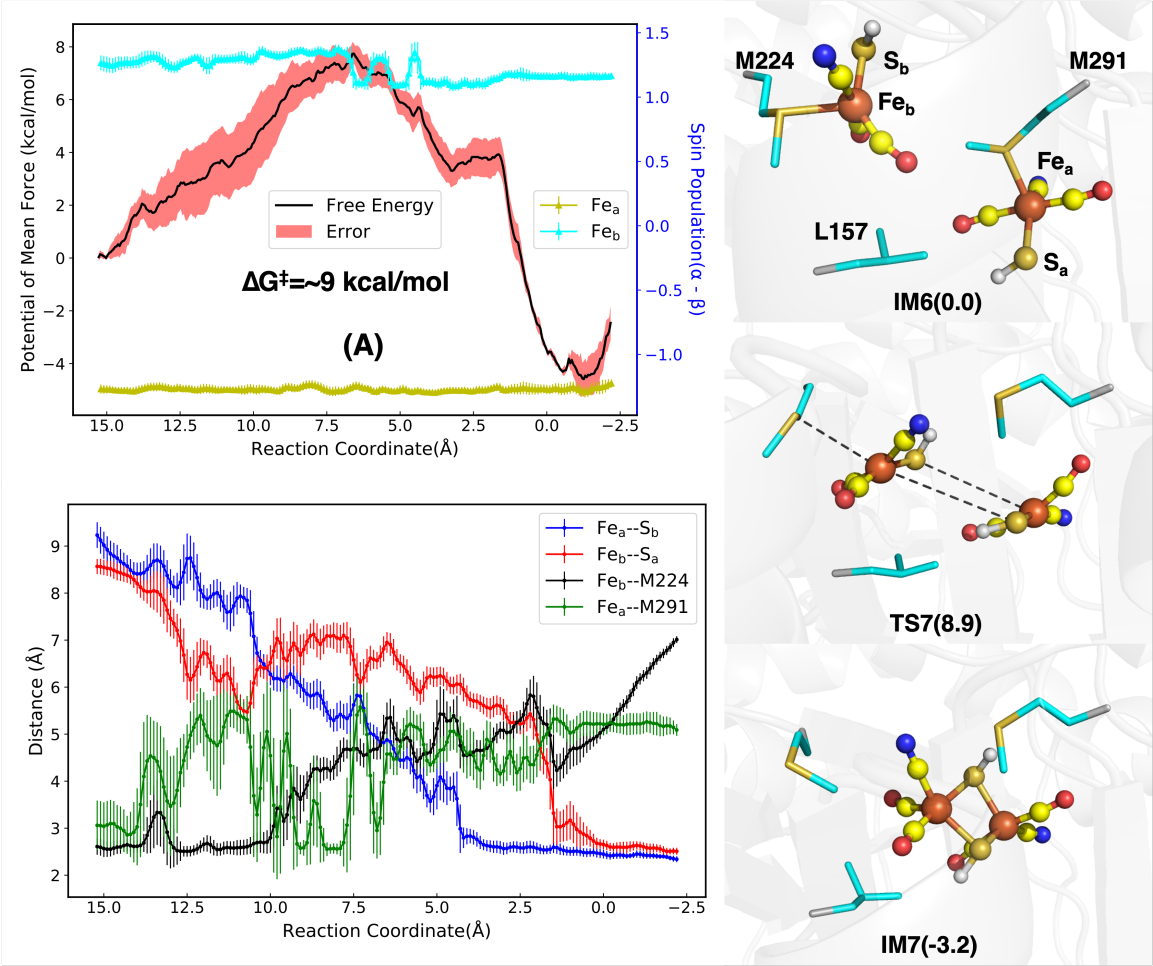


Figure 3.8: The dimerization mechanism of two $Fe(I)(CO)_2(CN)(SH)$ complexes resulting in $Fe(I)_2(SH)_2(CO)_4(CN)_2$. The free energy profile and Mulliken spin population are shown as a function of the reaction coordinate in the top left panel (A), showing that the product is an antiferromagnetically coupled species. The progression of key interatomic distances is plotted in (B) showing the dissociation from the Met side chains as the complexes are brought together. The initial, transition state and final structures are rendered in (C) with the same color scheme as previous figures.

3.3.3.2 Dimerization mechanism of Fe(I) complexes within HydE A short channel exists in the HydE structure between the Met291 and the Met224, providing enough space for the Fe(I)$(CO)_2(CN)(SH)$ species to be transferred from Met291 to Met224 and accommodating two Fe(I) clusters simultaneously. This indicates that HydE may be capable of carrying out the dimerization of the two clusters, yielding an antiferromagnetically coupled product with a $Fe(I)_2S_2$ core. Herein, the possible dimer initial structure is generated and displayed in Figure S10. To drive the dimerization process, the reaction coordinate was defined as the formation of two new Fe—S bonds while dissociating from the Met side chains as: d(Fe_a—S_b) + d(Fe_b—S_a) - d(Fe_b—S_{Met224}), where the subscripts "a" and "b" refer to the complexes initially coordinated to M291 and M224 respectively. This reaction coordinate has a large range due to the large distance changes and over 80 umbrella sampling windows were required. According to the energy profile in Figure 3.8, the dimerization process in HydE is kinetically and thermodynamically favorable with an energy barrier of less than 9 kcal/mol and releasing about 3 kcal/mol of free energy. As shown in Figure 3.8, the spin population of the two Fe atoms stays around -1.3 and +1.2 respectively across the entire reaction coordinate, showing that there is no qualitative change in the unpaired electron density and the final product contains two antiferromagnetically coupled Fe(I) atoms. Therefore, we can assert that our results are consistent with the EPR experimental results, in which there is no signal that would indicate a paramagnetic dimer product. The key structures and the key distances during the reactions are drawn in Figure 3.8(C). With the help of the distance changes in Figure 3.8(B), it is clear that the Fe_a-S_b and Fe_b-S_a distances decrease simultaneously at first, whereas the dissociation of Fe_b and Met224 occurs later. Along the reaction, the Met291 is one of the key residues that can coordinate with the Fe_a cluster around the transition state and stabilize the TS structure. After the TS, the Fe_a-S_b bond is formed followed

by the generation of the Fe_b-S_a bond. Finally, the reaction ends by elongating the Fe_b-S_{M224} distance. The dimerization product is located in the channel between M291 and M224. The small channel is relatively hydrophobic, preventing water molecules from entering the channel and binding to the unsaturated Fe clusters. The isobutyl side chains of the Leu157 and the dimethyl sulfide side chain of Met291 make a big contribution to constructing this hydrophobic channel, and the diatomic ligands on the Fe(I) cluster occupy much of the remaining space in the channel. Both of these reasons greatly reduce the number of water molecules in-between two Fe(I) clusters, and the water molecules in the pocket are mostly concentrated around the remaining 5′-dAdo. Moreover, our calculations showed that if the five-coordinate Fe(I) cluster were exposed to solvent, water forms a strong coordination bond that is difficult to release. This computational evidence supports the dimerization of the Fe(I) clusters in HydE before delivery to HydF.

3.4 Conclusions

In this article, we have proposed a complete mechanism for the catalytic cycle of HydE, in which two equivalents of a Fe(II)(cysteine)$(CO)_2(CN)$ substrate is converted to a diamagnetic $Fe(I)_2(SH)_2(CO)_4(CN)_2$ complex. The reaction is initialized with typical SAM decomposition, but immediately followed by an unconventional C-S radical addition to form a 10-s intermediate containing Fe(I)$(CO)_2(CN)$(SAC). The decomposition of the 10-s intermediate likely occurs through a radical-relay pathway involving homolytic cleavage of the Cβ—S bond on the SAC ligand, rather than a closed-shell proton transfer pathway, as the former has a much lower free energy barrier and preserves consistency with experiments. We further found that the decomposition involves the release of pyruvate, matching the experimental observation.

We have also proposed an energetically feasible mechanism for the dimerization of two Fe(I) complexes catalyzed by HydE. With the help of two methionines, namely Met224 and Met291, HydE is able to accommodate two single-Fe clusters simultaneously, and the dimerization was shown to be thermodynamically and kinetically favorable. We have also discussed other advantages of dimerization in HydE, such as the relative hydrophobic environment and the favorable relative positioning and orientation of the two within the protein. The species $[Fe_2(\mu\text{-}SH)_2(CN)_2(CO)_4]^{-2}$ is proposed to be the final product of HydE, and already contains several key elements of the final product of the HydG-HydE-HydF assembly line, that is, the H-cluster that forms the active site of [FeFe] hydrogenase.

References

[1]P. Dinis, D. L. Suess, S. J. Fox, J. E. Harmer, R. C. Driesener, L. De La Paz, J. R. Swartz, J. W. Essex, R. D. Britt, and P. L. Roach, "X-ray crystallographic and epr spectroscopic analysis of hydg, a maturase in [fefe]-hydrogenase h-cluster assembly", Proceedings of the National Academy of Sciences **112**, 1362–1367 (2015).

[2]D. L. Suess, I. Bürstel, L. De La Paz, J. M. Kuchenreuther, C. C. Pham, S. P. Cramer, J. R. Swartz, and R. D. Britt, "Cysteine as a ligand platform in the biosynthesis of the fefe hydrogenase h cluster", Proceedings of the National Academy of Sciences **112**, 11455–11460 (2015).

[3]D. L. Suess, C. C. Pham, I. Bürstel, J. R. Swartz, S. P. Cramer, and R. D. Britt, "The radical sam enzyme hydg requires cysteine and a dangler iron for generating an organometallic precursor to the [fefe]-hydrogenase h-cluster", Journal of the American Chemical Society **138**, 1146–1149 (2016).

[4]G. Rao, L. Tao, D. L. Suess, and R. D. Britt, "A [4fe–4s]-fe (co)(cn)-l-cysteine intermediate is the first organometallic precursor in [fefe] hydrogenase h-cluster bioassembly", Nature Chemistry **10**, 555–560 (2018).

[5]R. D. Britt, G. Rao, and L. Tao, "Bioassembly of complex iron-sulfur enzymes: hydrogenases and nitrogenases", Nature Reviews Chemistry **4**, 542–549 (2020).

[6]R. D. Britt, G. Rao, and L. Tao, "Biosynthesis of the catalytic h-cluster of [fefe] hydrogenase: the roles of the fe–s maturase proteins hyde, hydf, and hydg", Chemical Science **11**, 10313–10323 (2020).

[7]N. Chen, G. Rao, R. D. Britt, and L.-P. Wang, "Quantum chemical study of a radical relay mechanism for the hydg-catalyzed synthesis of a fe(ii)(co)2(cn)cysteine precursor to the h-cluster of [fefe] hydrogenase", Biochemistry **60**, 3016–3026 (2021).

[8]R. D. Britt, L. Tao, G. Rao, N. Chen, and L.-P. Wang, "Proposed mechanism for the biosynthesis of the [fefe] hydrogenase h-cluster: central roles for the radical sam enzymes hydg and hyde", ACS Bio & Med Chem Au **2**, 11–21 (2022).

[9]G. Rao, S. A. Pattenaude, K. Alwan, N. J. Blackburn, R. D. Britt, and T. B. Rauchfuss, "The binuclear cluster of [fefe] hydrogenase is formed with sulfur donated by cysteine of an [fe(cys)(co)(2)(cn)] organometallic precursor", Proceedings of the National Academy of Sciences **116**, 20850–20855 (2019).

[10]G. Rao, N. Chen, D. A. Marchiori, L.-P. Wang, and R. D. Britt, "Accumulation and pulse electron paramagnetic resonance spectroscopic investigation of the 4-oxidobenzyl radical generated in the radical s-adenosyl-l-methionine enzyme hydg", Biochemistry **61**, 107–116 (2022).

[11]G. Rao, L. Tao, and R. D. Britt, "Serine is the molecular source of the nh(ch2)2 bridgehead moiety of the in vitro assembled [fefe] hydrogenase h-cluster", Chem. Sci. **11**, 1241–1247 (2020).

[12]L. Tao, S. A. Pattenaude, S. Joshi, T. P. Begley, T. B. Rauchfuss, and R. D. Britt, "Radical sam enzyme hyde generates adenosylated fe(i) intermediates en route to the [fefe]-hydrogenase catalytic h-cluster", Journal of the American Chemical Society **142**, 10841–10848 (2020).

[13]R. Rohac, L. Martin, L. Liu, D. Basu, L. Tao, R. D. Britt, T. B. Rauchfuss, and Y. Nicolet, "Crystal structure of the [fefe]-hydrogenase maturase hyde bound to complex-b", Journal of the American Chemical Society **143**, 8499–8508 (2021).

[14]Y. Zhang, L. Tao, T. J. Woods, R. D. Britt, and T. B. Rauchfuss, "Organometallic fe2(μ-sh)2(co)4(cn)2 cluster allows the biosynthesis of the [fefe]-hydrogenase with only the hydf maturase", Journal of the American Chemical Society **144**, 1534–1538 (2022).

[15]C. H. Chang and K. Kim, "Density functional theory calculation of bonding and charge parameters for molecular dynamics studies on [fefe] hydrogenases", Journal of chemical theory and computation **5**, 1137–1145 (2009).

[16]L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martınez, and V. S. Pande, "Building a more predictive protein force field: a systematic and reproducible route to amber-fb15", The Journal of Physical Chemistry B **121**, 4023–4039 (2017).

[17]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field", Journal of Computational Chemistry **25**, 1157–1174 (2004).

[18]D. Case, K. Belfon, I. Ben-Shalom, S. Brozell, D. Cerutti, I. T.E. Cheatham, V. Cruzeiro, T. Darden, R. Duke, G. Giambasu, M. Gilson, H. Gohlke, A. Goetz, R. Harris, S. Izadi, S. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R. Walker, J. Wang, L. Wilson, R. Wolf, X. Wu, Y. Xiong, Y. Xue, D. York, and P. Kollman, Amber 2020, ucsf.

[19]Y. Shao, Z. Gan, E. Epifanovsky, A. T. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kuś, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. W. III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. D. Jr., H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. Hanson-Heine, P. H. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. D. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. Thom, T. Tsuchimochi,

V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. Chan, D. M. Chipman, C. J. Cramer, W. A. G. III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. S. III, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. V. Voorhis, J. M. Herbert, A. I. Krylov, P. M. Gill, and M. Head-Gordon, "Advances in molecular quantum chemistry contained in the q-chem 4 program package", Molecular Physics **113**, 184–215 (2015).

[20]Y. Zhou, S. Wang, Y. Li, and Y. Zhang, "Born–oppenheimer ab initio qm/mm molecular dynamics simulations of enzyme reactions", in Methods in enzymology, Vol. 577 (Elsevier, 2016), pp. 105–118.

[21]C. Zhao, Y. Li, C. Wang, and H. Chen, "Mechanistic dichotomy in the activation of sam by radical sam enzymes: qm/mm modeling deciphers the determinant", ACS Catalysis **10**, 13245–13250 (2020).

[22]Y. Zhang, "Pseudobond ab initio qm/mm approach and its applications to enzyme reactions", Theoretical Chemistry Accounts **116**, 43–50 (2006).

[23]B. Roux, "The calculation of the potential of mean force using computer simulations", Computer physics communications **91**, 275–282 (1995).

[24]A. Grossfield, Wham: an implementation of the weighted histogram analysis method.

4 Sequence-based Prediction of Metamorphic Behavior in Proteins[2]

4.1 Introduction

Christian Anfinsen was awarded a Nobel Prize in Chemistry in 1972 for his work on the apparent one-to-one relationship between the amino acid sequence of a protein and its three-dimensional fold,[1,2] giving rise to the classic paradigm: "one sequence, one fold". However, serendipitous discoveries in the past few decades have led to the identification of "metamorphic proteins"[3] that have the ability to jump reversibly between two distinctly different folds under native conditions. These proteins are fundamentally different[4] from intrinsically disordered proteins (IDPs)[5], morpheeins,[6] and moonlight proteins[7,8] which have been studied for a long time. Typical conformational changes in proteins often involve "shearing" or "hinge" behavior where entire protein subunits or secondary structure elements undergo relative motions without significantly altering the fold of the protein.[9,10] In contrast, the different folds/structures of a metamorphic protein are dissimilar on a more fundamental level, often involving changes such as the transformation of a whole α-helix into a β-strand (Figure 4.1). In this paper, we use significant changes in secondary structure as the key defining characteristic of metamorphic proteins.

Although the number of known examples of metamorphic proteins such as IscU[11], RfaH[12,13], Selecase[14], Mad2[15,16], Lymphotactin[17], CLIC1[18], KaiB[19,20] is relatively small, it is anticipated to increase steadily and populate the "Metamorphome". In all these metamorphs, the transition from one-fold to another takes place in response to environmental triggers like pH, temperature, salt concentrations, binding partners, redox state, or oligomerization. Uncovering the metamorphome is crucial as it is expected to

have a transformative effect on long-held concepts of protein structure and function. It could also lead to the engineering of metamorphic proteins, which are molecular switches, to act as sensors for small molecules or local environmental changes.

Traditional X-ray crystallography techniques, which account for solving 90% of the protein structures in the Protein Databank (PDB), are limited in their ability to identify metamorphism in proteins. These methods trap the protein in a minimum free energy structure in a specific crystallographic environment, thus they do not reveal the existence of alternate folds if the protein is metamorphic. A powerful method to detect protein metamorphism is solution-state NMR. However, high-throughput screening protein sequences for potential metamorphic behavior by NMR is not feasible. A realistic approach would be to identify metamorphic candidates using computational approaches, which would allow experimental verification to focus on a smaller set of candidate proteins.

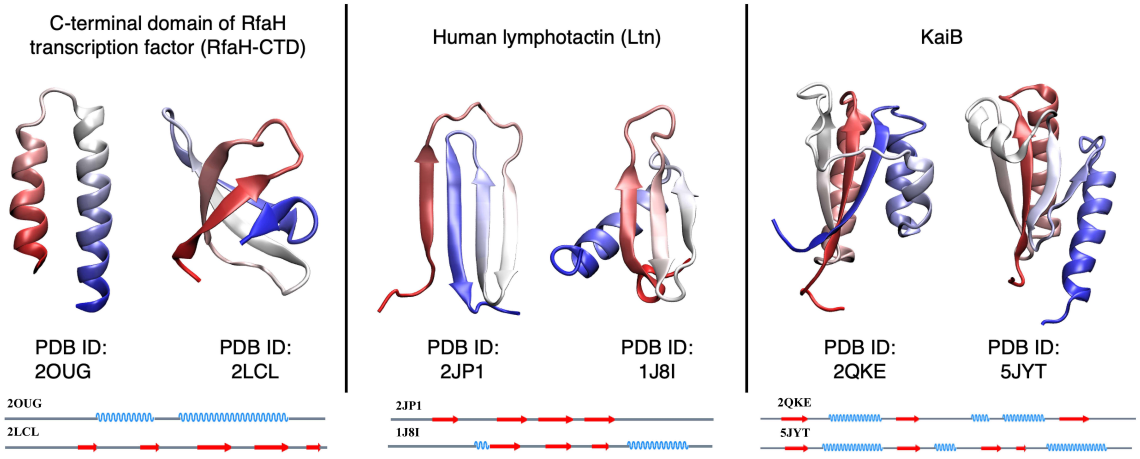


Figure 4.1: Representative examples of metamorphic proteins with 3-dimensional structures of both folds. The protein backbone is colored from N-terminal (red) to C-terminal (blue). Secondary structure diagrams corresponding to the 3D structures are shown at the bottom of each panel.

A recent computational study from Porter and Looger[21] identified 96 fold-switching

candidates in the Protein Data Bank. The study stated that two characteristics of metamorphic proteins include discrepancies between experimentally derived and computationally predicted secondary structures, and the occurrence of multiple independent subdomains that each fold cooperatively. Using these two metrics, they estimated that up to 4% of the proteins in the PDB may be metamorphic, which suggests that this class of proteins appears to be more common than those identified so far.

In this work, we propose a novel binary classifier for predicting protein metamorphism based on the diversity index, which takes advantage of the uncertainty in secondary structure prediction methods. This method has a unique advantage that it can predict metamorphic behavior in a protein of interest purely based on the amino acid sequence, without requiring a priori experimental knowledge of the three-dimensional structure. The classification method is trained using two reference datasets consisting of 200 manually annotated monomorphic and metamorphic sequences respectively. We found the robust performance of the diversity index-based classifier with a Matthews correlation coefficient of 0.4 (corresponding to 70% accuracy) that is largely insensitive to changes in the parameterization and training dataset.

The rest of this paper is organized as follows. We first give a brief overview of secondary structure prediction (SSP) methods, as they provide the essential inputs into our classifier. Next, we introduce the diversity index (DI) which measures the uncertainty of predicted secondary structure, and we outline how the DI is used to classify a protein sequence as metamorphic or monomorphic. This is followed by a description of the reference datasets containing known metamorphic and monomorphic sequences used to train our classifier. The performance of the classifier is discussed in detail using metrics such as the Matthews correlation coefficient, true positive rates, and true negative rates, and its robustness is tested using randomized cross-validation,

sensitivity analysis, and examining how performance varies with different input SSP programs. We include a discussion of "outlier" protein sequences that are consistently misclassified by the DI-based model, as well as how the performance depends on the sequence database for position-specific scoring matrix generation, an important auxiliary input for the SSP programs. The paper concludes with some promising future directions.

4.2 Theory

4.2.1 Secondary Structure Prediction (SSP)

Secondary structure (SS) is a property of amino acid residues within a protein structure that describes its local intrachain three-dimensional structure. Under the well-known DSSP system,[22] secondary structure may be classified into eight states, which can be further reduced down to three: α-helix (H), β-sheet (E), and random coil (C). In the years since the introduction of SS classification for known protein structure, several data-driven computational methods have emerged for secondary structure prediction (SSP) using only primary structure information, i.e. the amino acid sequence. Today, SSP is a vital part of the modern toolkit for protein structure prediction and design.

SSP methods can be understood in the conceptual framework of machine learning. The protein sequence is first processed into a feature vector consisting of information with structural relevance. Such features may include a PSSM (position-specific scoring matrix), which estimates the probability distribution of amino acid residues at each position in the sequence, and is computed by performing sequence alignments to a sequence database[23] using programs such as PSI-BLAST[24]. The feature vector is input into a neutral network model, which has significant flexibility in its internal

architecture, and provides three outputs representing the relative probabilities of helix (H), sheet (E), and random coil (C) at each position. The parameters of the neural network are trained to reproduce known secondary structures from widely available structural datasets. The accuracy of three-state SSP for modern methods has been reported to be as high as 82-84%.[25]

In this paper, four widely used SSP programs were applied to predict the secondary structure of every sequence in our datasets, namely Psipred[26], SPIDER2[27], SPIDER3[28], and Porter 5.0, denoted here as Porter5[29]. Psipred, developed in 1999, introduced the idea of using the PSSM generated by PSI-BLAST as input to a neural network for secondary structure prediction. SPIDER2 uses a deep neural network that incorporates the PSSM from PSI-BLAST along with amino acid physicochemical properties[30] to predict secondary structures and main chain dihedral angles. SPIDER3 is an updated version of SPIDER2 that incorporates hidden Markov model sequence profiles generated by the HHBlits program[31] as input to a bidirectional recurrent neural network architecture, effectively allowing the entire sequence to calculate SS prediction at each position instead of a sliding window as in SPIDER2. Porter5 is the latest version of a series of SSP programs and uses HHBlits-generated HMM sequence profiles and PSI-BLAST-generated PSSMs as input. In this paper, we used the UniProt90_2019_01 sequence database as the input to PSI-BLAST for PSSM generation, and the Uniclust30_2018_08 database was used as input for HHBlits. These published sequence alignment databases are distinct from the metamorphic and monomorphic reference datasets that were compiled as part of this work.

4.2.2 Metamorphic Proteins and Diversity Index

Metamorphic proteins can reversibly adopt multiple folded conformations for the same amino acid sequence under native conditions.[3,4] Moreover, representative examples of

metamorphic proteins are characterized by significant differences in secondary structure between folds (Figure 1), which is a distinct feature from more typical kinds of conformational change that generally preserve the secondary structure as described in the Introduction. Because these metamorphic proteins possess multiple stable folds with differences in secondary structure, our central hypothesis is that metamorphic protein sequences are able to "confuse" secondary structure prediction programs. According to this hypothesis, we defined one descriptor, the diversity index (DI):

$$DI = (P(E)^2 + P(H)^2 + P(C)^2)^{-1} \tag{2}$$

where P(E), P(H) and P(C) are output quantities from the SSP program representing the probabilities of strand, helix, and coil, respectively, for a single residue in the sequence. The DI for a residue is the reciprocal of the well-known Herfindahl and Simpson indices[32] for quantifying diversity in a probability distribution, and its value ranges from 1.0 (100% probability of one output, 0% for the other two) to 3.0 (equal probability of all three outputs). The value of the DI is also equivalent to the exponentiated Shannon entropy[33] in the above limiting cases, but takes on slightly different values for other distributions. High values of the DI indicate greater uncertainty in secondary structure prediction. Because metamorphic proteins tend to have contiguous portions of the sequence (or even the whole protein) capable of undergoing changes in secondary structure, we also hypothesized that the DI of metamorphic protein sequences are elevated in contiguous regions of the sequence. Therefore, we consider the maximum value of a moving average of the DI over the sequence as the main criterion to predict metamorphic behavior in a protein sequence. In other words,

a sequence will be classified as metamorphic if the following criterion is satisfied:

$$\max\left\{\frac{1}{CR}\sum_{j=0}^{j<CR} DI_{i+j}\right\}_{i=1}^{L-CR+1} > DI_{thre} \tag{3}$$

where the CR "number of consecutive residues" and DI_{thre} "diversity index threshold" are adjustable parameters. This binary classifier needs to be trained on reference or "manually annotated" datasets consisting of known-metamorphic and known-monomorphic (i.e. single fold) sequences. We will describe the construction of these datasets in the following sections.

4.3 Dataset Setup

4.3.1 Construction of the metamorphic reference dataset

In 2018, Porter et al published a paper listing 192 (96 pairs)[21] of existing metamorphic proteins21 in which most pairs have very high sequence similarity to one another (between 90% and 100%). Our metamorphic reference dataset, listed in Supplementary Table S1 in the publication, makes the following revisions to the listing in Ref. [21]. In eight cases, both folds of a metamorphic protein existed as different chains with identical sequences in a single structure, and these sequences are counted twice in our dataset (e.g. 5C1V). Among the original set of structures, one protein is no longer available from the database (PDB ID: 2A01). We also removed proteins where the fold switching region is contained within 20 residues of the N- and C-termini (4ZRB, counted twice) or if the sequence length is shorter than 40 residues (4FU4, 4G0D, 5K5G, and 2KB8); this is because our classifier requires taking a moving average of the diversity index, requiring a sequence that is longer than the largest window size (¿15 residues) plus the number of the removed terminal residues (¿5*2 residues). In

total, 8 proteins were removed from the list for the reasons above. We also added several proteins to the list, including the designed sequence pair GA/GB from protein G^{34} that was excluded from Ref. [21] (PDB ID: 2LHC and 2LHD) and 15 other possible metamorphic proteins which have experimental evidence for metamorphism but one solved structure, such as 2LSH. Therefore, our reference metamorphic dataset contains 192 – 8 + 15 + 2 = 201 metamorphic protein structures in total.

4.3.2 Construction of the monomorphic reference dataset

Our classification model for predicting protein metamorphism needs to be trained on proteins with known metamorphic behavior, as well as those with known single-fold (i.e. monomorphic) behavior. Although it is widely assumed that the PDB contains mostly monomorphic proteins, it is likely that a significant portion exhibits as-yet undiscovered metamorphic behavior. Therefore, we queried the PDB to obtain a set of protein structures that are highly likely to be monomorphic based on the following set of criteria: (1) The structure should be reported at least 10 years ago and has a good quality structure, in the sense that X-ray structures with resolution ¿ 2.2 Å were filtered out; (2) there must be ¿30 published structures with at least 50% sequence similarity with the structure of interest; (3) the sequence length is ¿ 40 and ¡ 250 residues, in order to meet the criteria of having a well-folded core while staying within the typical sequence lengths of globular proteins. Each structure found in the above manner is termed 'parent protein', and structures with high sequence similarity found in step (2) above are termed 'child proteins'. A total of 1387 'parent protein' structures with a maximum sequence similarity of 70% and more than 65000 'child protein' structures were downloaded along with their abstracts from the RCSB PDB web server using an automated crawler written in Python that uses the scrappy package. Two filtering rules were imposed in order to maximize the probability that

a ‘parent protein’ is monomorphic:

1. The root-mean-square deviation (RMSD) values were calculated for all pairs of structures after sequence alignment for a ‘parent protein’ and all of its ‘children’. The structure was excluded from the data set if any of the pairwise RMSD values exceeded 2.4 Å.

2. The mismatch in secondary structure (SS) was calculated for all pairs of structures after sequence alignment with a ‘parent protein’ and all of its ‘children’. A positional mismatch score is calculated by summation over aligned residues in a window of 30 residues in length, where “2” was assigned if one sequence is H and the other is E, and “1” was assigned if one sequence is C and the other is either E or H, then taking the maximum value over all window positions. The structure was excluded from the data set if the SS mismatch score between any pair of sequences exceeded 9.

Finally, the abstracts of the corresponding publications were checked for keywords such as ‘fold-switching’, ‘metamorphic’, ‘two-folds’, and other synonyms; if the abstract indicated possible metamorphic behavior, then it was excluded from this dataset as well. This procedure resulted in a total of 140 likely monomorphic proteins (Supplementary Table S2), including 4 proteins that we deemed to be monomorphic from reviewing the literature but did not meet the above criteria.

An example of a metamorphic protein (KaiB) and a monomorphic protein (1AB9) from our reference datasets is shown in Figure 4.2. The highest RMSD value in the KaiB (a typical example of metamorphic proteins) cluster exceeds 7.0 Å, and many pairs of sequences exhibit a secondary structure mismatch of 23 or greater. On the contrary, both the RMSD values and SS score are consistently low for the monomorphic protein, 1AB9.

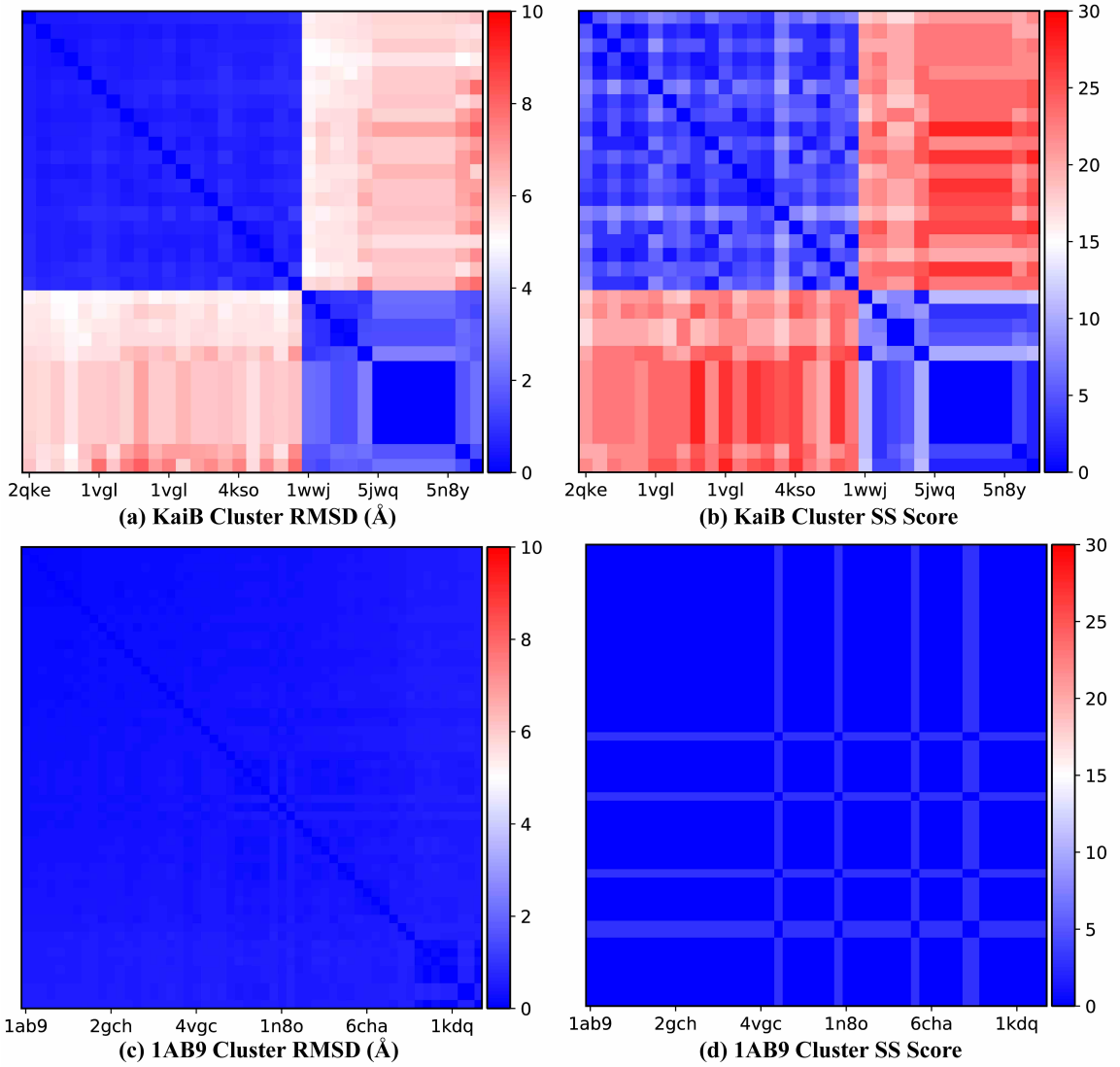


Figure 4.2: RMSD(a,c) and SS(b,d) score of KaiB and 1AB9, respectively. The highest RMSD value and highest SS scores of the KaiB cluster are much larger than those of the 1AB9 cluster

4.4 Results and Discussion

4.4.1 Behavior of the diversity index (DI)

According to Equation (1), the range of the diversity index (DI) is from 1 to 3, with larger values indicating greater uncertainty of SS prediction. Figure 3 plots

the SS and DI from the SPIDER2 program for a well-known metamorphic sequence (KaiB, left panel) and monomorphic sequence (ubiquitin, right panel) along with the experimentally derived secondary structure(s). As shown in the left panel, the DI of the KaiB sequence has several regions of elevated values in the metamorphic region that spans positions 60-90. On the other hand, DI of ubiquitin is relatively low for the whole sequence, with small jumps at the boundaries of different secondary structure domains that are smoothed out by taking the moving average. This example illustrates how diversity indices may be used to predict metamorphic behavior in proteins when the folds exhibit different secondary structure in the metamorphic regions.

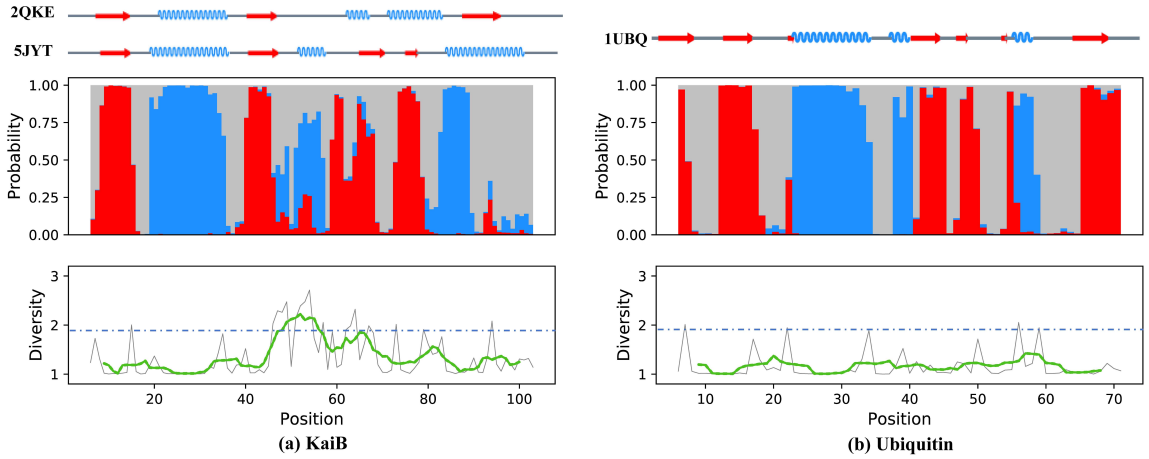


Figure 4.3: Secondary structure prediction results for KaiB (a) and Ubiquitin (b). Top: Sequences depict the experimentally derived SS of both KaiB structures (PDB ID: 2QKE and 5JYT, left) and Ubiquitin (1UBQ, right). Middle panel: Stacked bar plot showing predicted SS probabilities at each position in the sequence from SPIDER2 (red, strand (E); blue, helix (H); gray, coil (C)). Bottom panel: The diversity indices for each residue position in the sequence (gray), with a moving average with a window size of 14 (green) and DI threshold for metamorphic behavior (blue dotted line). The DI takes on higher values when the predicted SS is more evenly distributed between H, C, and E, indicating greater uncertainty. For KaiB, the higher DI regions coincide with the experimentally known metamorphic regions.

4.4.2 Diversity index-based classifier performance

The performance of our model is measured using the Matthews correlation coefficient (MCC), a well-established measure of the quality of binary classifications. For each combination of our parameters CR and DI_{thre}, the MCC is computed from a matrix of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), called a confusion matrix:[35]

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

The value of the MCC ranges from –1 to +1, where random classification gives a value of 0, perfect classification gives +1, and "perfectly wrong" classification gives –1 (equivalent to perfect classification if all predictions are reversed). For context, a recent review of machine learning methods for predicting disease in individuals reports MCC values ranging from –0.24 to +0.55.[36] An advantage of MCC is that the positive and negative data sets play equally important roles even if they are imbalanced in size. We also report simpler measures of true positive rate (TPR), true negative rate (TNR), and accuracy, defined as:

$$TPR = \frac{TP}{TP + FN},\ TNR = \frac{TN}{TN + FP},\ ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

which range from 0 to 1, and random classification gives 0.5. The accuracy is intuitive because it is simply the ratio of correct predictions to the total number of data points, but we do not use it in training because it can hide the effects of imbalanced performance for positive and negative cases.

Generally speaking, larger values of CR correspond to increased window size and tend to decrease the maximum value of the moving average. Larger values of DI_{thre}

also tend to decrease the probability that a sequence is classified as metamorphic. Thus, for increasing values of CR and DI_{thre}, the true negative and false negative rates both increase. In order to get a better understanding of the behavior of our model, we plotted heat maps of the MCC in our two-dimensional parameter space shown in Figure 4.4. Herein, we consider possible values of CR ranging from 6 to 15, and possible values of DI_{thre} ranging from 1.4 to 2.6 with a step size of 0.05.

The sensitivity of our model was tested by cross-validation. In general, cross-validation involves partitioning the dataset into the training set and test set, and verifying the model obtained from the training set by making predictions for the test set. Here we applied 6-fold cross-validation: The complete dataset including monomorphic and metamorphic proteins was randomly shuffled and split into six even-sized chunks. In each of the six trials, five selected chunks were treated as the training set and the remaining chunk as the test set. The parameters were determined by maximizing the MCC for the training set, then used to calculate the MCC for the test set.

Figure 4 and Table 1 show the main results for our DI-based classification using four SSP programs. Similar levels of performance for the training set were obtained using all four SSP programs as input to the DI-based classification. Among these methods, SPIDER2 had the highest average MCC value of 0.418 for the training set (Table 1), which was slightly higher than that of Porter5 (0.393), SPIDER3 (0.401), and Psipred (0.311); the differences were rather small and within the standard errors from randomized cross-validation trials. The parameters that maximized the MCC tended to appear in the middle of the parameter space, with significant regions of the parameter space exhibiting only minor variations from the optimum. For example, in the case of SPIDER2, the largest MCC value among all the trials was around CR 15 and DI_{thre} 2.1.

In terms of the test set, Porter5 and Psipred performed similarly with MCC values of 0.25–0.28, with differences being within the standard errors from randomized cross-validation trials. SPIDER2 has the highest average MCC value (0.355) for the test set. The small difference between the test set and training set MCCs, and the consistency of our results across several models, indicate that the DI-based classifier is a robust method for predicting metamorphic behavior. SPIDER3 showed moderate performance in the test set compared with the other three methods, with a MCC of 0.332. Figure 4.4(b) also shows that SPIDER2 has a broader range of parameter space with near-optimal performance as compared to Psipred, SPIDER3, and Porter5. The training and test results overall indicate that higher secondary structure prediction accuracy does not directly translate to better performance in metamorphic protein classification.

Although these methods had similar MCC values, their accuracy in terms of correctly predicting true positives and true negatives showed much greater variations. According to the data shown in Table 2, the true positive rate (TPR) is lower than the true negative rate (TNR) in all four methods for the optimum parameters that maximized the MCC. Among these methods, SPIDER3 has by far the highest TNR value (0.92) and lowest TPR (0.42). The other three methods had similar TPR ranging from 0.59–0.66 and TNR ranging from 0.78–0.83, which are within the limits of statistical errors from our cross-validation studies. We presumed that the large TNR values of SPIDER3 come from overall low values of the calculated diversity index, which possibly originates from higher SS prediction confidence levels as compared to other methods. We thus recommend SPIDER2 as the input method of choice for metamorphic protein classification, due to its consistently high MCC value for both training and test sets, balanced true positive and true negative rates, and wide regions of parameter space with near-optimal performance.

MCC	Psipred	SPIDER2	SPIDER3	Porter5
Training Set	0.311 (0.018)	**0.417 (0.015)**	0.401 (0.017)	0.393 (0.035)
Test Set	0.255 (0.091)	**0.355 (0.104)**	0.327 (0.124)	0.272 (0.096)

Table 1: The MCC results from four different SSP programs, including the training set results and the test set results. Numbers in parentheses are sample standard deviations over cross-validation trials.

Measure	Psipred	SPIDER2	SPIDER3	Porter5
True Positive Rate	0.598 (0.113)	0.633 (0.087)	0.411 (0.017)	0.581 (0.018)
True Negative Rate	0.707 (0.071)	0.782 (0.064)	0.918 (0.023)	0.763 (0.062)
Accuracy	0.649 (0.021)	0.698 (0.015)	0.633 (0.016)	0.645 (0.027)

Table 2: True positive rate, false positive rate, and accuracy of test set for four different SSP programs. Numbers in parentheses are sample standard deviations over cross-validation trials.

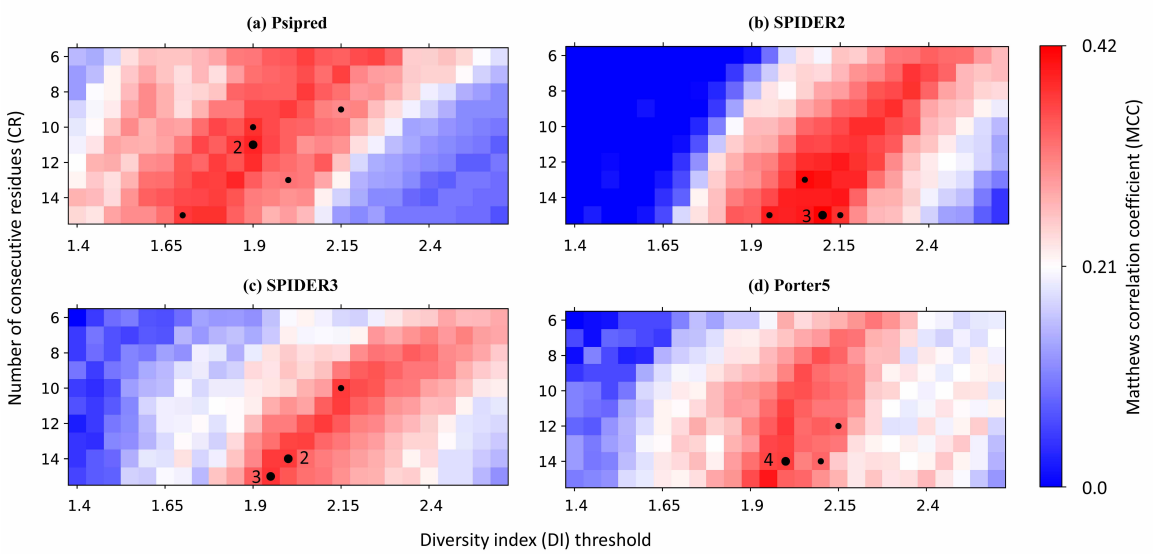


Figure 4.4: MCC heat maps for the diversity index-based classifier using predicted secondary structure from four programs, namely (a) Psipred, (b) SPIDER2, (c) SPIDER3, and (d) Porter5. The color map (blue ¡ white ¡ red) corresponds to MCC values computed for the full reference dataset. Each point indicates the optimized parameter value for a randomly selected training set (84.3% of the full dataset), with numbers indicating how many times each optimum was found.

4.4.3 Comparison with other methods

There currently exist a few methods in the literature for predicting metamorphic behavior in proteins.[21,37] To our knowledge, all existing methods require the knowledge of either the protein's three-dimensional structure or secondary structure information from the experiment. Porter et al. hypothesized that metamorphic proteins possess at least one independent folding domain with a 3-D structure that is largely independent of the rest of the sequence and proposed a method to predict metamorphic behavior based on the prediction of independent folding domains.[21,37] This method uses the protein's 3D structure as essential input data, and thus its predictions are based on existing structural knowledge.

More recently, Porter et al and coworkers reported that metamorphic proteins have lower SSP accuracy than monomorphic proteins or fragments,[38] which is similar to the ideas in our current work; however, the method they proposed requires prior knowledge of experimental secondary structure. A major differentiating feature of the diversity index-based classification method presented here is that it requires no experimental data for the sequence of interest. Thus, this method could be used to make predictions of metamorphism in protein sequences where there is no existing structural data.

4.4.4 Classification using multiple diversity indices

We also examined the possibility of obtaining an improved classification model based on a linear combination of DIs obtained from two SSP programs, essentially increasing the number of descriptors to two. The discriminant parameters (i.e. slope and intercept of the line) were optimized by maximizing the MCC.[35] Using a linear combination of the SPIDER2 DI and the Porter5 DI, we found the MCC value of the

optimal model increases to 0.45. Figure 4.5 plots the discriminant line and the descriptor values for each protein as a scatter plot. The diagonal shape of the distribution indicates a high degree of correlation between the two diversity indices (R^2=0.41), and most of the metamorphic proteins identified as true positives (TP) are located in the top-right corner of the figure. We found similar performance using some alternate approaches, for example, an "inconsistency index" to predict metamorphism using the level of disagreement between two SSP programs[39] (Supplementary Figure S1), and principal component analysis on the results of multiple SSP programs followed by K-means clustering (Supplementary Figure S2). These methods all yielded results with MCC values within 0.1 of the basic method using a single diversity index.

However, our analysis also revealed some false negatives (FN, open red circles) in the lower left of Figure 4.5; these are metamorphic proteins in our reference dataset but have very low diversity indices, and contradict our rationale for the DI-based classifier. The same applies to false positives (FP, open blue circles) in the upper right of Figure 4.5, as these are monomorphic proteins in the reference dataset with high diversity indices. In the following section, we provide a rationale to explain these outliers.

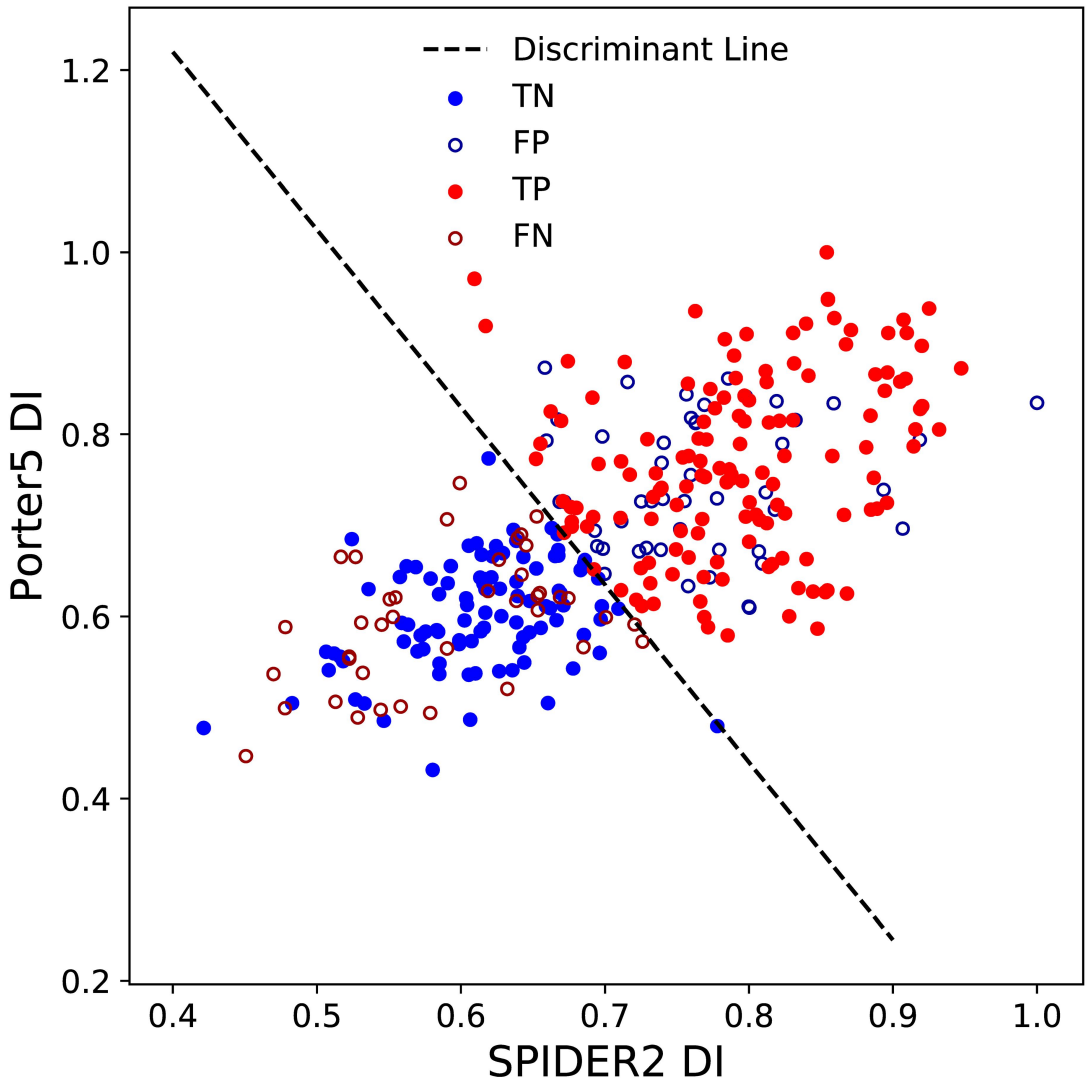


Figure 4.5: Two-parameter classification using diversity indices from SPIDER2 and Porter5. Here the true positives (metamorphic proteins in the reference dataset correctly classified as metamorphic) are represented by red-filled circles, and the false negatives (metamorphic proteins in the reference dataset incorrectly classified as monomorphic) are represented by red open circles. The blue-filled circles and blue open circles represent true negatives (monomorphic proteins in the reference dataset classified as monomorphic) and false positives (monomorphic proteins in the reference dataset classified as monomorphic), respectively.

4.4.5 Analysis of outliers in diversity index-based classification

Several proteins are consistently misclassified by the DI method using the predicted SS from all four programs. There are 22 metamorphic proteins in our reference dataset that are consistently misclassified as monomorphic proteins (false negatives), and 14 monomorphic proteins consistently misclassified as metamorphic proteins (false positives).

We examined the 22 "persistent" false negatives, i.e. metamorphic proteins from our reference dataset that are consistently misclassified as monomorphic, and generally found that their two folds did not satisfy our initial criterion of having significantly different secondary structures, and instead feature other kinds of conformational differences, which we discuss in the following examples. The three-dimensional structures of the false negatives are shown in Supplementary Figure S3.

1. 2LQW[40]:2BZY[41]. A closer examination reveals that these two structures have very similar secondary structures. As shown in Figure S2, 2LQW is a key signaling protein that exists as a monomer while 2BZY is a partial structure of a CrkL homo-dimer protein, in which the existing part has a similar SS to 2LQW. Moreover, the truncated CrkL monomer protein (PDB ID 2BZX) has a highly similar secondary structure to 2BZY. The high similarity in secondary structure is consistent with the classification assigned by our method, which is based on differences in secondary structure between folds.

2. 2NNT[42]:2MWF[43]. 2NNT is a tetramer amyloid protofilament that forms an extended β-sheet between multiple chains, whereas 2MWF is a mutant monomer that forms a β-sheet within the residues in one chain. Again, the highly similar secondary structure in both folds is consistent with the classification assigned by our method.

3. 4HDD[44]:2LEP[45]. The structures in this pair are similar in terms of secondary structure but have a large RMSD. 4HDD is a homodimer in which a β-sheet is formed between chains, whereas 2LEP uses the same domain to form a β-sheet within one chain.

4. 1G2C[46] is a truncated protein whose SS closely matches with the corresponding residues in 5C6B, which is a full structure. Strikingly, the other protein 5C6B36, which has a similar sequence to 1G2C, is correctly classified by SPIDER2 and PORTER5 as a metamorphic protein. The high-DI domain of 5C6B (residue 270 to 295) was not part of the 1G2C structure, which indicates the incorrect classification of 1G2C might be solely due to the truncation of the input sequence.

5. 4XWS[47]:4Y0M[48]. Upon examination of the structures, we think this structure pair had been incorrectly included in our reference metamorphic dataset, as these two structures are highly similar in terms of secondary structure as well as three-dimensional structure (RMSD: 1.561 Å). In fact, the text of Ref. [47] states that the metamorphic region of the protein could not be solved by X-ray crystallography.

1. 2UU8[49] is a concanavalin A protein and its DI value is relatively high in all the SSP programs, particularly in SPIDER2 (about 2.5). This structure possesses many short adjacent domains with different secondary structures, which leads to uncertainty in the SSP programs. Also, β-sheets are dominant in the SS of 2UU8, and we observed that the outermost strand of the β-sheet has a general tendency to have high uncertainty from SSP programs.

2. 3SEB[50] is a protein with several short SS domains (including α-helices and β-

sheets), leading to the high DI values for these domains, similar to the example above. More than half of the false positives follow the same trend, indicating that our DI-based classifier is biased to misclassify protein sequences that are monomorphic but intrinsically difficult for SSP programs due to having many short subdomains with distinct SS or outermost β-sheet among several (anti-)parallel β-sheet strands.

3. 2JE7[51], a recombinant lectin, has the same situation that the outermost strand of β-sheet and the short adjacent SS domains have the highest DI values. However, this protein is known to form either a dimer or a tetramer depending on the pH value. Although no direct evidence shows the SS change during this dimer-tetramer equilibrium process, it is possible that this process is associated with metamorphism not yet discovered.

4. 3CHB[52] is another labeled monomorphic protein whose DIs are very large in all the SSP programs. Unlike the other two false positive proteins above, 3CHB[52] has a long α-helix and five medium length β-sheets. Another short length α-helix is located at the N-terminal. According to the SPIDER2 prediction, three out of five β-sheets have relatively large DIs, resulting in a region with a high average DI value. So far, we do not have a good explanation for the reason for this false positive. One possibility is that other proteins in the PDB have highly similar sequences to these high-DI β-sheets but have different SS.

We note that it is possible for our reference monomorphic dataset to include proteins that are actually metamorphic, despite our efforts to minimize this occurrence. This is because our selection of monomorphic proteins was based on the analysis of known structures in the PDB, which by definition excludes alternate folds or structures that have not yet been discovered or deposited.

4.4.6 Dependence of results on sequence database

The performance of SSP programs relies on the non-redundant sequence database that is used to compute the position-specific scoring matrix. Figure 6 shows the differences in classification performance when SPIDER2 is used as the SSP program for different choices of the non-redundant sequence database. The Uniref-50, Uniref-90, and Uniref-100 databases have a progressively larger number of sequences and sequence identity among pairs of sequences. Figure 6 shows that Uniref-50 has a markedly lower performance for our classifier compared to Uniref-90 and Uniref-100, and it is currently unclear whether the poor classification performance is due to the smaller size of the sequence dataset or the more stringent threshold on sequence identity. Surprisingly, the modified non-redundant sequence dataset[53] from I-TASSER (PSSpred) gives a very high MCC value (0.457), even though it was released in 2014 and has not been continually updated as the other three. Thus, the DI-based classification performance depends on the sequence database in a nontrivial way and does not necessarily yield improved results for updated database versions.

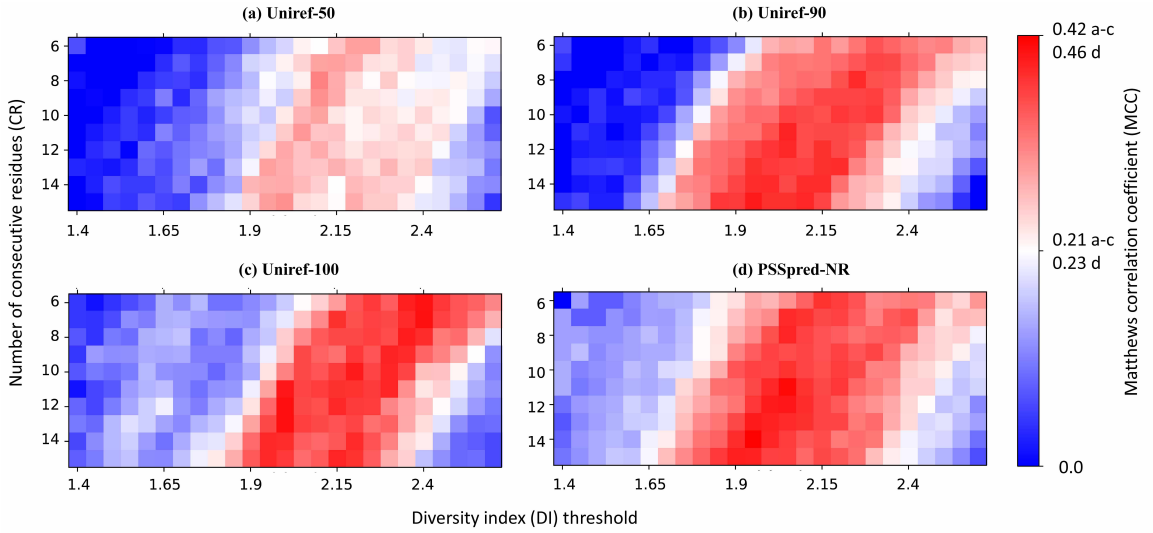


Figure 4.6: The different MCC values calculated based on: (a) 50% non-redundant sequence dataset, (b) 90% non-redundant sequence dataset, (c) 100% non-redundant sequence dataset and (d) PSSpred non-redundant sequence dataset.

4.5 Conclusions

In this paper, we described a diversity index-based classification model to predict metamorphic behavior in proteins solely based on the protein sequence. Our model was trained on a reference dataset consisting of 136 known monomorphic proteins and 201 known metamorphic proteins. Although the main purpose of SSP programs is to predict secondary structure, our results indicate that the "byproducts" of SSP, namely the alternate SS probabilities and the derived diversity index, can play a key role for predicting metamorphism in proteins. Among the four popular SSP programs, SPIDER2 has the overall best performance and robustness in classifying proteins as monomorphic vs. metamorphic. Further improvements in performance may be obtained by comparing the output of multiple SSP programs. Because all four SSP programs give similar MCC values when used in classification to within 10%, we think further improvements in predicting protein metamorphism will require SSP

methods that focus more on accurate quantification of uncertainty rather than yielding the best fit to experimental data. There is also potential for improvement in curating the annotated metamorphic and likely-monomorphic datasets; for example, the thermodynamic stability of the native state could be used as a criterion for a likely-monomorphic protein.[37]

Our examination of false positives and false negatives illustrates both the predictive potential and the limitations of the DI-based approach. In terms of false positives, we found some indications of undiscovered metamorphic behavior in the monomorphic dataset, possibly driven by pH-dependent changes in stoichiometry. On the other hand, the false negatives highlight fold-switching behavior in proteins that is not well-described by significant changes in secondary structure. This indicates that metrics going beyond SSP may be needed to predict certain kinds of protein metamorphism, which is a promising direction of future research.

References

[1]C. B. Anfinsen, E. Haber, M. Sela, and F. White Jr, "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain", Proceedings of the National Academy of Sciences **47**, 1309–1314 (1961).

[2]C. B. Anfinsen and E. Haber, "Studies on the reduction and re-formation of protein disulfide bonds", Journal of Biological Chemistry **236**, 1361–1363 (1961).

[3]A. G. Murzin, "Metamorphic proteins", Science **320**, 1725–1726 (2008).

[4]A. F. Dishman and B. F. Volkman, "Unfolding the mysteries of protein metamorphosis", ACS chemical biology **13**, 1438–1446 (2018).

[5]J. Ross, "Comments on the article "persistent confusion of total entropy and chemical system entropy in chemical thermodynamics"[(1996) proc. natl. acad. sci. usa 93, 7452–7453]", Proceedings of the National Academy of Sciences **93**, 14314–14314 (1996).

[6]E. K. Jaffe, "Morpheeins–a new structural paradigm for allosteric regulation", Trends in biochemical sciences **30**, 490–497 (2005).

[7]J. Piatigorsky, W. E. O'Brien, B. L. Norman, K. Kalumuck, G. J. Wistow, T. Borras, J. M. Nickerson, and E. F. Wawrousek, "Gene sharing by delta-crystallin and argininosuccinate lyase.", Proceedings of the National Academy of Sciences **85**, 3479–3483 (1988).

[8]C. J. Jeffery, "Moonlighting proteins", Trends in biochemical sciences **24**, 8–11 (1999).

[9]M. Gerstein, A. M. Lesk, and C. Chothia, "Structural mechanisms for domain movements in proteins", Biochemistry **33**, 6739–6749 (1994).

[10]S. Mitternacht and I. N. Berezovsky, "Binding leverage as a molecular basis for allosteric regulation", PLoS computational biology **7**, e1002148 (2011).

[11]Z. Dai, M. Tonelli, and J. L. Markley, "Metamorphic protein iscu changes conformation by cis–trans isomerizations of two peptidyl–prolyl peptide bonds", Biochemistry **51**, 9595–9602 (2012).

[12]F. Werner and D. Grohmann, "Evolution of multisubunit rna polymerases in the three domains of life", Nature Reviews Microbiology **9**, 85–98 (2011).

[13]B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Rösch, "An α helix to β barrel domain switch

transforms the transcription factor rfah into a translation factor", Cell **150**, 291–303 (2012).

[14]M. López-Pelegrın, N. Cerdà-Costa, A. Cintas-Pedrola, F. Herranz-Trillo, P. Bernadó, J. R. Peinado, J. L. Arolas, and F. X. Gomis-Rüth, "Multiple stable conformations account for reversible concentration-dependent oligomerization and autoinhibition of a metamorphic metallopeptidase", Angewandte Chemie **126**, 10800–10806 (2014).

[15]X. Luo and H. Yu, "Protein metamorphosis: the two-state behavior of mad2", Structure **16**, 1616–1625 (2008).

[16]M. Mapelli, L. Massimiliano, S. Santaguida, and A. Musacchio, "The mad2 conformational dimer: structure and implications for the spindle assembly checkpoint", Cell **131**, 730–743 (2007).

[17]R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman, "Interconversion between two unrelated protein folds in the lymphotactin native state", Proceedings of the National Academy of Sciences **105**, 5057–5062 (2008).

[18]D. R. Littler, S. J. Harrop, W. D. Fairlie, L. J. Brown, G. J. Pankhurst, S. Pankhurst, M. Z. DeMaere, T. J. Campbell, A. R. Bauskin, R. Tonini, et al., "The intracellular chloride ion channel protein clic1 undergoes a redox-controlled structural transition", Journal of Biological Chemistry **279**, 9298–9305 (2004).

[19]Y.-G. Chang, S. E. Cohen, C. Phong, W. K. Myers, Y.-I. Kim, R. Tseng, J. Lin, L. Zhang, J. S. Boyd, Y. Lee, et al., "A protein fold switch joins the circadian oscillator to clock output in cyanobacteria", Science **349**, 324–328 (2015).

[20]R. Tseng, N. F. Goularte, A. Chavan, J. Luu, S. E. Cohen, Y.-G. Chang, J. Heisler, S. Li, A. K. Michael, S. Tripathi, et al., "Structural basis of the day-night transition in a bacterial circadian clock", Science **355**, 1174–1180 (2017).

[21]L. L. Porter and L. L. Looger, "Extant fold-switching proteins are widespread", Proceedings of the National Academy of Sciences **115**, 5968–5973 (2018).

[22]W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features", Biopolymers: Original Research on Biomolecules **22**, 2577–2637 (1983).

[23]K. D. Pruitt, T. Tatusova, and D. R. Maglott, "Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins", Nucleic acids research **33**, D501–D504 (2005).

[24]S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool", Journal of molecular biology **215**, 403–410 (1990).

[25]Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, and Y. Zhou, "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?", Briefings in bioinformatics **19**, 482–494 (2018).

[26]D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices", Journal of molecular biology **292**, 195–202 (1999).

[27]Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, and Y. Zhou, "Spider2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks", Prediction of protein secondary structure, 55–63 (2017).

[28]R. Heffernan, Y. Yang, K. Paliwal, and Y. Zhou, "Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility", Bioinformatics **33**, 2842–2849 (2017).

[29]M. Torrisi, M. Kaleel, and G. Pollastri, "Porter 5: state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes", BioRxiv, 289033 (2018).

[30]J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke, "Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks", Molecular modeling annual **7**, 360–369 (2001).

[31]M. Remmert, A. Biegert, A. Hauser, and J. Söding, "Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment", Nature methods **9**, 173–175 (2012).

[32]E. H. Simpson, "Measurement of diversity", nature **163**, 688–688 (1949).

[33]C. E. Shannon, "A mathematical theory of communication", The Bell system technical journal **27**, 379–423 (1948).

[34]P. Kulkarni, T. L. Solomon, Y. He, Y. Chen, P. N. Bryan, and J. Orban, "Structural metamorphism and polymorphism in proteins on the brink of thermodynamic stability", Protein Science **27**, 1557–1567 (2018).

[35]S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy", Remote sensing of Environment **62**, 77–89 (1997).

[36]D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation", BMC genomics **21**, 1–13 (2020).

[37]L. L. Porter and G. D. Rose, "A thermodynamic definition of protein domains", Proceedings of the National Academy of Sciences **109**, 9420–9425 (2012).

[38]S. Mishra, L. L. Looger, and L. L. Porter, "Inaccurate secondary structure predictions often indicate protein fold switching", Protein Science **28**, 1487–1493 (2019).

[39]S. Mishra, L. L. Looger, and L. L. Porter, "A sequence-based method for predicting extant fold switchers that undergo α-helix↔ β-strand transitions", Biopolymers **112**, e23471 (2021).

[40]W. Jankowski, T. Saleh, M.-T. Pai, G. Sriram, R. B. Birge, and C. G. Kalodimos, "Domain organization differences explain bcr-abl's preference for crkl over crkii", Nature chemical biology **8**, 590–596 (2012).

[41]M. Harkiolaki, R. J. Gilbert, E. Y. Jones, and S. M. Feller, "The c-terminal sh3 domain of crkl as a dynamic dimerization module transiently exposing a nuclear export signal", Structure **14**, 1741–1753 (2006).

[42]N. Ferguson, J. Becker, H. Tidow, S. Tremmel, T. D. Sharpe, G. Krause, J. Flinders, M. Petrovich, J. Berriman, H. Oschkinat, et al., "General structural motifs of amyloid protofilaments", Proceedings of the National Academy of Sciences **103**, 16248–16253 (2006).

[43]R. Zhou, G. G. Maisuradze, D. Suñol, T. Todorovski, M. J. Macias, Y. Xiao, H. A. Scheraga, C. Czaplewski, and A. Liwo, "Folding kinetics of ww domains with the united residue force field for bridging microscopic motions and experimental measurements", Proceedings of the National Academy of Sciences **111**, 18243–18248 (2014).

[44]C. Lazareno-Saez, E. Arutyunova, N. Coquelle, and M. J. Lemieux, "Domain swapping in the cytoplasmic domain of the escherichia coli rhomboid protease", Journal of molecular biology **425**, 1127–1142 (2013).

[45]A. R. Sherratt, D. R. Blais, H. Ghasriani, J. P. Pezacki, and N. K. Goto, "Activity-based protein profiling of the escherichia coli glpg rhomboid protein delineates the catalytic core", Biochemistry **51**, 7794–7803 (2012).

[46]X. Zhao, M. Singh, V. N. Malashkevich, and P. S. Kim, "Structural characterization of the human respiratory syncytial virus fusion protein core", Proceedings of the National Academy of Sciences **97**, 14172–14177 (2000).

[47]A. Krarup, D. Truan, P. Furmanova-Hollenstein, L. Bogaert, P. Bouchier, I. J. Bisschop, M. N. Widjojoatmodjo, R. Zahn, H. Schuitemaker, J. S. McLellan, et al., "A highly stable prefusion rsv f vaccine derived from structural analysis of the fusion mechanism", Nature communications **6**, 8143 (2015).

[48]I. Jo, I.-Y. Chung, H.-W. Bae, J.-S. Kim, S. Song, Y.-H. Cho, and N.-C. Ha, "Structural details of the oxyr peroxide-sensing mechanism", Proceedings of the National Academy of Sciences **112**, 6443–6448 (2015).

[49]H. Ahmed, M. Blakeley, M. Cianci, D. Cruickshank, J. Hubbard, and J. Helliwell, "The determination of protonation states in proteins", Acta Crystallographica Section D: Biological Crystallography **63**, 906–922 (2007).

[50]A. C. Papageorgiou, H. S. Tranter, and K. R. Acharya, "Crystal structure of microbial superantigen staphylococcal enterotoxin b at 1.5 Å resolution: implications for superantigen recognition by mhc class ii molecules and t-cell receptors", Journal of molecular biology **277**, 61–79 (1998).

[51]C. S. Nagano, J. J. Calvete, D. Barettino, A. Pérez, B. S. Cavada, and L. Sanz, "Insights into the structural basis of the ph-dependent dimer–tetramer equilibrium through crystallographic analysis of recombinant diocleinae lectins", Biochemical Journal **409**, 417–428 (2008).

[52]E. A. Merritt, P. Kuhn, S. Sarfaty, J. L. Erbe, R. K. Holmes, and W. G. Hol, "The 1.25 Å resolution refinement of the cholera toxin b-pentamer: evidence of peptide backbone strain at the receptor-binding site", Journal of molecular biology **282**, 1043–1059 (1998).

[53]R. Yan, D. Xu, J. Yang, S. Walker, and Y. Zhang, "A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction", Scientific reports **3**, 2619 (2013).

A Supporting Information for Chapter 2: Quantum chemical study of a radical relay mechanism for the HydG-catalyzed synthesis of a Fe(II)$(CO)_2$(CN)cysteine precursor to the H-cluster of [FeFe] hydrogenase

A.1 Computational methods

Structural modeling. The initial HydG structure for the simulations was obtained from the published crystal structure, PDB ID 4WCX.[1] Herein, only Chain A of the protein dimer was employed in the study as it included the dangler iron. When preparing the simulation input structure, five missing residues in the N-terminal region were omitted and seven missing residues (345-351) were manually added in the middle of the chain. Because the SAM cofactor was missing from Chain A, we modeled its structure by substituting the SAM structure from Chain C for the corresponding methionine residue in chain A by superposition. The crystal structure included a free alanine (ALA) and sulfur atom (H2S) bonded to the auxiliary cluster, which we interpreted to be identical to the cysteine ligand to the dangler Fe from experimental evidence.[2] In order to add the missing tyrosine substrate, the crystal structure of the tryptophan lyase NosL (pdb ID: 4R34)[3] was used as a template to place the tyrosine substrate into HydG by superposition with the tryptophan substrate in NosL; the interactions between tyrosine and HydG were then optimized using a docking calculation.[4] The protonation of all residues was decided based on comparing the experimental pH value (7.0) with standard side chain pKa values, and His265 was

protonated as HID (Nδ is protonated) as Nϵ is coordinated to the dangler iron. Given the tyrosine substrate, two protonation states were considered in this study, shown in Scheme S1.

Details of the MM molecular dynamics simulations. Before starting any QM/MM simulations, the structure was relaxed by running classical molecular dynamics (MD) using molecular mechanics (MM) force fields in the AMBER software package.[5] The AMBER-FB15 protein force field[6] and TIP3P-FB water model[7] were used for the protein and water molecules in the system, the GAFF small molecule force field[8] was used to model the SAM cofactor, and the Fe-S clusters used a force field model published previously for MD simulations of [FeFe] hydrogenases.[9] These force fields are mutually compatible based on past studies that used the Fe-S cluster force field together with AMBER-family force fields,[10,11] and the high accuracy of AMBER-FB15 and TIP3P-FB when used together.[6]

The MM MD simulations used a simulation time step of 1 fs. Harmonic energy restraints were added to selected interatomic distances in order to ensure the force field does not change the coordination environment around the transition metal centers. The force constant was set to 50 kcal/mol/Å^2 and the restrained distances include the distances between the dangler iron and the atoms coordinated to it (two oxygen atoms from two water molecules, the O, N, and S atoms from the cysteine ligand, and the Nϵ in His265). The cutoff values for short-range electrostatics and van der Waals interactions were set to 12 Å, and the particle-mesh Ewald method was used for long-range summation of electrostatic interactions.[12] Covalent bond lengths involving hydrogen were constrained using the SHAKE algorithm.[13] A Langevin thermostat algorithm with a collision frequency of 1.0 ps^{-1} was employed for temperature control. In the simulation procedure, energy minimization was carried out first, followed by gradual heating from 0 K to 300 K using a 200 ps MD simulation at constant volume

(NVT), and this was followed by a 200 ps equilibration simulation at constant temperature and pressure (NPT) where a Berendsen barostat was added. Next, a 50 ns MD simulation was carried out under NVT conditions, and the final structure was used as the starting point of hybrid QM/MM simulations.

Hybrid QM/MM simulations at the canonical cluster. The QM/MM simulations were carried out using the Q-Chem and AMBER software packages.[5,14,15] Due to the complexity of the HydG catalytic reaction, different QM regions were chosen based on which reaction step was studied. Figure S1, the left panel shows the selection of QM regions for two reactions occurring at the rSAM Fe-S cluster; in these reaction steps, the broken symmetry approximation was used to model the high spin states and antiferromagnetic coupling of the Fe atoms. Standard electrostatic embedding was applied for the electrostatic interactions between the QM and MM regions, and a pseudo-bond and pseudo-atom approach[16] was used to treat the covalent bonds between the QM and MM regions.

The QM region was treated using density functional theory (DFT) using the unrestricted B3LYP density functional approximation, which we deemed appropriate for the canonical cluster as the reaction steps here involved mostly organic species. A hybrid basis set was used comprising the LANL2DZ basis set and pseudopotential for Fe atoms and the 6-31G* basis set for all other atoms. The choice of a relatively small basis was necessary in order to enable the QM/MM umbrella sampling described later, which involved running >10,000 serial individual calculations. To validate the accuracy of using this basis, we carried out potential energy scans using the larger def-TZVP basis; a comparison of energy profiles shows that the choice of basis set affects the barrier height by 0-2 kcal/mol (Figure S2).

The QM/MM free energy profiles were generated using an umbrella sampling approach where 15 ps of MD simulations were carried out at multiple windows along

the reaction coordinate. The cutoff values for the nonbonded interactions were set to 12 Å, and the time step was set to 1 fs. In order to compute QM/MM free energy profiles, an umbrella sampling approach was adopted. Multiple independent simulations were carried out corresponding to values of the chosen reaction coordinate; in this study, most of the spacing between simulations was set to 0.2 Å. Individual umbrella sampling runs were modified in order to maximize the thermodynamic overlap between windows while keeping the computational cost affordable; 20 windows were used for tyrosine radical generation, 45 windows for the 2-D umbrella sampling of tyrosine decomposition, 20 windows for DHG radical formation, and 12 windows for DHG decomposition.

The initial structures of each umbrella sampling window were determined using a series of constrained energy minimizations. A harmonic potential was added to each simulation to ensure the simulation trajectory remains close to the reaction coordinate. The umbrella sampling QM/MM MD simulations were carried out with an added harmonic potential to ensure the simulation trajectory remains close to the reaction coordinate. The force constants of the harmonic potentials were chosen according to the slope of the energy profile from the constrained minimizations and ranged from 5 to 80 kcal/mol/Å^2. QM/MM MD simulations were carried out for 15 ps for each window, and then the weighted histogram analysis method (WHAM) procedure[17,18] was used for data in the last 10 ps to determine the free energy profile from the biased trajectories. This procedure produces relative free energies with a statistical error on the order of 1 kcal/mol.

Cluster model calculations at the auxiliary cluster. The reactions at the auxiliary cluster required a greater number of atoms, electrons, and basis functions to be treated simultaneously at the QM level. Because this increased the computational cost significantly, we could not carry out QM/MM umbrella sampling calculations for

these reaction steps, and instead used a cluster model that included the atoms shown in blue in Figure S1, right panel. In this model, the dangler Fe is coordinated to the tridentate cysteine ligand, 5-methylimidazole (as a model for the His265 side chain), and two water molecules. The cysteine S atom bridges the dangler Fe and the Fe_4S_4 auxiliary cluster, and coordination of cysteine residues to the other three Fe atoms are modeled as MeS (methanethiol) groups. To initiate the catalytic cycle, the initial equivalent of $COOH^{\bullet}$ and CN^{-} ligands are placed in close proximity to the dangler iron. Figure S3 indicates that the dangler Fe is adjacent to a large cavity in the TIM barrel normally occupied by water molecules, indicating there is sufficient space for the movement and reorientation of ligands.

The QM calculations at the auxiliary cluster were carried out using the TeraChem package,[19,20] which includes graphics processing unit (GPU)-accelerated implementations of density functional theory (DFT) and implicit solvent models. During geometry optimizations, the B3LYP functional and mixed LANL2DZ ECP/6-31G* basis set was adopted, the same as the QM/MM calculations. A few optimizations focusing on spin crossover employed the B3LYP* functional[21] instead, which reduced the percentage of Hartree-Fock exchange from 20% to 15% and improved accuracy for spin crossover enthalpies of iron-containing complexes.[22] After optimizing the geometries, single point energies were computed along the minimum energy pathway to further improve accuracy; these employed a hybrid functional that combines 5% HF / 95% B88 exchange and P86 correlation, here called BP86x5,[23] and a larger triple-ζ basis set called ma-def2-TZVP(-f)_LTZ+[24]. This basis combines def2-TZVP[25] with $l \geq 3$ basis functions removed for non-Fe atoms, augmented by a minimal set of diffuse functions[26], and the LANL2TZ+ ECP/basis set for Fe atoms.[27] Empirical dispersion corrections of the D3(BJ) form were used, adopting the model parameters developed for BP86.[28] A switching Gaussian polarizable continuum model[29–31] was used with

standard Bondi radii and a dielectric constant of 78.4 (equivalent to water), because we observed the auxiliary cluster in our MM MD simulations to be solvated by 15-20 water molecules.

Equilibrium geometries and transition states were optimized using the geomeTRIC software package,[32] which uses a translation-rotation internal coordinate system to efficiently optimize the geometries of multi-molecular systems. After optimization of the transition states, an approximate minimum energy path was obtained by minimizing the energy starting from the TS structure along the imaginary mode with step sizes restricted to < 0.01Å. The energy corrections for the TS were obtained the IRCMax approach[33] by tracing over this path with single-point calculations at the BP86x5/ma-def2-TZVP-f_LTZ+ level of theory and choosing the highest energy on the path. The Gibbs free energies of these reaction steps were estimated using vibrational analysis carried out at the reaction endpoints and the transition state and applying the rigid rotor/harmonic oscillator approximation. Although these free energy corrections model the translational entropy using an ideal gas, which is highly approximate and for which numerous corrections have been proposed,[34–36] we did not find the role of translational entropy to be significant in any reaction steps studied here.

In order to compute standard redox potentials corresponding to electron transfer (ET) and proton-coupled electron transfer (PCET) steps, the free energy change of the reaction was computed in solution. The free energy of the proton at pH 0.0 in aqueous solution was taken to be -11.803 eV, following previous studies,[37,38] and adjusted to -11.390 eV by adding 59 meV per pH unit according to the Nernst equation. The redox potential was then computed as:

$$E^\circ = \frac{\Delta G(\text{reduced - oxidized})}{nF} - 4.43 \text{ V} \tag{6}$$

where n is the number of electrons transferred (1 in this study), F is the Faraday constant (1 eV V^{-1}), and 4.43 V is the absolute potential of the standard hydrogen electrode.[39] In reactions where a chemical reducing agent (dithionite) was used, the overpotential was computed as $\eta = -(E^{\circ} - 0.66V)$ and the resulting value is converted back to a free energy difference in kcal/mol, allowing for the inclusion of electrochemical steps on a reaction free energy diagram.[40]

EPR properties, in particular the g-tensor and hyperfine tensor (A) eigenvalues, were computed for the structures **5a**, **10** and **11** using the ORCA software package.[41] These calculations used the BP86x5 functional (same as the energies) with the zeroth-order regular relativistic approximation (ZORA) Hamiltonian[42], and a mixed basis set consisting of EPR-III[43] for first-row elements and "ZORA-def2-TZVP" for Fe and S, a reconstructed version of def2-TZVP[44] for ZORA calculations. The RIJCOSX method consisting of density fitting for Coulomb and chain-of-spheres approximation for exchange integrals was employed to speed up the calculations.[45] The g-tensor was computed using a coupled-perturbed SCF approach[46] and the hyperfine values were computed from spin-orbit couplings[47]. The 3D structure of **5a** was used as-is, whereas the **10** and **11** were modified by removing the spectator CH_3SH and HCN ligands (respectively) and re-optimizing the structure at the BP86x5/def2-SV(P) level of theory.

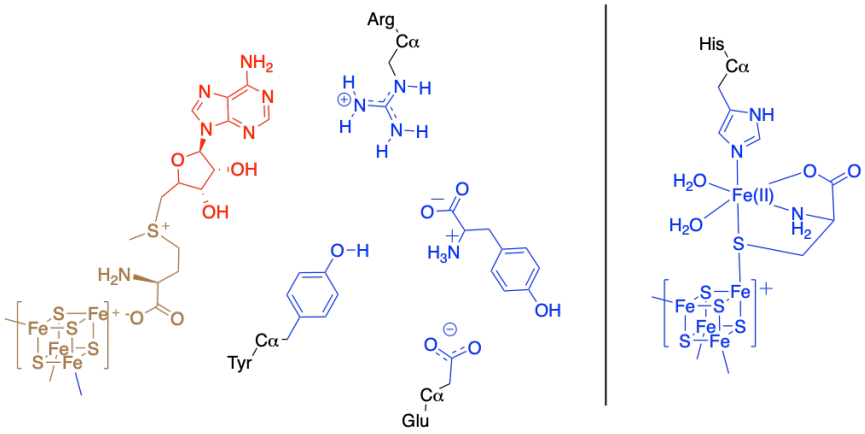


Figure S1: QM regions used for various elementary steps at the rSAM and auxiliary clusters. Left: Atoms shown in brown or red comprise the QM/MM region for the SAM decomposition step, and atoms shown in blue or red comprise the QM/MM region for other reaction steps in the canonical cluster. Right: Atoms in blue comprise the auxiliary cluster model.

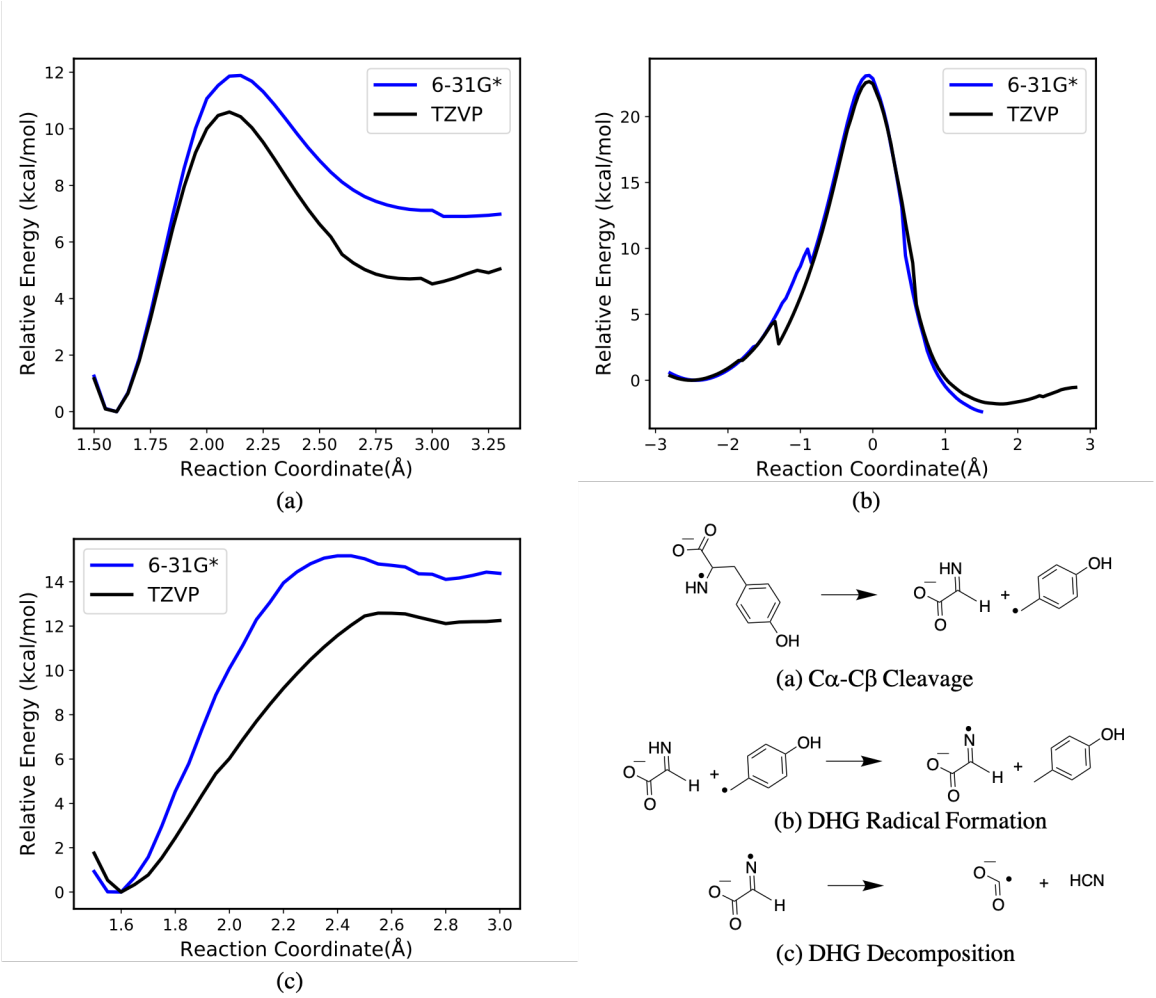


Figure S2: The basis set dependence of the three reactions studied using QM/MM, corresponding to Figures 2-4 in the main text, is investigated by driving the reaction coordinate. The larger def-TZVP basis set (blue curve) predicts barrier heights that are 2 kcal/mol lower for both the Cα-Cβ cleavage (top left) and DHG decomposition reactions (bottom left) compared to the 6-31G* basis used in QM/MM umbrella sampling studies. By contrast, the two basis sets give nearly identical energy profiles in DHG radical formation (top right). Schemes of the reactions are shown at the bottom right.

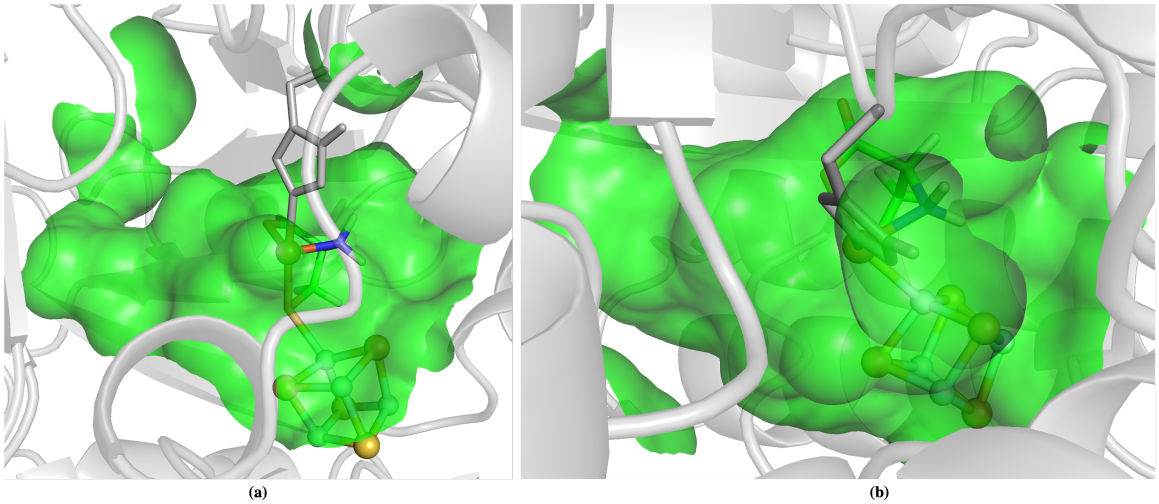


Figure S3: The position of the auxiliary cluster and dangler Fe at the end of the HydG TIM barrel. The green surface indicates the interior space of the barrel indicating sufficient space for ligand substitutions to be carried out. The solvent-accessible surface is drawn with a probe radius of 1.4Å.

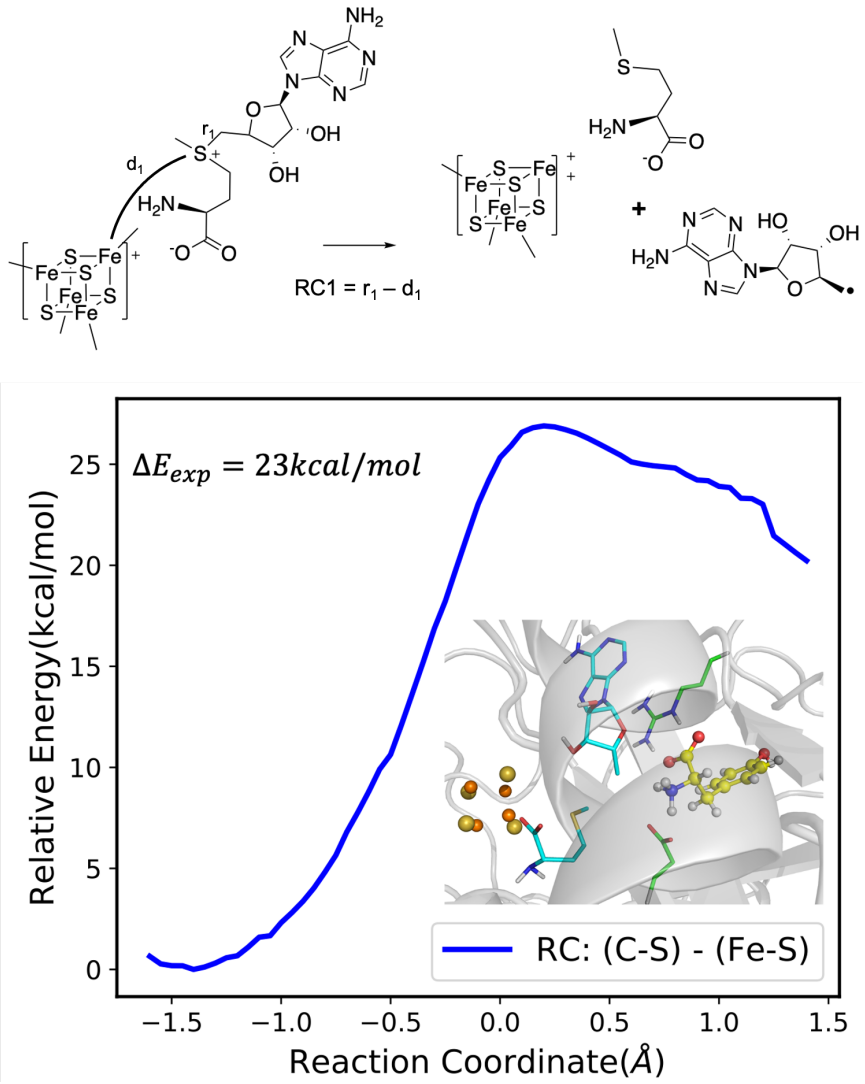


Figure S4: Energy profile of 5-Ado radical generation, generated by driving the reaction coordinate, defined as RC1 = d(C…S) – d(Fe…S). Here the activation energy is ≈ 26.3 kcal/mol, which is comparable to the experimental data (23 kcal/mol).

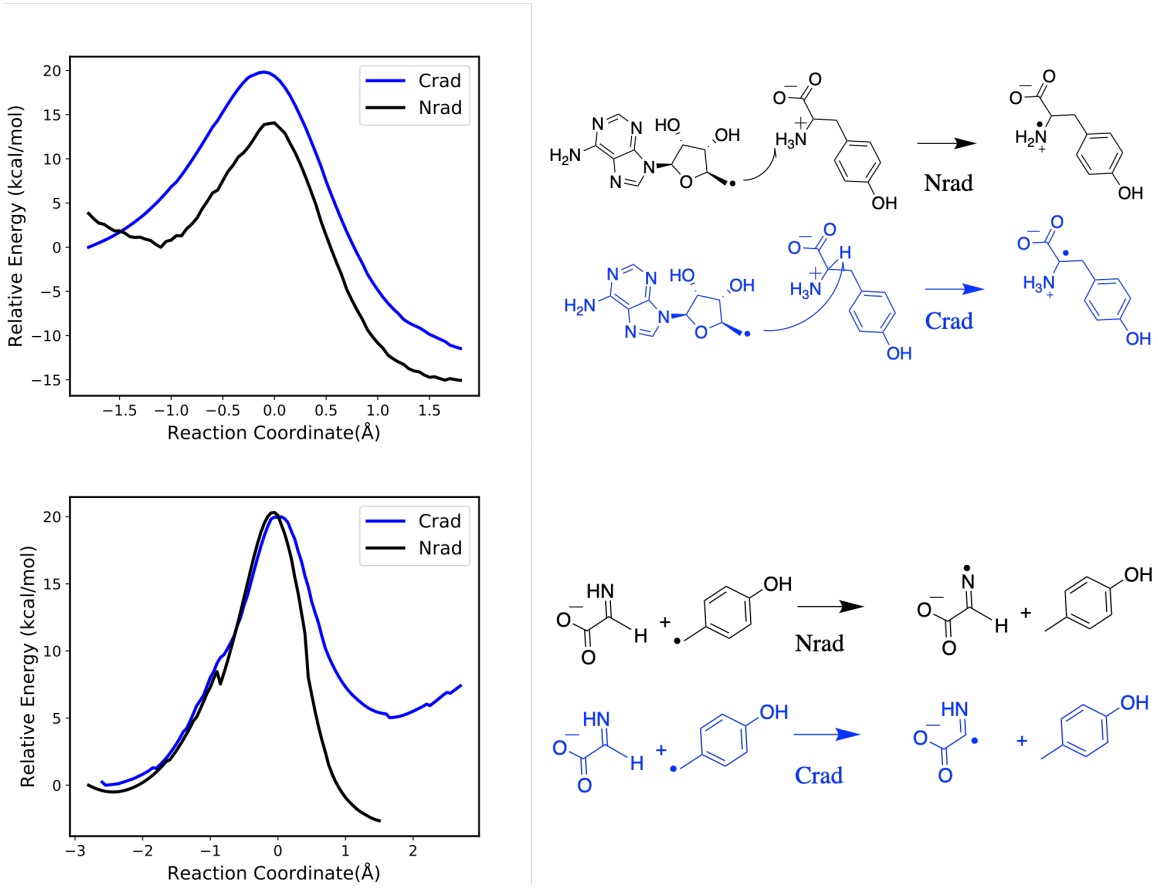
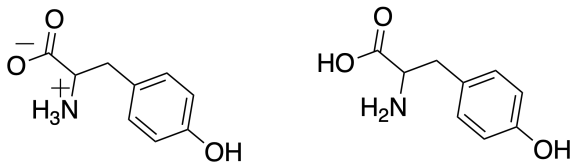


Figure S5: Possible alternative pathways in tyrosine radical formation and the DHG radical formation, (the main mechanism is Figure 2 and 4 in the main text respectively). As for the tyrosine radical formation, the experimental data supports the mechanism that the hydrogen abstraction occurred in the amine hydrogen. Here our calculations also support this mechanism since the other possible H resource, which is the Cα, cannot transfer its hydrogen with an energy barrier that as low as the amino group. Regarding the DHG radical formation, the abstraction of the H from amino group gives a more stable product than that of the H connected to the sp3 hybridization carbon.



TYH **TYY**

Scheme 2: Possible protonation states of the tyrosine substrate.

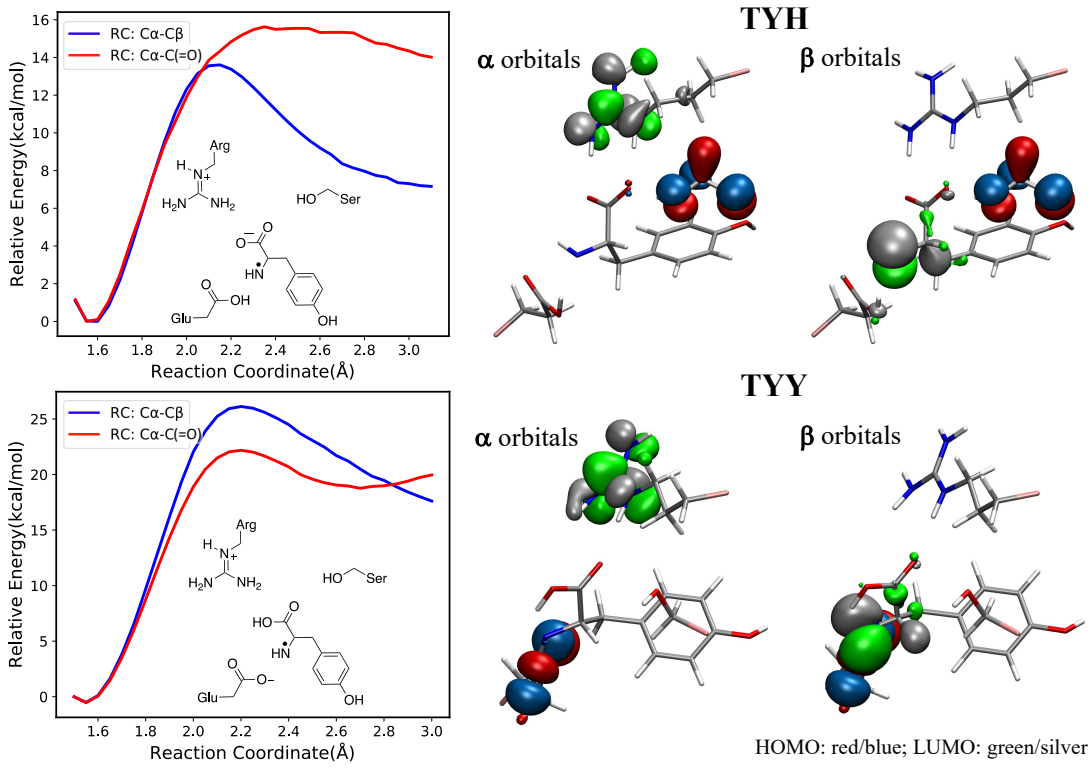


Figure S6: Comparison of alternative tyrosine decomposition mechanisms (the main mechanism is Figure 3 in the main text). Two possible protonation states of tyrosine, denoted TYH and TYY, are shown on the top and bottom respectively. To facilitate direct comparison, we assume the radical $NH2^{\bullet +}$ transfers a proton to the Glu side chain prior to C—C cleavage, which could also be observed in our QM/MM simulations. The energy profiles of the two C—C cleavage mechanisms for each protonation state are shown on the left; blue curves represent the Cα—Cβ cleavage and red curves represent the Cα—C(=O) cleavage. The frontier orbitals of these two tyrosine models are shown on the right. In the TYH model, the LUMO of the β electron is shared between the $NH^{\bullet}$ and the Cα—Cβ σ bond, indicating that the H-atom abstraction may have the effect of weakening the Cα—Cβ bond order. This correlates with the lowered barrier of Cα—Cβ cleavage in TYH and is not observed in TYY, where the barrier to Cα—Cβ cleavage is significantly higher.

Figure S7: Several possible pathways of the reduction of COOH radical to CO in the canonical rSAM Fe_4S_4 cluster pocket. All of the activation energies are in excess of 30 kcal/mol, which indicates that COOH does not decompose in the cluster pocket. Instead, we propose that $COOH^{\bullet}$ diffuses to the auxiliary cluster where reduction to CO occurs at the dangler iron.

State Label	Charge	Spin Mult.	Fe_4S_4 State	Dangler Fe State	N($COOH^{\bullet}$)	N(CN^-)
1	-3	5	+	Fe(II)(Cys)(5-MIm) $(H_2O)_2$	1	1
2	-3	1	+	Fe(II)(Cys)(5-MIm) (CN)(H_2O)	1	1
3	-3	1	2+	Fe(II)(Cys)(5-MIm) (CN)(COOH)	1	1
4	-2	1	2+	Fe(II)($CysH^+$)(5-MIm) (CN)(COOH)	1	1
5	-2	1	2+	Fe(II)(Cys)(5-MIm) (CN)(CO)	1	1
4a	-3	2	+	Fe(II)($CysH^+$)(5-MIm) (CN)(COOH)	1	1
5a	-3	2	+	Fe(II)(Cys)(5-MIm) (CN)(CO)	1	1
6	-3	1	+	Fe(III)(Cys)(5-MIm) (CN)(COCOOH)	2	1
7	-2	1	2+	Fe(II)($CysH^+$)(5-MIm) (CN)(COCOOH)	2	1
8	-2	1	2+	Fe(II)(Cys)(5-MIm) (CN)(CO)	2	1
7a	-3	2	+	Fe(II)($CysH^+$)(5-MIm) (CN)(COCOOH)	2	1
8a	-3	2	+	Fe(II)(Cys)(5-MIm) (CN)(CO)	2	1
9	-3	2	+	Fe(II)(Cys) (CN)$(CO)_2$	2	1
10	-4	2	+	Fe(II)(Cys) (CN)$(CO)_2$	2	2
10.S	-1	1	N/A	Fe(II)(Cys) (CN)$(CO)_2$	2	1
10.C	-3	2	+	N/A	0	1
11	-3	2	+	N/A	0	1

Table S1: The key properties of the various calculated states in the catalytic cycle at the auxiliary cluster. The state labels correspond to Fig. 5 and Fig. 6 in the main text. The charge and spin multiplicity given is for the whole calculated system, including any added $COOH^{\bullet}$ and CN^- ligands. The state labeled 10 is actually two calculations consisting of a synthon fragment and Fe_4S_4—CN fragment labeled 10.S and 10.C respectively.

	$Fe(H_2O)_2$	$Fe(CO)(H_2O)$	$Fe(CN)(H_2O)$	$Fe(CO)(CN)$
Active Space	(20, 18)	(20, 18)	(20, 18)	(22, 20)
E(HS)-E(LS)	-68.2	28.3	25.1	50.2

Table S2: The energy differences between different iron cluster structures calculated by density matrix renormalization group (DMRG).

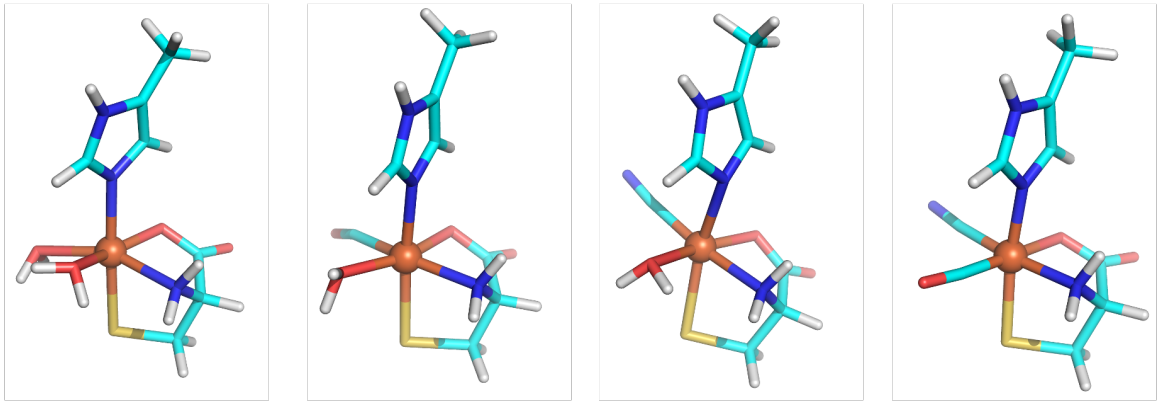


Figure S8: Structures of the Fe(cys)(5-MIm) complex used in the multireference DMRG studies.

Multireference DMRG calculation methods. All four structures were optimized under B3LYP//6-31G*/LANL2DZ(Fe) in Q-Chem, and the DMRG single-point energies were calculated using the TZV basis set in PySCF. The active spaces in the calculations were selected from localized molecular orbitals by including the five orbitals with $3d$ character on the Fe atom and the orbitals with p character on the coordinating ligands ($2p$ for O, C, N and $3p$ for S) that had an overlap integral of > 0.3 with any of the d orbitals; the orbital selections were then confirmed by visual inspection of isosurface plots. The total number of orbitals and electrons in the DMRG calculations is provided in the table above as (active electrons, active orbitals). The bond dimension of the DMRG calculations is set to 1000, and the number of sweeps is set to the default value of 4. Energies are in units of kcal/mol.

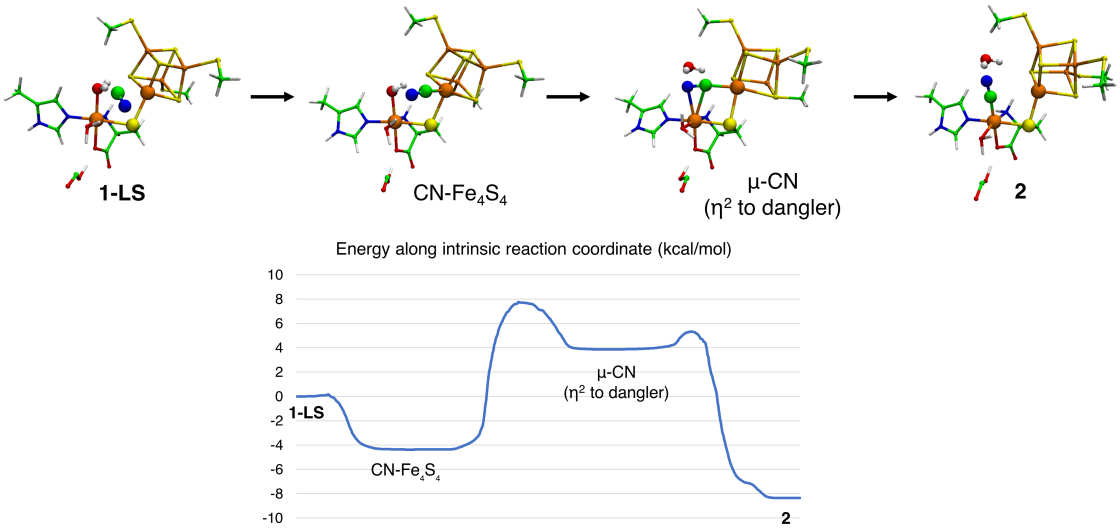


Figure S9: 3D renderings and plot of electronic energies (`B3LYP/6-31G*_LANL2DZ/PCM`) along the minimum energy path of the reaction **1-LS** + CN^- → **2** + H_2O where a CN^- ligand displaces an aquo ligand. A spectator $COOH^\bullet$ species is present in the system.

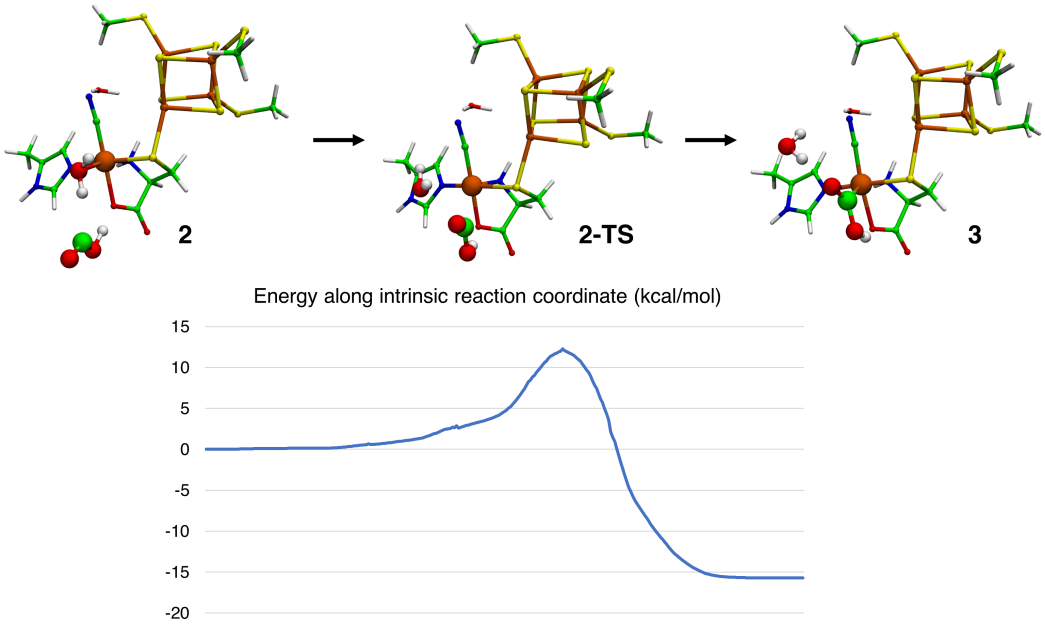


Figure S10: 3D renderings and plot of electronic energies (B3LYP/6-31G*_LANL2DZ/PCM) along the minimum energy path of the reaction **2-LS** + $COOH^{\bullet}$ $\rightarrow$ **3** + H_2O where $COOH^{\bullet}$ displaces the second aquo ligand. Following this, **3** is protonated to form **4**.

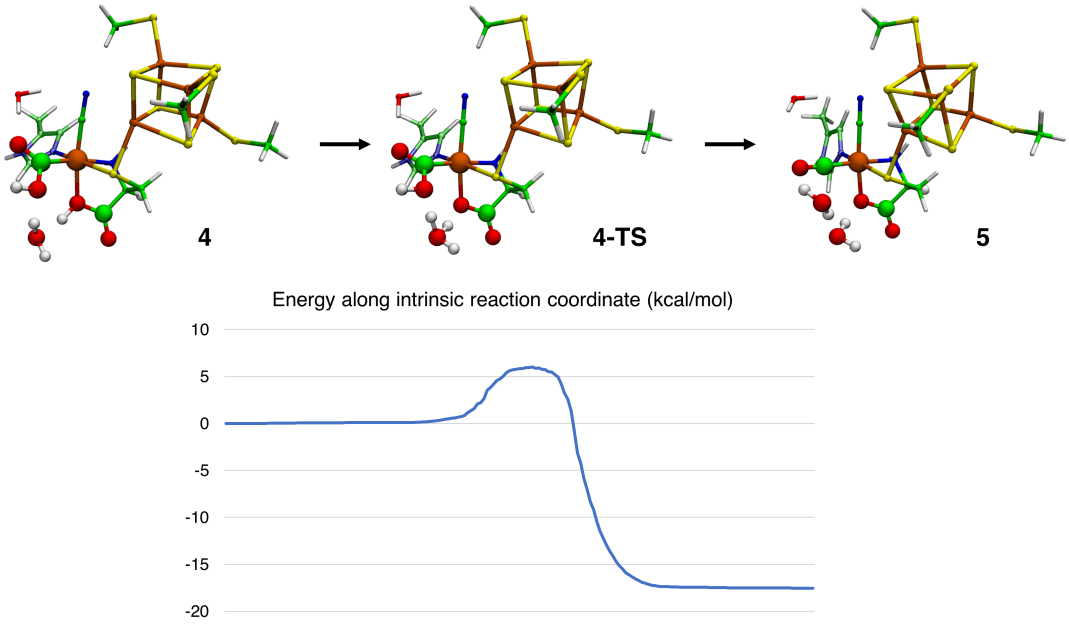


Figure S11: 3D renderings and plot of electronic energies (B3LYP/6-31G*_LANL2DZ/PCM) along the minimum energy path of the reaction **4** $\rightarrow$ **5** + H_2O where the COOH ligand accepts a proton from the cysteine oxygen, then is decomposed to CO and H_2O.

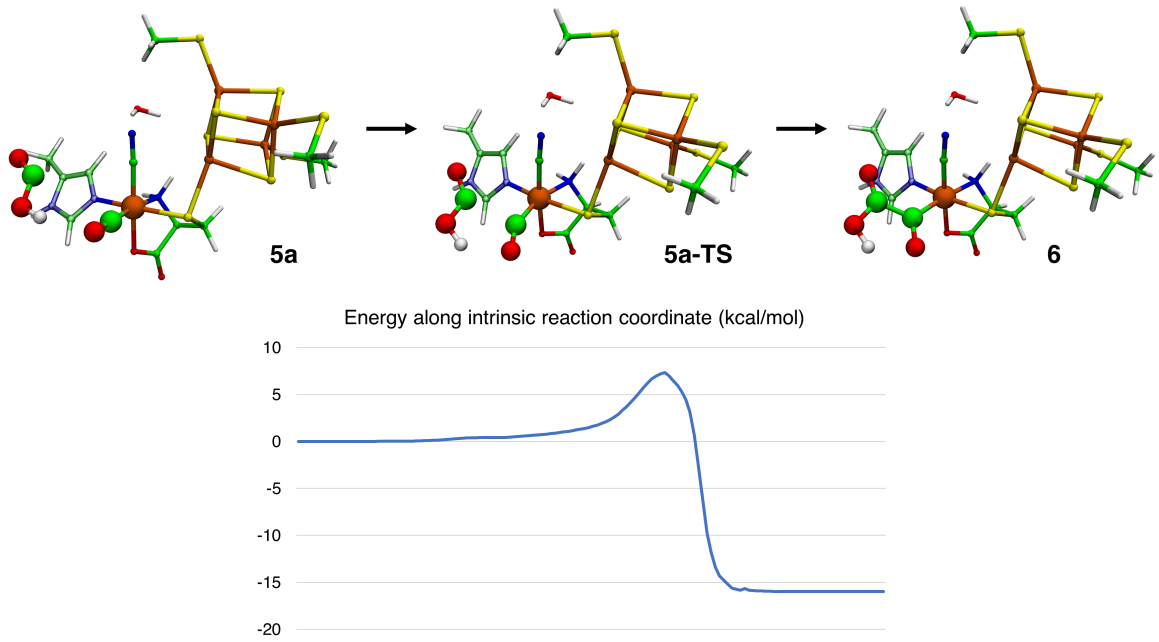


Figure S12: 3D renderings and plot of electronic energies (B3LYP/6-31G*_LANL2DZ/PCM) along the minimum energy path of the reaction **5a** + $COOH^{\bullet}$ → **6** where the added $COOH^{\bullet}$ forms a C—C bond with the first CO ligand to form a COCOOH ligand to the dangler Fe. Following this, **6** is protonated to form **7**.

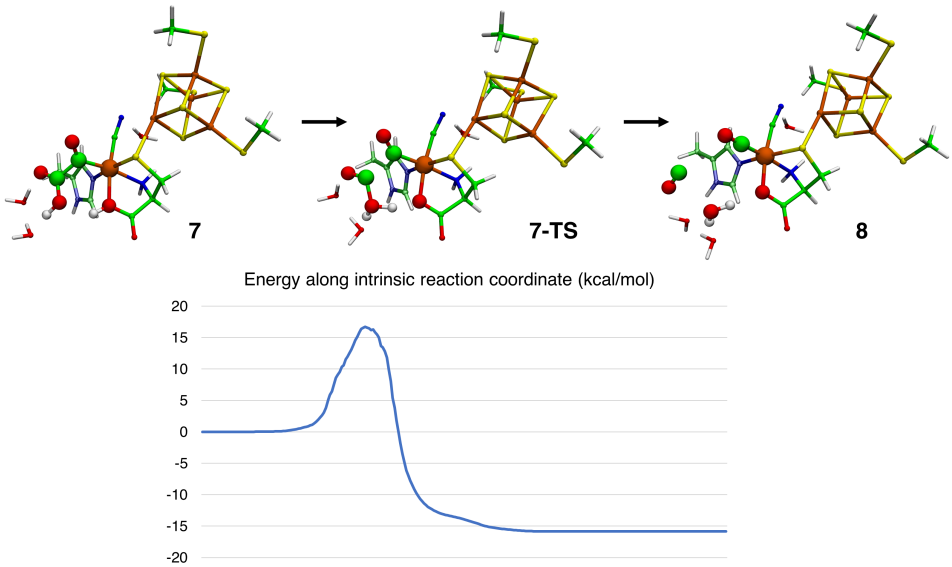


Figure S13: 3D renderings and plot of electronic energies (B3LYP/6-31G*_LANL2DZ/PCM) along the minimum energy path of the reaction **7** $\rightarrow$ **8** + CO + H_2O where the COCOOH ligand accepts a proton from the cysteine oxygen, then is decomposed to a free CO + CO ligand + H_2O.

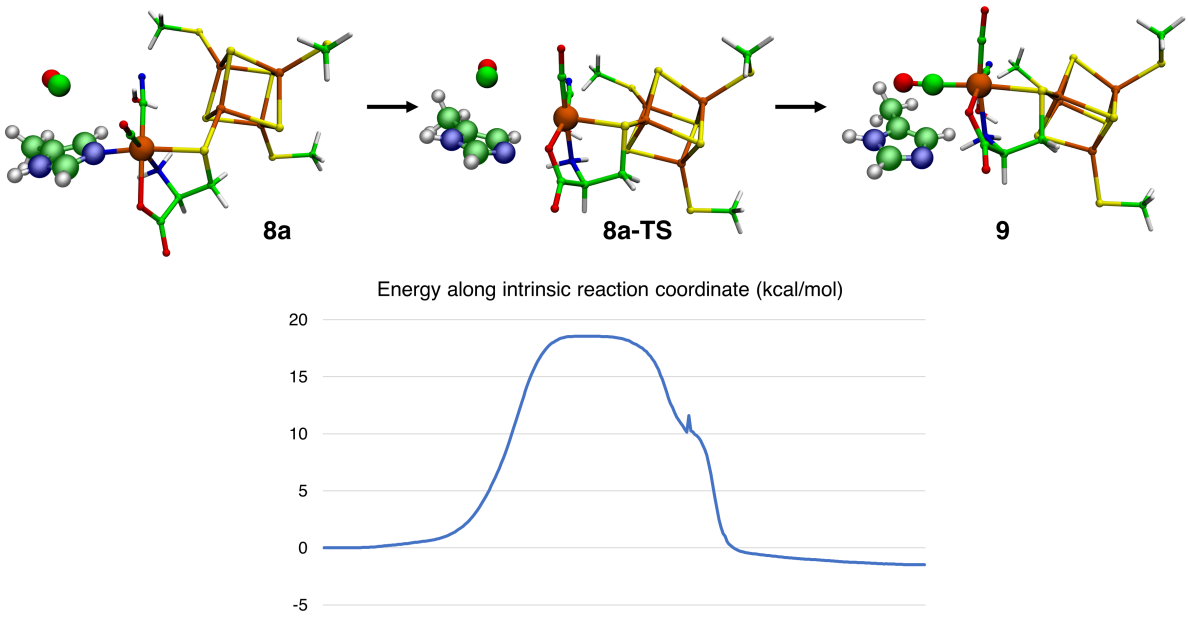


Figure S14: 3D renderings and plot of electronic energies (B3LYP/6-31G*_LANL2DZ/PCM) along the minimum energy path of the reaction **8** + CO → **9** + 5-MIm where the second CO displaces the 5-methylimidazole ligand.

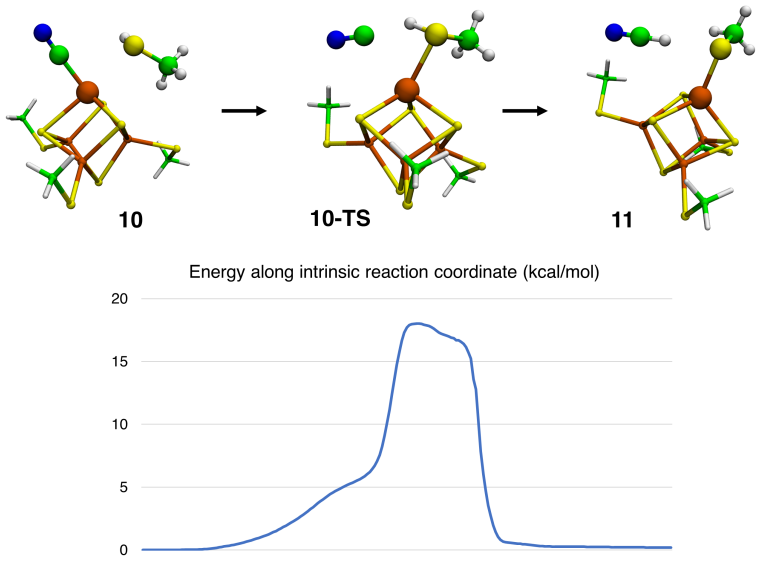


Figure S15: 3D renderings and plot of electronic energies (`B3LYP/6-31G*_LANL2DZ/PCM`) along the minimum energy path of the reaction **10** + CH_3SH → **11** + HCN where a CH_3SH model of the cysteine side chain displaces the CN ligand to the auxiliary cluster, releasing HCN and turning over the catalytic cycle.

References

[1]P. Dinis, D. L. Suess, S. J. Fox, J. E. Harmer, R. C. Driesener, L. De La Paz, J. R. Swartz, J. W. Essex, R. D. Britt, and P. L. Roach, "X-ray crystallographic and epr spectroscopic analysis of hydg, a maturase in [fefe]-hydrogenase h-cluster assembly", Proceedings of the National Academy of Sciences **112**, 1362–1367 (2015).

[2]D. L. Suess, I. Bürstel, L. De La Paz, J. M. Kuchenreuther, C. C. Pham, S. P. Cramer, J. R. Swartz, and R. D. Britt, "Cysteine as a ligand platform in the biosynthesis of the fefe hydrogenase h cluster", Proceedings of the National Academy of Sciences **112**, 11455–11460 (2015).

[3]Y. Nicolet, L. Zeppieri, P. Amara, and J. C. Fontecilla-Camps, "Crystal structure of tryptophan lyase (nosl): evidence for radical formation at the amino group of tryptophan", Angewandte Chemie International Edition **53**, 11840–11844 (2014).

[4]G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "Autodock4 and autodocktools4: automated docking with selective receptor flexibility", Journal of computational chemistry **30**, 2785–2791 (2009).

[5]D. Case, K. Belfon, I. Ben-Shalom, S. Brozell, D. Cerutti, I. T.E. Cheatham, V. Cruzeiro, T. Darden, R. Duke, G. Giambasu, M. Gilson, H. Gohlke, A. Goetz, R. Harris, S. Izadi, S. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R. Walker, J. Wang, L. Wilson, R. Wolf, X. Wu, Y. Xiong, Y. Xue, D. York, and P. Kollman, Amber 2020, ucsf.

[6]L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martınez, and V. S. Pande, "Building a more predictive protein force field: a systematic and reproducible route to amber-fb15", The Journal of Physical Chemistry B **121**, 4023–4039 (2017).

[7]L.-P. Wang, T. J. Martinez, and V. S. Pande, "Building force fields: an automatic, systematic, and reproducible approach", The Journal of Physical Chemistry Letters **5**, 1885–1891 (2014).

[8]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field", Journal of Computational Chemistry **25**, 1157–1174 (2004).

[9]C. H. Chang and K. Kim, "Density functional theory calculation of bonding and charge parameters for molecular dynamics studies on [fefe] hydrogenases", Journal of chemical theory and computation **5**, 1137–1145 (2009).

[10]H. Long, P. W. King, and C. H. Chang, "Proton transport in clostridium pasteurianum [fefe] hydrogenase i: a computational study", The Journal of Physical Chemistry B **118**, 890–900 (2014).

[11]A. Kubas, C. Orain, D. De Sancho, L. Saujet, M. Sensi, C. Gauquelin, I. Meynial-Salles, P. Soucaille, H. Bottin, C. Baffert, et al., "Mechanism of o 2 diffusion and reduction in fefe hydrogenases", Nature chemistry **9**, 88–95 (2017).

[12]U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh ewald method", The Journal of Chemical Physics **103**, 8577–8593 (1995).

[13]H. C. Andersen, "Rattle: a "velocity" version of the shake algorithm for molecular dynamics calculations", Journal of Computational Physics **52**, 24–34 (1983).

[14]Y. Shao, Z. Gan, E. Epifanovsky, A. T. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kuś, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. W. III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. D. Jr., H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. Hanson-Heine, P. H. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowal-

czyk, C. M. Krauter, K. U. Lao, A. D. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. Thom, T. Tsuchimochi, V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. Chan, D. M. Chipman, C. J. Cramer, W. A. G. III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. S. III, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. V. Voorhis, J. M. Herbert, A. I. Krylov, P. M. Gill, and M. Head-Gordon, "Advances in molecular quantum chemistry contained in the q-chem 4 program package", Molecular Physics **113**, 184–215 (2015).

[15]Y. Zhou, S. Wang, Y. Li, and Y. Zhang, "Born–oppenheimer ab initio qm/mm molecular dynamics simulations of enzyme reactions", in Methods in enzymology, Vol. 577 (Elsevier, 2016), pp. 105–118.

[16]Y. Zhang, "Pseudobond ab initio qm/mm approach and its applications to enzyme reactions", Theoretical Chemistry Accounts **116**, 43–50 (2006).

[17]B. Roux, "The calculation of the potential of mean force using computer simulations", Computer physics communications **91**, 275–282 (1995).

[18]A. Grossfield, Wham: an implementation of the weighted histogram analysis method.

[19]S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, S. I. L. Kokkila-Schumacher, N. Luehr, J. W. Snyder, C. Song, A. V. Titov, I. S. Ufimtsev, and T. J. Martínez, "Terachem: accelerating electronic structure and ab initio molecular dynamics with graphical processing units", The Journal of Chemical Physics **152**, 224110 (2020).

[20]S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, C. M. Isborn, S. I. L. Kokkila-Schumacher, X. Li, F. Liu, N. Luehr, J. W. Snyder Jr., C. Song, A. V. Titov, I. S. Ufimtsev, L.-P. Wang, and T. J. Martínez, "Terachem: a graphical processing unit-accelerated electronic structure package for large-scale ab initio molecular dynamics", WIREs Computational Molecular Science **n/a**, e1494.

[21]O. Salomon, M. Reiher, and B. A. Hess, "Assertion and validation of the performance of the b3lyp* functional for the first transition metal row and the g2 test set", The Journal of Chemical Physics **117**, 4729–4737 (2002).

[22]O. S. Siig and K. P. Kepp, "Iron(ii) and iron(iii) spin crossover: toward an optimal density functional", The Journal of Physical Chemistry A **122**, 4208–4217 (2018).

[23]R. K. Szilagyi and M. A. Winslow, "On the accuracy of density functional theory for iron—sulfur clusters", Journal of Computational Chemistry **27**, 1385–1397 (2006).

[24]H. Jang, Y. Qiu, M. E. Hutchings, M. Nguyen, L. A. Berben, and L.-P. Wang, "Quantum chemical studies of redox properties and conformational changes of a four-center iron co2 reduction electrocatalyst", Chem. Sci. **9**, 2645–2654 (2018).

[25]F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: design and assessment of accuracy", Phys. Chem. Chem. Phys. **7**, 3297–3305 (2005).

[26]J. Zheng, X. Xu, and D. G. Truhlar, "Minimally augmented karlsruhe basis sets", Theoretical Chemistry Accounts **128**, 295–305 (2011).

[27]L. E. Roy, P. J. Hay, and R. L. Martin, "Revised basis sets for the lanl effective core potentials", Journal of Chemical Theory and Computation **4**, 1029–1031 (2008).

[28]S. Grimme, S. Ehrlich, and L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory", Journal of Computational Chemistry **32**, 1456–1465 (2011).

[29]M. Cossi, N. Rega, G. Scalmani, and V. Barone, "Energies, structures, and electronic properties of molecules in solution with the c-pcm solvation model", Journal of Computational Chemistry **24**, 669–681 (2003).

[30]A. W. Lange and J. M. Herbert, "A smooth, nonsingular, and faithful discretization scheme for polarizable continuum models: the switching/gaussian approach", The Journal of Chemical Physics **133**, 244111 (2010).

[31]F. Liu, N. Luehr, H. J. Kulik, and T. J. Martínez, "Quantum chemistry for solvated molecules on graphical processing units using polarizable continuum models", Journal of Chemical Theory and Computation **11**, 3131–3144 (2015).

[32]L.-P. Wang and C. Song, "Geometry optimization made simple with translation and rotation coordinates", The Journal of Chemical Physics **144**, 214108 (2016).

[33]D. K. Malick, G. A. Petersson, and J. A. Montgomery, "Transition states for chemical reactions i. geometry and classical barrier height", The Journal of Chemical Physics **108**, 5704–5713 (1998).

[34]M. Mammen, E. I. Shakhnovich, J. M. Deutch, and G. M. Whitesides, "Estimating the entropic cost of self-assembly of multiparticle hydrogen-bonded aggregates based on the cyanuric acid · melamine lattice", The Journal of Organic Chemistry **63**, 3821–3830 (1998).

[35]R. E. Plata and D. A. Singleton, "A case study of the mechanism of alcohol-mediated morita baylis–hillman reactions. the importance of experimental observations", Journal of the American Chemical Society **137**, 3811–3826 (2015).

[36]V. P. Tuguldurova, A. V. Fateev, O. K. Poleshchuk, and O. V. Vodyankina, "Theoretical analysis of glyoxal condensation with ammonia in aqueous solution", Phys. Chem. Chem. Phys. **21**, 9326–9334 (2019).

[37]M. D. Tissandier, K. A. Cowen, W. Y. Feng, E. Gundlach, M. H. Cohen, A. D. Earhart, J. V. Coe, and T. R. Tuttle, "The proton's absolute aqueous enthalpy and gibbs free energy of solvation from cluster-ion solvation data", The Journal of Physical Chemistry A **102**, 7787–7794 (1998).

[38]L.-P. Wang, Q. Wu, and T. Van Voorhis, "Acid-base mechanism for ruthenium water oxidation catalysts", Inorganic Chemistry **49**, 4543–4553 (2010).

[39]H. Reiss and A. Heller, "The absolute potential of the standard hydrogen electrode: a new estimate", The Journal of Physical Chemistry **89**, 4207–4213 (1985).

[40]K. S. Exner, I. Sohrabnejad-Eskan, and H. Over, "A Universal Approach To Determine the Free Energy Diagram of an Electrocatalytic Reaction", ACS Catalysis **8**, Publisher: American Chemical Society, 1864–1879 (2018).

[41]F. Neese, F. Wennmohs, U. Becker, and C. Riplinger, "The orca quantum chemistry program package", The Journal of Chemical Physics **152**, 224108 (2020).

[42]C. van Wullen, "Molecular density functional calculations in the regular relativistic approximation: method, application to coinage metal diatomics, hydrides, fluorides and chlorides, and comparison with first-order relativistic calculations", The Journal of Chemical Physics **109**, 392–399 (1998).

[43]V. Barone, A. Bencini, and P. Fantucci, Recent advances in density functional methods (WORLD SCIENTIFIC, 2002).

[44]F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: design and assessment of accuracy", Phys. Chem. Chem. Phys. **7**, 3297–3305 (2005).

[45]R. Izsak and F. Neese, "An overlap fitted chain of spheres exchange method", The Journal of Chemical Physics **135**, 144105 (2011).

[46]F. Neese, "Prediction of electron paramagnetic resonance g values using coupled perturbed hartree–fock and kohn–sham theory", The Journal of Chemical Physics **115**, 11080–11096 (2001).

[47]F. Neese, "Metal and ligand hyperfine couplings in transition metal complexes: the effect of spin–orbit coupling as studied by coupled perturbed kohn–sham theory", The Journal of Chemical Physics **118**, 3939–3948 (2003).

B Supporting Information for Chapter 3: How does HydE work? A Comparison Between A Radical Mechanism and A Proton-Transfer Mechanism

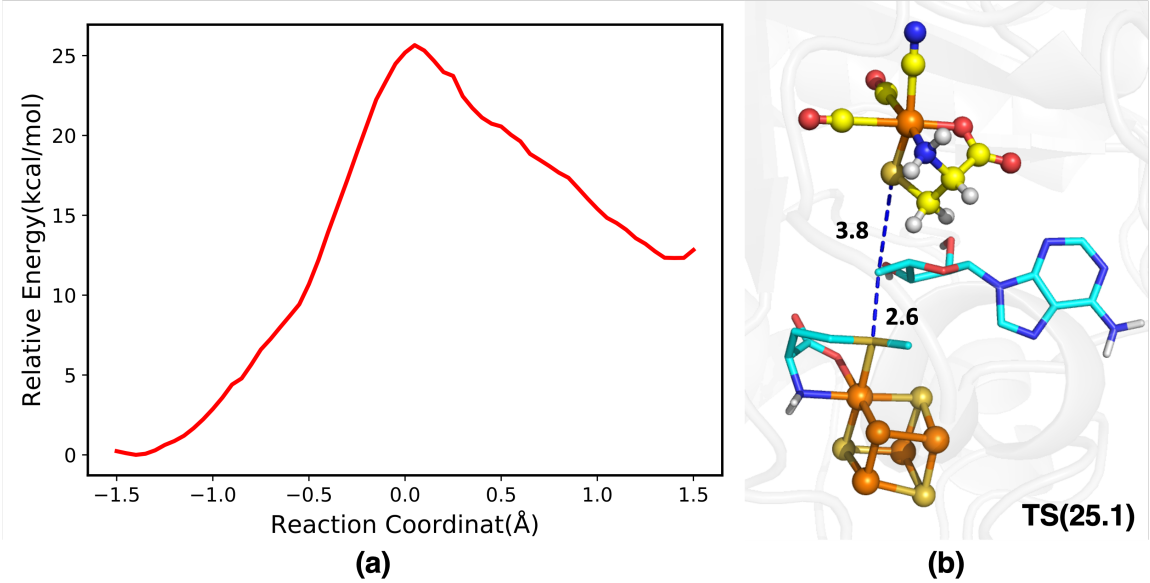


Figure S1: The relative energy profile of the SAM decomposition (a) with its corresponding transition state (b).

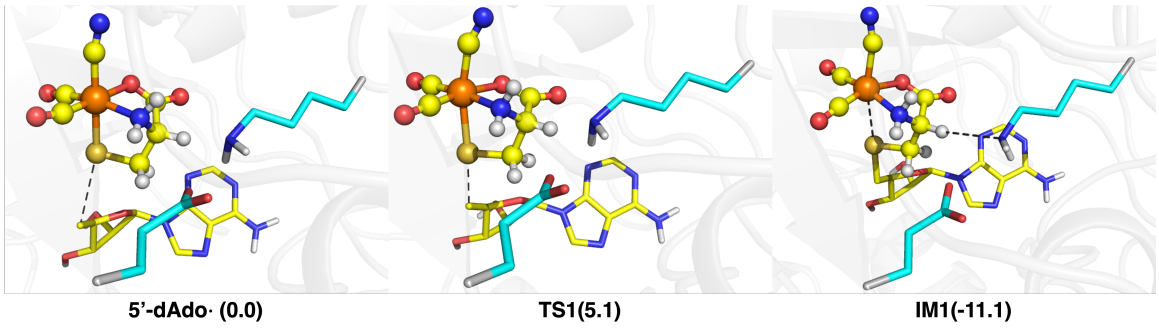


Figure S2: The key structures of the C-S radical addition. The color scheme is exactly the same as the figures in the main text.

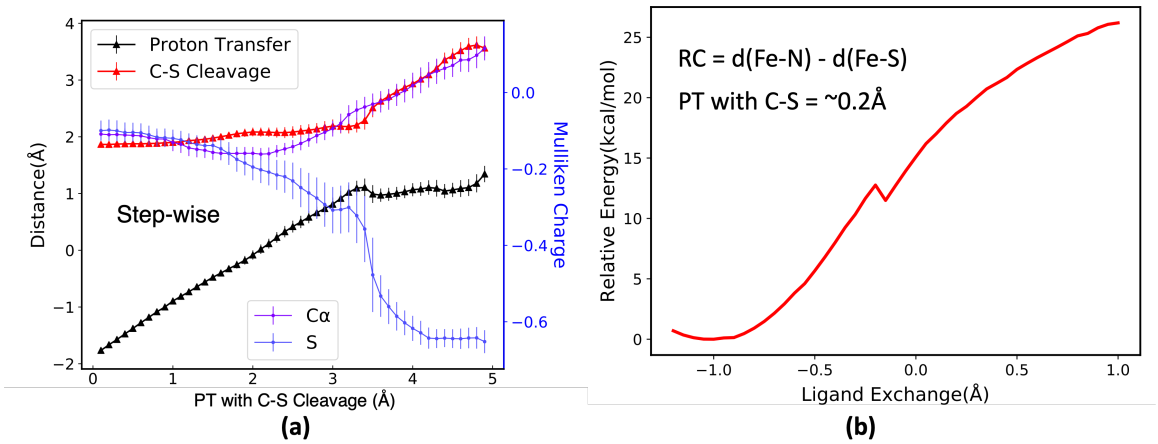


Figure S3: The change of the key distances and Mulliken charges of key atoms along the reaction coordinate(a), and the relative energy profile for the ligand exchange without the proton transfer and the C-S bond cleavage.

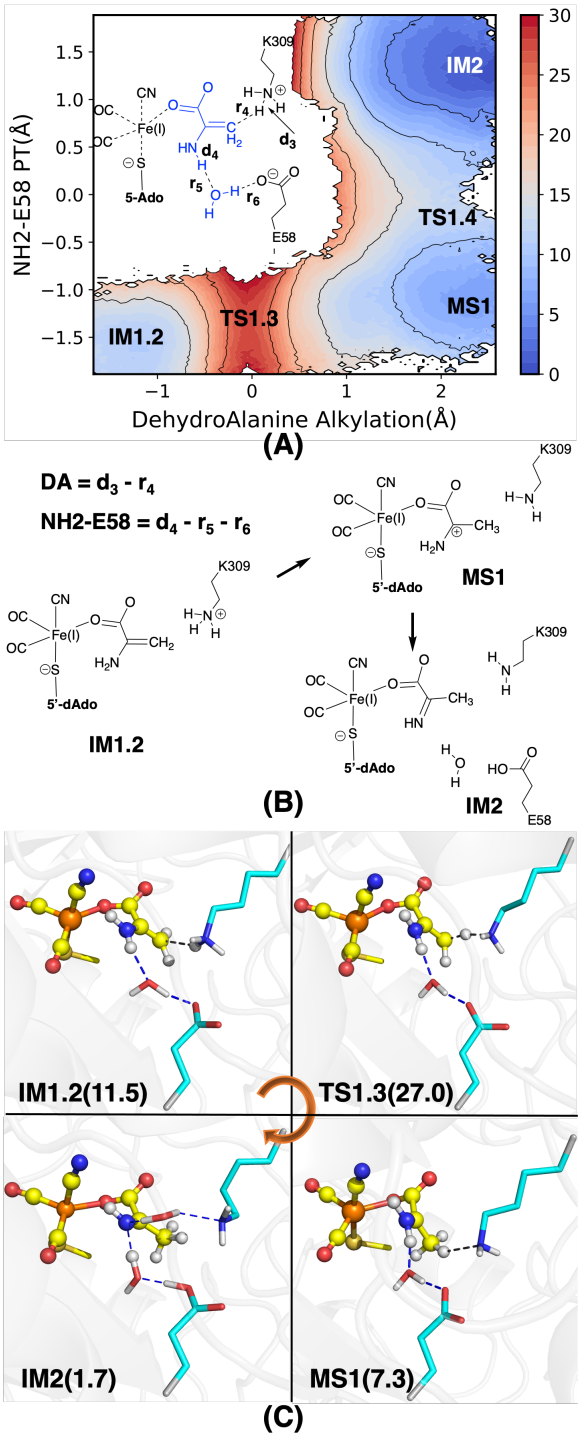


Figure S4: The free energy map of the dehydroalanine alkylation with the proton transfer (a) and the corresponding key structures (b,c).

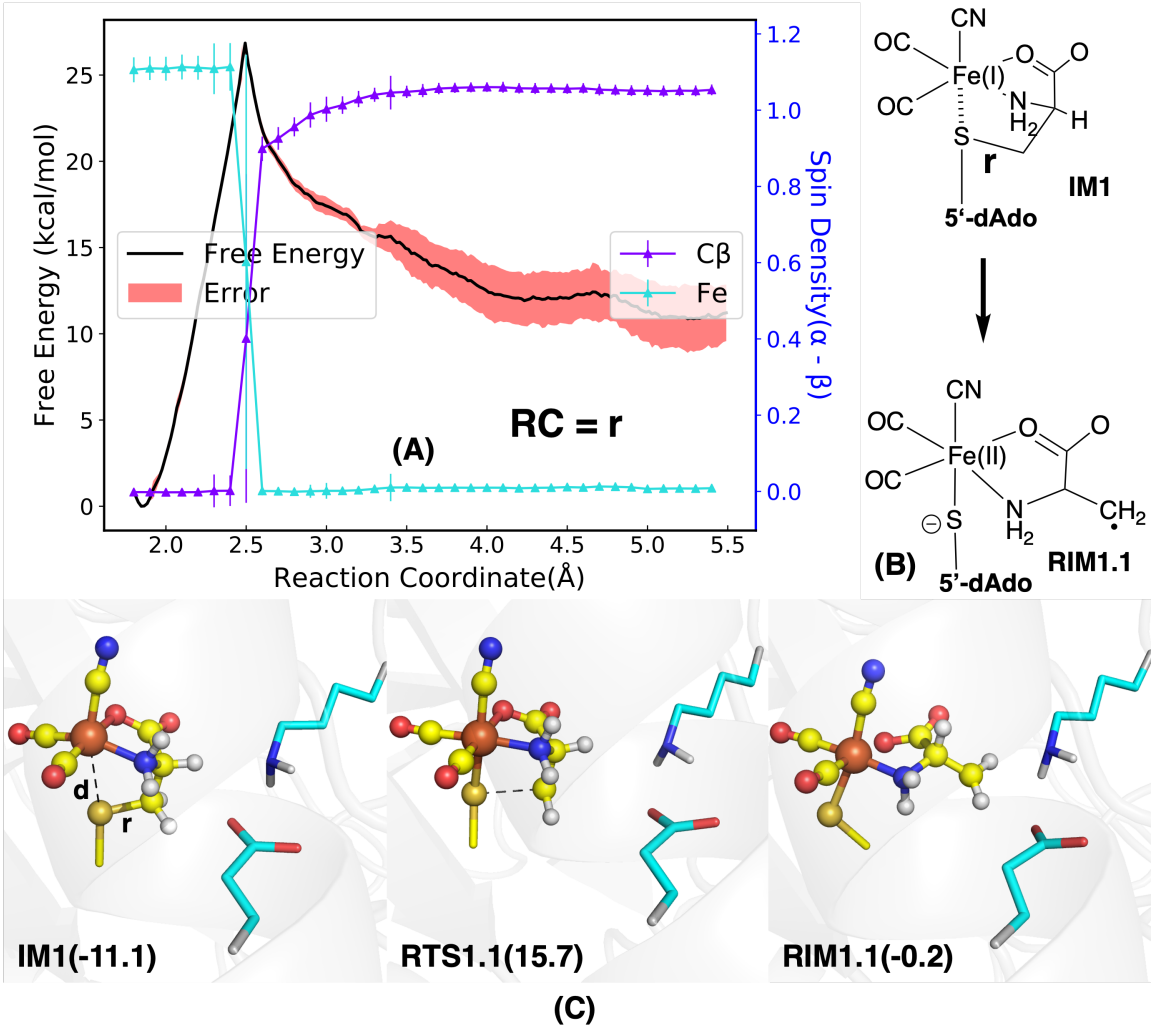


Figure S5: The free energy profile of the homolytic C-S bond cleavage with the spin density changes (A) and its key structures (B, C).

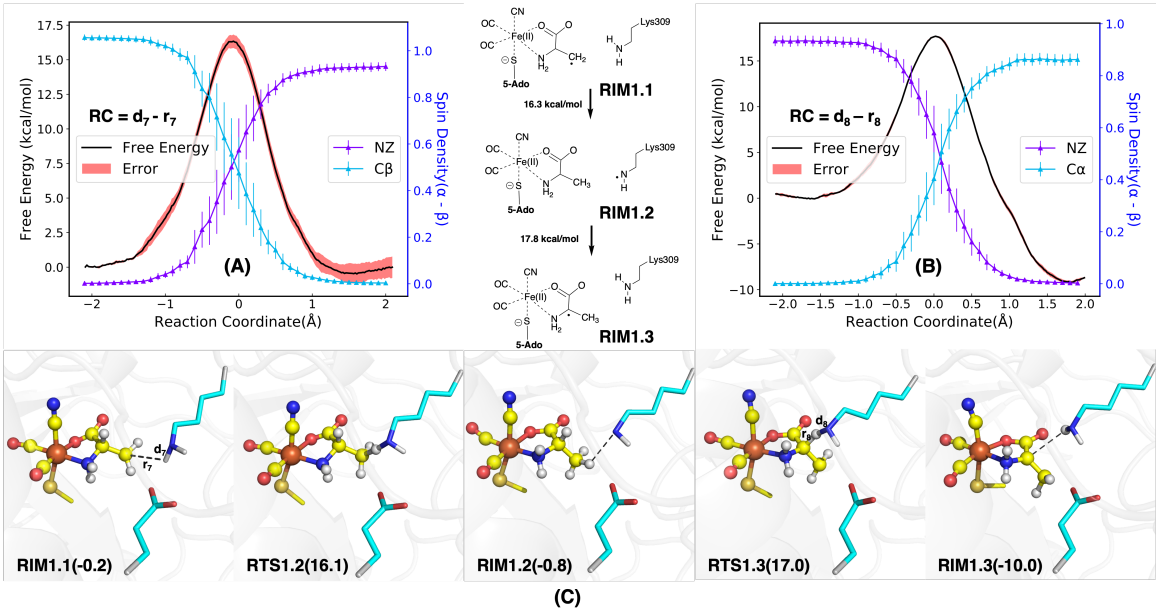


Figure S6: The free energy profile of the radical pathway (A,B) and the key structures (C).

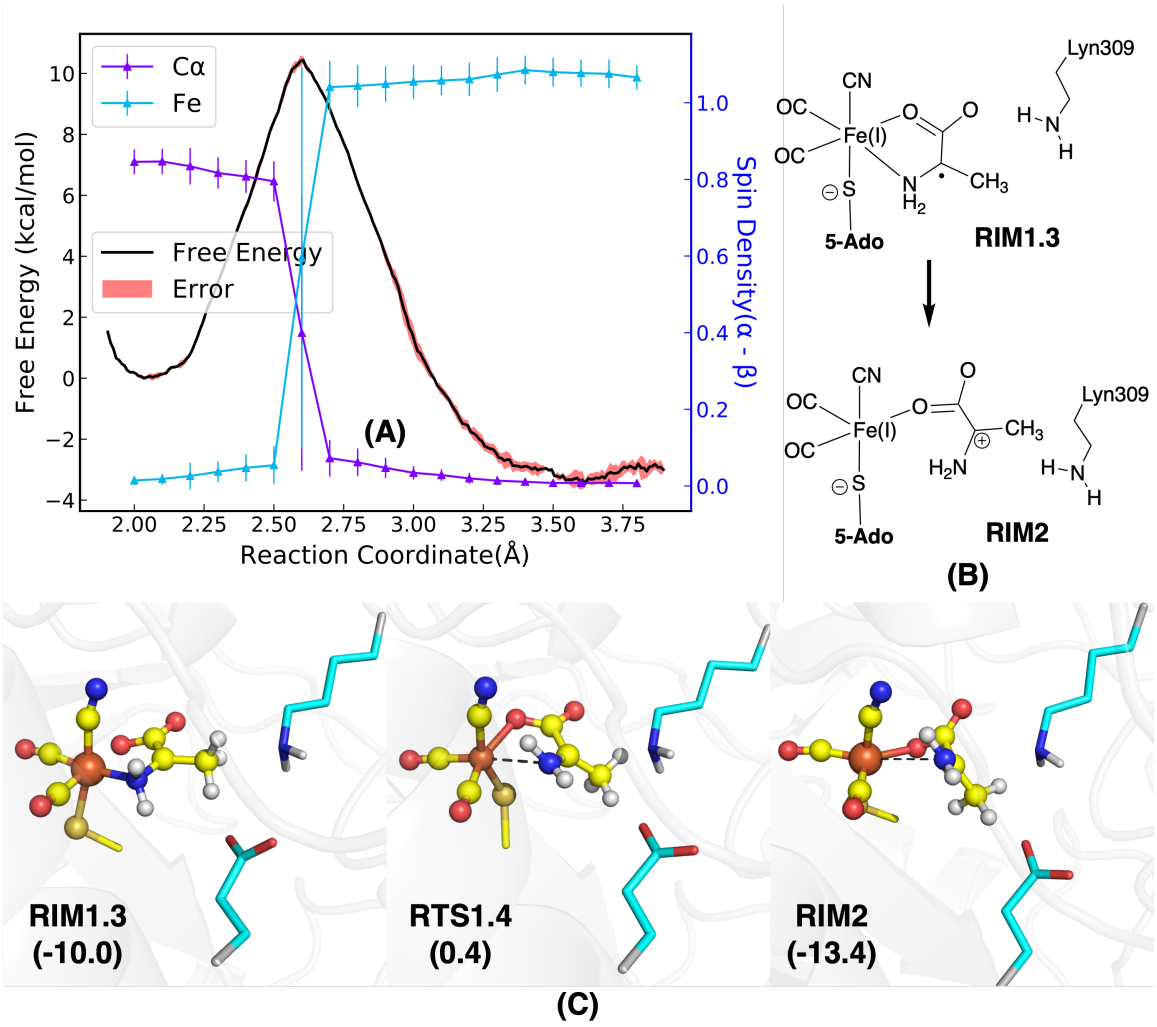


Figure S7: The free energy profile of the Fe-N dissociation (A), and the corresponding key structures in 2D (B) and 3D (C).

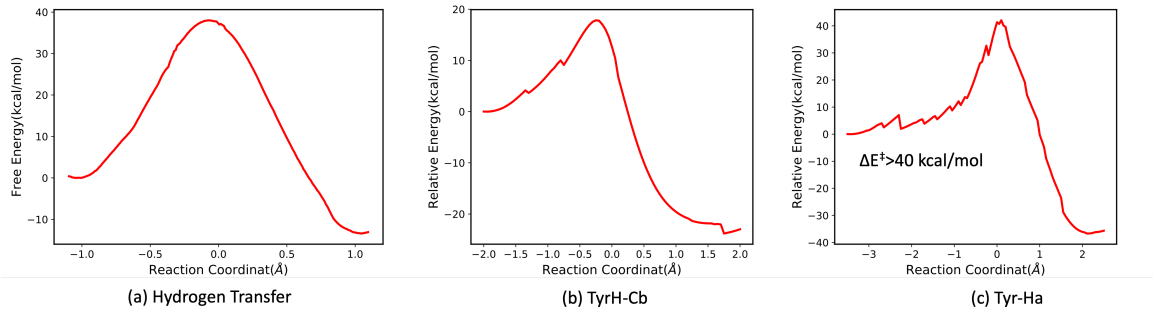


Figure S8: The relative energy profile of the radical transfer via Tyr306. The hydrogen transfer from Cα to Cβ is in (a). The Tyr306$^{\bullet}$ generation is in (b), while the following step Hα transfer is in (c)

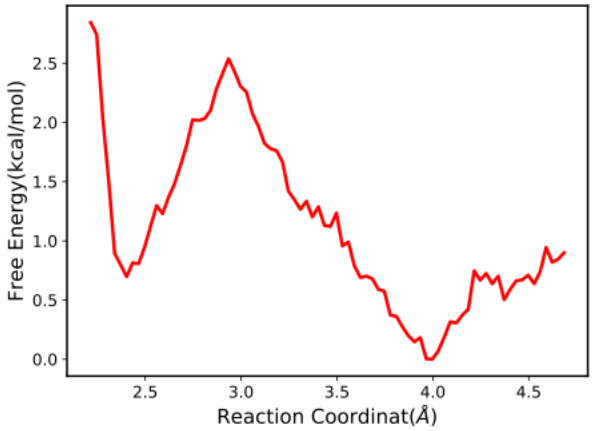


Figure S9: The free energy profile of the ligand exchange between M291 and 5-Ado.

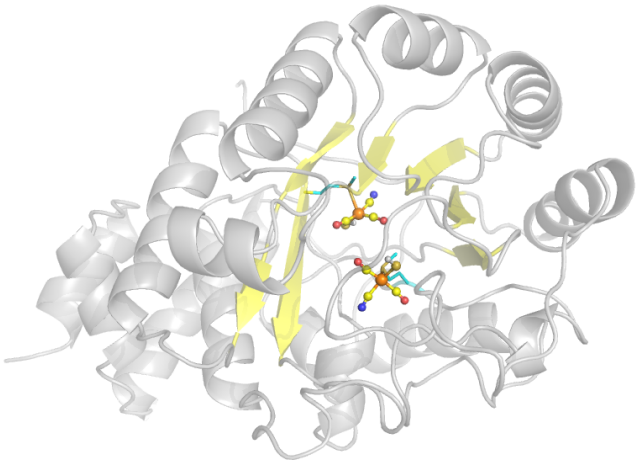


Figure S10: The initial structure of two Fe(I) clusters in HydE β-extent barrel.

C Supporting Information for Chapter 4: Sequence-based Prediction of Metamorphic Behavior in Proteins

Inconsistency Index. Besides the diversity index, there is another index describing the 'confusion' of the SSP programs, named the inconsistency index (iCI) because it evaluates the differences in predictions between two SSP programs. When a residue in a sequence has identical SSP results from two programs, then we assign an inconsistency value of 0 to that position. The assigned value is 1 when one prediction is random coil (C) and the other is either helix (H) or sheet (E). The assigned value is 3 when one prediction is helix (H) and the other is sheet (E). Similar to the DI, two variables are used to optimize the performance of the iCI, which are the number of consecutive residues in the moving window (CR, vertical axis) and the threshold value of the iCI (horizontal axis). The results are plotted as a heat map similar to Figure 4 in the main text. The SS comparison between Porter5 and Psipred has the MCC value as large as 0.3951, which is comparable to the MCC value obtained by the DI descriptor.

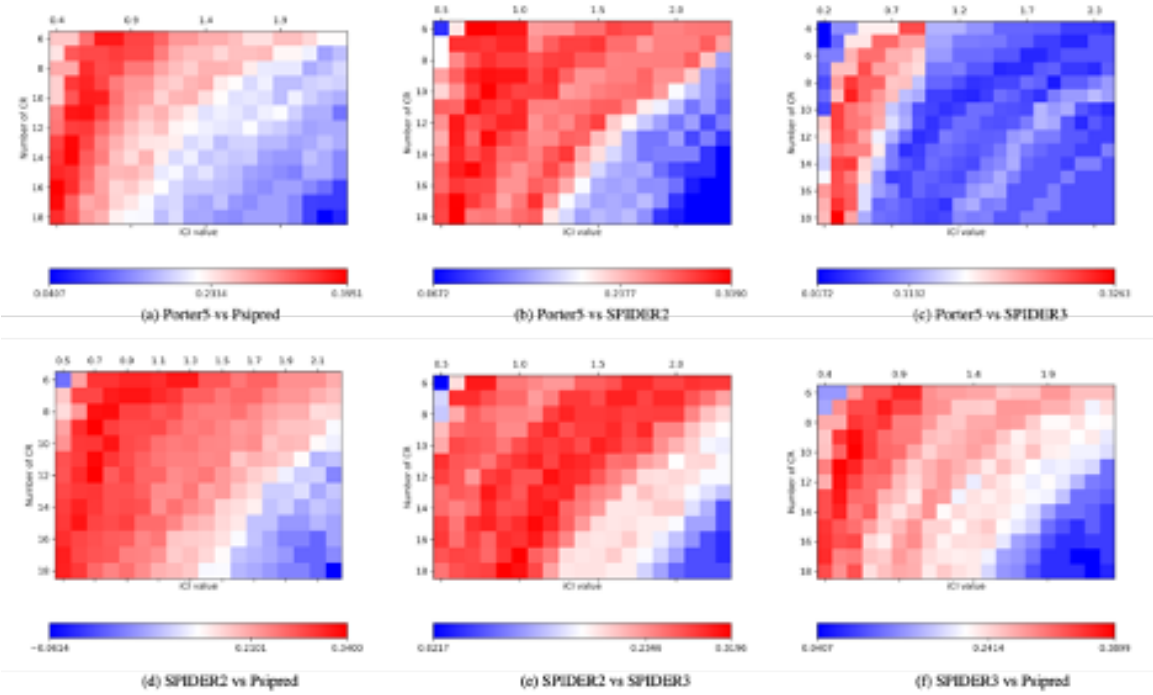


Figure S1: The MCC map of the IC descriptor of six SSP programs comparison, including the SS comparison between (a) Porter5 and Psipred, (b) Porter5 and SPIDER2, (c) Porter5 and SPIDER3, (d) SPIDER2 and Psipred, (e) SPIDER2 and SPIDER3, and (f) SPIDER3 and Psipred.

Principal Component Analysis (PCA). The unsupervised PCA method followed by K-means clustering was used to separate all the data into two groups, namely metamorphic (positive) and non-metamorphic groups (negative). After clustering, we calculate the MCC value as 0.41. This result indicates that without knowing the metamorphic property of proteins, the DIs are still appropriate for separating metamorphic proteins and non-metamorphic proteins.

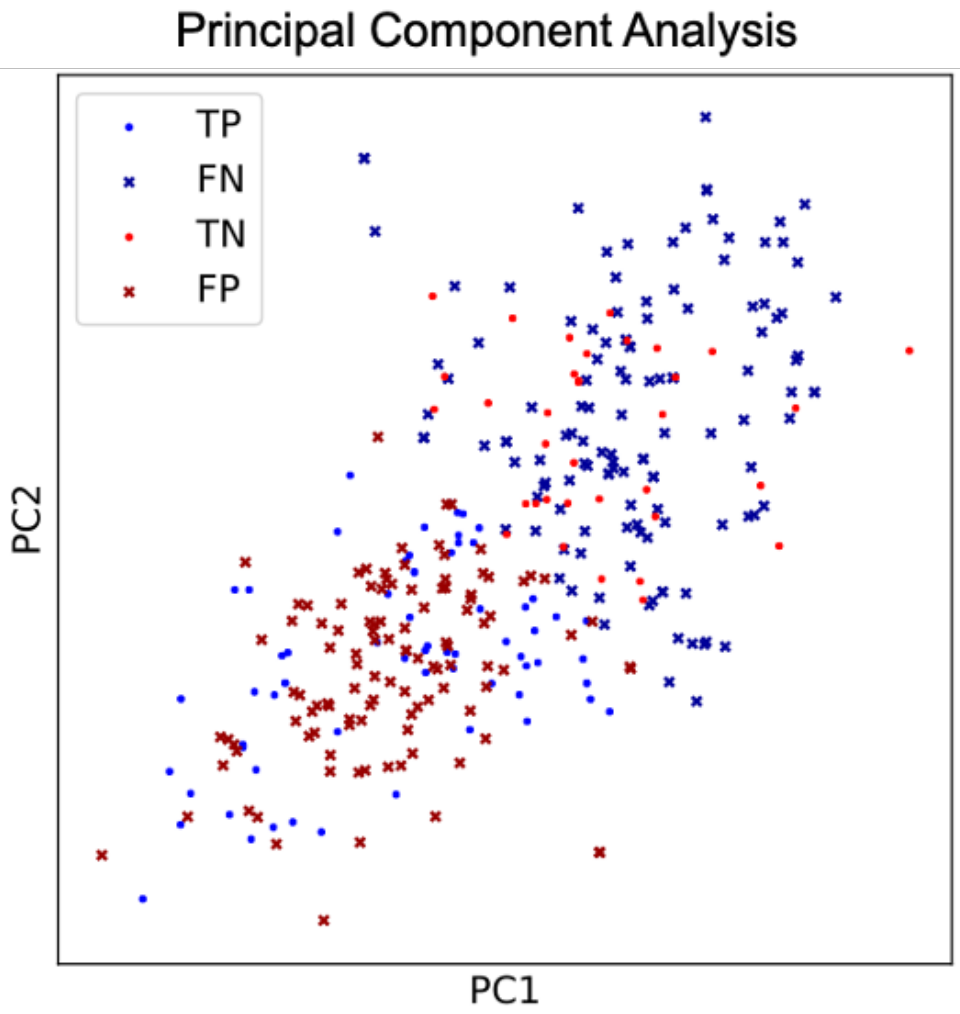
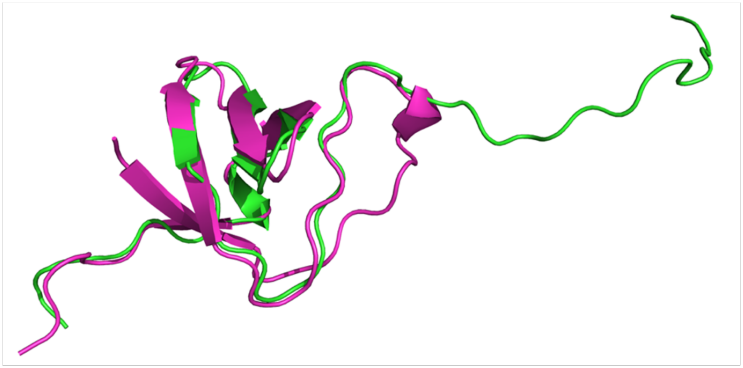
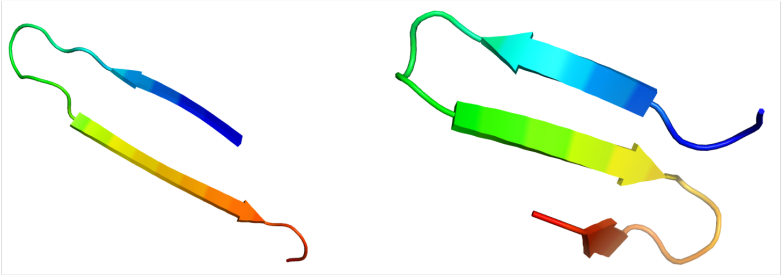


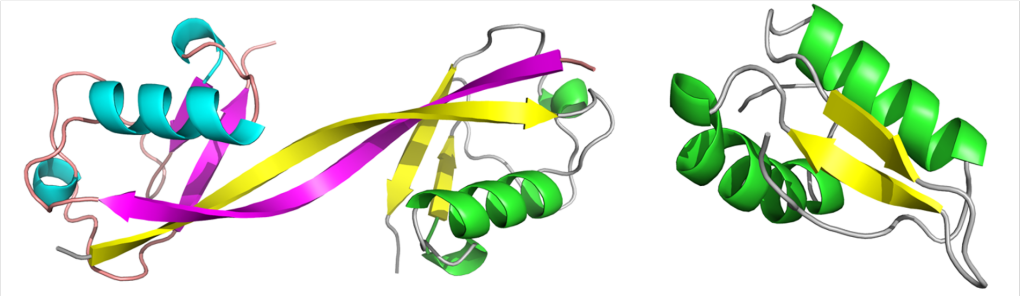
Figure S2: The PCA result of two most important principal components.



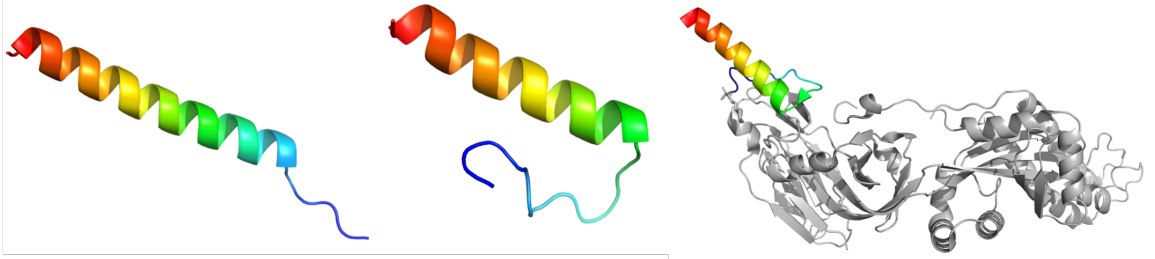
(a)



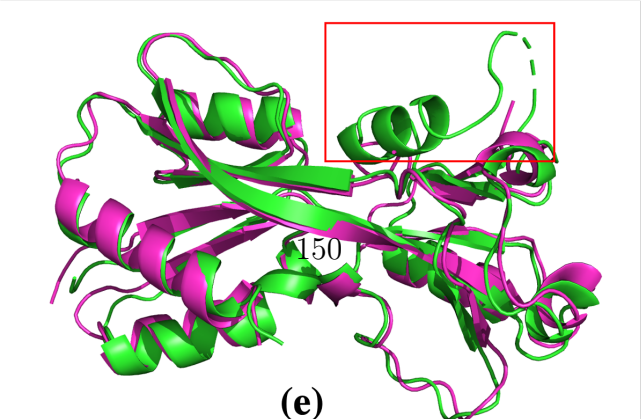
(b)



(c)



(d)



(e)

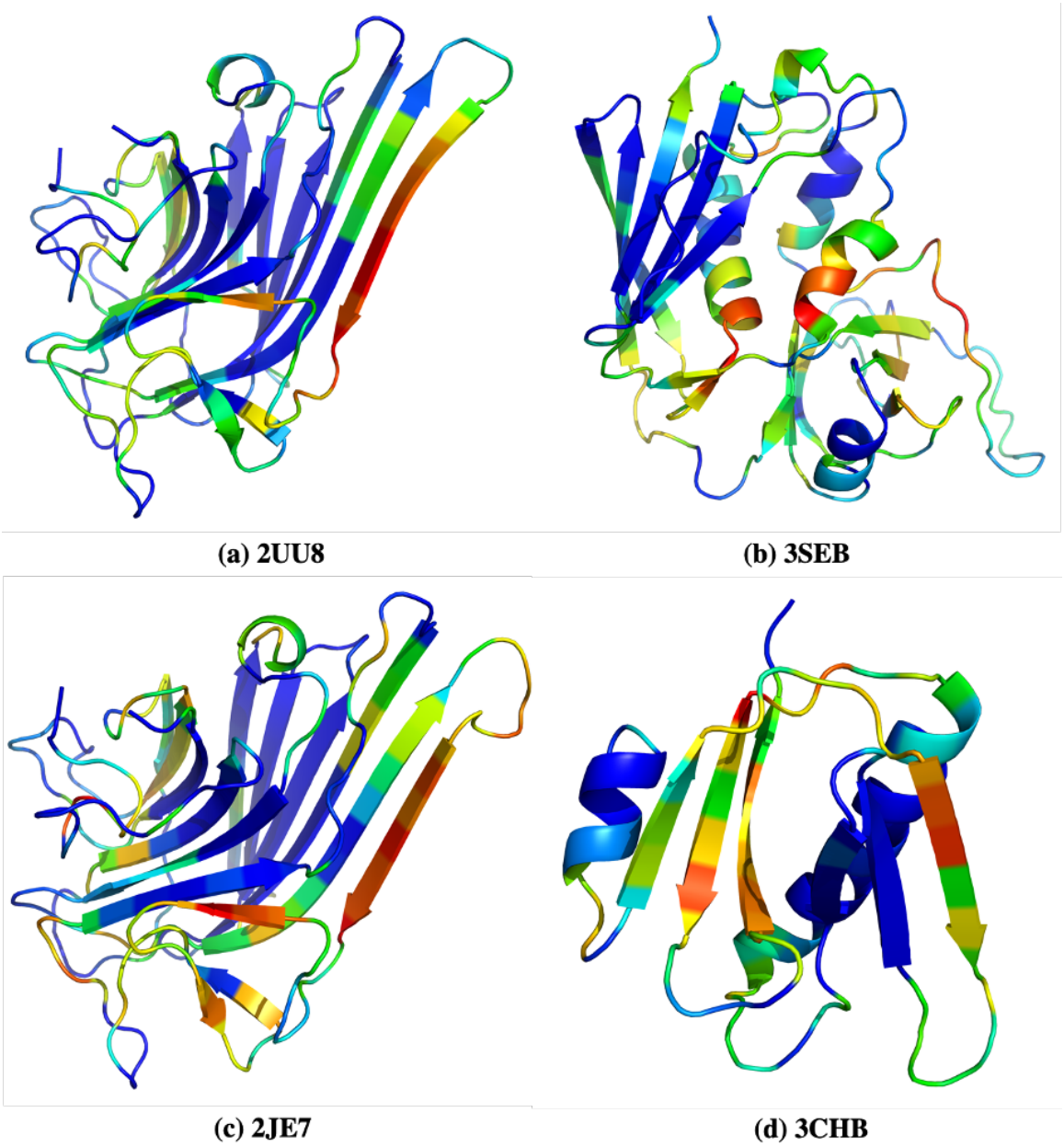


Figure S4: Four of the persistent 'false positive' results for diversity index-based metamorphic protein classification: (a) 2UU8, (b) 3SEB, (c) 2JE7 and (d) 3SEB. The secondary structures of the proteins are colored based on diversity indices at each position, ranging from 1.0 (blue) to 3.0 (red).