DEVELOPMENT AND APPLICATION OF MULTISCALE MODELLING TECHNIQUES FOR BIOCHEMICAL PROESSES

Neil Rory MCFARLANE

Supervisor(s):
Prof. Dr. Jeremy N. Harvey
Prof. Dr. Julianna Oláh

Members of the Examination Committee:
Prof. Dr. Koen Clays, Chair
Prof. Dr. Thomas Jagau
Prof. Dr. Peter Dedecker
Prof. Dr. Daniel Escudero

Prof. Dr. Tell Tuttle
(University of Strathclyde)

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Science (PhD): Chemistry

September 2024

© 2024 Neil Rory McFarlane

Uitgegeven in eigen beheer, Neil Rory McFarlane, Leuven

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaandelijke schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher

Acknowledgements

My previous professors and Master's thesis colleagues told me before I started my PhD that these four years would go by in a flash, and at the time I didn't believe them. As it happens, four years on and they were absolutely right. There are so many people to thank, each of which have provided support in various forms over the years, so if you are missing from these acknowledgements, it does not mean that I am not thinking of you.

My first thanks go to my supervisor, mentor, and good friend, Jeremy Harvey. After I finished my Master's thesis, I simply sent Jeremy an email asking if he had any opportunities opening up. Not only did he reply quickly, but when we talked, I could immediately tell that he would be a great person to work with. Soon after, he offered me a four-year position starting in October 2020. While the first months were undoubtedly hard due to global circumstances, we made the most of it, and it has all come together into a thesis that I am proud of. I have learned so much from Jeremy, especially since I started at KU Leuven with almost zero computational chemistry or biochemistry experience. From day one he has always appreciated my scientific suggestions, even if in reflection they were not always the most insightful. In the spirit of the idiom, Jeremy's office door truly is always open for discussion, and he has always made time for me despite his busy schedule. I've had many extracurricular opportunities over the years, including a visit to Vietnam teaching a course on the very subject of this thesis, all of which have contributed in various ways to my personal and professional development. Jeremy's attitude towards teaching at the Bachelor's and Master's level has been refreshing, and I have had many opportunities to contribute. Teaching subsequently became one of my favourite parts of my calendar which I looked forward to every year. I will always look fondly back on our relationship, so I thank him for everything.

Similarly, my co-supervisor Julianna Oláh has been a source of tremendous support, particularly in the early years. During the first year, I asked Julianna countless basic questions, and she was never patronizing or condescending. I affectionately look back on mine and Jeremy's short visit to Budapest in November 2022, from which I vividly remember Julianna organizing dinner at her house, ensuring that all the food was vegetarian for me – such thoughtfulness meant a lot to me. It was always a pleasure to talk with Julianna, so I thank her for the support over the years, and I sincerely hope that we will stay in touch.

While research groups are by nature somewhat of a revolving door, there are many colleagues and friends from Jeremy's group over the years who have had a lasting impact on me, and so deserve recognition. Thank you to Tolga (for Counter-Strike 1.6 games), Carlos (for constant positivity), Xiaoyong (for QM/MM expertise), Quico (for GROMACS skills), Peter (for interesting insight into many fields of science), Charles-Alexandre (for career advice and friendly chats), Hani (for friendly discussion and support), and Janko. I would like to extend a special thanks to Janko, as we have worked together a lot in the final year of my PhD. Every time we talk, I feel like I learn something new. We have made a great team, have produced some interesting (although perhaps unsatisfying...) results, and I consider us to be good friends. From Julianna's group, I would be remiss if I didn't give a special thanks to Ahmed. In the early years we worked together rather a lot and produced some nice results, including my first publication. I consider us to be good friends and hope that we will remain connected.

There are many other colleagues and friends from the division outwith the research groups I am a part of who deserve gratitude. Thank you to Koen V., Prashant, Mariana, Maarten, Charlotte T. (special thanks to her for abstract translation), and Quyen. Thank you also to the other members of our division: Daniel Escudero, Thomas Jagau, Jérôme Loreau, Natalie Beckers, and Hans

Vansweevelt. The culture which you have all made in the division is a pleasure to be a part of. A special thanks to Daniel and Thomas who have extended their responsibilities to also being a part of my examination committee. On that note, I would like to extend my thanks to the other members of my examination committee: Prof. Koen Clays, Prof. Peter Dedecker, and Prof. Tell Tuttle.

Besides the PhD students and postdoctoral colleagues, I have worked with many Master's students over the years who deserve just as much acknowledgement. Thank you to Savannah, Fileto, Jiangli, Mariia, Gautam, Xevi, Santi, Jingyi, and Deepti. Similarly, I feel the need to express my thanks to the Master's in Theoretical Chemistry and Computational Modelling students which are intrinsically connected to our division through the internship and Master's programme. Not only have many of these students provided interesting results, but more personally, they have shown me that you can learn a lot by guiding budding researchers.

Outside of KU Leuven, there are many friends in Belgium who deserve some thanks. Thank you firstly to my expat friends: Kyle, Tin, Feiyang, Hrvoje, Steven, Robin, and Màrcio. It's been a great pleasure learning about all your cultural experiences as far reaching as China and Uganda. A special thanks to Kyle for driving us to various countries with our bikes crammed into the boot, in search of the biggest cols we can find. Another thanks goes out to all of my Belgian friends: Koen K., Jelle, Dominik, Wouter, Alice, Lola, Louana, Tom, Eva, Benja, Aylin, Mehdi, Charlotte G., and Romain. You are all part of the reason that I wish to remain in Belgium in the coming years. *Bedankt et merci à tous !*

My friends all the way back from secondary school (some even from primary school and nursery!) in Scotland always have been and always will be my closest friends and greatest source of joy. Thanks to Alistair, James, Aaron, Dylan, Megan, Khadijah, Ruaraidh, and Arion. Even though we don't see each other in person that often these days, we continue to play Counter-Strike

almost every week (even though we're old and slow…), and always try to see one-another in person at least a couple of times a year. Our camping trips are always one of the highlights of my year.

Another of my greatest thanks goes out to my parents, Alison and Nelson. Given the size of our family, it really is in many ways just the three of us. Without you both, I would never be where I am as you have always supported my choices, no matter the associated difficulties. Thanks for everything, both now and for the future.

I have been lucky enough to have a family in Belgium as well as a family back home. *Un grand remerciement à Africa, Ignazio, Esteban, Elena, Rafael, Maeva, Oceania, Marvin, Valentina, et Nesta. Vous êtes une grande source de soutien et comme ma deuxième famille. Merci à tous.*

My final and greatest thanks go to my partner in life, Victoria. Without her, I would never have found myself doing a PhD in Belgium, and now we're living together and creating a life here. She is always able to cheer me up even if it's been a difficult and long day, and for that I am forever grateful. I can say with complete certainty that she is my best friend, and that we laugh every day. Thanks for everything, and here's to the next seven years.

I will be forever fond of this period of my life, in no small part due to all the people listed above.

> *"No one who achieves success does so without acknowledging the help of others. The wise and confident acknowledge this help with gratitude."*
>
> *- Alfred North Whitehead*

Abstract

In every organism, all occurring biochemical processes are in some way interconnected. Such complexity means that studying a given organism in its complete form is typically unfeasible, and so great effort goes into constructing small models which represent the most interesting parts of a given organism. This thesis is concerned with the use of computational modelling techniques to aid in understanding biochemical processes. Due to the different sizes and rates on which biochemical processes occur, a multiscale modelling approach is necessitated.

Chapter 1 explores the use of multiscale modelling techniques for the study of biochemical processes in a general sense and provides the biochemical context for the systems studied in detail in this thesis. Chapter 2 summarises the necessary theory behind the computational chemistry methods used in this thesis. The theory is not provided exhaustively, but at a level of detail appropriate for this thesis.

In Chapter 3, existing multiscale modelling techniques are used to study the only known NO receptor in humans, soluble guanylate cyclase, which upon binding NO sees a large increase in catalytic activity. The exact details behind NO binding to the regulatory domain of soluble guanylate cyclase remain poorly understood by experiment, and moreover the atomistic nature of the signalling pathway caused by NO binding is unclear. In this chapter, the binding of NO gas to the regulatory part of soluble guanylate cyclase is studied by high-level quantum mechanical cluster calculations on a small model, and with hybrid quantum mechanics/molecular mechanics calculations on the complete regulatory domain. In conjunction, the conformational changes caused by NO binding and the diffusion of NO gas into the regulatory domain are studied by molecular dynamics simulations. Ultimately, these calculations and simulations are

considered in a biological context, enabling a more complete understanding of this regulatory protein.

Chapter 4 explores another protein in which NO gas binds, cystathionine β-synthase. In this case, CO may also bind, and instead of increasing activity as for soluble guanylate cyclase, CO/NO binding results in complete inhibition of enzymatic activity. Experimentally, several models associated with CO/NO binding have been proposed, but the conformational changes following CO/NO binding causing inhibition remain ambiguous. Initially, high-level quantum mechanical cluster calculations on a small model are used to verify the spectroscopic models proposed for CO/NO binding. Following this, the conformational changes caused by CO/NO binding on the inhibited forms of cystathionine β-synthase are investigated with extended molecular dynamics simulations of the complete protein, enabling an understanding of signalling pathways of up to 25 Å in length.

In addition to the application of existing methodologies in Chapters 3 and 4, Chapter 5 is concerned with the development and implementation of a new technique for studying biochemical reactivity – the single-ended growing string method. While the algorithm itself is not novel, this represents the first implementation into the hybrid quantum mechanics/molecular mechanics framework. Following the implementation, the algorithm has been verified against a biochemical test case, the Claisen rearrangement catalysed by chorismate mutase. For the initial test case, the algorithm works effectively, with as much as 88% of the reaction pathways produced deemed to be of publication quality, and can be generated quickly and efficiently. Further testing on more complicated biochemical systems is required, but this initial case study shows promising results. On the relatively large dataset produced, a statistical analysis has been performed on the underlying distribution of activation energies calculated by hybrid quantum mechanics/molecular

mechanics. Previously, the distribution was considered to be Gaussian in nature, but the analysis performed here reveals that a slight positive skewness is more likely.

Beknopte Samenvatting

In elk organisme zijn alle biochemische processen op de een of andere manier met elkaar verbonden. Deze complexiteit betekent dat het bestuderen van een gegeven organisme in zijn volledige vorm meestal onhaalbaar is, en daarom wordt er veel moeite gedaan om kleine modellen te construeren die de meest interessante delen van een gegeven organisme voorstellen. Dit proefschrift gaat over het gebruik van computationele modelleringstechnieken om biochemische processen te begrijpen. Vanwege de verschillende lengteschalen en tijdschalen waarop biochemische processen zich afspelen, is een modelbenadering op meerdere schalen noodzakelijk.

Hoofdstuk 1 verkent het gebruik van multischaal modelleringstechnieken voor de studie van biochemische processen in algemene zin en biedt de biochemische context voor de systemen die in detail bestudeerd worden in dit proefschrift. Hoofdstuk 2 geeft een samenvatting van de noodzakelijke theorie achter de computationeel-chemische methoden die in dit proefschrift worden gebruikt. De theorie wordt niet uitputtend gegeven, maar op een detailniveau dat noodzakelijk is voor dit proefschrift.

In hoofdstuk 3 worden bestaande multischaalmodelleringstechnieken gebruikt om de enige bekende NO-receptor bij mensen te bestuderen, oplosbaar guanylaatcyclase, dat na binding van NO een grote toename in katalytische activiteit vertoont. De exacte details achter NO-binding aan het regulatiedomein van oplosbaar guanylaatcyclase worden experimenteel nog steeds niet goed begrepen en bovendien is de atomistische aard van de signaalroute die door NO-binding wordt veroorzaakt onduidelijk. In dit hoofdstuk wordt de binding van NO-gas aan het regulatiedomein van oplosbaar guanylaatcyclase bestudeerd door middel van

kwantummechanische clusterberekeningen op hoog niveau op een klein model, en met hybride kwantummechanische/moleculaire mechanica-berekeningen op het volledige regulatiedomein. Daarnaast worden de allosterische veranderingen veroorzaakt door NO-binding en de diffusie van NO-gas in het regulatiedomein bestudeerd met moleculaire dynamica simulaties. Uiteindelijk worden deze berekeningen en simulaties in een biologische context geplaatst, waardoor een completer begrip van dit regulerende eiwit mogelijk wordt.

Op vergelijkbare wijze wordt in hoofdstuk 4 een ander eiwit onderzocht waaraan NO-gas bindt, cystathionine-β-synthase. In dit geval kan CO-gas ook binden, maar in plaats van de activiteit te verhogen zoals bij oplosbaar guanylaatcyclase, resulteert CO/NO-binding in volledige remming van de enzymactiviteit. Experimenteel werden er reeds verschillende modellen voorgesteld die geassocieerd zijn met CO/NO binding, maar de conformationele veranderingen na CO/NO binding die remming veroorzaken blijven onduidelijk. In eerste instantie worden kwantummechanische clusterberekeningen op hoog niveau op een klein model gebruikt om de spectroscopische modellen die zijn voorgesteld voor CO/NO-binding te verifiëren. Vervolgens worden de conformationele veranderingen veroorzaakt door CO/NO binding op de geïnhibeerde vormen van cystathionine β-synthase, onderzocht met uitgebreide moleculaire dynamica simulaties van het complete eiwit, waardoor allosterische paden met een lengte tot 25 Å diepgaand begrepen kunnen worden.

Naast de toepassing van bestaande methodologieën in hoofdstuk 3 en 4, beschrijft hoofdstuk 5 de ontwikkeling van een nieuwe techniek voor het bestuderen van biochemische reactiviteit - de single-ended growing string methode. Hoewel het algoritme zelf niet nieuw is, werd het nog niet geïmplementeerd in het hybride kwantummechanica/moleculaire mechanica schema. Na de implementatie werd het algoritme geverifieerd aan de hand van

een biochemische testcase, de Claisen-omlegging gekatalyseerd door chorismaatmutase. Voor de eerste testcase werkt het algoritme effectief, waarbij maar liefst 88% van de geproduceerde reactiepaden van publicatiekwaliteit wordt geacht. Dergelijke routes kunnen snel en efficiënt worden gegenereerd. Hoewel het verstandig zou zijn verder te testen op gecompliceerdere biochemische systemen, laat deze eerste casestudie reeds veelbelovende resultaten zien. Op de relatief grote dataset die geproduceerd werd, werd een statistische analyse uitgevoerd op de onderliggende verdeling van activeringsenergieën berekend door hybride kwantummechanica/moleculaire mechanica. Eerder werd de verdeling als normaal beschouwd, maar de hier uitgevoerde analyse laat zien dat een lichte positieve scheefheid waarschijnlijker is.

List of Abbreviations

AdoMet – S-adenosyl-methionine

ATP – adenosine triphosphate

BsCM – Bacillus subtilis chorismate mutase

CADD – computer-aided drug design

CBS – cystathionine β-synthase

CC – coupled-cluster

CM – chorismate mutase

CSE – cystathionine γ-lyase

DE-GSM – double-ended growing string method

DLC – delocalised internal coordinate

DLPNO – domain-based localised pair natural orbital

DFT – density functional theory

DoF – degrees-of-freedom

GGA – generalised gradient approach

GSM – growing string method

GTO – Gaussian-type orbital

GTP – guanosine triphosphate

hCBS – human cystathionine β-synthase

HF – Hartree-Fock

HK – Hohenberg-Kohn

H-NOX – Heme-nitric-oxide/oxygen binding

KS – Kohn-Sham

LCAO – linear combination of atomic orbitals

LDA – local density approximation

LDSA – local spin-density approximation

MD – molecular dynamics

MEP – minimum energy pathway

MM – molecular mechanics

MO – molecular orbital

NAC – near-attack conformer

NEB – nudged elastic band

NOS – nitric oxide synthase

PAS – Per-Arnt-Sim

PBC – periodic boundary condition

PDB – Protein Data Bank

PDF – probability density function

PLP – pyridoxal 5'-phosphate

PNO – pair natural orbital

QM – quantum mechanics

QM/MM - quantum mechanics/molecular mechanics

RAS – rat sarcoma virus

RMSD – root-mean-squared-deviation

RMSF – root-mean-squared-fluctuation

SCF – self-consistent field

SE-GSM – single-ended growing string method

sGC – soluble guanylate cyclase

STO – Slater-type orbital

TS – transition state

TSA – transition state analogue

ZORA – zeroth-order relativistic approximation

ZVPE – zero-point vibrational energy

Table of Contents

Acknowledgements ..... *I*

Abstract ..... *V*

Beknopte Samenvatting ..... *VIII*

List of Abbreviations ..... *XI*

Table of Contents ..... *XIII*

1 Introduction *1*

1.1 The Multiple Scales of Biochemical Processes **2**

1.2 The 3-Dimensional Structure Revolution **3**

1.3 Computational Modelling of Biochemical Processes **4**

1.3.1 Static Methodologies 5

1.3.2 Dynamic Methodologies 10

1.4 Biochemical Processes and Systems Studied **12**

1.4.1 Human Biochemical NO-signalling: Soluble Guanylate Cyclase 14

1.4.2 Inhibition of the Transsulfuration Pathway: Cystathionine β-synthase 17

1.4.3 Pericyclic Claisen Rearrangement: Chorismate Mutase 23

1.5 Conclusions and Objectives **25**

1.6 References **28**

2 Theoretical Background *33*

2.1 Molecular Mechanics **33**

2.2 Quantum Mechanics **37**

2.2.1 Hartree-Fock Method 39

2.2.2 Electron Correlation 43

2.2.3 Density Functional Theory 46

2.3 Hybrid QM/MM **53**

2.3.1 QM and MM Region Coupling 54

2.3.2 QM and MM Region Boundary 57

2.4 Molecular Dynamics **58**

2.4.1 Trajectory Propagation 58

2.4.2 The Molecular Dynamics Method 62

2.4.3 Periodic Boundary Conditions 63
2.4.4 Thermodynamic Ensembles 65

2.5 Free energy 66
2.5.1 Statistical Mechanical Methods 66
2.5.2 Sampling Methods 68

2.6 Delocalised Internal Coordinates 69
2.6.1 Generation of Delocalised Internal Coordinates 70
2.6.2 Geometry Optimisation in Delocalised Internal Coordinates 72
2.6.3 Imposing Delocalised Internal Coordinate Constraints 74

2.7 References 76

3 Human Biochemical NO-signalling: Soluble Guanylate Cyclase 79

3.1 Introduction 79

3.2 Computational Details 84
3.2.1 Quantum Mechanical Calculations 84
3.2.2 Molecular Dynamics Simulations 85
3.2.3 QM/MM Calculations 87

3.3 Results and Discussion 88
3.3.1 Quantum Mechanical calculations 88
3.3.2 Molecular Dynamics Simulations 91
3.3.3 QM/MM Calculations 99

3.4 Overall Findings 103

3.5 Conclusions and Future Work 106

3.6 References 108

4 Inhibition of the Transsulfuration Pathway: Cystathionine β-synthase 112

4.1 Introduction 112

4.2 Computational Details 118
4.2.1 Quantum Mechanical Calculations 118
4.2.2 Molecular Dynamics Simulations 121

4.3 Results and Discussion 124
4.3.1 Quantum Mechanical Calculations 124
4.3.2 Molecular Dynamics Simulations 130

4.4 Conclusions and Future Work 136

4.5 References 139

5 Development of a Hybrid QM/MM Growing String Method 143

5.1 Introduction 143

5.2 Algorithm Details 151

5.2.1 Hybrid Delocalised Internal Coordinates 151
5.2.2 Single-ended Growing String Method 153

5.3 Case Study: Chorismate Mutase 156

5.3.1 Introduction 156
5.3.2 Computational Details 159
5.3.3 Results and Discussion 161

5.4 Conclusions and Future Work 173

5.5 References 175

6 Summary and Outlook 180

6.1 References 186

7 List of Publications 187

7.1 Journal Articles 187

7.2 Conferences 187

8. Appendices 188

A. Supporting Information for Chapter 3 188

MD Simulations 188
QM/MM Calculations 191

B. Supporting Information for Chapter 4 201

C. Supporting Information for Chapter 5 224

1 Introduction

In recent years, computational modelling has become a pillar of research and development, allowing for numerically complex physical problems to be addressed. Such physical problems may be those of fundamental importance, practical relevance, or a combination of both. With the golden era described by Moore's law seemingly not yet over, albeit perhaps slowing down, the scale of computational modelling is constantly increasing, enabling larger and more complex problems to be investigated. A large part of computational modelling is dedicated to investigating natural phenomena, which can range from enormous-scale astrophysical events to microscopic biochemical processes, both benefitting from increasing computational power. In this thesis, the focus is on biochemical processes, and how these can be modelled using computational methodologies at multiple scales.

Biochemistry is a discipline which rationalises biological observations with chemical concepts.[1] More specifically, a biochemical process is defined as a process which naturally occurs within living organisms, and ultimately provides some integral function to the organism. Some examples occurring in humans may be as broad as oxygen transport by hemoglobin, or as multi-faceted as the role of adenosine triphosphate (ATP) in numerous biological processes. Computational modelling techniques based on theoretical chemistry may be used to investigate such processes, giving insights ranging from the undergraduate-level understanding of inter- and intramolecular forces to complex molecular orbital descriptions of reactivity. In this thesis, it will be demonstrated that it can be necessary to employ a multiscale approach when studying such systems as the features of a given process occur at variable scales of time and/or space, as will be explored in more detail in Section 1.1.

While computational modelling can be a very powerful tool that enables great insight into biochemical processes, choosing exactly how a process will be modelled remains a challenge. If the model is poorly considered, one may fall victim to the principle of *"Garbage In, Garbage Out"*, implying that the significance of the model is only as valuable as the care taken by the user. Throughout this doctorate, this principle has been borne in mind.

> *"On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question."*
>
> *- Charles Babbage, Passages from the Life of a Philosopher, 1864*

The aim of this opening chapter is not to detail the multiscale computational modelling techniques used but instead is to underscore their importance in understanding biochemical processes in a general sense with relevant examples. Then, the biochemical context for the processes studied herein will be described, with the specific details of each process being left to the relevant chapters.

1.1 The Multiple Scales of Biochemical Processes

Due to the large size of biochemical systems, the modelling of a biochemical process is intrinsically a high-dimensionality problem. Moreover, the strong correlation with other biochemical processes, along with the wide range of time-scales for individual biochemical reactions, which can span from femtoseconds to hours, further contributes to the complexity of the problem. Even with Moore's law holding steadfast, effectively modelling a given biochemical process which considers possible correlation with

other processes at the limit of the expected time-scale is typically not possible, so some simplification must be made.

The history of this simplification points to the seminal work by Michaelis and Menten,[2] where the concept of time-scale separation is introduced. This concept suggests that the biochemically-relevant part of a system occurs significantly faster than the remainder of the system, such that it can be isolated and studied independently to make meaningful conclusions. Extrapolating this idea further, this suggests biochemical processes can be effectively modelled at multiple scales, whether this be on a time- or size-scale, supposing that an appropriate model is chosen to describe the process.[3] Conceptually, this opens the door for the computational modeller, allowing for effectively any biochemical problem to be studied.

1.2 The 3-Dimensional Structure Revolution

While the concepts described above which enable the multiscale modelling of large and complex biochemical processes are powerful, to model such a process, 3-dimensional structures of the proteins, cofactors, lipids, nucleic acids and carbohydrates involved should be available. Perhaps the most complex of these to obtain are proteins, due to the 20 essential amino acids giving an almost limitless number of polypeptide sequences which fold in complex ways to generate secondary, tertiary and quaternary structures. Such proteins may be bound to a variety of cofactors in complex and cooperative ways, complicating the problem yet further.

Established in 1971 with only seven 3-dimensional structures, the Protein Data Bank (PDB) has since solved this problem in part by allowing for these structures to be deposited by experimentalists worldwide.[4] This database enables the study of many biochemical processes, supposing that relevant and quality 3-dimensional structures have been deposited. Several techniques are used to

obtain these structures, but a common factor between them is that they are invariably laborious, meaning that if a structure has not yet been deposited, then it may take rather a long time for it to be available.

Fortunately, we are currently in the midst of a protein structure prediction revolution,[5] due to the advent of machine-learning structure prediction tools such as the globally-popular AlphaFold,[6–8] enabling efficient and accurate prediction of the protein fold of any polypeptide sequence. While in this thesis machine learned structure prediction tools were not used and all 3-dimensional structures were from the PDB, it is important to emphasize their relevance for the future, firmly placing us in a golden era of structural biology and computational modelling of biochemical processes.[9]

1.3 Computational Modelling of Biochemical Processes

> *"If we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."*
>
> *- Richard P. Feynman, The Feynman Lectures on Physics, 1964*

The fundamental concepts behind time- and size-scale separation have been outlined above, but in this section, the detail of how this relates to computational chemistry with relevant examples of biochemical use cases will be given. Computationally, a chemical system may be studied in either a 'static' or 'dynamic' fashion. When using the former, the system is not subject to time-evolution and so cannot represent the dynamics of a given process, and any calculated properties are, by definition, time-independent. On the

other hand, the latter considers the effect of time on the behaviour of a system, so that dynamic processes may be modelled, and any calculated quantities are time-averaged.

Static techniques are certainly valuable largely due to their significantly lower computational cost when levels of theory are compared like-for-like to their dynamic counterpart. Moreover, carefully-selected static models which describe key components of overall system behaviour can be very useful. However, in its purest form, a static technique does not sample configurational space in any way, meaning that deriving chemically-relevant conclusions can be challenging. By using statistical mechanical methods like those described in Section 2.5.1, one can obtain quasi time-averaged properties in an efficient manner, however, for various practical reasons, these methods are typically only applied to small molecules.

For the study of large biochemical processes in which overall structural dynamics are important, it is logical that a dynamic technique is necessitated. Besides this requirement for a dynamic technique is certain cases, another improvement gained is best exemplified by the concept of sampling. Almost all experiments and computations fall victim to the sampling problem to varying extents, which is a statistical error that occurs when the selected sample does not represent the population. Due to their time-dependence, dynamic protocols intrinsically tackle the sampling problem. When studying large biochemical systems, appropriate sampling becomes increasingly important due to the large number of degrees-of-freedom (DoF).

1.3.1 Static Methodologies

The simplest of static methods described herein is the molecular mechanics (MM) method, which uses a classical description of molecular interactions to efficiently and often quite accurately describe their interactions (see Section 2.1 for theoretical details).

The most important ingredient in a successful MM calculation is that the force field must be well-selected, which is a set of parameters that describe the magnitude and type of interactions between atoms. Accordingly, there has been extensive work into developing widely-applicable and self-consistent force fields to improve the accuracy of MM calculations. Several options exist to effectively model proteins, lipids, small organic molecules, and materials, among others.[10–13]

Traditional, or non-polarisable, force fields define the charges on atoms once and only once. However, in reality, local electrostatic environments in large chemical systems are constantly changing. In order to improve upon non-polarisable force fields, there has been recent growth in the development of polarisable force fields, where local charge densities are allowed to vary, giving only a modest increase in computational expense and more accurately describing non-bonded interactions.[14] Particularly for when MM is used in the context of dynamics, systems with strong polarisation effects such as in ionic liquids[15,16] are described well by polarisable force fields. Nevertheless, for the study of most biomolecular processes, non-polarisable force fields remain the dominant method of choice due to the limited portability associated with polarisable force fields at this stage in their development.[17]

Perhaps the foremost weakness of MM is that the making and breaking of chemical bonds cannot be captured, and so biochemical reactivity cannot be modelled. Even though, in principle, some relatively novel reactive MM force fields allow for modelling of chemical reactivity,[18] the increased computational cost in addition to other conceptual limitations has meant they have seen limited use thus far.[19]

Notwithstanding the limitations outlined above, MM has seen much use in biochemical modelling, largely because the MM potential energy function can be calculated quickly. These days,

MM is usually not used exclusively, but rather as a key component of many other techniques or workflows. MM is often used for Computer-Aided Drug Design (CADD),[20] where it is used to generate poses by molecular docking in the virtual screening of novel compounds.[21] MM is also extensively applied to energy minimisation procedures on biochemical structures, where many MM calculations are performed with geometry optimisation steps between cycles. Perhaps the principal application of MM is in biochemical simulation, where it defines the potential energy function in classical molecular dynamics (MD) simulations, as will be explored in Section 1.3.2.

For any biochemical reaction, it is well-established that the catalytic effect of an enzymatic environment has a large influence on reaction rate, stereoselectivity, and enantioselectivity.[22] Biochemical reactivity is typically studied using one of two techniques: quantum mechanical (QM) cluster models and hybrid quantum mechanics/molecular mechanics (QM/MM). The QM (see Section 2.2 for theoretical details) cluster approach requires careful construction of a small- to medium-sized model of the active site of the enzyme under study.[23–27] This approach represents both size- and time-scale separation, as a small part of the system is selected where chemical reactivity occurs at a much greater speed than the conformational changes found in the biochemical system following reactivity. A great advantage of this approach is that accurate QM methods can be used, thereby giving an excellent description of the enzyme's reactivity. Of course, the choice of accurate QM methods will be limited by even the smallest of models, so model choice is critically important.

The QM cluster must be constructed in such a way that enough of the enzyme environment is included so that its (de)stabilising effect is accurately captured. The point at which this is achieved is often known as cluster size convergence. As an example of the identification of this point, take the mechanism of 4-

oxalocrotonate tautomerase, which catalyses the conjugation of α-keto acids. Since monomeric 4-oxalocrotonate tautomerase is a very small enzyme,[28] it represents a good test case for systematic improvement of cluster size. Sevastik and Himo have previously studied the mechanism with two different cluster sizes, where they test the cluster both without and with implicit solvation representative of bulk protein electrostatics. If the difference in relative energies between the implicitly solvated and solvent-free cluster is below some threshold, then cluster convergence is said to be reached.[29] They find that the larger of the two clusters shows cluster convergence, but it is worth highlighting that the larger model already includes a relatively large number of the only 64 amino acids found per monomer. In cases with larger substrates and/or longer amino acid sequences, larger clusters may be necessary due to the long-range (de)stabilising effect of amino acid residues.[30]

Another illustrative example of the identification of cluster convergence is in the study of haloalcohol dehalogenase, an enzyme which catalyses the dehalogenation of haloalcohols, forming an epoxide product. Hopmann and Himo have studied the reaction catalysed by haloalcohol dehalogenase, constructing three different clusters of increasing size.[31] In this work, the authors find that cluster convergence is achieved at a rather large 161 atoms. If one wishes to use more accurate QM methods, a cluster size of 161 atoms could quickly become cost-prohibitive.

In any case, in both examples outlined above, the authors do indicate that the use of carefully-selected cluster model complemented with an implicit solvation model to account for the remaining protein and solvent environment can also provide cluster convergence. However, such a procedure requires some experience, and if one must perform tests on larger models to verify a smaller model, then the cost benefit of the smaller model may become less important.

Usefully, hybrid QM/MM (see Section 2.3 for theoretical details) methods enable inclusion of the entire protein environment (modelled as MM) surrounding a chosen active site (modelled as QM), and usually at a lesser cost than the quantum mechanical cluster equivalent with the same level of theory.[32] Typically, the QM method chosen is one of relatively modest cost, as there can be many cycles of optimisation associated with a QM/MM calculation due to the increased number of DoF introduced with the inclusion of the entire protein environment.[33–38] In any case, the QM region would typically be of a smaller size than its QM cluster counterpart.

It has been suggested by many that to approach chemical accuracy, the QM region selected should be as large as computationally feasible.[39–41] However, it has also been suggested that the size of the QM region is less important as the hybrid QM/MM approach naturally captures charge transfer processes.[42] In practice, choosing the QM region is non-trivial, and is in-fact problem dependent.

QM/MM has become the state-of-the-art for the modelling of reactivity in biochemical systems,[43] which will be explored in greater detail in Chapter 5. However, with any QM/MM system, there are associated difficulties in model construction. For the above example of 4-oxalocrotonate tautomerase, the mechanism was studied using hybrid QM/MM using various models and it was found that if the model was poorly constructed (in this case the orientation of the substrate in the QM region was unfavourable for a given model),[44] then the QM/MM approach was found to provide unsatisfactory results, highlighting the importance of user experience and care. Evidently then, the QM cluster approach can certainly still be valuable, giving an overall simpler system preparation procedure. Some authors have even suggested protocols that use both methods,[45] exploiting the computational efficiency of QM/MM in generating structures and then the

accuracy of a large QM-only model to describe electronic structure.

1.3.2 Dynamic Methodologies

As indicated above, in the context of studying biochemical processes, many interesting properties can only be illuminated using a dynamic protocol. In this thesis, the dynamic method which was used extensively is the molecular dynamics (MD) approach (see Section 2.4 for theoretical details).[46] In this approach, since the process under study is subject to time-evolution, particles can translate based on the forces calculated by the potential energy function (usually based on a MM force field). This approach allows for structural dynamics to be studied, such as conformational changes following reactions, protein folding, and substrate binding.

Since experimentally-determined 3-dimensional biomolecular structures provide an average of the atomic positions, complex structural dynamics are not well-captured by experimental procedures. Using an experimental (or indeed machine-learned) 3-dimensional structure as a starting point, MD simulations can simulate the behaviour of a given system along some selected time-scale. A common consideration is how to determine whether a simulation has been running long enough. Many measures such as the root-mean-squared-deviation (RMSD), root-mean-squared-fluctuation (RMSF), and radius of gyration can be used to continuously monitor simulation progress, but these are in many cases not sufficient and need to be supplemented with some knowledge of the system under study.

In certain circumstances, short simulation times are sufficient as the conformational changes are known in an 'end-to-end' fashion. The ends of such a conformational change should be structurally characterised (i.e., 3-dimensional structures of the starting and end point should be available). As an example, only 50 ns of

simulation time (very short in terms of modern computing) was needed to simulate the expected conformational shift in a lactate dehydrogenase.[47] In this case, the simulation was guided by experimental structure restraints, reducing simulation time. The purpose of such simulations is to characterise a known conformational change in atomistic detail. In most cases, however, a given conformational change is not known in an 'end-to-end' fashion, where only one 3-dimensional structure is available. The extent of simulation required in these cases is in general much less clear. While a helpful rule-of-thumb is that the simulation should be at least 10-times longer than the time-scale of the slowest biochemical process,[48] it is of course difficult to know how long such a process is.

An illustrative example of how long MD simulations can become is in the form of the superfamily of rat sarcoma virus (RAS) proteins, of which there are more than 150 members.[49] These proteins are GTPases, which are proteins that have an increased affinity to bind to effector proteins when GTP is bound. RAS proteins are highly-flexible, and so are a good candidate for study by MD simulations. Recently, a study into all KRAS (a sub-family of RAS proteins) G12 missense mutations was performed, where an aggregate of 170 μs of simulation time was required to elucidate the complex structural dynamics.[50]

In another recent study, an astonishing 1.45 ms of simulation time was required to study the dynamic orientation of a wild-type KRAS (another sub-family of RAS proteins) against membranes of various lipid compositions.[51] In practice, the extent of simulation required is best considered in the context of what conformational changes are hypothesised, although how long a simulation should be remains a challenging question with no distinct answer.

Many other computational techniques are supplemented with a time-dependent methodology such as MD. A usual procedure when studying biochemical reactivity with hybrid QM/MM is to

start with an ensemble of structures generated by a MD simulation and then perform calculations on all of these structures.[52] Exactly how many structures should be sampled remains a questions, but this will be explored in more detail in Chapter 5. In any case, conformational sampling is absolutely essential in studying biochemical reactivity, making MD simulations a useful and rather efficient tool for exploring conformational space prior to reactivity studies. Such a conformational ensemble is also very useful from a CADD perspective, where molecular docking can be applied to the ensemble in order to consider drug-receptor binding in a quasi-dynamic fashion.[53]

1.4 Biochemical Processes and Systems Studied

In this thesis, two biochemical problems were investigated. Biochemical problems are those for which computational methodology was used to gain insight into a biochemical process which has not yet been fully understood using experimental methods and/or previous computational investigations. The varying size and rate of the different processes being studied for each of these problems meant that the specific choice of computational modelling technique(s) was problem-dependent, which will be explored in the relevant chapters.

Both biochemical problems investigated in this thesis included proteins which are a member of a class of proteins called heme proteins, or hemoproteins. Hemoproteins encompass all proteins which contain one or more heme prosthetic group(s) – a common enzyme cofactor that consists of a substituted porphine ring with chelated iron (see Figure 1.1).[54] The three types of heme shown in Figure 1.1 vary depending on their porphine substitution, and the oxidation state of the central iron. This oxidation state is dependent on the reduction/oxidation potential of the particular hemoprotein, which is defined by a number of factors.[54]

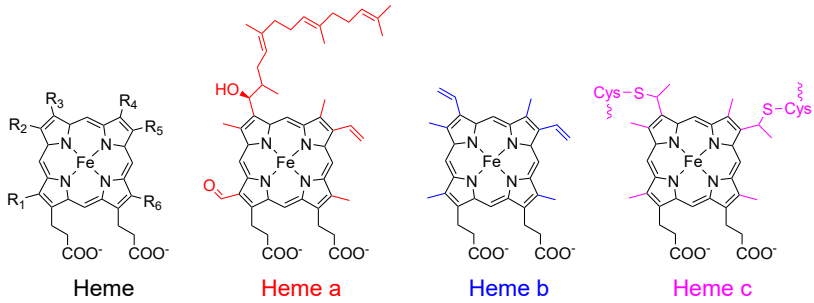

Figure 1.1: *The three most common types of heme (RHS: heme a, heme b, and heme c) shown alongside the common features shared between them (LHS: heme). Shown in red, blue, and pink are the different R-groups for heme a, heme b, and heme c, respectively. The oxidation state of the iron atoms are not given.*

Based on a search of the PDB curated in 2008,[55] heme b and heme c are (by a relatively large margin) the most common types of heme groups found in hemoproteins, with a comparatively small number of heme a. Of the three heme groups shown in Figure 1.1, heme a and heme b are non-covalently bound to their respective proteins, whereas heme c is covalently bound to the protein *via* cysteine residues.[56,57] It is worth highlighting that the set of heme groups shown in Figure 1.1 is not exhaustive, and there are many other lesser-occurring types of heme such as heme o, heme d_1, mesoheme, and siroheme, to name a few.[54] Irrespective of the heme group, the central iron atom is frequently coordinated to either one or two amino acids, with the particular amino acid(s) coordinated usually being related to the function of the heme group. Histidine is the most common amino acid ligand, with cysteine, methionine, and tyrosine being other candidates.[55]

Hemoproteins have diverse physiological functions including oxygen transport (myoglobin,[58] hemoglobin[59]), catalysis (heme peroxidase,[60] cytochrome P450[61]), electron transport (cytochromes[62]), intracellular signalling (soluble guanylate cyclase[63]), and metabolism (cystathionine β-synthase[64]). In this thesis, soluble guanylate cyclase (sGC) and cystathionine β-synthase (CBS) were the heme proteins investigated. The diverse

properties and functionality of heme proteins has cemented their ubiquity in research, both experimentally and computationally.

In addition to the biochemical problems, a biochemical case study was also selected, which was a reaction that was relatively well-understood, had quality 3-dimensional structures available, and had extensive previous computational studies. The case study was used for development of a new QM/MM method for obtaining reaction pathways, which can ultimately be used to obtain insight into the free energy of reaction. Specifically, the aim was to implement an algorithm into the QM/MM code, *QoMMMa*,[65] with the goal of facilitating the generation of quality reaction pathways for biochemical reactions. The case study chosen was the Claisen rearrangement catalysed by chorismate mutase, largely due to the detailed understanding of this reaction already provided by previous computations.

This subsection outlines biochemical background behind the biochemical problems and the case study investigated in this thesis. The specific details of the processes are left to the relevant chapters, here only a summary of their biochemical function is given.

1.4.1 Human Biochemical NO-signalling: Soluble Guanylate Cyclase [I]

In humans, biochemical NO signalling begins with activation of nitric oxide synthases (NOSs), which are hemoproteins that endogenously generate NO. Of the NOSs, there are three distinct isoforms of NOS – eNOS, iNOS, and nNOS.[66,67] Interestingly, while the three homologous enzymes have different sequences, they

[I] Part of the text in this subsection has been taken, expanded, and modified from: A. M. Rozza, M. Papp, N. R. McFarlane, J. N. Harvey, J. Oláh, *Chem. Eur. J.* **2022**, 28, e202200930. The original text in the publication was written by N. R. McFarlane and A. M. Rozza but has been sufficiently edited such that it does not greatly resemble the original text.

show a high degree of conservation in their active sites containing heme b groups.[68] NOSs catalyse the production of NO from l-arginine with citrulline by-product (see Figure 1.2).[69]

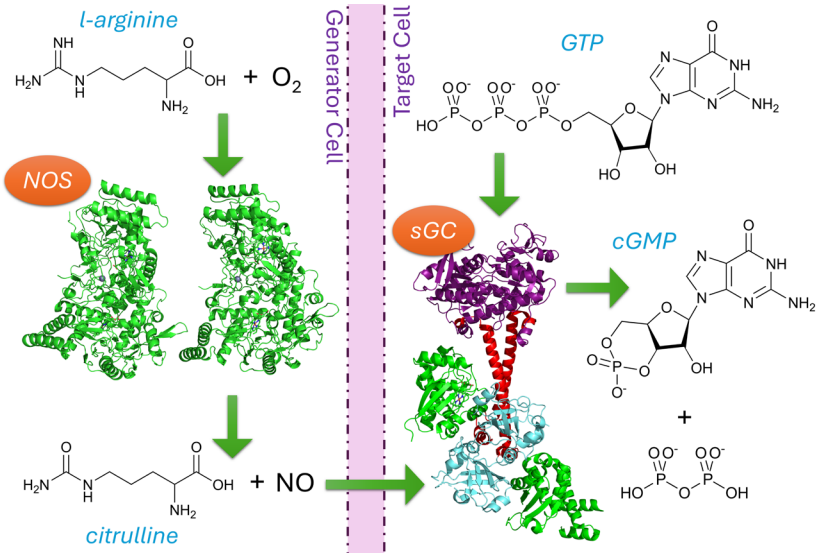

Figure 1.2: *A simplified illustration of the relationship between NOSs and the NO-sGC-cGMP pathway. The 3D structure of NOS was based on pdb-id:4NOS[68] which is an iNOS isoform of NOSs, and the 3D structure of sGC was based on pdb-id:6JT0[70]. Both were generated using PyMOL.[71] Enzymes and substrates are labelled with orange and blue, respectively. Other than the phosphate groups, all species are given in their neutral form.*

Following its production, NO gas may then diffuse very quickly across the membrane of the target cell due to its small size.[72] In humans, the only known NO-activated protein is sGC. Activation of sGC caused by binding of NO to a peripheral heme b leads to increased production of cyclic guanosine monophosphate (cGMP) from guanosine triphosphate (GTP).[73] While this conversion does occur regardless of the presence of NO, NO binding gives a 200-fold increase in cGMP production, categorising sGC as an NO-sensor protein. CO may also bind to sGC, giving an increase in activity of only a few-fold.[74]

Because of the link between NO binding to sGC and cGMP production, it is often called the NO-sGC-cGMP pathway; this pathway plays a key role in a downstream signalling cascade resulting in regulation of a host of physiological functions including vasodilation, wound healing, neurotransmission and effective immune response, among others.[66,75–77] Given that there is currently only one FDA-approved sGC stimulator drug, *riociguat*,[78] there is evidently much that remains to be understood in sGC.

Human sGC is the largest protein investigated in this thesis and has an overall mass of ~ 149 kDa, with a residue count of 1309.[70] It has a dimeric (alpha and beta), four-domain structure, with the binding of NO occurring at the beta N-terminal β H-NOX (Heme-nitric-oxide/oxygen binding) domain. The remaining three domains are the Per-Arnt-Sim (PAS) domain, a coiled-coil domain, and the C-terminal catalytic cyclase domain (see Figure 1.3).[73]

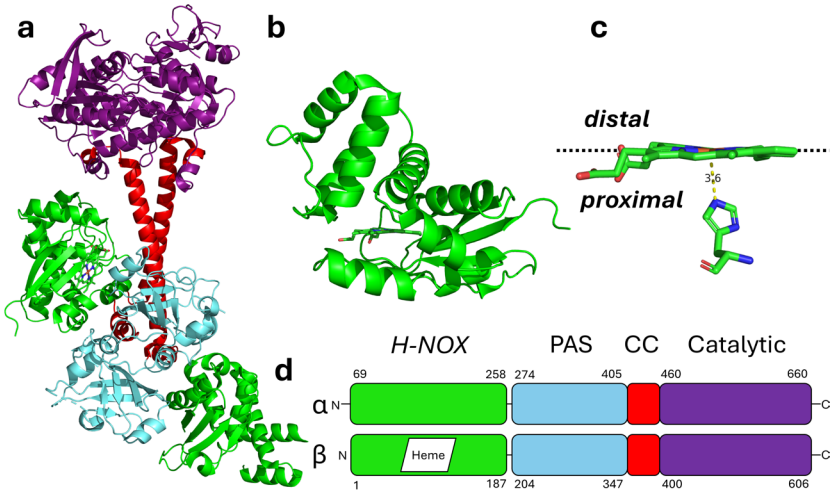

***Figure 1.3**: **a.** overall structure of human sGC in the basal state with the 4 domains having the colours shown in figure inset d; **b.** structure of the β H-NOX domain of human sGC showing the heme group; **c.** the heme group found in the β H-NOX domain of human sGC showing the proximally-coordinated histidine residue; **d.** the domain organisation of the α and β monomers of human sGC. The figures were based on pdb-id:6JT0[70] and generated using PyMOL.[71]*

In the catalytic cyclase domain, GTP is cyclised to form cGMP.[63] The dynamics and mechanism of this reaction are well-understood,[79] but the details of how NO binds to the β H-NOX domain continues to be ambiguous. NO binding occurs at the heme group, which is ferrous heme b with a proximally coordinated histidine residue. What the signalling mechanism is, and the number of NO equivalents required to initiate the activation process remains vague.

1.4.2 Inhibition of the Transsulfuration Pathway: Cystathionine β-synthase [II]

In mammals, the transsulfuration pathway plays a key role in sulfur metabolism and in intracellular redox regulation.[80] At the heart of the transsulfuration pathway is the regulation of homocysteine and cysteine concentrations. Their concentrations are crucial from two opposing perspectives: homocysteine in high concentrations is toxic in mammals,[81] and cysteine plays an important part in many biochemical processes including the production of the antioxidant glutathione,[82] the biogenesis of iron-sulfur clusters,[83] and in the folding and stability of cellular proteins *via* structural disulfide bonds,[84] among many others. There is even some relatively new, albeit contentious, evidence that cysteine has a cytoprotective function on tumour growth.[85]

The transsulfuration pathway formally begins with homocysteine, but as an overall metabolic pathway, it is mediated by the methionine cycle (see Figure 1.4). The methionine cycle regulates the conversion of dietary methionine to homocysteine. The first step of the methionine cycle is the reaction of methionine with ATP to form S-adenosyl-methionine (AdoMet) catalysed by methionine

[II] Part of the text in this subsection has been taken, expanded, and modified from N. R. McFarlane, J. Gui, J. Oláh, J. N. Harvey, *Phys. Chem. Chem. Phys*, **2024**, 26, 16579-16588. The original text in the publication was written by N. R. McFarlane but has been sufficiently edited such that it does not greatly resemble the original text.

adenosyltransferase. The second step is a demethylation reaction, forming S-adenosylhomocysteine, which, completing the cycle, can be hydrolysed back to homocysteine catalysed by S-adenosylhomocysteinase.[86–88] The now-formed homocysteine has two possible destinations, either back into the methionine cycle and converted back to methionine *via* catalysis by betaine homocysteine methyltransferase, or to be taken as the first metabolite in the transsulfuration pathway.

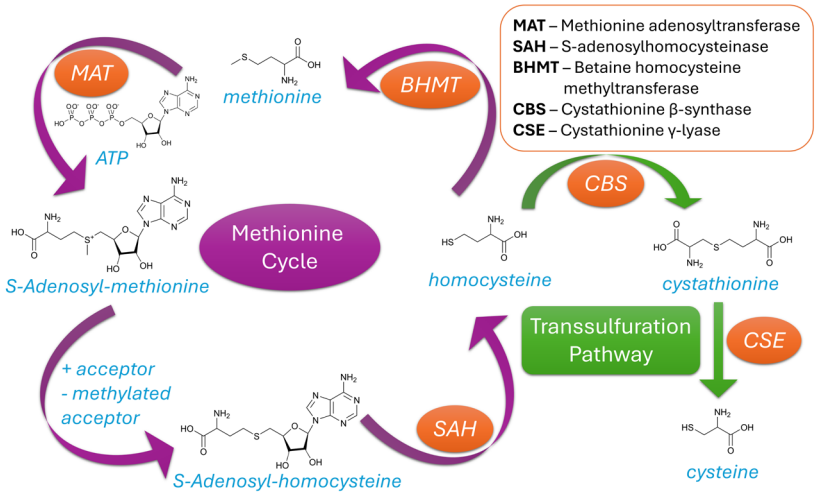

Figure 1.4: *A simplified scheme of the relationship between the methionine cycle and the transsulfuration pathway. Enzymes and substrates are labelled with orange and blue, respectively. The key boxed in orange provides the full names of all enzymes involved in the scheme.*

Starting with homocysteine, the canonical reactions of the transsulfuration pathway involve the reaction of homocysteine with serine to form cystathionine catalysed by of CBS, followed by the conversion of cystathionine to cysteine catalysed by cystathionine γ-lyase (CSE) (see Figure 1.4).[89] Both CBS and CSE are pyridoxal 5'-phosphate (PLP) dependent enzymes, which is to say that the reactions catalysed by the enzymes use PLP as a reactive cofactor. Despite their similarities in reactivity, CBS is a hemoprotein, while CSE is not. The presence of heme in CBS

ultimately has implications for regulation of reactivity, as will be explored in detail below. Shown in Figure 1.5 is the set of reactions catalysed by CBS and CSE, as well as reactions catalysed by both. The chemistry in both CBS and CSE is relatively similar, with the main difference being that in CSE, γ-carbon based elimination and substitution reactions can occur, whereas in CBS only β-carbon based chemistry is observed. In fact, in CSE, β-carbon based chemistry may also occur, and these plausible side-reactions are reflected in Figure 1.5. From a mechanistic perspective, PLP-dependent enzymes have previously been studied in detail and are largely well understood.[90–92]

Besides the regulation of homocysteine and cysteine concentrations by the canonical reactions of CBS and CSE, the transsulfuration pathway also leads to the production of H_2S, *via* a number of side-reactions found in both enzymes – see Figure 1.5.[93,94] For the sake of completeness, the only other enzyme known to synthesise H_2S in humans is 3-mercaptopyruvate sulfurtransferase,[95] but where CBS and CSE both contain PLP at the heart of their catalysis,[89] 3-mercaptopyruvate sulfurtransferase does not. While traditionally considered as purely a toxic gas,[96] there is an increasing wealth of evidence to suggest that H_2S plays a regulatory role in a number of critical biochemical processes such as in vasodilation and neuromodulation.[97–102] The importance of H_2S as an intracellular signalling molecule is a subject of increasing interest, so much so that H_2S has been suggested to join the esteemed company of CO and NO as the third endogenous gas signalling molecule.[103]

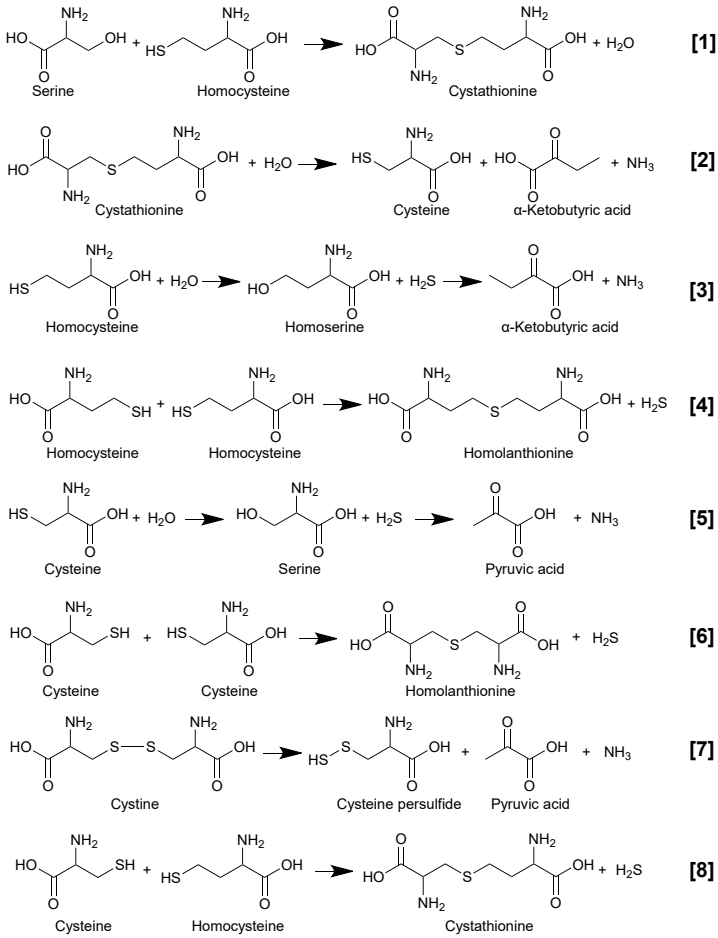

Figure 1.5: *The set of reactions found in the transsulfuration pathway. Reaction [1] is the canonical reaction found in CBS, reaction [2] is the canonical reaction found in CSE, and reactions [3] and [4] are two additional side-reactions found exclusively in CSE. Reactions [5-8] are the side reactions which can be found in both enzymes. All species are given in their neutral form.*

Despite the functional similarity between CBS and CSE, they are differentiated by their mechanisms for inhibition and regulation. Total, albeit reversible, inhibition of CSE has been shown to be caused by binding of dl-propargylglycine to the active PLP cofactor.[64,104] This dl-propargylglycine-dependent mechanism of CSE inhibition is certainly useful from a rational design standpoint,

but is less informative from an endogenous regulation perspective. On the other hand, CBS exhibits complex and even cooperative mechanisms of allosteric regulation and inhibition, based on binding of commonly-occurring intracellular molecules.[105]

There are three known regulatory/inhibitory pathways observed in in CBS. The first is the binding of AdoMet to the so-called AdoMet domain (highlighting again the intrinsic relationship between the transsulfuration pathway and the methionine cycle, which produces AdoMet – see Figure 1.4). The latter two are both related to the heme group in the heme domain: redox switching of the oxidation state of the heme group, and the binding of small gas molecules to the heme group.

Turning first to the AdoMet domain, upon binding of AdoMet, the activity of the CBS has been reported to increase up to five-fold.[106] The increase in activity is explained by improved access for the reactants shown in Figure 1.5 to the PLP pocket of the enzyme *via* motion of secondary structures of up to 8 Å.[106,107] This increased rate of diffusion of reactants has not yet been investigated computationally. These are relatively large-scale structural changes and substrates, and the distance between the AdoMet binding pocket and the catalytic PLP core is approximately 40 Å.[106] Bearing this in mind, these motions could be rather challenging to study computationally, perhaps other than for very long simulation times or accelerated simulations. In contrast to the structural changes caused by AdoMet binding, the allosteric changes localised around the heme group are expected to be on a smaller scale, with the distance between the heme group and the PLP catalytic core being only 20 Å.[108]

Human CBS (hCBS) is a biological dimer with an overall mass of ~ 127 kDa, with a residue count of 1,126.[107] The monomers are sequence-identical, with minor differences in their secondary structures. Found in the heme domain, the heme group is heme b, and is in a hydrophobic pocket near the surface of the enzyme (see

Figure 1.6). The heme group is 6-coordinate, with a rather unusual coordination sphere of a distal histidine and a proximal cysteine. Electronically, the heme may exist as either ferric or ferrous, but is ferric in the basal state.[105]

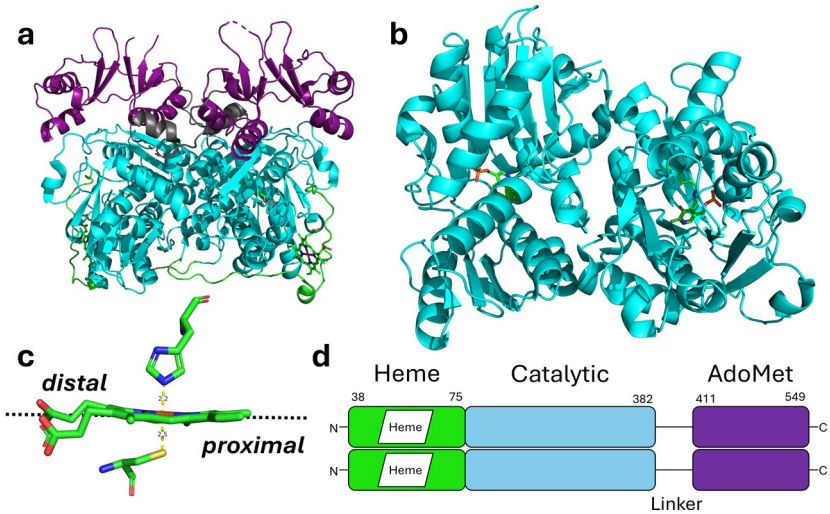

Figure 1.6: a. *overall structure of hCBS with the 3 domains having the colours shown in figure inset d;* ***b.*** *structure of the catalytic domain of hCBS showing the PLP catalytic site and how it is covalently bound to the backbone via a lysine residue;* ***c.*** *the heme group found in the heme domain of hCBS showing the proximally-coordinated cysteine and distally-coordinated histidine residues;* ***d.*** *the domain organisation of the two monomers of hCBS, which are sequence-identical. The figures were based on pdb-id:4COO[107] and generated using PyMOL.[71]*

Returning to the heme domain, it is known that the role of the heme cofactor is not directly catalytic, so its role in hCBS has been proposed to be regulatory.[108] This regulation is realised in the rate of the PLP catalysed reactions in hCBS shown Figure 1.5. As mentioned above, there have been two proposed methods for this regulation. The first is by redox regulation, whereby ferrous heme has been found to be approximately 2-fold less active than ferric heme based on comparison of UV-Vis. Soret peaks between oxidised and reduced hCBS by titanium citrate.[109] The heme group in hCBS is therefore redox-active, thus hCBS can be considered as

a heme redox sensor protein.[109–111] While the redox regulation of the heme is certainly important in intracellular modulation of catalytic activity, it is quite well-understood. On the other hand, the binding of small gas molecules CO and NO binding to the heme group, and what their relative binding affinities are, is less accurately known.

1.4.3 Pericyclic Claisen Rearrangement: Chorismate Mutase [III]

Chorismate mutase (CM) is involved in the final step of the shikimate metabolic pathway, leading to the production of aromatic amino acids phenylalanine and tyrosine (see Figure 1.7).[112,113] The complete shikimate pathway is a seven-step pathway which is found in fungal, bacterial, and higher-plant systems. Here, however, only the final step of the shikimate pathway (chorismate to prephenate) and the subsequent aromatic amino acid biosynthetic pathways are illustrated in Figure 1.7.

Given that the shikimate pathway exists only in fungal, bacterial, and higher-plant systems, a greater understanding of the CM catalysed step could in principle lead to development of novel and selective fungicides, antibiotics, and herbicides.[114] Accordingly, the mechanism of the Claisen rearrangement has been well characterised for a long time.[115] The mechanism does not involve any covalent bond formation between the substrate and the enzyme (making it well-suited for QM/MM studies). CM is a member of the rare class of pericyclic catalysing enzymes.[116,117] The reaction mechanism is illustrated in Figure 1.8.

[III] The text here has been taken, expanded, and modified from N. R. McFarlane, J. N. Harvey, *Phys. Chem. Chem. Phys.*, **2024**, 26, 5999-6007. The original text in the publication was written by N. R. McFarlane, but has been sufficiently edited such that it does not greatly resemble the original text.

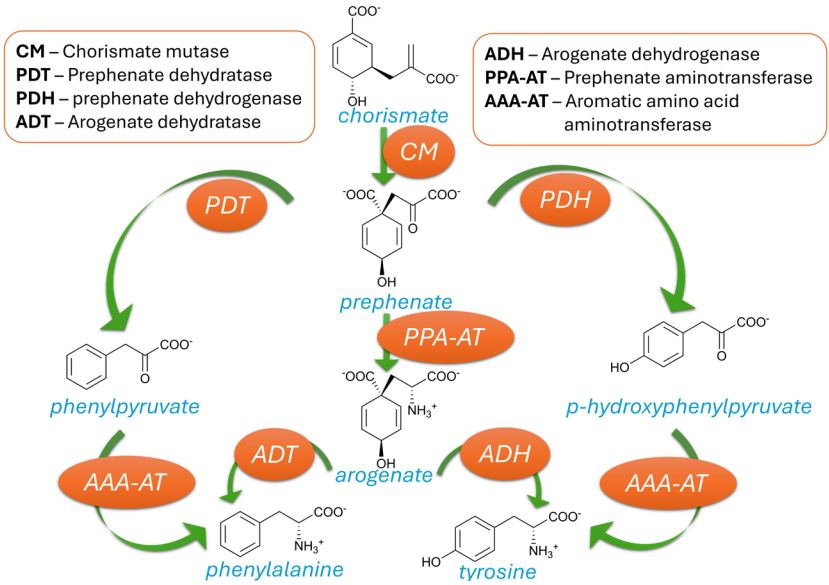

Figure 1.7: *A simplified scheme of the final step of the shikimate pathway (chorismate to prephenate) and the subsequent biosynthetic pathways leading to amino acids phenylalanine and tyrosine. Enzymes and substrates are labelled with orange and blue, respectively. The keys boxed in orange provides the full names of all enzymes involved in the scheme. All substrates are given in their charged forms.*

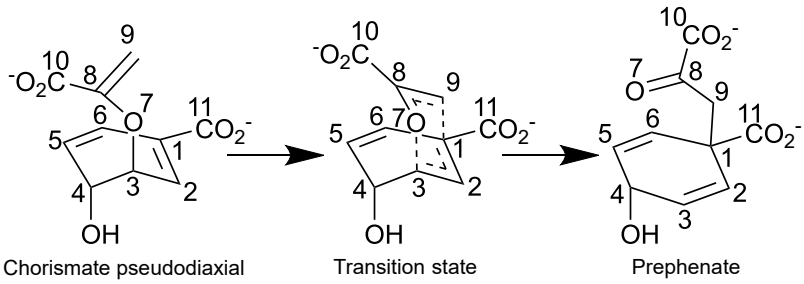

Figure 1.8: *The reaction mechanism found in CM. In this reaction, the reactant, chorismate, (in its pseudodiaxial conformation) converts to the product, prephenate, through the indicated transition state via a pericyclic rearrangement. Atom numbering (omitting the carboxylate and hydroxyl oxygen(s)) is also given.*

CM is expressed in many systems, and accordingly there are crystal structures available for *Bacillus subtilis*,[118,119] *Escherichia coli*,[120] *Mycobacterium tuberculosis*,[121] and yeast,[122] among

others. Perhaps the most computationally investigated form of chorismate mutase is that from *Bacillus subtilis* (BsCM).[118] BsCM is a homotrimer, where the pericyclic rearrangement occurs at the interface between two monomers of the homotrimer. The homotrimer is relatively small, with an overall mass of ~ 45 kDa, and a residue count of 381 (see Figure 1.9). BsCM has been the system-of-choice for many technique validations presumably due to the availability of high-resolution crystal structures containing transition state analogues (TSA) and other experimental data.

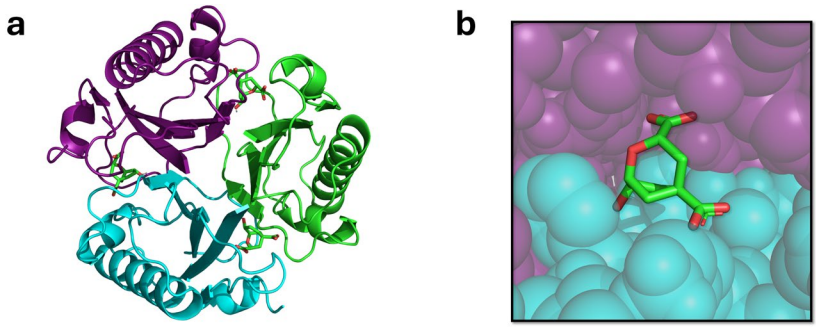

Figure 1.9: a. *overall structure of BsCM, which is a homotrimer, with each of the monomers shown with blue, purple, and green, respectively;* ***b.*** *the transition state analogue sandwiched between chains A and B, shown in blue and purple, respectively. The figures were based on pdb-id:2CHT[118] and generated using PyMOL.[71]*

1.5 Conclusions and Objectives

From a general perspective, this thesis discusses the use of multiscale modelling approaches for studying biochemical processes, evidencing the requirement for complementary techniques to fully understand biochemistry. The objectives of this thesis are split into three parts based on Chapters 3, 4, and 5 because there is not a distinct overarching goal of this work.

In Chapter 3, to understand how NO gas binds to the β H-NOX domain of sGC, static methodologies were applied. Accurate QM cluster calculations on various heme derivatives were performed,

thereby obtaining energetics for the several plausible NO-activation routes. These calculations were coupled with the use of QM/MM calculations sampled from a MD simulation to obtain reaction pathways for rate-limiting steps of the NO-activation routes.

The energetics and reaction pathways, while useful, do not allow insight into the dynamic nature of sGC, so MD simulations were applied to investigate (i) the nature of NO diffusion into the heme region of the β H-NOX domain, and (ii) the conformational changes following NO binding ultimately leading to sGC activation. For the former, MD simulations of the β H-NOX domain saturated with NO gas were used to give insight into the diffusibility of NO gas. For the latter, extended MD simulations of the same domain enabled an understanding of the conformational changes in the amino acid hydrogen bonding network caused by NO binding

In Chapter 4, the objectives were rather similar to that of Chapter 3, but for a different hemoprotein. Using static techniques, the first goal was to explore the free energy landscape associated with CO/NO binding to the heme group in hCBS. Accurate QM cluster calculations on the CO/NO bound heme derivatives were used to obtain the free energy surface of the inhibitory pathways. These free energies allow for estimation of the probability that a given species would be found at standard state, and more importantly allow verification of the experimentally proposed models.

A second and more important goal was to understand what conformational changes are associated with CO/NO binding to hCBS. Based on the verified spectroscopic models of the inhibited forms, extended MD simulations were performed to study the conformational changes caused by CO/NO binding and how this causes inhibition of the activity of hCBS.

In Chapter 5, the details of how the QM/MM algorithm that was implemented is provided. In this context, the relevance of other

algorithms used to study of biochemical reactivity is also detailed. The Claisen rearrangement catalysed by chorismate mutase was used as an initial case study for algorithm validation for several reasons, but primarily, the abundance of QM/MM calculations performed on this system mean that comparative assessment of the new QM/MM algorithm developed herein was enabled. In addition, from a practical perspective, the system preparation and setup for the Claisen rearrangement catalysed by chorismate mutase is relatively easy, having only a small number of atoms and a relatively simple reaction.

The goal of this chapter was to develop a new QM/MM methodology to study biochemical reactivity, and to use the reaction described above as a case study. In addition to the main goal of the chapter, in the algorithm validation process, since many data points were generated, some statistical analysis into QM/MM sampling was also afforded. In this analysis, the underlying distribution of QM/MM-calculated activation energies has been investigated, providing insight into the extent of sampling required in these types of studies.

1.6 References

1 A. Wright, J. Provost, J. A. Roecklein-Canfield and E. Bell, *Biochem. Mol. Biol. Educ.*, 2013, **41**, 302–308.

2 K. A. Johnson and R. S. Goody, *Biochemistry*, 2011, **50**, 8264–8269.

3 J. Gunawardena, *FEBS J.*, 2014, **281**, 473–488.

4 S. K. Burley, H. M. Berman, J. M. Duarte, Z. Feng, J. W. Flatt, B. P. Hudson, R. Lowe, E. Peisach, D. W. Piehl, Y. Rose, A. Sali, M. Sekharan, C. Shao, B. Vallat, M. Voigt, J. D. Westbrook, J. Y. Young and C. Zardecki, *Biomolecules*, 2022, **12**, 1425.

5 J. Jumper and D. Hassabis, *JAMA*, 2023, **330**, 1425–1426.

6 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

7 M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis and S. Velankar, *Nucleic Acids Res.*, 2022, **50**, D439–D444.

8 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, 1–3.

9 O. Carugo and K. Djinović-Carugo, *PLOS Biol.*, 2023, **21**, e3002187.

10 D. Petrović, X. Wang and B. Strodel, *Proteins*, 2018, **86**, 935–944.

11 A. P. Lyubartsev and A. L. Rabinovich, *Biochim. Biophys. Acta BBA - Biomembr.*, 2016, **1858**, 2483–2497.

12 F.-Y. Lin and A. D. MacKerell, *Methods Mol. Biol. Clifton NJ*, 2019, **2022**, 21–54.

13 J.-R. Hill, C. M. Freeman and L. Subramanian, in *Reviews in Computational Chemistry*, John Wiley & Sons, Ltd, 2000, pp. 141–216.

14 Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J.-P. Piquemal and P. Ren, *Annu. Rev. Biophys.*, 2019, **48**, 371–394.

15 K. Goloviznina, J. N. Canongia Lopes, M. Costa Gomes and A. A. H. Pádua, *J. Chem. Theory Comput.*, 2019, **15**, 5858–5871.

16 O. Borodin, *J. Phys. Chem. B*, 2009, **113**, 11463–11478.

17 V. S. Inakollu, D. P. Geerke, C. N. Rowley and H. Yu, *Curr. Opin. Struct. Biol.*, 2020, **61**, 182–190.

18 T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama and A. C. T. van Duin, *Npj Comput. Mater.*, 2016, **2**, 1–14.
19 Y. Han, D. Jiang, J. Zhang, W. Li, Z. Gan and J. Gu, *Front. Chem. Sci. Eng.*, 2016, **10**, 16–38.
20 B. D. Bekono, A. N. Sona, D. B. Eni, L. C. O. Owono, E. Megnassan and F. Ntie-Kang, *Phys. Sci. Rev.*, 2023, **8**, 457–477.
21 V. Salmaso and S. Moro, *Front. Pharmacol.*, 2018, **9**, 923.
22 T. Borowski, M. Quesne and M. Szaleniec, in *Advances in Protein Chemistry and Structural Biology*, Elsevier, 2015, vol. 100, pp. 187–224.
23 X. Sheng, M. Kazemi, F. Planas and F. Himo, *ACS Catal.*, 2020, **10**, 6430–6449.
24 F. Himo, *J. Am. Chem. Soc.*, 2017, **139**, 6780–6786.
25 M. R. A. Blomberg, T. Borowski, F. Himo, R. Z. Liao and P. E. M. Siegbahn, *Chem. Rev.*, 2014, **114**, 3601–3658.
26 S. B. Han, H. S. Ali and S. P. de Visser, *Inorg. Chem.*, 2021, **60**, 4800–4815.
27 M. G. Quesne and S. P. de Visser, in *Challenges and Advances in Computational Chemistry and Physics*, Springer, 2019, vol. 29, pp. 439–462.
28 L. H. Chen, G. L. Kenyon, F. Curtin, S. Harayama, M. E. Bembenek, G. Hajipour and C. P. Whitman, *J. Biol. Chem.*, 1992, **267**, 17716–17721.
29 R. Sevastik and F. Himo, *Bioorganic Chem.*, 2007, **35**, 444–457.
30 P. E. M. Siegbahn and F. Himo, *WIREs Comput. Mol. Sci.*, 2011, **1**, 323–336.
31 K. H. Hopmann and F. Himo, *J. Chem. Theory Comput.*, 2008, **4**, 1129–1137.
32 L. Hu, P. Söderhjelm and U. Ryde, *J. Chem. Theory Comput.*, 2011, **7**, 761–777.
33 M. J. Field, P. A. Bash and M. Karplus, *J. Comput. Chem.*, 1990, **11**, 700–733.
34 A. Warshel and M. Levitt, *J. Mol. Biol.*, 1976, **103**, 227–249.
35 M. G. Quesne, T. Borowski and S. P. De Visser, *Chem. - Eur. J.*, 2016, **22**, 2562–2581.
36 H. Hirao, in *Challenges and Advances in Computational Chemistry and Physics*, Springer, 2019, vol. 29, pp. 463–486.
37 F. Claeyssens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel and H. J. Werner, *Angew. Chem. - Int. Ed.*, 2006, **45**, 6856–6859.
38 R. Lonsdale, J. N. Harvey and A. J. Mulholland, *Chem. Soc. Rev.*, 2012, **41**, 3025–3038.
39 H. J. Kulik, J. Zhang, J. P. Klinman and T. J. Martínez, *J. Phys. Chem. B*, 2016, **120**, 11381–11394.
40 L. Hu, J. Eliasson, J. Heimdal and U. Ryde, *J. Phys. Chem. A*, 2009, **113**, 11793–11800.
41 D. Flaig, M. Beer and C. Ochsenfeld, *J. Chem. Theory Comput.*, 2012, **8**, 2260–2271.
42 G. Jindal and A. Warshel, *J. Phys. Chem. B*, 2016, **120**, 9913–9921.
43 H. M. Senn and W. Thiel, *Angew. Chem. Int. Ed.*, 2009, **48**, 1198–1229.
44 T. Tuttle and W. Thiel, *J. Phys. Chem. B*, 2007, **111**, 7665–7674.
45 L. Hu, P. Söderhjelm and U. Ryde, *J. Chem. Theory Comput.*, 2013, **9**, 640–649.
46 S. A. Hollingsworth and R. O. Dror, *Neuron*, 2018, **99**, 1129–1143.

47 A. Hospital, J. R. Goñi, M. Orozco and J. L. Gelpí, *Adv. Appl. Bioinforma. Chem.*, 2015, **8**, 37–47.
48 D. M. Zuckerman, *Annu. Rev. Biophys.*, 2011, **40**, 41–62.
49 K. Wennerberg, K. L. Rossman and C. J. Der, *J. Cell Sci.*, 2005, **118**, 843–846.
50 T. Pantsar, S. Rissanen, D. Dauch, T. Laitinen, I. Vattulainen and A. Poso, *PLOS Comput. Biol.*, 2018, **14**, e1006458.
51 C. Neale and A. E. García, *Biophys. J.*, 2020, **118**, 1129–1141.
52 M. Sanchez-Martinez, E. Marcos, R. Tauler, M. Field and R. Crehuet, *J. Phys. Chem. B*, 2013, **117**, 14261–14272.
53 M. De Vivo, M. Masetti, G. Bottegoni and A. Cavalli, *J. Med. Chem.*, 2016, **59**, 4035–4061.
54 J. Everse, in *Encyclopedia of Biological Chemistry (Second Edition)*, eds. W. J. Lennarz and M. D. Lane, Academic Press, Waltham, 2013, pp. 532–538.
55 C. J. Reedy, M. M. Elvekrog and B. R. Gibney, *Nucleic Acids Res.*, 2008, **36**, D307–D313.
56 T. Li, H. L. Bonkovsky and J. Guo, *BMC Struct. Biol.*, 2011, **11**, 13.
57 V. Guallar and B. Olsen, *J. Inorg. Biochem.*, 2006, **100**, 755–760.
58 J. B. Wittenberg, *Physiol. Rev.*, 1970, **50**, 559–636.
59 E. Antonini and M. Brunori, *Annu. Rev. Biochem.*, 1970, **39**, 977–1042.
60 J. Vlasits, C. Jakopitsch, M. Bernroitner, M. Zamocky, P. G. Furtmüller and C. Obinger, *Arch. Biochem. Biophys.*, 2010, **500**, 74–81.
61 T. Omura, *Biochem. Biophys. Res. Commun.*, 1999, **266**, 690–698.
62 F. Hannemann, A. Bichet, K. M. Ewen and R. Bernhardt, *Biochim. Biophys. Acta BBA - Gen. Subj.*, 2007, **1770**, 330–344.
63 E. R. Derbyshire and M. A. Marletta, *Annu. Rev. Biochem.*, 2012, **81**, 533–559.
64 Q. Sun, R. Collins, S. Huang, L. Holmberg-Schiavone, G. S. Anand, C.-H. Tan, S. van-den-Berg, L.-W. Deng, P. K. Moore, T. Karlberg and J. Sivaraman, *J. Biol. Chem.*, 2009, **284**, 3076–3085.
65 J. N. Harvey, *Faraday Discuss.*, 2004, **127**, 165.
66 C. Farah, L. Y. M. Michel and J.-L. Balligand, *Nat. Rev. Cardiol.*, 2018, **15**, 292–316.
67 U. Förstermann and W. C. Sessa, *Eur. Heart J.*, 2012, **33**, 829–837.
68 T. O. Fischmann, A. Hruza, X. D. Niu, J. D. Fossetta, C. A. Lunn, E. Dolphin, A. J. Prongay, P. Reichert, D. J. Lundell, S. K. Narula and P. C. Weber, *Nat. Struct. Biol.*, 1999, **6**, 233–242.
69 D. J. Stuehr, *Annu. Rev. Pharmacol. Toxicol.*, 1997, **37**, 339–359.
70 Y. Kang, R. Liu, J.-X. Wu and L. Chen, *Nature*, 2019, **574**, 206–210.
71 L. Schrödinger and W. DeLano, PyMOL (version 2.4.0) http://www.pymol.org/pymol 2020.
72 D. D. Thomas, L. A. Ridnour, J. S. Isenberg, W. Flores-Santana, C. H. Switzer, S. Donzellie, P. Hussain, C. Vecoli, N. Paolocci, S. Ambs, C. Colton, C. Harris, D. D. Roberts and D. A. Wink, *Free Radic. Biol. Med.*, 2008, **45**, 18–31.
73 W. R. Montfort, J. A. Wales and A. Weichsel, *Antioxid. Redox Signal.*, 2017, **26**, 107–121.
74 G. Wu, E. Martin, V. Berka, W. Liu, E. D. Garcin and A.-L. Tsai, *J. Inorg. Biochem.*, 2021, **214**, 111267.
75 K. M. Sanders and S. M. Ward, *Am. J. Physiol.*, 1992, **262**, G379-392.

76 N. B. Fernhoff, E. R. Derbyshire and M. A. Marletta, *Proc. Natl. Acad. Sci.*, 2009, **106**, 21602–21607.
77 K. Kajiya, R. Huggenberger, I. Drinnenberg, B. Ma and M. Detmar, *FASEB J.*, 2008, **22**, 530–537.
78 C. Koress, K. Swan and P. Kadowitz, *Curr. Hypertens. Rep.*, 2016, **18**, 42.
79 R. R. Khalid, A. Maryam, O. U. Sezerman, E. Mylonas, A. R. Siddiqi and M. Kokkinidis, *Sci. Rep.*, 2020, **10**, 9488.
80 J. I. Sbodio, S. H. Snyder and B. D. Paul, *Br. J. Pharmacol.*, 2019, **176**, 583–593.
81 J. Perła-Kaján, T. Twardowski and H. Jakubowski, *Amino Acids*, 2007, **32**, 561–572.
82 Y. Chen, H. G. Shertzer, S. N. Schneider, D. W. Nebert and T. P. Dalton, *J. Biol. Chem.*, 2005, **280**, 33766–33774.
83 R. Lill and U. Mühlenhoff, *Annu. Rev. Cell Dev. Biol.*, 2006, **22**, 457–486.
84 C. S. Sevier and C. A. Kaiser, *Nat. Rev. Mol. Cell Biol.*, 2002, **3**, 836–847.
85 H.-F. Zhang, R. I. K. Geltink, S. J. Parker and P. H. Sorensen, *Trends Cell Biol.*, 2022, **32**, 800–814.
86 J. L. Palmer and R. H. Abeles, *J. Biol. Chem.*, 1979, **254**, 1217–1226.
87 G. L. Cantoni, *J. Am. Chem. Soc.*, 1952, **74**, 2942–2943.
88 G. D. Markham and M. A. Pajares, *Cell. Mol. Life Sci. CMLS*, 2009, **66**, 636–648.
89 S. M. Aitken, P. H. Lodha and D. J. K. Morneau, *Biochim. Biophys. Acta BBA - Proteins Proteomics*, 2011, **1814**, 1511–1517.
90 A. Salvà, J. Donoso, J. Frau and F. Muñoz, *J. Phys. Chem. A*, 2003, **107**, 9409–9414.
91 E. F. Oliveira, N. M. F. S. A. Cerqueira, P. A. Fernandes and M. J. Ramos, *J. Am. Chem. Soc.*, 2011, **133**, 15496–15505.
92 J. Liang, Q. Han, Y. Tan, H. Ding and J. Li, *Front. Mol. Biosci.*, 2019, **6**, 4.
93 A. P. Landry, J. Roman and R. Banerjee, *Curr. Opin. Struct. Biol.*, 2021, **71**, 27–35.
94 T. Chiku, D. Padovani, W. Zhu, S. Singh, V. Vitvitsky and R. Banerjee, *J. Biol. Chem.*, 2009, **284**, 11601–11612.
95 N. Shibuya, Y. Mikami, Y. Kimura, N. Nagahara and H. Kimura, *J. Biochem. (Tokyo)*, 2009, **146**, 623–626.
96 R. J. Reiffenstein, W. C. Hulbert and S. H. Roth, *Annu. Rev. Pharmacol. Toxicol.*, 1992, **32**, 109–134.
97 K. Abe and H. Kimura, *J. Neurosci. Off. J. Soc. Neurosci.*, 1996, **16**, 1066–1071.
98 R. Hosoki, N. Matsuki and H. Kimura, *Biochem. Biophys. Res. Commun.*, 1997, **237**, 527–531.
99 W. Zhao, J. Zhang, Y. Lu and R. Wang, *EMBO J.*, 2001, **20**, 6008–6016.
100 W. Zhao and R. Wang, *Am. J. Physiol.-Heart Circ. Physiol.*, 2002, **283**, H474–H480.
101 M. Whiteman, J. S. Armstrong, S. H. Chu, S. Jia-Ling, B.-S. Wong, N. S. Cheung, B. Halliwell and P. K. Moore, *J. Neurochem.*, 2004, **90**, 765–768.
102 C. Szabo, *Biochem. Pharmacol.*, 2018, **149**, 5–19.
103 R. Wang, *FASEB J.*, 2002, **16**, 1792–1798.

104 C. Steegborn, T. Clausen, P. Sondermann, U. Jacob, M. Worbs, S. Marinkovic, R. Huber and M. C. Wahl, *J. Biol. Chem.*, 1999, **274**, 12675–12684.
105 M. Meier, M. Janosik, V. Kery, J. P. Kraus and P. Burkhard, *EMBO J.*, 2001, **20**, 3910–3916.
106 J. Ereño-Orbea, T. Majtan, I. Oyenarte, J. P. Kraus and L. A. Martínez-Cruz, *Proc. Natl. Acad. Sci.*, 2014, **111**, E3845–E3852.
107 T. J. McCorvie, J. Kopec, S.-J. Hyung, F. Fitzpatrick, X. Feng, D. Termine, C. Strain-Damerell, M. Vollmar, J. Fleming, J. M. Janz, C. Bulawa and W. W. Yue, *J. Biol. Chem.*, 2014, **289**, 36018–36030.
108 S. Singh, P. Madzelan and R. Banerjee, *Nat Prod Rep*, 2007, **24**, 631–639.
109 S. Taoka, S. Ohja, X. Shan, W. D. Kruger and R. Banerjee, *J. Biol. Chem.*, 1998, **273**, 25179–25184.
110 R. Banerjee and C. Zou, *Arch. Biochem. Biophys.*, 2005, **433**, 144–156.
111 S. Taoka, B. W. Lepore, Ö. Kabil, S. Ojha, D. Ringe and R. Banerjee, *Biochemistry*, 2002, **41**, 10454–10461.
112 V. Tzin and G. Galili, *Arab. Book Am. Soc. Plant Biol.*, 2010, **8**, e0132.
113 E. Pérez, M. B. Rubio, R. E. Cardoza, S. Gutiérrez, W. Bettiol, E. Monte and R. Hermosa, *Front. Microbiol.*, 2015, **6**, 1181.
114 P. M. Dewick, *Nat. Prod. Rep.*, 1998, **15**, 17–58.
115 H. Goerisch, *Biochemistry*, 1978, **17**, 3700–3705.
116 P. R. Andrews, G. D. Smith and I. G. Young, *Biochemistry*, 1973, **12**, 3492–3498.
117 S. D. Copley and J. R. Knowles, *J. Am. Chem. Soc.*, 1987, **109**, 5008–5013.
118 Y. M. Chook, H. Ke and W. N. Lipscomb, *Proc. Natl. Acad. Sci. U. S. A.*, 1993, **90**, 8600–8603.
119 Y. M. Chook, J. V. Gray, H. Ke and W. N. Lipscomb, *J. Mol. Biol.*, 1994, **240**, 476–500.
120 A. Y. Lee, P. A. Karplus, B. Ganem and J. Clardy, *J. Am. Chem. Soc.*, 1995, **117**, 3627–3628.
121 R. Qamra, P. Prakash, B. Aruna, S. E. Hasnain and S. C. Mande, *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.*, 2005, **61**, 473–475.
122 N. Sträter, G. Schnappauf, G. Braus and W. N. Lipscomb, *Structure*, 1997, **5**, 1437–1452.

2 Theoretical Background

"It is nice to know that the computer understands the problem. But I would like to understand it too."

- Eugene Wigner, unknown

In this chapter, the theoretical details behind the computational modelling techniques which have been used will be given. Unlike in other chapters, at the start of each major subsection, a small number of selected textbooks and key references are given which were used to inform the background material throughout. In-text references are also used, but these are for specific reference material rather than general background material.

2.1 Molecular Mechanics

The background material in this subsection is based on the following textbooks: A. Leach, Molecular Modelling: Principles and Applications, 2nd Edition, ***2001****; D. Frenkel & B. Smit, Understanding Molecular Simulation, 2nd Edition,* ***2002****; A. Hinchliffe, Molecular Modelling for Beginners, 2nd Edition,* ***2003****.*

MM has been a long-serving workhorse for the study of chemical systems, largely due to the simplicity of its formulation. Perhaps due to this simplicity, "precomputer" MM does not point to one highly influential work, but rather to several communications published in the late 1940s and early 1950s by Hill, Westheimer, Mayer, and Barton.[1] The MM method is concerned with the calculation of chemical system properties *via* a classical description of their molecular interactions. Within this method, a principal approximation is that the total system energy is calculated as an additive sum of several terms which account for the molecular interactions, expressed in a simplified form in Equation (2.1) as a function of the nuclear coordinates, $\boldsymbol{R}$, and a set of predefined constants, $\boldsymbol{k}$.

$$E(\boldsymbol{R},\boldsymbol{k}) = \sum E_{bonds} + \sum E_{valence\ angles} + \sum E_{improper\ dihedral\ angles} + \sum E_{proper\ dihedral\ angles} + \sum E_{non-bonded} \quad \textbf{(2.1)}$$

The first four energy terms are the energy contributions defined by bonded, intramolecular interactions, and the last term are those defined by non-bonded, intermolecular interactions, as illustrated in Figure 2.1.

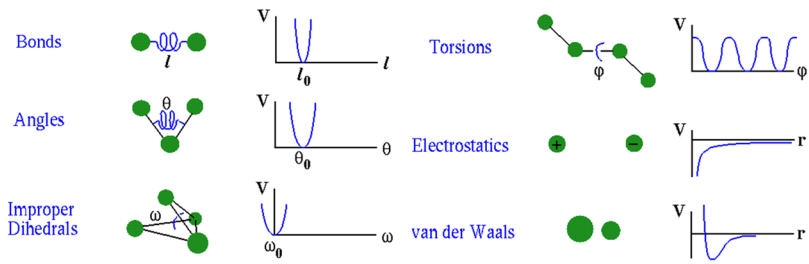

Figure 2.1: *Illustration of the types of interactions and associated potential energy functions found in a molecular mechanics calculation. Figure adapted from: Steinbach, P. (2019).* The Empirical Potential Energy Function. *National Institute of Allergy and Infectious Diseases. (accessed June 27, 2024).*

In addition to the nuclear coordinates, $\boldsymbol{R}$, and some definition of their atomic connectivity, all terms are dependent on many constants, $\boldsymbol{k}$, defined in a so-called force field, $\boldsymbol{V}(\boldsymbol{R})$, which describes the type and magnitude of the interactions between atom types. Looking more closely into the terms shown in Equation (2.1), several of the bonded interactions are predicated on the assumption that vibrational oscillations of molecules follow a harmonic potential which obeys Hooke's law, $V(x) = \frac{1}{2}kx^2$, where x represents the displacement, and k is a force constant describing how readily a given displacement returns to its equilibrium value. This assumption can be readily extended to the bond (r_{ij}) and valence angle (θ_{ijk}) energy contributions – see

Equations (2.2) and (2.3) for the bond and valence angle contributions, respectively.

$$E_{ij} = \frac{1}{2} k_{ij} \left(r_{ij} - r_{ij,\rightleftharpoons} \right)^2 \qquad \textbf{(2.2)}$$

$$E_{ijk} = \frac{1}{2} k_{ijk} \left(\theta_{ijk} - \theta_{ijk,\rightleftharpoons} \right)^2 \qquad \textbf{(2.3)}$$

Equations (2.2) and (2.3) require only two constants in the force field: force constants, k_{ij} and k_{ijk}, and the equilibrium values, $r_{ij,\rightleftharpoons}$ and $\theta_{ijk,\rightleftharpoons}$. If all equilibrium angles, $\theta_{ijk,\rightleftharpoons}$, are accurately parametrized in the force field, planar systems should be correctly described. Nevertheless, in many force fields, an additional term used to enforce planarity based on the form of improper dihedrals (ω_{ijkl}) using the harmonic potential assumption is often also included in the potential energy function– see Equation (2.4).[2]

$$E_{ijkl,improper} = \frac{1}{2} k_{ijkl} \left(\omega_{ijkl} - \omega_{ijkl,\rightleftharpoons} \right)^2 \qquad \textbf{(2.4)}$$

Like the bonds and valence angles, Equation (2.4) requires only force constants, k_{ijkl}, and equilibrium values, $\omega_{ijkl,\rightleftharpoons}$. Proper dihedral angles (ϕ_{ijkl}) are, however, often not described by the harmonic potential assumption, but instead by a sum of periodic solutions, where cosine functions are used – see Equation (2.5).

$$E_{ijkl,proper} = \sum_{n=1}^{m} k_{ijkl,n} (1 + \cos(n\phi_{ijkl} - \delta_{ijkl,n})) \qquad \textbf{(2.5)}$$

Where in Equation (2.5), $k_{ijkl,n}$ are the force constants, and $\delta_{ijkl,n}$ are the cosine phase shifts, which vary subject on the multiplicity, m. Depending on the type of dihedral angle, various values of m can be used, thereby giving different sums of cosine multiplicities. In practice, many force fields employ a “triple-cosine” form of

Equation (2.5), where $i = 3$. Of course, each multiplicity requires appropriate parametrization of $k_{ijkl,n}$ and $\delta_{ijkl,n}$, so an accurate description of proper dihedral terms can be challenging to achieve.

The final term shown in Equation (2.1) encompasses all non-bonded interactions, which includes the electrostatic (the first term in brackets) and van der Waals interaction (the second term in brackets) between atoms – see Equation (2.6).

$$E_{ij,non-bonded} = \left(\frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}}\right) + \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) \qquad \textbf{(2.6)}$$

The electrostatic term is described by classical Coulomb attraction and repulsion, and the van der Waals interactions are described using a Lennard-Jones potential. Unlike those for the bonded interactions, Equation (2.6) requires no force constants, but several other parameters are needed, including: the dielectric permittivity in vacuo, ε_0, the effective atomic charges, q_i and q_j, and two parameters which account for the dispersion interaction and short-range repulsion, A_{ij} and B_{ij}, respectively.

Practically speaking, calculating the potential energy of a given system using Equation (2.1) requires summation of a relatively simple, albeit large, set of classically defined equations containing a set of constants. This simplicity means that MM potential energy functions can be calculated exactly, and rather quickly. However, the accuracy is largely dependent on the force-field, $\boldsymbol{V}(\boldsymbol{R})$, and, to a lesser extent, the nuclear coordinates, $\boldsymbol{R}$. If the selected force field does not describe the system accurately, then the results of the MM calculation will reflect this poor choice.

2.2 Quantum Mechanics

The background material in this subsection is based on the following textbooks: C. J. Cramer, Essentials of Computational Chemistry, 2nd Edition, ***2004****; T. Helgaker, P. Jorgensen, & J. Olsen, Molecular Electronic-Structure Theory, 1st Edition,* ***2000****; A. Szabo, & N. S. Ostlund, Modern Quantum Chemistry, rev. 1st Edition,* ***1989****; R. M. Martin, Electronic Structure, 1st Edition,* ***2004****; D. S. Sholl, & J. A. Steckel, Density Functional Theory, 1st Edition,* ***2009****.*

Postulated as early as 1925, and subsequently leading to the Nobel Prize in Physics in 1933, the Schrödinger equation lies at the heart of QM. Expressed in a somewhat loose fashion, the formation of this equation is a result of the notion that quantum-scale objects cannot be fully described by continuous waves, or even as particles, but rather as a wave-particle duality. Therefore, an equation which can reliably predict the properties of a quantum system must include a wave-like description of particles, i.e., a wavefunction. This equation is expressed in bra-ket notation in its time-independent non-relativistic form in Equation (2.7).

$$\hat{H}|\Psi\rangle = E|\Psi\rangle \tag{2.7}$$

Where in Equation (2.7), $\hat{H}$ is the Hamiltonian operator (which will be explained in more detail below), E is the total system energy, and $|\Psi\rangle$ is the wavefunction. In the QM formalism, it is a primary assumption that the state of a given (chemical) system can be completely specified by a wavefunction.

Putting the fundamentals aside, in practice, solving Equation (2.7) usually requires working under the Born-Oppenheimer approximation where the motion of electrons is taken to be uncoupled to the motion of nuclei (due to the much lesser mass of electrons compared to nuclei). This assumption enables construction of the electronic, or 'clamped-nuclei', Schrödinger equation, as shown in Equation (2.8), where $\boldsymbol{r}$ and $\boldsymbol{R}$ represent the electron and nuclear coordinates, respectively.

$$\hat{H}_e|\Psi_e(\boldsymbol{r},\boldsymbol{R})\rangle = E_e|\Psi_e(\boldsymbol{r},\boldsymbol{R})\rangle \tag{2.8}$$

Under this approximation, the wavefunction reduces to the electronic wavefunction, $|\Psi_e(\boldsymbol{r},\boldsymbol{R})\rangle$, dependent on the coordinates of both the nuclei ($\boldsymbol{R}$) and electrons ($\boldsymbol{r}$). Similarly, the Hamiltonian operator, $\hat{H}$, reduces to the electronic Hamiltonian operator, $\hat{H}_e$. Expressed in atomic units for a system of n electrons and N nuclei, this reduction is shown in going from Equation (2.9) to Equation (2.10).

$$\hat{H} = \underbrace{-\sum_{I}^{N}\frac{\nabla_I^2}{2M_I}}_{\substack{nuclear\\ kintetic\\ energy}} \underbrace{-\sum_{i}^{n}\frac{\nabla_i^2}{2}}_{\substack{electronic\\ kinetic\\ energy}} \underbrace{+\sum_{I}^{N}\sum_{J>I}^{N}\frac{Z_I Z_J}{|R_I - R_J|}}_{\substack{nuclear\\ repulsion\\ energy}} \underbrace{-\sum_{I}^{N}\sum_{i}^{n}\frac{Z_I}{|R_I - r_i|}}_{\substack{electronic-nuclear\\ attraction\\ energy}} \underbrace{+\sum_{I}^{n}\sum_{j>i}^{n}\frac{1}{|r_i - r_j|}}_{\substack{electronic\\ repulsion\\ energy}} \tag{2.9}$$

$$\hat{H}_e = -\sum_{i}^{n}\frac{\nabla_i^2}{2} - \sum_{I}^{N}\sum_{i}^{n}\frac{Z_I}{|R_I - r_i|} + \sum_{I}^{n}\sum_{j>i}^{n}\frac{1}{|r_i - r_j|} \tag{2.10}$$

Where in Equation (2.9) and (2.10), ∇^2 is the Laplacian operator, Z and M represent the charge and mass of the nuclei, R and r represent the coordinates of nuclei and electrons, respectively, counters i and j are used for electrons, and counters I and J are used for nuclei. Evidently, in the electronic Hamiltonian operator (Equation (2.10)), the nuclear kinetic energy term has been eliminated since the nuclei are static. Moreover, nuclear repulsion energy is a constant which is simply added to the total system energy after obtaining the electronic energy, simplifying the problem yet further.

Despite the above simplifications, due to the simultaneous dependence of $|\Psi_e\rangle$ on all electron coordinates, Equation (2.8) cannot be solved exactly for anything more than a single electron containing system. As such, to obtain reasonable solutions for so-called many-body molecular systems, an approximate wavefunction is necessitated. It is, however, important to be able to assess whether this approximate wavefunction is a good representation of the 'true' wavefunction. This is often evaluated subject to the variational theorem, a systematic approach used to improve the method from which the approximate wavefunction was constructed. Under this theorem, the lowest energy possible of a given system is the ground state energy, so then the exact wavefunction will always give a lower energy than an approximate wavefunction. Exploiting this idea, one can be sure that if the energy is lower for a given approximate wavefunction than a previous approximation, then the current approximate wavefunction is superior to the one previously used.

2.2.1 Hartree-Fock Method

An early attempt to solve the electronic Schrödinger equation is the Hartree-Fock (HF) approach, which aims to approximate the wavefunction as a single Slater determinant of spin-orbitals (see Equation (2.11)) such that several of the fundamental postulates of quantum mechanics are satisfied: Ψ must be continuous, normalisable, anti-symmetric and square-integrable.

$$\Psi_{HF}(x_1, x_2, \dots, x_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_n(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_n(x_n) \end{vmatrix} \quad \textbf{(2.11)}$$

Where in Equation (2.11), each element in the determinant, $\psi_i(x_i)$, represents a spin-orbital for the spatial (r_i) and spin (σ_i) coordinates of electron i (where r_i and σ_i are grouped together as x_i), and n represents the total number of electrons. In correspondence with the Pauli exclusion principle, the formulation

of Ψ_{HF} intrinsically accounts for anti-symmetry as exchange of any two rows in the determinant shown in Equation (2.11) results in $-\Psi_{HF}$.

Exploring further the ideas presented in Equation (2.11), the HF method is based on the idea that $\widehat{H}_e$ can be decoupled into one-electron operators in a mathematically satisfying manner. With reference to Equation (2.10), it is evident that the electronic kinetic energy and the electronic-nuclear repulsion energy (the first two terms) depend only on the coordinates of one electron at a time. On the other hand, the electronic repulsion energy (the final term) is simultaneously dependent on the coordinates of two electrons and so is intrinsically inseparable. These ideas result in the formation of the one-electron Fock operator, $f(x_i)$, as shown in Equation (2.12).

$$f(x_i) = -\frac{\nabla_i^2}{2} - \sum_{I}^{N} \frac{Z_I}{|R_I - r_i|} + \upsilon_{HF}(x_i) \qquad \textbf{\textit{(2.12)}}$$

Where in Equation (2.12), $f(x_i)$ is the one-electron Fock operator for electron i, ∇^2 is the Laplacian operator, Z represents the charge of the nuclei, R and r represent the coordinates of nuclei and electrons, respectively, and I is used as a counter for nuclei. $\upsilon_{HF}(x_i)$ represents the effect that electron i experiences because of the positions of all other electrons approximated by an average mean-field approach. This term is formed of two components (not mathematically shown here): the Coulomb interaction which accounts for electron-electron repulsion, and the exchange interaction which is a direct result of the anti-symmetry principle used to construct the Slater determinant given in Equation (2.11). Equation (2.12) can be recast into a set of Schrödinger-like equations (Equation (2.13)).

$$f(x_i)|\psi_i\rangle = \varepsilon_i|\psi_i\rangle \quad \textbf{(2.13)}$$

Where in Equation (2.13), ψ_i is the ith HF orbital and ε_i is the corresponding orbital energy. While a given formulation of Equation (2.13) is indeed a one-electron problem, since $\upsilon_{HF}(x_i)$ depends on the positions of all electrons and hence the entire wavefunction, $\Psi_{HF}(x_1, x_2, \dots, x_n)$, the problem is non-linear and so necessitates an iterative solution. Determination of the set of spin-orbitals which correspond to the minimum energy wavefunction makes use of the self-consistent field (SCF) method, where an initial set of guess spin-orbitals is iteratively optimised such that the electronic energy decreases – exploiting the previously mentioned variational theorem.

Such an iterative SCF procedure is often achieved using the Roothaan-Hall method,[3] where molecular orbitals (MOs) are expressed as linear combinations of basis functions,[IV] each with associated coefficients, as shown in Equation (2.14).

[IV] In QM calculations, the basis functions, φ_k, represent atomic orbitals centred on the nuclei, so taking linear combinations of such basis functions is the linear combination of atomic orbitals (LCAO) approach. The questions of which type of atomic orbitals to use remains, however. From a fundamental perspective, Slater-type orbitals (STOs) are the obvious choice, since these are exact solutions to the Schrödinger equation for hydrogen-like atoms. However, due to the computational cost associated with calculating two-electron integrals based on STOs, it was proposed that STOs could be approximated as linear combinations of Gaussian-type orbitals (GTOs). Most basis sets are based on linear combinations of GTOs, and one can keep improving upon a simple linear combination by adding polarisation and diffuse functions to improve the description of highly polarised or delocalised electronic structures. The mathematical details of basis sets are not explored herein, but in general, if the number of basis functions K is increased, then the accuracy of the calculation will increase. There is always the consideration that increasing the number of basis functions also incurs a larger virtual space, which of course increases the computational cost, specifically when using correlated methods.

$$\psi_i = \sum_{k=1}^{K} c_{ik}\varphi_k \qquad \textbf{(2.14)}$$

Where in Equation (2.14), for a given orbital i, spin-orbitals are given as ψ_i, the basis functions (of which there are K in total) are given as φ_k, and c_{ik} represent the associated expansion coefficients which are iteratively optimised throughout the SCF procedure. The Roothaan-Hall equations may then be solved, as expressed in matrix form in Equation (2.15).

$$\boldsymbol{FC} = \boldsymbol{SC}\varepsilon \qquad \textbf{(2.15)}$$

Where in Equation (2.15), matrix $\boldsymbol{F}$ contains elements $F_{pq} = \langle\varphi_p|\hat{F}|\varphi_q\rangle$, matrix $\boldsymbol{C}$ includes the expansion coefficients shown in Equation (2.14), matrix $\boldsymbol{S}$ is an overlap matrix with entries $S_{pq} = \langle\varphi_p|\varphi_q\rangle$, and ε is a diagonal matrix whose entries represent the eigenvalues corresponding to the energy of the occupied orbitals. Since the matrix $\boldsymbol{F}$ is dependent on the expansion coefficient matrix $\boldsymbol{C}$, Equation (2.15) must be solved iteratively.

Perhaps the foremost weakness of the HF approach is that electron correlation is not accounted for, which is defined as the difference between the HF solution (assuming that it has been solved at the complete basis set limit) and the exact solution to the Schrödinger equation.[4] In practice, the HF approach can make good estimations of properties which have no significant changes in electronic structure such as relative conformer stabilities. However, if one wishes to evaluate complex electronic structures, then an improved approach which explicitly includes electron correlation is necessitated.

2.2.2 Electron Correlation

From a broad perspective, for a system of n electrons, the motion of electron i will be influenced by the position of all $n-1$ remaining electrons due to Coulomb repulsion. There are many methods which build on the fundamentals of the HF approach but also include this electron correlation explicitly. Some examples of these so-called post-HF methods include: configuration interaction theory,[5] multi-configurational self-consistent field methods,[6] and coupled-cluster (CC) methods,[7] among several others. Most of these methods will not be discussed in detail; the discussion will begin with the fundamentals of CC and build to local-CC methods due to their relevance in this thesis.

2.2.2.1 Coupled-cluster Methods

CC methods express the wavefunction, Ψ, as an exponential ansatz, where the so-called CC operator, $\hat{T}$, (shown in Equation (2.17)) acts upon a reference wavefunction, Ψ_0, as shown in bra-ket notation in Equation (2.16).

$$|\Psi\rangle = e^{\hat{T}}|\Psi_0\rangle \tag{2.16}$$

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 \ldots \hat{T}_n \tag{2.17}$$

Where in Equation (2.17), n is the number of electrons in the system. Using a complete CC operator, $\hat{T}$, in which all possible excitation operators are accounted for would give the limit of a full configuration interaction method, but this is typically not possible due to the large computational cost, so the CC operator is usually truncated to include only some of these operators. For example, the CCSD method (where $\hat{T} = \hat{T}_1 + \hat{T}_2$, and only single and double excitations are included) and the CCSDT method (where $\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3$, and single, double, and triple excitations are included) see some use, but the CCSD method is often not accurate enough, and

the CCSDT method can quickly become prohibitively expensive. Perhaps the most extensively used variant is the CCSD(T) method, bridging the gap between CCSD and CCSDT. In the CCSD(T) method,[8] single and double excitations are included in the same way as the CCSD method, but the triple excitations are included in a non-iterative fashion based on perturbation theory. The CCSD(T) method is often referred to as the gold standard for main-group QM calculations, being able to accurately calculate molecular properties and bond energies at a lesser cost than CCSDT, but with sufficient accuracy in most cases.

Other than the degree of excitation operators included in the exponential ansatz given in Equation (2.16) above, since CC methods are, by definition, single-reference, the choice of reference wavefunction, Ψ_0, can have a large impact on the quality of the results. Traditionally, CCSD(T) calculations use a HF reference wavefunction, but density functional theory (see Section 2.2.3) reference wavefunctions may also be used, in some cases with equivalent performance and are able to sidestep the common complications associated with HF solutions.[9] On the other hand, it has been shown that in other systems the choice of Ψ_0 can have a large impact,[10] so density functional theory reference wavefunctions may not always be suitable.

2.2.2.2 Local Coupled-cluster Methods

First proposed by Pulay in 1983,[11] the concept behind local correlation methods is that correlation energy between orbitals decreases rapidly as the distance between them increases, so by using localised MOs, the computational cost can be greatly reduced while still recovering most of the correlation energy. Transformation of delocalised canonical MOs to localised MOs can be achieved in several ways, but the occupied and virtual orbitals are usually transformed in different ways.

The occupied orbital space can be relatively easily localised using techniques such as Foster-Boys[12] or Pipek-Mezey,[13] among many

other options. Ultimately, these techniques produce a set of localised MOs which can be attributed to individual atoms or bonds.

Transformation of virtual orbital space is in general rather more involved, and the virtual orbitals can be compressed with various techniques such as non-orthogonal projected atomic orbitals, orbital specific virtual orbitals, or as pair natural orbitals (PNOs). Moving to CC calculations with higher quality basis sets, the extent of the virtual orbital space becomes increasingly larger than the occupied orbital space, and so much of the computational expense and memory consumption can be attributed to the cluster amplitudes involving this virtual space. Therefore, compression of virtual orbital space can be an effective way to reduce computational cost. In general, PNOs provide the most compact description of virtual orbital space, and so their use can greatly reduce the number of cluster amplitudes and ultimately simplify the CC wavefunction shown in Equation (2.16).[14,15]

In addition to the decrease in cost associated with compression of the delocalised canonical MOs, another commonly-used simplification used in localised correlated methods is to organise the localised orbital pairs into domains,[16] such that they are separated into strong and weak pairs, usually based on some distance or atomic connectivity criteria. The strong pairs can then be calculated in a fully canonical fashion as outlined above, and the weak pairs using a lower level of theory such as MP2.

A problem associated with local correlation methods is that when transforming the virtual orbital space from canonical MOs into localised MOs, due to the compression described above, the resulting virtual orbitals may be non-orthogonal with respect to one another. As a result, the non-iterative approach to calculating the triple excitations used in the CCSD(T) method described above is not suitable, and so an iterative approach must again be used. While one could, in principle, use the fully iterative method (or

LCCSDT) to overcome this problem, even with the reduced orbital space and hierarchical domain structure, this may be prohibitively expensive. Accordingly, the (T_0) and (T_1) approximations have been developed. In the (T_0) approximation, off-diagonal Fock matrix elements are neglected (i.e., removing non-orthogonal virtual orbitals), meaning that the triples correction can be calculated in a similar manner to the CCSD(T) method. This method has been reported to recover up to 97% of the triples correction in a very efficient manner.[17] In the (T_1) approximation, the first iteration of the fully iterative T is used to approximate the triples correction, providing an improved estimation of the triples correction relative to the (T_0) approach, albeit at a higher computational cost.

In practice, one of the most widely used variants of local CC is the Domain-based Localised Pair Natural Orbital (DLPNO) approach,[17] where PNO localised orbitals are used and organised into domains, as outlined above. In this thesis, two variants of the DLPNO method were used: the cheaper DLPNO-CCSD(T_0), and the more expensive DLPNO-CCSD(T_1).

2.2.3 Density Functional Theory

The discussion has thus far focused on Wave Function Theory, but what if instead of describing the electrons as discrete entities, an electron density could be used? A theory based on this idea offers an obvious simplification of the many-body problem, where the requirement for a complex and elusive wavefunction can be evaded.

This concept is largely ascribed to the seminal work of Hohenberg and Kohn in 1964,[18] where they propose two theorems which to-this-day form the basis of density functional theory (DFT). The first theorem that they propose is that the ground state energy is uniquely a function of the ground state electron density, $E_0(\rho_0(r))$, where $\rho_0(r)$ describes the ground state electron density in space (Equation (2.18)).

$$E_0(\rho_0(r)) = \underbrace{T(\rho_0(r))}_{\substack{electronic\\kinetic\\energy}} + \underbrace{V_{ee}(\rho_0(r))}_{\substack{electronic\\repulsion\\energy}} + \underbrace{V_{en}(\rho_0(r))}_{\substack{electronic-nuclear\\attraction\\energy}} \qquad \textbf{(2.18)}$$
$$= F_{HK}(\rho_0(r)) + V_{en}(\rho_0(r))$$

In other words, the ground state energy is described by a set of functions, or so-called functionals, that only take $\rho_0(r)$ as an argument. As shown in Equation (2.18), the Hohenberg-Kohn (or HK) functional, $F_{HK}(\rho_0(r))$, is often written as a sum of $T(\rho_0(r))$ and $V_{ee}(\rho_0(r))$. This grouping is because finding a solution to this functional requires further development of the ideas presented by Hohenberg and Kohn, where it must be in some way approximated. On the other hand, $V_{en}(\rho_0(r))$ can be calculated in a straightforward fashion ($\int \rho_0(r) v_{ext.}(r)\, dr$) using a potential arising from an external field, $v_{ext.}(r)$, which corresponds to the nuclei in this case.

The second theorem of Hohenberg and Kohn relates to deriving a variational principle analogous to the one used in the HF approach, enabling systematic improvement of the system energy. Taking a given electron density, $\tilde{\rho}$, it must be that the ground state energy, E_0, will only be found if the exact ground state electron density, ρ_0, is used. Otherwise, the energy obtained using F_{HK} will always be an upper bound to ρ_0, as expressed in Equation (2.19).

$$E_0 \leq E(\tilde{\rho}) = V_{en}(\tilde{\rho}) + F_{HK}(\tilde{\rho}) \qquad \textbf{(2.19)}$$

In principle, Equation (2.19) only holds for an exact definition of F_{HK} (demonstrating that if the exact ground state electron density is known, then the exact ground state energy can be obtained), and it is possible that the various definitions F_{HK} which will be explored below may give energies lower than the ground state energy. Nevertheless, the iterative minimisation procedure allowed by the variational theorem is practically useful and so is still used.

At this point it is worth highlighting that the above discussion of Hohenberg and Kohn's work is pure DFT which, as a result of their second theorem, allows the exact ground state energy to be found if the exact ground state electron density is known. However, other than for a select set of small systems for which the exact wavefunction can be obtained, a central problem relates to finding the best approximation of F_{HK}. Even in those cases where F_{HK} can be found exactly, it is not transferable to other systems. An approximate expression for F_{HK} enables an approximation to the ground state electron density to be found, but since the ground state density will be an approximation, then so will the ground state energy. In this way, various formulations of F_{HK} are strictly speaking density functional approximations. Nevertheless, for the sake of convenience and convention, these formulations of F_{HK} will be referred to as DFT functionals throughout this thesis.

2.2.3.1 Kohn-Sham Density Functional Theory

The ideas presented by Kohn and Sham in 1965[19] form the basis of most modern DFT approaches which aim to approximate F_{HK} in the most cost-effective and accurate way. In this work, they propose to use a fictitious, non-interacting system of electrons (a so-called Kohn-Sham density, ρ_s). To understand their ideas better, it is useful to break F_{HK} back down into its component parts: $V_{ee}(\rho)$ and $T(\rho)$ (see Equation (2.18)). Starting with $V_{ee}(\rho)$, it can be written in the form of a classical-type electronic repulsion term, and a correction term, $\Delta V_{ee}(\rho)$, as shown in Equation (2.20).

$$V_{ee}(\rho) = \frac{1}{2}\iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + \Delta V_{ee}(\rho) \qquad \textbf{(2.20)}$$

As mentioned above, the first term in Equation (2.20) (henceforth termed $J(\rho)$) is grounded in classical mechanics and so can be calculated, but the correction term, $\Delta V_{ee}(\rho)$, is strictly non-classical and has contributions from electronic self-interaction,

exchange, and correlation. The latter component of F_{HK}, $T(\rho)$, is tackled in a HF-adjacent approach, where ρ_s is calculated in terms of Kohn-Sham orbitals, ψ^{KS} – expressed in Equation (2.21) for a system of n particles.

$$\rho_s(r) = \sum_{i=1}^{n} \left|\psi_i^{KS}\right|^2 \qquad \textbf{(2.21)}$$

Using these orbitals, a term for the kinetic energy of the system of n non-interacting particles can be expressed as shown in Equation (2.22).

$$T_s(\rho) = \sum_{i=1}^{n} \left\langle \psi_i^{KS} \middle| -\frac{\nabla_i^2}{2} \middle| \psi_i^{KS} \right\rangle \qquad \textbf{(2.22)}$$

Since these Kohn-Sham orbitals have no physical meaning, they do not strictly minimise the ground state energy, but instead build up the ground state density. In this way, to obtain the true kinetic energy, much like was the case for Equation (2.20), a correction term, $\Delta T(\rho)$, must be added, as shown in Equation (2.23).

$$T(\rho) = T_s(\rho) + \Delta T(\rho) \qquad \textbf{(2.23)}$$

Grouping Equations (2.20), (2.22) and (2.23) together into the Hohenberg-Kohn functional yields Equation (2.24).

$$F_{HK} = J(\rho) + T_s(\rho) + \underbrace{(\Delta V_{ee}(\rho) + \Delta T(\rho))}_{\substack{exchange-correlation \\ functional}} \qquad \textbf{(2.24)}$$

As outlined above, the first two terms in Equation (2.24) can be calculated exactly due to the assumption of the non-interacting system of electrons, but the final two terms remain unknown. These final two terms are often grouped together into a so-called

exchange-correlation functional, $E_{XC}(\rho)$, which can be calculated with many approximate approaches.

2.2.3.2 Exchange-correlation Functionals

In this subsection, some effort will be made to summarise and give appropriate examples of the main types of exchange-correlation functional (in increasing order of accuracy), but it is important to emphasize that this is by no means an exhaustive summary. It is also worth making clear that there is not a known form for $E_{XC}(\rho)$ and all the various types outlined below offer varying levels of accuracy for different systems and properties. In practice, if one wishes to obtain accurate properties with DFT for a given system, many different DFT functionals are tested and compared to a reference calculation or high-quality experimental data – a process usually known as benchmarking.

The local density approximation (LDA) is the simplest of the exchange-correlation functionals, in which a uniform electron gas expression is used to calculate the exchange and correlation terms of $E_{XC}(\rho)$, as shown in Equation (2.25).

$$E_{XC}{}^{LDA}(\rho) = \int \rho(r)\varepsilon_{XC}(\rho)\, dr \qquad \textbf{(2.25)}$$

Where in Equation (2.25), ε_{XC} represents the change and correlation energy per electron in a uniform electron gas of density, ρ. A related exchange-correlation functional is the local spin-density approximation (LSDA), which enables the treatment of open-shell molecules by separate treatment of α and β electrons. Due to the localisation of electrons around nuclei, electron clouds are not uniform in chemical systems, so LDA/LSDA is a crude approximation which is seldom used. Nevertheless, LDA/LSDA form the basis of many other functionals.

Clearly, the electron density is not homogeneous, so the LDA method can be improved upon by also including the gradient of the

electron gas in the integrand of $E_{XC}(\rho)$, known as a generalised gradient approach (GGA), as shown in Equation (2.26).

$$E_{XC}{}^{GGA}(\rho) = \int f_{XC}\big(\rho(r), \nabla\rho(r)\big)\, dr \qquad \textbf{(2.26)}$$

Where in Equation (2.26), f_{XC} is a function of both the electron density, ρ, and the gradient of the electron density, $\nabla\rho$, used to describe the exchange-correlation. In practice, these methods are widely used due to the excellent balance of performance and accuracy, assuming that appropriate benchmarking has been performed.[20] Some examples of GGA functionals that have been used in this thesis are BP86,[21,22] BLYP,[23] OLYP,[24] and PBE.[22]

A further improvement upon GGA functionals was to include a kinetic energy density, τ, to the exchange-correlation functional, enabling a more accurate treatment of the various types of chemical bonds. These types of functionals are of the meta-generalised gradient approximation (meta-GGA) family, as expressed in Equation (2.27).

$$E_{XC}{}^{meta-GGA}(\rho, \tau) = \int f_{XC}(\rho(r), \nabla\rho(r), \tau(r))\, dr \qquad \textbf{(2.27)}$$

Where in Equation (2.27), the kinetic energy, τ, is defined as a summation over all occupied Kohn-Sham orbitals, as shown in Equation (2.28).

$$\tau(r) = \frac{1}{2}\sum_{i}^{N_{occ.}} \left|\nabla\psi_i^{KS}(r)\right|^2 \qquad \textbf{(2.28)}$$

The meta-GGA approach offers yet another improvement over the previously mentioned functionals. The most popular example of a meta-GGA functional is TPSS,[25] which has been used in this thesis.

Since the exchange energy is very important for accurate calculations, yet another type of DFT functional was proposed, where a combination of exact exchange calculated by HF theory and a semi-local exchange and correlation calculated by DFT is used as the exchange-correlation functional. Theoretically grounded in the adiabatic connection,[26] these types of functionals are perhaps the most widely used branch of DFT functionals and are called hybrid functionals, as expressed in a general sense in Equation (2.29).

$$E_{XC}{}^{hybrid} = \alpha E_X{}^{HF} + (1-\alpha)E_X{}^{DFT} + E_C{}^{DFT} \qquad \textbf{(2.29)}$$

Where in Equation (2.29), α is a parameter which determines the extent of exact exchange which will be used in the functional. This family of functionals also includes the most popular of all exchange-correlation functionals, B3LYP, as shown in Equation (2.30).

$$E_{XC}{}^{B3LYP} = E_{XC}{}^{LSDA} + 0.2\left(E_X{}^{HF} - E_X{}^{LSDA}\right) + 0.72E_X{}^{B88} + 0.81E_c{}^{LYP} \qquad \textbf{(2.30)}$$

Where in Equation (2.30), B88 is an exchange functional developed by Becke,[27] and LYP is Lee, Yang, and Parr's correlation functional.[23] The three numerical parameters shown in Equation (2.30) are experimentally derived, meaning that B3LYP is strictly not *ab initio*, but does still provide good results in most cases. In this thesis, B3LYP has been used in many cases, due to the excellent balance of computational cost and accuracy.

2.2.3.3 A Note on Dispersion Correction

Mean-field electronic structure methods such as Kohn-Sham DFT (and indeed the HF approach) do not account well for London dispersion interactions, and in particular are unable to capture the long-range dependence of interaction energies between neutral

partners. This also has consequences for molecular systems, where the dispersion effects between atoms a few Angstrom from each other are poorly described in standard Kohn-Sham approaches. For all Kohn-Sham DFT calculations in this thesis, methods to correct for the lack of dispersion effects were included.

The most widely used dispersion correction methodology is Grimme's D3 model, which uses only the nuclear coordinates to accurately correct for the lack of dispersion.[28] The use of a correction such as the D3 form used here also requires damping, and in this thesis the Becke-Johnson damping function has been used as this yields a good description of medium-range interactions.[29] The derivations of these functions are not given here, but are detailed in an excellent review by the father of dispersion correction, S. Grimme.[30]

2.3 Hybrid QM/MM

The background material in this subsection is based on the following publications: 31–33

Large chemical systems can be modelled accurately and efficiently with the MM method, but a primary limitation lies is its inability to model chemical reactivity. Careful application of selected QM cluster models solve this problem in part,[34] allowing for bonds to be freely broken and formed, however, computational limitations dictate that these models are sometimes not sufficiently large to accurately model large chemical systems.

Accordingly, in 1976, Warshel and Levitt introduced the hybrid QM/MM approach, exploiting the speed of MM and the precision of QM,[35] and ultimately leading to the Nobel prize in chemistry in 2013 along with Karplus. Application of the QM/MM approach to a chemical system requires a sensibly chosen QM region which describes the part of the system where any important changes in electronic structure occur, and a MM region which describes the

remaining 'spectator' atoms that may provide medium- and long-range (de)stabilising effects to the selected QM region (see Figure 2.2).

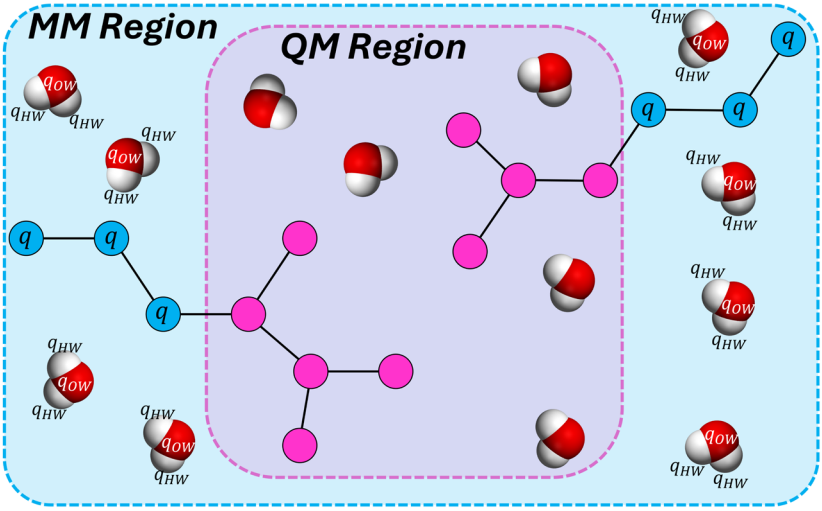

Figure 2.2: *Simplified representation of a QM/MM calculation, showing the MM region in blue, and the QM region in pink. For the MM region, point charges represented by q are shown, highlighting the specific case of electrostatic embedding.*

With this type of partitioning scheme, in the QM/MM method, the accurate and expensive QM calculation(s) is/are only performed for a small portion of the complete system, and the MM calculation(s) can sufficiently evaluate the properties of the remaining system. Consequently, the QM/MM approach provides an elegant solution to describe the large chemical systems in a rigorous and economical way.

2.3.1 QM and MM Region Coupling

A proper QM/MM calculation is not simply the sum of entirely separate QM and MM calculations on two different system regions. To provide meaningful results, the QM and MM region must be in some way coupled, which can be achieved in two ways: subtractive and additive. In principle, the choice of subtractive or

additive is not critical, and if set up correctly, should produce very similar to identical results, as shown by Ryde.[36] For the discussion of the two schemes below, the distinction will be made in terms of Hamiltonians describing the various subsystems necessitated by the coupling. The subtractive scheme is relatively simple, where three Hamiltonians are required: a QM-derived Hamiltonian for the QM region ($\hat{H}^{QM}_{QM\ region}$), a MM-derived Hamiltonian for the QM region ($\hat{H}^{MM}_{QM\ region}$), and a MM-derived Hamiltonian for the entire system($\hat{H}^{MM}_{system}$), as shown in Equation (2.31).

$$\hat{H}^{subtractive}_{QM.MM} = \left(\hat{H}^{QM}_{QM\ region} + \hat{H}^{MM}_{system}\right) - \hat{H}^{MM}_{QM\ region} \qquad \textbf{(2.31)}$$

Providing its namesake, the subtraction of $\hat{H}^{MM}_{QM\ region}$ from the sum of $\hat{H}^{QM}_{QM\ region}$and $\hat{H}^{MM}_{system}$ yields the QM/MM Hamiltonian which has no double-counted interactions. This method does require, however, that reasonable force field parameters are available for all atoms in the system – a practical limitation in many cases. On the other hand, the additive scheme negates the requirement for QM region force field parameters, as QM ($\hat{H}^{QM}_{QM\ region}$) and MM ($\hat{H}^{MM}_{MM\ region}$) Hamiltonians are only required for the respective regions. However, an additional QM-MM region coupling Hamiltonian ($\hat{H}^{QM-MM}$) must be introduced, as shown in Equation (2.32).

$$\hat{H}^{additive}_{QM.MM} = \left(\hat{H}^{QM}_{QM\ region} + \hat{H}^{MM}_{MM\ region}\right) + \hat{H}^{QM-MM} \qquad \textbf{(2.32)}$$

Despite what Equation (2.32) may imply, $\hat{H}^{QM-MM}$ is not explicitly a separate term, but instead represents any modifications that need to be made to the QM and/or MM Hamiltonians to account for the cross-boundary interactions. The exact definition of $\hat{H}^{QM-MM}$ varies somewhat depending on the implementation, but typically

includes van der Waals and electrostatic interactions. The van der Waals interactions are usually calculated quite accurately using well-parametrized MM methods, whereas several forms of charge embedding exist.

For most use cases, there are two types of embedding: mechanical and electrostatic. In the former, an additional set of point charges is included in the MM Hamiltonian to describe the QM region. These charges may be simply taken from the force field used, or alternatively can be QM-derived. The benefit of this scheme lies in its simplicity, but since the QM region is usually the reactive part of the system, one would expect that the polarisation of the QM region by the MM region is more important. Electrostatic embedding is exactly this, where the MM region point charges are included in the QM Hamiltonian as illustrated in Figure 2.2, and mathematically in Equation (2.33),[37] where N_{QM}, N_{MM}, and N_e are the number of QM atoms, MM atoms, and electrons in the QM region, respectively, Q are the MM point charges, q is the charge of a QM nucleus, and $\boldsymbol{R}$ and $\boldsymbol{r}$, are the nuclear and electron coordinates, respectively.

$$\hat{H}^{QM-MM} = \sum_{A}^{N_{MM}} \sum_{B}^{N_{QM}} \frac{Q_A q_B}{|\boldsymbol{R}_A - \boldsymbol{R}_B|} - \sum_{A}^{N_{MM}} \sum_{i}^{N_e} \frac{Q_A}{|\boldsymbol{R}_A - \boldsymbol{r}_i|} \qquad \textbf{(2.33)}$$

Broadly speaking, electrostatic embedding is the most widely used, due to the improved description of the reactive region. There is also an increasing emphasis on the use of yet a third type of charge embedding, polarised embedding,[38] whereby both the QM and MM regions are continuously polarised by one another by the inclusion of a polarisable force field to describe the MM region. However, static force fields remain dominant in QM/MM applications.

2.3.2 QM and MM Region Boundary

In hybrid QM/MM calculations in solution, it relatively easy to define a QM and MM region which involves no covalent bond breakage. However, for biochemical or oligomeric systems, it is often necessary to define the QM-MM boundary across covalent bonds. Ultimately, this results in a QM region with an unsatisfied valence, which is problematic for electronic structure calculations. There are several approaches to solving this problem, of which two will be described here: the link atom and the local self-consistent field approaches.

The link atom approach is the simpler of the two, where the unsatisfied valence in the QM region is saturated by adding an additional atom, usually a hydrogen atom, subject to some geometric criteria. In this way, the electronic structure of the QM region is not greatly modified. Of course, an unsatisfied valence remains in the MM region, but since MM methods are purely classical, the effect is less problematic. In any case, the unphysical change in local charge density is usually perturbed by setting the charge of the frontier MM atom to zero and distributing the zeroed charge among neighbouring MM atoms to not collapse the overall self-consistency of the force field. The link atom approach, while effective, can lead to problems with the subtractive QM/MM scheme if the user is not careful with their implementation (i.e., $\hat{H}^{MM}_{system}$ will not contain link atoms, but $\hat{H}^{MM}_{QM\,region}$ will).[36] Moreover, when performing geometry optimisations, link atoms may lead to distorted geometries due to the addition of artificial nuclear coordinates.

In the initial QM/MM composition by Warshel and Levitt,[35] the local self-consistent field approach is used, whereby the electron density along the frontier bond is described by a single-electron hybrid sp^2 orbital added at the QM frontier. This atomic orbital is frozen and so is not included in the SCF procedure, but instead behaves in a similar way to point charges on the MM atoms,

allowing for QM region polarisation. Beneficially, adding no new nuclear coordinates preserves the geometry optimisation procedure, but the local self-consistent field approach has been shown to be rather sensitive to the QM region size, and the local MM charge environment.[39]

Supposing that higher order or strongly polarised bonds are not broken across the QM-MM boundary, then both schemes work well in practice.[39] However, due to the relative simplicity in implementation, the link atom approach with hydrogen atoms is more widely used.

2.4 Molecular Dynamics

The background material in this subsection is based on the following textbooks: A. Leach, Molecular Modelling: Principles and Applications, 2nd Edition, ***2001****; D. Frenkel & B. Smit, Understanding Molecular Simulation, 2nd Edition,* ***2002****; A. Hinchliffe, Molecular Modelling for Beginners, 2nd Edition,* ***2003****.*

For the time-dependent study of chemical systems, the most prevalent technique used is the molecular dynamics (MD) method. Formally proposed in 1957 by Alder and Wainwright,[40] the concept behind MD is very intuitive: using some definition of the potential energy of the system, solve Newton's equations of motion for the particles to obtain velocities, and repeat. Calculation of velocities allows for the position of every particle to be updated subject to a selected time-step, thereby enabling time-evolution.

2.4.1 Trajectory Propagation

Prediction of new nuclear coordinates, $\boldsymbol{R}_{i+1}$, from a given set of nuclear coordinates, $\boldsymbol{R}_i$, requires prediction of velocities of all nuclear coordinates, $\frac{d\boldsymbol{R}_i}{dt}$. To obtain these velocities, a set of forces acting on all nuclear coordinates, $\boldsymbol{F}_i^j$, as well as an appropriately small time-step, Δt, are required. A remaining requirement is that the initial velocities, $\frac{d\boldsymbol{R}_0}{dt}$, must be selected, which are sometimes

simply taken as random, or more often sampled from a Boltzmann distribution. The forces, $\boldsymbol{F}_i^j$, can be computed as the negative of the gradient of the potential energy function, $V(\boldsymbol{R})$, with respect to the nuclear coordinates, as shown in Equation (2.34).

$$\boldsymbol{F}_i^j = -\frac{\partial V(\boldsymbol{R}_i^j)}{\partial \boldsymbol{R}^j} \tag{2.34}$$

Substitution of these forces into Newton's second law yields Equation (2.35), which enables the calculation of the nuclear coordinates and velocities *via* first-order Taylor expansions. Taylor expansions for the positions and velocities are shown in Equations (2.36) and (2.37), respectively.

$$\boldsymbol{F}_i^j = -\frac{\partial V(\boldsymbol{R}_i^j)}{\partial \boldsymbol{R}^j} = m^j . \frac{d^2 \boldsymbol{R}_i^j}{dt^2} \tag{2.35}$$

$$\boldsymbol{R}_{i+1} = \boldsymbol{R}_i + \frac{d\boldsymbol{R}_i}{dt} . \Delta t \tag{2.36}$$

$$\frac{d\boldsymbol{R}_{i+1}}{dt} = \frac{d\boldsymbol{R}_i}{dt} + \frac{d^2 \boldsymbol{R}_i^j}{dt^2} . \Delta t \tag{2.37}$$

If the velocities, $\frac{d\boldsymbol{R}_i}{dt}$, are known, and the time-step, Δt, has been defined, the new set of nuclear coordinates, $\boldsymbol{R}_{i+1}$, can be calculated rather easily using Equation (2.36). Calculation of the velocities, $\frac{d\boldsymbol{R}_{i+1}}{dt}$, on the new set of nuclear coordinates, $\boldsymbol{R}_{i+1}$, necessitates substitution of Equation (2.35) into Equation (2.37), as shown in Equation (2.38).

$$\frac{d\boldsymbol{R}_{i+1}}{dt} = \frac{d\boldsymbol{R}_i}{dt} - \left(\frac{\partial V(\boldsymbol{R}_i^j)}{\partial \boldsymbol{R}^j} . \frac{1}{m^j} . \Delta t \right) \tag{2.38}$$

Now that the velocities, $\frac{d\boldsymbol{R}_i}{dt}$, are known, the gradient of the potential energy function, $\frac{\partial V\left(R_i^j\right)}{\partial R^j}$, has been computed and the time-step, Δt, and mass factor, m^j, have been selected, the velocities can be obtained. Altogether, the procedure outlined above forms a step of trajectory propagation, albeit a rather naïve trajectory propagation method. In this case, a first-order Taylor expansion (the Euler algorithm) was used in an illustrative capacity, however due to the harsh truncation, this kind of trajectory propagation method can quickly lead to numerical drift after extended periods of propagation. In practice, more sophisticated methods of trajectory propagation are often used.

Two of the most used are the Verlet algorithm and the improved velocity Verlet algorithm. The fundamental idea of the Verlet algorithm is to construct two fourth-order Taylor expansions for the nuclear coordinates, one taking one Δt forward (Equation (2.39)) and one taking one Δt backward (Equation (2.40)).

$$\boldsymbol{R}_{i+1} = \boldsymbol{R}_i + \frac{d\boldsymbol{R}_i}{dt}.\Delta t + \frac{1}{2!}.\frac{d^2\boldsymbol{R}_i}{dt^2}.\Delta t^2 + \frac{1}{3!}.\frac{d^3\boldsymbol{R}_i}{dt^3}.\Delta t^3 + \frac{1}{4!}.\frac{d^4\boldsymbol{R}_i}{dt^4}.\Delta t^4 \tag{2.39}$$

$$\boldsymbol{R}_{i-1} = \boldsymbol{R}_i - \frac{d\boldsymbol{R}_i}{dt}.\Delta t + \frac{1}{2!}.\frac{d^2\boldsymbol{R}_i}{dt^2}.\Delta t^2 - \frac{1}{3!}.\frac{d^3\boldsymbol{R}_i}{dt^3}.\Delta t^3 + \frac{1}{4!}.\frac{d^4\boldsymbol{R}_i}{dt^4}.\Delta t^4 \tag{2.40}$$

Addition of Equations (2.39) and (2.40) with some rearranging and substitution of Equation (2.35) yields the Verlet algorithm for obtaining the nuclear coordinates, as shown in Equation (2.41). The fourth-order term is neglected as the use of a small Δt means that this term approaches zero.

$$\boldsymbol{R}_{i+1} = 2\boldsymbol{R}_i - \boldsymbol{R}_{i-1} - \left(\frac{\partial V\left(\boldsymbol{R}_i^j\right)}{\partial \boldsymbol{R}^j} . \frac{1}{m^j} . \Delta t^2\right) + \frac{1}{12}\frac{d^4\boldsymbol{R}_i}{dt^4} . \Delta t^4 \qquad \textbf{(2.41)}$$

Similarly, the velocities may be obtained by subtraction of Equations (2.39) and (2.40), yielding Equation (2.42).

$$\frac{d\boldsymbol{R}_i}{dt} = \frac{\boldsymbol{R}_{i+1} - \boldsymbol{R}_{i-1}}{2\Delta t} \qquad \textbf{(2.42)}$$

Equation (2.42) exposes the shortcomings of the Verlet algorithm, as the velocities are computed one Δt behind the nuclear coordinates, which requires that nuclear coordinates are stored for three consecutive time-steps: $\boldsymbol{R}_{i-1}$, $\boldsymbol{R}_i$, and $\boldsymbol{R}_{i+1}$. For simple cases with relatively few atoms where a suitably small Δt is selected, this procedure can be very effective. However, storing three consecutive time-steps can quickly consume memory in larger systems.

A further development of the Verlet algorithm, the velocity Verlet algorithm, has been developed (derivation not shown here, but it is similarly constructed from Taylor expansions). In this algorithm, the positions and velocities are obtained using Equations (2.43) and (2.44), respectively (with higher-order terms truncated).

$$\boldsymbol{R}_{i+1} = \boldsymbol{R}_i + \frac{d\boldsymbol{R}_i}{dt} . \Delta t - \left(\frac{\partial V\left(\boldsymbol{R}_i^j\right)}{\partial \boldsymbol{R}^j} . \frac{1}{2m^j} . \Delta t^2\right) \qquad \textbf{(2.43)}$$

$$\frac{d\boldsymbol{R}_{i+1}}{dt} = \frac{d\boldsymbol{R}_i}{dt} - \frac{1}{2m^j} . \left(\frac{\partial V\left(\boldsymbol{R}_{i+1}^j\right)}{\partial \boldsymbol{R}^j} + \frac{\partial V\left(\boldsymbol{R}_i^j\right)}{\partial \boldsymbol{R}^j}\right) . \Delta t \qquad \textbf{(2.44)}$$

The velocity Verlet algorithm also has the benefit of computing velocities and nuclear coordinates at the same time-step, meaning that various properties can be calculated in real-time.

The velocity Verlet algorithm has accordingly become one of the most widely used trajectory propagation methods, giving a good balance of speed and reliability. Another extensively used method is the leap-frog algorithm, but the essence of this algorithm is similar to the Verlet and velocity Verlet algorithms, so it has not been discussed here.

2.4.2 The Molecular Dynamics Method

Irrespective of the trajectory propagation method used, the steps described above form the mathematical basis of the molecular dynamics method, whereby a trajectory is propagated with a selected time-step for as long as the desired simulation time, as illustrated in Figure 2.3.

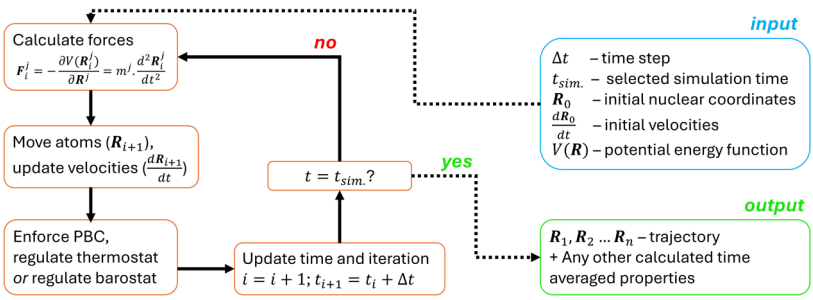

Figure 2.3: *Schematic representation of the molecular dynamics method, in a simplified form. The blue box shows the minimum required inputs; the green box shows the typical outputs (which may vary depending on the application); the orange boxes show the main steps of the molecular dynamics method.*

The choice of simulation time is not necessarily trivial, as the question of how long a simulation should be has no definitive answer. In practice, several measurements (RMSD, RMSF, radius of gyration, among others) are continuously taken along the course of a trajectory to determine if the simulation is 'long enough', but from a statistical standpoint, a longer simulation is always desirable. The choice of a suitable time-step has a slightly more analytical solution, where a useful rule-of-thumb is that it should be one to two orders of magnitude smaller than the fastest motion

in the system. For chemical systems, this is often the carbon-hydrogen bond vibrational frequency at room temperature, which is $\approx 3000\ cm^{-1} \approx 0.01\ ps$. Therefore, a recommended time-step would be in the range of 0.1 – 1 fs, depending on the thermodynamic ensemble used.

Besides the simulation time and time-step, another critical choice is the potential energy function used to evaluate the forces acting on each of the nuclear coordinates. This function can be defined in many ways, such as by a QM calculation (*ab initio* MD), a MM calculation (classical MD), or even a QM/MM calculation (*ab initio* QM/MM MD). A given MD simulation is ultimately subject to any of the limitations of the function chosen, and it is broadly-speaking limited by the speed of the potential energy function evaluation, as the trajectory propagation steps are typically very fast. In practice, the selection of the potential energy function is largely dependent on whether one is modelling electronic structure changes (*ab initio* MD or *ab initio* QM/MM MD), or dynamic systems which have no changes in bonding (classical MD).

2.4.3 Periodic Boundary Conditions

Considering a set of particles placed in a cubic box, many of the particles will be exposed to vacuum, due to then being on the surface of the cube. Chemical systems do not exist in a vacuum, so an MD simulation on a set of N nuclear coordinates contained within a cubic box would be expected to behave unphysically.

To mitigate these finite system effects, in most MD simulations, so-called periodic boundary conditions (PBCs) are enforced (see Figure 2.3). Placing the N nuclear coordinates in a box (e.g., a cube, a hexagonal prism, rhombic dodecahedron) which can be stacked in a fully space-filling way enables this box to be surrounded in every dimension in infinitely repeating identical unit cells (see Figure 2.4).

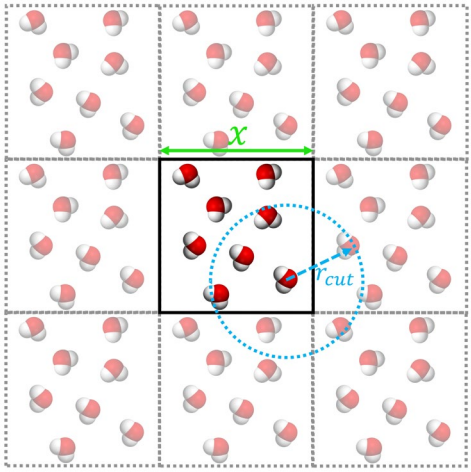

Figure 2.4: *Simplified 2-dimensional illustration of periodic boundary conditions in molecular dynamics simulations. The box (with dimensions $x \times x$) in the centre represents the primary unit cell where the simulation is performed, using an r_{cut} (shown in blue) to truncate intermolecular interactions. All periodic cells (also with dimensions $x \times x$) are shown as slightly faded and are identical to the primary unit cell.*

If a molecule exits one face of the box, then it enters the opposite face of the box, thereby preserving the N nuclear coordinates. It should be noted that the trajectory is only propagated for the primary unit cell, where the periodic cells are simply replicas of the unit cell, but intermolecular interactions are still included across the periodic boundary. These intermolecular interactions are calculated based on two criteria: an interaction is always calculated only with the closest periodic replica (the minimum image convention), and the interactions are truncated with some pre-defined cut-off distance which is great enough that an atom cannot interact with both the replica cell and the primary unit cell (i.e., $r_{cut} > 0.5x$), as illustrated in Figure 2.4. However, Coulombic interactions are important at ranges longer than r_{cut}, so are very often calculated using the Ewald summation method. In this method, the short-range interactions are calculated in real space using Coulomb's law, and long-range interactions are calculated by way of Fourier transforms.[41]

2.4.4 Thermodynamic Ensembles

The explanation of trajectory propagation and the MD method so far have not explored the effects of thermodynamics, which play a large part in a time-averaged system. The discussion above has considered a conventional MD simulation, where a system of N nuclear coordinates is contained in a defined volume, V. By their very nature, Newton's equations of motion preserve the total system energy, and so the energy, E, is a constant of motion. Therefore, these fully adiabatic simulations are part of the microcanonical statistical ensemble, often termed the NVE-ensemble.

In practice, other than for particularly small time-steps and/or systems with a large number of particles, molecular dynamics simulations in the NVE-ensemble can present issues with temperature stability. In order to stabilise the temperature, since the equipartition theorem holds in the NVE-ensemble, one could in principle scale the velocities. However, this breaks the Newtonian nature of the simulations and can lead to statistical mechanical artefacts. A more elegant solution which is widely applied is to maintain contact with a constant temperature heat bath, or a thermostat (as shown on Figure 2.3). Simulations of this type are part of the canonical statistical ensemble, or the NVT-ensemble.

Simulations in the NVT-ensemble offer less sensitivity to system size and time-step so are widely used, however, since chemically-relevant systems are typically studied at constant pressure and (relatively) invariant temperature, a barostat is often also used (as shown on Figure 2.3). In this case, the temperature is regulated with a heat bath as with the NVT-ensemble, and the pressure is regulated by scaling the volume (i.e., scaling the unit cell size in a system with PBCs). These simulations are part of the isothermal-isobaric statistical ensemble, or the NPT-ensemble.

2.5 Free energy

The background material in this subsection is based on the following textbooks: D. Frenkel & B. Smit, Understanding Molecular Simulation, 2nd Edition, ***2002****; C. J. Cramer, Essentials of Computational Chemistry, 2nd Edition,* ***2004****.*

Broadly-speaking, chemistry is concerned with understanding the macroscopic behaviour of interacting atoms. These ensembles of molecules are fundamentally governed by the laws of thermodynamics, arising from the fact that chemistry occurs above absolute zero. Consequently, calculating microscopic quantities on single molecules (by QM, MM, QM/MM, etc.) without consideration given to the laws of thermodynamics is not consistent with the goal of predicting macroscopic properties.

2.5.1 Statistical Mechanical Methods

In quantum-mechanical systems, moving from a microscopic to a macroscopic description first necessitates the calculation of a zero-point vibrational energy (ZPVE). Since Heisenberg's Uncertainty Principle dictates that the momentum of a particle cannot be zero, then there must always be some residual vibrational energy remaining. To correct this, the familiar harmonic oscillator model is applied as sum of all the molecular vibrations (shown in Equation (2.45)) which are calculated by a second derivative calculation of the electronic energy with respect to the nuclear coordinates.

$$E_{ZPVE} = \sum_{i=1}^{vibrations} \frac{1}{2} h\omega_i \qquad \textbf{(2.45)}$$

Where in Equation (2.45), h is Planck's constant, and ω_i is a given molecular vibration. While in this context the harmonic oscillator model is inherently an approximation, it works well for computing the ZPVE correction, assuming that the molecular vibrations are calculated on a true stationary point. To complete the

thermodynamic description of a microscopic system, a statistical mechanical approach under the assumption of the ideal gas law is often used, enabling the prediction of macroscopic behaviour based on probability theory. Such thermodynamic properties are calculated using the fundamental thermodynamic potentials shown in Equations (2.46) and (2.47) for the enthalpy, H, and Gibbs free energy, G, respectively (noting that the second term in Equation (2.46) is expressed as $k_B T$ rather than pV due to the assumption of the ideal gas law).

$$H = U - k_B T \tag{2.46}$$

$$G = H - TS \tag{2.47}$$

Calculation of H and G requires an expression for the internal energy, U, (or equivalently E) and entropy, S. To achieve this, the molecular partition function, Q, is used, which can be expressed as a product of its component parts: translational (q_{trans}), rotational (q_{rot}), vibrational (q_{vib}), and electronic (q_{elec}) DoF. The result of this separation is that U and S can be evaluated as a sum of contributions from each of these types of DoF, as shown in Equations (2.48) and (2.49) for U and S, respectively.

$$U = E_{trans} + E_{rot} + E_{vib} + E_{elec} \tag{2.48}$$

$$S = S_{trans} + S_{rot} + S_{vib} + S_{elec} \tag{2.49}$$

The detailed mathematics behind the form of the terms given in Equations (2.48) and (2.49) is not given here, but as for the ZPVE correction, the harmonic oscillator model is again applied here for the vibrational correction, so the importance of computing accurate molecular vibrations is emphasised.

2.5.2 Sampling Methods

The discussion above has been focused on single point calculations, because the application of statistical mechanics in the way described above on an ensemble of structures produced by molecular simulation can be problematic for several reasons. Namely, molecular simulation often invokes the use of a force field and since the energy produced by a force field calculation is not a zero energy, but instead a potential energy, applying the methods described above is not strictly possible. As a result, methods based on sampling the configurational space generated by molecular simulation are a useful alternative for evaluating macroscopic properties.

Taking the example of obtaining a minimum energy reaction pathway (MEP), which is the least energy path connecting two minima on a molecular potential energy surface. There may be one or more saddle points (transition states), and indeed one or more minima between the two minima (intermediates). The intricate details of how MEPs may be obtained as well as the common difficulties associated with them will be explored in detail in Chapter 5, but in general, one may obtain such a MEP as either a free energy surface by using dynamic techniques, or as a potential energy surface using static techniques.

While not explored in any detail in this thesis, dynamic techniques such as umbrella sampling,[42–44] free energy perturbation methods,[45–48] and thermodynamic integration[49–53] do provide an attractively systematic approach to describing reactivity. These methods are, however, quite demanding in a QM context (which is often necessary when studying reactivity), since they require extensive sampling of the potential energy surface at the QM level, often limiting the choice of accurate QM methods due to increasing computational cost.[49]

Alternatively, static techniques supplemented by sampling are interesting in that they do not require QM simulation. To obtain a free energy barrier, starting from an ensemble of structures, a number of MEPs equal to the number of structures is calculated. In principle, each MEP is approximated as a set of discrete structures leading from reactants to products, and the activation free energy of the reaction is estimated by averaging the ensemble of potential energy barriers obtained.[54] There exist several methods which can be used to calculate a MEP for a given chemical reaction from a reactant and/or product structure, such as adiabatic mapping,[55] the nudged elastic band (NEB) method,[56–58] or the growing string method (GSM).[59–64] The growing string method will be explored in much greater detail in Chapter 5.

It is worth noting that the estimate of the free energy barrier obtained by averaging potential energy barriers is not exact, and, as will be discussed in detail in Chapter 5, the extent of sampling required to obtain meaningful estimation of the barrier is a non-trivial question, somewhat restricting the relative computational efficiency compared to dynamic approaches.

2.6 Delocalised Internal Coordinates

The background material in this subsection is based on the following publications: 65–68

When performing calculations or simulations on nuclear coordinates, Cartesian coordinates are the obvious choice since they can be easily defined for all systems, are often the default format for many structural databases (e.g., the PDB), and provide a certain convenience due to ease of intercommunication between programs. Despite these practical benefits, Cartesian coordinates are a strongly coupled coordinate system. Generally, less coordinate coupling is desirable, as any coordinate change should have minimal impact on any other coordinates. For cases like Cartesian coordinates with a high degree of coupling,

geometry optimisation convergence can be slow and in particularly bad cases, fail.[69] Another critical factor for studying chemical reactivity is that constraints on for example interatomic distances are often necessary, and imposing such constraints in Cartesian space, while workable, can be problematic.[70,71]

Unlike Cartesian coordinates, internal coordinates are based on the atomic connectivity in a given system, and thus offer a more chemically intuitive description of a geometry. Besides the improved speed offered by these coordinates due to having fewer DoF, more importantly, they are less strongly coupled than Cartesian coordinates and so typically offer more robust geometry optimisation convergence. A number of internal coordinate systems exist, but the most appropriate choice for the study of chemical reactivity is delocalised internal coordinates (DLCs), since multi-component constraints can be easily imposed.[65]

2.6.1 Generation of Delocalised Internal Coordinates

Assume a system described by Cartesian coordinates, $\boldsymbol{x}$, with n atoms and hence $n_x = 3n$ Cartesian coordinates. The same system can equivalently be described by primitive internal coordinates, $\boldsymbol{q}$, of which there is a variable number n_q. The exact definition of $\boldsymbol{q}$ is not critical and can be any combination of bonds, valence angles, or dihedral angles. However, to generate a set of DLC from these coordinates, they must span at least $n_x - 6$ DoF. Small displacements in $\boldsymbol{q}$ can be converted to small displacements in $\boldsymbol{x}$ by way of the Wilson $\boldsymbol{B}$ matrix, as defined in Equation (2.50).

$$\boldsymbol{B} = \begin{pmatrix} \frac{\partial q_1}{\partial x_1} & \cdots & \frac{\partial q_1}{\partial x_{n_x}} \\ \vdots & \ddots & \vdots \\ \frac{\partial q_{n_q}}{\partial x_1} & \cdots & \frac{\partial q_{n_q}}{\partial x_{n_x}} \end{pmatrix} \qquad \textbf{(2.50)}$$

Where a given element, B_{ij}, of this rectangular $n_q \times n_x$ matrix is $\frac{\partial q_i}{\partial x_j}$, which is known and can be evaluated analytically from the definition of $\boldsymbol{q}$ and $\boldsymbol{x}$. Since $\boldsymbol{q}$ is typically generated automatically based on atomic connectivity, there is very likely some redundancy in the coordinate set. However, this is not of great concern as these redundancies can be entirely removed through a diagonalisation procedure. First, the square $n_q \times n_q$ matrix $\boldsymbol{G}$ must be defined, as shown in Equation (2.51).

$$\boldsymbol{G} = \boldsymbol{B}\boldsymbol{B}^T \tag{2.51}$$

The square, symmetric (and, by definition, singular) matrix $\boldsymbol{G}$ as defined in Equation (2.51) can be diagonalised, creating two sets of eigenvectors, $\boldsymbol{UR}$, and a set of eigenvalues $\boldsymbol{\Lambda}$ corresponding to $\boldsymbol{U}$ and $\{0\}$ corresponding to $\boldsymbol{R}$, as expressed in Equation (2.52).

$$\boldsymbol{G} = (\boldsymbol{UR}) \begin{pmatrix} \boldsymbol{\Lambda} & 0 \\ 0 & 0 \end{pmatrix} (\boldsymbol{UR})^{-1} \tag{2.52}$$

The matrix of eigenvectors $\boldsymbol{U}$ is rectangular with $n_x - 6$ columns and n_q rows, and the matrix $\boldsymbol{R}$, while also rectangular, has $n_q - (n_x - 6)$ columns and again n_q rows. The eigenvectors $\boldsymbol{R}$ represent the redundancy within $\boldsymbol{q}$ and are discarded, while the eigenvectors $\boldsymbol{U}$ contain all the weight of the non-redundant $\boldsymbol{q}$ and thus make the basis for the formation of the DLC. A given element U_{ij} is therefore the weight of the jth primitive internal coordinate in the ith DLC. Lastly, the DLC, $\boldsymbol{s}$, of which there are $n_x - 6$, are formed by linear combinations of $\boldsymbol{q}$ with eigenvectors $\boldsymbol{U}$, as shown in Equation (2.53).

$$\boldsymbol{s} = \boldsymbol{U}^T \boldsymbol{q} \tag{2.53}$$

2.6.2 Geometry Optimisation in Delocalised Internal Coordinates

The $n_x - 6$ DoF required to describe a given system without redundancy have thus been isolated in the form of $\boldsymbol{s}$, and it can be taken as the working coordinate set with usual optimisation techniques performed in this space. However, for appropriate optimisation steps to be taken, the Cartesian gradient, $\boldsymbol{g}_{cart}$, and (optionally) Hessian, $\boldsymbol{H}_{cart}$, must be updated to DLC subspace. Firstly, as shown in Equation (2.54), the original, primitive Wilson $\boldsymbol{B}$ matrix (hereafter $\boldsymbol{B}^{prim}$) must be transformed to the basis of the active coordinate set.

$$\boldsymbol{B} = \boldsymbol{U}^T \boldsymbol{B}^{prim} \tag{2.54}$$

The DLC-subspace rectangular matrix $\boldsymbol{B}$ has dimensions $n_x \times (n_x - 6)$. To facilitate the transformation of $\boldsymbol{g}_{cart}$ and/or $\boldsymbol{H}_{cart}$, $\boldsymbol{B}^T$ must be inverted. To form such an inverse matrix, a pseudo-inverse methodology is necessitated, as $\boldsymbol{B}^T$ is not a square matrix, and thus usual inversion techniques cannot be used. To facilitate this inversion procedure, reformation of $\boldsymbol{G}$ in DLC-subspace is required. $\boldsymbol{G}$ is constructed in the same fashion using Equation (2.51), but since $\boldsymbol{B}$ is now in DLC-subspace, is it not of exactly the same form as described previously. $\boldsymbol{G}$ is still a square (and so can be inverted using usual techniques), symmetric, and singular matrix, but now has dimensions $(n_x - 6) \times (n_x - 6)$. Now, using the left Moore-Penrose pseudo-inverse, $(\boldsymbol{B}^T)^{-1}$, can be constructed as shown in Equation (2.55).

$$(\boldsymbol{B}^T)^{-1} = \boldsymbol{G}^{-1}\boldsymbol{B} \tag{2.55}$$

$(\boldsymbol{B}^T)^{-1}$ may now be used to update $\boldsymbol{g}_{cart}$ and/or $\boldsymbol{H}_{cart}$ into DLC subspace through Equations (2.56) and (2.57), forming the DLC gradient, $\boldsymbol{g}$, and Hessian, $\boldsymbol{H}$, respectively.

$$\boldsymbol{g} = (\boldsymbol{B}^T)^{-1}\boldsymbol{g}_{cart} \tag{2.56}$$

$$\boldsymbol{H} = (\boldsymbol{B}^T)^{-1}\left(\boldsymbol{H}_{cart} - \boldsymbol{g}\frac{d\boldsymbol{B}}{d\boldsymbol{x}}\right)[(\boldsymbol{B}^T)^{-1}]^T \tag{2.57}$$

Equation (2.57) is seldom used as an accurate $\boldsymbol{H}_{cart}$ is often challenging to obtain. Moreover, the method used to find $\frac{d\boldsymbol{B}}{d\boldsymbol{x}}$ can result in changes to the Hessian eigenvalue structure, and thus lead to an inconsistency in the definition of the DLC. In practice, an approximate initial primitive internal coordinate Hessian, $\boldsymbol{H}_{prim}$, is used and updated to DLC subspace.

As previously mentioned, the usual input for computational chemistry software is Cartesian coordinates, so typically the final step in every optimisation cycle in DLC subspace is to convert the DLCs back to Cartesian coordinates. While in principle one could use the matrix $\boldsymbol{B}$ to convert between $\boldsymbol{s}$ and $\boldsymbol{x}$ in a similar fashion to converting between $\boldsymbol{q}$ and $\boldsymbol{x}$ with $\boldsymbol{B}^{prim}$, since the relationship $\boldsymbol{s}$ and $\boldsymbol{x}$ is curvilinear, using a linear transformation matrix such as $\boldsymbol{B}$ would give poor transformations, perhaps other than for very small changes in $\boldsymbol{s}$. In practice, a known set of DLC ($\boldsymbol{s}$) is back transformed to Cartesian coordinates, using an iterative procedure shown in Equation (2.58).

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + [(\boldsymbol{B}^T)^{-1}]^T(\boldsymbol{s} - \boldsymbol{s}_k) \tag{2.58}$$

This transformation typically converges very quickly, with $(\boldsymbol{s} - \boldsymbol{s}_k) < 10^{-10}$ and hence $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ achieved in only a few cycles.

$\boldsymbol{x}_{k+1}$ can then be taken as the starting Cartesian coordinates for the next cycle of optimisation.

2.6.3 Imposing Delocalised Internal Coordinate Constraints

A significant benefit of DLCs is the ease of imposing standard geometrical constraints. Take the illustrative example of fixing one primitive internal coordinate, which may be a bond, valence angle or dihedral angle. First generate a unit vector, $\boldsymbol{C}_{prim}$, of length n_q with unit component corresponding to the constrained primitive internal coordinate. As this vector is currently in primitive internal coordinate subspace, in order for constraints to be satisfied in DLC subspace, it must be projected onto $\boldsymbol{U}$ to form $\boldsymbol{C}$, as shown in Equation (2.59).

$$\boldsymbol{C} = \sum_{k=1}^{n_x-6} \langle \boldsymbol{C}_{prim} | \boldsymbol{U}_k \rangle \boldsymbol{U}_k \qquad \textbf{(2.59)}$$

To ensure unit length, $\boldsymbol{C}$ is first normalised, and it is stacked with the active DLC coordinate transformation set, $\boldsymbol{U}$, forming a $(n_x - 6) + 1$ dimensional vector space, $\boldsymbol{V}$, as shown in Equation (2.60).

$$\boldsymbol{V} = \{\boldsymbol{C}, \boldsymbol{U}_k \qquad k = 1 \ldots (n_x - 6)\} \qquad \textbf{(2.60)}$$

Performing a Gram-Schmidt orthogonalisation procedure on $\boldsymbol{V}$, where the first vector taken is $\boldsymbol{C}$, yields the set of vectors $\boldsymbol{V}'$, where a given vector in $\boldsymbol{V}'$ is calculated using Equation (2.61).

$$\boldsymbol{V}'_k = \boldsymbol{V}_k - \sum_{l=1}^{k-1} \langle \boldsymbol{V}_k | \boldsymbol{V}'_l \rangle \boldsymbol{V}'_l \qquad \textbf{(2.61)}$$

Due to the linear dependence on all other vectors in $\boldsymbol{V}'$, the last vector taken will drop to zero, and leave a fully orthonormal set of active vectors and one constraint vector, $\boldsymbol{C}$. In this way, the fixed primitive has been isolated from the active vector set, and optimisation on the $(n_x - 6) - 1$ orthonormal active vectors may proceed as normal, leaving $\boldsymbol{C}$ fixed. Combined primitive internal coordinate constraints may also be just as easily imposed on DLC subspace, where $\boldsymbol{C}_{prim}$ contains contributions from multiple primitive internal coordinates, with the overall combined coordinate chosen to remain constant. In fact, with this procedure, linear combinations of any and all primitive internal coordinates can be imposed easily.

2.7 References

1 V. Poltev, in *Handbook of Computational Chemistry*, ed. J. Leszczynski, Springer Netherlands, Dordrecht, 2016, pp. 1–48.
2 K. Vanommeslaeghe, O. Guvench and A. D. MacKerell, *Curr Pharm Des*, 2014, **20**, 3281–3292.
3 C. C. J. Roothaan, *Rev. Mod. Phys.*, 1951, **23**, 69–89.
4 D. P. Tew, W. Klopper and T. Helgaker, *J Comput Chem*, 2007, **28**, 1307–1320.
5 C. David Sherrill and H. F. Schaefer, in *Advances in Quantum Chemistry*, eds. P.-O. Löwdin, J. R. Sabin, M. C. Zerner and E. Brändas, Academic Press, 1999, vol. 34, pp. 143–269.
6 B. O. Roos, in *Lecture Notes in Quantum Chemistry: European Summer School in Quantum Chemistry*, ed. B. O. Roos, Springer, Berlin, Heidelberg, 1992, pp. 177–254.
7 J. Čížek, *The Journal of Chemical Physics*, 1966, **45**, 4256–4266.
8 J. D. Watts, J. Gauss and R. J. Bartlett, *The Journal of Chemical Physics*, 1993, **98**, 8718–8733.
9 Z. Benedek, P. Timár, T. Szilvási and G. Barcza, *J Comput Chem*, 2022, **43**, 2103–2120.
10 M. Feldt, Q. M. Phung, K. Pierloot, R. A. Mata and J. N. Harvey, *J. Chem. Theory Comput.*, 2019, **15**, 922–937.
11 P. Pulay, *Chemical Physics Letters*, 1983, **100**, 151–154.
12 S. F. Boys, *Rev. Mod. Phys.*, 1960, **32**, 296–299.
13 J. Pipek and P. G. Mezey, *Journal of Chemical Physics*, 1989, **90**, 4916–4926.
14 F. Neese, F. Wennmohs and A. Hansen, *The Journal of Chemical Physics*, 2009, **130**, 114108.
15 F. Neese, A. Hansen and D. G. Liakos, *J Chem Phys*, 2009, **131**, 064103.
16 H.-J. Werner and K. Pflüger, *Annual Reports in Computational Chemistry*, 2006, **2**, 53–80.
17 C. Riplinger, B. Sandhoefer, A. Hansen and F. Neese, *The Journal of Chemical Physics*, 2013, **139**, 134101.
18 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B871.
19 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133–A1138.
20 A. D. Becke, *The Journal of Chemical Physics*, 2014, **140**, 18A301.
21 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
22 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
23 C. Lee, W. Yang and R. G. Parr, *Phys Rev B Condens Matter*, 1988, **37**, 785–789.
24 N. C. HANDY and A. J. COHEN, *Molecular Physics*, 2001, **99**, 403–412.
25 J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401.
26 A. D. Becke, *The Journal of Chemical Physics*, 1993, **98**, 1372–1377.
27 A. D. Becke, *The Journal of Chemical Physics*, 1993, **98**, 5648–5652.
28 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *The Journal of Chemical Physics*, 2010, **132**, 154104.
29 S. Grimme, S. Ehrlich and L. Goerigk, *Journal of Computational Chemistry*, 2011, **32**, 1456–1465.

30 S. Grimme, A. Hansen, J. G. Brandenburg and C. Bannwarth, *Chem. Rev.*, 2016, **116**, 5105–5154.
31 H. Lin, *QM/MM Methods*, American Chemical Society, 2023.
32 G. Groenhof, in *Biomolecular Simulations*, eds. L. Monticelli and E. Salonen, Humana Press, Totowa, NJ, 2013, vol. 924, pp. 43–66.
33 H. M. Senn and W. Thiel, *Angew. Chem. Int. Ed.*, 2009, **48**, 1198–1229.
34 M. R. A. Blomberg, T. Borowski, F. Himo, R.-Z. Liao and P. E. M. Siegbahn, *Chem. Rev.*, 2014, **114**, 3601–3658.
35 A. Warshel and M. Levitt, *Journal of Molecular Biology*, 1976, **103**, 227–249.
36 L. Cao and U. Ryde, *Front. Chem.*, 2018, **6**, 89.
37 A. Laio, J. VandeVondele and U. Rothlisberger, *The Journal of Chemical Physics*, 2002, **116**, 6941–6947.
38 A. Kumar, A. Simmonett and A. D. MacKerell, *Biophysical Journal*, 2021, **120**, 175a–176a.
39 N. Reuter, A. Dejaegere, B. Maigret and M. Karplus, *J. Phys. Chem. A*, 2000, **104**, 1720–1735.
40 B. J. Alder and T. E. Wainwright, *The Journal of Chemical Physics*, 1957, **27**, 1208–1209.
41 B. A. Wells and A. L. Chaffee, *J. Chem. Theory Comput.*, 2015, **11**, 3684–3695.
42 J. Kästner, *WIREs Computational Molecular Science*, 2011, **1**, 932–942.
43 W. You, Z. Tang and C. A. Chang, *J. Chem. Theory Comput.*, 2019, **15**, 2433–2443.
44 S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *Journal of Computational Chemistry*, 1992, **13**, 1011–1021.
45 H. Hu and W. Yang, *Annu. Rev. Phys. Chem.*, 2008, **59**, 573–601.
46 H. Hu, Z. Lu, J. M. Parks, S. K. Burger and W. Yang, *The Journal of Chemical Physics*, 2008, **128**, 034105.
47 H. Hu, Z. Lu and W. Yang, *J. Chem. Theory Comput.*, 2007, **3**, 390–406.
48 Y. Zhang, H. Liu and W. Yang, *The Journal of Chemical Physics*, 2000, **112**, 3483–3492.
49 J. Kästner, H. M. Senn, S. Thiel, N. Otte and W. Thiel, *J. Chem. Theory Comput.*, 2006, **2**, 452–461.
50 J. G. Kirkwood, *The Journal of Chemical Physics*, 2004, **3**, 300–313.
51 M. Sprik and G. Ciccotti, *The Journal of Chemical Physics*, 1998, **109**, 7737–7744.
52 J. Schlitter and M. Klähn, *The Journal of Chemical Physics*, 2003, **118**, 2057–2060.
53 W. K. den Otter and W. J. Briels, *The Journal of Chemical Physics*, 1998, **109**, 4139–4146.
54 A. M. Cooper and J. Kästner, *ChemPhysChem*, 2014, **15**, 3264–3269.
55 R. Lonsdale, J. N. Harvey and A. J. Mulholland, *Chem. Soc. Rev.*, 2012, **41**, 3025.
56 H. Jónsson, G. Mills and K. W. Jacobsen, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, WORLD SCIENTIFIC, LERICI, Villa Marigola, 1998, pp. 385–404.
57 G. Henkelman and H. Jónsson, *The Journal of Chemical Physics*, 2000, **113**, 9978–9985.

58 S. A. Trygubenko and D. J. Wales, *The Journal of Chemical Physics*, 2004, **120**, 2082–2094.
59 B. Peters, A. Heyden, A. T. Bell and A. Chakraborty, *The Journal of Chemical Physics*, 2004, **120**, 7877–7886.
60 P. M. Zimmerman, *J. Comput. Chem.*, 2015, **36**, 601–611.
61 P. Zimmerman, *J. Chem. Theory Comput.*, 2013, **9**, 3043–3050.
62 P. M. Zimmerman, *The Journal of Chemical Physics*, 2013, **138**, 184102.
63 M. Jafari and P. M. Zimmerman, *J. Comput. Chem.*, 2017, **38**, 645–658.
64 A. Behn, P. M. Zimmerman, A. T. Bell and M. Head-Gordon, *J. Chem. Theory Comput.*, 2011, **7**, 4019–4025.
65 J. Baker, A. Kessi and B. Delley, *Journal of Chemical Physics*, 1996, **105**, 192–212.
66 J. Baker, *J. Comput. Chem.*, 1997, **18**, 1079–1095.
67 J. Baker, A. Kessi and B. Delley, *The Journal of Chemical Physics*, 1996, **105**, 192–212.
68 J. Baker, D. Kinghorn and P. Pulay, *The Journal of Chemical Physics*, 1999, **110**, 4986–4991.
69 H. B. Schlegel, *Int. J. Quantum Chem.*, 1992, **44**, 243–252.
70 J. Baker, *Journal of Computational Chemistry*, 1992, **13**, 240–253.
71 J. Baker and D. Bergeron, *Journal of Computational Chemistry1*, 1993, **14**, 1339–1346.

3 Human Biochemical NO-signalling: Soluble Guanylate Cyclase[V]

As described in Chapter 1, the NO-sGC-cGMP pathway is mediated by a conformational change induced by NO binding to the heme group of the β-subunit H-NOX domain, but the details of how NO binds to the heme group and what conformational changes occur remain elusive. In this chapter, the energetics of NO binding to the heme group is addressed using QM and hybrid QM/MM calculations. The diffusion of NO into sGC and the conformational changes induced by NO binding to the heme group is analysed by way of MD simulations. The findings from all calculations and simulations are then considered in a biochemical context to propose the route to biochemical NO-signalling in humans.

3.1 Introduction

In the resting state of the sGC, the central ferrous iron is five-coordinate and is in the high-spin state (Figure 3.1(B), species **A**). It is axially coordinated by a histidine sidechain on the proximal side of the heme group.[1] Originally, the NO-sGC activation process

---

[V] This chapter is based on the following publication: A. M. Rozza, M. Papp, N. R. McFarlane, J. N. Harvey, J. Oláh, *Chem. Eur. J.*, **2022**, 28, e202200930. For the purposes of this thesis, sections of the publication have been slightly modified. Specifically, sections of the introduction have been removed as they have been instead included in the Chapter 1.

The calculations and simulations were performed by A. M. Rozza, M. Papp, N. R. McFarlane, and J. Oláh. Results were interpreted by all co-authors of the paper. The first draft of the paper was mostly written by A. M. Rozza and N. R. McFarlane, where A. M. Rozza wrote Section 3.3.2, N. R. McFarlane wrote Sections 3.3.1 & 3.3.3, and all authors contributed to the first draft of Sections 3.1, 3.2, & 3.4. All authors made contributions to editing the final version of the manuscript. In the text below, those parts of the work not performed by N. R. McFarlane will be explicitly identified, but these are included in this thesis as they form an integral part of the work.

was thought to result from a single NO molecule binding to the distal side of the heme iron, leading to the formation of a six-coordinate low-spin heme group (Figure 3.1(B), species **B**) followed by rupture of the Fe-His bond (Figure 3.1(B), species **D***). Based on subsequent experimental studies it has, however, been proposed that a single NO can only partially activate sGC and that an additional NO is required for the full NO activation mechanism.[2–4]

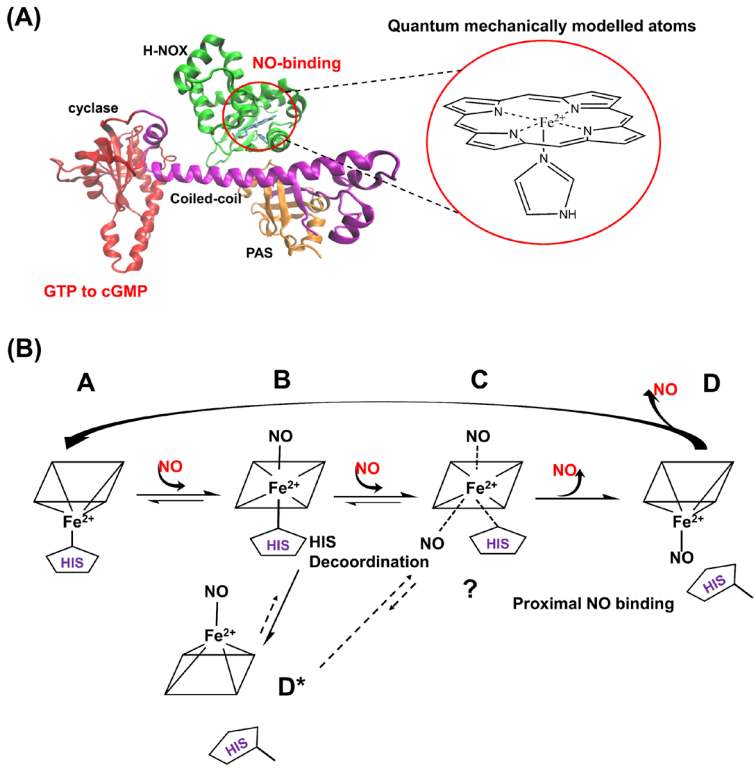

Figure 3.1: (A) *Quaternary structure of the β-subunit of sGC and the overall signalling process. Domain names are given in black and chemical reactions are shown in red. The model of the enzyme used in QM calculations and as the QM region of QM/MM calculations is circled in red.* ***(B)*** *Proposed mechanism of sGC under exposure to NO.* ***A, B, C, D*** *and* ***D**** *are the resulting species of the proposed activation pathways. Apolar hydrogens and side chains of the porphyrin ring have been omitted for clarity. The axial histidine ring is represented by an imidazole ring. Figure (A) was based on pdb-id:6JT0[5] and generated using VMD.[6]*

Although experimental studies imply that sGC undergoes multi-step NO binding, still, the exact binding site of the NO molecule(s) remains unclear. Marletta et al. found that in the presence of excess NO, S-nitrosylation of Cys122 of the β H-NOX leads to full activation of the enzyme.[7] In another study, however, S-nitrosylation of the same residue by a nitrosylating agent resulted in desensitization of sGC and loss of sensitivity to NO.[8] Such a thiol-NO reaction does not take place anaerobically, and it is too slow under normal oxidative conditions in the human body to be a factor in the observed fast NO binding.[9,10] However, it may occur under physiological conditions of oxidative stress.[7]

Further experimental investigations indicated that the sGC heme iron itself could also be the locus of binding of the second equivalent of NO. This requires the preceding or concurrent fission of the Fe-His bond. The binding of NO as a proximal ligand was suggested to strongly depend on NO concentration and as a final product **D**, involving a five-coordinate heme group with proximal NO was hypothesized to be the predominant form of the maximally activated sGC.

Conversion of **B** to **D** implies the formation of a very transient intermediate, the "bis-NO-sGC quaternary complex" (Figure 3.1(B), species **C**), which is extremely difficult to detect leaving the intricate details of the reaction mechanism elusive. Such a multi-step process of NO binding was found to occur in the highly NO-selective cytochrome c' – a microbial protein from *Alcaligenes xylosoxidans* – using stopped-flow Fourier transform infrared spectroscopy.[11] This protein is smaller than sGC and can form a stable five coordinate complex with NO bound to the proximal face of the heme (pdb-id:1E85). A sub-family of cytochrome P450s called TxtE has been studied to investigate the presence of Fe(II)-nitrosyl intermediate as a part of its catalytic cycle of nitrating cyclic substrates such as, L-tryptophan.[12] Experimental results generated from stopped-flow spectroscopic analysis showed that

the rate of NO binding to Fe(II)-thiolate of the TxtE protein is 2.5 slower than the binding of O_2. In addition, DFT calculations supported this finding, where O_2 binds first to form a Fe(III)-superoxo complex and NO later reacts with this species to give the active intermediate Fe(III)-peroxynitrite. In turn, this splits homolytically to yield Fe(IV)-oxo heme and a free NO_2 radical *via* a small free energy of activation. Thus, NO doesn't bind to the TxtE protein in the same way that it does in H-NOX. A similar family of heme-thiolate proteins are nitric oxide synthases. These also detoxify the excess amount of peroxynitrite anion; however, an Fe(III)-NO complex is the result of peroxynitrite activation instead of the oxoferryl complex.[13]

Until recently,[VI,14] quality crystal structures of the H-NOX domains of human sGC were not available, so bacterial homologues of the H-NOX domains of sGC have been thoroughly researched to gain valuable information on the characteristics of gas binding.[15] Bacterial H-NOX sensors are grouped into NO- and O_2-sensors based on their affinities to gaseous molecules. Using stopped-flow UV-Vis and MCD spectroscopy, the kinetic parameters of NO, CO and O_2 binding – k_{on}, k_{off} and K_D – have been determined for H-NOX proteins from various species.[16] One particular H-NOX protein, the *Shewanella oneidensis* H-NOX (*So* H-NOX), has been suggested to exhibit characteristics of gas binding resembling human sGC: a capacity to break the Fe-histidine bond upon exposure to NO leading to the same five-coordinate heme iron with proximal NO as final product (**D**), presumably formed *via* an equivalent

---

[VI] In reference 14, the first crystal structure of human sGC in the NO activated state with the NO molecule resolved was published. This was published after the paper that this chapter is based on (A. M. Rozza, M. Papp, N. R. McFarlane, J. N. Harvey, J. Oláh, *Chem. Eur. J.* **2022**, 28, e202200930), so the findings in this chapter do not build upon the findings of reference 14.

That being said, reference 14 provides further justification to the findings in this publication, which is that NO has a higher propensity to bind to the distal side of the heme group (i.e., formation of **D*** is more facile than **D**).

unobservable intermediate (**C**). Such a similarity holds out the promise that this protein serves as a good model of human sGC.[16]

However, conclusions of a recent study appear to contradict the hypothesis involving the formation of **C**.[17] Through the use of UV-Vis and EPR spectroscopy, *So* H-NOX was investigated under limiting and excess NO conditions, and in both cases the product seemed to be the monoligated NO complex with distal coordination (**D***). While the observed breaking of the iron-histidine bond emphasizes the similarity between sGC and *So* H-NOX, it also reignites the debate on the activation mechanism and on the nature of the active species. Especially the events following the binding of the first NO molecule are disputable, whether the scission of the Fe-histidine linkage occurs independently or assisted by the binding of a second NO molecule, and which of **D*** or **C** (which could lead to the generation of **D**) are more easily accessible.

Spectroscopically, the various species shown in Figure 3.1(B) are typically identified using UV-Vis spectroscopy. In the resting state of the enzyme, the UV-Vis spectra are characterised by a Soret peak at ~430 nm and a single broad α/β peak at ~555 nm. These features are representative of a 5-coordinate ferrous heme. Upon binding of one NO equivalent, the Soret peak moves to ~400 nm and a shoulder at ~485 nm emerges.[18] However, an experimental challenge found in these systems is that active states **D*** and **D** present the same UV-Vis spectra. Bearing this in mind, for **D** to be formed, **C** must first be visited. Given **C** is predicted to be a short-lived intermediate, it is difficult to say unambiguously which active state is the most prominent.

In this chapter, several computational modelling techniques are used to explore key aspects of signalling by sGC and to assess possible reaction pathways in *So* H-NOX as a model of human sGC. QM methods are used to assess the energetics of the different structures shown in Figure 3.1(B) using a model system,

consisting of an Fe-porphin ring (i.e., a porphyrin ring without sidechains), an axial imidazole ring representing the histidine sidechain and NO molecules, where appropriate. Next, starting from the crystal structures of **A**, **B** and **D** from *So* H-NOX,[19] MD simulations have been used to address the following questions: (1) Does NO diffuse into the proximal pocket? (2) How does the tertiary structure change when the proximal His decoordinates upon rupture of the Fe-His bond? (3) Is the movement of the His into/out of the proximal pocket facile? Finally, coordination of the first and second equivalent of NO as well as decoordination of the His residue have been addressed using QM/MM calculations in a large model of the protein. With the aid of previous literature and the calculations and simulations performed herein, three core questions related to *So* H-NOX activation by NO-binding were set out to be resolved: (i) Is it plausible that a second NO equivalent binds to heme yielding intermediate **C**? (ii) Are pathways to **C** and **D*** energetically accessible starting from the NO-unbound structure **A**? (iii) Of **C** and **D***, which is the most easily accessible? All of the computational data generated herein is then used in conjunction with estimated physiological concentrations of NO gas to reach a conclusion about the favoured pathway.

3.2 Computational Details

3.2.1 Quantum Mechanical Calculations[VII]

A Fe-porphin model of the heme ring was used to model species **A**, **B**, **C**, and **D/D*** (in such a reduced model, species **D** and **D*** are identical) in vacuum. Structures were optimized with the Gaussian 16[20] program package employing the B3PW91 functional with Grimme's D3 dispersion correction[21] and the BJ damping function[22] together with the 6-311+G* basis set on Fe and the axial

[VII] The QM calculations were mostly performed by co-author J. Oláh, though N. R. McFarlane also contributed to discussion and analysis of the results.

NO ligand(s), and the 6-31G* basis set on all other atoms. Similar basis set compositions were already shown to be useful for the description of binding of NO and CO to Fe(II) and Fe(III) systems.[23,24] Second-derivative calculations ensured that the located minima corresponded to stable structures on the potential energy surface.

Additional single-point energies were calculated on the optimised geometries, at the DFT (BP86[25,26] and hybrid B3LYP[27], B3LYP*[28], and B3PW91[29] functionals, again including D3 dispersion correction[21] with the BJ damping function[22]) and natural pair orbital-based local DLPNO-CCSD(T_0)[30] levels using the Orca 4.2.1 program package.[31,32] All single-point calculations used the zeroth-order relativistic approximation (ZORA)[33] using the ZORA-def2-TZVP basis set on all atoms. For the DLPNO-CCSD(T_0) calculations, the reference wavefunction was based on the Kohn-Sham orbitals obtained with the BP86 functional. Test calculations using instead B3PW91 orbitals led to almost identical relative energies. Density fitting (RI) approximation was applied with the SARC/J auxiliary basis set in DLPNO-CCSD(T_0) calculations.

3.2.2 Molecular Dynamics Simulations[VIII]

The crystal structure coordinates of *So* H-NOX protein for **A**, **B**, and **D** were extracted from the PDB with pbd-id:4U99, pbd-id:4U9G and pbd-id:4U9B, respectively. They were used as a starting point for MD simulations of the four key structures in the proposed reaction mechanism: **A**, **B**, **D***, and **D**. The structures for **A** and **B** are based on proteins with some point mutations; for the MD study, the corresponding residues were back-mutated to the wild-type amino acids: Q154A/Q155A/K156A, using the Maestro software package. Under neutral pH, all titratable amino acids were preserved in the same protonation states as predicted by the

[VIII] The MD simulations were performed by co-author A. Rozza, though N. R. McFarlane also contributed to discussion and analysis of the results.

authors,[19] with the exception of His103 in **D**, where the proton was placed on δN instead of εN of the histidine ring, due to its coordination with the heme group. pbd-id:4U9G representing **B** includes carbon monoxide instead of NO, which was changed manually to NO. In order to simulate structure **B,** a bond between the nitrogen of the distal NO and Fe(II) of the heme was created manually by inclusion of a specbond.dat file. In structures **D*** and **D**, where there is no bond between the εN of the proximal His103 and Fe(II), the His103-Fe(II) bond was manually broken again with use of a specbond.dat file. Iterative solvation by the optimal point charge water model[34] was performed for all structures; water molecules were arranged in a rhombic dodecahedron box, where the protein was placed at least 12 Å from the box edge. The complete system was set to realistic physiological concentration of NaCl (0.15 M).[35]

The system was sequentially energy-minimized to diminish unfavourable initial contacts, followed by sequential relaxation of constraints on protein atoms in three NVT steps (100 ps each) and an additional NPT (100 ps) simulation to initiate 1 bar pressure. Finally, 150-450 ns MD simulations were performed in the NPT ensemble at 310 K and 1 bar, and without any positional restraints. Snapshots were taken at 4 ps intervals for further analysis. All MD simulations used the CHARMM27 force field and the GROMACS package.[36] The stability of the MD simulations was evaluated with the RMSD of mainchain heavy atoms (backbone heavy atoms and carbonyl oxygen) along the whole trajectory as well as obtaining its cumulative average.

The diffusion of NO within the protein matrix was investigated in the case of **A** using a previously validated methodology.[37] First, structures visited along the MD simulation of **A** were clustered based on the conformation of the mainchain heavy atoms. The mid-structure of the most populated cluster for **A** served as the starting structure for simulation of NO diffusion. 40 NO molecules,

modelled with a three-site model,[38] were placed into solvent-free holes around the protein system and an MD simulation was performed for the duration of 400 ns. Finally, the occupancy of the distal and proximal pockets by NO was calculated from the equilibrated portion of the trajectory.[37]

3.2.3 QM/MM Calculations[IX]

The calculations for the QM and MM regions of the system were performed with two different packages – Gaussian 09[39] and Tinker,[40] respectively. To combine the QM and MM parts of the system as a QM/MM calculation, the *QoMMMa* program was used.[41] In this implementation, *QoMMMa* uses an additive QM/MM scheme, and at the QM/MM boundary, the valences in the QM region are satisfied using the hydrogen link atom procedure.[42] To mitigate non-physical effects caused by this bridging, the charge on the neighbouring MM atom was set to zero, and its original charge was distributed across adjacent atoms to preserve a neutral charge. For the QM region, all calculations were performed using the BP86 or B3LYP functionals with D3 dispersion correction with the BJ damping function.[21,22] All calculations used the 6-311+G*/6-31G* basis set combination, with 6-311+G*being used to describe the atoms with the greatest electronic structure change – Fe, the four pyrrole nitrogens, and any incoming or bound NO molecules. The MM part of the system was described by the same CHARMM27 forcefield as was used in MD simulations.

The reaction paths leading from **B** to **C** and from **B** to **D*** (see Figure 3.1(B)) were obtained using the adiabatic mapping approach. Three snapshots taken from the MD simulation of **B** were used as starting structures. In order to model the conversion of **B** to **C**, an additional NO molecule was manually placed in the proximal pocket. The reaction coordinate used to drive the transformation

[IX] Some preliminary QM/MM calculations, not discussed here, were performed by co-author M. Papp, but all QM/MM calculations reported here were performed by N. R. McFarlane.

of **B** to **D*** was the distance between the heme iron and the εN of His103 ($d(Fe - \varepsilon N)$); all species were modelled as doublets using unrestricted DFT. The reaction coordinate for the reaction path from **B** to **C** was the difference between two distances: (1) the distance between the heme iron and the εN of His103 ($d(Fe - \varepsilon N)$); (2) the distance between the heme iron and the N of the incoming NO ligand ($d(Fe - NO)$). This reaction was studied in the singlet state with care taken to identify broken symmetry open-shell singlet solutions where they lay lower in energy than the closed-shell solution.

3.3 Results and Discussion

3.3.1 Quantum Mechanical calculations

While numerous previous studies have modelled bonding of NO to the imidazole-ligated heme group (**B**) with QM methods,[24,43–47] for consistency and in order to obtain near-benchmark quality results, a new set of calculations was performed. These calculations included a broad overview of DFT functionals, as well as some benchmarking DLPNO-CCSD(T_0) which, despite known caveats about accuracy,[48,49] should provide a reasonable benchmark for the DFT functionals. Only results obtained with the ZORA-def2-TZVP basis set on all atoms are presented for consistency between DFT and DLPNO-CCSD(T_0) calculations. This subsection will start by considering the spin state of the ground state of the studied model complexes (**A**, **B**, **C** and **D**/**D*** depicted in Figure 3.1(B)), followed by the discussion of the overall energy changes associated with NO and histidine coordination to the heme group.

Firstly, it is important to establish the electronic ground states of the species shown in Figure 3.1(B). The ground state of **A** is well-known to be a high-spin quintet[50,51] which can react in a fast, almost barrierless spin-forbidden reaction with NO to yield **B**, which has a doublet ground state.[24,52] As in previous work,[53] the quartet and sextet states of **B** are found to lie higher in energy. **C** is

an elusive intermediate and as such there is no experimental or previous computational data available on its ground spin state. Using the BP86-D3 functional, **C** is predicted to have a singlet ground state lying 16.7 kcal/mol below the triplet, while the hybrid functionals predict close-lying triplet and singlet states, the triplet being more stable by 2-3 kcal/mol. It has previously been shown that gradient corrected methods such as BP86 tend to favour low-spin states, while hybrid functionals often yield small energy differences between high- and low- spin states.[54,55] Bearing this discussion in mind, due to the minimal energy difference found with hybrid functionals, **C** was modelled in the singlet state. Decoordination of the imidazole ring gives **D**/**D***, which has previously been shown to have a doublet ground state.[56]

Table 3.1: *QM relative energies (kcal/mol, including the ZORA with the ZORA-def2-TZVP basis set) for all structures studied herein –* ***A****,* ***B****,* ***C****, and* ***D****/****D****. The relative energy changes for the two reaction pathways which are examined in more detail using QM/MM methods are also shown.*

Method	E_{rel} **(kcal/mol)**					
	A	**B**	**C**	**D/D***	ΔE_{BC}	ΔE_{BD*}
BP86-D3	0.0	-56.6	-62.9	-41.4	-6.3	15.2
B3LYP-D3	0.0	-22.2	-6.7	-8.6	15.4	13.5
B3LYP*-D3	0.0	-32.7	-24.0	-18.0	8.7	14.7
B3PW91-D3	0.0	-24.5	-9.7	-8.7	14.8	15.8
DLPNO-CCSD(T_0)	0.0	-13.3	3.3	5.9	16.7	19.3

Turning now to the relative energies of the studied species (Table 3.1), the initial structure, **A**, is the relative energy baseline from which other energies are calculated. In QM calculations with a small model not including any amino acid residues, **D** and **D*** are equivalent, so are considered as one in the following discussion.

Inspection of the data in Table 3.1 gives information on two important aspects: (1) how comparable are the results obtained with various DFT functionals with the benchmark DLPNO-CCSD(T_0) calculations?; (2) What kind of chemistry related to sGC activation do the obtained numbers suggest? Firstly, the dissociation of imidazole from **B** to yield **D*** is consistently endothermic for all functionals. The computed relative energies for substituting imidazole by NO (going from **B** to **C**) vary more significantly, with the hybrid functionals predicting an endothermic process, while BP86-D3 finds this step to be exothermic, implying very strong binding of the second NO molecule. Given the lower admixture of exact exchange, B3LYP*-D3 yields a result closer to BP86-D3 than do the other two functionals.

The DLPNO-CCSD(T_0) relative energies do not perfectly match those obtained with any of the functionals, but bearing in mind the expected inaccuracies in DLPNO-CCSD(T_0) for open-shell systems,[48,49] there is acceptable agreement with B3LYP-D3 and B3PW91-D3. The BP86-D3 functional agrees much less closely, leading to the conclusion that the predicted high stability of species **C** with this functional is an artefact.

Looking now at the predicted chemistry, NO binding to heme is energetically favourable (forming **B** from **A**), but being aware of the unfavourable entropic contribution not included here, the binding is still predicted to be reversible, especially for low NO concentration. Next, dissociation of the imidazole group to yield **D** is somewhat endothermic, but the energy change is relatively modest, consistent with relatively facile dissociation of the

histidine sidechain from intermediate **B**. Finally, except for BP86-D3, coordination of a second NO group to yield **C** is endothermic. This prediction is consistent with all hybrid functionals and with DLPNO-CCSD(T_0) so even allowing for the errors in these methods, direct formation of **C** from **B** is unlikely to happen in sGC and related systems, especially considering the unfavourable entropy of binding (not included in the relative energies of Table 3.1) and the low physiological concentration of NO of 100 pM – 5 nM.[57] As **C** lies slightly lower in energy than **D**/**D***, it does however represent a plausible intermediate in an exchange process that would convert **D*** to **D**, as will be explored in more detail in Section 3.4.

3.3.2 Molecular Dynamics Simulations

Before analysing the MD trajectories, it is important to assess the stability of the MD simulations for **A**, **B**, **D***, and **D**. This was achieved by measuring the RMSD of the protein's mainchain (backbone plus the carbonyl oxygen). Table A1 shows the average values of the RMSDs, which range from 0.12 to 0.16 nm, and Figure A1(A) shows the RMSD with respect to time. This deviation from the zero ns reference structure appears to be acceptable. The cumulative average of the mainchain RMSD, as shown in Figure A1(B), also shows flattened curves, highlighting the stability of the simulations.

3.3.2.1 NO Diffusion Process

NO binding to heme proteins involves two key stages: (1) ligand diffusion to the active site; (2) bond formation between the ligand and the protein, whose description requires fundamentally different approaches.[24] In the previous section (as well in Section 3.3.3), the latter is considered, and this section will consider the NO diffusion process.

Proper modelling of ligand diffusion into proteins necessitates long MD simulations where the diffusion events from the solvent phase into the active site and vice versa are well-equilibrated.

Therefore, a 400 ns MD simulation of **A** with 40 NO molecules added to examine the probability of NO diffusion into both distal and proximal pockets was performed. The distal pocket (the pocket defined by the upper face of the heme group, with respect to Figure 3.2) is surrounded by residues**:** Met1, Ile5, Leu73, Leu138, Cys141, and Leu145; while the proximal pocket (the pocket defined by the lower face of the heme group, with respect to Figure 3.2) is defined by residues**:** His103, Asn107, Ser114, and Pro116.

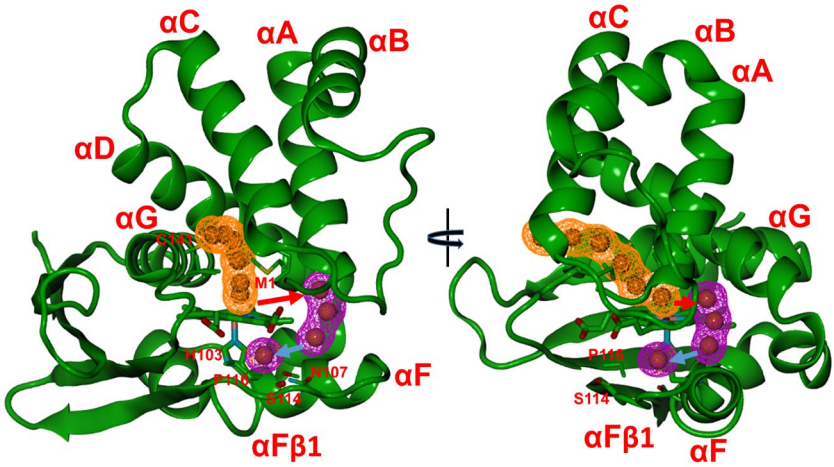

Figure 3.2: *Identified gas migration tunnels towards the distal and proximal pockets in So H-NOX protein. The orange mesh depicts the tunnel that transports the NO gas molecules from the external solvent environment to the protein matrix and finally to the distal pocket. The red and light blue arrows show the movement of NO gas molecules from the distal face to the proximal pocket which is represented by purple mesh. The figures were generated using Chimera.*[58]

Based on Table 3.2 – which provides a general overview of some characteristic properties of NO diffusion into each pocket – comparison of the properties of *So* H-NOX with other NO-sensing bacterial H-NOX systems studied for NO diffusion can be achieved.[37] The average time that NO can occupy the distal pocket of *So* H-NOX was found to be significantly shorter than that of *Nostoc sp.* H-NOX (920 ps) but of the same order of magnitude as *Kordia algicida* (116 ps). However, the percentage of the distal

pocket occupancy with at least one NO molecule was found to be reasonably comparable between these three H-NOXs. On the proximal side, NO was trapped within this pocket of *So* H-NOX for a seemingly shorter time than in both *Nostoc sp.* and *Kordia algicida*, 212.2 ps and 7940 ps, respectively, with a much lower number of in-and-out-events in the two H-NOXs. The same pattern was seen for the occupancy percentage with at least one NO molecule being inside the proximal pocket.

Table 3.2: *Ligand-binding properties of the distal and proximal pockets derived from 400 ns MD simulations of* ***A****.*

Binding Pockets	**Distal**	**Proximal**
Average time trapped [a]	58.2 ps	9.7 ps
Number of in-and-out events [b]	6358	6975
Occupancy with at least one NO molecule [c]	90.5%	16.6%
Maximum number of gas molecules [d]	3	1

[a] The average amount of time that NO gas molecules spent inside a given pocket before exiting again. ***[b]*** The number of in-and-out diffusion events made by NO gas molecules, where "protein core → geminate pair" represents the in-event path, and the reversible direction of this route represents the out-event path.[24,37] ***[c]*** Occupancy of the pockets with at least one NO molecule during the simulation time. ***[d]*** Maximum number of NO gas molecules found inside both pockets over the course of the MD simulation.

In the simulations performed herein, NO molecules tend to smoothly diffuse into the distal pocket and at least one is present there during most of the simulation time, sometimes indeed with more than one NO present (Table 3.2). The proximal pocket, on the other hand, is only ever occupied by one gas molecule at a time, and this for only 16% of the MD trajectory. Thus, NO appears to be able to enter both pockets, so that reaction from both sides might occur, though geminate recombination in the distal pocket, which is expected to be barrierless, should be more facile. In a previous experimental study on sGC using time-resolved spectroscopy,[59] the time-scale for NO rebinding from the protein core to a 5-

coordinate His-bound heme group (i.e. to **A**) was found to be as short 6.5 ns.

Inspection of the tunnelling system that the gas ligand uses to access the distal and proximal pockets enables an understanding of why the distal pocket has a higher occupancy of NO molecules than the proximal pocket. After visually scrutinizing the inward/outward migration events of the gas ligands (i.e., from the solvent to the protein matrix and finally the pocket to the protein matrix then to the solvent environment), one short route was recognized. This tunnel was observed to be most likely taken by the gas molecules to first reach the distal pocket and then bounce horizontally over the heme distal face towards a cavity next to the heme (shown in purple mesh in Figure 3.2) that lies between helices αA, αG, and αF, leading to the proximal pocket. Only one of the 40 NO gas molecules found that pathway into the proximal pocket, and because there was no simple exit pathway, it remained in the proximal pocket (and no other NO molecules exchanged positions with it). This is presumably due to the apparent short length of the tunnel, meaning that NO molecules can easily stream though and ultimately making the distal pocket occupied much of the MD simulation time. This short tunnel (shown in orange mesh in Figure 3.2) is lined by helices αC, αD and αG, allowing the gas ligands to move directly from the external solvent environment to the vicinity of the heme propionate sidechains and subsequently to heme group. This tunnel has been observed as the shortest route in two other bacterial H-NOX proteins from *Nostoc sp.* and *Caldanaerobacter subterraneus*, making it highly conserved and typical molecular tunnel in H-NOX proteins.[37,60] In addition, another longer access tunnel (between helices αA and αD) that gas molecules only rarely used to migrate to the distal pocket was identified. Previously, a third tunnel in H-NOX proteins was found,[37] however, it is blocked by two aromatic amino acids (His81 and Tyr110), which likely form a steric barrier

mostly preventing NO molecules from diffusing into the H-NOX protein matrix.

3.3.2.2 Proximal Histidine Swing-Out

Dissociation of the histidine proximal ligand from the heme group to generate **D*** or **D** is known to allow the proximal histidine to move away from the heme partner, and to be able to lead to pronounced structural changes that are believed to be important for signal transmission between sGC domains. A crystal structure for **D** confirms this conformational change. However, the position of the proximal histidine is unknown in the case of the distal NO complex **D*** for which no crystal structure currently exists.[61] What is also unknown is the ease with which the protein changes conformation upon His103 decoordination, and the relative free energy of the "inward" and "outward" conformations of the His103 sidechain, in which it is respectively "close" and "distant" to the heme group. In this context, MD trajectories of **D*** and **D** complexes (which differ in whether the NO ligand is attached to the distal or proximal side of the heme group) were thoroughly investigated for the motion of the proximal histidine after dissociation from the proximal face of heme, how it affects the proximal subdomain of H-NOX protein, and what its favourable location is in both species.

MD simulations for **D*** and **D** complexes show closely similar behaviour for the motion of the proximal histidine. Along the simulation trajectory of both complexes, two distinct conformations (inward & outward) for the proximal His103 were observed. The outward conformation was found to be the dominant state in **D** and **D*** complexes, but the inward conformation was also seen in a non-negligible portion of the simulations. Figure 3.3 shows the transition events between the two conformations, which are represented by a change in $\mathrm{Fe-His}_{\varepsilon\mathrm{N}}^{103}$ distance over the simulation time. It is also clear that at an early stage of both complexes' simulations, the transition from

inward to outward in **D***- or from outward to inward in **D** took nearly the same amount of time (~50 ns). Then, for the next 100 ns of the simulation of **D**, the transition occurred four times, with an average of 11 ns of being trapped in the inward conformation. In the **D*** simulation, on the other hand, the His103 sidechain remains in the outward position, with no further transitions. These differences may simply be a product of insufficient sampling, and further trajectory replicas may show the same behaviour.

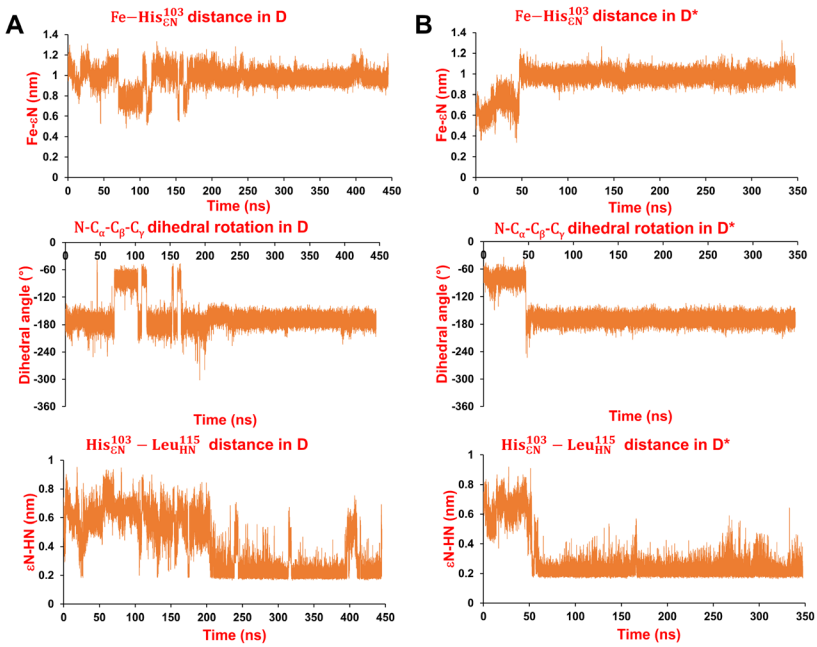

Figure 3.3: *Time evolution (ns) of the calculated distances (nm) and dihedral rotation around Cα – Cβ bond in both complexes (A)* ***D*** *and (B)* ***D****. The distances are between εN of His103 and both heme-Fe & amide hydrogen of Leu115. The N-Cα-Cβ-Cγ dihedral angle of His103 is calculated in both* ***D*** *and* ***D****.*

The two locations of the proximal histidine in both NO-bound forms were visually investigated to identify the set of interactions within each conformation: (a) Histidine was observed to be in the heart of the proximal pocket (inward) – with an average $\mathrm{Fe} - \mathrm{His}_{\varepsilon\mathrm{N}}^{103}$ distance of ~7.5 Å – due to rotation of the αF helix followed by

motion of the αFβ1 loop (Figure 3.4,A), with an average dihedral angle of ~ -79° around $C_\alpha - C_\beta$ (χ_1) bond of His[103]. The rotation of the His103 imidazole ring may also contribute to the inward displacement. Upon His103 decoordination, the imidazole ring of His99 rotates outward to the solvent area, eliminating the H-bond between the carbonyl oxygen of His99 and H-δN of His103, allowing the His103 side chain to move freely. This is clearly seen in the simulation of **D*** (see Figure A3,B), where the His103 was in closer contact than in the initial conformation of **D**. Thus, the side chain of His103 may be positioned away from the heme proximal face, and instead within plausible hydrogen-bonding distance of either or both of Ile98 or Ile118, (δN-H of His103 with carbonyl oxygen "O=C" of Ile98 & εN of His103 with amide hydrogen "H-N" of Ile118) (see Figure A2,A & B); (b) Swinging-out to the solvent-accessible void (outward), which, as was shown above, appeared to be the most favourable position for the proximal His103 (Figure 3.3,A and Figure 3.4,A), and corresponds to its position in the crystal structure of **D**.[19] The "swinging out" motion is associated with change in the dihedral rotation around the $C_\alpha - C_\beta$ bond (χ_1) by ~90° from the inward state[19] – which was characterized by an average torsion angle of ~ -169° – and with rotation of the αF helix, which leads to changes in positions of the backbone atoms of the proximal histidine by ~ 1 Å. Consequently, the side chain of His103 was then trapped in a new set of polar interactions with Asn107, His99, and water molecules (see Figure A3,A & B). This group of interactions is noticeably wider than the set established in the inward location, which may be the key to the stabilization of the outward conformation for most of the simulation time. The rotation of the αF helix also could aid in the formation of an H-bond between the carbonyl oxygen of Asp100 and the amide hydrogen of Leu104, which might keep the histidine side chain to be exposed to the solvent area, and this has previously been shown in a study on the same system.[19] While histidine was in the outward state, another H-bond was observed that may establish an interaction

between αF helix and αFβ$_1$ loop, particularly between the εN of His103 and the amide hydrogen of Leu115 (Figure 3.3,A & B and Figure 3.4,A)

Overall, the fluctuation between the inward and outward states of the histidine side chain was found to occur on a relatively short time-scale, consistent with the short time constant (70 ps) that has been observed for the rebinding step of the proximal histidine to the heme moiety in sGC.[59] This short time-scale might be interpreted as corresponding to the easiest route for return to the starting species (**A**) from either **D*** or **D**, but this assignment will be discussed in more detail in Section 3.3.3.

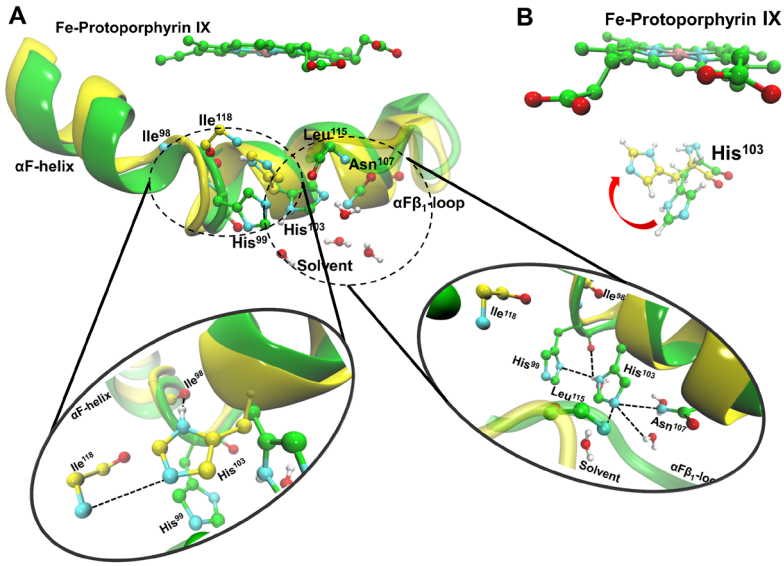

Figure 3.4: *(A) Structural fluctuations of the proximal subdomain (αF-helix & αFβ1-loop) between the inward (helix, loop, and carbon are coloured in yellow) and the outward (helix, loop, and carbon are coloured in green) conformations – It shows the hydrogen bond network in both states. (B) Shows the rotation the dihedral around Cα-Cβ bond of the proximal His103. The figure was generated using VMD.*[6]

3.3.3 QM/MM Calculations

This subsection will consider the predicted energies for bonding and dissociation events at the heme group, based on new QM/MM calculations and on reference to previous QM calculations[24] and those reported here (see Section 3.3.1). Geminate binding leading from **A** to **B** was predicted to be facile once the NO ligand is present in the distal pocket in previous work[24] and has not been considered again here. Here the focus is on whether dissociation of the proximal histidine occurs directly (**B** to **D***) or is assisted by the coordination of a second NO molecule on the proximal side of the heme (**B** to **C**). Both processes were predicted to be endothermic by QM calculations; the present QM/MM calculations aim to investigate their feasibility in the protein environment.

To study the pathway from **B** to **D***, any changes in electronic spin should first be considered. In this case, both **B** and **D*** are found to have a doublet ground state, so the transition is formally spin-allowed. To generate reaction pathways, the adiabatic mapping approach was used on three starting structures of **B** using BP86-D3 and B3LYP-D3 to describe the QM region. While the BP86-D3 functional is overall not very accurate for sGC energetics, it reproduces the relative energy of **B** and **D/D*** quite accurately so it can be used here (see Table 3.1). Average bond lengths of reactant and product structures obtained from mapping procedures with both methods are given in Figure 3.5(I). Turning first to BP86-D3, the changes in bond lengths tell a relatively simple story whereby $d(Fe - NO_{distal})$ remains unchanged at ~1.70 Å and $d(Fe - \varepsilon N)$ lengthens from 2.50 Å to a non-bonded distance of 4.30 Å. The BP86-D3 and B3LYP-D3 predicted structures are also very similar to one another, with the only noteworthy difference being a slight increase in $d(Fe - NO_{distal})$, but this is comfortably within expected error margins. Both methods are in accordance with what is expected to happen in this reaction pathway, and the

changes in bond lengths agree favourably with previous QM/MM studies on a related system.[24]

Moving now to the energy changes associated with these conversions, energy profiles from both methods show no additional barrier above their endothermicity, and accordingly, no local minima for the product, **D***, could be identified. Hence, the predicted relative energy of this species depends in a somewhat arbitrary way on the point along the calculated dissociation curves chosen to correspond to "product". Using a $d(Fe - \varepsilon N)$ of 4.3 Å, **D*** to was found to lie 17.2 and 19.1 kcal/mol above **B** with BP86-D3 and B3LYP-D3, respectively, as compared to 15.2 and 13.5 kcal/mol obtained in the QM calculations, respectively. These findings are expected, given that it was found in Section 3.3.1 that BP86-D3 could compare to more accurate methods when modelling the change from **B** to **D**/**D***. Energy profiles explicitly showing the QM and MM contributions with associated structures for these calculations, as well as reactant and product structures for all snapshots with the comparison between levels of theory are given in Appendix A.

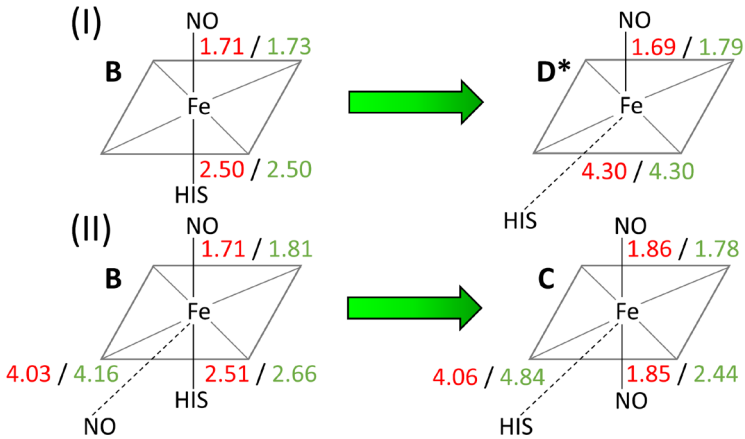

Figure 3.5: *Average bond lengths of reactant and product structures for QM/MM modelling of the transition from (I)* ***B*** *to* ***D****, and (II)* ***B*** *to* ***C****. In red are the bond lengths associated with BP86-D3 modelling, and in green are those associated with B3LYP-D3 modelling.*

Dissociation of His leads to a five-coordinate heme group with a distal NO ligand, **D***. It is plausible that this could convert to **D** *via* **C** by entrance of a second NO ligand in the proximal pocket, binding of NO to the proximal face of the heme group, and dissociation of the distal NO molecule. These steps have not been considered in detail, but given the relatively low energy of the putative **C** intermediate in the QM calculations, this inversion could presumably occur relatively easily. Of course, to a first approximation, neglecting interactions with the protein environment, **D*** to **D** conversion should be thermoneutral, but considering these interactions, one of the two may in fact be somewhat favoured in the protein. Given that there are experimental crystal structures of a bacterial homologue of **D** but not of **D***, ostensibly the former is more stable. This finding is also supported by the more recent publication of human sGC where **D** is found to be the primary NO-bound species.[14]

For the study of the pathway leading from **B** to **C** upon binding of a second NO equivalent, the starting structure of **B** has an additional NO ligand added in the proximal pocket. Inclusion of the additional NO means that the starting structure now has two possible ground states: a triplet or an open-shell singlet. As it was found in Section 3.3.1 that the singlet and triplet states of **C** with two bound NO ligands lie very close in energy, the reaction pathway was modelled in the singlet state only. As the reaction progresses from **B** to **C**, the obtained electron distribution changed from an open-shell singlet to a closed-shell singlet. The spin population of atoms was continuously monitored, and great care was taken to locate the lower-lying of the open-shell and closed-shell states at each point. As above, adiabatic mapping procedures were performed using the BP86-D3 and B3LYP-D3 functionals for the QM region. In both cases, the reaction was followed by constraining a reaction coordinate corresponding to the difference between the two proximal Fe-ligand distances. It is worth highlighting that the constraint used is different to that used in the study of the route

leading from **B** to **D*** and allowed more degrees of freedom. The average bond lengths of reactant and product structures using both methods are given in Figure 3.5(II).

As it does not directly participate in the reaction, $d(Fe - NO_{distal})$ is expected to change relatively minimally – as was observed in the **B** to **D*** pathways. In this case, $d(Fe - NO_{distal})$ changes from averages of 1.71 Å to 1.86 Å for BP86-D3, and 1.81 Å and 1.78 Å for B3LYP-D3. Particularly for BP86-D3, the change is more notable. The exact reason for this increase is not clear, but it is, overall, a relatively small change so is assumed to have a negligible effect on the reaction pathway. When using BP86-D3, $d(Fe - NO_{proximal})$ converges to an average distance of 1.85 Å, which is relatively similar to $d(Fe - NO_{distal})$. On the other hand, with B3LYP-D3, $d(Fe - NO_{proximal})$ converges to an average bond length of 2.44 Å – significantly greater than $d(Fe - NO_{distal})$. This effect could be a result of the constraint used in this pathway, which allows simultaneous variation of both $d(Fe - NO_{proximal})$ and $d(Fe - \varepsilon N)$. It is plausible that the energy decrease associated with moving the histidine further away from the metal centre is greater than bringing the proximal NO closer, so the constraint choice may not be entirely suitable, here. In any case, overall, the changes in bond lengths along the reaction paths make chemical sense and, overall, align well with the expected reaction scheme.

Considering now the relative energy changes, when using BP86-D3, there is an energy barrier of around 6.4 kcal/mol. For the case of B3LYP-D3, the corresponding step was found to have no additional barrier above the endothermicity. Therefore, for B3LYP-D3, much like in the case of **B** to **D***, defining when the product has been reached is somewhat arbitrary. Here, the product was taken to be reached once $d(Fe - NO_{proximal})$ stopped changing to any significant degree. The relative energy changes for the transition from **B** to **C** are −1.1 and +10.2 kcal/mol for BP86-D3 and B3LYP-D3, respectively. These QM/MM calculated reaction energies

agree in broad terms with those reported in Section 3.3.1, where the same process is predicted to have an energy change of -6.3 and 15.4 kcal/mol for BP86-D3 and B3LYP-D3, respectively. Energy profiles explicitly showing the QM and MM contributions with associated structures for these calculations, as well as reactant and product structures for all snapshots with the comparison between levels of theory are given in Appendix A.

As discussed in Section 3.3.1, the B3LYP-D3 results are considered to be more accurate, so conversion of **B** to **C** is most likely endothermic, with a reaction energy similar to that for conversion of **B** to **D***. Based on these QM and QM/MM energies, and given the fact that NO binding to form **C** is entropically unfavourable (an effect not included here), formation of **D*** should be more favourable than formation of **C**. However, **D*** can most likely convert to **D** with **C** as intermediate.

Considering now the overall process in the protein: in the presence of NO, **A** can be expected to convert ultimately to **D**, accompanied by histidine dissociation. Under low NO partial pressures, with the fact that **D** is similar in stability to **A**, **D** and **D*** are expected to slowly convert back to **A**. Conversion of **D*** back to **A** should be straightforward, simply involving recoordination of the histidine sidechain and dissociation of NO.[24] On the other hand, conversion of **D** to **A** in the absence of excess NO should be less facile as NO is required to initiate conversion of **D** to **C**. No other route leading from **D** to **A** is apparent, so at low NO partial pressures, conversion of **D** to **A** is predicted to be slow.

3.4 Overall Findings

Based on these QM, MD, and QM/MM results, and on their confrontation to experimental observations, an overall scenario for the chemical and conformational changes occurring upon interaction between NO and sGC will be suggested in this section.

For this discussion, it is useful to first consider the relevant physiological concentrations of NO gas leading to switching signalling 'on' and 'off'. NO is synthesised by the NO synthase complex in endothelial cells and diffuses into neighbouring smooth muscle cells, where it binds the heme group of the H-NOX domain to activate the sGC receptor, leading to cGMP accumulation. Oxidation of the heme group or heme loss can cause NO insensitivity, and this can also be seen as an outcome of heterodimer dissociation.[62]

Based upon an assortment of data, it has been suggested that the physiological concentration of NO gas lies in the range of 100 pM – 5 nM.[57] This is consistent with previous studies of the activity of sGC under increasing concentrations of NO,[63] with the corresponding activity plots showing simple saturation behaviour, which can be interpreted as meaning that sGC binds only 1 gas equivalent in the final activated state. It should also be noted that the activity reaches saturation at approximately 10 nM, similar to the upper bound of the proposed physiological concentration. Higher NO concentrations may occur in different cell types under oxidative conditions, such as leukocytes, endothelial cells, cardiac myocytes, nerve cells, and fibroblasts.[64] These conditions stimulate the expression of cytosolic iNOS to produce high level of NO. Excess NO may react with superoxide at a diffusion-controlled rate, resulting in the formation of the peroxynitrite anion, which causes cell toxicity.[65]

The initial route followed upon NO reaction with sGC certainly involves NO binding to form six-coordinate intermediate **B**. Given the relatively low binding energy, and the entropic cost of binding, however, as the partial pressure of NO drops, the thermodynamics of binding will become less favourable. In this case, dissociation to regenerate **A** and NO is expected to occur, with the equilibrium presumably shifting as the NO concentration varies from the low to high ends of its physiological range. Species **B** also has a

relatively low binding energy for the imidazole sidechain of the proximal histidine, so that decoordination of this group to yield initially **D*** should be feasible. The MD simulations performed herein show that the His sidechain in **D*** can quite easily rotate outward into solvent, with this motion being accompanied by shifts in helix positions that are consistent with this decoordination leading to significant conformational change in the protein. The time-scale of the MD simulations performed here as well as the size of the modelled system do not enable analysis of the large-scale conformational changes in detail, but the role of the NO-binding induced His decoordination in 'triggering' it is supported by the present results.

In principle, NO could also bind to the heme group on the proximal side, leading to conversion of **B** to the bis-nitrosyl species **C**. Both the DFT and benchmark DLPNO-CCSD(T_0) results however show that NO binding in species **C** is relatively weak, which, together with the corresponding entropic cost, suggests that this species will not be formed directly from **B** as a stable intermediate in sGC at physiological concentrations of NO.

The calculations performed herein also suggest a possible role for **D**, with a five-coordinate heme group and a proximal NO ligand. This species could be accessed from **D***, in presence of elevated NO partial pressures, through diffusion of a second NO ligand into the proximal pocket and exchange through **C**. In this exchange process, as mentioned above, **C** is not suggested to be a stable long-lived species but is stable enough to be a plausible intermediate *en route* to **D**. The existence of a crystal structure of **D** suggests that the inversion process from **D*** to **D** may be somewhat thermodynamically favourable. The relevance of this conversion is as follows: while the intermediate **D*** can readily convert back to the resting state, **A**, through reversal of the conformational change, rebinding of the proximal His sidechain, and NO dissociation, there is no obvious route for conversion of **D**

to **A**. Hence at low NO partial pressure, **D*** would become thermodynamically disfavoured and should disappear. On the other hand, if the exchange from **D*** to **D** can take place during the period of elevated NO concentration associated with signalling, then reversal to the resting state, **A**, will become much slower, because at low NO concentration, **D** can only switch rather slowly to **D***, presumably again through **C**.

3.5 Conclusions and Future Work

Gas-molecule mediated signalling is key to the proper functioning of organisms. sGC is the sole receptor of nitric oxide in the human body, and it contributes to versatile physiological processes. Although it has been thoroughly investigated experimentally and is an important drug target, there has been a long-standing debate on its reaction mechanism.

Using a combination of computational chemistry tools (QM calculations, MD simulations, and QM/MM calculations) and the *So* H-NOX model protein, this work suggests that after the initial binding of NO to the distal side of the heme group, the iron-histidine bond undergoes facile scission and a 5-coordinate heme-NO species (**D***) is formed. The decoordination of histidine is facilitated by its rotation outward to the solvent, leading to slight conformational changes in the protein. Under elevated concentrations of NO (i.e., during signalling) the distal five-coordinate heme-NO complex could be converted *via* the energy-rich bis-NO-heme intermediate (**C**) to the five-coordinate proximal heme-NO adduct (**D**). This species (**D**) with proximal-NO coordination may be thermodynamically somewhat more stable than **D*** and may be the major species involved in signalling due to its slower conversion back to the resting state of the enzyme (**A**) compared to **D***.

sGC is a rather complex protein, containing 4 domains, each with a respective purpose. Until recently, the complete structure of

mammalian sGC was not available, therefore opening the door to more extensive computational modelling. While crystal structures of NO-activated and basal sGC exist,[5,14] the atomistic detail into the dynamics connecting the two states remains unclear. In this work, this problem was tackled in part by investigating the motion of histidine following its decoordination, but the use of accelerated MD simulations could provide some interesting insight into the interplay between activated and inactivated sGC. Notably, for sGC, these two states are rather different, with the PAS domains remaining relatively fixed,[66] while the catalytic domains show significant structural perturbation. The exact reasoning for such large conformational changes remains unclear, at present.

3.6 References

1 Y. Zhao, P. E. Brandish, D. P. Ballou and M. A. Marletta, *Proceedings of the National Academy of Sciences of the United States of America*, 1999, **96**, 14753–14758.

2 E. R. Derbyshire, A. Gunn, M. Ibrahim, T. G. Spiro, R. D. Britt and M. A. Marletta, *Biochemistry*, 2008, **47**, 3892–3899.

3 E. Martin, V. Berka, I. Sharina and A. L. Tsai, *Biochemistry*, 2012, **51**, 2737–2746.

4 S. P. L. Cary, J. A. Winger and M. A. Marletta, *Proceedings of the National Academy of Sciences of the United States of America*, 2005, **102**, 13064–13069.

5 Y. Kang, R. Liu, J.-X. Wu and L. Chen, *Nature*, 2019, **574**, 206–210.

6 W. Humphrey, A. Dalke and K. Schulten, *Journal of Molecular Graphics*, 1996, **14**, 33–38.

7 N. B. Fernhoff, E. R. Derbyshire and M. A. Marletta, *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**, 21602–21607.

8 N. Sayed, P. Baskaran, X. Ma, F. Van Den Akker and A. Beuve, *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**, 12312–12317.

9 S. VS. Kharitonov VG, Sundquist AR, *J Biol Chem*, 1995, **270**, 28158−28164.

10 E. Martin, V. Berka, A. L. Tsai and F. Murad, *Methods in Enzymology*, 2005, **396**, 478–492.

11 D. M. Lawson, C. E. M. Stevenson, C. R. Andrew, S. J. George and R. R. Eady, *Biochemical Society Transactions*, 2003, **31**, 553–557.

12 S. Louka, S. M. Barry, D. J. Heyes, M. Q. E. Mubarak, H. S. Ali, L. M. Alkhalaf, A. W. Munro, N. S. Scrutton, G. L. Challis and S. P. De Visser, *Journal of the American Chemical Society*, 2020, **142**, 15764–15779.

13 J. Lang, A. Maréchal, M. Couture and J. Santolini, *Biophysical Journal*, 2016, **111**, 2099–2109.

14 R. Liu, Y. Kang and L. Chen, *Nitric Oxide*, 2023, **134–135**, 17–22.

15 E. M. Boon, J. H. Davis, R. Tran, D. S. Karow, S. H. Huang, D. Pan, M. M. Miazgowicz, R. A. Mathies and M. A. Marletta, *Journal of Biological Chemistry*, 2006, **281**, 21892–21902.

16 G. Wu, W. Liu, V. Berka and A. lim Tsai, *Biochimie*, 2017, **140**, 82–92.

17 Y. Guo, D. L. M. Suess, M. A. Herzik, A. T. Iavarone, R. D. Britt and M. A. Marletta, *Nature Chemical Biology*, 2017, **13**, 1216–1221.

18 J. R. Stone and M. A. Marietta, *Biochemistry*, 1994, **33**, 5636–5640.

19 M. A. Herzik, R. Jonnalagadda, J. Kuriyan and M. A. Marletta, *Proceedings of the National Academy of Sciences*, 2014, **111**, E4156–E4164.

20 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T.

Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, Gaussian16 Gaussian, Inc., Wallingford CT 2016.

21 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *Journal of Chemical Physics*, 2010, **132**, 154104.

22 S. Grimme, S. Ehrlich and L. Goerigk, *Journal of Computational Chemistry*, 2011, **32**, 1456–1465.

23 J. Oláh and J. N. Harvey, *The Journal of Physical Chemistry A*, 2009, **113**, 7338–7345.

24 A. Lábas, D. K. Menyhárd, J. N. Harvey and J. Oláh, *Chemistry - A European Journal*, 2018, **24**, 5350–5358.

25 Alex D. Becke, *Physical Review A*, 1988, **38**, 3098–3100.

26 J. P. Perdew, K. Burke and M. Ernzerhof, *Physical Review Letters*, 1996, **77**, 3865–3868.

27 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *Journal of Physical Chemistry®*, 1994, **98**, 11623–11627.

28 M. Reiher, O. Salomon and B. A. Hess, *Theoretical Chemistry Accounts*, 2001, **107**, 48–55.

29 Alex D. Becke, *The Journal of Chemical Physics*, 1993, **98**, 5648–5656.

30 C. Riplinger and F. Neese, *The Journal of Chemical Physics*, 2013, **138**, 034106.

31 F. Neese, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2012, **2**, 73–78.

32 F. Neese, *WIREs Computational Molecular Science*, 2018, **8**, 1327.

33 E. Van Lenthe, E. J. Baerends and J. G. Snijders, *The Journal of Chemical Physics*, 1993, **99**, 4597–4610.

34 S. Izadi, R. Anandakrishnan and A. V. Onufriev, *Journal of Physical Chemistry Letters*, 2014, **5**, 3863–3871.

35 G. A. Ross, A. S. Rustenburg, P. B. Grinaway, J. Fass and J. D. Chodera, *J Phys Chem B*, 2018, **122**, 5466–5486.

36 S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845–854.

37 A. M. Rozza, D. K. Menyhárd and J. Oláh, *Molecules*, 2020, **25**, 1–19.

38 M. Meuwly, O. M. Becker, R. Stote and M. Karplus, *Biophysical Chemistry*, 2002, **98**, 183–207.

39 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam,

M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian09 Gaussian, Inc., Wallingford CT 2009.

40 J. A. Rackers, Z. Wang, C. Lu, M. L. Laury, L. Lagardère, M. J. Schnieders, J. P. Piquemal, P. Ren and J. W. Ponder, *Journal of Chemical Theory and Computation*, 2018, **14**, 5273–5289.

41 J. N. Harvey, *Faraday Discussions*, 2004, **127**, 165–177.

42 N. Reuter, A. Dejaegere, B. Maigret and M. Karplus, *Journal of Physical Chemistry A*, 2000, **104**, 1720–1735.

43 V. K. K. Praneeth, C. Näther, G. Peters and N. Lehnert, *Inorganic Chemistry*, 2006, **45**, 2795–2811.

44 A. Ghosh, *Accounts of Chemical Research*, 2005, **38**, 943–954.

45 M. Ibrahim, C. Xu and T. G. Spiro, *Journal of the American Chemical Society*, 2006, **128**, 16834–16845.

46 M. S. Liao, M. J. Huang and J. D. Watts, *Journal of Physical Chemistry B*, 2013, **117**, 10103–10114.

47 L. M. Blomberg, M. R. A. Blomberg and P. E. M. Siegbahn, *Journal of Inorganic Biochemistry*, 2005, **99**, 949–958.

48 M. Feldt, Q. M. Phung, K. Pierloot, R. A. Mata and J. N. Harvey, *Journal of Chemical Theory and Computation*, 2019, **15**, 922–937.

49 D. G. Liakos, Y. Guo and F. Neese, *Journal of Physical Chemistry A*, 2020, **124**, 90–100.

50 L. Pauling and C. D. Coryell, *Proceedings of the National Academy of Sciences*, 1936, **22**, 210–216.

51 M. Radoń, *J. Chem. Theory Comput.*, 2014, **10**, 2306–2321.

52 N. Strickland and J. N. Harvey, *The journal of physical chemistry. B*, 2007, **111**, 841–52.

53 A. Lábas, D. K. Menyhárd, J. N. Harvey and J. Oláh, *Chemistry - A European Journal*, 2018, **24**, 5350–5358.

54 C. Xu and T. G. Spiro, *Journal of Biological Inorganic Chemistry*, 2008, **13**, 613–621.

55 D. M. A. Smith, M. Dupuis, E. R. Vorpagel and T. P. Straatsma, *Journal of the American Chemical Society*, 2003, **125**, 2711–2717.

56 C. Rovira, K. Kunc, J. Hutter, P. Ballone and M. Parrinello, *Journal of Physical Chemistry A*, 1997, **101**, 8914–8925.

57 C. N. Hall and J. Garthwaite, *Nitric Oxide - Biology and Chemistry*, 2009, **21**, 92–103.

58 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J Comput Chem*, 2004, **25**, 1605–1612.

59 B. K. Yoo, I. Lamarre, J. L. Martin, F. Rappaport, M. Negrerie and W. A. Eaton, *Proceedings of the National Academy of Sciences of the United States of America*, 2015, **112**, E1697–E1704.

60 Rahul Banerjee and John D. Lipscomb, *Accounts of Chemical Research*, 2021, **54**, 2185–2195.

61 Y. Guo, D. L. M. Suess, M. A. Herzik, A. T. Iavarone, R. D. Britt and M. A. Marletta, *Nature Chemical Biology*, 2017, **13**, 1216–1221.
62 D. J. Stuehr, S. Misra, Y. Dai and A. Ghosh, *Journal of Biological Chemistry*, 2021, **296**, 100336.
63 B. Roy, E. J. Halvey and J. Garthwaite, *Journal of Biological Chemistry*, 2008, **283**, 18841–18851.
64 C. Farah, L. Y. M. Michel and J. L. Balligand, *Nature Reviews Cardiology*, 2018, **15**, 292–316.
65 R. Radi, *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**, 4003–4008.
66 W. R. Montfort, *Journal of Molecular Biology*, 2024, **436**, 168235.

4 Inhibition of the Transsulfuration Pathway: Cystathionine β-synthase[X]

As outlined in Chapter 1, hCBS may bind to either CO or NO, causing complete inhibition of the enzyme's activity. Various experimental studies have investigated the binding of both gas molecules. The aim of this chapter is to first investigate the spectroscopically proposed models for CO and NO binding using high-level QM calculations, and then to investigate the conformation changes caused by CO or NO binding by way of MD simulations, ultimately improving understanding of the inhibition.

4.1 Introduction

Upon CO binding, UV-Vis. spectroscopy shows a blue shift of the Soret absorption maximum from 450 nm to 422 nm.[1,2] Similarly, NO binding causes a larger UV-Vis. blue shift from 450 nm to ~390 nm.[3–5] Compared to those reported for CO binding, the equilibrium dissociation constants for NO binding were initially predicted to be rather a lot higher than for CO binding.[3,4] However, subsequently, these values were suggested to be artificially high due to using a relatively high pH of 8.6 and excess dithionite.[4] Note, however, that the effect of pH on ligand binding in hCBS has been previously investigated and claimed to be insignificant within the range of 7.9 – 10.5, albeit with respect to CO binding.[2] The effect is therefore

---

[X] This chapter is based on the following publication: N. R. McFarlane, J. Gui, J. Oláh, J. N. Harvey, *Phys. Chem. Chem. Phys.*, **2024**, 26, 16579-16588. For the purposes of this thesis, sections of the publication have been slightly modified. Specifically, sections of the introduction have been removed as they have been instead included in the Chapter 1.

The calculations and simulations were performed by N. R. McFarlane and J. Gui, a Master's student who worked under N. R. McFarlane's supervision. Results were interpreted collectively by all co-authors. The first draft of the paper was entirely written by N. R. McFarlane, with J. N. Harvey and J. Oláh editing it and contributing to the final version.

more likely to do with the excess dithionite used in the initial study.[3] More recent refinement shows that the equilibrium dissociation constants for both gases are similar in magnitude, at $K_D = 0.8\,\mu M$ and $K_D = 0.23\,\mu M$ for CO and NO, respectively.[5] In this study, the authors use authentic NO gas, rather than an NO releaser, such as DEA-NONOate as was used in previous studies.[3,4] The authors do note, however, that this K_D may be slightly artificially low, although certainly more in the expected range as compared to previous studies given the known physiological role of NO.[5] All reported equilibrium dissociation constants for hCBS are summarised in Table 4.1.

Table 4.1: *Chronological summary of all equilibrium dissociation constants reported in the literature for CO and NO binding to hCBS.*

Equilibrium Dissociation Constant, K_D (μM)		**Reference**
CO	*NO*	
1.5 +/- 0.1	-	1
-	281 +/- 50	3
3.9 +/- 2.0	-	2
-	30 +/- 5	4
-	0.23	5
0.8	-	5

Relative to another heme signalling protein which binds both CO and NO, sGC (see Chapter 3), these equilibrium dissociation constants are slightly lower for CO ($K_D = 260\,\mu M$ for sGC), and many orders of magnitude higher for NO ($K_D = 53\,nM - 850\,pM$ for sGC), but this is expected given the known fast and strong binding of NO to sGC.[6] The dissociation constants reported in Table 4.1 do not take into account the fact that hCBS is a homodimer, with binding to the heme group in one chain potentially being able to affect the binding strength in the other chain. While there is evidence of such allosteric effects for bonding of CO, with different K_{D1} and K_{D2} being reported,[1,2,5] these effects have not been modelled in this work. In any case, whether

NO or CO binds to hCBS, the heme group is quickly oxidised back to the ferric state by O_2, resulting in total recovery of enzymatic activity.[7]

As shown by EPR spectroscopy, CO binding is associated with decoordinatation of proximal cysteine in a two-step fashion (**A→D→E**), where the initial loss of cysteine (**A→D**) is the rate-limiting step (see Figure 4.1). The final inhibitory state, **E**, is found to be hexa-coordinate with proximal CO and distal histidine.[2,8] Also based on EPR spectroscopy, the NO binding pathway has been shown to occur in two steps (**A→B→C**), starting with concerted displacement of the distal histidine by incoming NO (**A→B**), followed by (again rate-limiting) decoordination of proximal cysteine (**B→C**). In this case, the final inhibitory state, **C**, is determined to be penta-coordinate, with a proximal NO, and both amino acids decoordinated (see Figure 4.1).[2,3,8] Given the mechanism for CO binding, it is feasible that **D** could bind to NO rather than CO equivalent, forming a different hexa-coordinate heme group with proximal NO and distal histidine (**F**) (see Figure 4.1).

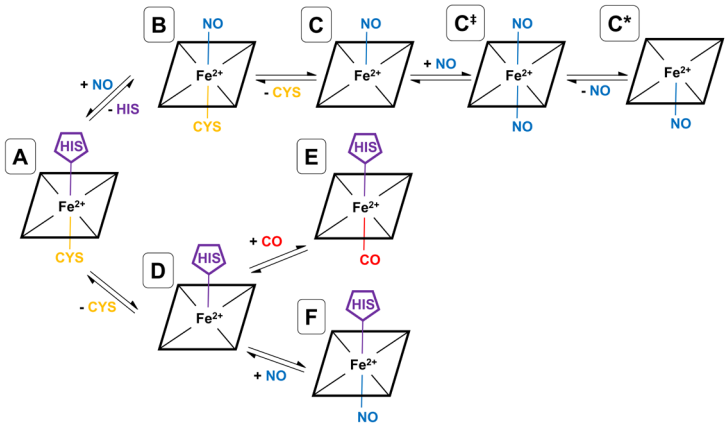

Figure 4.1: *Gaseous inhibitory pathways shown in hCBS. Species **A, B, C, D,** and **E** have been experimentally observed, and **C** and **E** are the final inhibitory states for NO and CO, respectively. Species **C‡, C*** and **F** are not experimentally observed, but are computationally and biochemically suggested in this work.*

As discussed in Chapter 3, NO binds to sGC, and in that case, the NO-bound form is also suggested to be pentacoordinate, with NO binding being accompanied by histidine sidechain decoordination. There is also some evidence that the NO-bound form involves a proximal NO ligand, i.e., bound on the same side as the departing histidine sidechain. In Chapter 3,[9] it was suggested that this proximal NO state could be formed by reaction of a second NO equivalent with the initially formed penta-coordinate Fe(II)-NO system with a distal NO, to form a transient hexa-coordinate NO-Fe(II)-NO system (**C‡**) that would then lose the first NO to yield the isomeric penta-coordinate NO-Fe(II) (**C***) (see Figure 4.1). The proximal NO binding (whose formation might be plausible during a short period of elevated NO partial pressure) could be associated with slower return to the initial state, as under low NO conditions, access to hexa-coordinate **C‡** would be difficult. It is possible that a similar mechanism occurs also for hCBS, where experiment does not allow a conclusive assignment of the bound NO state to a **C** or **C*** form. If **C*** is indeed formed, this could explain the reported lower rate of NO dissociation (~0.003 s^{-1}) compared to that of CO (~0.5 s^{-1}) in hCBS, as has already been suggested.[5]

Both CO or NO binding to the heme group completely inhibit hCBS, and it has been shown that this is due to conformational changes in the PLP-dependent active site upon gas binding to the heme site, ultimately leading to mechanical signalling between the two sites. However, the detailed way in which this signalling occurs and the way in which it causes the inhibition are not yet fully understood. The environment of the heme group and how it is structurally related to the PLP environment is shown in Figure 4.2.

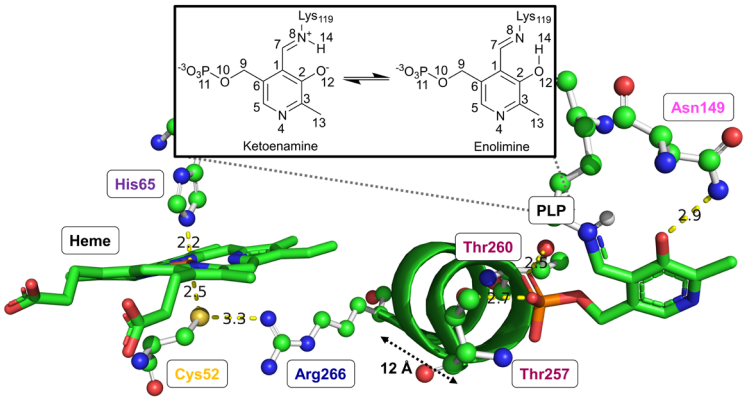

***Figure 4.2**: Environments of the heme group and active PLP site. All relevant amino acids are shown, alongside the α-helix 8 which links the two sites. α-helix 8 is approximately 12 Å in length. For simplicity, hydrogen atoms are omitted, so intermolecular interactions (shown by dotted lines) have the positions of hydrogen atoms implied. The figure inset shows the tautomerisation of the PLP active site between the active ketoenamine form and inactive enolimine form. The figure was based on pdb-id:4COO[10] and generated using PyMOL.[11]*

The proximal Cys52 forms a salt bridge with Arg266, where Arg266 is at one end of a short α-helix 8, and at the other end, Thr260 and Thr257 exhibit strong hydrogen bonding with the phosphate group of PLP.[10,12–14] Mutation of both Arg266 and Thr260/Thr257 has been shown to reduce enzymatic activity.[13,15–17] As well as the suggested mechanical signalling between the heme group and PLP cofactor, a form of electronic coupling between the heme group, α-helix 8 and the PLP cofactor has been suggested[18] to play a role in this inhibition, but the nature of this coupling is unclear. Bhatt et al. indeed show very large changes in electron density distribution when performing comparative density functional theory calculations on the heme group, α-helix 8 and the PLP cofactor as separate fragments and in structures as found in hCBS, but these large changes seem to be due to artefacts associated with the gas-phase nature of the DFT calculations performed, which involve a fragment with large negative charge (the PLP cofactor) being brought into close proximity of the model for α-helix 8. Given these

possible errors, the focus here will be on the mechanical signalling model.

In the mechanical signalling model,[19] the active ketoenamine form of PLP and its protonated Schiff base nitrogen is stabilised by hydrogen bonding between Asn149 and the pyridoxine ring oxygen atom, as shown in Figure 4.2. The protonation of the Schiff base nitrogen atom makes the imine carbon atom more susceptible to nucleophilic attack.[19–21] In case the PLP centre undergoes tautomerisation, an inactive enolimine form with a neutral imine nitrogen is obtained.[19] Fluorescence spectroscopy experiments show that upon binding of CO/NO, the PLP cofactor changes character from the ketoenamine to enolimine form.[19] This was suggested to be due to α-helix 8 motion caused by breaking of the Cys52–Arg266 salt bridge, with the strong hydrogen bonds between Thr257/Thr260 and the phosphate group of PLP leading to movement of the PLP group that cause disruption of the Asn149–O^-(PLP) hydrogen bond and hence the shift of the tautomerisation equilibrium towards the inactive enolimine form. The experimental technique does not however provide access to detailed atomistic insight into the structural changes in α-helix 8 and the PLP site.

The goal of this chapter is to investigate the structural changes occurring upon CO or NO binding to the heme group. To do so, the free energy landscape associated with CO and NO binding is first evaluated using quantum-chemical methods so as to obtain accurate microscopic models of the **C**, **E** and **F** inhibited forms of the heme group. Based on these models, extended MD simulations of the inhibited forms were performed to study the motion of α-helix 8 following breakage of the Cys52–Arg266 salt bridge, and what effect this has on the structure surrounding the PLP active site.

4.2 Computational Details

4.2.1 Quantum Mechanical Calculations[XI]

In all calculations, the heme ring was modelled as an Fe-porphin ring,[22,23] and cysteine and histidine were modelled as hydrosulfide and imidazole, respectively. To provide adiabatic spin-state energetics,[22] each of the heme derivatives shown in Figure 4.1 (excluding **C‡** and **C***, which is identical to **C** with the modelling approach used) were geometry optimised in their low-, medium-, and high-spin states using the hybrid B3LYP functional[24,25] with the TZVP/SVP basis set combination[26] (see Table 4.2).

Table 4.2: *The three basis set combinations used throughout the QM calculations. Basis set combination TZVP/SVP was used for all geometry optimisations, and basis set combinations TZVPP/TZVP/SVP and QZVPP/TZVP/SVP were used for basis set extrapolations on DLPNO-CCSD(T_1) calculations.*

Basis Set Combination	**Atom**					
	Fe	*N*	*O*	*S*	*C*	*H*
TZVP/SVP	Def2-TZVP	Def2-SVP	Def2-SVP	Def2-SVP	Def2-SVP	Def2-SVP
TZVPP/TZVP/SVP	ZORA-Def2-TZVPP	ZORA-Def2-TZVP	ZORA-Def2-TZVP	ZORA-Def2-TZVP	ZORA-Def2-TZVP	ZORA-Def2-SVP
QZVPP/TZVP/SVP	ZORA-Def2-QZVPP	ZORA-Def2-TZVP	ZORA-Def2-TZVP	ZORA-Def2-TZVP	ZORA-Def2-TZVP	ZORA-Def2-SVP

Grimme's D3 dispersion correction with the Becke-Johnson damping function was used for all optimisations.[27,28] Using second-derivative calculations, no imaginary frequencies were found, thus all heme derivatives were found to be stable minima on their corresponding potential energy surface. These optimisations and frequency analyses were performed with the Gaussian 16 program.[29]

[XI] All QM calculations were performed by N. R. McFarlane.

To identify the electronic ground states, single-point DLPNO-CCSD(T_1) calculations were performed on the optimised geometries of the heme derivatives using the ORCA quantum chemistry program package.[30] The coupled-cluster expansion was based on B3LYP Kohn-Sham orbitals[24,25] obtained using the TZVP/SVP basis set combination (see Table 4.2).[26] The density fitting approximation was used to compute Coulomb integrals, and numerical chain-of-sphere integration was used for the Hartree-Fock exchange integrals.[31]

As previously shown,[32] test calculations using Hartree-Fock orbitals for the coupled-cluster expansion showed very similar relative energies. Scalar relativistic effects were also considered through the zero-order regular approximation (ZORA).[33–35] The ZORA recontracted versions of the def2 basis sets were used within this approximation.[36] As auxiliary basis sets, the normal def2 family of basis sets and the segmented all-electron relativistically contracted (SARC) were used.[26,36]

On the calculated electronic ground state for each species, a higher-level DLPNO-CCSD(T_1) calculation was performed, where two-point complete basis set limit extrapolation[37] was applied with respect to the heme iron using two basis set combinations, TZVPP/TZVP/SVP & QZVPP/TZVP/SVP (see Table 4.2).[26] The CBS limit extrapolations for the electronic and correlation energy were calculated using Equation (4.1) and Equation (4.2), respetively, and the total CBS limit extrapolated DLPNO-CCSD(T_1) electronic energy was calculated with Equation (4.3). Here, there are two constants, $a = 7.88$ and $b = 2.97$, and $X = 3$ and $Y = 4$ represent the cardinal numbers of the def2 basis set hierarchies.[37]

$$E_{elec.}^{CBS} = \frac{e^{a\cdot\sqrt{X}} \cdot E_{elec.}^{(X)} - e^{a\cdot\sqrt{Y}} \cdot E_{elec.}^{(Y)}}{e^{a\cdot\sqrt{X}} - e^{a\cdot\sqrt{Y}}} \quad \textbf{(4.1)}$$

$$E_{corr.}^{CBS} = \frac{X^b \cdot E_{corr.}^{(X)} - Y^b \cdot E_{corr.}^{(Y)}}{X^b - Y^b} \quad \textbf{(4.2)}$$

$$E_{DLPNO-CCSD(T1)}^{CBS} = E_{elec.}^{CBS} + E_{corr.}^{CBS} \quad \textbf{(4.3)}$$

All other parameters for the DLPNO-CCSD(T_1) calculations remained the same as already outlined above in the context of the ground electronic spin state determination.

To evaluate the Gibbs free energy changes associated with the steps of the reaction pathways, two terms were added to the complete basis set extrapolated DLPNO-CCSD(T_1) electronic energy. The first is the statistical mechanics Gibbs free energy correction for each of the species at standard state (1M) taken from a frequency analysis of a hybrid B3LYP-D3 calculation performed using the TZVPP/TZVP/SVP basis set combination. The translational contribution to the Gibbs free energy was removed for decoordinated amino acid models, hydrosulfide and imidazole, since these remain bound to the protein even upon dissociation from the heme group. The second is the Gibbs free energy of solvation for each of the species, which was calculated by the difference in hybrid B3LYP-D3 electronic energy between a vacuum and implicitly solvated model using a CPCM implicit solvent model with a protein-representative dielectric constant of $\varepsilon = 4$.

In addition to the B3LYP and DLPNO-CCSD(T_1) calculations, a series of single-point energy calculations on a selection DFT functionals were performed on the electronic ground state of each species using the ORCA quantum chemistry package.[30] Among the

DFT calculations were 4 GGA methods (BP86,[38,39] BLYP,[25] OLYP,[40] & PBE[39]), 1 meta-GGA method (TPSS[41]), 3 hybrid-GGA functionals (B3PW91,[24] B3LYP*,[42] & PBE0[43,44]), and 1 hybrid meta-GGA functional (TPSSh[41,45]). As with the DLPNO-CCSD(T_1) calculations, scalar relativistic effects were considered through ZORA[33–35], the ZORA-recontracted versions of the def2 basis sets were used (using the TZVPP/TZVP/SVP basis set combination – see Table 4.2).[36] As auxiliary basis sets, the normal def2 family of basis sets and the segmented all-electron relativistically contracted (SARC) were used.[26,36] Grimme's D3 dispersion correction with the Becke-Johnson damping function was applied in all cases.[27,28] For pure DFT calculations, the density fitting approximation was used to solve Coulomb integrals.[46–48] For hybrid functionals, the density fitting approximation was supplemented with the chain-of-sphere integration method to account for the Hartree-Fock exchange integrals.[31,49,50] The Gibbs free energy correction was calculated with the same method as outlined above.

4.2.2 Molecular Dynamics Simulations[XII]

There exist 8 crystal structures of hCBS in the basal state (pdb-id: 1JBQ, 1M54, 4COO, 4L0D, 4L3V, 4L27, 4L28, 5MMS).[10,12,51–53] On the basis of having the best resolution (2.00 Å), being the wild-type form of hCBS, and having all residues modelled in the relevant domains, 4COO was used as the starting point in this work. For a detailed comparison of all available crystal structures for hCBS, see Table 4.3.

[XII] The initial MD simulations were performed and interpreted by Master's student J. Gui under N. R. McFarlane's supervision. Further MD simulations and interpretation was performed by N. R. McFarlane.

Table 4.3: *Summary of important parameters relating to the available crystal structures for hCBS. In order, the columns are the pdb-id, the resolution (in Angstroms), the R-Value Free, any comments relating to the choice of crystal structure in this work, and the relevant reference. Shown in bold are the details of the crystal structure used in this work.*

pdb-id	*Resolution (Å)*	*R-Value Free*	*Comments*	*Ref.*
1JBQ	2.60	0.296	Wild-type, reduced model (only heme and catalytic domain), missing residues in catalytic domain	12
1M54	2.90	0.365	Mutated, reduced model (only heme and catalytic domain), missing residues in catalytic domain	51
4COO	**2.00**	**0.201**	**Wild-type, all four domains, only some missing residue atoms in catalytic domain**	10
4L0D	2.97	0.284	Mutated, all four domains, only some missing residue atoms in catalytic domain	53
4L3V	3.63	0.232	Wild-type, all four domains, no missing atoms or residues in catalytic domain	53
4L27	3.39	0.244	Mutated, all four domains, no missing atoms or residues in catalytic domain	53
4L28	2.63	0.281	Mutated, all four domains, missing residues in catalytic domain	53
5MMS	2.80	0.221	Mutated, reduced model, missing residues in catalytic domain	52

hCBS is a biological homodimer and was modelled as such. 4COO contains all three domains of hCBS: the N-terminal heme domain, the central catalytic domain, and the C-terminal regulatory

domain.[10] Since the C-terminal regulatory domain is not the main focus of this work and its inclusion would surely increase computational expense, it was excluded from modelling – a previously successful methodology for catalytic domain focused modelling work.[54] There are 64 missing residues at the N-termini for both chains which were not added due to uncertainty in their secondary structures. With these considerations in mind, both chains, residues 43 – 381, were used for modelling throughout, giving a total of 676 residues. Within this selection, there are a few residues with missing atoms – these were added back using the SCWRL program.[55]

The parameters for the covalently-bound PLP cofactor were generated using the CGenFF web app,[56] and deemed to be satisfactory for the purposes of this work. To include the heme group in the system topology, the native support within CHARMM36 was used.[57] Where relevant, the distal His65 and proximal Cys52 were bonded to the heme group by inclusion of a specbond.dat file. All other atoms also used the native support within CHARMM36.[57] The protonation states of all titratable amino acid side-chains were assigned based on their pKa predicted by PropKa[58] and set accordingly within the system topology.

Prior to production MD, equilibration was performed for species **A** (see Figure 4.1), using the TIP3P water model,[59] and periodic boundary conditions with a rhombic dodecahedron unit cell, where the protein was at least 20 Å from the box edge. The total charge of the system was neutralised and set to a realistic physiological concentration of NaCl (0.15 M).[60] To remove any initial bad contacts, an energy minimisation was performed and NVT and NPT equilibrations were performed for 1000 ps each, using the velocity-rescaling thermostat[61] and Parrinello-Rahman barostat,[62] thus equilibrating the system at 300 K and 1 bar. Lastly, positional restraints on all atoms were sequentially relaxed, and two 1000 ns simulations were performed using different starting

velocities in the NPT ensemble with Parrinello-Rahman pressure coupling.[62]

In order to simulate a given inhibitory pathway (**A→C**, **A→F**, and **A→E** – see Figure 4.1), the final geometry from the trajectory of **A** was extracted and structurally modified to represent the relevant bonding for **C**, **E**, or **F**, and the specbond.dat file was updated appropriately. Any CO or NO molecules were introduced appropriately by making slight structural modifications to the amino acids they displace to avoid any severe steric clashes and were bonded to the heme group using the specbond.dat file. These modifications were performed for both monomers, so for each trajectory, there are two heme and PLP groups to be modified. In all cases, the proximal Cys52 was protonated upon decoordination from the heme iron. An energy minimisation was used to mitigate any bad contacts caused by the crude structural modifications. Following this, NVT and NPT equilibrations were performed for 1000 ps each, using the velocity-rescaling thermostat[61] and Parrinello-Rahman barostat.[62] Once each of the systems were sufficiently equilibrated, all positional restraints were sequentially relaxed, and simulations for each inhibitory state in the NPT ensemble with Parrinello-Rahman pressure coupling[62] were performed. For each pathway, two 1000 ns simulations were performed using different starting velocities. All MD simulations were performed with GROMACS.[63,64]

4.3 Results and Discussion

4.3.1 Quantum Mechanical Calculations

4.3.1.1 Heme Derivatives Ground Spin States

Many of the heme derivatives which have been studied in this work have previously been investigated in detail, in some cases with more accurate methods as compared to DLPNO-CCSD(T_1). Nevertheless, for the sake of having a consistent dataset, and to identify the ground spin states of species **A** and **B** (which have not

been studied previously), the ground spin states of all species **A** – **F** have been studied here (see Table B2 for data).

In good agreement with previous *ab initio* and DFT studies,[65–68] **F** has a clear doublet ground state, and this is also the case for **C**, again in agreement with previous work.[69] As already mentioned, there are no computational studies of species **B**, though the corresponding ferric species has been studied, and found to have a clear low-spin form (in that case, a singlet).[70] This suggests that **B** should be low-spin (i.e. a doublet), and this is consistent with this work, with a large spin state gap to the closest-lying spin state (sextet) of 20.5 kcal/mol. A stable singlet ground state is found for species **E**, in positive agreement with numerous previous computational studies.[65,68,71]

Moving now to the gas-free heme adducts, it is well established from previous experimental[72] and computational[22] work that **D** has a quintet ground state with a close-lying triplet, and this is also found here, albeit with a larger quintet-triplet gap than in some previous studies,[22] perhaps due to the use of local coupled-cluster with a relatively small basis set. For species **A** there is seemingly no previous experimental or computational work on the electronic ground state. The ground spin state of **A** is found to be a singlet herein, albeit with a relatively small spin state gap to the triplet of 4.2 kcal/mol.

4.3.1.2 Gibbs Free Energy of Reactions in hCBS

Having established the ground spin states of the various heme adducts, the Gibbs free energy changes associated with the inhibitory pathways calculated using the ground spin states for the electronic structure calculations will be compared to experimental observations. As described in Section 4.2.1, the free energies calculated herein are composed of complete basis set limit extrapolated DLPNO-CCSD(T_1) electronic energies and a Gibbs free energy correction. Before discussing the results, there are several possible sources of error in these calculations. First,

DLPNO-CCSD(T_1) approach used is not exact,[73] especially for transition metal complexes.[74] No small effort has been made to minimise these errors by including the improved iterative T_1 method as well as complete basis set extrapolation with respect to heme iron.[75] Pantazis et al. recommend the use of complete PNO space extrapolation to approach the limit of fully canonical calculations,[75] however this was computationally unfeasible in the present study, so some uncertainty remains.

As well as the inherent uncertainties coming from the level of electronic structure theory used, the approach used here is limited in accuracy due to three aspects. First, dissociation of histidine and/or cysteine sidechains from the heme group does not involve formation of a separate molecule, since the sidechains remain part of the whole protein-heme system. Hence, within the ideal-gas-like treatment of statistical mechanics, it is not clear how to treat the dissociated species, as this will not have a large relative translational partition function. In the calculations herein, somewhat arbitrarily, these amino acids are treated as having no translational freedom. The rotational and vibrational parts of the partition function are computed using the usual rigid-rotor/harmonic oscillator approaches. Since the dissociated sidechain does have some range of possible motion, this may underestimate the entropy (and hence overestimate the Gibbs energy) of the dissociated states, but this may be partly compensated for by including the full rotational partition function (the sidechain will in reality have some degree of constraint on its rotational motion).

A second approximation involves the use of a continuum solvent with a generic dielectric permittivity of 4 to describe the medium surrounding the heme model. The initial species **A** is an anion, but dissociation of cysteine leads to neutral species, so solvation does affect the calculated Gibbs energies to some extent, and the polarity of the medium may not be well captured by a single

dielectric constant. Varying the dielectric constant to that of water does not impact the calculated relative Gibbs energies in a major way (test calculations show that species **C** moves from -0.4 to -2.9 kcal/mol, with the difference being of the order of the estimated error in the quantum-chemical method used), but this approach does still introduce some additional error in the estimated Gibbs energies. It is worth highlighting that the departing cysteine sidechain is treated as anionic HS^- in the calculations performed here. The cysteine sidechain has a pK_a near 7, so that when using a standard state for the proton concentration corresponding to pH 7 (i.e., using $\Delta G°'$ rather than $\Delta G°$, see reference 76) the standard Gibbs energies of the anionic thiolate and neutral thiol forms are almost equal, and so the solvation energies of the two forms should be equivalent. As an additional test for the accuracy of this assumption and of the solvation treatment, the pK_a of H_2S was computed at the same level of theory used here, obtaining a value of 4.7, reasonably close to the experimental value of 7. Cysteine residue have pK_a values ranging from 3.5 to 12 depending on the protein environment and reactivity,[77] so such a prediction is in reasonable qualitative agreement.

A third aspect relates to the conformational rearrangement of the overall protein structure, α-helix 8, and the PLP cofactor following CO or NO coordination, that will be discussed below. This rearrangement is spontaneous and will hence lead to a favourable contribution to the reaction free energy. As the inhibition by CO and NO is observed to be total, the free energy gain associated with relaxation of the structure after CO or NO coordination must be significant, since otherwise a small portion of the protein would remain in its active form. Assuming inhibition by a factor of 100, the Gibbs energy change should be $-RT \ln 100$ or 2.7 kcal/mol, but the precise value is unclear. Of course, relaxation cannot be too favourable, because otherwise the protein would not be stable in the active form, even in the absence of CO or NO binding.

These provisos noted, the calculations performed in this work will now be compared to experiment. The pathway leading to inhibition by CO is believed to follow the route **A→D→E**. According to our calculations, the first step is endergonic in Gibbs energy terms (+11.2 kcal/mol), and the second step is exergonic (-15.1 kcal/mol) (see Figure 4.3). While these calculations do not address energy barriers, as previously argued,[66] gaseous ligand binding, as in the second step, can be expected to be quite fast so will have a low barrier. Decoordination of cysteine (with concomitant protonation) in the first step will likely have a larger barrier, consistent with the slow observed binding of CO to ferrous hCBS.[8] The net Gibbs free energy change from **A** to **E** is calculated to be −3.9 kcal/mol. Given the many sources of uncertainty mentioned above, this is consistent with the observed binding affinity and K_D = 0.8 μM, which implies $\Delta G°'$ = −8.3 kcal/mol.

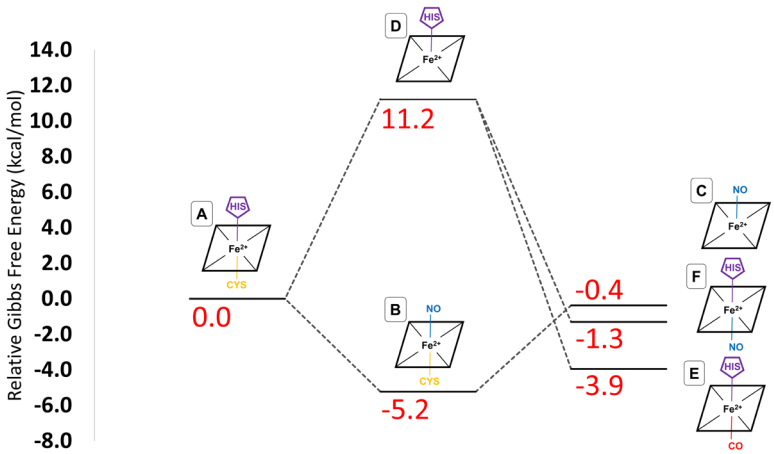

Figure 4.3: *Gibbs Free energy changes associated with the inhibitory pathways caused by binding of CO/NO to hCBS. Relative free energies are calculated relative to the basal species,* ***A****. All values are given in kcal/mol.*

Experimentally, NO has been observed to inhibit hCBS by binding through the pathway, **A→B→C**. The alternative pathway *via* intermediate **D** as for CO is presumably much less favourable, given the difference in the observed kinetics. The first step **A** -> **B** is

calculated to be exergonic (-5.2 kcal/mol) and will have a barrier which has not been studied here. Presumably, given the observed fast binding of NO, this barrier is relatively low, but the reason for the difference with CO, for which this pathway is not observed, is less clear.

The second step is predicted as endergonic (+4.8 kcal/mol), leading to an overall predicted binding free energy of −0.4 kcal/mol, which differs by 8.7 kcal/mol from the value indicated by the experimental dissociation constant of K_D = 0.23 μM, which means that $\Delta G°'$ = −9.1 kcal/mol. Formation of **C** requires dissociation of both the cysteine and histidine sidechains, so is twice subject to the uncertainty mentioned above relating to the translational entropy of the dissociated ligand. In case some portion of the translational entropy does still contribute to the dissociation Gibbs energy, this in addition to the possible inaccuracies of local CC may account for some of the mismatch between the calculated and experimental $\Delta G°'$. In addition to the possible errors in the electronic structure method used, using a QM cluster model may be less suitable here as the conformational change following dissociation of cysteine from heme may also bring the calculated value closer to the experimental $\Delta G°'$.

The suggested endpoint of NO binding based on experiment is species **C**, while the *a priori* plausible **F** species not experimentally observed. The calculations here suggest that **F** would be more stable (see Figure 4.3), but given the possible sources of error, and the small difference in computed Gibbs energies, this is not a major inconsistency.

In addition to the DLPNO-CCSD(T_1)/B3LYP Gibbs free energy changes associated with the steps of the reaction pathways, the same free energy changes have been evaluated using a selection of DFT functionals (see Table B1*)*. Generally speaking, the hybrid functionals perform better for the various reaction pathways. Similar trends were observed in Chapter 3 in the discussion of

sGC. However, in this case, most functionals perform qualitatively reasonably well, suggesting that for this system, the choice of DFT functional is less imperative.

To conclude this section, the QM calculations are consistent with the experimental observations and provide a solid basis for assigning the bonding pattern to be used in the MD simulations addressing the conformational changes in structure upon NO or CO coordination, which will be discussed next.

4.3.2 Molecular Dynamics Simulations

The primary goal of these MD simulations is to understand the changes in structure around the heme group, α-helix 8 and the PLP cofactor upon gas binding to the heme group. There is no experimental structural data on this inhibited form of hCBS is available. Based on the atomistic structures for the heme group in the inhibited form, together with extensive MD simulation, exploratory studies can be performed to predict the nature of the structural changes. Note that as hCBS is a homodimer, all simulations provide duplicate structures for the two chains. In principle, some coupling between motion in the two monomers could occur, and could account for the mentioned difference in affinity between the first and second CO binding affinity,[1,2,5] but such allosteric effects have not been modelled here.

To confirm the quality of the CGenFF parameters for PLP generated herein with respect to their validity for studying the inhibitory pathways **A→C**, **A→E**, and **A→F**, simulations of the resting state, **A**, were performed first. The core of the mechanical signalling hypothesis is that when Cys52 decoordinates from the heme group, it is quickly protonated, and the strong salt bridge between Cys52 and Arg266 (see Figure 4.2) would be broken as a direct consequence. Rupture of this salt bridge may cause some mechanically induced signalling pathway to PLP *via* α-helix 8 and enable the tautomerisation to the enolimine form of PLP, thus

inhibiting the enzyme's activity. For this hypothesis to be supported, the salt bridge must of course be stable in the basal state **A**, so two separate simulations (1000 ns each) were performed for this state, with the Cys52 still coordinated to heme. These trajectories indeed show stable structures, as evidenced by a flat overall root mean square deviation (RMSD) for the α-carbon backbone (Figure B1, a & b), an intact Cys52-Arg266 hydrogen bond throughout the simulation (Figure B3), and a largely unbroken hydrogen-bonding network around the PLP cofactor. For this latter aspect, a criterion for a hydrogen bond will first be defined: the distance between the heavy atoms of the donor and acceptor must be < 3.5 Å, and the donor-hydrogen-acceptor angle must be > 100°.[78] Using this criterion, returning to the simulations of the basal state **A**, a stable Asn149–O^-(PLP) hydrogen bond is observed throughout the simulations, with the hydrogen bond being observed for an average of 84% of the simulation time (Figures B7, B8, B15, B16 and Table B4). There is some slight weakening occurring during one of the replica simulations for one of the homodimeric chains, where the hydrogen bond is observed for 73% of simulation time (Figures B7, B8, B16 and Table B4). Besides this key hydrogen bond, strong hydrogen bonds are expected to be observed between the PLP phosphate group and the residues at the N-terminus of α-helix 8, ultimately coupling the positioning of α-helix 8 to the PLP cofactor. For these hydrogen bonds, since there are a large number of possible donor-acceptor pairs complicating analysis, the above H-bond criterion is not used and instead only the donor-acceptor distances are used. However, the stability and minimal fluctuation of these donor-acceptor distances indicates that these hydrogen bonds are, as expected, very stable throughout the simulations (Figures B23, B24, B25, B26).

Turning now to the simulations of the inhibited forms **C** and **E** resulting from binding of NO and CO, respectively, as well as of the possible modified inhibited form **F**, in which NO binds to the heme

group but the histidine residue also remains bonded to the heme iron. Two 1000 ns simulations of each of the final inhibitory states (**C**, **E**, and **F**), representing the pathways **A→C**, **A→E**, and **A→F**, have been performed. From a MD standpoint, the topologies of **E** and **F** are very similar and are generated by modification of the topology of **A**, such that heme-Cys52 bond is broken, and a distal CO or NO (for **E** and **F**, respectively) molecule is introduced. While similar, the forcefield parameters are also slightly different for the hexa-coordinate heme groups. Again, from modification of a topology of **A**, generation of a topology for **C** requires that both the heme-Cys52 and heme-His65 bonds are broken, and that a distal NO molecule is introduced, resulting in a penta-coordinate heme group. For all simulations of **C**, **E**, and **F**, Cys52 has been protonated given the proximity to bulk solvent and the typical pK_a of cysteine sidechains. Moreover, the experimental study of the coupling between the heme group and the PLP cofactor suggests that cysteine protonation is an essential step in the inhibition pathway.[19]

As no structures are available for the inhibited form, it was not clear which structural changes compared to species **A** would be likely to be observed, nor on which time-scale they could be expected to occur. However, minor conformational change could certainly be expected to be visible on the microsecond timescale of the MD simulations. Such changes were analysed by monitoring a range of structural coordinates relating to α-helix 8, the heme group and the PLP cofactor for each of the species **C**, **E** and **F**, each of the chains in the homodimers, and each of the independent simulations.

In all simulations, the Cys52-Arg266 interaction is found to break almost immediately upon modifying the heme group topology. This is expected since the cysteine is protonated as part of the change in topology (see Figures B4, B5 & B6 for **C**, **E**, and **F**, respectively). The forcefield partial charge on the coordinated sulfur atom in

species **A** is −0.8 but is only −0.23 in the neutral cysteine form, leading to much weaker electrostatic interactions.

After the initial instantaneous increase in d($S\gamma_{Cys52}$-$N1\eta_{Arg266}$), slower and higher-amplitude variations are also observed, associated with the fact that the proximal Cys52 is part of a loop region of the secondary protein structure. Consequently, upon decoordination from heme, Cys52 (and His65 for the case of **C**) is free to move as its position is no longer defined by the coordination sphere of the heme group. On the other hand, the position of the heme group is in some way locked by interaction of the propionate side-chains with Arg51 and Arg224,[79] which are not part of the flexible loop region, so that despite the loss of the ligand, the heme group remains reasonably well anchored.

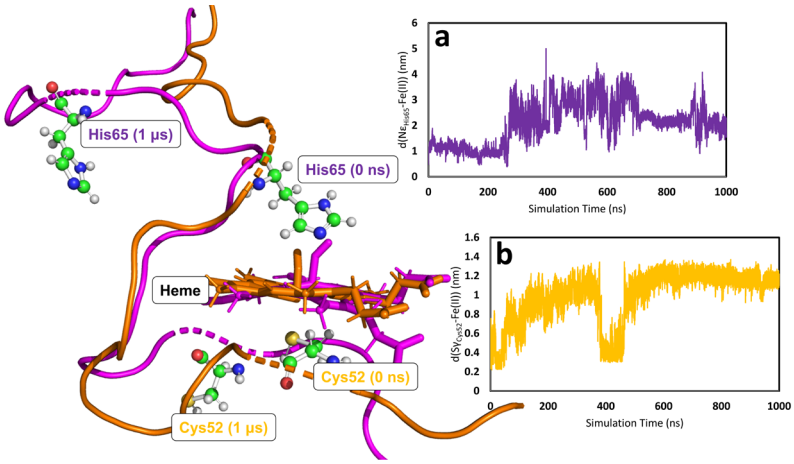

Figure 4.4: *Large-scale motions of Cys52 and His65 during one of the 1000 ns simulations for species **C** for chain B. The heme group and flexible loop region is shown in orange and purple for 0 and 1000 ns, respectively. Figure insets **a** and **b** show the motion of His65 and Cys52, respectively, relative to their initial coordination with heme Fe(II).* The figures were generated using PyMOL.[11]

Motion of Cys52 following decoordination occurs in all simulations of **C**, **E**, and **F**, but the direction of translation of Cys52 is not systematic, presumably due to the flexibility of the loop region. Similar freedom of movement is seen for His65 for simulations of

C, again with no discernible pattern. Figure 4.4 shows the types of large-scale motions which are observed for both Cys52 and His65 following decoordinated from heme to form **C**.

None of the simulations of species **C**, **E** and **F** show rapid or major changes in the positions of the heme group, α-helix 8 or the PLP cofactor, as shown by the moderate overall RMSD values for the trajectories, with respect to their species **A**–derived reference points (Figures B1 & B2). Nevertheless, close inspection of these trajectories does show evidence for structural change consistent with the mechanical signalling hypothesis. The main metric showing this is the previously mentioned Asn149–O^-(PLP) hydrogen bond. Considering the three species **C**, **E** and **F**, the two chains of the homodimer, and the two independent MD simulations for each case, there are 12 time series to be analysed. Again, using the criterion defined above, in a way that is not seen for the simulations of **A**, in three of these time series, the Asn149–O^-(PLP) hydrogen bond is unambiguously disrupted in a major way, with it being observed for 38, 25, and 13% of simulation times (Figures B9, B10, B13, B14, B18, B21, B22 and Table B4). In a further three time series, the hydrogen bond shows some weakening in a similar fashion to that which was observed in one of the replica simulations of **A**, where the hydrogen bond is observed for 75, 71, and 79% of simulation times (Figures B9, B10, B13, B14, B17, B18, B21 and Table B4). The remaining six simulations do not show convincing evidence of hydrogen bond elongation. Since the traditional, non-reactive MD simulations performed here cannot account for the tautomeric change in the PLP cofactor, any weakening of the hydrogen bonding must be due to the mechanical signalling through α-helix 8. The lack of clear changes in hydrogen bonding in some of the time series may simply be due to insufficient sampling, evidenced by the fact that there is no clear pattern concerning the differences between species **C**, **E** and **F**. In fact, the species showing the least evidence for weakening of the interaction, **E** (Figures B11, B12, B19, B20 and

Table B4), is treated in an extremely similar way within the forcefield to the related species **F**, which shows clear elongation of the Asn149–O^-(PLP) hydrogen bond in some of the time series (Figures B13, B14), strengthening the suggestion that the differences between simulations are indeed due to insufficient sampling.

Other structural metrics relating to the PLP cofactor are stable throughout the simulation. For example, the strong hydrogen bonds between the PLP phosphate group and the α-helix 8 N-terminus are largely unbroken throughout (Figures B23 – B38); some of the changes in distance shown in these plots arise due to switching of hydrogen bonds between the different phosphate oxygens). The α-helix is thereby firmly attached to the PLP cofactor, and motion of the α-helix leads to the above-mentioned weakening of the interaction of the PLP cofactor with Asn149.

The elongation of the Asn149–O^-(PLP) hydrogen bond is associated with other slight structural changes relating to the position of α-helix 8 and the PLP cofactor. Some insight into these changes can be obtained by comparing a typical species **A** structure and a typical structure in which the hydrogen bond is surely broken (Figure 4.5). This comparison is somewhat artificial since the choice of the two structures is necessarily arbitrary, but it has the benefit of providing an overall picture of the changes arising with the mechanical coupling mechanism. In Figure 4.5, the two structures are aligned based on the positions of the α-carbons of all amino acids, so that their protein backbones are maximally close to each other. As can be seen, there is a clear change in the relative position of α-helix 8 and the PLP cofactor with respect to Asn149, confirming that indeed α-helix 8 does play a role in mechanically induced signalling between the heme group and PLP cofactor.

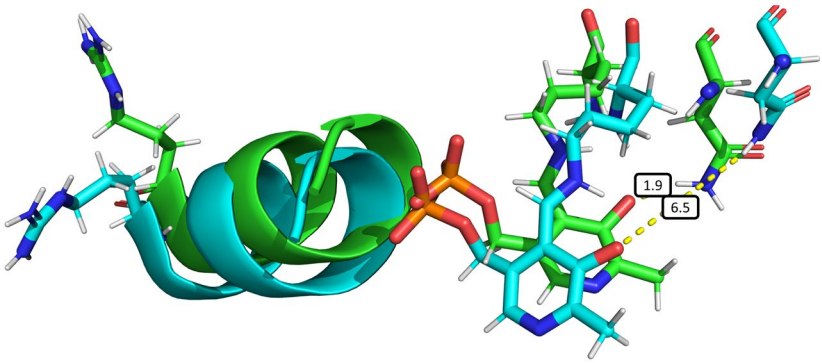

***Figure 4.5**: Overall structural changes following decoordination of Cys52 and rotation of Arg266. The initial structure at 0 ns is shown in green and a snapshot after 1000 ns is shown in blue (species **F**, chain A). The bond length d(Nδ_{Asn149}-O12$_{PLP}$) is highlighted in the figure, increasing from 1.9 Å (at 0 ns) and 6.5 Å (at 1000 ns). This figure was generated by α-carbon backbone alignment in PyMOL.*[11]

In principle, biased simulation techniques together with free energy evaluation methods could be used to quantify the differences arising when switching from the Cys-heme bound species **A** to the free Cys species **C**, **E** and **F**. This would however be rather challenging to model, given the extensive small structural differences between the various species, and the lack of a clearly identified reaction coordinate associated with the mechanical coupling. Hence this has not been attempted in this work.

4.4 Conclusions and Future Work

In this work, the role of a peripheral heme group found in hCBS that does not directly engage in catalytic activities, but instead exerts regulatory control over the enzyme has been investigated in detail. This regulatory function manifests when CO or NO gas molecules bind to the heme group, leading to complete inhibition of PLP-dependent reactions.

Initially, the CO/NO bonding models proposed by EPR spectroscopy were validated through detailed QM studies,

analysing the Gibbs free energy changes associated with the inhibitory pathways. Despite noted limitations in computational accuracy, the findings demonstrated positive agreement with previous calculations on related heme derivatives, often employing more sophisticated QM methods, and the calculated Gibbs energies are consistent with observed thermodynamics, giving confidence in the atomistic models for the basal and ligand-bound states.

Using forcefields with bonding patterns identified using the QM calculations, several 1000 ns molecular dynamics simulations on the various inhibitory species were performed. Through these simulations, the proposed mechanical signalling model was studied, which suggests that the rupture of the Cys52-Arg266 salt bridge, triggered by gas binding, initiates a cascade of conformational changes. It was found that motion of α-helix 8 (which Arg266 is a part of) which happens following breakage of the Cys52-Arg266 salt bridge leads to destabilisation of the active ketoenamine form of PLP. This destabilization arises from the elongation of a crucial hydrogen bond between Asn149 and a pyridoxine ring oxygen atom, suggesting a lowered barrier to tautomerization which produces the inactive enolimine form of PLP.

These findings underscore the intriguing regulatory role of CO and NO in hCBS, indicating that both gases have the potential to inhibit the enzyme both experimentally and theoretically. Moreover, reactions catalysed by hCBS at the PLP active site cause production of H_2S, yet further indicating the emerging theme of complex interplay between CO, NO and H_2S as signalling molecules.

One interesting aspect of hCBS which was briefly alluded to in Section 4.1 is that the equilibrium dissociation constants for CO binding to the two monomers are not equivalent, suggesting that some cooperativity exists between the monomers. The biological

relevance of this is that NO gas is often endogenously synthesized in pulses, as discussed in more detail in Chapter 3. If NO gas were also to bind in a cooperative fashion to hCBS, then there may be implications for the biochemical function of hCBS. In this work, while in the MD simulations both monomers were studied in the CO/NO bound forms simultaneously, the relationship between them was not studied in any detail.

Moreover, it is known that binding of AdoMet not only increases the activity of hCBS *via* increased accessibility of substrates to the active site, but also increases the binding affinity of CO/NO gas to the heme domain. These complex cooperative effects could, in principle, be studied by significantly longer or accelerated MD simulations and be used to understand the intricacies of the multiple domains of hCBS better.

4.5 References

1 S. Taoka, M. West and R. Banerjee, *Biochemistry*, 1999, **38**, 2738–2744.
2 M. Puranik, C. L. Weeks, D. Lahaye, Ö. Kabil, S. Taoka, S. B. Nielsen, J. T. Groves, R. Banerjee and T. G. Spiro, *Journal of Biological Chemistry*, 2006, **281**, 13433–13438.
3 S. Taoka and R. Banerjee, *Journal of Inorganic Biochemistry*, 2001, **87**, 245–251.
4 C. Gherasim, P. K. Yadav, O. Kabil, W.-N. Niu and R. Banerjee, *PLOS ONE*, 2014, **9**, e85544.
5 J. B. Vicente, H. G. Colaço, M. I. S. Mendes, P. Sarti, P. Leandro and A. Giuffrè, *Journal of Biological Chemistry*, 2014, **289**, 8579–8587.
6 G. Wu, I. Sharina and E. Martin, *Front Mol Biosci*, 2022, **9**, 1007768.
7 S. Carballal, P. Madzelan, C. F. Zinola, M. Graña, R. Radi, R. Banerjee and B. Alvarez, *Biochemistry*, 2008, **47**, 3194–3201.
8 S. Carballal, E. Cuevasanta, I. Marmisolle, O. Kabil, C. Gherasim, D. P. Ballou, R. Banerjee and B. Alvarez, *Biochemistry*, 2013, **52**, 4553–4562.
9 A. M. Rozza, M. Papp, N. R. McFarlane, J. N. Harvey and J. Oláh, *Chemistry – A European Journal*, 2022, **28**, e202200930.
10 T. J. McCorvie, J. Kopec, S.-J. Hyung, F. Fitzpatrick, X. Feng, D. Termine, C. Strain-Damerell, M. Vollmar, J. Fleming, J. M. Janz, C. Bulawa and W. W. Yue, *Journal of Biological Chemistry*, 2014, **289**, 36018–36030.
11 L. Schrödinger and W. DeLano, PyMOL (version 2.4.0) http://www.pymol.org/pymol 2020.
12 M. Meier, M. Janosik, V. Kery, J. P. Kraus and P. Burkhard, *The EMBO Journal*, 2001, **20**, 3910–3916.
13 P. K. Yadav, P. Xie and R. Banerjee, *Journal of Biological Chemistry*, 2012, **287**, 37611–37620.
14 R. Evande, S. Ojha and R. Banerjee, *Archives of Biochemistry and Biophysics*, 2004, **427**, 188–196.
15 S. Singh, P. Madzelan, J. Stasser, C. L. Weeks, D. Becker, T. G. Spiro, J. Penner-Hahn and R. Banerjee, *Journal of Inorganic Biochemistry*, 2009, **103**, 689–697.
16 A. T. Smith, Y. Su, D. J. Stevens, T. Majtan, J. P. Kraus and J. N. Burstyn, *Biochemistry*, 2012, **51**, 6360–6370.
17 A. Bhatt and Md. E. Ali, *Journal of Biomolecular Structure and Dynamics*, 2022, **40**, 12690–12698.
18 A. Bhatt, A. Mukhopadhyaya and Md. E. Ali, *J. Phys. Chem. B*, 2022, **126**, 4754–4760.
19 C. L. Weeks, S. Singh, P. Madzelan, R. Banerjee and T. G. Spiro, *J. Am. Chem. Soc.*, 2009, **131**, 12809–12816.
20 A. Salvà, J. Donoso, J. Frau and F. Muñoz, *J. Phys. Chem. A*, 2003, **107**, 9409–9414.
21 E. F. Oliveira, N. M. F. S. A. Cerqueira, P. A. Fernandes and M. J. Ramos, *J. Am. Chem. Soc.*, 2011, **133**, 15496–15505.
22 M. Radoń, *J. Chem. Theory Comput.*, 2014, **10**, 2306–2321.
23 Md. E. Ali, B. Sanyal and P. M. Oppeneer, *J. Phys. Chem. B*, 2012, **116**, 5849–5859.

24 A. D. Becke, *The Journal of Chemical Physics*, 1993, **98**, 5648–5652.
25 C. Lee, W. Yang and R. G. Parr, *Phys Rev B Condens Matter*, 1988, **37**, 785–789.
26 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
27 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *The Journal of Chemical Physics*, 2010, **132**, 154104.
28 S. Grimme, S. Ehrlich and L. Goerigk, *Journal of Computational Chemistry*, 2011, **32**, 1456–1465.
29 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, Gaussian16 Gaussian, Inc., Wallingford CT 2016.
30 F. Neese, F. Wennmohs, U. Becker and C. Riplinger, *The Journal of Chemical Physics*, 2020, **152**, 224108.
31 B. Helmich-Paris, B. de Souza, F. Neese and R. Izsák, *The Journal of Chemical Physics*, 2021, **155**, 104109.
32 Z. Benedek, P. Timár, T. Szilvási and G. Barcza, *J Comput Chem*, 2022, **43**, 2103–2120.
33 E. van Lenthe, J. G. Snijders and E. J. Baerends, *The Journal of Chemical Physics*, 1996, **105**, 6505–6516.
34 E. van Lenthe, E. J. Baerends and J. G. Snijders, *The Journal of Chemical Physics*, 1993, **99**, 4597–4610.
35 E. van Lenthe, E. J. Baerends and J. G. Snijders, *The Journal of Chemical Physics*, 1994, **101**, 9783–9792.
36 D. A. Pantazis, X.-Y. Chen, C. R. Landis and F. Neese, *J. Chem. Theory Comput.*, 2008, **4**, 908–919.
37 F. Neese and E. F. Valeev, *J. Chem. Theory Comput.*, 2011, **7**, 33–43.
38 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
39 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
40 N. C. HANDY and A. J. COHEN, *Molecular Physics*, 2001, **99**, 403–412.
41 J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401.
42 M. Reiher, O. Salomon and B. Artur Hess, *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 2001, **107**, 48–55.
43 J. P. Perdew, M. Ernzerhof and K. Burke, *The Journal of Chemical Physics*, 1996, **105**, 9982–9985.

44 C. Adamo and V. Barone, *The Journal of Chemical Physics*, 1999, **110**, 6158–6170.
45 V. N. Staroverov, G. E. Scuseria, J. Tao and J. P. Perdew, *The Journal of Chemical Physics*, 2003, **119**, 12129–12137.
46 J. L. Whitten, *The Journal of Chemical Physics*, 1973, **58**, 4496–4501.
47 O. Vahtras, J. Almlöf and M. W. Feyereisen, *Chemical Physics Letters*, 1993, **213**, 514–518.
48 B. I. Dunlap, J. W. D. Connolly and J. R. Sabin, *The Journal of Chemical Physics*, 1979, **71**, 3396–3402.
49 F. Neese, F. Wennmohs, A. Hansen and U. Becker, *Chemical Physics*, 2009, **356**, 98–109.
50 R. Izsák and F. Neese, *The Journal of Chemical Physics*, 2011, **135**, 144105.
51 S. Taoka, B. W. Lepore, Ö. Kabil, S. Ojha, D. Ringe and R. Banerjee, *Biochemistry*, 2002, **41**, 10454–10461.
52 J. B. Vicente, H. G. Colaço, F. Malagrinò, P. E. Santo, A. Gutierres, T. M. Bandeiras, P. Leandro, J. A. Brito and A. Giuffrè, *Oxidative Medicine and Cellular Longevity*, 2017, **2017**, e8940321.
53 J. Ereño-Orbea, T. Majtan, I. Oyenarte, J. P. Kraus and L. A. Martínez-Cruz, *Proceedings of the National Academy of Sciences*, 2013, **110**, E3790–E3799.
54 S. Gupta, S. Kelow, L. Wang, M. D. Andrake, R. L. Dunbrack and W. D. Kruger, *Journal of Biological Chemistry*, 2018, **293**, 13921–13931.
55 G. G. Krivov, M. V. Shapovalov and R. L. Dunbrack Jr., *Proteins: Structure, Function, and Bioinformatics*, 2009, **77**, 778–795.
56 K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. MacKerell, *J Comput Chem*, 2010, **31**, 671–690.
57 J. Huang and A. D. MacKerell, *J Comput Chem*, 2013, **34**, 2135–2145.
58 M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 525–537.
59 P. Mark and L. Nilsson, *J. Phys. Chem. A*, 2001, **105**, 9954–9960.
60 G. A. Ross, A. S. Rustenburg, P. B. Grinaway, J. Fass and J. D. Chodera, *J Phys Chem B*, 2018, **122**, 5466–5486.
61 G. Bussi, D. Donadio and M. Parrinello, *J Chem Phys*, 2007, **126**, 014101.
62 M. Parrinello and A. Rahman, *Journal of Applied Physics*, 1981, **52**, 7182–7190.
63 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1–2**, 19–25.
64 S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845–854.
65 N. Strickland and J. N. Harvey, *J. Phys. Chem. B*, 2007, **111**, 841–852.
66 A. Lábas, D. K. Menyhárd, J. N. Harvey and J. Oláh, *Chem. Eur. J.*, 2018, **24**, 5350–5358.
67 J. Oláh and J. N. Harvey, *J. Phys. Chem. A*, 2009, **113**, 7338–7345.
68 M. Radoń and K. Pierloot, *J. Phys. Chem. A*, 2008, **112**, 11824–11832.
69 C. Rovira, K. Kunc, J. Hutter, P. Ballone and M. Parrinello, *J. Phys. Chem. A*, 1997, **101**, 8914–8925.

70 D. A. Scherlis, C. B. Cymeryng and D. A. Estrin, *Inorg. Chem.*, 2000, **39**, 2352–2359.
71 S. Liu, S. Xia, D. Yue, H. Sun and H. Hirao, *Inorg. Chem.*, 2022, **61**, 17494–17504.
72 L. Pauling and C. D. Coryell, *Proceedings of the National Academy of Sciences*, 1936, **22**, 210–216.
73 D. G. Liakos, Y. Guo and F. Neese, *J. Phys. Chem. A*, 2020, **124**, 90–100.
74 M. Feldt, Q. M. Phung, K. Pierloot, R. A. Mata and J. N. Harvey, *J. Chem. Theory Comput.*, 2019, **15**, 922–937.
75 M. Drosou, C. A. Mitsopoulou and D. A. Pantazis, *J. Chem. Theory Comput.*, 2022, **18**, 3538–3548.
76 M. R. Gunner, T. Murakami, A. S. Rustenburg, M. Işık and J. D. Chodera, *J Comput Aided Mol Des*, 2020, **34**, 561–573.
77 D. W. Bak, T. J. Bechtel, J. A. Falco and E. Weerapana, *Curr Opin Chem Biol*, 2019, **48**, 96–105.
78 E. N. Baker and R. E. Hubbard, *Progress in Biophysics and Molecular Biology*, 1984, **44**, 97–179.
79 V. Guallar and B. Olsen, *Journal of Inorganic Biochemistry*, 2006, **100**, 755–760.

5 Development of a Hybrid QM/MM Growing String Method[XIII]

5.1 Introduction

Successfully describing biochemical reaction pathways and, crucially, how enzymes accelerate the rate of these reactions is central both from a fundamental point of view,[1,2] and for applications ranging from drug development[3,4] to industrial biocatalysis.[5,6] However, due to the short-lived nature of a transition state (TS), these pathways can be challenging to characterise experimentally. Allowing the modelling of reaction pathways and TSs, computational methodology is an attractive alternative which is widely used.[7–9]

In studying these reaction pathways, it is often desirable to include in the model not only the substrates and the catalytic site amino acid residues, but also the wider protein and solvent environment. Hybrid QM/MM is one of the commonly used methods that can satisfy this.[10–14] However, by its very nature, a hybrid QM/MM model invokes a significant number of DoF, and thus finding the MEP between reactant and product is seldom straightforward. In the QM/MM formalism, one may obtain the MEP as either a free energy surface by using dynamic techniques, or as a potential energy surface through the use of static techniques.[15] The differences

---

[XIII] This chapter is based on the following publication: N. R. McFarlane, J. N. Harvey, *Phys. Chem. Chem. Phys.*, **2024** 26, 5999-6007. For the purposes of this thesis, sections of the publication have been slightly modified. Specifically, the order of sections has been moved around slightly. Also, Section 5.3.3.4, which was not included in the published paper, but may be published later, has been added.

All calculations, simulations and code implementation were performed by N. R. McFarlane. Results were interpreted by both authors. The first draft of the paper was entirely written by N. R. McFarlane and was then edited by both authors.

between these types of method have been explored in the context of free energy prediction in Section 2.5.2, and in practice both are used. In this chapter, the focus is on static techniques, and improved algorithms for identifying MEPs by static techniques are desirable.

At this point, it is important to clearly define what a MEP is in the context of a molecular potential energy surface. A MEP can be defined as the pathway of least energy connecting two minima on a potential energy surface. At every point along the MEP, the energy is minimised in directions orthogonal to the path, and any maxima or minima correspond to TSs or intermediates, respectively. Since one cannot reasonably minimise every point along a MEP, it is often approximated as a discrete set of structures which can be generated in several ways, which will be explored below. For the sake of convention, in this chapter, even though the MEPs obtained are strictly speaking an approximation to the true MEP, they will be referred to as the MEP.

Notwithstanding the above simplification, calculation of a MEP is seldom trivial due to the high number of DoF ($3N - 6$, where N is the total number of atoms in the system) giving a potential energy surface with dimensionality equivalent to this number of DoF. Accordingly, there are many, many plausible TSs and/or intermediates connecting two minima. Complicating the problem yet further, since the many possible TSs and intermediates are closely related, during the course of MEP determination, it is not impossible that from one discrete structure to the next along the pathway, one undergoes a 'jump' from one MEP to another one, meaning that the pathway obtained may be structurally and energetically discontinuous.

In any case, two classes of static technique to find a MEP exist: double- and single-ended methodologies. Double-ended methodologies start from a pair of structures, often corresponding to reactant and product minima, and then connect them *via* a

MEP.[16] Usefully, double-ended methods do not require a definition of the reaction coordinate to function, as the structural differences between the two 'ends' are used to generate the MEP linking them. Before presenting the various double-ended algorithms which are available, regardless of which algorithm is used, there is an important caveat which must be expressed. This is that if one wishes to obtain a MEP for a simple chemically activated transformation, the two 'ends' (typically reactant and product structures) must be configurationally coupled, i.e. the two input structures need to be as geometrically similar as possible given the required changes in chemical bonding occurring during the reaction. This can be viewed as an application of the principle of least motion.[17] This is problematic because as system size increases, so to do the number of DoF, and so independently-generated reactant and product structures are very likely to differ in many 'spectator' degrees of freedom.[18] This requirement for configurationally coupled reactant and product structures means that double-ended reaction pathway optimisation techniques are not always readily applicable to biochemical systems, as the 3-dimensional structures (such as those obtained from the PDB) of reactant and product are very often not configurationally coupled in the sense described above. In principle, a MEP can still be calculated, but it may involve structural rearrangement as well as chemical changes, and it may also prove to be particularly susceptible to the MEP "jumps" mentioned above.

Caveats aside, the most widely used double-ended method in the QM/MM framework is the nudged elastic band method (NEB).[19–23] The NEB method works on the principle of first creating a set of system replicas, or so-called images, between the reactant and product by linear interpolation, usually in Cartesian coordinates, between their geometries. Then, the pathway is iteratively optimised, but subject to a restoring spring force defined between neighbouring images. This force preserves the even spacing of the

images that was initially created by interpolation between reactant and product.

If one generates the initial structures for the images by simple linear interpolation of Cartesian coordinates, the created images may have unfavourable structures, with close contacts between different atoms. This will be particularly so for reactions with significant structural changes between reactant and product. Such close contacts can result in problems with convergence of electronic structure calculations and could also cause MEP "jumping" to occur. To sidestep these difficulties, Smidstrup et al. have proposed alternative strategies to generate the images by interpolating between distances between pairs of atoms and then optimising these in a least-squares sense.[24] More recently, Zhu et al. have demonstrated that the images should be generated non-linearly, and instead that a geodesic curve is a better approximation to chemical space.[25]

Nonetheless, the NEB method has seen extensive use in modelling biochemical reactivity. As an example, the NEB method with a QM/MM model has been used to successfully describe the hydrolysis of ATP in solution,[26] a critically important biochemical process resulting in the release of stored energy. Previously, using user-defined reaction coordinates,[27] this reaction had been predicted to be endothermic. The authors of the NEB study argue that the NEB method, by not pre-assuming a reaction coordinate, yields a more accurate MEP and thereby also yields a more meaningful reaction energy, which they find to be in quantitative agreement with experiment.

Another illustrative example of the NEB method in a QM/MM model is in the study of a glycosyltransferase enzyme, ppGalNAcT2.[28] In contrast to the previous in-solution study of ATP hydrolysis, this study investigates an enzymatic mechanism using the NEB method. The authors successfully calculate a MEP for their reactant and product pair of structures, however, to obtain such

structures, extensive and error-prone optimisation procedures needed to be performed.[29] The resultant MEP differs by 8 kcal/mol from its minimum free energy pathway counterpart, which may be in part due to how the reactant and product structures were generated. This, however, is plausibly simply a result of insufficient sampling – a concept that will be explored later in this chapter.

In response to the difficulties associated with linearly interpolated system replicas in the NEB method, in 2004, Peters et al. suggested that instead of generating the entire reaction pathway at the onset of the search, that the discretised points along the MEP could be generated sequentially. This procedure also provided the name for this new approach, with the algorithm being termed the growing string method (GSM). In this approach, the pathway connecting reactant and product is 'grown' one image at a time from both sides until both path fragments meet. The algorithmic details of the GSM and how it is useful will be explored below, and this double-ended variant of the GSM (DE-GSM) has been used in several works.[30–33]

Despite the usefulness of double-ended methods, the need for a pair of structures greatly limits their use in biochemical modelling. On the other hand, single-ended methods only require a single starting structure, typically a reactant minimum, but product minima can also be used. This requirement for only one structure largely eliminates the configurational coupling problem, as one can generate the other structure from the one which is already available, minimising possible structural differences in 'spectator' atoms between the two. However, in order to reach the desired product (or reactant), one also needs to provide some definition of the reaction coordinate, which is often a challenging task.[7,16]

The natural example of a single-ended method is the adiabatic mapping approach (often also called a relaxed scan).[7] The approach is conceptually simple, where a reaction coordinate describing the major changes in chemical bonding for a reaction is

defined and iteratively incremented. Each newly generated structure is then constrained subject to this reaction coordinate, and the procedure stops when a structure resembling the expected product is reached. A primary limitation of this approach is that the expected structural changes between reactant and product should be relatively minimal, as the requirement for a user defined reaction coordinate becomes increasingly complex with more complicated reactions.

While it is true that careful application of the adiabatic mapping approach has the advantage that the obtained MEP is more likely to connect two structures that are configurationally coupled, the large number of DoF means that during optimisation of the MEP, structural changes in 'spectator' degrees of freedom are frequently observed, ultimately leading to MEPs that include one or several discontinuities in the energy profile.[7] Another contributing factor to this behaviour is that the reaction coordinate is defined once and only once at the start of the pathway optimisation. As described above, it may be that for complex reactions, one cannot easily define an appropriate reaction coordinate. In addition, there are also cases with conformationally flexible groups or solvent molecules near the reacting site where many DoF can couple strongly to the bonding change and accurately defining the reaction coordinate along the whole course of the pathway can be challenging.[7]

Nevertheless, the adiabatic mapping approach is still extensively used in biochemical modelling. As an example, this approach was used in Chapter 3, where the binding of NO gas to heme groups in sGC was studied.[34] In this case, the reaction coordinates chosen were mostly suitable, but in some cases it was observed that certain bonds had elongated more than they were expected to, illustrating that the choice of a reaction coordinate can be hard even for reactions with minimal structural changes.

Another example of the limitation of coordinate-driving approaches such as adiabatic mapping is in the study of the reaction found in chorismate synthase, where 5-enolpyruvylshikimate-3-phosphate is converted to chorismate in two steps.[35] In this study, the authors use the adiabatic mapping approach to generate an approximate pathway and then use the NEB method to improve the quality of the pathway. Ultimately, they find that while the adiabatic mapping pathway was of reasonable quality considering how easily it was generated, if it was used as-is without further refinement by the NEB method, then the TS would have been wrongly characterised. This is due to the fact that when using adiabatic mapping in its purest form, the TS obtained is strictly an approximation. NEB methods, however, are often supplemented with an exact eigenvector-following TS search algorithm used to characterise the TS more accurately. In any case, the synergistic approach using adiabatic mapping followed by the NEB method is a common one.[14] By using both, one can ensure that reactant and product are (at least mostly) configurationally coupled since the initial pathway is generated from a single structure, and then the quality of the pathway can be improved with the NEB method.

While the adiabatic mapping/NEB approach outlined above is effective, avoiding the difficulties associated with both techniques, it does of course incur a higher computational cost as well as being laborious. There is therefore the scope for a single-ended QM/MM technique which directly improves upon the method for construction of a reaction coordinate. In this way, one can avoid the configurational coupling problem, and also not encounter the problems associated with defining a single reaction coordinate to describe the whole pathway. Ultimately, a method which satisfies these requirements would be expected to improve MEP quality and not require any remapping or successive NEB pathway refinement, reducing labour and computational cost.

In 2015, Zimmerman developed a single-ended variant of the GSM (SE-GSM).[36] It is broadly similar to the DE-GSM approach, with both algorithms having a two-stage methodology in which the pathway is first 'grown' by adding discrete structures along the pathway, and then optimised with an additional reparameterization step between optimisation cycles.[32,36] The method uses a DLC scheme, which allows for the reaction coordinate constraints to be expressed in a flexible way as a combination of all relevant internal coordinates.[37–39] Zimmerman's implementation is limited to purely QM reaction pathway calculations, and does not have functionality for QM/MM reactivity, as is the focus of this chapter.

In this chapter, the first implementation of the SE-GSM within the QM/MM framework is outlined, using a hybrid coordinate system, where the QM region is described by DLCs, and the MM region by Cartesian coordinates.[40] As mentioned above, the SE-GSM algorithm has two stages, where the initial reaction coordinate used during the growth phase is user-defined (this phase is in many ways similar to the adiabatic mapping method[14]), and the reaction coordinates used in the subsequent optimisation phase are allowed to vary along the pathway. This variable reaction coordinate should, in principle, have an improved propensity to obtain continuous MEPs.

The newly-implemented algorithm is verified against a well-understood test case: the Claisen rearrangement catalysed by chorismate mutase. Not only is this reaction well-characterised, but there is quality 3-dimensional structure data available, and extensive previous QM/MM studies have been performed.

This study also represents the greatest number of QM/MM pathways that have been generated for the Claisen rearrangement catalysed by chorismate mutase. Due to the large number of obtained pathways, an investigation into the underlying

distribution of QM/MM activation energies was enabled, providing an opposing view to the previously suggested distribution.

5.2 Algorithm Details

5.2.1 Hybrid Delocalised Internal Coordinates

The optimisation implementation uses hybrid DLCs as these yield faster optimisation than Cartesian coordinates, have fewer degrees of freedom,[41,42], have a lesser degree of coordinate coupling, and they also facilitate the implementation of multi-component constraints.[37–39] In addition, DLCs can be generated completely automatically from a set of Cartesian coordinates. These advantages outweigh the computational expense required to generate them for large systems, and the need for an iterative back-transformation procedure between optimisation cycles to Cartesian coordinates that does not always converge.[37–39] By using a hybrid DLC scheme where DLCs are used for the QM region, and Cartesian coordinates for the MM region,[40] the expense is in any case negligible, and convergence of the back-transformation can be ensured by taking relatively small geometry update steps with a robust algorithm.

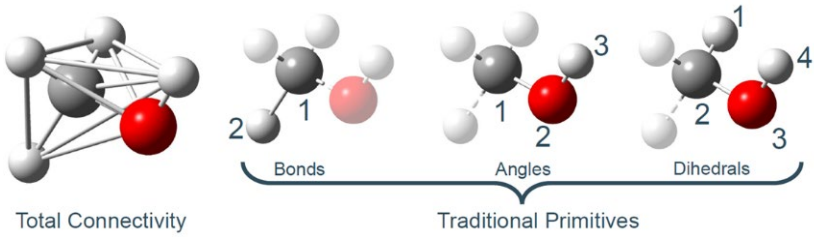

Figure 5.1: *Allowed definitions of primitive internal coordinates implemented in QoMMMa. On the left is the 'total connectivity' scheme where every atom is connected to every atom, and on the right is the 'traditional primitives' scheme which is a covalently defined set of bonds, and valence and dihedral angles.*

The process of generating and constraining a set of DLCs has been described in Section 2.6, here only a brief outline is given with relevance to a hybrid DLC QM/MM implementation. The primitive internal coordinates are first generated completely automatically and allows for either a complete set of interatomic distances (a 'total connectivity' scheme) or a covalently defined set of bonds, and valence and dihedral angles (see Figure 5.1).

The primitive internal coordinates and the subsequently formed DLCs are only generated for the QM region – a previously validated partitioning scheme for QM/MM calculations.[40] In an additive and electrostatic embedding scheme such as the one in *QoMMMa*, the QM and MM regions are independently optimised, and the QM region is polarised *via* MM region point charges. This independent optimisation means that treating the two regions with different coordinate types is straightforward. Any link atoms introduced by scission of a covalent bond across the QM-MM boundary can be seamlessly included in the generation of the primitive internal coordinates and hence DLCs. Moreover, the microiterative scheme as implemented in *QoMMMa* where the entire MM region is optimised at each macrocyle[43] is preserved whilst still optimising the QM region with DLCs, thereby improving performance at low computational expense.[44,45] Optimisation steps involve updating the DLCs, which can be seamlessly converted to a step in Cartesian space using the methods described in Section 2.6.

For reaction pathway optimisation, constraints are needed, and their introduction in a DLC-based algorithm is relatively straightforward, as outlined in Section 2.6.3. The constraint imposed can, in principle, be applied to any linear combination of several primitive internal coordinates,[37,38] and can be imposed exactly. This is beneficial for the SE-GSM algorithm, as reaction coordinates can be many-component, as will be explored below.

5.2.2 Single-ended Growing String Method

As described in Section 5.1 single-ended methodologies begin with just one structure, and incrementally adjust the structure subject to some reaction coordinate.[16] However, a poorly-defined reaction coordinate can result in a discontinuous reaction pathway. The SE-GSM partly mitigates this problem due to its two-stage nature, whereby the user-defined reaction coordinate is only used initially in the growth phase, while a structure-dependent reaction coordinate is used during the optimisation phase. Furthermore, during each cycle of the optimisation phase, a reparameterization step is carried out, where the set of discrete structures (or so–called nodes) are equally distributed along the reaction path – thus ensuring even spacing of said nodes.[36]

In the growth phase, the user defines an initial, guess reaction coordinate known as the primitive driving coordinate, $\boldsymbol{q}_{init}$. $\boldsymbol{q}_{init}$ may be a combination of any number of primitive internal coordinates, and it does not need to describe the reaction perfectly. It should, however, capture the anticipated changes in chemical bonds, angles, and dihedral torsions. Unlike in usual single-ended methodologies, the range over which the reaction coordinate should be incremented is not required as input but handled internally. In order to generate a new node (i.e., to 'grow' the string), for each element of $\boldsymbol{q}_{init}$, linear steps of 0.1 Å and 5° are taken for bonds and valence and/or dihedral angles, respectively. These steps may be either in the positive or negative direction depending on the exact definition of $\boldsymbol{q}_{init}$. While the size of these steps is the default, for cases where $\boldsymbol{q}_{init}$ contains many components, it may be desirable to reduce the magnitude of the overall step taken so that the structure of the newly generated node is not severely different from that of the previous one. For the case study described below in Section 5.3, the magnitude of these steps has been verified by performing five pathway optimisations using half the step-size. The five pathways obtained were similar in

shape and showed the same barrier height (see Figure C7). It may be beneficial to take smaller steps for cases where the TS is structurally very similar to either reactant or product.

In detail, the growth phase operates as follows: starting with primitive internal coordinate structure $\boldsymbol{q}_i$, structure $\boldsymbol{q}_{i+1}$ is formed by taking a linear displacement along $\boldsymbol{q}_{init}$. Subsequently, $\boldsymbol{q}_{i+1}$ is converted to DLCs (forming $\boldsymbol{s}_{i+1}$), constrained subject to $\boldsymbol{q}_{init}$, and QM/MM optimisation cycles are performed, where the convergence criteria used here are relatively 'loose', and a maximum of 5 optimisation steps per node is taken. This cycle is repeated until the QM/MM energy of structure $\boldsymbol{q}_{i+1}$ is lower than the QM/MM energy of structure $\boldsymbol{q}_i$ (i.e., a barrier has been surmounted). In this way, it is worth mentioning that this algorithm is limited to reactions with a measurable potential energy barrier.

The growth phase is completed by first adding an equal number of nodes on the other side of the barrier and performing a constrained optimisation on each of these nodes, also following $\boldsymbol{q}_{init}$. However, as postulated by Hammond,[46] other than for thermoneutral reactions, barriers are not symmetric. Therefore, to ensure that the product has been reached, an unconstrained optimisation is performed on the final node in the string. In the following optimisation phase, this final node is fixed, meaning that the node density along the string may change in the following stages of the algorithm. In the description thus far, the SE-GSM bears resemblance to adiabatic mapping, but the pathway obtained at this stage is much more approximate than with adiabatic mapping,[14] and the pathway will subsequently be refined in the optimisation phase.

In the optimisation phase, $\boldsymbol{q}_{init}$ is discarded and, instead, the change in $\boldsymbol{q}$ for each pair of neighbouring nodes $\boldsymbol{q}_i$ and $\boldsymbol{q}_{i+1}$ is taken and used to define the reaction coordinate, henceforth $\boldsymbol{\tau}_{i,i+1}$, as shown in Equation (5.1).

$$\boldsymbol{\tau}_{i,i+1} = \begin{cases} q_{i+1} - q_i \ , & E_{i+1} > E_i \\ q_i - q_{i+1} \ , & E_{i+1} < E_i \end{cases} \qquad \textbf{(5.1)}$$

Not every single element of $\boldsymbol{q}$ will be included in $\boldsymbol{\tau}_{i,i+1}$, but only those which exceed an internally-defined cut-off threshold going from structure $\boldsymbol{q}_i$ to $\boldsymbol{q}_{i+1}$. By default, these cut-offs are defined as 0.05 Å and 2.5° for bonds and valence and/or dihedral angles, respectively. To minimise any possibility of kinks in the string, Henkelman's upwinding principle is obeyed, whereby $\boldsymbol{\tau}_{i,i+1}$ is defined to point in the direction of increasing energy,[19] as indicated in Equation (5.1). This definition of the reaction coordinate is dynamic, and captures aspects of the reaction coordinate which were not predicted initially in $\boldsymbol{q}_{init}$. In this phase, no new nodes are added, but each node is constrained subject to $\boldsymbol{\tau}_{i,i+1}$, and optimised this time with usual convergence criteria and now a maximum of 25 optimisation steps. Since the reaction coordinate is no longer equivalent at each node, the spacing of nodes will not be even, so a reparameterization step is taken. During this step, the overall difference in $\boldsymbol{q}$ between the reactant and (now fixed) product node is calculated, and each node is linearly spaced along this tangent. The cycle of optimisation phase and reparameterization step is repeated until the constrained QM/MM optimisation of every node on the string is converged, and the SE-GSM algorithm is complete.

With the algorithm complete and a MEP obtained, the activation energy can be easily approximated. It is worth noting that, in its current implementation, the activation energy and hence the TS geometry are strictly-speaking an approximation as an exact TS search algorithm has not yet been implemented.[22,33,40] Given the

algorithmic similarity,[32,33,47] the DE-GSM algorithm has also been implemented into *QoMMMa*. However, due to the aforementioned reasons, it is not necessarily useful for biomolecular QM/MM reaction pathway calculations. It may, however, be useful in other QM/MM reaction pathway applications.

5.3 Case Study: Chorismate Mutase

5.3.1 Introduction

As described in Chapter 1, as an initial test case, a well-understood biomolecular reaction, the Claisen rearrangement catalysed by CM was selected. In this reaction, the substrate chorismate is converted to prephenate in a single step (as shown in Figure 1.8).[48–51]. This choice was based upon several factors, including ease of system preparation, system size, and wealth of computational investigations already performed.

Extensive QM/MM studies have been carried out on this reaction using BsCM as the starting structure, so the barrier height and nature of the TS is now well-characterised.[52–63] QM/MM investigations into the Claisen rearrangement in BsCM are chronologically given in Table 5.1, including also the experimental energy barrier for the sake of comparison.[64]

Overall, five QM methods have been used since the original study, as well as some minor variations in MM forcefield choice. The barrier has been most accurately calculated using LCCSD(T_0), where the energy barrier is predicted to be 13.1 ± 1.1 kcal/mol – in excellent agreement with experiment. However, the high computational expense means that this method has only been used for single-point calculations on a B3LYP optimised pathway.

Table 5.1: *Chronological list of QM/MM studies of reactivity in BsCM. For each study, the starting PDB structure(s), the QM method and basis set, the MM forcefield, the barrier height (in kcal/mol) with standard deviation or range (where available) are provided.*

pdb-id	QM method/Basis set	MM force field	Barrier height (kcal/mol)	Reference
2CHT	AM1	CHARMM	17.8	52
2CHT, 1COM	HF/6-31G(d)	Amber	24.8	53
2CHT, 1COM	MP2/6-31G(d)	Amber	6.4	53
N/R	B3LYP/6-31G(d)	Amber	1.4	54
1COM	AM1	CHARMM	35.7	55
2CHT	HF/4-31G//cc-pVDZ	CHARMM	19.7	56
2CHT	B3LYP/6-311+G(2d,p)//6-31G(d)	CHARMM	2.7 – 11.9	57
2CHT	AM1	CHARMM	24.5 – 31.6	57
2CHT	MP2/6-31+G(d)//6-31G(d)	CHARMM	7.4 – 11.0	58
2CHT	B3LYP/6-311+G(2d,p)//6-31G(d)	CHARMM	12.7 – 16.1	58
2CHT	B3LYP/6-31G(d)	CHARMM	12.0 ± 1.7	59
2CHT	B3LYP/aug-cc-pVTZ//cc-pVTZ	CHARMM	10.2 ± 1.8	60
2CHT	LCCSD(T_0)/aug-cc-pVTZ//cc-pVTZ	CHARMM	13.1 ± 1.1	60
2CHT	B3LYP/6-31G(d)	CHARMM	11.3 ± 1.8	61
2CHT	B3LYP/6-31G(d)	Amber	12.1 ± 0.2	62
2CHT	B3LYP/def2-svp	Amber	15.1	63
N/A	Exp.		12.7 ± 0.4	64

Taking the averages of the remaining four methods (using the upper end of the ranges from references 50 and 51), barriers of 8.7 kcal/mol (MP2), 28.4 kcal/mol (AM1), 22.3 kcal/mol (HF), and 11.3 kcal/mol (B3LYP) are predicted. Compared to the experimental value, HF and AM1 clearly overestimate the barrier height, and have not been used in more recent studies of this enzyme. On the other hand, MP2 underestimates the barrier by a few kcal/mol. When comparing to the known experimental value for this reaction, the best-known choice for QM/MM reaction pathway generation is B3LYP. Removing the outlier predicted by Hall 2000, who calculate the barrier to be unreasonably small presumably due to insufficient system preparation and/or sampling, reducing the average barrier to be 12.7 kcal/mol – very similar to the experimental value. It is worth noting that the size of the QM region, methods for system preparation, and the extent of sampling is not consistent across these studies, so these factors may also have some effect on the energy barriers.

The origin of the catalytic effect in CM has been much discussed, with two dominant theories. The first is the so-called near-attack conformation ('NAC') theory, whereby catalysis occurs due to the enzyme preferentially binding to the substrate in a conformation that is close in structure to the TS.[65] A frequently-used definition of a 'NAC' in CM is a reactant structure in which the van der Waals radii of the reacting atoms (C_1 & C_9 in Figure 1.8) overlap, so that $d(C_1 - C_9) < 3.7$ Å.[61] A particular 'NAC' valence angle[65] as well as two particularly important dihedral torsions have also been used to measure the distance between reactant structures and the TS and thereby identify a 'NAC'.[66]

While the proportion of 'NAC'-like structures found in the reactant complex varies somewhat depending on the exact definition used,[48] an increased population of pre-reactive structures compared to aqueous solution is found, and is argued to explain catalysis. The second theory argues that the key factor is TS

stabilisation by the enzyme environment *via* intermolecular interactions.[57–59,67] Arguments in favour of each of the theories have been provided, and for the present purposes, the distinction is perhaps less important.

5.3.2 Computational Details

For modelling the Claisen rearrangement in BsCM,[68] the starting point was a crystal structure containing a TSA was used (pdb-id:2CHT),[69] where Bartlett's inhibitor is used as the TSA.[70] 2CHT is a homotrimer with three active sites each situated between two of the monomers, and the site at interface between chains A and B was modelled here. Missing residues at both the C- and N-terminus of chains A and B were ignored due to their distance from the active site. As in previous work,[57] a gas-phase optimised (B3LYP/6-311G[71–75] with Grimme's D3 dispersion correction and the BJ damping function[76,77]) structure of chorismate was placed in the active site after aligning it to the TSA. The DFT calculations were performed using Gaussian16.[78] Forcefield parameters for the chorismate co-factor were generated using CGenFF[79] with the geometry obtained from the gas-phase optimisations of chorismate; the initially suggested parameters were used without additional validation. Protonation states of all titratable amino acid side-chains were predicted using PropKa.[80] The system (protein and chorismate) was solvated using the TIP3P model[81] and placed in a rhombic dodecahedron with the protein at least 10 Å from the box edge. The total charge of the system was neutralised with Na^+ and Cl^- ions, and set to a physiological concentration of NaCl (0.15 M).[82] To remove any initial bad contacts, an energy minimisation procedure was performed. Following this, NVT and NPT simulations were performed for 100 ps each, using Berendsen thermostats and pressure couplings with positional restraints on all heavy atoms, thus equilibrating the system at 300 K and 1 bar. Lastly, positional restraints were sequentially relaxed, and a 200 ns simulation was performed in the

NPT ensemble with Parrinello-Rahman pressure coupling. All MD simulations were performed within GROMACS[83,84] and used the CHARMM36/CGenFF forcefield.[79,85] Along the course of the simulation, the RMSD of the polypeptide α-carbons was monitored (see Figure C1), and stabilised after approximately 68 ns at 0.098 nm. Starting from 68 ns, 132 snapshots were selected after each ns of simulation and, without any further pre-screening, used as starting points for QM/MM calculations.

In preparation for the QM/MM calculations, water molecules further than 15 Å from the centre-of-mass of the chorismate co-factor were deleted giving a total of 6010 atoms. An energy minimisation (with an RMS gradient criterion of 0.001 kcal/mol/Å) was performed on the resulting large cluster model using the same CHARMM36/CGenFF[79,85] forcefield in the Tinker program package,[86] with atoms > 10 Å from the centre-of-mass of the chorismate frozen. For all QM/MM calculations, the calculations for QM and MM regions of the system were performed with two different packages – Gaussian09[87] and Tinker,[86] respectively. The QM and MM parts of the system are combined as a QM/MM calculation using *QoMMMa*.[43] The chorismate co-factor was chosen as the QM region, and the remaining protein and solvent was the MM region – resulting in no link atoms between the QM and MM regions, and a QM region size of 24 atoms. For the QM region, all calculations used B3LYP/6-31G(d)[71–75] and Grimme's D3 dispersion correction with the BJ damping function.[76,77] The MM region was described using the CHARMM36 forcefield.[85] For each snapshot, an unconstrained optimisation using DLC for the QM region was performed. The primitive driving coordinate, $\boldsymbol{q}_{init}$, selected was the difference between the forming bond and the breaking bond, i.e. $\boldsymbol{q}_{init} = d(C_3 - O_7) - d(C_1 - C_9)$ (see Figure 1.8). To minimise any structural changes in distant MM atoms, and ultimately improve reaction pathway continuity, any atoms which were > 10 Å from the centre-of-mass of chorismate were frozen.[7]

5.3.3 Results and Discussion

5.3.3.1 Pathway Quality with the QM/MM SE-GSM

Of the 132 pathways obtained, a number were found to be discontinuous based on an analysis of the QM, MM and QM/MM energies for neighbouring nodes along the MEP. If the QM, MM, or QM/MM energy difference to one side of a given node differed too strongly from that on the other side (i.e., if $E_{k-1} - E_k$ differed by more than 6 kcal/mol from $E_k - E_{k+1}$), this was considered to possibly be a discontinuous path to be verified by visual inspection. The threshold of 6 kcal/mol was based on inspection of the visibly discontinuous paths in their QM/MM energy profiles, and a sensitivity analysis of this threshold where it was changed by ± 1 kcal/mol did not cause any variation in the number of discarded pathways. For examples of reaction pathways with discontinuous QM/MM energy profiles, see Figure C8. Ultimately, 15 of the pathways were discarded using this analysis. The remaining 117 pathways were considered to be sufficiently smooth, although minor discontinuities may still be present. While the reaction catalysed by CM is known to be a favourable case for optimisation of continuous or near-continuous reaction paths, which have been obtained in many studies using adiabatic mapping,[52–63] the fact that starting from 132 randomly-selected initial structures and using an automated protocol, a success rate of over 88% in generating near-continuous pathways is encouraging in terms of the wider applicability of the SE-GSM.

To understand the accuracy and efficiency of the QM/MM SE-GSM versus adiabatic mapping, five reaction pathways using the adiabatic mapping approach were generated and compared to the SE-GSM reaction pathways which used the same starting geometry. Looking at the barrier heights, both techniques predict approximately the same barrier height, with average difference in barrier heights across the five pathways being 0.6 kcal/mol. Turning to the question of efficiency, the average number of

gradient calls required to reach a smooth pathway were 745 and 627 for the SE-GSM and adiabatic mapping, respectively, so the latter is slightly more efficient.

Gradient calls aside, the SE-GSM is a more streamlined procedure as one does not need to manually define a reaction coordinate with appropriate range, but instead one provides a guess of the reaction coordinate components. Also, the five pathways generated from the SE-GSM were already smooth and therefore meaningful, whereas two of the pathways generated using adiabatic mapping were initially not continuous and required refining. This is a common experience with adiabatic mapping studies of reaction pathways and may be because the constraint is applied along the same pre-assumed reaction coordinate along the whole reaction pathway. In any case, discontinuities in QM/MM energy profiles are quite common when using adiabatic mapping, so that to achieve a smooth profile, one needs to perform quite tedious reverse mapping.[7] Hence, even though the total number of gradient calls required to generate smooth pathways was slightly lower with adiabatic mapping, the ease of generating smooth pathways with the SE-GSM is a considerable benefit.

In general, the quality of pathways generated by the SE-GSM can be in part ascribed to the relatively 'loose' convergence criteria used in the growth phase, meaning that discontinuities in the QM energy profile are less likely to occur as compared to adiabatic mapping. Moreover, any discontinuities in the QM energy profile which emerge following the growth phase should be smoothed-out by the flexibility of the optimisation phase reaction coordinate. The primary limitation in generating quality pathways lies in the choice of $\boldsymbol{q}_{init}$, as it must represent the main changes in chemical bonding for the reaction under study – something which is not always trivial to understand. However, the selection of an appropriate $\boldsymbol{q}_{init}$ can in principle be done automatically.[88–90]

5.3.3.2 Prediction of the TS in BsCM

As the potential energy surface in BsCM was sampled extensively using the MD snapshot methodology, there is a degree of variation in predicted barrier heights from different starting snapshots (see Figure 5.3 for a histogram of all barrier heights). Of the 117 reaction pathways, the standard deviation of the barrier height was ± 2.8 kcal/mol and the maximum and minimum barrier heights were 9.3 kcal/mol and 24.2 kcal/mol, respectively. This discrepancy in barrier heights is not unusual for static QM/MM reaction pathway calculations and highlights the importance of appropriate sampling.[7] In the case herein, given the relatively modest standard deviation, the extent of sampling seems to be sufficient.

There are two commonly used methods to evaluate the average barrier height, the arithmetic and exponential average. As applied to the 117 pathways generated herein: the arithmetic average of all barrier heights 16.8 kcal/mol, while exponential averaging yields a value of 12.0 kcal/mol. The exponential averaging technique has been shown to be preferable for approximating the free energy of the barrier, as lower energy barriers contribute more to the average barrier height,[91] hence the discussion below will focus on exponential averaging (even though this does not yield a rigorous estimate of the activation free energy). The calculated barrier height of 12.0 kcal/mol calculated herein agrees well with both the average of previous DFT-based computational investigations of 12.7 kcal/mol, and with the experimental barrier of 12.7 kcal/mol – as summarised in Table 5.1. The key bond lengths (including their standard deviations) for the reactant, TS, and product are summarised in Figure 5.2.

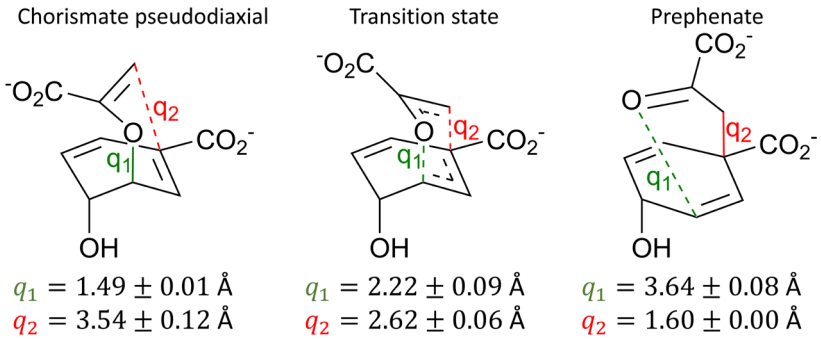

Figure 5.2: *The average and standard deviation of selected interatomic distances across all 117 snapshots for the reactant (left), TS (centre), and product (right) for the two key interatomic distances in the Claisen rearrangement in BsCM. When < 0.00 Å, standard deviations are rounded to 0.00 Å.*

The predicted TS structures are rather similar to those reported previously, e.g. $q_1 = 2.14 \pm 0.05$ Å and $q_2 = 2.64 \pm 0.05$ Å from reference [61], or $q_1 = 2.28$ Å and $q_2 = 2.58$ Å from reference [58]. This favourable agreement, and the rather low standard deviations of the key bond lengths around the TS, gives confidence that the QM/MM SE-GSM with hybrid DLCs approximates the TS well for this reaction. With reference to Figure 5.2, one may note that the standard deviation of q_2 for the reactant structures is larger than that of any other primitive internal coordinate, indicating that some reactant structures are considerably 'closer' to the TS than others. This reflects the unbiased nature of the sampling of reactant structure from the MD trajectory. As a result, the variance of the obtained barrier heights and reactant structures is greater than in previous investigations of this system, which typically involved some degree of manual selection of initial structures.

For all pathways, the so-called 'NAC' angle[65] for the reactant geometry and plotted this against the barrier height has been calculated – see Figure C2. The activation energy has been suggested to be directly dependent on this 'NAC' angle, where a lower 'NAC' angle suggests that the reactant geometry is predisposed to form the TS, thus lowering the barrier height. In this

work, there is some (albeit quite small) correlation between the ’NAC’ angle and the barrier height, where a lower angle gives a lower barrier height. Since all the reactant conformations used in this work were in the pseudodiaxial form of chorismate, and therefore all more-or-less a ’NAC’, seeing only a minor correlation is not unexpected. An alternative theory argues that the primary effect in lowering the barrier height is TS stabilisation by the enzyme environment *via* electrostatic preorganisation.[57–59,67] For the present purposes, whether the ‘NAC’ or electrostatic preorganisation theory is the main barrier lowering effect is unclear given that the structural features leading to TS stabilisation are also expected to stabilise ‘NAC’s and hence lead to their observation.

5.3.3.3 Hybrid DLCs and Dynamic Reaction Coordinate

The hybrid DLC scheme as applied to additive QM/MM calculations is not new,[40] but specifically in the case of the GSM where the reaction coordinate does not change linearly, it was not clear whether this scheme could be seamlessly implemented. In the current implementation, the QM region is the DLC region by default. However, in principle, the choice of DLC region in a hybrid DLC scheme need not be limited to simply the QM region, but could also include particularly important MM atoms, or indeed a smaller subset of the QM region atoms. Including MM region atoms requires modification of the Hessian matrices used for optimisation,[40] and it is not clear if there is a significant benefit to using a scheme such as this given that the most significant structural changes typically occur in the QM region. Moreover, selection of important MM atoms requires more user intervention – something which should be minimised in the QM/MM SE-GSM methodology. However, in cases where it is desirable to use a particularly large QM region,[92] but only one part of it undergoes any significant structural changes, it could be useful to use DLC for only this part. Therefore, the repeated diagonalisation procedure required to generate the DLC could be made faster.

When compared to adiabatic mapping, the implementation of the SE-GSM offers a dynamic definition of reaction coordinate during the optimisation phase, and the improvement over adiabatic mapping can largely be ascribed to this. This benefit is less important in a simple case study as described herein, but for calculations on more complex biochemical systems, this may be critical to achieve a smooth MEP. The effect of the dynamic reaction coordinate can be clearly seen in the components of the reaction coordinate as the optimisation phase proceeded, where several additional interatomic distances emerged as important. These are summarised alongside the original primitive internal coordinates within $\boldsymbol{q}_{init}$ in Table 5.2.

Table 5.2: *Summary of the interatomic distances in the reaction coordinate during the optimisation phase. The first two columns give the name and definition of the coordinates, where q_1 and q_2 are the primitive internal coordinates within $\boldsymbol{q}_{init}$. The final three columns give the average interatomic distances with standard deviations across all 117 snapshots at reactant, TS, and product. When < 0.00 Å, standard deviations are rounded to 0.00 Å.*

Internal Coordinate	**Definition**	**Distance (Å)**		
		Reactant	*TS*	*Product*
q_1	$d(C_3 - O_7)$	1.49 ± 0.01	2.22 ± 0.09	3.63 ± 0.08
q_2	$d(C_1 - C_9)$	3.54 ± 0.12	2.62 ± 0.06	1.60 ± 0.00
q_3	$d(C_1 - C_{11})$	1.35 ± 0.00	1.40 ± 0.01	1.51 ± 0.00
q_4	$d(C_2 - C_3)$	1.50 ± 0.01	1.38 ± 0.01	1.34 ± 0.00
q_5	$d(O_7 - C_8)$	1.39 ± 0.01	1.28 ± 0.01	1.22 ± 0.00
q_6	$d(C_8 - C_9)$	1.34 ± 0.00	1.39 ± 0.01	1.53 ± 0.01

From Table 5.2 it is clear that not only do the components of the initial driving coordinate play a part in the reaction, but so do some other relevant distances, albeit to a lesser extent. In this way, other reaction coordinate components have been included entirely automatically. The components of the initial primitive driving coordinate (q_1 and q_2) continue to have the largest weight in the reaction coordinate, but this is to be expected given the relatively

simple nature of this reaction. However, for more complex biochemical reactions in which one is unsure of which internal coordinates aid in reaction catalysis, their contribution could be captured during pathway optimisation. There are no discontinuities in q_1 and q_2, and the same is true for all other interatomic distances shown in Table 5.2 (see Figures C3 & C5). Figure C4 shows the variation of q_1 and q_2 for four sample pathways, highlighting the smoothness of the reaction coordinate for individual pathways and that the smoothness of Figure C3 is not simply a result of averaging. Given that this was generated only from an initial guess of the driving coordinate, not a specific constrained range nor any biased selection of MD snapshots, the result is quite pleasing. Besides these distances, a large number or valence and dihedral angles also played a part in the reaction coordinate. Regardless of the reaction, this will always happen as there are many valence and dihedral angles dependent on the position of a given atom. Due to their large number, they are not all reported here, but the two known important dihedral angles found in this reaction are shown in Figure C6, aligning well with previous investigations.[61]

5.3.3.4 The Underlying Distribution of QM/MM Barrier Heights

Given the number of pathways generated, there is scope for some statistical analysis on the reaction pathways generated. Specifically, in QM/MM studies, the degree of sampling of the different MD snapshot conformations required to obtain converged barrier heights is unclear.[93] Ryde has suggested that many independent pathways are needed to obtain a reasonably converged overall barrier based on the assumption that the energies of the individual barriers obtained along each of the pathways are normally distributed, and that the overall barrier is obtained by taking an exponential average.[94] The motivation for taking an average of this type is that it will most closely approximate an observed effective barrier height given a set of starting points that are randomly chosen from the reactant state

distribution, and an exponential dependence of reactivity for each pathway on the local barrier height.[91] While the assumption that the individual barriers obtained by sampling pathways will have a normal distribution seems a reasonable one, close inspection of the datasets from which this assumption was predicated on (Figure C9) reveals that the distributions are not normal.

In order to measure deviation from a normal distribution, the Fisher-Pearson skewness (g) and kurtosis (k) can be used.[95–97] These are defined using Equations (5.2) and (5.3), respectively, where x represents the sample, μ is the mean of the sample, σ is the sample standard deviation, and N is the size of the sample.

$$g = \sum_{i=1}^{N} \frac{(x_i - \mu)^3}{N} / \sigma^3 \tag{5.2}$$

$$k = \sum_{i=1}^{N} \frac{(x_i - \mu)^4}{N} / \sigma^4 \tag{5.3}$$

Returning to Figure C9, the skewness and kurtosis vary rather significantly, with both positive- and negative-skewed datasets being observed. This large variation is presumably due to insufficient sampling, and/or poor input structure preparation. It is also worth noting that the metrics used herein are heavily susceptible to outliers in the sample, particularly for smaller datasets, of which there are a number in the data reported by Ryde. Ultimately, these datasets are too small and/or poorly generated to ponder what their distribution is and does not give great insight into the overall distribution of QM/MM barrier heights. The relatively large number of pathways obtained here enable some investigation into this assumption (see histogram in Figure 5.3).

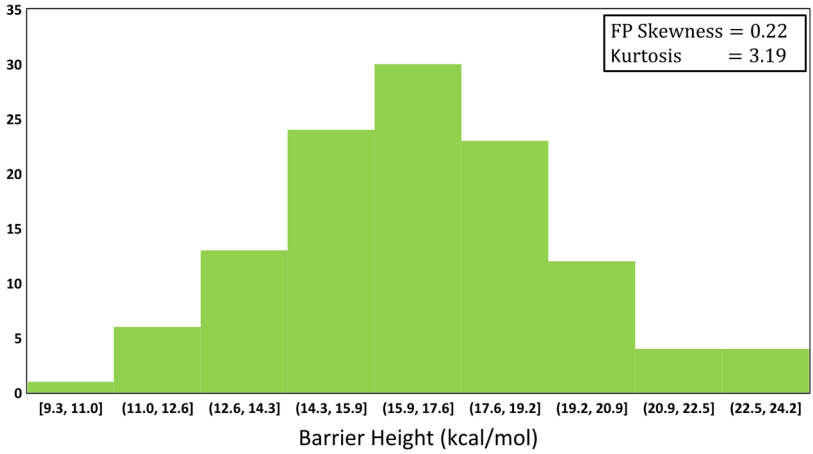

Figure 5.3: *Histogram of the distribution of barrier heights for the 117 snapshots obtained herein. Also given is the Fisher-Pearson skewness and kurtosis for the dataset, showing a slight positive skewness and slightly leptokurtic distribution.*

The obtained distribution appears to be roughly normal, as assumed by Ryde, but the Fisher-Pearson skewness, and kurtosis indicate some deviation from a Gaussian distribution. The distribution is slightly positively skewed ($g > 0.0$), and there is some leptokurtosis ($k > 3.0$). To investigate the relevance of this departure from a Gaussian distribution, the dataset has been fitted to various probability density functions (PDFs) using the residual sum-of-squares method as implemented in distfit to score the fit.[98] These PDFs used were: Gaussian, t-distribution, gamma, beta, log-normal, generalised extreme value, Weibull, continuous uniform, Pareto, and exponential. Ultimately, many of the functions showed similar scores (except the uniform, Pareto, exponential distributions), so were taken onto the next step of the fitting procedure.

For the remaining 7 PDFs, they were fit against the sample using SciPy to obtain their respective shape parameters.[99] Using the shape parameters, one can recover μ, σ, g, and k and compare these to the values calculated using the sample. Of the remaining 7 PDFs, the gamma distribution recovered μ, σ, g, and k with the

best accuracy, and was therefore taken as providing the best description of the actual distribution within this sample. For the sake of comparison, the shape parameters for the Gaussian distribution were also retained. The gamma distribution is plotted against the dataset in Figure 5.4(a), and the (by definition unskewed) Gaussian fit is plotted in Figure 5.4(b).

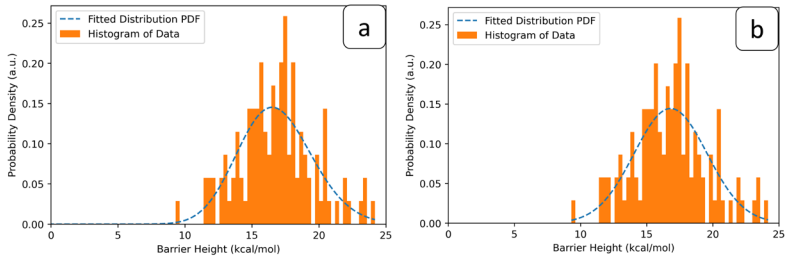

Figure 5.4: (a) *Gamma distribution fitted to the distribution of barrier heights over the 117 snapshots;* ***(b)*** *Gaussian distribution fitted to the distribution of barrier heights over the 117 snapshots.*

Having established that the barriers obtained along a large number of individual MEPs for one particular enzyme reaction are not quite normally distributed, one can return to the question of how many MEPs are needed in order to obtain an overall exponentially-averaged barrier height that is well-enough numerically averaged.[94] Assuming that QM/MM individual barriers follow a particular distribution, one can generate random sets of barrier heights that mimic the results that would be obtained with QM/MM calculations. To investigate what the effect of the distribution is on the required degree of sampling is, a Monte Carlo sampling algorithm on the gamma and Gaussian PDFs shown in Figure 5.4 was used. Sample sizes of 10, 100, 1000, 10000, 100000, 1000000, 10000000 were taken for both samples. To ensure statistical significance, each sample size was randomly taken 1000 times and averaged. To understand if the sample size is large enough, the exponentially averaged mean of the dataset was taken for each of the sample sizes, and the standard deviation was calculated. A

lower standard deviation implies that the distribution has been better sampled. The results of this procedure are given in Table 5.3.

Table 5.3: *The standard deviation of the exponentially averaged barrier with logarithmically increasing sample size. A Monte Carlo sampling along the gamma and Gaussian distributions obtained 1000 datasets for each of the sample sizes and the standard deviation reported is the average of these datasets.*

Sample Size	**Exponentially Averaged Barrier Height Standard Deviation (kcal/mol)**	
	Gamma Distribution	*Gaussian Distribution*
10	1.927	2.172
100	0.556	1.099
1000	0.231	0.537
10000	0.080	0.296
100000	0.021	0.171
1000000	0.005	0.117
10000000	0.001	0.039

The estimated standard deviation based on barrier heights selected from the gamma distribution is always lower than when using the Gaussian distribution for an equivalent sample size. Even though the skewness and kurtosis are quite small in this case, if the overall distribution of QM/MM barrier heights is in fact more like a skewed gamma distribution, then obtaining converged exponentially averaged barrier heights requires less sampling than predicted by Ryde.

Considering the nature of the distribution in some more detail, a slightly positively skewed dataset aligns with what one would predict for QM/MM-calculated barrier heights. One would expect that, for most reactions, the majority of pathways will be on the order of 10^1 kcal/mol, and thus relatively low, with some outliers at (sometimes significantly) higher energy. However, there should be fewer major outliers at lower barrier heights since barriers lower than 0 kcal/mol cannot exist. Therefore, even a handful of larger energy pathways will mean the data is, by its very nature, positively

skewed. The data reported in this work reflect this gentle positive skewness, and this behaviour could be expected to be found in other QM/MM reaction pathway studies. Note also that the Claisen rearrangement in BsCM is quite a well-behaved reaction, with a rigid active site, where the barrier height is not as sensitive to changes in conformational space as many other reactions. With more complex reactions, the positive skewness may be even greater.

The sample size used here (117 MD snapshots) is perhaps only just large enough to prove positive skewness of the distribution. Taking 1000 random half-size sub-samples of the dataset, 85% are found to yield positive skewness, while 15% are negatively skewed, though with a small negative skewness, suggesting that the conclusion is fairly robust, though a more systematic study would clearly be beneficial.

This study thereby does not finally resolve the question of how many QM/MM reaction pathways are needed to obtain reasonably converged estimates of barrier height. The results do suggest, however, that barrier heights are not normally distributed, with even small degrees of positive skewness leading to lower sensitivity of average barriers to the number of pathways. However, this conclusion is system-dependent, and a normal distribution does provide a useful initial guideline.

As a practical recommendation: when trying to evaluate an overall energy barrier using multiple reaction pathways, the high sensitivity of the result to the degree of sampling should always be borne in mind, and this can to some extent be monitored by computing the estimated error of the exponentially averaged barrier as one generates additional pathways. When this is done, the static approach to predicting reactivity using QM/MM reaction pathways can be usefully complementary to dynamic approaches such as umbrella sampling, thermodynamic integration, or free energy perturbation methodologies. Static approaches have the

benefit of providing TS structures which can be used to examine chemical bonding effects and their impact on reactivity, also allowing for comparison of bonding effects in competing reaction paths.

5.4 Conclusions and Future Work

In this chapter, the single-ended growing string method as implemented in *QoMMMa* has been described in detail and a relevant case study given. The implementation of this method represents the first in the QM/MM framework. As a validation case study, the algorithm was successfully applied to a well-known biochemical system – the Claisen rearrangement of chorismate to prephenate catalysed by chorismate mutase. The algorithm generates useful pathways for the reaction and aligns well with previous computational investigations. Moreover, the algorithm appears to more reliably generate continuous pathways than the usually employed single-ended methods, with this improvement owed to the fact that constraints applied to the discrete structures along the reaction coordinate are better able to adapt to the local nature of chemical changes taking place.

In the algorithm validation process, many continuous reaction pathways were generated, possibly the most that have been generated for this system in a given study. Accordingly, the large number of reaction pathways enabled a statistical analysis investigating the relationship between predicted QM/MM barrier height and the required degree of sampling. In contrast to what was previously assumed, the underlying distribution appears to be somewhat positively skewed, which suggests that a converged estimate of the average barrier height can be obtained with a lower extent of sampling required than has previously been thought.

With regards to future work, while the algorithm has been successfully validated, this was against a relatively simple test case. Testing against more complicated biochemical systems is

ongoing. As a second case study, an investigation into peptide bond cleavage catalysed by carboxypeptidase alpha is currently underway. Previously, this reaction was very challenging to characterise using adiabatic mapping approaches,[100] presumably due to the complexity of the reaction coordinate. While early calculations have shown that the QM/MM SE-GSM algorithm does not completely side-step the issues associated with this complex reaction, it does seem to produce pathways of reasonable quality quite readily.

A third case study involves the study of an enzyme called CutC, which catalyses the cleavage of choline to form trimethylamine and acetaldehyde.[101] This is not one reaction in this case, but rather a cycle which leads from choline to the products. Only one previous computational study has investigated CutC, but the complete catalytic cycle was not investigated.[102] Early calculations on the various steps of the catalytic cycle reveals that the QM/MM SE-GSM algorithm seems to work quite effectively for this system, although further validation is certainly required.

5.5 References

1 G. G. Hammes, S. J. Benkovic and S. Hammes-Schiffer, *Biochemistry*, 2011, **50**, 10422–10430.
2 P. K. Agarwal, *Biochemistry*, 2019, **58**, 438–449.
3 G. A. Holdgate, T. D. Meek and R. L. Grimley, *Nat Rev Drug Discov*, 2018, **17**, 115–132.
4 A. C. Rufer, *Drug Discovery Today*, 2021, **26**, 875–886.
5 E. M. M. Abdelraheem, H. Busch, U. Hanefeld and F. Tonin, *React. Chem. Eng.*, 2019, **4**, 1878–1894.
6 A. Guarneri, W. J. van Berkel and C. E. Paul, *Current Opinion in Biotechnology*, 2019, **60**, 63–71.
7 R. Lonsdale, J. N. Harvey and A. J. Mulholland, *Chem. Soc. Rev.*, 2012, **41**, 3025.
8 S. Ahmadi, L. Barrios Herrera, M. Chehelamirani, J. Hostaš, S. Jalife and D. R. Salahub, *Int J Quantum Chem*, 2018, **118**, e25558.
9 A. J. Mulholland, *Drug Discovery Today*, 2005, **10**, 1393–1402.
10 R. P. Magalhães, H. S. Fernandes and S. F. Sousa, *Isr. J. Chem.*, 2020, **60**, 655–666.
11 H. M. Senn and W. Thiel, *Angew. Chem. Int. Ed.*, 2009, **48**, 1198–1229.
12 C. E. Tzeliou, M. A. Mermigki and D. Tzeli, *Molecules*, 2022, **27**, 2660.
13 M. G. Quesne, T. Borowski and S. P. de Visser, *Chem. Eur. J.*, 2016, **22**, 2562–2581.
14 K. E. Ranaghan and A. J. Mulholland, *International Reviews in Physical Chemistry*, 2010, **29**, 65–133.
15 G. Groenhof, in *Biomolecular Simulations*, eds. L. Monticelli and E. Salonen, Humana Press, Totowa, NJ, 2013, vol. 924, pp. 43–66.
16 A. L. Dewyer, A. J. Argüelles and P. M. Zimmerman, *WIREs Comput Mol Sci*, 2018, **8**, 1–20.
17 M. L. Sinnott, in *Advances in Physical Organic Chemistry*, ed. D. Bethell, Academic Press, 1988, vol. 24, pp. 113–204.
18 P. E. M. Siegbahn and T. Borowski, *Faraday Discuss.*, 2010, **148**, 109–117.
19 G. Henkelman and H. Jónsson, *The Journal of Chemical Physics*, 2000, **113**, 9978–9985.
20 H. Jónsson, G. Mills and K. W. Jacobsen, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, WORLD SCIENTIFIC, LERICI, Villa Marigola, 1998, pp. 385–404.
21 S. A. Trygubenko and D. J. Wales, *The Journal of Chemical Physics*, 2004, **120**, 2082–2094.
22 G. Henkelman, B. P. Uberuaga and H. Jónsson, *The Journal of Chemical Physics*, 2000, **113**, 9901–9904.
23 G. Mills, H. Jónsson and G. K. Schenter, *Surface Science*, 1995, **324**, 305–337.
24 S. Smidstrup, A. Pedersen, K. Stokbro and H. Jónsson, *The Journal of Chemical Physics*, 2014, **140**, 214106.
25 X. Zhu, K. C. Thompson and T. J. Martínez, *J Chem Phys*, 2019, **150**, 164103.
26 C. Wang, W. Huang and J.-L. Liao, *J. Phys. Chem. B*, 2015, **119**, 3720–3726.
27 J. Akola and R. O. Jones, *J. Phys. Chem. B*, 2003, **107**, 11774–11783.

28 P. Janoš, T. Trnka, S. Kozmon, I. Tvaroška and J. Koča, *J. Chem. Theory Comput.*, 2016, **12**, 6062–6076.
29 T. Trnka, S. Kozmon, I. Tvaroška and J. Koča, *PLOS Computational Biology*, 2015, **11**, e1004061.
30 A. Behn, P. M. Zimmerman, A. T. Bell and M. Head-Gordon, *J. Chem. Theory Comput.*, 2011, **7**, 4019–4025.
31 M. Jafari and P. M. Zimmerman, *J. Comput. Chem.*, 2017, **38**, 645–658.
32 P. M. Zimmerman, *The Journal of Chemical Physics*, 2013, **138**, 184102.
33 P. Zimmerman, *J. Chem. Theory Comput.*, 2013, **9**, 3043–3050.
34 A. M. Rozza, M. Papp, N. R. McFarlane, J. N. Harvey and J. Oláh, *Chemistry – A European Journal*, 2022, **28**, e202200930.
35 N. Lawan, K. E. Ranaghan, F. R. Manby and A. J. Mulholland, *Chemical Physics Letters*, 2014, **608**, 380–385.
36 P. M. Zimmerman, *J. Comput. Chem.*, 2015, **36**, 601–611.
37 J. Baker, A. Kessi and B. Delley, *The Journal of Chemical Physics*, 1996, **105**, 192–212.
38 J. Baker, *J. Comput. Chem.*, 1997, **18**, 1079–1095.
39 J. Baker, D. Kinghorn and P. Pulay, *The Journal of Chemical Physics*, 1999, **110**, 4986–4991.
40 S. R. Billeter, A. J. Turner and W. Thiel, *Phys. Chem. Chem. Phys.*, 2000, **2**, 2177–2186.
41 H. B. Schlegel, *Int. J. Quantum Chem.*, 1992, **44**, 243–252.
42 E. F. Koslover and D. J. Wales, *The Journal of Chemical Physics*, 2007, **127**, 234105.
43 J. N. Harvey, *Faraday Disc.*, 2004, **127**, 165.
44 R. B. Murphy, D. M. Philipp and R. A. Friesner, *Journal of Computational Chemistry*, 2000, **21**, 1442–1457.
45 T. Vreven, K. Morokuma, Ö. Farkas, H. B. Schlegel and M. J. Frisch, *J. Comput. Chem.*, 2003, **24**, 760–769.
46 G. S. Hammond, *J. Am. Chem. Soc.*, 1955, **77**, 334–338.
47 B. Peters, A. Heyden, A. T. Bell and A. Chakraborty, *The Journal of Chemical Physics*, 2004, **120**, 7877–7886.
48 X. Zhang, X. Zhang and T. C. Bruice, *Biochemistry*, 2005, **44**, 10443–10448.
49 C. C. Galopin, *Tett. Lett.*, 1996, **37**, 8675–8678.
50 A. L. Lamb, *Biochemistry*, 2011, **50**, 7476–7483.
51 N. A. Khanjin, J. P. Snyder and F. M. Menger, *J. Am. Chem. Soc.*, 1999, **121**, 11831–11846.
52 P. D. Lyne, A. J. Mulholland and W. G. Richards, *J. Am. Chem. Soc.*, 1995, **117**, 11345–11350.
53 M. M. Davidson, I. R. Gould and I. H. Hillier, *Journal of the Chemical Society, Perkin Transactions 2*, 1996, 525–532.
54 R. J. Hall, S. A. Hindle, N. A. Burton and I. H. Hillier, *J. Comput. Chem.*, 2000, **21**, 1433–1441.
55 S. Martí, J. Andrés, V. Moliner, E. Silla, I. Tuñón, J. Bertrán and M. J. Field, *J. Am. Chem. Soc.*, 2001, **123**, 1709–1712.
56 Y. S. Lee, S. E. Worthington, M. Krauss and B. R. Brooks, *J. Phys. Chem. B*, 2002, **106**, 12059–12065.

57 K. E. Ranaghan, L. Ridder, B. Szefczyk, W. A. Sokalski, J. C. Hermann and A. J. Mulholland, *Molecular Physics*, 2003, **101**, 2695–2714.
58 K. E. Ranaghan, L. Ridder, B. Szefczyk, W. A. Sokalski, J. C. Hermann and A. J. Mulholland, *Org. Biomol. Chem.*, 2004, **2**, 968.
59 F. Claeyssens, K. E. Ranaghan, F. R. Manby, J. N. Harvey and A. J. Mulholland, *Chem. Commun.*, 2005, 5068.
60 F. Claeyssens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel and H.-J. Werner, *Angew. Chem. Int. Ed.*, 2006, **45**, 6856–6859.
61 F. Claeyssens, K. E. Ranaghan, N. Lawan, S. J. Macrae, F. R. Manby, J. N. Harvey and A. J. Mulholland, *Org. Biomol. Chem.*, 2011, **9**, 1578.
62 X. Pan, R. Van, E. Epifanovsky, J. Liu, J. Pu, K. Nam and Y. Shao, *J. Phys. Chem. B*, 2022, **126**, 4226–4235.
63 S. A. Siddiqui and K. D. Dubey, *Physical Chemistry Chemical Physics*, 2022, **24**, 1974–1981.
64 P. Kast, M. Asif-Ullah and D. Hilvert, *Tetrahedron Letters*, 1996, **37**, 2691–2694.
65 S. Hur and T. C. Bruice, *J. Am. Chem. Soc.*, 2003, **125**, 10540–10542.
66 S. Hur and T. C. Bruice, *J. Am. Chem. Soc.*, 2003, **125**, 1472–1473.
67 M. Štrajbl, A. Shurki, M. Kato and A. Warshel, *J. Am. Chem. Soc.*, 2003, **125**, 10228–10237.
68 Y. M. Chook, H. Ke and W. N. Lipscomb, *Proc Natl Acad Sci U S A*, 1993, **90**, 8600–8603.
69 Y. M. Chook, J. V. Gray, H. Ke and W. N. Lipscomb, *J Mol Biol*, 1994, **240**, 476–500.
70 P. A. Bartlett and C. R. Johnson, *J. Am. Chem. Soc.*, 1985, **107**, 7792–7793.
71 A. D. Becke, *The Journal of Chemical Physics*, 1993, **98**, 5648–5652.
72 C. Lee, W. Yang and R. G. Parr, *Phys Rev B Condens Matter*, 1988, **37**, 785–789.
73 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.
74 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.
75 R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, *J. Chem. Phys.*, 1980, **72**, 650–654.
76 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *The Journal of Chemical Physics*, 2010, **132**, 154104.
77 S. Grimme, S. Ehrlich and L. Goerigk, *Journal of Computational Chemistry*, 2011, **32**, 1456–1465.
78 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand,

K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, Gaussian16 Gaussian, Inc., Wallingford CT 2016.

79 K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. MacKerell, *J Comput Chem*, 2010, **31**, 671–690.

80 M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, *J. Chem. Theory Comput.*, 2011, **7**, 525–537.

81 P. Mark and L. Nilsson, *J. Phys. Chem. A*, 2001, **105**, 9954–9960.

82 G. A. Ross, A. S. Rustenburg, P. B. Grinaway, J. Fass and J. D. Chodera, *J Phys Chem B*, 2018, **122**, 5466–5486.

83 S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845–854.

84 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1–2**, 19–25.

85 J. Huang and A. D. MacKerell, *J Comput Chem*, 2013, **34**, 2135–2145.

86 J. A. Rackers, Z. Wang, C. Lu, M. L. Laury, L. Lagardère, M. J. Schnieders, J.-P. Piquemal, P. Ren and J. W. Ponder, *J. Chem. Theory Comput.*, 2018, **14**, 5273–5289.

87 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian09 Gaussian, Inc., Wallingford CT 2009.

88 E. Martínez-Núñez, G. L. Barnes, D. R. Glowacki, S. Kopec, D. Peláez, A. Rodríguez, R. Rodríguez-Fernández, R. J. Shannon, J. J. P. Stewart, P. G. Tahoces and S. A. Vazquez, *J Comput Chem*, 2021, **42**, 2036–2048.

89 P. M. Zimmerman, *J. Comput. Chem.*, 2013, **34**, 1385–1392.

90 A. Nakao, Y. Harabuchi, S. Maeda and K. Tsuda, *Phys. Chem. Chem. Phys.*, 2022, **24**, 10305–10310.

91 A. M. Cooper and J. Kästner, *ChemPhysChem*, 2014, **15**, 3264–3269.

92 H. Kang and M. Zheng, *Computational and Theoretical Chemistry*, 2021, **1204**, 113399.

93 Q. Cui, T. Pal and L. Xie, *J. Phys. Chem. B*, 2021, **125**, 689–702.

94 U. Ryde, *J. Chem. Theory Comput.*, 2017, **13**, 5745–5752.

95 R. A. Groeneveld and G. Meeden, *Journal of the Royal Statistical Society. Series D (The Statistician)*, 1984, **33**, 391–399.

96 D. N. Joanes and C. A. Gill, *Journal of the Royal Statistical Society. Series D (The Statistician)*, 1998, **47**, 183–189.
97 K. P. Balanda and H. L. MacGillivray, *The American Statistician*, 1988, **42**, 111–119.
98 E. Taskesen, distfit is a python library for probability density fitting. (version 1.4.0) https://erdogant.github.io/distfit 2020.
99 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, *Nature Methods*, 2020, **17**, 261–272.
100 M. W. Y. Szeto, J. I. Mujika, J. Zurek, A. J. Mulholland and J. N. Harvey, *Journal of Molecular Structure: THEOCHEM*, 2009, **898**, 106–114.
101 S. Craciun and E. P. Balskus, *Proc Natl Acad Sci U S A*, 2012, **109**, 21307–21312.
102 Z. Yang, R. Mehmood, M. Wang, H. W. Qi, A. H. Steeves and H. J. Kulik, *React. Chem. Eng.*, 2019, **4**, 298–315.

6 Summary and Outlook

"...all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind..."

- George Box, Empirical Model-Building and Response Surfaces, 1987

This thesis has explored the use of multiscale modelling techniques for the study of several biochemical processes. The processes have ranged from intracellular gas signalling to enzymes which catalyse steps in metabolic pathways. In every case, due to the complex interplay between biochemical processes, appropriate models at various time- and size-scales have been selected. As clearly portrayed by George Box, a model is inherently an approximation, and the problem is complicated further given that the selection of an appropriate model is a non-trivial task. Nevertheless, the work presented in this thesis has always been conscious of the model chosen, such that, speaking metaphorically, *the dog wags the tail, and that the tail does not wag the dog*.

Given the variety of the biochemical processes studied in this thesis, there is not one overarching conclusion, but rather a set of conclusions and recommendations relating to the multiscale modelling of biochemical processes in general.

A first conclusion relates to whether static or dynamic techniques are more suitable for the study of these processes. In Chapters 3 and 4, the static QM cluster modelling approach was used, where heme groups were modelled accurately with DLPNO-CCSD(T_x) calculations. Calculations on this level of theory have the benefit of providing accurate electronic energies for the heme derivatives

studied at a modest computational cost, though they remain generally computationally demanding. However, increasing computational power coupled with further development into accurate and cost-effective methods such as DLPNO-CCSD(T_x) is continuously reducing this expense.

In Chapters 3 and 5, static hybrid QM/MM methods were used. These approaches have the benefit of including the enzyme environment, which has been shown to be critical in obtaining an accurate description of reactivity.[1] Since the computational cost of QM/MM methods tends to scale linearly with the QM method used, the question emerges of whether one should always use a QM/MM model to describe biochemical processes? In this thesis, this question has not been explored in detail, but some conclusions can be drawn in any case.

With QM/MM methods, there is an associated labour-cost which is not immediately apparent. This in addition to the difficult and sensitive system setup can make the benefit of QM/MM models quickly vanish. In this context, it is worth highlighting that for Chapters 3 and 4, the QM cluster model was used. In Chapter 3, the results obtained agreed favourably with previous studies on related systems, indicating that the QM cluster model can often be a cost-effective choice. For the case of Chapter 4, however, while the predicted free energies did agree quite well with previous computations on related systems, they not agree particularly well with the experimental free energies. This discrepancy could be a compounding result of many factors, but it may be that a QM/MM model would have been more suitable to capture the (de)stabilising effect of the amino acid environment. In any case, hybrid QM/MM models still have their place, particularly in relation to the next conclusion related to calculating reaction pathways.

Returning to the question of whether dynamic or static techniques are more appropriate, in all results chapters, MD simulations were used. The MD method is extremely powerful and is very often the

most appropriate way to describe dynamic processes in biochemical systems. In Chapter 3, such simulations were used to investigate small gas molecule diffusion, and to investigate protein conformational changes following the binding of these small gas molecules. Similarly, in Chapter 4, larger scale protein conformational changes were studied using extended MD simulations, where it was shown that conformational coupling of sites as far as 25 Å away from one another could be studied.

In Chapters 3 and 5, MD simulations were also applied in a preparatory way, so as to provide an ensemble of structures to be used in QM/MM calculations. This is a common approach to describe reactivity of a large system in a statistically satisfactory way. The key conclusion is that there is an inherent complementarity between static and dynamic approaches in studying biochemical processes. Certain facets of a given process require a static model, and others require a dynamic model.

A second conclusion is not unrelated to the first one, but is with specific reference to calculating reaction pathways with hybrid QM/MM. In the generation of biochemical reaction pathways, inclusion of the amino acid environment is critically important due to the (de)stabilising effect of amino acid residues. In these cases, a QM/MM approach is often necessitated as generating reaction pathways with a sufficiently large QM cluster model could be computational unfeasible. Accordingly, in Chapter 5, a novel algorithm for calculating reaction pathways was presented: the SE-GSM with hybrid DLCs in the QM/MM framework. In this chapter, this technique was found to be effective for a case study of the Claisen rearrangement catalysed by chorismate mutase.

The algorithm presented could be classified as static, where interpolation across a reaction coordinate generates a series of geometries along a reaction pathway. However, since biochemical environments are constantly changing, if one wishes to obtain an

accurate free energy of reaction, some degree of conformational sampling is required.

Certain dynamic techniques, such as umbrella sampling,[2] do enable dynamic sampling of conformational space at the QM-level so that accurate free energies of reaction can be obtained. However, these techniques are computationally rather costly. On the other hand, a particular challenge related to calculating static reaction pathways is that individual pathways do not address sampling in any way. As explored by a statistical analysis in Chapter 5, the use of many starting geometries can in part solve this problem, enabling an approximation to the free energy of reaction to be obtained.

In practice, reaction pathways are calculated *via* both static and dynamic methodologies. Due to the increased conformational sampling, dynamic techniques are, in principle, a more accurate way to calculate the free energy of reaction. However, static techniques enable the computationally inexpensive generation of a set of discrete structures tracing chemical reactivity. As such, static techniques place themselves as a qualitative tool used to study competing pathways. Moreover, the findings in Chapter 5 suggest a positive future for quantitative static QM/MM calculations, implying that less sampling is required to obtain converged free energies of reaction than was previously suggested.

As a final note on QM/MM, an increasing field of interest in QM/MM studies is to use polarisable force fields for the MM region,[3] allowing mutual polarisation of the QM and MM regions. Conceptually, such an idea makes biochemical sense given the (de)stabilising nature of amino acids in the context of biochemical reactivity. These types of force field are expected to have increased use in QM/MM studies in the coming years, but challenges remain with polarisable forcefield portability.

A final and rather general conclusion relates to whether experimental or computational techniques are most suitable for the study of biochemical processes. Logically, wet-lab techniques should be used since biochemical processes are naturally occurring. However, computational techniques usually enable faster and more flexible modelling. The reality is that both techniques are interconnected for the study of biochemical processes. Computational modelling often relies on experimentally-derived 3-dimensional structural data of the systems under study. With the advent of the popular AlphaFold[4–6] this requirement is waning, however since the models with which AlphaFold is trained on are themselves experimentally-derived, it would be naïve to discount the value of experimental structures. On the other side of the coin, computational techniques are very often used to supplement experimental data, with a particular emphasis on CADD.

In summary, this thesis has exemplified the use of multiscale modelling techniques in the modelling of biochemical processes. This was achieved by way of application of existing techniques, and through development of new ones. Choosing exactly how a process will be modelled remains a central problem, but with constant development of novel and highly-specific programs and techniques, the future of biochemical modelling is positive. Ultimately, biochemistry can be analysed in detail *via* computational modelling in a complementary way to experiment.

"In general we look for a new law by the following process. First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right. Then we compare the result of the computation to nature, with experiment or experience, compare it directly with observation, to see if it works. If it disagrees with experiment it is wrong. In that simple statement is the key to science. It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is – if it disagrees with experiment it is wrong. That is all there is to it."

- Richard P. Feynman, Seeking New Laws, 1964

6.1 References

1 V. Guallar and F. H. Wallrapp, *Biophys. Chem.*, 2010, **149**, 1–11.

2 J. Kästner, *WIREs Comput. Mol. Sci.*, 2011, **1**, 932–942.

3 R. P. Magalhães, H. S. Fernandes and S. F. Sousa, *Isr. J. Chem.*, 2020, **60**, 655–666.

4 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, 1–3.

5 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

6 M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis and S. Velankar, *Nucleic Acids Res.*, 2022, **50**, D439–D444.

7 List of Publications

7.1 Journal Articles

[1] A. M. Rozza, M. Papp, **N. R. McFarlane**, J. N. Harvey, J. Oláh, The Mechanism of Biochemical NO-Sensing: Insights from Computational Chemistry, *Chem. Eur. J.* **2022**, 28, e202200930

DOI: https://doi.org/10.1002/chem.202200930

[2] **N. R. McFarlane**, J. N. Harvey, Exploration of Biochemical Reactivity with a QM/MM Growing String Method, *Phys. Chem. Chem. Phys.*,**2024**, 26, 5999-6007

DOI: https://doi.org/10.1039/D3CP05772K

[3] J. Čivić, **N. R. McFarlane**, J. Masschelein, J. N. Harvey, Exploring the Selectivity of Cytochrome P450 for Enhanced Novel Anticancer Agent Synthesis, *Faraday Discuss.*, **2024**

DOI: https://doi.org/10.1039/D4FD00004H

[4] **N. R. McFarlane**, J. Gui, J. Oláh, J. N. Harvey, Gaseous Inhibition of the Transsulfuration Pathway by Cystathionine β-Synthase, *Phys. Chem. Chem. Phys*, **2024**, 26, 16579-16588

DOI: https://doi.org/10.1039/D4CP01321B

7.2 Conferences

CECAM Workshop, Lecco, Italy, June 2022, Exploring biomolecular reactions with hybrid QM/MM single-ended growing string method

ChemCYS, Blankenberge, Belgium, October 2022, unbiased approach to exploring biomolecular reactions with hybrid QM/MM

EuChemS CompChem, Thessaloniki, Greece, August 2023 Gaseous Inhibition of the Transsulfuration Pathway by Cystathionine β-synthase: CO, or NO?

Faraday Discussions Biocatalysis, London, United Kingdom, May 2024

8. Appendices

A. Supporting Information for Chapter 3

MD Simulations

Table A1. *The average values of mainchain (backbone plus the carbonyl oxygen) RMSD calculated from the MD trajectories of the* ***A, B, D*,*** *and* ***D*** *simulations with respect to the zero ns structure.*

H-NOX	A	B	D*	D
Av. RMSD [nm]	0.120	0.150	0.164	0.135

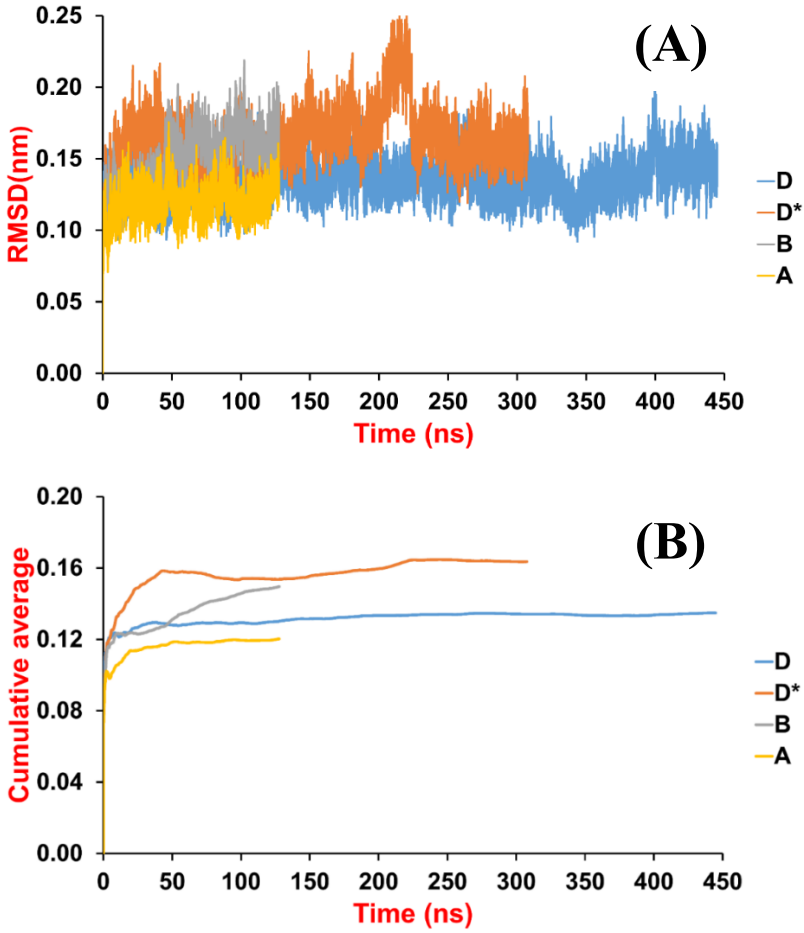

Figure A1: ***(A)*** *Mainchain (backbone plus the carbonyl oxygen) RMSD calculated from MD trajectories of* ***A, B, D****, *and* ***D*** *with respect to the zero ns structure.* ***(B)*** *cumulative average of the mainchain (backbone plus the carbonyl oxygen) RMSD calculated from MD trajectories of* ***A, B, D****, *and* ***D*** *with respect to the zero ns structure.*

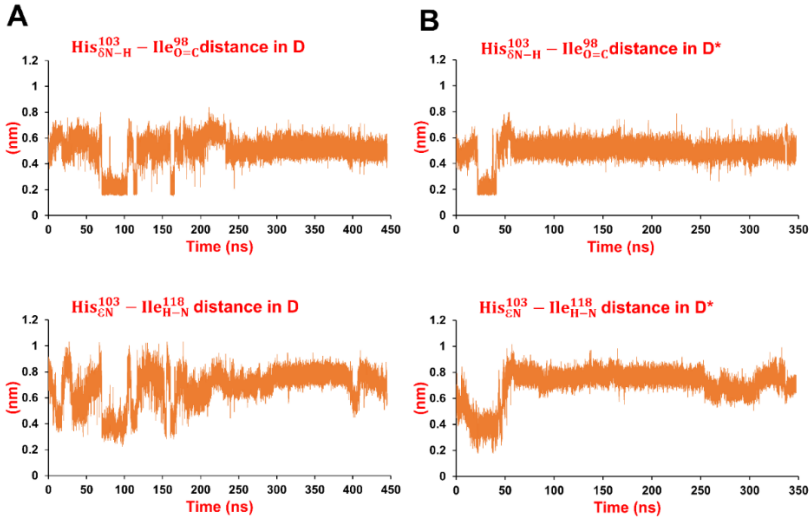

Figure A2. *Time-dependent plots showing distance between hydrogen bonding partners in nanometres during the inward conformation in both (A)* ***D*** *and (B)* ***D****. The engaged partners are His103 with Ile98 & His103 with Ile118.*

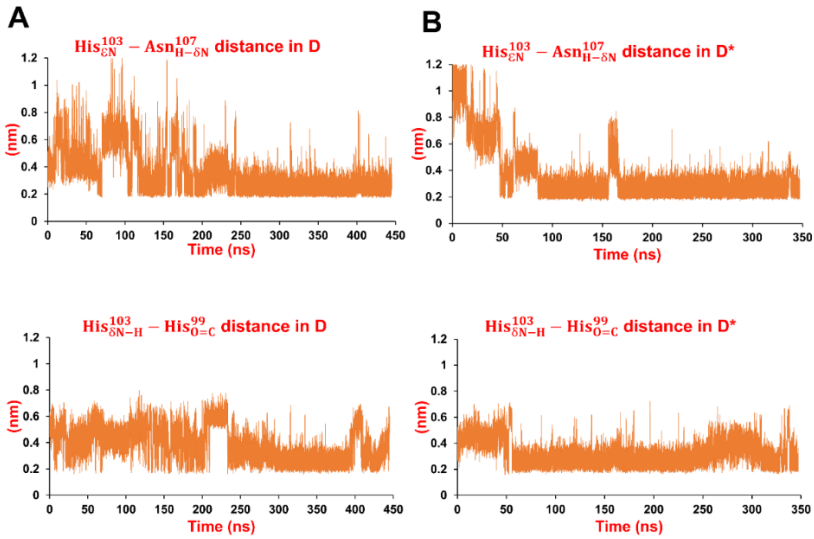

Figure A3. *Time-dependent plots showing distance between hydrogen bonding partners in nanometres during the outward conformation in both (A)* ***D*** *and (B)* ***D****. The engaged partners are His103 with Asn107 & His103 with His99.*

QM/MM Calculations

B to D Reaction Pathways*

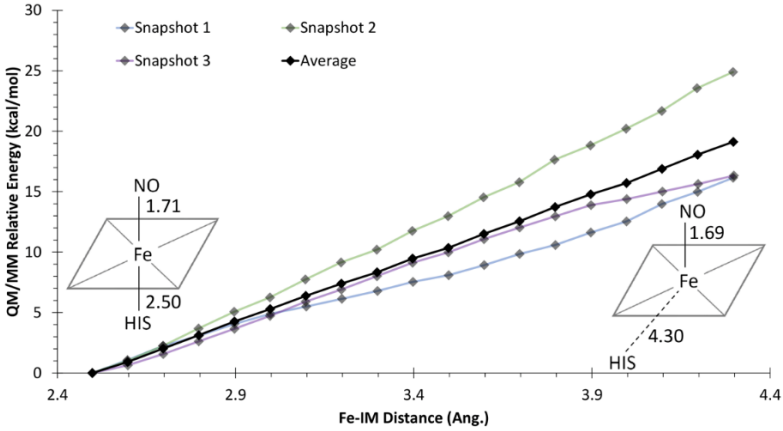

Figure A4: *Energy profile for BP86-D3 QM/MM modelling of the transition from* ***B*** *to* ***D****. This represents the change from 6-coordinate heme (****B****) to 5-coordinate heme (****D****). Insets (a) and (b) show* ***B*** *and* ***D****, respectively. The reaction shows the Fe-His distance is elongating from ~2.50 Å to ~4.30 Å. Meanwhile, the Fe-NO distance remains relatively unchanged at ~1.70 Å.*

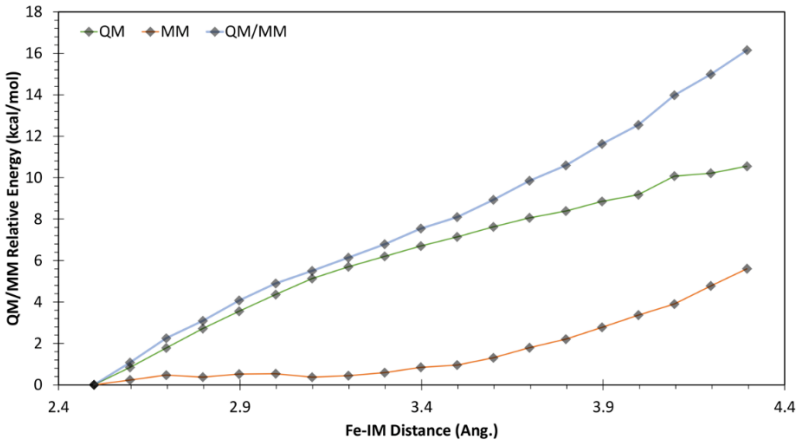

Figure A5: *Energy profile for BP86-D3 QM/MM modelling of the transition from* ***B*** *to* ***D**** *for snapshot 1. This represents the change from 6-coordinate heme (****B****)*

*to 5-coordinate heme (**D***). The QM, MM and QM/MM energy contributions are explicitly shown.*

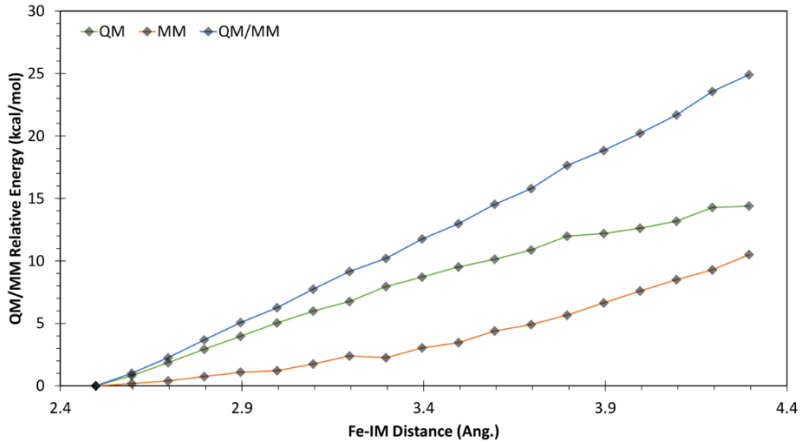

Figure A6: *Energy profile for BP86-D3 QM/MM modelling of the transition from **B** to **D*** for snapshot 2. This represents the change from 6-coordinate heme (**B**) to 5-coordinate heme (**D***). The QM, MM and QM/MM energy contributions are explicitly shown.*

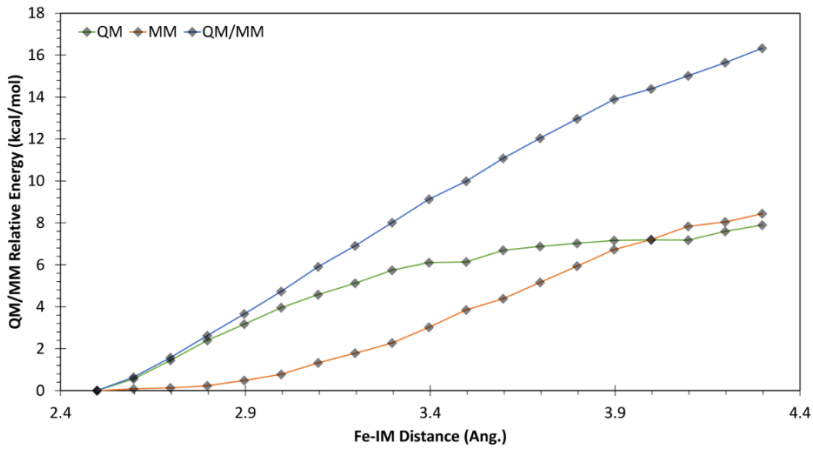

Figure A7: *Energy profile for BP86-D3 QM/MM modelling of the transition from **B** to **D*** for snapshot 3. This represents the change from 6-coordinate heme (**B**) to 5-coordinate heme (**D***). The QM, MM and QM/MM energy contributions are explicitly shown.*

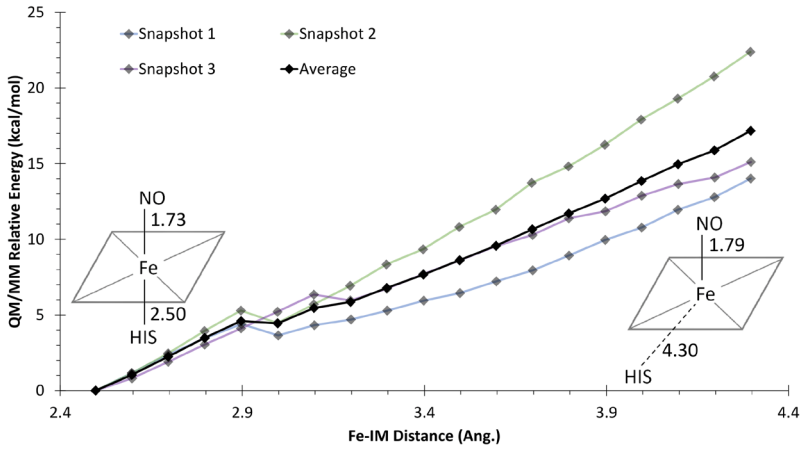

Figure A8: *Energy profile for B3LYP-D3 QM/MM modelling of the transition from **B** to **D***. This represents the change from 6-coordinate heme (**B**) to 5-coordinate heme (**D***). Insets (a) and (b) show **B** and **D***, respectively. The reaction shows the Fe-His distance is elongating from ~2.50 Å to ~4.30 Å. Meanwhile, the Fe-NO distance changes minimally from ~1.73 Å to ~1.79 Å.*

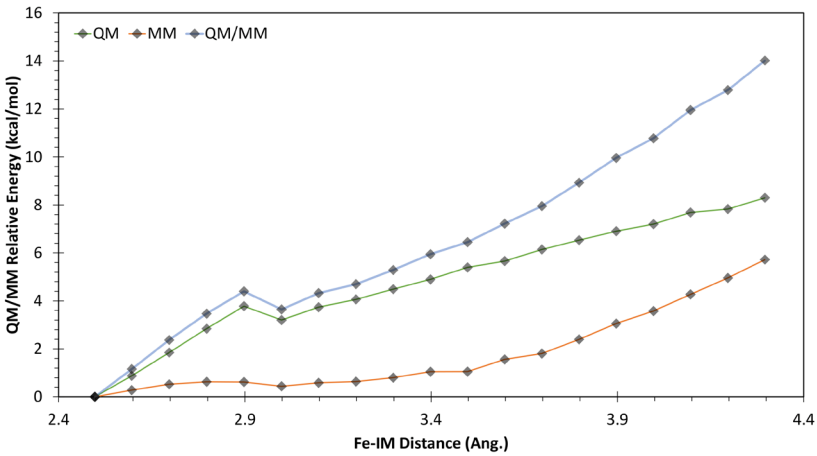

Figure A9: *Energy profile for B3LYP-D3 QM/MM modelling of the transition from **B** to **D*** for snapshot 1. This represents the change from 6-coordinate heme (**B**) to 5-coordinate heme (**D***). The QM, MM and QM/MM energy contributions are explicitly shown.*

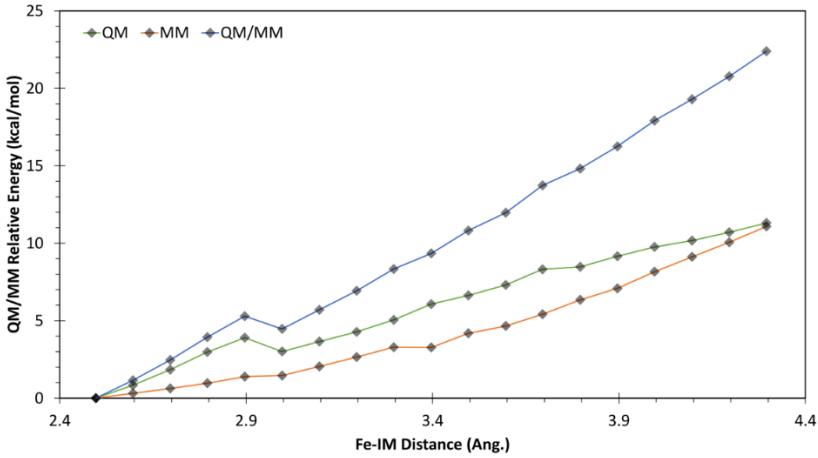

Figure A10: *Energy profile for B3LYP-D3 QM/MM modelling of the transition from **B** to **D*** for snapshot 3. This represents the change from 6-coordinate heme (**B**) to 5-coordinate heme (**D***). The QM, MM and QM/MM energy contributions are explicitly shown.*

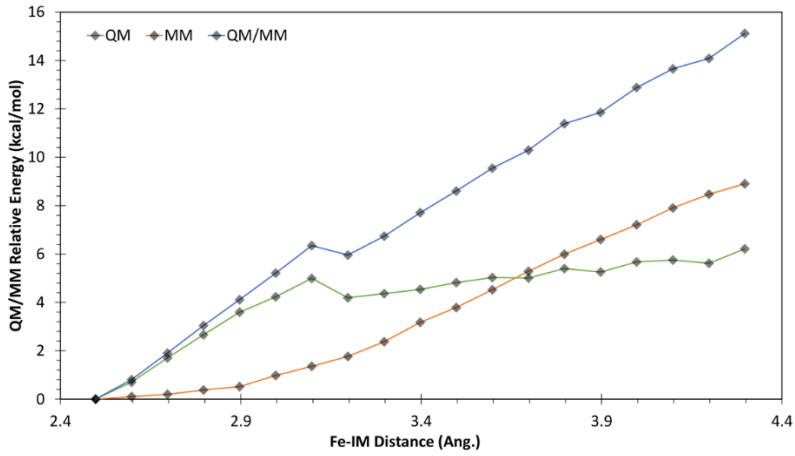

Figure A11: *Energy profile for B3LYP-D3 QM/MM modelling of the transition from **B** to **D*** for snapshot 2. This represents the change from 6-coordinate heme (**B**) to 5-coordinate heme (**D***). The QM, MM and QM/MM energy contributions are explicitly shown.*

Table A2: *The bond lengths for each of the insets found in* ***Figure A12****. For comparison between levels of theory, the average bond lengths are also given.*

Bond	Bond Length (Angstroms)							
	B3LYP				*BP86*			
	(a)	(b)	(c)	Average	(d)	(e)	(f)	Average
NO_{distal} --- Fe	1.73	1.72	1.74	**1.73**	1.71	1.70	1.71	**1.71**
Imidazole --- Fe	2.50	2.50	2.50	**2.50**	2.50	2.50	2.50	**2.50**

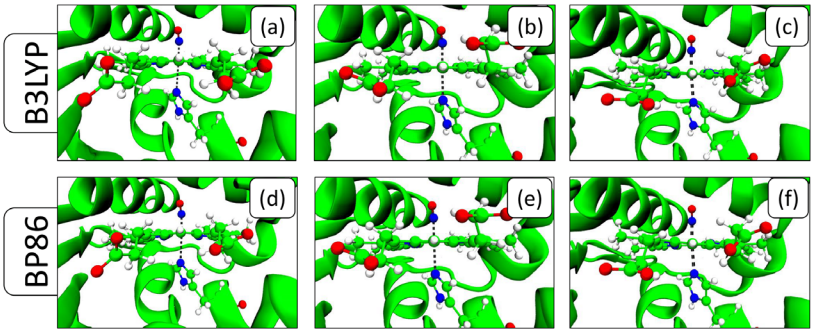

Figure A12: *Structures of species* ***B*** *for each snapshot for both levels of theory in the study of the pathway* ***B*** *to* ***D****.*

Table A3: *The bond lengths for each of the insets found in* ***Figure A13****. For comparison between levels of theory, the average bond lengths are also given.*

Bond	Bond Length (Angstroms)							
	B3LYP				*BP86*			
	(a)	(b)	(c)	Average	(d)	(e)	(f)	Average
NO_{distal} --- Fe	1.79	1.79	1.79	**1.79**	1.69	1.68	1.69	**1.69**
Imidazole --- Fe	4.30	4.29	4.30	**4.30**	4.30	4.29	4.30	**4.30**

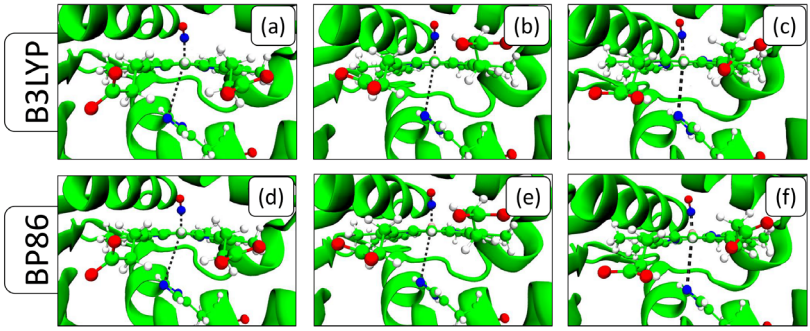

Figure A13: *Structures of species* ***D**** *for each snapshot for both levels of theory in the study of the pathway* ***B*** *to* ***D****.*

B to C Reaction Pathways

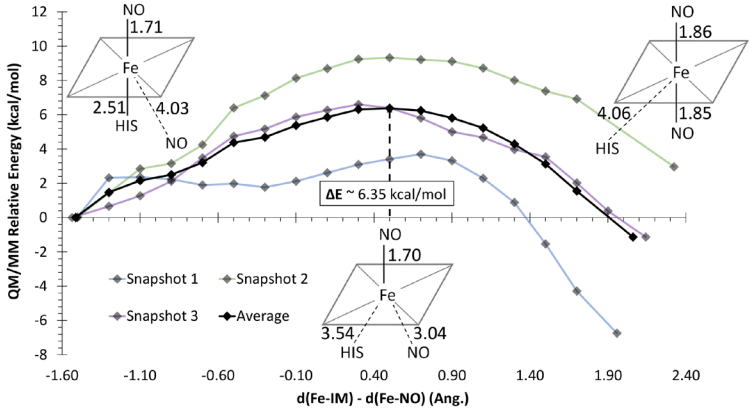

Figure A14: *Energy profile for BP86-D3 QM/MM modelling of the transition from **B** to **C**. This represents the change from 6-coordinate heme (**B**) to 6-coordinate heme (**C**), with switching of proximal ligand. Insets (a) and (b) show **B** and **C**, respectively. The reaction shows a low energy barrier of ~ 6.35 kcal/mol, with the Fe-NO(prox.) distance shortening from ~4.00 Å to ~1.80 Å. Meanwhile, the Fe-histidine distance elongates from ~2.50 Å to ~4.10 Å.*

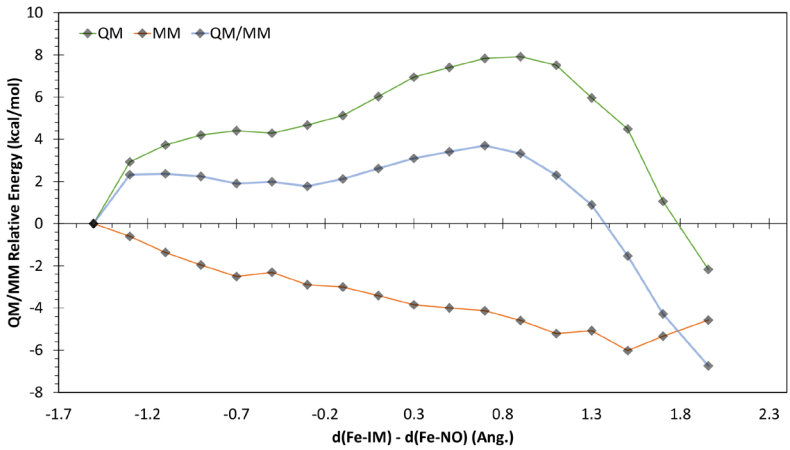

Figure A15: *Energy profile for BP86-D3 QM/MM modelling of the transition from **B** to **C** for snapshot 1. This represents the change from 6-coordinate heme (**B**) to 6-coordinate heme (**C**), with switching of proximal ligand. The QM, MM and QM/MM energy contributions are explicitly shown.*

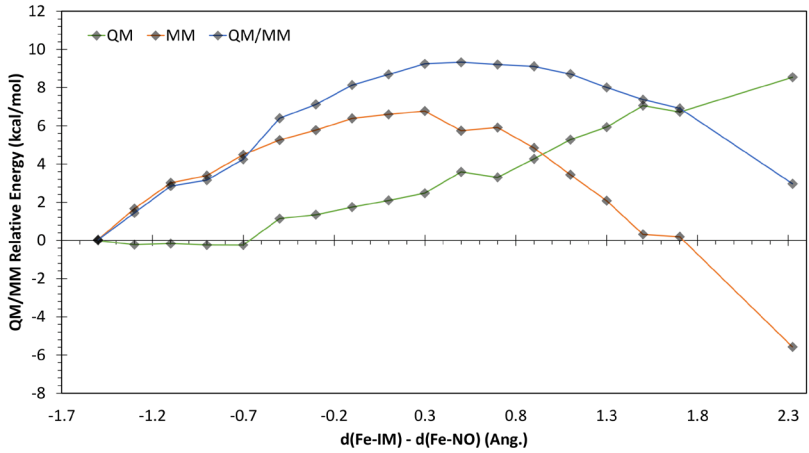

Figure A16: *Energy profile for BP86-D3 QM/MM modelling of the transition from **B** to **C** for snapshot 2. This represents the change from 6-coordinate heme (**B**) to 6-coordinate heme (**C**), with switching of proximal ligand. The QM, MM and QM/MM energy contributions are explicitly shown.*

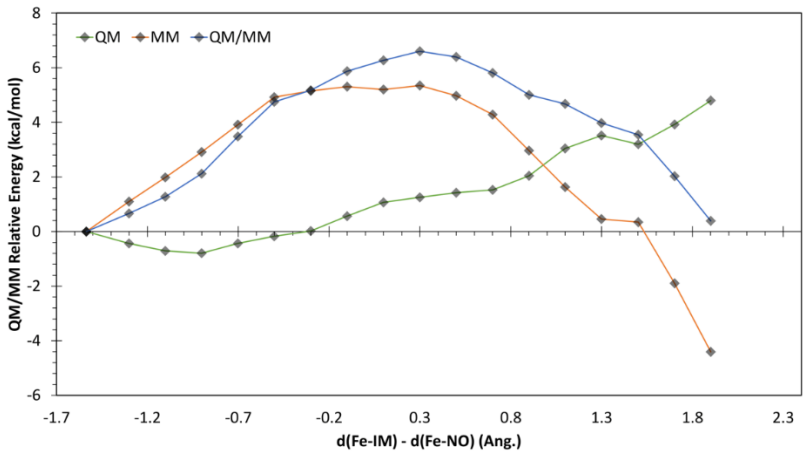

Figure A17: *Energy profile for BP86-D3 QM/MM modelling of the transition from **B** to **C** for snapshot 3. This represents the change from 6-coordinate heme (**B**) to 6-coordinate heme (**C**), with switching of proximal ligand. The QM, MM and QM/MM energy contributions are explicitly shown.*

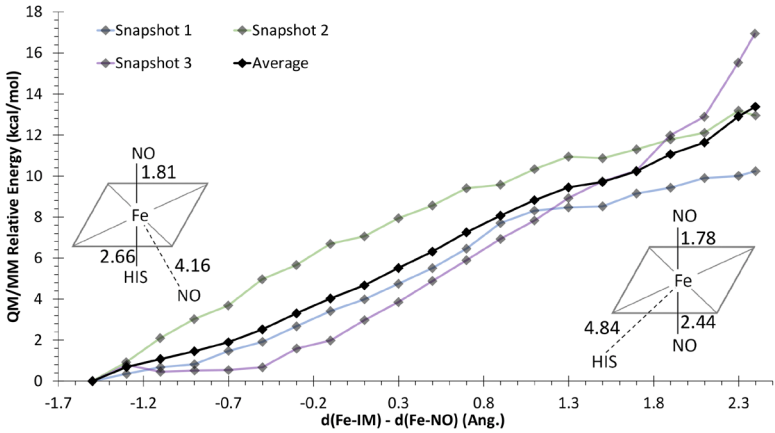

Figure A18: *Energy profile for B3LYP-D3 QM/MM modelling of the transition from **B** to **C**. This represents the change from 6-coordinate heme (**B**) to 6-coordinate heme (**C**), with switching of proximal ligand. Insets (a) and (b) show **B** and **C**, respectively. The reaction shows the Fe-NO(prox.) distance shortening from ~4.00 Å to ~2.45 Å. Meanwhile, the Fe-histidine distance elongates from ~2.50 Å to ~4.85 Å.*

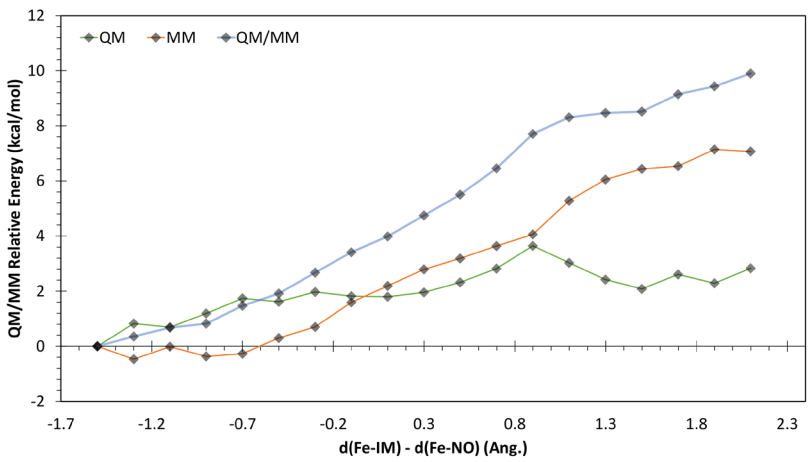

Figure A19: *Energy profile for B3LYP-D3 QM/MM modelling of the transition from **B** to **C** for snapshot 1. This represents the change from 6-coordinate heme (**B**) to 6-coordinate heme (**C**), with switching of proximal ligand. The QM, MM and QM/MM energy contributions are explicitly shown.*

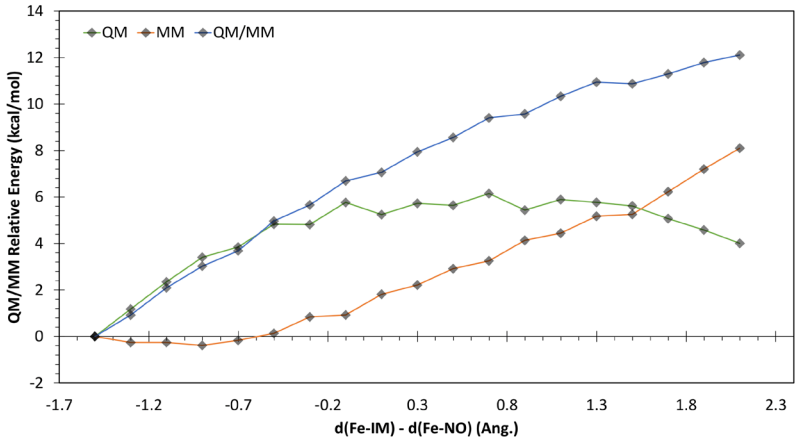

Figure A20: *Energy profile for B3LYP-D3 QM/MM modelling of the transition from **B** to **C** for snapshot 2. This represents the change from 6-coordinate heme (**B**) to 6-coordinate heme (**C**), with switching of proximal ligand. The QM, MM and QM/MM energy contributions are explicitly shown.*

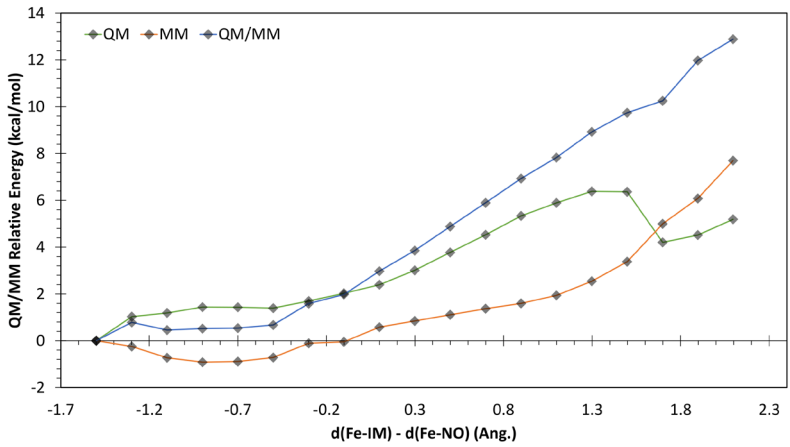

Figure A21: *Energy profile for B3LYP-D3 QM/MM modelling of the transition from **B** to **C** for snapshot 3. This represents the change from 6-coordinate heme (**B**) to 6-coordinate heme (**C**), with switching of proximal ligand. The QM, MM and QM/MM energy contributions are explicitly shown.*

Table A4: *The bond lengths for each of the insets found in* ***Figure A22****. For comparison between levels of theory, the average bond lengths are also given.*

Bond	Bond Length (Angstroms)							
	B3LYP				*BP86*			
	(a)	(b)	(c)	Average	(d)	(e)	(f)	Average
NO_{distal} --- Fe	1.84	1.74	1.85	**1.81**	1.72	1.71	1.71	**1.71**
$NO_{proximal}$ --- Fe	4.22	4.01	4.25	**4.16**	4.01	4.01	4.06	**4.03**
Imidazole --- Fe	2.72	2.51	2.74	**2.66**	2.51	2.51	2.52	**2.51**

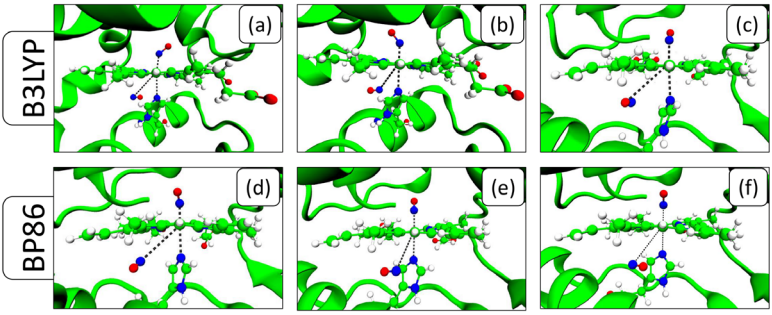

Figure A22: *Structures of species* ***B*** *with an additional NO atom added to the proximal pocket structures for each snapshot for both levels of theory in the study of the pathway* ***B*** *to* ***C****.*

Table A5: *The bond lengths for each of the insets found in* ***Figure A23****. For comparison between levels of theory, the average bond lengths are also given.*

Bond	Bond Length (Angstroms)							
	B3LYP				*BP86*			
	(a)	(b)	(c)	Average	(d)	(e)	(f)	Average
NO_{distal} --- Fe	1.81	1.72	1.8	**1.78**	1.91	1.82	1.84	**1.86**
$NO_{proximal}$ --- Fe	2.51	2.46	2.36	**2.44**	1.96	1.8	1.79	**1.85**
Imidazole --- Fe	4.91	4.85	4.75	**4.84**	4.13	4.13	3.93	**4.06**

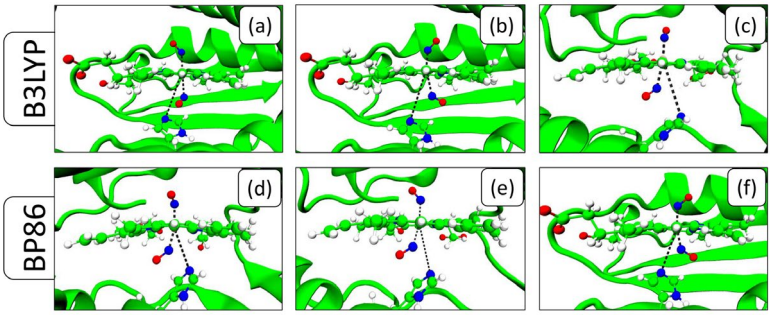

Figure A23: *Structures of species* ***C*** *structures for each snapshot for both levels of theory in the study of the pathway* ***B*** *to* ***C****.*

B. Supporting Information for Chapter 4

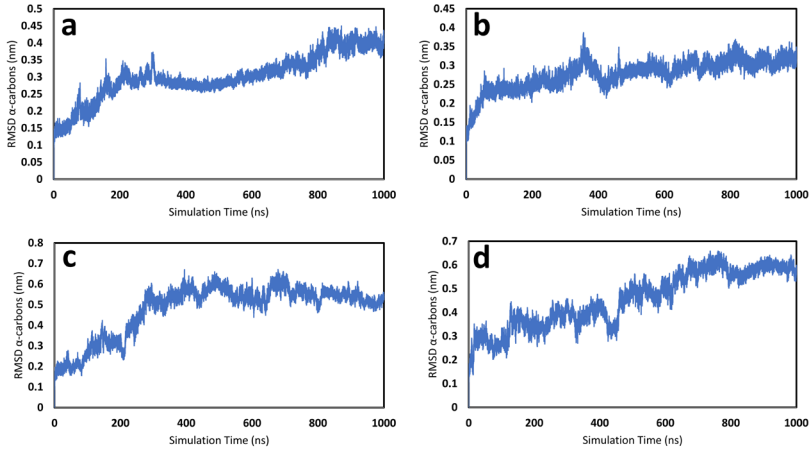

***Figure B1**: **a.** RMSD of the α-carbons averaged across both chains A and B for the first 1000 ns simulation of **A**; **b.** RMSD of the α-carbons averaged across both chains A and B for the second 1000 ns simulation of **A**; **c.** RMSD of the α-carbons averaged across both chains A and B for the first 1000 ns simulation of **C**; **d.** RMSD of the α-carbons averaged across both chains A and B for the second 1000 ns simulation of **C**.*

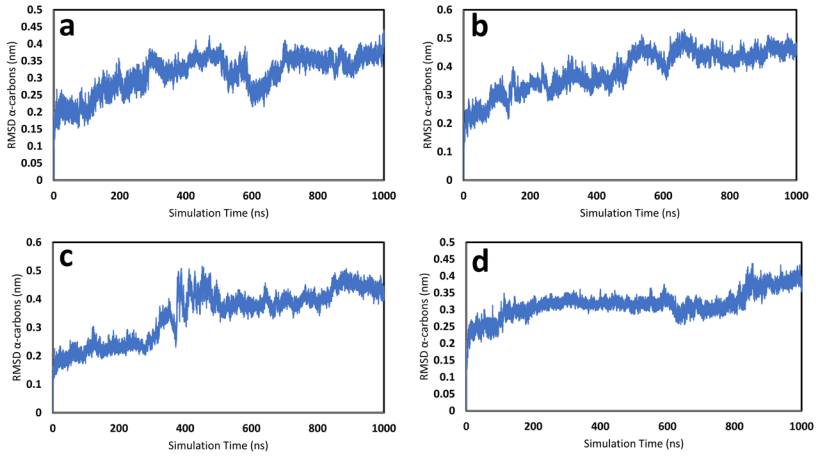

***Figure B2: a.** RMSD of the α-carbons averaged across both chains A and B for the first 1000 ns simulation of **E**; **b.** RMSD of the α-carbons averaged across both chains A and B for the second 1000 ns simulation of **E**; **c.** RMSD of the α-carbons averaged across both chains A and B for the first 1000 ns simulation of **F**; **d.** RMSD of the α-carbons averaged across both chains A and B for the second 1000 ns simulation of **F**.*

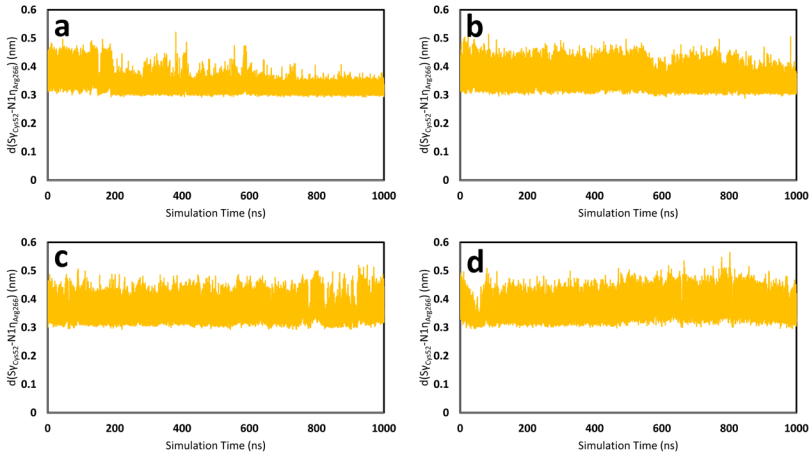

Figure B3: a. *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain A for the first 1000 ns simulation of* ***A; b.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain B for the first 1000 ns simulation of* ***A; c.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain A for the second 1000 ns simulation of* ***A; d.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain B for the second 1000 ns simulation of* ***A****.*

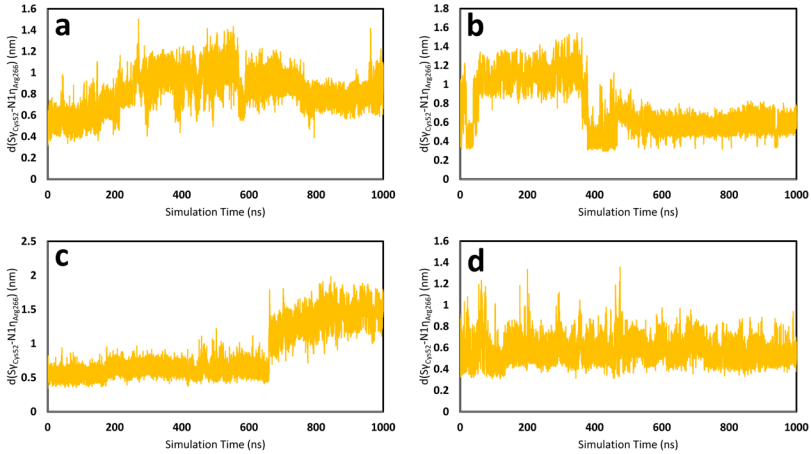

Figure B4: a. *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain A for the first 1000 ns simulation of* ***C; b.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain B for the first 1000 ns simulation of* ***C; c.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain A for the second 1000 ns simulation of* ***C; d.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain B for the second 1000 ns simulation of* ***C****.*

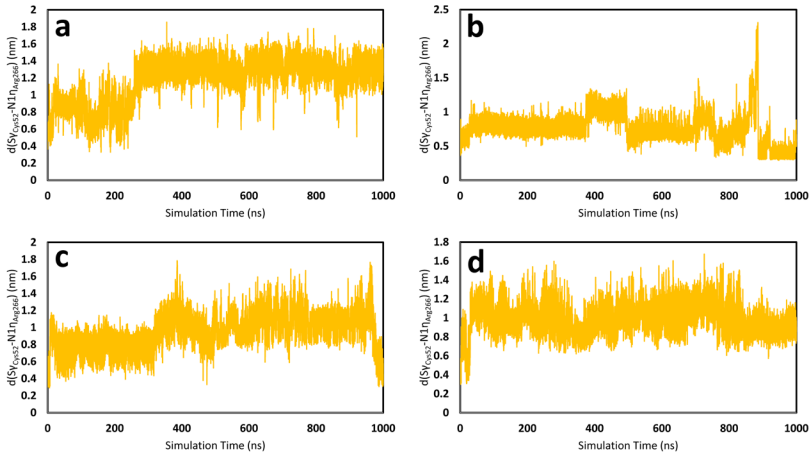

Figure B5: a. *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain A for the first 1000 ns simulation of* ***E; b.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain B for the first 1000 ns simulation of* ***E; c.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain A for the second 1000 ns simulation of* ***E; d.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain B for the second 1000 ns simulation of* ***E****.*

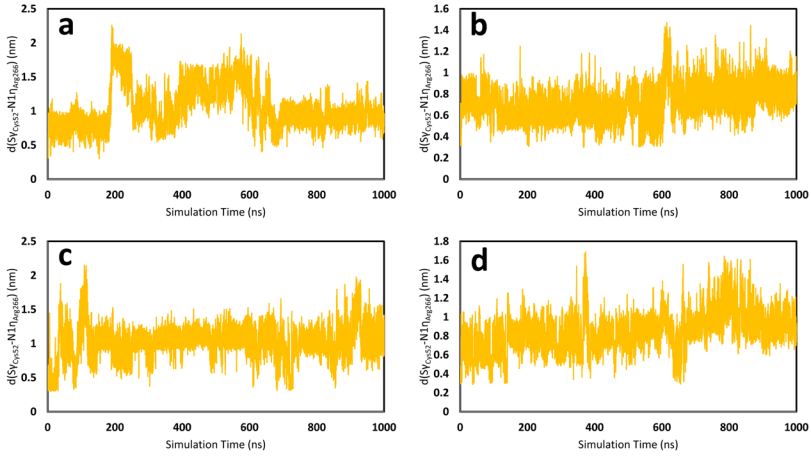

Figure B6: a. *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain A for the first 1000 ns simulation of* ***F; b.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain B for the first 1000 ns simulation of* ***F; c.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain A for the second 1000 ns simulation of* ***F; d.*** *variation of d(Sγ$_{Cys52}$-N1η$_{Arg266}$) for chain B for the second 1000 ns simulation of* ***F****.*

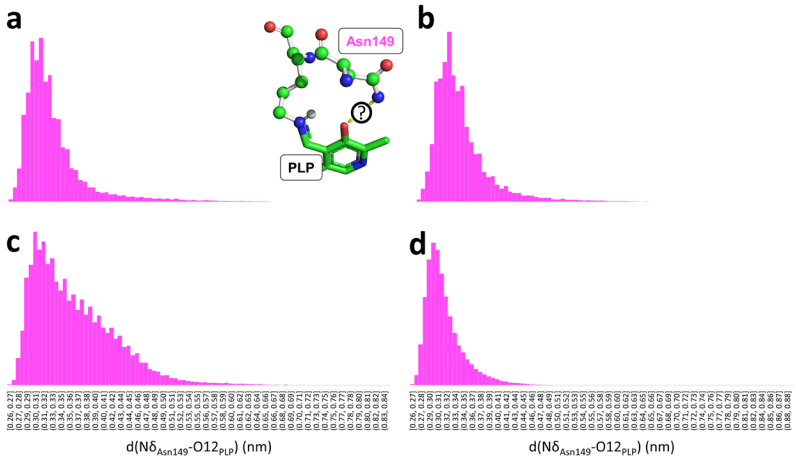

Figure B7: a. *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***A; b.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***A; c.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***A; d.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***A****. The figure inset visually shows d(Nδ$_{Asn149}$-O12$_{PLP}$) and how it relates to PLP and Asn149 in human CBS.*

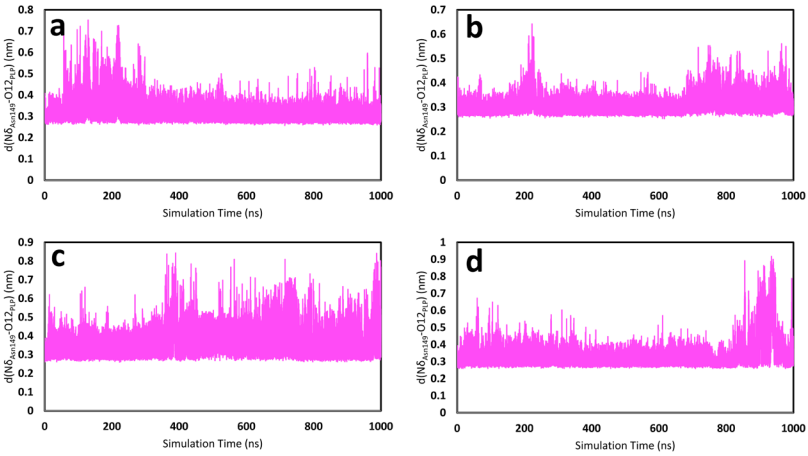

Figure B8: a. *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***A; b.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***A; c.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***A; d.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***A****.*

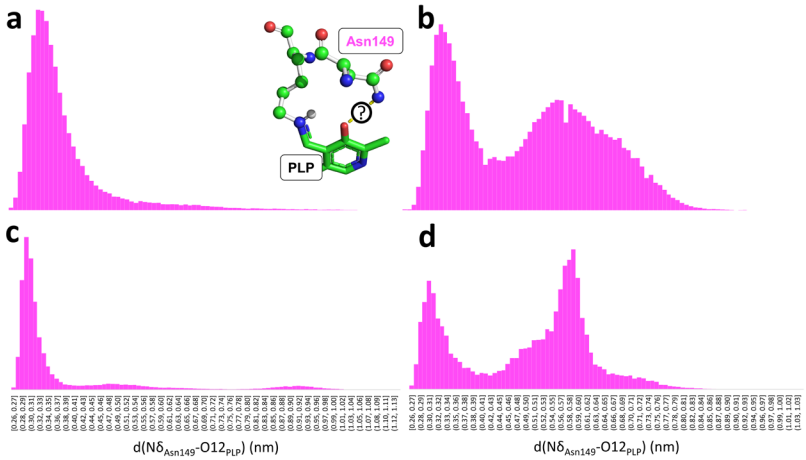

Figure B9: a. *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***C; b.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***C; c.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***C; d.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***C****. The figure inset visually shows d(Nδ$_{Asn149}$-O12$_{PLP}$) and how it relates to PLP and Asn149 in human CBS.*

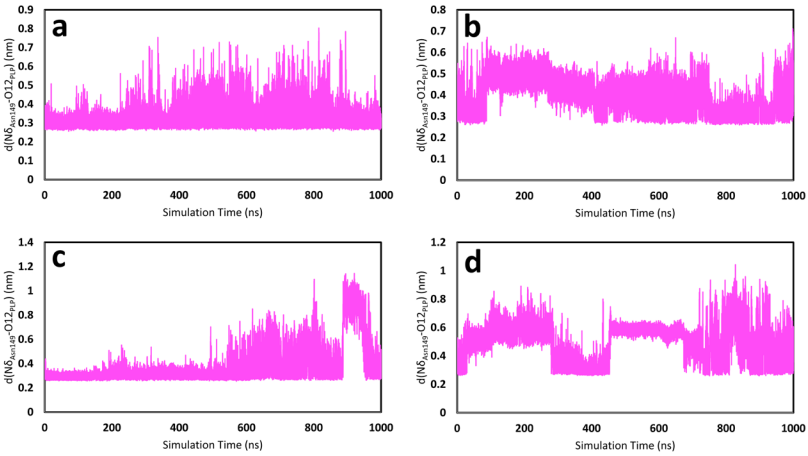

Figure B10: a. *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***C; b.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***C; c.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***C; d.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***C****.*

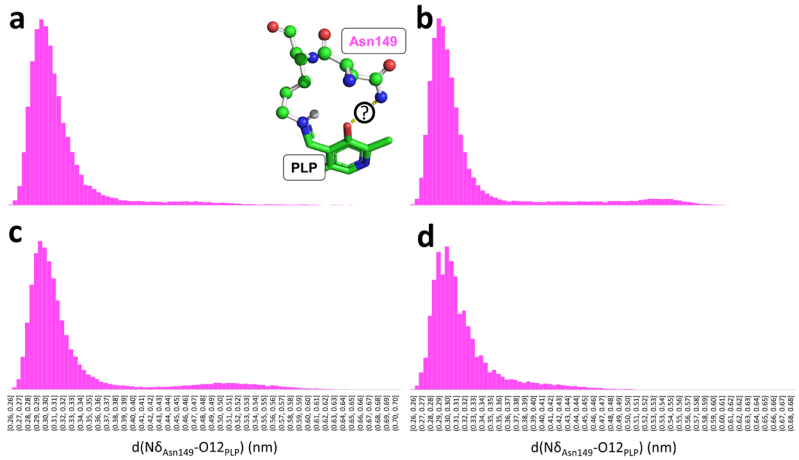

Figure B11: ***a.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***E;*** ***b.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***E;*** ***c.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***E;*** ***d.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***E****. The figure inset visually shows d(Nδ$_{Asn149}$-O12$_{PLP}$) and how it relates to PLP and Asn149 in human CBS.*

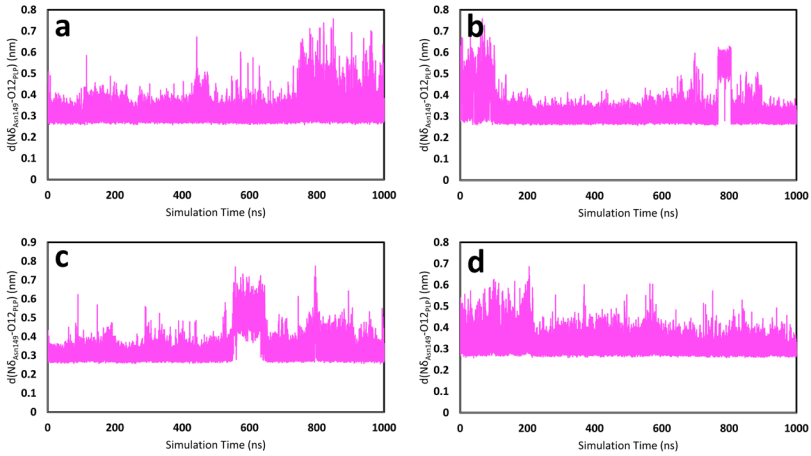

Figure B12: ***a.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***E;*** ***b.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***E;*** ***c.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***E;*** ***d.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***E****.*

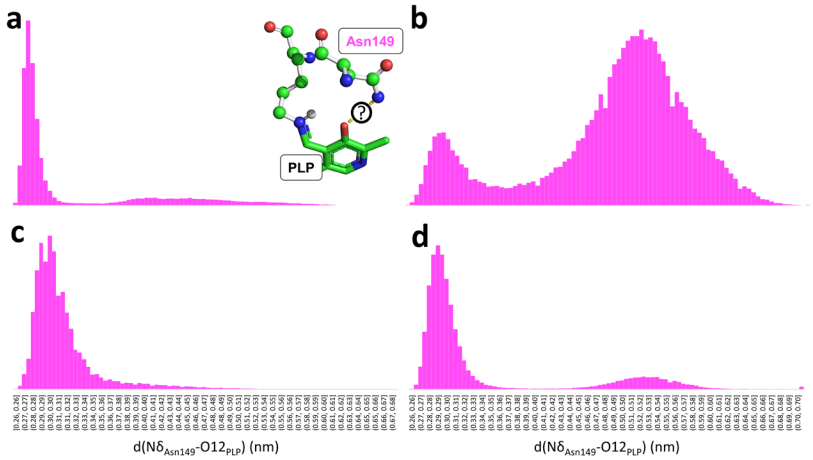

Figure B13: ***a.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***F;*** ***b.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***F;*** ***c.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***F;*** ***d.*** *histogram of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***F****. The figure inset visually shows d(Nδ$_{Asn149}$-O12$_{PLP}$) and how it relates to PLP and Asn149 in human CBS.*

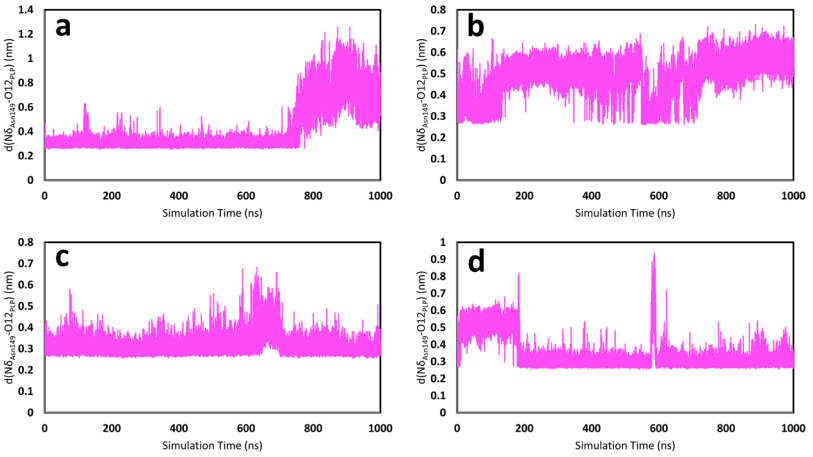

Figure B14: ***a.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***F;*** ***b.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***F;*** ***c.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***F;*** ***d.*** *variation of d(Nδ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***F****.*

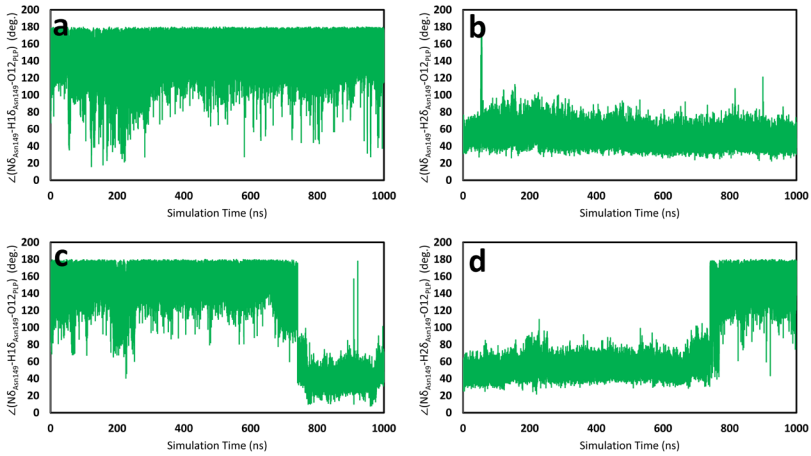

Figure B15: a. *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***A; b.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***A; c.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***A; d.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***A.***

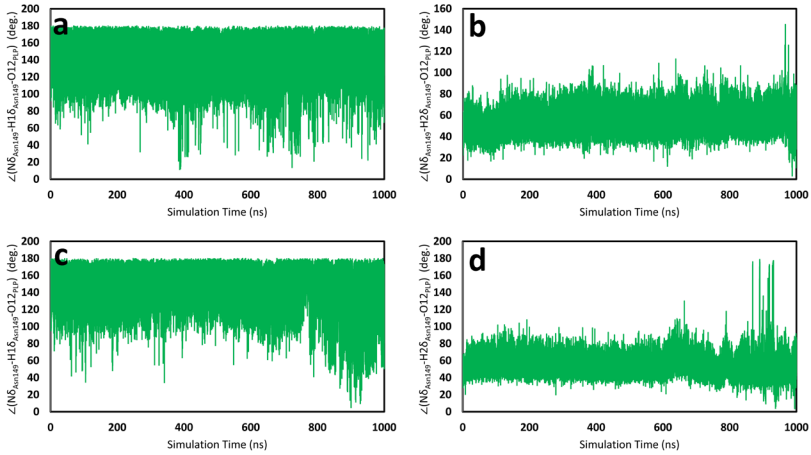

Figure B16: a. *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***A; b.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***A; c.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***A; d.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***A.***

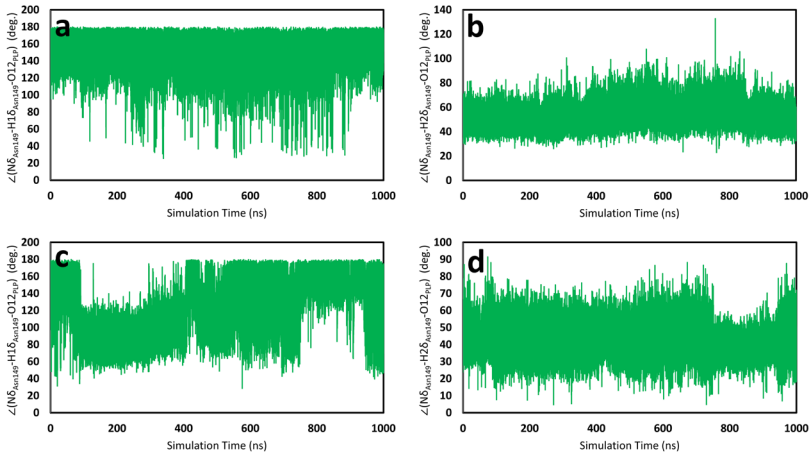

Figure B17: a. *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***C; b.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***C; c.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***C; d.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***C****.*

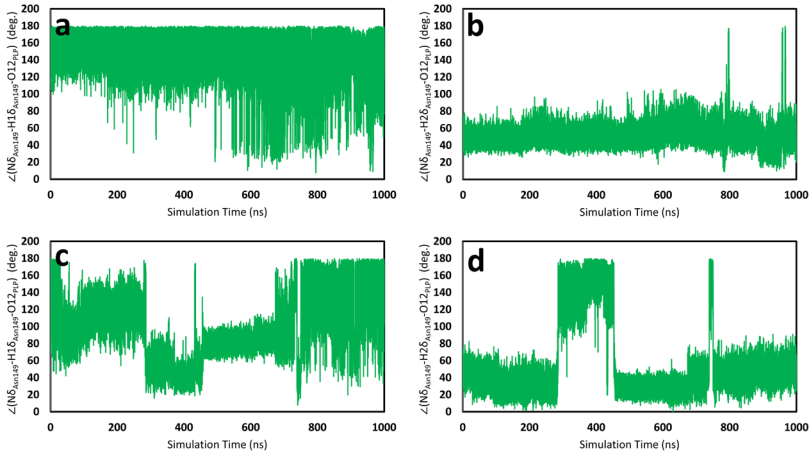

Figure B18: a. *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***C; b.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***C; c.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***C; d.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***C****.*

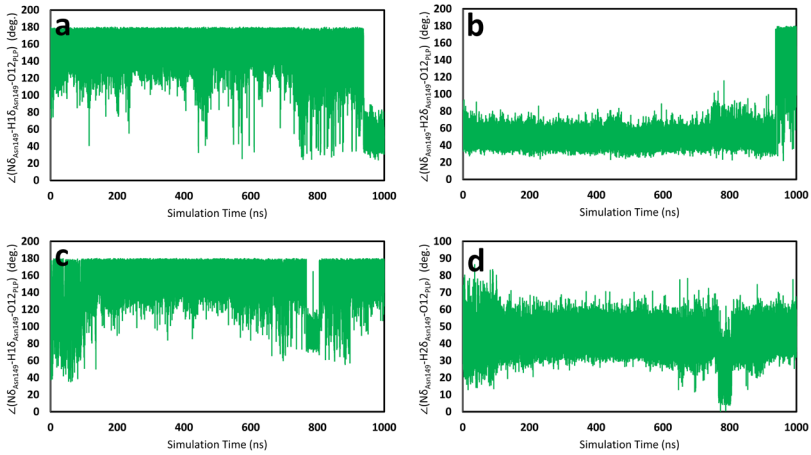

Figure B19: a. *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***E; b.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***E; c.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***E; d.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***E****.*

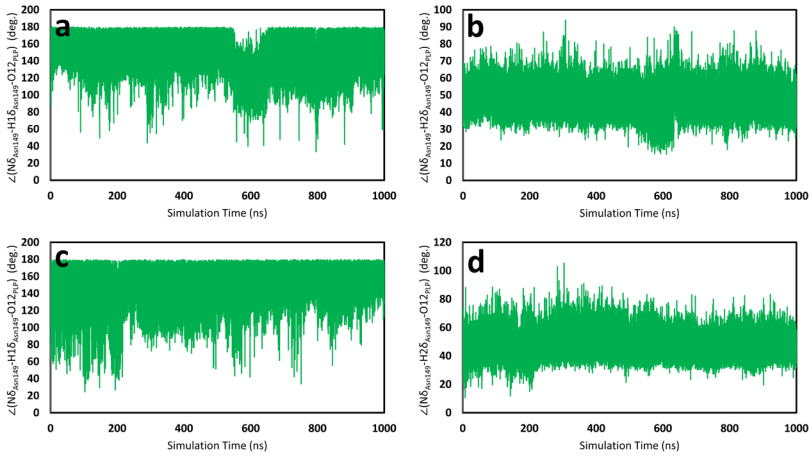

Figure B20: a. *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***E; b.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***E; c.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***E; d.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***E****.*

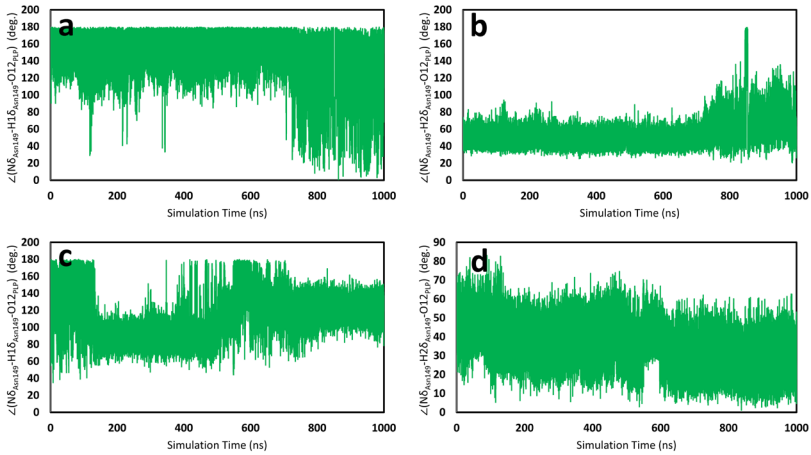

Figure B21: ***a.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***F;*** ***b.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain A for the first 1000 ns simulation of* ***F;*** ***c.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***F;*** ***d.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain B for the first 1000 ns simulation of* ***F****.*

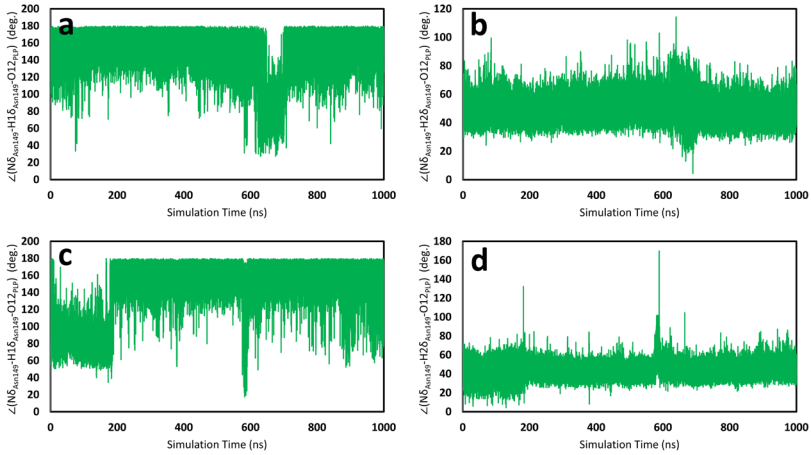

Figure B22: ***a.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***F;*** ***b.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain A for the second 1000 ns simulation of* ***F;*** ***c.*** *variation of ∠(Nδ$_{Asn149}$-H1δ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***F;*** ***d.*** *variation of ∠(Nδ$_{Asn149}$-H2δ$_{Asn149}$-O12$_{PLP}$) for chain B for the second 1000 ns simulation of* ***F****.*

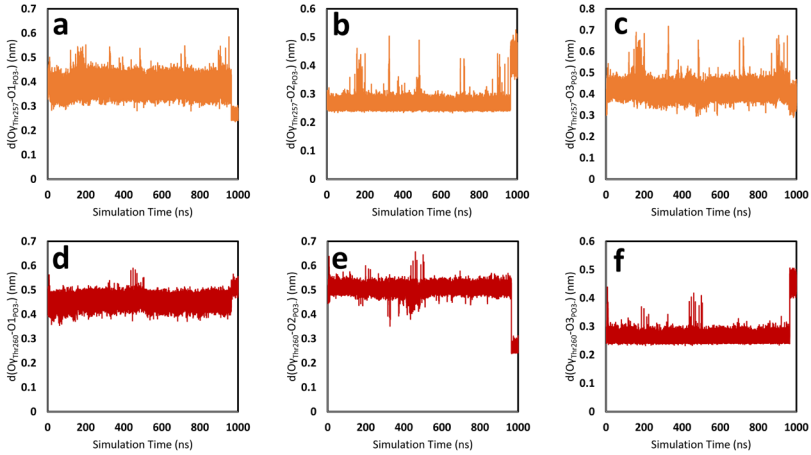

Figure B23: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain A for the first 1000 ns simulation of* ***A; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain A for the first 1000 ns simulation of* ***A****.*

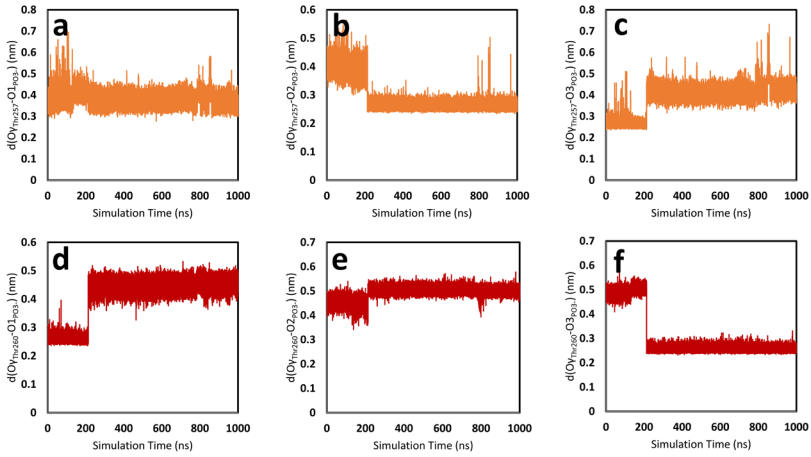

Figure B24: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain B for the first 1000 ns simulation of* ***A; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain B for the first 1000 ns simulation of* ***A****.*

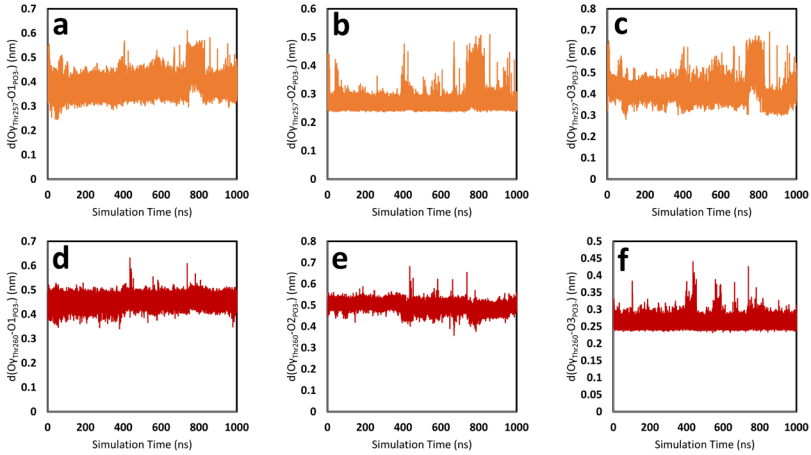

Figure B25: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain A for the second 1000 ns simulation of* ***A; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain A for the second 1000 ns simulation of* ***A****.*

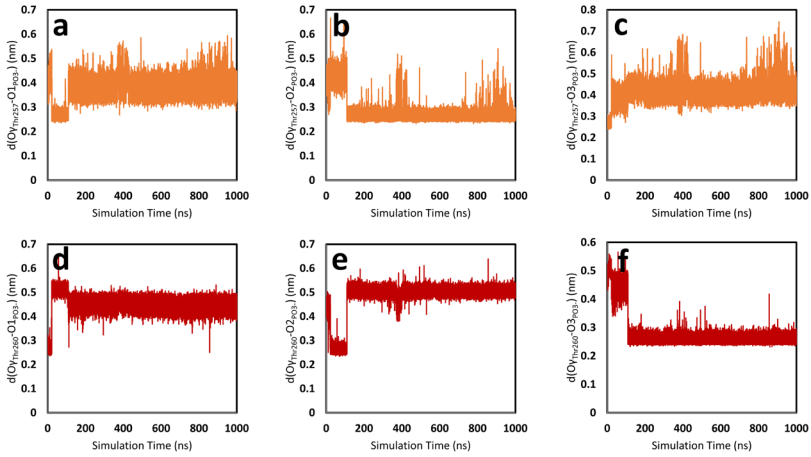

Figure B26: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain B for the second 1000 ns simulation of* ***A; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain B for the second 1000 ns simulation of* ***A****.*

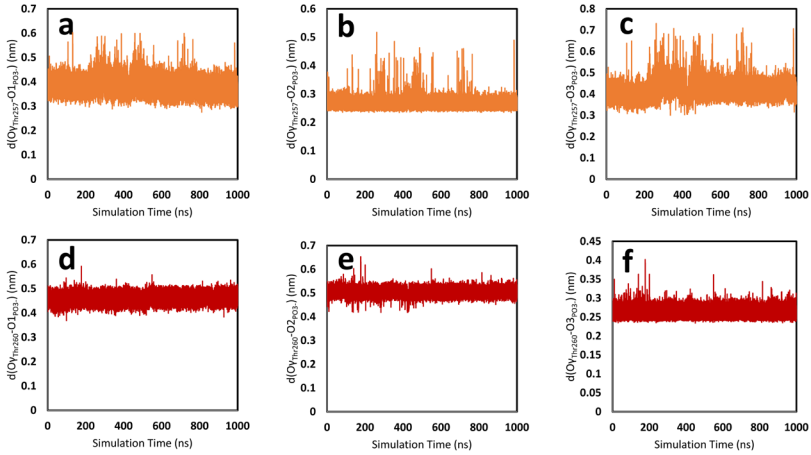

Figure B27: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain A for the first 1000 ns simulation of* ***C; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain A for the first 1000 ns simulation of* ***C****.*

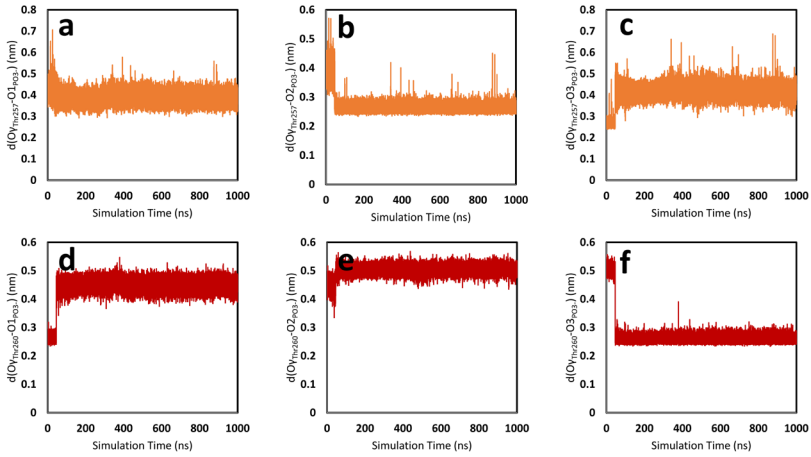

Figure B28: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain B for the first 1000 ns simulation of* ***C; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain B for the first 1000 ns simulation of* ***C****.*

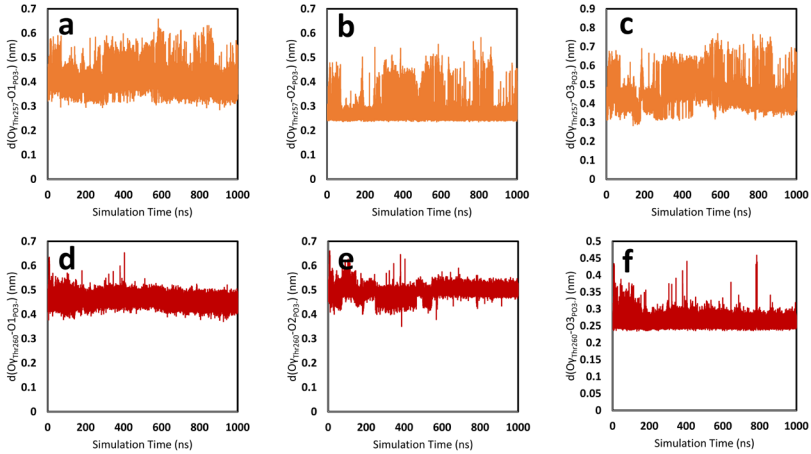

Figure B29: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain A for the second 1000 ns simulation of* ***C; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain A for the second 1000 ns simulation of* ***C****.*

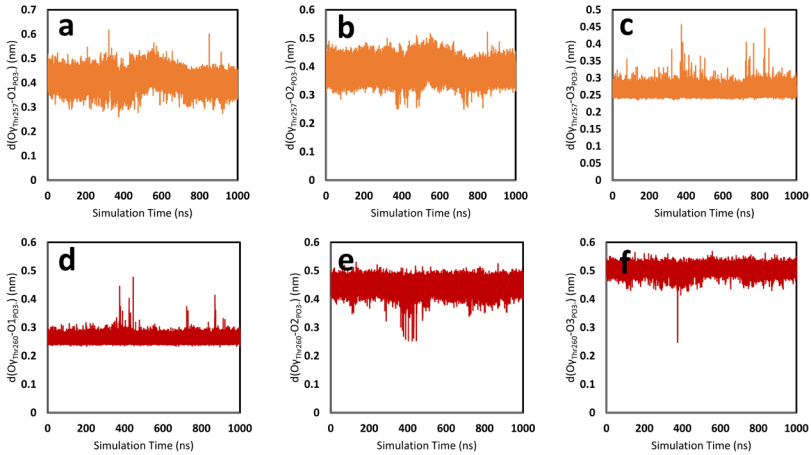

Figure B30: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain B for the second 1000 ns simulation of* ***C; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain B for the second 1000 ns simulation of* ***C****.*

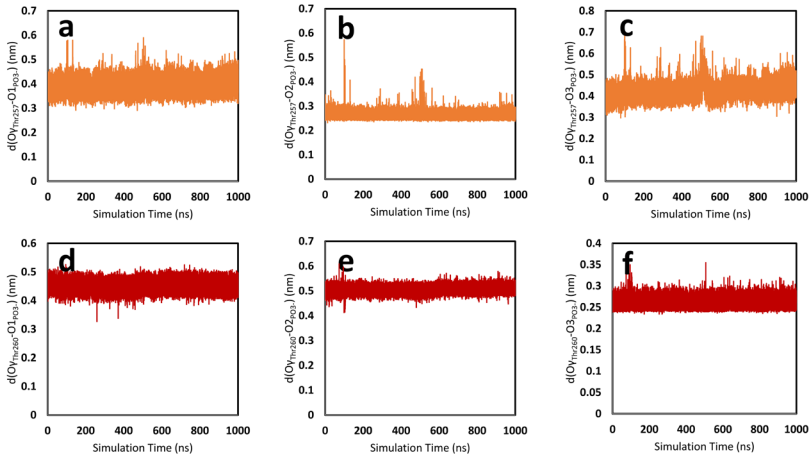

Figure B31: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain A for the first 1000 ns simulation of* ***E; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain A for the first 1000 ns simulation of* ***E****.*

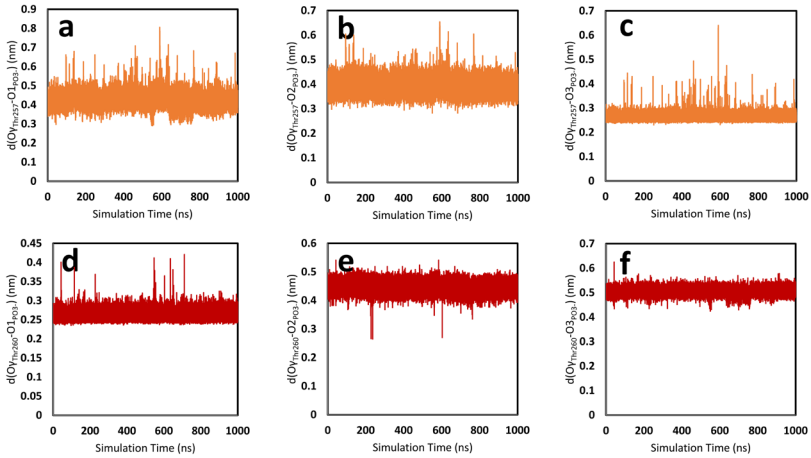

Figure B32: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain B for the first 1000 ns simulation of* ***E; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain B for the first 1000 ns simulation of* ***E****.*

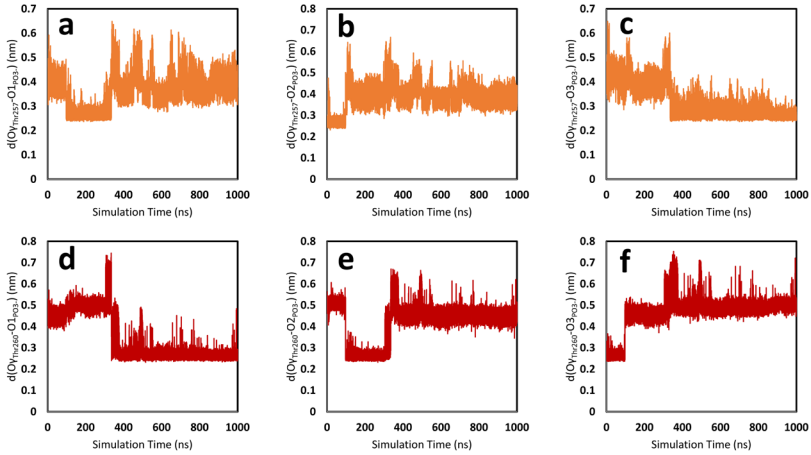

Figure B33: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain A for the second 1000 ns simulation of* ***E; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain A for the second 1000 ns simulation of* ***E****.*

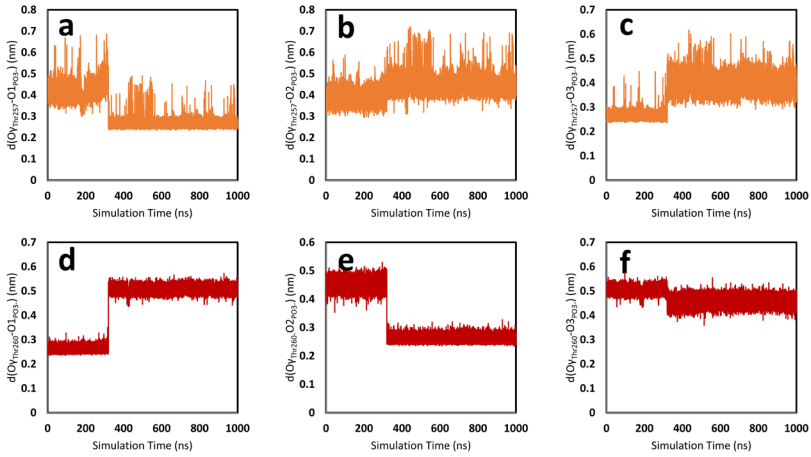

Figure B34: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain B for the second 1000 ns simulation of* ***E; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain B for the second 1000 ns simulation of* ***E****.*

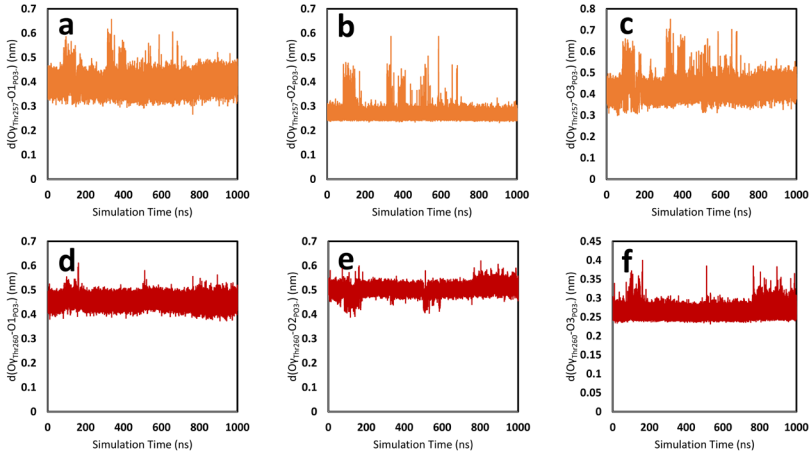

Figure B35: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain A for the first 1000 ns simulation of* ***F; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain A for the first 1000 ns simulation of* ***F****.*

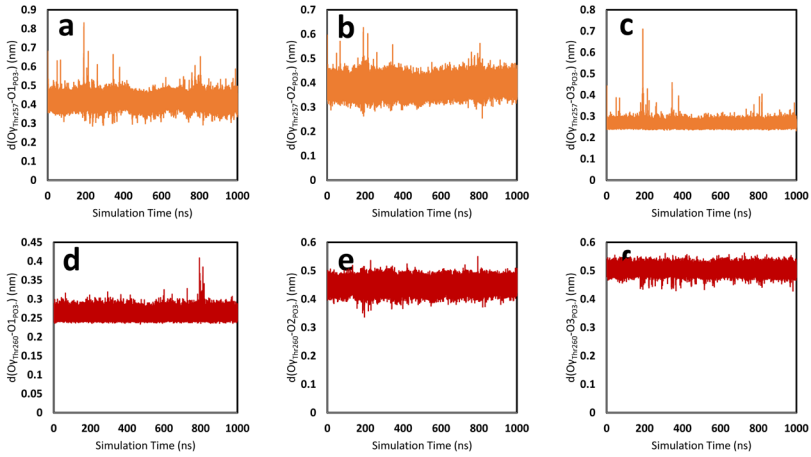

Figure B36: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain B for the first 1000 ns simulation of* ***F; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain B for the first 1000 ns simulation of* ***F****.*

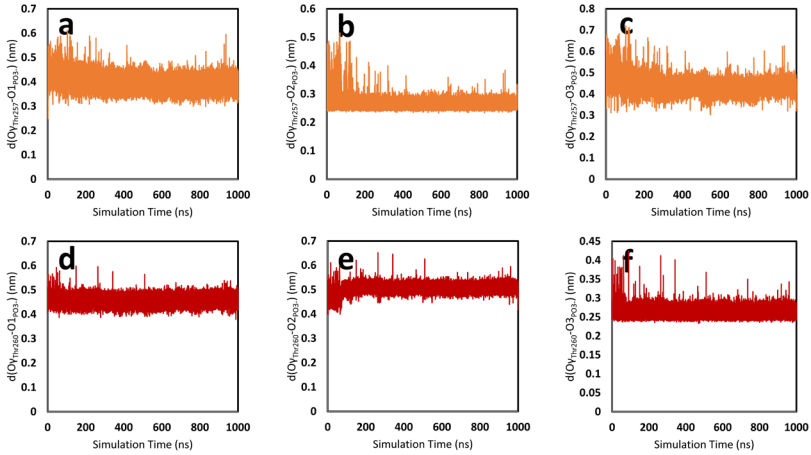

Figure B37: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain A for the second 1000 ns simulation of* ***F; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain A for the second 1000 ns simulation of* ***F****.*

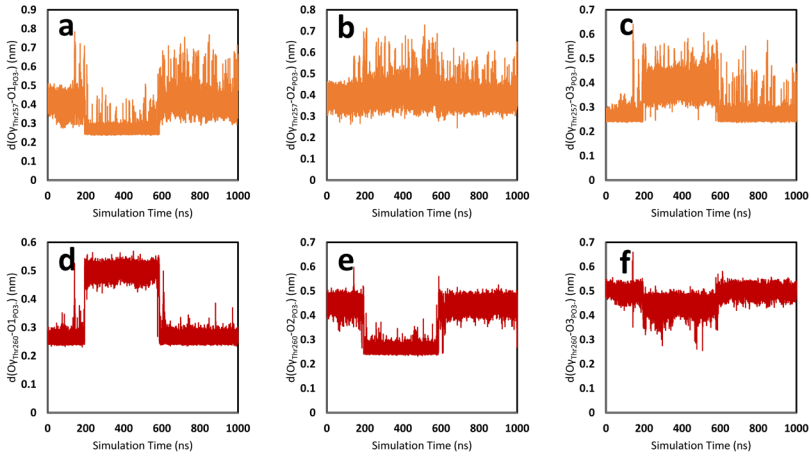

Figure B38: a,b,c. *variation of d(Oγ$_{Thr257}$-O1/2/3$_{PO3-}$) for chain B for the second 1000 ns simulation of* ***F; d,e,f.*** *variation of d(Oγ$_{Thr260}$-O1/2/3$_{PO3-}$) for chain B for the second 1000 ns simulation of* ***F****.*

Table B1: *Relative Gibbs free energy changes calculated with several DFT functionals as well as the DLPNO-CCSD(T_1) calculations. The left-hand side of the table shows the Gibbs free energy relative to the basal state,* ***A****, depending on whether CO or NO is the inhibitory gas. The right-hand side of the tables shows the Gibbs free energy changes associated with the steps of the inhibitory pathways.*

Functional/ Method	**A+CO**	**A+NO**	**D(CO)**	**D(NO)**	**B**	**E**	**C**	**F**
BP86	0.0	0.0	26.2	26.2	-18.2	-9.9	-11.6	-11.8
B3PW91	0.0	0.0	-1.0	-1.0	-8.3	-9.0	-5.6	-5.0
B3LYP	0.0	0.0	0.4	0.4	-8.2	-7.5	-7.6	-5.1
B3LYP*	0.0	0.0	5.0	5.0	-10.7	-8.6	-8.2	-7.2
BLYP	0.0	0.0	22.3	22.3	-18.2	-8.4	-15.6	-11.8
OLYP	0.0	0.0	7.9	7.9	-19.2	-10.2	-19.8	-12.4
PBE	0.0	0.0	25.6	25.6	-20.7	-10.6	-14.0	-12.9
TPSS	0.0	0.0	25.0	25.0	-17.2	-10.7	-10.3	-11.7
PBE0	0.0	0.0	-5.2	-5.2	-8.3	-8.6	0.0	-4.3
TPSSh	0.0	0.0	11.1	11.1	-12.1	-10.2	-5.6	-8.4
DLPNO-CCSD(T_1)	**0.0**	**0.0**	**11.2**	**11.2**	**-5.2**	**-3.9**	**-0.4**	**-1.3**

Table B2: *Summary of raw data from ground spin state calculations on the various heme derivatives. In order, the columns are the species, the relevant spin state, the spin state gaps relative to the singlet or doublet depending on the species, and the obtained ground spin state based on these calculations.*

Species	Spin State	Spin State Gaps (kcal/mol)	Ground Spin State
	Singlet	0	
A	**Triplet**	4.237814347	**Singlet**
	Quintet	8.72320129	
	Doublet	0	
B	**Quartet**	21.66251233	**Doublet**
	Sextet	20.47532124	
	Doublet	0	
C	**Quartet**	22.14982757	**Doublet**
	Sextet	14.92007064	
	Singlet	0	
D	**Triplet**	-6.878271598	**Quintet**
	Quintet	-18.80561276	
	Singlet	0	
E	**Triplet**	50.49830911	**Singlet**
	Quintet	34.17459941	
	Doublet	0	
F	**Quartet**	23.8340426	**Doublet**
	Sextet	11.571415	

Table B3: *Summary of the RMSF based on the position of the α-carbons of the main residues in the signalling pathway found in human CBS. In order, the columns are the species, which simulation run it was, the chain (A or B), and the RMSF values for the relevant residues for each of the species.* ***A, C, E,*** *or* ***F****.*

Species	Run	Chain	RMSF (nm)					
			Cys52	*His65*	*Asn149*	*Thr257*	*Thr260*	*Arg266*
A	*1*	A	0.046	0.190	0.095	0.075	0.063	0.074
		B	0.043	0.158	0.116	0.089	0.063	0.065
	2	A	0.033	0.115	0.087	0.105	0.068	0.061
		B	0.037	0.105	0.128	0.082	0.069	0.068
C	*1*	A	0.388	1.057	0.145	0.107	0.102	0.119
		B	0.484	1.117	0.153	0.111	0.104	0.111
	2	A	0.410	1.156	0.151	0.105	0.086	0.132
		B	0.225	0.859	0.122	0.094	0.075	0.090
E	*1*	A	0.310	0.250	0.073	0.078	0.064	0.073
		B	0.243	0.149	0.072	0.081	0.071	0.075
	2	A	0.219	0.229	0.074	0.078	0.069	0.092
		B	0.177	0.162	0.074	0.084	0.068	0.087
F	*1*	A	0.437	0.165	0.141	0.126	0.108	0.098
		B	0.430	0.319	0.120	0.083	0.072	0.096
	2	A	0.414	0.161	0.089	0.076	0.062	0.089
		B	0.244	0.160	0.089	0.086	0.078	0.075

Table B4: *Hydrogen bond percentage for the Asn149–O^-(PLP) hydrogen bond. The percentage is calculated on the basis of the following hydrogen bond assignment criteria: ∠($N\delta_{Asn149}$-$H\delta_{Asn149}$-$O12_{PLP}$) > 100 degrees, and d($N\delta_{Asn149}$-$O12_{PLP}$) < 0.35 nm.*[26] *Plots of the data used to calculate these percentages are shown in* ***Figures S8 – S23****. In order, the columns are the species, which simulation run it was, the chain (A or B), and hydrogen bond percentage.*

Species	Run	Chain	Hydrogen Bond %
A	*1*	A	87
		B	92
	2	A	73
		B	81
C	*1*	A	83
		B	38
	2	A	75
		B	25
E	*1*	A	89
		B	86
	2	A	83
		B	86
F	*1*	A	71
		B	13
	2	A	88
		B	79

C. Supporting Information for Chapter 5

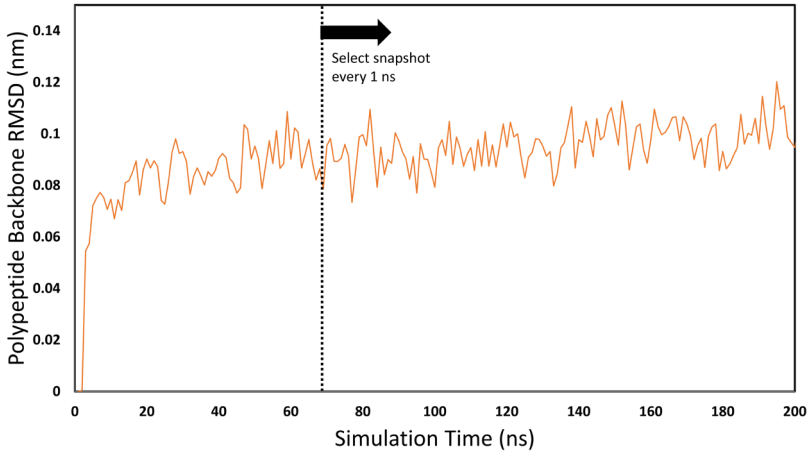

Figure C1: *RMSD of the polypeptide backbone α-carbons with respect to the starting configuration over 200 ns of simulation time, calculating the RMSD at every ns. From 68 ns, MD snapshots were taken systematically every 1 ns, as illustrated.*

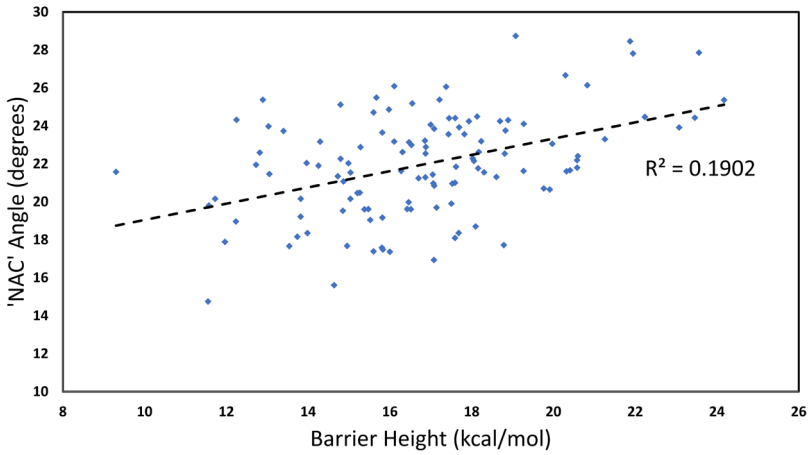

Figure C2: *The dependency of the 'NAC' angle on the barrier height of the 117 pathways calculated herein. For a definition of the 'NAC' angle, please refer to S. Hur and T. C. Bruice, J. Am. Chem. Soc., 2003,* ***125****, 10540–10542.*

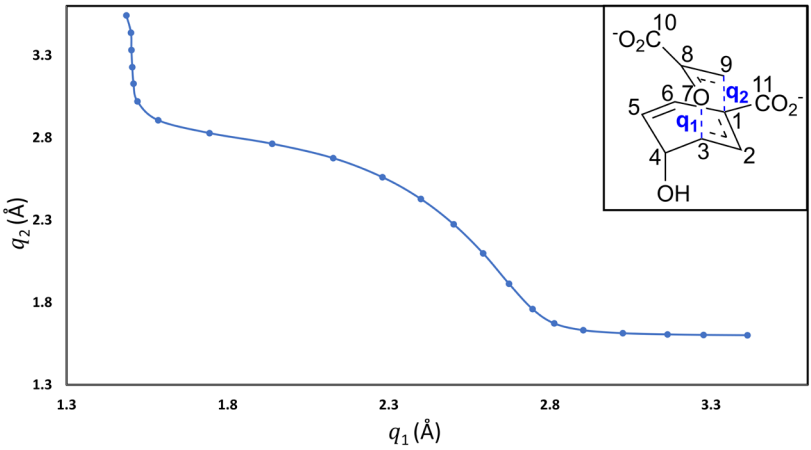

Figure C3: *Variation of internal coordinates q_1 and q_2 averaged along 117 pathways generated herein. q_1 is defined as $d(C_3 - O_7)$, and q_2 is defined as $d(C_1 - C_9)$. An illustration of the relevant internal coordinates is given in the figure inset.*

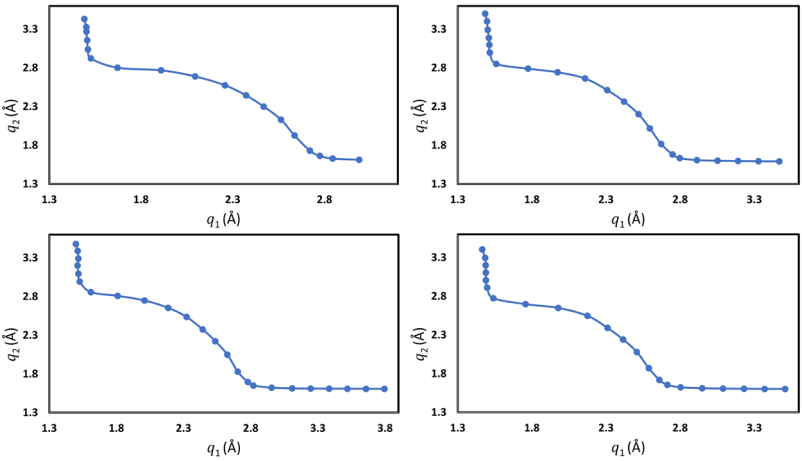

Figure C4: *Variation of internal coordinates q_1 and q_2 for four of the 117 pathways generated herein. q_1 is defined as $d(C_3 - O_7)$, and q_2 is defined as $d(C_1 - C_9)$. An illustration of the relevant internal coordinates is given in the figure inset of* ***Figure C3****.*

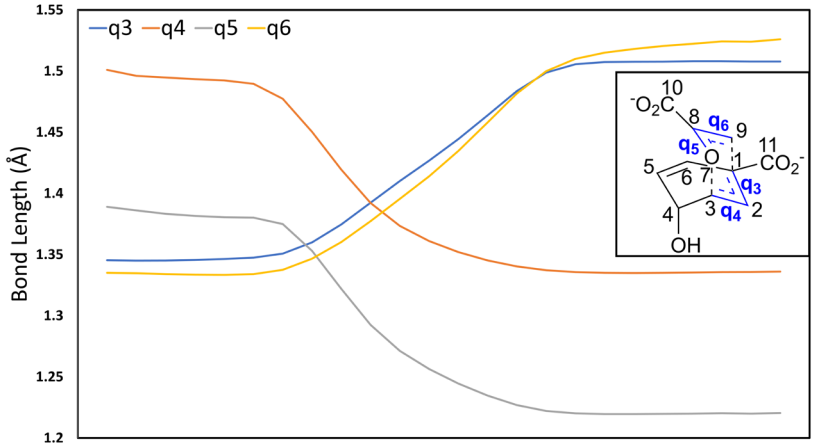

Figure C5: *Variation of internal coordinates q_3, q_4, q_5, and q_6 averaged along 117 pathways generated herein. q_3 is defined as $d(C_1 - C_{11})$, q_4 is defined as $d(C_2 - C_3)$, q_5 is defined as $d(O_7 - C_8)$, and q_6 is defined as $d(C_8 - C_9)$. An illustration of the relevant internal coordinates is given in the figure inset.*

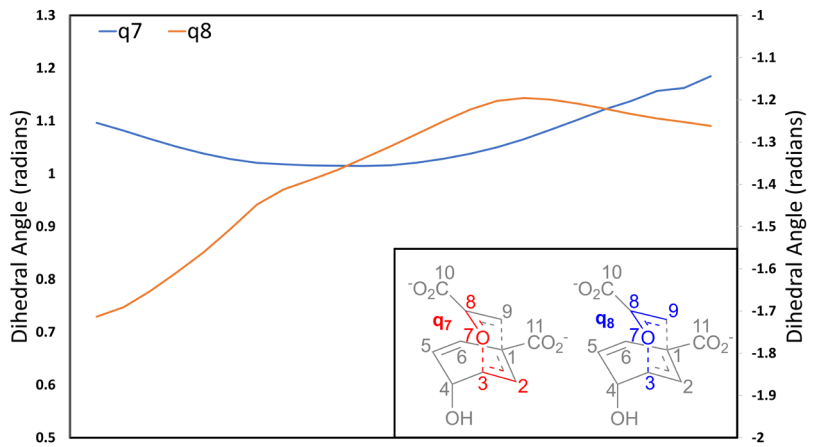

Figure C6: *Variation of internal coordinates q_7 and q_8 averaged along 117 pathways generated herein. q_7 is defined as $\varphi(C_2 - C_3 - O_7 - C_8)$, and q_8 is defined as $\varphi(C_3 - O_7 - C_8 - C_9)$. An illustration of the relevant internal coordinates is given in the figure inset.*

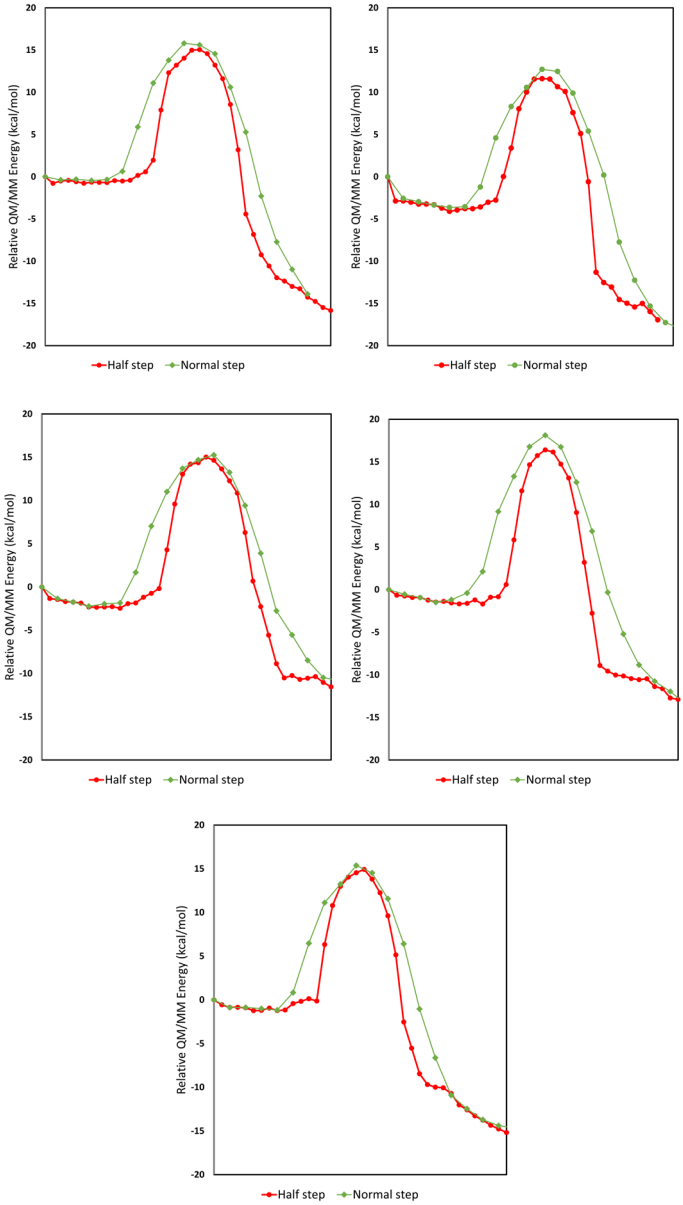

Figure C7: *Comparison of using half-step or normal-step in the growth-phase of the QM/MM SE-GSM algorithm for five of the 117 pathways generated herein. A normal step is defined as steps of 0.1 Å and 5° for bonds and valence and/or dihedral angles, respectively. The half-step is half of these values.*

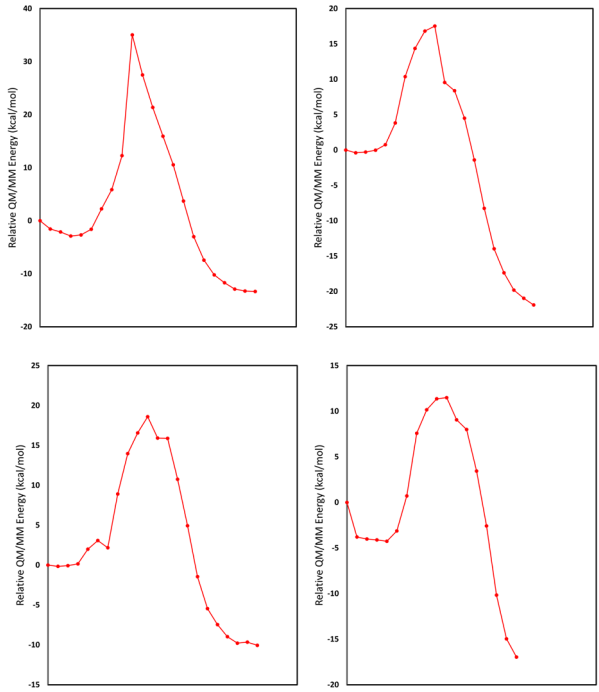

Figure C8: *Four examples of QM/MM energy profiles which were deemed as discontinuous and therefore discarded from further analysis.*

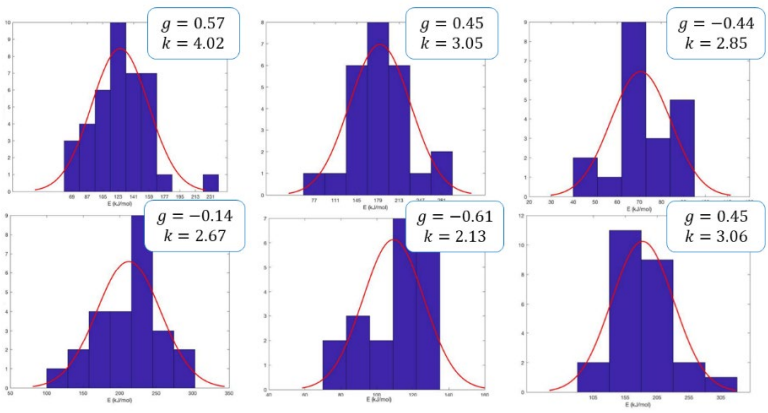

Figure C9: *The calculated skewness (g) and kurtosis (k) for 6 of the datasets reported in the Supporting Information of Ryde, U., J. Chem. Theory Comput.* ***2017****, 13 (11), 5745–5752.. The histograms themselves were taken directly from said Supporting Information, but the skewness (g) and kurtosis (k) were calculated herein with reference to the source of each of these datasets.*