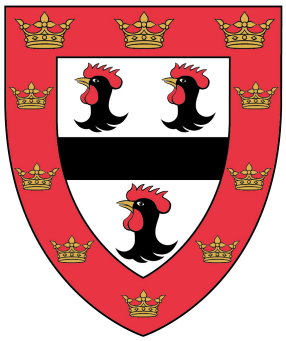


Department of Chemistry
University of Cambridge

Anaerobic, NADH-Dependent Haem Breakdown in a Family of Haemoproteins

—

An Experimental and Computational Study

This dissertation is submitted to the University of Cambridge for the degree Doctor of Philosophy.

Alasdair Donald Keith
Jesus College

September 2021

Declaration

The work described in this dissertation was carried out by the author in the Department of Chemistry at the University of Cambridge between October 2017 and September 2021. The contents are the original work of the author except where otherwise indicated and contain nothing that is the outcome of collaboration. The contents have not previously or concurrently been submitted for any other degree or qualification at the University of Cambridge or any other institution. All resulting publications and the research data are made publicly available as required by the Open Access policies of the University of Cambridge and the Engineering and Physical Sciences Research Council. The number of words in this dissertation does not exceed 60000.

Alasdair Donald Keith

September 2021

Abstract

Anaerobic, NADH-Dependent Haem Breakdown in a Family of Haemoproteins

Alasdair Donald Keith

Many pathogens function by internalising the haem molecules of their host organism and breaking down the porphyrin scaffold to sequester the Fe^{2+} ion. Typically, this breakdown mechanism is mediated by a haem oxygenase. However, a novel class of reaction has been discovered, which can be performed anaerobically using 'nature's reductant', NADH, and the *Yersinia enterocolitica* protein, HemS. To study the features of this reaction in more detail, conventional experimental methods were combined with Energy Landscape Theory. Deuterium labelling demonstrated that the reaction was initiated by hydride transfer and stopped-flow spectroscopy showed that the reaction proceeded *via* a short-lived intermediate. Since no structural information regarding NADH-binding to HemS was available, computational calculations were used to sample the conformational space around possible NADH-protein binding sites and to construct kinetic transition networks. From these networks, pathways showing the unfolding and approach of NADH to haem inside the pocket were determined. These pathways highlighted the roles of various residues, thus allowing for a targeted mutagenesis study. This study, carried out using both computation and laboratory-based experimentation, was especially focussed on a double phenylalanine gate located in the centre of the main cavity. Key insight concerning how this feature regulates the access of NADH to haem was gained.

Computational results suggested that the HemS homologues, HmuS, ChuS and ShuS, were also capable of promoting anaerobic haem breakdown, but that catalysis by ChuS and ShuS may be limited by competing functions. Bioinformatics was used to gauge what these possible alternative functions could be, and to place HemS within its wider phylogenetic context. The computational predictions were

then tested in the laboratory. The three homologues were all shown to engage in the reductive haem breakdown process but to varying degrees of efficacy. These findings demonstrate that this novel haem breakdown reaction is not unique to HemS, but instead is a feature of a wider class of haemoproteins. A subset of these haemoproteins are known to bind certain DNA promoter regions, suggesting not only that they can catalytically degrade haem, but that they are also involved in transcriptional modulation responding to haem flux. Many of the bacterial species responsible for this class of protein (including those that produce HemS, ChuS and ShuS) are known to specifically target oxygen-depleted regions of the gastrointestinal tract. A deeper understanding of anaerobic haem breakdown processes engaged in by these pathogens could therefore prove useful in the development of future strategies for disease prevention.

Acknowledgements

I have always been an unashamedly sentimental type, so this acknowledgements page shall go on for a while...

I thank my two supervisors, Professor David Wales and Dr Paul Barker, for their wisdom, guidance and forbearance. I thank them for suggesting this interesting, enthralling project and I hope I've been of help in unravelling the mysteries of HemS! Much of this project took me out of my comfort zone – I'm glad they were both there to guide me along the way.

I thank Dr Konstantin Röder, Daniel Sharpe and Luke Dicks from the Wales group. Konstantin has proven to be a good friend, mentor and, most importantly, drinking buddy, whilst I have thoroughly enjoyed these last four years with Daniel and Luke as fellow PhD students.

I thank Dr Sally Boss, Dr George Biggs, Jamie Klein and Victoria Daramy-Williams from the Barker group. Their presence certainly made for some interesting group meetings, not least Sally's cat – a regular and popular attendee over Zoom. I am indebted to both George and Jamie for their regular collaborations in the lab, in particular the many mass spec samples they submitted for me!

I thank James Cole and Yuhang Xie, my two Part III students. Both proved to be model students and excellent companions in the lab, each with that slight hint of eccentricity required to work with me. I will never forget James' face (nor he mine I should imagine!) when he lost half his protein sample down the sink, nor will I forget Yuhang's many zany (yet always somehow relevant) cartoons in his group presentations.

Thank you also to the many others I have gotten to know in Cambridge. Dr Alex Thom most definitely deserves a mention, especially considering the many bar tabs he's insisted on paying over the years. I must owe him a fortune. I'm sure Alex would put my reticence to pay down to my being a Scot.

I thank the crowd from Dundee, including Vera Ross, Jack & Margaret Scott and Philma Fotheringham, all of whom unfortunately are long gone but always took a keen interest in my education and motivated me to succeed when I was young.

I must also thank Sheila Whammond, who likewise encouraged me to stick in at school, and John Smith, who has always been available for advice, a cup of coffee and a good laugh.

I thank my friends from school, including but not limited to Stuart Cant, Cameron Ireland, Chris Milne, Junaid Rasul, Jamie Stewart, Paul Whelan, Ryan Barnett and James Todd. I also thank my friends from undergraduate at Heriot-Watt University, in particular those who I lived with, Ryan Crowe, Callum Keanie and Sam Penman.

I thank Nikol Kadeřábková for the years we had together, and wish her every success in the future. I also wish Archie good health and much love.

I thank my uncle, Brian Keith, his wife Marian, and my two cousins Claire and Niall. Here's to more visits to Ibrox together!

I thank my uncle, Derek Keith, who has often been more than just an uncle to me. His staunch defence of the rights of the people of Scotland to walk freely in their own country and to fish its rivers has always served as an inspiration to me.

I thank those who I lived with in Malcolm Street and Wesley House over the last fifteen months of my PhD. The crowd at Wesley House, in particular, managed to temper my pre-viva nerves and grumpiness with much fun and laughter, including but not limited to THAT Halloween prank, heated debates about Kenyan cats, and enthusiastic Christmas decorating & carolling.

From Wesley House, I especially thank Allison Burnette. From talks over cuppas to late-night laughter, I've loved spending so much time together. I've never known someone with so much positivity, energy, and zeal for 'making merry.' *I feel like we fell out of a lucky tree, hit every branch on the way down, and ended up in a pool full of cash and Sour Patch Kids.*

I thank my 'adopted Grandad', Magne Røinås. I don't think I have ever met anyone more generous. Nor have I known someone so close to nature. When I saw him swimming with seals in the North Sea, I realised then that he was a true Viking! I'm not sure whether Jennifer was so pleased – she is sorely missed.

I thank my Nana, Phyllis McLaren. From her I learned the wonderful Doric word, drochle, which makes for a great insult. She was of a generation of Scots long since gone – a stern Presbyterian on the surface, but a vivacious and loving woman to those who knew her best. I also thank her husband and my Great-Grandfather, Donald McLaren, Spitfire pilot during World War II, who made the ultimate sacrifice so we could live in peace and security today.

I never got to know my Grandad, James Martin (1941-1965), nor my Granny, Catherine Keith (1931-1993). Both died far too young. However, through their examples, both taught me to love the people of this country. My Grandad died as

an RAF officer, whilst my Granny spent much of her life standing up for ordinary working-class people. I know she would have been proud of her four grandchildren going to university, and especially of her youngest son being elected a Labour Party councillor.

I thank my Step-Granny, Jean Keith, who sadly passed away the day I submitted this thesis for examination. There is no doubt she was a blessing to my Grandad in his later years. The way she told the story about my Grandad singing at their wedding always made me laugh. Jean was always so kind to my brother and I. She truly was part of the family, and will be sorely missed.

I thank my brother, Martin. I could not wish for a better brother. He is not only that. He is my best friend. Though nearly 6 years younger than me, he's taught me so much (not least because he's far more 'with it' than me!). Though I was the brother blessed with the superior looks and charm (of course!), his wit has always cracked me up. I'll never tire of Mum and I in hysterics at something Martin has said. Poor Dad, who is usually the brunt of it all!

Yes, my Dad, Alexander, has to put up with a lot. I've never known somebody so selfless. He has always put his family first, his community second, and himself firmly last. It was he who taught me, as a young boy, the importance of serving others. As a local councillor, he never once wavered in his principles, which was to stand up for the poor and speak truth to power. Sometimes I wish I just had an ounce of the steel and determination he has. I could not have wished for a better role model.

Lastly, I thank my Mum, Susan. She, like my Dad, spent many years in service to the community. As a social worker, she has seen the very best and very worst of society, and she can be proud of what she did to make it that bit better. But it's always been clear, to me at least, that my Mum's first love is her family. She has often been a crutch for me to lean on. She has always known what to say and how to make things better. Though I hate to say it, I think it may just be true that 'Mums are always right.' She has never let me down.

This thesis is dedicated to my Grandad, Alexander Keith (1928-2015), and to my Granny, Anne Martin (1940-2019). Both inspired me deeply. My Grandad imparted two important life lessons – the virtue of patience and contentment, and a love for Rangers FC. I was especially close to my Granny, and I miss her every day. She had a hard life but remained happy and was able to light up any room. She showed you could make something out of nothing. She always used to say of her own Grandmother, 'She had a heart of gold. She would have given a beggar her last penny.' The same was true of herself. I will always look up to her.

Glossary of Abbreviations

5′-dA•	5′-deoxyadenosyl-5′-radical
ABC	ATP-binding cassette
AMBER	Assisted model building with energy refinement
AMBER12	AMBER package version 12
AMBER16	AMBER package version 16
ATP	Adenosine triphosphate
BH	Basin-hopping
BTP	Bis-tris propane
BVR	Biliverdin reductase
CCDC	Cambridge Crystallographic Data Centre
CDC	The Centers for Disease Control and Prevention
CIP	Calf intestinal alkaline phosphatase
CPR-NADPH	Cytochrome P450 Reductase-NADPH
CPU	Central processing unit
Cryo-EM	Cryogenic electron microscopy
DAB	Deuteroanaerobilin
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNEB	Doubly-nudged elastic band
DPS	Discrete path sampling
dsDNA	Double-stranded DNA
DTT	Dithiothreitol

EF	Eigenvector-following
ELT	Energy landscape theory
EPR	Electron paramagnetic resonance
ESI-MS	Electrospray ionisation mass spectrometry
FBS	Fur-binding site
FES	Free energy surface
ff99SB	Force field 99 with structure balance
FFS	Forward flux sampling
Fur	Ferric uptake regulator
GB	Generalised Born
GPU	Graphical processing unit
HBP	Haem breakdown product
HEF	Hybrid eigenvector-following
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
His	Histidine
HIV-1	Human immunodeficiency virus 1
HO	Haem oxygenase
HPLC	High performance liquid chromatography
IDA	Iminodiacetic acid
IPTG	Isopropyl-β-D-thiogalactopyranoside
IR	Infrared
I-TASSER	Iterative threading assembly refinement
IUPAC	International Union of Pure and Applied Chemistry
LB	Lysogeny broth
L-BFGS algorithm	Limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm
LCMS	Liquid chromatography mass spectrometry
LEaP	Link, edit and parm.
LED	Light-emitting diode
MAB	Mesoanaerobilin

MBH	Mutational basin-hopping
MCM	Monte Carlo plus energy minimisation
MD	Molecular dynamics
min	Minimum/minima
mRNA	Messenger RNA
MS	Mass spectrometry
MS^n	Tandem mass spectrometry
NAD^+	Nicotinamide adenine dinucleotide (oxidised)
NADD	Deuterated NADH
NADH	Nicotinamide adenine dinucleotide (reduced)
$NADP^+$	Nicotinamide adenine dinucleotide phosphate (oxidised)
NADPH	Nicotinamide adenine dinucleotide phosphate (reduced)
ndMS	Non-denaturing mass spectrometry
NGT	New graph transformation
NIR	Near-infrared
NMR	Nuclear magnetic resonance
OD	Optical density
ORF	Open reading frame
PBT	Periplasmic binding-protein-dependent transport
PCR	Polymerase chain reaction
PDA	Photodiode-array
PDB	Protein data bank
PEG	Polyethylene glycol
PES	Potential energy surface
Phe-	Phenylalanine
PPI	Protein-protein interaction
PPIX	Protoporphyrin IX
QM/MM	Quantum mechanics/molecular mechanics
RD	Restriction digest

RMS	Root mean square
RNA	Ribonucleic acid
RNAP	RNA polymerase
RRKM theory	Rice-Ramsperger-Kassel-Marcus theory
RSMT	Radical SAM methyltransferase
SAM	*S*-adenosylmethionine
SASA	Solvent accessible surface area
SD	Sub-database
SDS-PAGE	Sodium dodecyl sulphate-polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
SOC	Super optimal broth with catabolite repression
SP	Stationary point
SPR	Surface plasmon resonance
sRNA	Small RNA
ssDNA	Single-stranded DNA
SVD	Singular value decomposition
ToF	Time of flight
TPS	Transition path sampling
TS	Transition state
TST	Transition state theory
UPLC	Ultra performance liquid chromatography
UV-Vis	Ultraviolet-visible
WT	Wild type

Glossary of Symbols

A, B and I	Reactant, product and intervening sets from DPS
A_{ij}	Electrostatic force constant of repulsion between i and j
A_λ	Absorbance at a given wavelength, λ
B_{ij}	Electrostatic force constant of attraction between i and j
C_b^A / C_a^B	Committor probabilities
$C^H(T)$	Heat capacity
c	Concentration
$D(i,j)$	Euclidean distance between i and j
$\mathbf{E}$	Potential energy
E_{AMBER}	Energy derived from AMBER force field
E_n	Energy at which the superbasin analysis, n, was performed
E_s	Set of edges in a weighted, directed graph
$F_i^E(T)$	Free energy of minimum i
F_i	Component of the gradient along each eigenvector
F_α	Force on particle α
$\mathbf{G}$	Gradient of the potential energy surface
G	Magnitude of $\mathbf{G}$
g_i	Change in occupation probability of minimum i
$\mathbf{g}$	Gradient of the 'true' potential energy (PE) (DNEB)
$\tilde{\mathbf{g}}$	Gradient of the spring PE (DNEB)
g^{DNEB}	Gradient of the doubly-nudged elastic band
g^{NEB}	Gradient of the nudged elastic band

$\mathbf{g}^{\parallel}$	Gradient of the parallel component of the 'true' PE
$\mathbf{g}^{\perp}$	Gradient of the perpendicular component of the 'true' PE
$\tilde{\mathbf{g}}^{\parallel}$	Gradient of the parallel component of the spring PE
$\tilde{\mathbf{g}}^{\perp}$	Gradient of the perpendicular component of the spring PE
$\mathbf{H}$	Hessian matrix
h	Planck constant
K_{a}	Acid dissociation constant
K_{D}	Dissociation constant
K_{M}	Michaelis constant
K_r	Force constant (AMBER bonds)
K_θ	Force constant (AMBER angles)
k_{AB} / k_{BA}	Rate constants from region B to A / A to B
k_{AB}^{NSS} / k_{BA}^{NSS}	Non-steady state rate constants from region B to A / A to B
k_{ab} / k_{ba}	Rate constants from minimum b to a / a to b
k_B	Boltzmann constant
k_{spr}	Spring constant
$k_i^{\dagger}$	Unimolecular rate constant through transition state † from minimum i
l	Optical path length
M_α	Mass of particle α
M_s	Set of minima in a weighted, directed graph
min	Local energy minimisation
N	Number of atoms
n	Multiplicity
n_i	Number of distinct permutational isomers in minimum i
n_{max}	Maximum number of allowed connection attempts
n_{spr}	Number of springs
n_u	Number of connection attempts
P_a / P_b / P_i	Occupation probability of minimum a / b / i
P_A^{eq} / P_B^{eq}	Equilibrium occupation probability of region A / B

P_a^{eq} / P_b^{eq}	Equilibrium occupation probability of minimum a / b
$\mathbf{P}(t)$	Occupation probability vector for all states at time t
$\mathbf{P}^{eq}(t)$	Equilibrium occupation probability vector at time t
q_i	Charge on atom i
R_i / R_j	Effective Born radii
r_b	Bond length
r_{eq}	Equilibrium bond length
r_{ij}	Distance between atoms i and j
s	Integrated path length
T	Temperature
t	Time
t_a / t_b	Mean waiting times under the non-steady state condition
V_i	Potential energy for minimum i
V_n	Force constant (AMBER dihedrals)
V_t	'True' potential in DNEB
$V(\mathbf{X})$	Potential energy surface
$V^*(\mathbf{X})$	Transformed potential energy surface
$\tilde{V}$	Spring potential in DNEB
$\mathbf{W}$	Transition matrix
$w(i,j)$	Edge weight of minima pair (i,j) in a weighted, directed graph
$\mathbf{X}$	Nuclear coordinates
$\mathbf{X}_\alpha$	Nuclear coordinates at particle α
$\mathbf{x}$	Small displacement from $\mathbf{X}$
$\mathbf{x}_{\mathrm{NR}}$	Newton-Raphson step
$Z(T)$	Canonical partition function
$Z_i(T)$	Contribution to canonical partition function from minimum i
$Z^\dagger(T)$	Modified canonical partition function of the TS
β	Thermodynamic beta, $1/k_BT$
γ	Phase angle

δE	Potential energy change
ΔG_{el}	Electrostatic solvation free energy
ΔG_{np}	Non-polar solvation free energy
ΔG_{solv}	Total solvation free energy
δt	Short time step
$\Delta V_i^{\dagger}$	Energy difference between minimum i and TS †
ϵ	Relative permittivity of medium
$\epsilon_{protein}$	Relative permittivity of protein
ϵ_{SM}	Extinction coefficient at the Soret maximum
ϵ_{solv}	Relative permittivity of solvent
ϵ_{λ}	Extinction coefficient at a given wavelength, λ
ζ	Weighted, directed graph
θ	Bond angle
θ_{eq}	Equilibrium bond angle
κ	Debye-Hückel screening parameter
$\tilde{\lambda}_i$	Eigenvalue i
λ	Wavelength
$\boldsymbol{\nu}_i$	Eigenvector i
$\overline{\nu_i}$	Geometric mean vibrational frequency of minimum i
ρ_i	Van der Waals radius of atom i
$\tilde{\rho}_i$	Intrinsic radius of atom i
ϕ	Torsional angle
χ	Number of vibrational degrees of freedom
$\Omega(E)$	Density of states
$\Omega_i(E)$	Contribution to density of states from minimum i

Contents

Chapter 1

Introduction

"Life itself is a school, and Nature always a fresh study." Hugh Miller

In 1997, leading geneticists Craig Venter and Daniel Cohen famously stated that, 'If the 20th century was the century of physics, the 21st century will be the century of biology.'[1] In many respects, this prediction has been borne out, with significant advances, for example, in epigenetics,[2–4] directed evolution,[5;6] genome editing[7–9] and cryo-electron microscropy.[10–12] Chemistry has often acted as an essential bridge so that past advances in physics can be translated through to those in biology. A key aspect of this process is the interplay between theoretical, computational modelling, and bench-top experimentation. Such an approach is especially useful for understanding protein folding and protein-ligand interactions, since these techniques can access different realms of information, and so can be used to complement each other in a wider research context.

Examples of successful collaborations between experimentalists and theoreticians within biology include the use of a computationally designed protein probe to isolate a broadly neutralising antibody found in HIV-1-infected patient serum,[13] and to better understand tumour growth in hypoxic tissues.[14–16] A further example would be the construction of protein-protein interaction (PPI) networks to chart the dynamic interactions between proteins in malaria-spreading *Plasmodium falciparum*.[17] These examples were all deep-level collaborations but, more broadly, advances made in modelling packages, analysis software and database organisation have greatly expanded experimental horizons.

In this thesis, both experiment and theory were used to help uncover the behaviour of the haem-binding protein, HemS, and a selection of its homologues and mutants. The enzymatic reactivity of haem with the biomolecule, NADH, first dis-

covered in HemS by the Barker Group, was investigated in detail using both of these approaches.

In this Introduction, a background to these proteins, their context within their wider operons, and how these systems fulfil certain aspects of bacterial haem acquisition and utilisation strategies, will be discussed. This presentation will be supplemented by a discussion of the principles behind computational biochemistry, and its expanding role in life sciences research.

1.1 Iron in Biology

1.1.1 The 'Iron Paradox'

The primary isotope of iron (^{56}Fe) has the greatest nuclear stability of any known element in the universe,[18] and is the fourth most abundant species in the Earth's crust.[19] These features, coupled with its chemical versatility, made it a logical, and perhaps even ideal, candidate for use in early biological systems. Its ability to act as a Lewis acid and to occupy a variety of oxidation (primarily II and III) and spin (high and low) states under physiological conditions[20;21] render it suitable for a wide range of biochemical processes.

There are certain limitations to using free iron in biological systems, however, which can partly be explained by evolutionary history. In the era of the 'primordial soup', when life first began on Earth, iron primarily resided in its stable ferrous form, Fe^{2+}. However, the expansion of oxygenic photosynthesis, which the evolution of early microorganisms afforded, led to the 'pollution' of the atmosphere with molecular oxygen. This negatively impacted iron-based biochemistry in two ways. Firstly, it resulted in a gradual shift in iron's natural state, from the ferrous to the ferric form ($Fe^{2+} \rightarrow Fe^{3+}$). This shift had serious implications for availability in biological systems as the solubility in aqueous conditions and pH 7.0 for ferrous and ferric iron are 0.1 M and 10^{-18} M respectively.[20] Secondly, the use of ferrous iron became compromised due to its ability to coordinate O_2 and initiate Fenton chemistry.[20;22]

$$\begin{aligned} Fe^{2+} + O_2 &\longrightarrow Fe^{3+} + O_2^- \\ 2\,O_2^- + 2\,H^+ &\longrightarrow H_2O_2 + O_2 \\ Fe^{2+} + H_2O_2 &\longrightarrow Fe^{3+} + OH^- + OH^{\cdot} \end{aligned}$$

Scheme 1: Iron oxidation, followed by Fenton chemistry.

This process produces highly reactive hydroxyl radicals, which are destabilising for cellular environments and can cause extensive tissue damage.

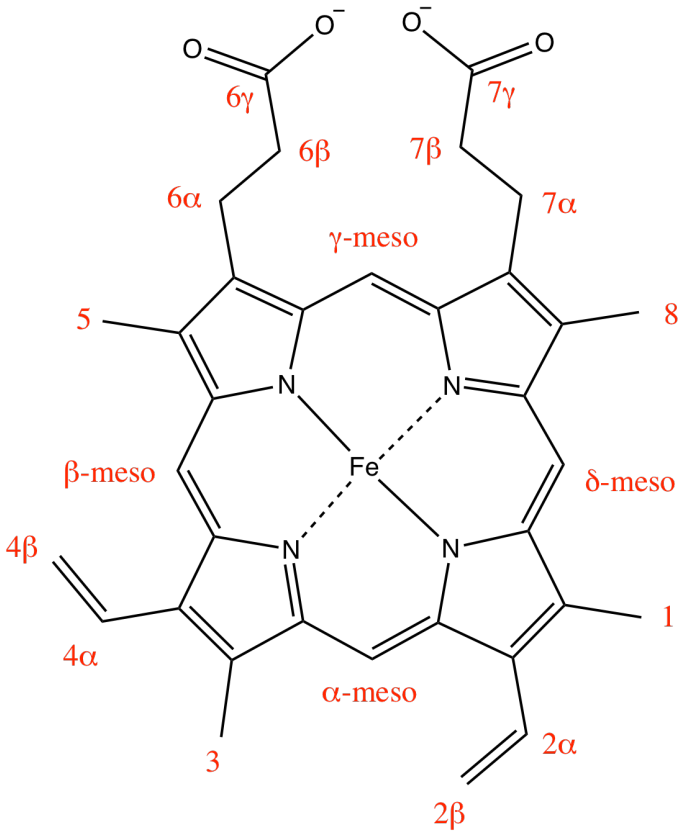


Figure 1.1: Labelled structure of haem b.

This dilemma is commonly known as the 'iron paradox' – life on Earth is reliant on iron, and yet its two main ions are either toxic or insoluble inside cells. Nature has developed a variety of methods to overcome this problem, such as storage as a crystallite in lactoferrin,[23] or being bound as a single ion (primed to react) within a lipoxygenase.[24] Another solution is to store the iron in a porphyrin scaffold, an examples of which, haem, has become pervasive across the biological realm.

1.1.2 Haem

Haem is thought to be a vital component in every living organism. It can exist in a variety of subtly different forms, the most common of which is haem b, which is the focus of this thesis. This form consists of an iron ion coordinated *via* four chelating nitrogen atoms to the cyclic tetrapyrrole, protoporphyrin IX (Fig. 1.1).

The ring surrounding the iron ion is conjugated and planar, leaving two free axial sites, which proteins often make use of to bind as a Lewis base to the iron. Such coordination can significantly disrupt the planarity of the ring. Additionally, the ring itself is functionalised by two propionate groups, two vinyl groups and four methyl groups, in an arrangement that makes the molecule asymmetrical about any axis.

These groups afford further opportunities for proteins to bind to haem. Examples include salt bridge formation between cationic residues (such as arginine, histidine or lysine) and the propionates,[25] and covalent attachment of cysteine residues to the vinyl groups,[26] as well as less rigid π–π aromatic interactions, polar contacts and dispersion forces.

A significant complicating factor, however, is that free haem is cytotoxic.[27] This is due to its hydrophobicity, causing it to separate into membranes or even stack together at high concentrations. The first of these possibilities can lead to lipid peroxidation due to the ability of the iron ion to initiate the Fenton reaction, despite being in a scaffold.[28] It is therefore essential for haem to be contained at all times. To achieve this containment, haemoproteins have evolved to fulfil three basic functions: acquisition, transportation and utilisation.

1.2 Haemoproteins

1.2.1 Protein Evolution, Structure and Function

Proteins are remarkable chemical machines. They are recognised as the 'workhorses of the cell', due to their abundance (comprising 55% of the dry mass of a typical bacterial cell[29]) and diversity in structure and function. Proteins are constructed from only six elements (carbon, hydrogen, oxygen, nitrogen and small amounts of sulfur and phosphorus), which are arranged into approximately 20[a] different amino acid building blocks, and are made *via* condensation polymerisation of these building blocks in a tightly controlled reaction directed by DNA-derived RNA templates. This control over the sequence is essential for the reliable synthesis of proteins, which typically range from 50-2000 amino acid units.[30] Assuming all sequential arrangements over this range to be possible gives a theoretical number of configurations of:

$$\sum_{n=50}^{2000} 20^n = 1.209\times10^{2602}. \tag{1.1}$$

This value vastly exceeds the number of atoms ($\sim10^{80}$) in the known universe. In reality, however, only a small subset of these configurations are ever accessed. This situation reflects the evolutionary history of protein, whereby they (and their associated information storage partners, DNA and RNA) had to adapt to given geological

[a]There was a long-established consensus that 20 amino acids were required to make all proteins. However, in recent years, selenocysteine and pyrrolysine have come to be regarded as the 21st and 22nd amino acids. There are perhaps more to be discovered.

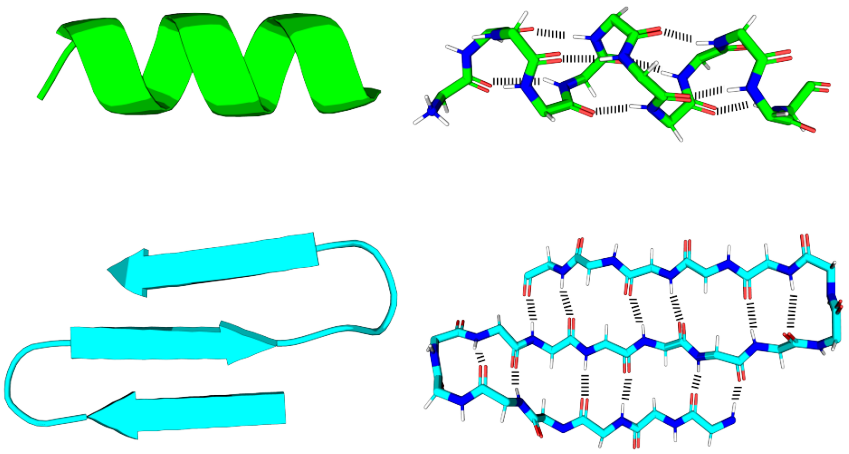


Figure 1.2: Top: A representative α-helix, taken from the first twelve residues (SIYEQYLQAKAD) of HemS. Bottom: A representative β-sheet, taken from residues 259-278 (KVTPHQDWINVFNQRFTLHL) of HemS. In the ball-and-stick models, the side chains have been removed for clarity. Dashed black lines show hydrogen-bonding between backbone atoms. There is a classic $i + 4 \rightarrow i$ pitch for the α-helix, and the β-sheet is in an antiparallel arrangement.

conditions. As life proliferated, competitive pressures (particularly due to viruses and their hosts) further necessitated protein diversification and specialisation.

To fulfil their required functions, different proteins have adopted a wide range of structural conformations. Each type of amino acid possesses a different side chain, which can broadly be categorised as charged, polar or non-polar. Most of the non-polar side chains are hydrophobic, and so a protein in aqueous solution will preferentially fold to bury as many of them as possible. This condition, combined with salt-bridge formations between the charged groups and polar-polar/hydrogen-bond contacts between the polar groups, gives a protein its overall conformation. Two main motifs that arise from such contacts are α-helices and β-sheets, and are depicted in Fig. 1.2.

Protein structure is essential for function, and the organisation of its sequence determines how a protein folds in aqueous solution. Different folds are necessary for different purposes. There are four main classes of protein – intrinsically disordered, fibrous, membrane and globular – within which there are also significant variations. Roughly speaking, fibrous proteins, such as collagen and elastin, tend to be rich in glycine and proline. Their hydrophobicity governs the formation of long, rigid regions which can aggregate, and are relatively resistant to denaturation, two features necessary for structural integrity in cartilage and tendons, for example. Transmembrane proteins (one of many membrane sub-classes), on the other hand, require

a large concentration of the residues necessary for α-helix and β-sheet formation, since their membrane-spanning domains require these motifs.[31] Such proteins often produce ion channels, thus necessitating a large hydrophobic region through which the ion can travel. Globular proteins, meanwhile, tend to be more flexible in their conformations and have a degree of water solubility. They tend to be more spherical in shape, with hydrophobic residues pointing towards their interior regions, and hydrophilic residues comprising their surfaces. This arrangement is typically necessary to store or utilise small biomolecules in the interior, whilst also allowing them to travel throughout the cell. As such, many globular proteins have roles in acquisition, transport and/or catalysis.

Proteins that specifically interact with haem are known as haemoproteins. They can be split into different classes, as discussed over the following sections.

1.2.2 Haem Acquisition

The first class of haemoproteins engage in haem acquisition. They are particularly prevalent in pathogens, as they do not typically synthesise their own haem, and so have to 'steal' it from other species. There are two main strategies a pathogen engages in to scavenge haem. The first is the secretion of siderophores, which are small biomolecules that chelate iron with high-affinity, thus solubilising it and transporting the iron to specific membrane receptors. Siderophores that acquire haem specifically are known as haemophores. Though widely used by pathogens, such a strategy is costly and even wasteful – there is an energy penalty for producing such molecules and only a fraction will return following secretion.[32]

The second uptake strategy is to have proteins capable of direct contact with exogenous iron-containing compounds (typically haem) distributed across the outside surface of the cell. These proteins are anchored in the membrane, and belong to a class known as cell surface receptors. Those proteins within this class that acquire haem typically have a β-barrel motif running through the membrane, as well as flexible extracellular loops, which can bind to haem, host haemoproteins or pathogenic haemophores.[33;34] Most also have characteristic FRAP/NPNL domains[34;35] and a histidine residue to coordinate to iron directly. These features are illustrated in Fig. 1.3. Pumping molecules across membranes requires energy and so all haemoproteins require a TonB motif, which is discussed in more detail in Section 1.3.

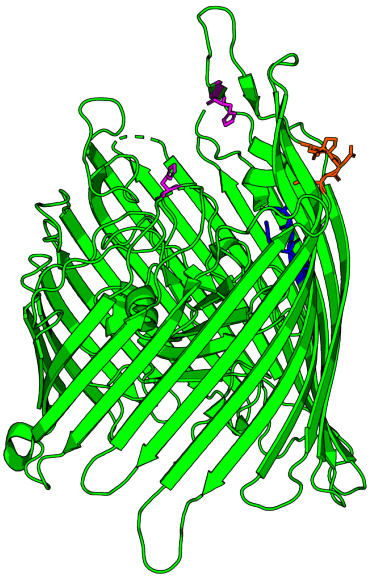


Figure 1.3: Example of a bacterial haem acquisition protein. H86 and H428 are highlighted in magenta – these residues are important for initial coordination to haem. Also highlighted are the conserved FRAP (blue) and NPNL (orange) domains. The β-barrel structure allows for haem to be protected from the outer membrane these proteins are typically embedded in. This structure is of *Shigella dysenteriae* ShuA (aka ShuR, PDB Code 3FHH),[36] an important protein which is discussed further in Section 1.3.

1.2.3 Haem Transport

The location of haem once it has been acquired by a pathogen rarely corresponds to the region where it is ultimately utilised. Generally, in a Gram-negative pathogen the haem needs to be transported from the outer membrane through the periplasm and to the cytosol for use. The process has been coined the 'periplasmic binding-protein-dependent transport (PBT)' system.[37] A complicated system of proteins is typically required to pump haem across the inner membrane. To shuttle haem through the periplasm or cytoplasm, haem transport proteins (otherwise known as haem chaperones) are required. These proteins need to be conformationally flexible to be effective at binding and releasing haem. Haem chaperones, particularly Gram-negative pathogenic varieties, are still poorly understood, with most of the current knowledge having been derived from proteins produced by the operons to be discussed in Section 1.3. Periplasmic haem chaperones are those encountered once a haem molecule has been transported through the outer membrane. As the name would imply, a periplasmic haem chaperone then shuttles the haem through the periplasm, to be deposited at the inner cytoplasmic membrane surface. Due to the relaxed porosity in the outer membrane, which allows for the influx of small, hydrophilic molecules, the periplasm tends to equilibrate to the pathogen's surrounding environment. Therefore, the periplasmic environment is highly variable in pH and

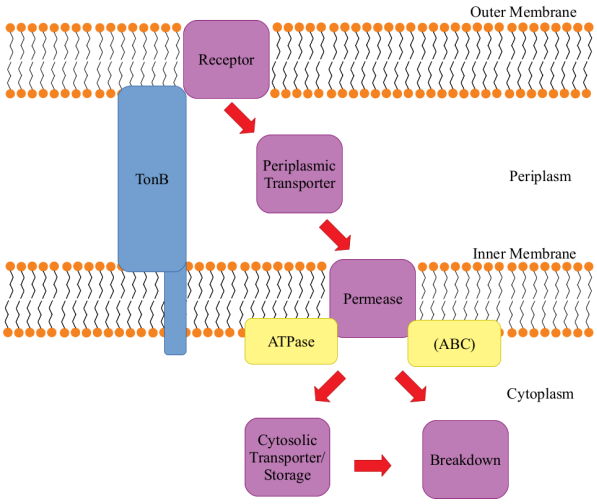


Figure 1.4: Representation of a typical bacterial haem-uptake system. A purple box indicates a haemoprotein, whereas the other proteins are auxiliaries. Haem progress through the system is charted by red arrows. ABC is the ATP-binding cassette.

salt levels. In addition, the periplasm contains a layer of peptidoglycan and has a high concentration of other proteins.[38] Periplasmic haem chaperones therefore tend to have robust yet flexible structures, making them relatively resistant to unfolding whilst tolerating small degrees of conformational change. Also, because of the dearth of adenosine 5′ triphosphate (ATP) in the periplasm and therefore of ready external energy, many chaperones are able to store energy upon substrate-binding, which is then used to power transport.[38] Though the study of these proteins has been limited, a common feature seems to be that haem is bound *via* a conserved tyrosine residue, which binds to Fe with its hydroxyl group.[39] The haem molecule is typically bound in a cleft between two subdomains of the protein,[38] and is oriented to face the relevant channel once at the inner membrane.

Once at the inner membrane, haem encounters a permease. Typically, this is quite a complex setup, involving not only the permease, but also a partner protein to induce ATP hydrolysis for the energy required to pump haem across the membrane. As with haem acquisition proteins found at the outer membrane, these permeases tend to have a large funnel running through them, thus providing a channel for the haem to travel through. They also typically have a series of residues pointing into the funnel, which are capable of binding haem *via* its iron ion, such as arginine or histidine, facilitating translocation through the membrane.[40] The ATP-binding protein, meanwhile, typically binds non-covalently to the cytosolic region of the permease. The energy it produces from ATP hydrolysis induces conformational

changes in the permease, driving haem movement through the membrane.

Once in the cytosol, haem is picked up either by an enzyme (to be discussed in Section 1.2.4) or another haem chaperone. Cytosolic haem chaperones are commonly found in eukaryotic cells. Haem synthesis in these cells is a complex process, requiring some steps to be undertaken in the mitochondrion and some in the cytoplasm. Once haem is synthesised, haem chaperones are then required to shuttle the haem to their required destination, whether that is somewhere else in the cytoplasm or to a target beyond the cell. As mentioned above, pathogens do not produce their own haem, and so there are fewer clear roles for haem chaperones to play in the cytoplasm. It is common for pathogenic cytosolic haemoproteins to be inconsistently labelled in the literature, as there is often uncertainty surrounding their function. This shall be a major topic of discussion in this thesis, where it shall be argued that the cytosolic recipients of haem derived from the *Hem*, *Hmu*, *Chu* and *Shu* operons show enzymatic behaviour, despite them commonly being classified as simple haem chaperones. The ambiguity arises because these proteins do not engage in typical haem oxygenase activity. Furthermore, there is evidence that at least some of these proteins or their close relatives engage in chaperoning haem to other haem breakdown proteins as well.

1.2.4 Haem Breakdown

Pathogens require haem for a number of processes. Some bacterial proteins require the haem ligand itself in order to function properly, and so the harvested haem is transferred directly to them. Many other processes require the iron instead, and so the haem molecule is broken down. This extracted iron could then be reconstituted for use in other iron-containing proteins.

Therefore, to extract iron, various haem breakdown strategies have been developed. Current research would suggest that the vast majority of pathogenic haem breakdown enzymes follow what is commonly known as the 'canonical haem oxygenase (HO) pathway'. This complex mechanism, which sequesters iron and produces biliverdin with the release of carbon monoxide (CO), is shown in Fig. 1.5. HOs themselves are found in almost all living organisms. In mammals, for example, the production of biliverdin from HO activity accounts for the greenish-blue colour of bruises.[41] As such, the mechanism is well-characterised – the breakdown of one haem molecule requires the consumption of three O_2 molecules, proceeding through the intermediates shown in Fig. 1.5.[42–44]

Despite their prevalence, the first crystal structure of a HO was only solved in 1999.[45] This result sparked an interest in solving such structures, and now certain

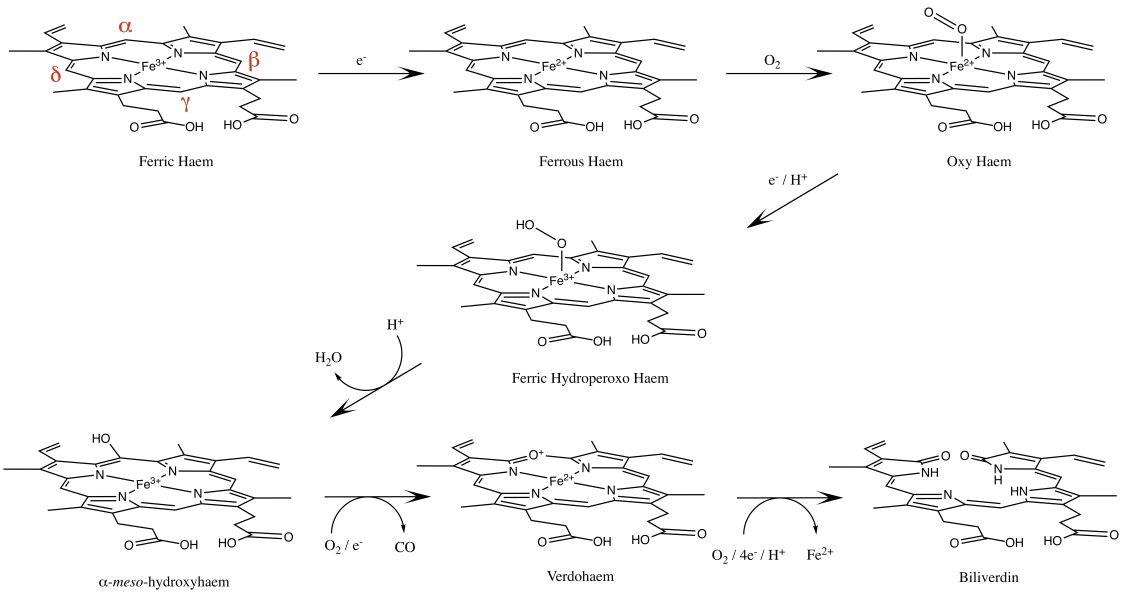


Figure 1.5: Schematic of the canonical HO mechanism, inspired by Unno *et al.*[44] Ferric haem is first reduced to ferrous haem so that it can coordinate O_2. Further reduction gives a ferric hydroxoperoxo haem. The hydroxyl group is then transferred to the porphyrin ring itself at the α-*meso*-position. Oxidation at this position yields verdohaem and expels CO. Verdohaem can then be cleaved, releasing iron and producing biliverdin.

HO motifs have been discovered and scrutinised, from a wide range of sources. One such motif is GXXXG, a conserved monomeric α-helical fold, which, in combination with a proximal histidine residue, is needed for catalysis.[46] As noted by Sawyer, the fact that these features are so well conserved across such a diverse range of sources (including bacteria, plants and humans) would suggest that other haem-degrading enzymes without these motifs are not strictly haem oxygenases.[42]

Quite apart from the useful extraction of iron, the other products from this mechanism have been shown to have beneficial properties over a wide range of downstream processes. For example, the released CO, as well as being an essential cell signalling molecule in neurons, is a potent antioxidant in higher order species such as humans. Some strains of pathogenic bacteria, such as *E. coli*, have been shown to take advantage of this effect by appropriating and sabotaging this inflammatory response.[47]

The biliverdin product can also engage in important biological processes. For example, it has been shown to inhibit viral replication.[48;49] Biliverdin can also be converted to bilirubin by biliverdin reductase (BVR). Bilirubin is an effective antimutagen, as it can scavenge hydroperoxyl radicals,[50;51] thus making it a powerful antioxidant.[52;53]

The canonical haem oxygenase mechanism is unusual in a number of ways. Firstly, it utilises haem both as a substrate and a cofactor.[54] Its controlled re-

lease of CO is also an important feature to prevent the inhibition of further haem breakdown since CO binds to ferrous haem with a higher affinity than O_2. Self-hydroxylation of the α-*meso*-carbon is also unusual as other haem proteins, such as cytochrome P450s and NO synthases, tend to heterolytically cleave a ferryl (i.e. Fe^{4+}) intermediate instead.[42]

Considering the prevalence of this reaction across such a wide range of species as one piece of evidence, and the product utility as another, it would appear as if this canonical HO mechanism is the preferred option for pathogenic haem breakdown. However, one of the main contentions of this thesis is that some pathogens may have developed alternative pathways, which can act as a 'backup' when the conditions for the canonical HO mechanism are not favourable. These conditions could include either a lack of oxygen or a lack/excess of haem.

1.3 Operons

1.3.1 Operon Structure

Proteins tend not to operate in isolation. Almost always, they are essential components within some wider biochemical framework, operating together to fulfil a specific function. For example, some enzymes need co-enzymes or chaperones to fulfil their function. Often, protein expression is therefore controlled, not on an individual basis, but by clusters of genes, which are transcribed and translated together. One example of this is an operon, which is a cluster of genes controlled by a single promoter.

More precisely, operons consist of three DNA components: the structural genes, which code for the non-regulatory proteins required by the cell; the promoter, which is recognised by RNA polymerase, initiating transcription; and the operator, to which a repressor protein can bind. Regulation of the operator varies between operons. For some, the presence of the repressor is controlled by another region in the DNA known as the regulator gene. Other biochemical stimuli control whether this gene is switched on or off, thus controlling whether the repressor is produced. In other operons, it is possible for an inducer (i.e. a small biomolecule) to displace the repressor from the operator. In both cases, the removal of the repressor allows for the transcription of the DNA.

Since the key goal of any organism is to thrive and replicate, often in harsh conditions or with a lack of resources, adaptability and energy/material conservation confer an advantage. This condition explains the utility behind the ability to switch particular operons on and off upon certain environmental signals. Protein synthesis

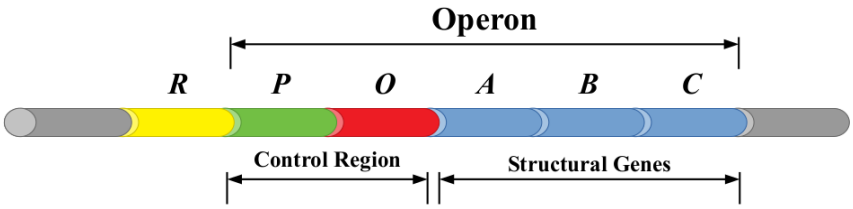


Figure 1.6: Diagram of a typical operon. The promoter, *P*, and operator, *O*, control the expression of the structural genes, *A*, *B*, *C*, etc. A regulator gene, *R*, controls synthesis of a repressor protein, which can bind to the operator. This gene is not necessarily situated near the operon.

can be a costly process, both in terms of energy and materials, and so it would be wasteful to produce some non-essential molecules during periods of scarcity. Alternatively, during periods of glut, it can be important to down-regulate the production of certain proteins, often because they interfere with or direct resources away from other proteins dealing with the excess.

Pathogens, by their very nature, are often exposed to harsh and changeable environments. They have therefore evolved a series of sophisticated operons to deal with these conditions. One such class of operon is to regulate iron content within the cell, often comprising many of the protein types discussed in Section 1.2. Bacterial iron homeostasis, as described by Andrews *et al.*, is governed by five strategies. An operon could cover some or all of these strategies, which are described in Table 1.1.[20;42]

The haemoproteins of interest in this thesis – HemS, HmuS, ChuS and ShuS – are all coded for as part of wider operons. These operons, along with the one that codes for PhuS, are all related, yet display interesting variations in their constructions. Their properties, similarities and differences shall now be described over the following sections. The *hem* operon of *Yersinia enterocolitica* shall be discussed in most detail since it codes the Hem family of proteins, which includes HemS, the most studied protein in this work.

1.3.2 The *Hem* Operon of *Yersinia enterocolitica*

1.3.2.1 The Bacterium

Yersinia enterocolitica is a Gram-negative, pathogenic bacterium. It is capable of infecting a wide range of species, including human beings, other mammals, reptiles and birds. Typically, it enters at the gastrointestinal tract and so the main cause of in-

Strategy	Problem	Response
Acquisition	Pathogen requires exogenous iron	Iron scavenged using high-affinity transport system
Storage	Exogenous iron supplies are limited	Internal deposits that were made during glut can be utilised
Consumption	External and internal iron supplies are limited	Down-regulation of certain, non-essential iron-containing proteins
Protection	Too many iron-induced reactive oxygen species	Up-regulation of proteins to store iron / degrade these species
Regulation	Combination of the above	Overall strategy to maintain iron homeostasis captured in operon

Table 1.1: Strategies undertaken by bacterial cells to maintain iron homeostasis.

fection for human beings is through the ingestion of uncooked meat or contaminated water. *Y. enterocolitica* is the strain chiefly responsible for yersiniosis, an infectious diarrhoeal disease which most commonly affects children. Though symptoms tend to be relatively benign (for example, abdominal pain and fever) these can develop into more serious conditions, such as acute gastroenteritis,[55] endocarditis,[56] mesenteric lymphadenitis,[57] or fulminant septicemia.[58] The *Centers for Disease Control and Prevention* (*CDC*), a US Federal Agency, estimates that *Y. enterocolitica* causes 117,000 illnesses, 640 hospitalisations and 35 deaths in America every year.[59] Research into this bacterium is therefore not only of academic interest but of medical importance as well.

As detailed in the sections below, access to exogenous iron is critical for the success and survival of the bacterium. Therefore, a deeper understanding of its haem-uptake and breakdown strategy could prove useful in future efforts to fight infection.

1.3.2.2 The Operon

The *hem* operon of *Yersinia enterocolitica* was the first operon that codes for haem-sequestering proteins to be discovered.[60] It consists of 9 genes, *hemWXYPRSTUV*, although two intergenic regions separate them into *hemWXY*, *hemPR* and *hemSTUV*. A restriction map showing this arrangement is given in Fig. 1.7.

It was not realised in the early studies that the *hemWXY* section was part of the operon. Therefore, it was thought that the first open reading frame (ORF) was for *hemP*, which codes for the poorly understood protein, HemP. In their original studies, Stojiljkovic & Hantke[60] were unable to isolate this protein, which is 81 amino acids long and approximately 8.5 kDa in weight, by sodium dodecylsulfate-

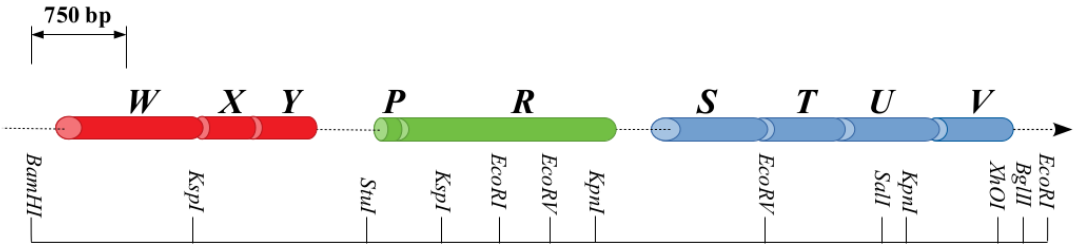


Figure 1.7: Restriction map of the *hem* operon in *Yersinia enterocolitica*. Genes are shown as colour-coded cylinders, separated by intergenic regions. Cylinder width corresponds to the number of base pairs in the gene.

polyacrylamide gel electrophoresis (SDS-PAGE). However, they did demonstrate that this protein is needed (along with HemR and HemS) for *Yersinia* to grow successfully under iron-limiting conditions.[42;60] Since then, it does not seem as if any studies have been carried out to uncover more of the behaviour of HemP in *Y. enterocolitica*. However, studies on other pathogenic hemin uptake systems have suggested that their versions of HemP/HmuP are involved in transcriptional activation of the genes encoding the outer membrane haem receptors.[61–63]

It is hypothesised by this author that HemP operates in conjunction with the Fur box, to better control haem homeostasis. Before examining this idea, the role of the Fur box must therefore be explained.

The Ferric Uptake Regulator (Fur) protein is a transcription factor that binds Fe^{2+}, which causes its affinity for certain regions of DNA, known as Fur-binding sites (FBSs), to increase 1000-fold.[42;64] These FBSs overlap with the -35 and -10 regions within the promoter. Therefore, when Fur binds, RNA polymerase is prevented from gaining access, and so transcription of the downstream genes is suppressed. The Fur box is therefore an effective feedback loop. When Fe^{2+} is relatively scarce, the Fur box sequence is free and so transcription proceeds, producing HemR, to acquire haem, and downstream proteins, to process it once in the cell. When the surroundings are rich in Fe^{2+}, the Fur box sequence is blocked and transcription is repressed. This effect limits the amount of haem entering the pathogen, mitigating against the production of reactive oxygen species. A typical Fur box is shown in Fig. 1.8.

It is possible that HemP adds another layer of complexity to the Fur box. Indeed, the overall Fur box overlaps the gene for HemP: in their original study of the *hem* system, Stojiljkovic & Hantke showed that the Fur box spans 412 nucleotides, starting at nucleotide 379, whereas the ORF for HemP ranged between nucleotides 373-615.[60] Though they did not consider HemP in their studies, Jacobi

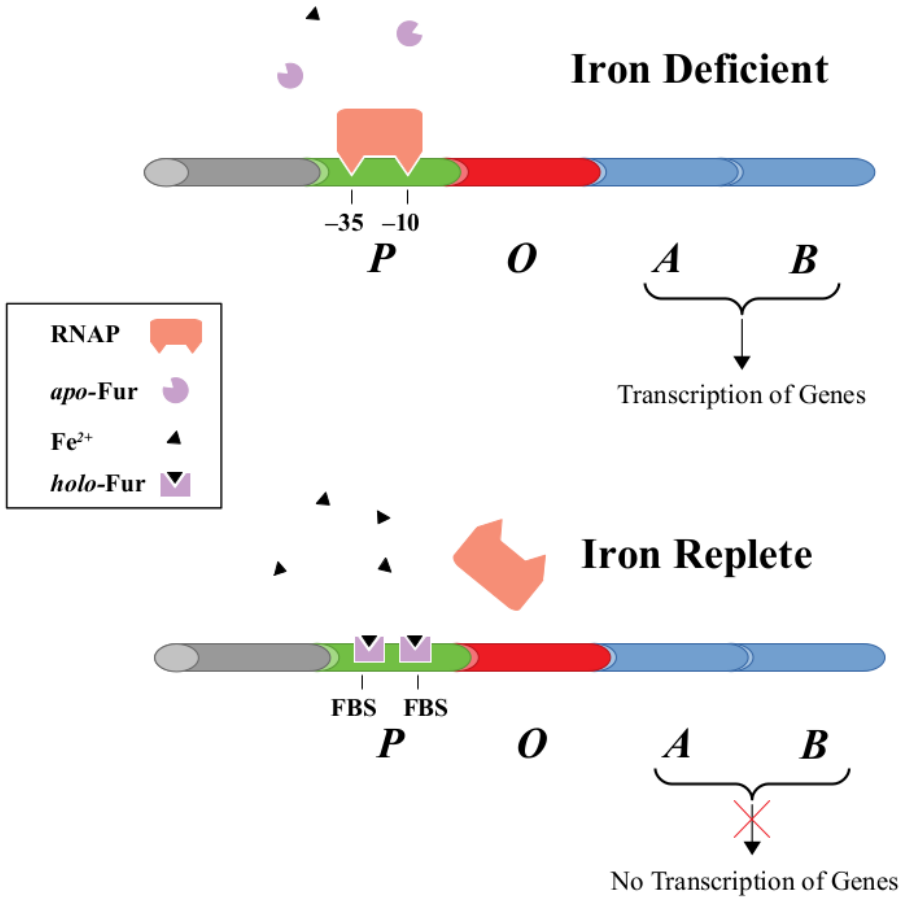


Figure 1.8: Model of Fur repression. Under iron limiting conditions, the RNA polymerase (RNAP) is free to bind to the −35 and −10 sites of the promoter, which allows the transcription of genes. Under iron replete conditions, the coordination of Fe^{2+} to Fur causes a conformational change, significantly increasing its affinity for the Fur-binding sites (FBS). As these overlap with the −35 and −10 sites, this effect prevents the coordination of RNAP to the promoter, and hence the transcription of downstream genes. The colour-coding and labelling of the genes is as in Fig. 1.6.

et al. observed three different levels of HemR expression when *Y. enterocolitica* was placed in various mouse tissues.[65] In tissues from the liver and intestinal lumen, expression was weak, reflecting the high iron content of these organs (to reiterate, *Y. enterocolitica* cannot become overloaded with iron, and so HemR production is repressed to avoid excessive absorption of haem under these conditions). Meanwhile, in splenic tissues, which have moderate iron content, there is a moderate expression of HemR. Interestingly, however, in peritoneal tissue, where there is a dearth of iron, HemR is 'hyperexpressed'. Jacobi *et al.* commented that '*Yersiniae* carrying *fyuA* or *hemR* reporter fusions exhibited threefold-stronger signals when grown in the peritoneal cavity of mice than those growing under iron derepression *in vitro*'.[65] They proposed that such an increase could be controlled by additional activators besides the Fur box. Considering the recent findings that HemP in related systems could be involved in transcriptional activation, this author tentatively speculates that the *hemPR* operon is controlled in the manner described in Fig. 1.9. Studies will be required to determine if this idea is correct or not.

The *hemR* gene produces HemR, a 78 kDa outer membrane haem acquisition protein.[60] Unfortunately, structural characterisation of HemR has been limited. LaCross *et al.* investigated HemR of another Gram-negative bacterium, Nontypeable *Haemophilus influenzae*,[66] using the I-TASSER server for protein structure and function prediction.[67–69] They even compared this result against the predicted structures of related proteins, some with very low sequence homologies, showing that they were very consistent. This observation suggests that HemR in *Yersinia enterocolitica* contains the motifs shown in these structures, namely those that are standard for haem acquisition proteins, as described in Section 1.2.2. These motifs are a β-barrel running through the membrane the protein is embedded in, as well as extracellular dangling loops, which act as receptors.

HemR in *Y. enterocolitica* is TonB-dependent,[60;70] a feature found in all iron acquisition systems described from Gram-negative bacteria. TonBs are a remarkable family of proteins, which span the periplasmic space of such bacteria. Their large, β-barrel structures allow them to transmit signals from the outer membrane to the cytoplasm. In *Y. enterocolitica*, the TonB box is located at the amino-terminus of HemR,[60] and so signalling to the cytoplasm (and therefore haemoprotein production) is intimately related to the concentration of extracellular haem being sensed by HemR.

Between *hemR* and the next ORF (required for *hemSTUV*) is an intergenic region of 120 nucleotides. Part of this region constitutes an inverted repeat,[60] probably to terminate *hemR* transcription. As with *hemR*, there is a Shine-Dalgarno

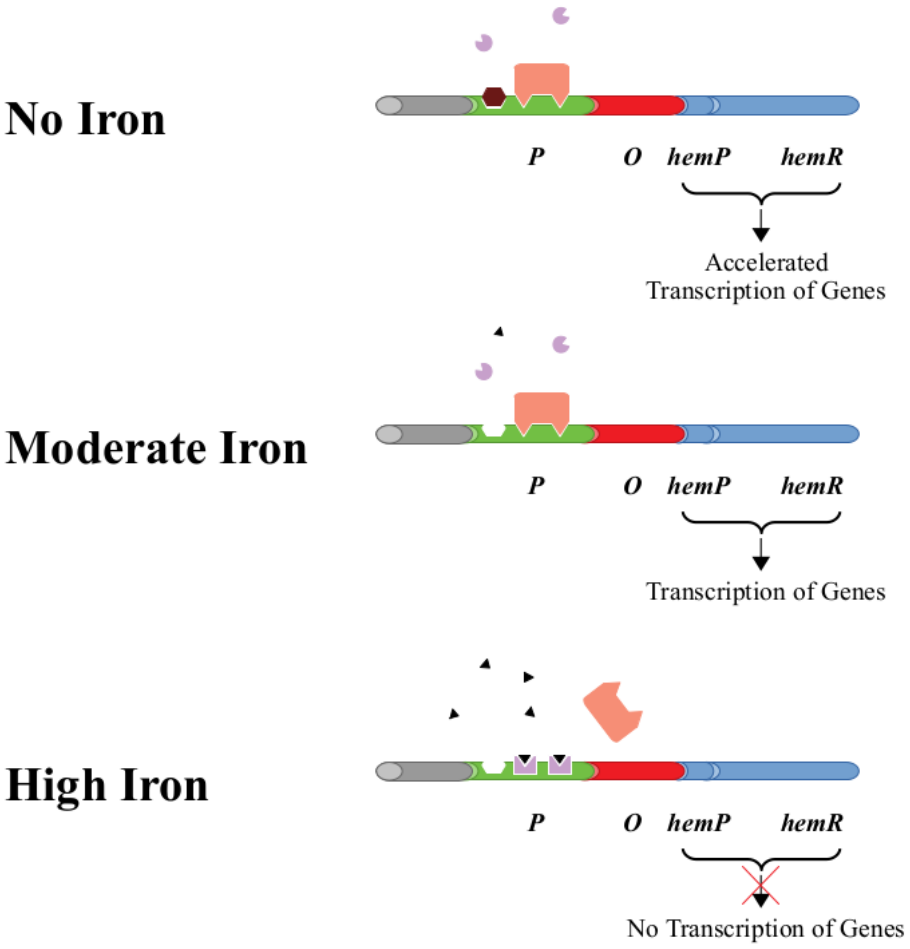


Figure 1.9: Proposed model for HemR expression based on iron availability. With moderate and high iron, expression behaves similarly to a standard Fur model, as described in Fig. 1.8. Symbols also correspond to that figure. However, at very low iron concentrations, HemR is hyperexpressed. It is thought this effect could be due to transcriptional activation by HemP (brown hexagon), although how such activation can only occur at very low iron concentrations is not properly understood.

sequence[71] approximately 10 nucleotides upstream of the start codon. Due to this intergenic region, this operon can be considered as a separate entity from the *hemPR* operon. Whereas *hemPR* is concerned with haem acquisition, *hemSTUV* is regarded as a haem-specific periplasmic binding-protein-dependent transport (PBT) system. Indeed, *Y. enterocolitica hemSTUV* was the first PBT to be characterised from any Gram-negative bacterium.[37;42] PBTs contain at least one of each of the following three classes of protein: a periplasmic binding protein; a hydrophobic protein, for spanning the inner membrane; and a hydrophilic protein which can bind ATP. Typically, they also have a fourth: a cytoplasmic protein, implicated either in haem transport or haem utilisation. In the case of *hemSTUV*, this is HemS. The structure and function of HemS shall be the focus of much of this thesis, and an introduction to this haemoprotein is presented in Section 1.4.1.

The last codon of *hemS* overlaps with the first of *hemT*, implying that they are translationally coupled. The protein, HemT, is poorly characterised in the literature. Its sequence, however, suggests it has a hydrophobic α-helical core. Its homologues HmuT, ShuT and PhuT have all been crystallised successfully. Of these proteins, HmuT, which was derived from *Y. pestis*, has the closest sequence homology (91%). This is such a close homology, it is assumed that HemT will behave in a very similar manner to HmuT. HmuT (and therefore presumably HemT) is different from ShuT and PhuT in that it can bind two stacked haem molecules in its large central cleft.[39] This dimer binds securely *via* a tyrosine (which is conserved across all periplasmic haem transport proteins) at the free axial site of one of the Fe^{2+} ions. A further histidine residue (less well conserved, reflecting the fact that not all of its homologues bind two haems) lies at the free axial site of the Fe^{2+} ion on the other haem. This structure provides a secure method with which to shuttle haem to the inner membrane.

Once at the inner membrane, haem is taken up by HemU, the next protein coded for in the *hemSTUV* operon. HemU is a permease, and as such is very hydrophobic. As with HemT, this is a poorly characterised protein, but the close homologue, HmuU from *Y. pestis* has been more deeply studied. In 2012, Woo *et al.* solved the structure for HmuU,[72] revealing that it exists as a heterodimer with HmuV, where the fully assembled transporter has stoichiometry $HmuU_2V_2$. The two HmuU subunits occupy the transmembrane domain whereas the two HmuV subunits are situated at the nucleotide-binding domain.

As clearly shown in Fig. 1.10, HmuU is rich in α-helices. Indeed, each subunit has ten transmembrane helices, giving a total of twenty in the fully assembled transporter. This structure provides the large, membrane-spanning funnel required of a

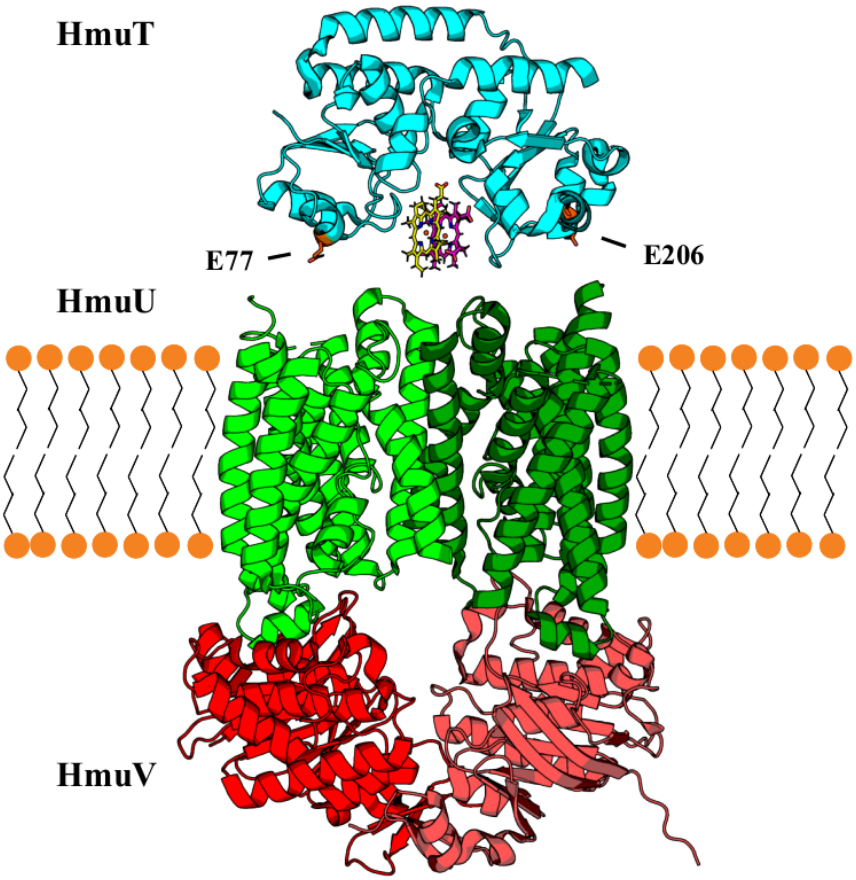


Figure 1.10: Proposed alignment of the periplasmic haem transporter, HmuT (PDB 3MD9),[39] with the membrane-spanning downstream transport system, $HmuU_2V_2$ (PDB 4G1U), as according to Woo *et al.*[72] Glu77 and Glu206, thought to be important to the docking event, are highlighted. It is thought that two stacked haem molecules can be transported through the membrane together.

permease, as described in Section 1.2.3

The crystal structures obtained by Woo *et al.* strongly reinforced some proposals concerning permease-ATPase (i.e. HemU/HmuU–HemV/HmuV) interactions made by Stojiljkovic & Hantke when they first examined the sytem.[37;72] Namely, it had been noticed that HemU, in common with other PBT permeases, such as the vitamin B_{12} transporter BtuC, has a conserved $EAAX_3GX_9LLLL$ sequence. It was suggested that this motif interacts closely with the ATPase of the PBT (i.e. HemV/HmuV). Though not remarked upon by Woo *et al.*, it can be seen from their crystal structure of the overall $HmuU_2V_2$ system, that each of the HmuU units has a sequence very nearly corresponding to this motif, of EAHYLGVNVRQAKLRLLLL. Furthermore, these motifs are intimately associated with the respective HmuV units, including a salt bridge between the first glutamic acid of the quoted sequence above and an arginine on HmuV.

This salt bridge, plus other noncovalent bonds, is typical of ATP-binding proteins in haem-uptake systems (c.f. Section 1.2.3), which are required to associate with the cytosolic region of the permease to provide energy. HemV itself is highly homologous to other known PBT ATP-binding proteins. In order to generate energy, HemV is required first to bind ATP and prime it for hydrolysis. It achieves this goal through two Walker motifs[73], which are prevalent across nucleotide-binding proteins: the GX_4GK Walker A motif, expressed as GPNGAGK in HemV, which is important for phosphoryl-binding; and the hhhhDE (where h is a hydrophobic residue) Walker B motif, expressed as WLFLDE in HemV, as aspartate, D, is required to bind the Mg^{2+} and glutamate, E, is required to effect the hydrolysis.[74]

Stojiljkovic & Hantke conducted studies to determine the importance of the *hemTUV* genes. They found that inactivation of *hemTUV* limited, but did not stop, bacterial growth. As the *hemPR* operon was left unaffected, HemR was still able to transport haem into the periplasm. Though the standard mechanism for transporting haem across the inner membrance was removed, at high enough haem concentrations it was possible for haem to seep into the cytoplasm regardless. In another study, Stojiljkovic & Hantke demonstrated that removal of HemP, HemR and HemS each caused cell death under iron-limiting conditions.[60] It does not seem as if an equivalent study was carried out in iron-replete conditions. An overall picture of the Hem set of proteins is provided in Fig. 1.11

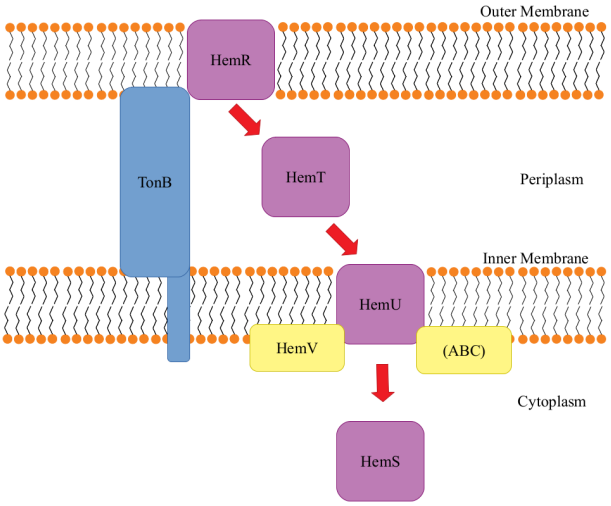


Figure 1.11: Adaptation of Fig. 1.4, applied to the Hem system of proteins. For clarity, HemP,W,X,Y are not included.

1.3.3 The *Hmu* Operon of *Yersinia pestis*

1.3.3.1 The Bacterium

Notorious for spreading the Black Death, or Bubonic Plague, *Yersinia pestis* is the most deadly pathogen in human history. It is a close relative of *Yersinia enterocolitica* and *Yersinia pseudotuberculosis*, the only other two bacteria that are pathogenic to humans from the wider *Yersinia* family. Whereas the other two target the intestine of their host and so are largely restricted to that region, *Y. pestis* developed a distinct inoculation route during the course of its divergence from an ancestral *Y. pseudotuberculosis*.[75] This distinct route is, famously, the flea bite. In common with the other pathogenic *Yersinia*, *Y. pestis* is lymphotropic[75] (i.e. it can quickly travel to and colonise lymphoid tissues) but, unlike its close relatives, is not largely localised to the intestine, thereby making it a more potent pathogen.

1.3.3.2 The Operon

The *hmu* (short for 'hemin utilisation') operon is the closest known relative of *hem*. Indeed, they and their proteins are so close it could be argued they should be regarded as interchangeable, as they often are in online databases such as Uniprot.[76] As detailed above, the *hmu* operon studied in this work comes from *Y. pestis* and, as such, has a few subtle differences from the *hem* operon of *Y. enterocolitica*. This difference is despite the operons being structured in the same way, i.e. both have three independent promoter regions, effectively splitting the operon into *hem*/*hmuWXY*,

hem/hmuPR and *hem/hmuSTUV* (c.f. Fig. 1.7). Both also have the same Fur box setup. The first piece of evidence for subtle divergences came when Perry *et al.* showed that there was only modest hybridisation betweeen *Y. pestis hmu* and *Y. enterocolitica hem* DNA.[77] In a follow-up study, they showed that these differences were mainly manifested in the HmuP and HmuR proteins.[78] In the case of HmuP, it was found to be only 41 amino acids long, nearly half the length of HemP (at 81 amino acids). This length gives credence to the findings made by Amarelle *et al.*, when investigating HmuP from *Sinorhizobium meliloti*, that this set of proteins operate as transcriptional regulators,[61] as protein length tends to be less important for function as it is in other protein classes. HmuR, meanwhile, was found to have a deleted region with respect to HemR, although in common with ChuA and ShuA, towards the carboxy-terminal region.[78] It is thought this deletion could affect the outer membrane receptors. Differences in the other proteins do not appear to be significant: in each case, the main motifs, as described in Section 1.3.2.2, are all retained.

1.3.4 The *Chu* Operon of *Escherichia coli*

1.3.4.1 The Bacterium

Escherichia coli has been studied more than any other bacterium, due to the ease with which it can be inexpensively grown and cultured in the laboratory. *E. coli* exists in many strains, most of which are harmless. Like *Y. enterocolitica*, they are Gram-negative and typically reside in the intestines of endotherms. Unlike *Y. enterocolitica*, however, *E. coli* tends not to be pathogenic, and indeed can have a symbiotic relationship with its host to prevent against pathogenic infection. The *chu* operon has been most widely studied from the *E. coli* O157:H7 serotype, however, which is a pathogenic variety.

1.3.4.2 The Operon

The *chu* (from *E. coli* haem-utilisation) operon shows both similarities and striking differences to *hem* or *hmu*. This operon is able to produce homologues of HemR (ChuA), HemS (ChuS), HemT (ChuT), HemU (ChuU) and HemV (ChuV). An equivalent for HemP is not included, but there are three other genes present, which code for ChuW, ChuX and ChuY, respectively. The overall operon structure is also controlled by four promoters, not three, and could therefore be considered as being split into *chuA*, *chuS*, *chuTWXY* and *chuUV*, respectively.

The expression of ChuA is, like HemR, regulated by a Fur box.[79] Without an

equivalent for HemP (which is also thought to be involved in HemR regulation, see Section 1.3.2.2) it does not appear as if transcription to ChuA is as tightly controlled.

The members of the periplasmic-to-cytoplasmic haem transport system, ChuT, ChuU and ChuV, have not been studied rigorously, although it is commonly accepted that they operate in a similar manner to HemT, HemU and HemV.[80] The transported haem is then picked up by ChuS in the cytoplasm. The role of ChuS, like HemS and its other homologues, is controversial, as discussed in Section 1.4.4.

The genes for ChuW, ChuX and ChuY differ from those for their homologues in *Y. enterocolitica* or *Y. pestis* in that they are clearly integrated within the overall *chu* operon. In the two *Yersinia* species, these genes precede *hemP* and, as such, have traditionally not been considered as part of the *hem*/*hmu* operons. Their characterisation is very limited; indeed, it was only through work done by the present author and, in particular, Yuhang Xie (also of the Barker Group), that it was realised that these proteins are homologous to ChuW, ChuX and ChuY.[81] The obvious names to give these proteins would be HemW, HemX and HemY in *Y. enterocolitica* and HmuW, HmuX and HmuY in *Y. pestis*. However, to confuse matters, there are other proteins coded for at different parts of the *Y. enterocolitica*, *Y. pestis* and *E. coli* genomes, respectively, which are often designated HemW, HemX and HemY. These proteins, though also involved in haem utilisation, do not have any significant homologies with those associated with the *hem*/*hmu*/*chu* operons. To avoid confusion between these proteins, and those associated with the *hem* operon in *Y. enterocolitica*, the latter shall be referred to in the rest of this report as HemW′, HemX′ and HemY′ respectively. Finer details concerning this topic, as well as a more detailed investigation into homologies between the proteins coded by the *hem*, *hmu*, *chu*, *shu* and *phu* operons, are in Appendix A.

It is perhaps because of their central position in the operon that the ChuW, ChuX and ChuY proteins have been studied more rigorously than any of their homologues. In 2009, Suits *et al.* investigated ChuX.[82] Their analysis suggested that ChuX is a cytosolic haem transport protein, much like ChuS was thought to be. Indeed, as noted by Suits *et al.*, ChuS and ChuX bear a structural resemblance, despite having low sequence homologies. ChuX (approximately half the size of ChuS) was found to dimerise, creating a fold similar to the ChuS monomer. However, they also showed that haem could bind to ChuX in a 1:1 ratio, meaning therefore that the homodimer could bind two haem molecules, rather than the one that ChuS typically accommodates. Unlike ChuS (which is known to be capable of breaking down haem using ascorbate or Cytochrome P450-Reductase, as shall be described in Section 1.4.4), ChuX was found not to have any haem breakdown capabilities. As ChuX

was shown to bind a molar equivalent of haem, and given its high K_D value, Suits *et al.* noted that such features had been observed in other related cytosolic proteins which had been determined to be involed in haem storage. They therefore suggested a similar function for ChuX.[82]

In 2016, LaMattina *et al.* built on this research, this time considering ChuW.[83] This work shifted the paradigm concerning bacterial haem catabolism, as they demonstrated a novel mechanism whereby this process could occur anaerobically. Key to this process were the ChuW, ChuX and ChuY proteins. Firstly, they noted that ChuW was a member of the radical *S*-adenosylmethionine (SAM) superfamily. Such enzymes coordinate redox active [4Fe-4S] clusters. These clusters are then used to reductively cleave SAM, creating a highly oxidative 5′-deoxyadenosyl-5′-radical (5′-dA•) that can then be used in catalysis. More precisely, ChuW belongs to the radical SAM methyltransferase (RSMT) family, which is known to transfer methyl groups to otherwise unreactive carbon atoms. Sometimes this transfer is accompanied by appreciable chemical rearrangements.[84;85] In the case of ChuW, LaMattina *et al.* found that the transfer involved cleavage of the haem porphyrin ring. The full assay, which included *E. coli* flavodoxin, *E. coli* oxidoreductase and NADPH, produced a novel compound, which LaMattina *et al.* called 'anaerobilin'. Its proposed structure is shown in Fig. 1.12, along with those for 'deuteroanaerobilin' (DAB) and 'mesoanaerobilin' (MAB). DAB and MAB were produced from the reduction of deuterohaem and mesohaem respectively.[83] Due to its greater solubility than haem under these conditions, deuterohaem proved to be a useful alternative, as it allowed for some downstream processes to be investigated in more detail.

Tandem mass spectrometry (MS^2) studies on DAB and MAB (and therefore, by implication, on anaerobilin) suggested that the porphyrin had been cleaved at the α-meso position, similar to the canonical HO mechanism described in Fig. 1.5. The observed increase in labile iron would seem to support this conclusion.

LaMattina *et al.* then introduced DAB (still working as a surrogate for anaerobilin) to ChuY. Based on its sequence, ChuY is considered to be part of the NAD(P)H oxidoreductase family. Therefore, in the assay, NADPH was also included. UV-Vis results showed the signature peaks for DAB decreasing over time. Furthermore, increasing ChuY concentration led to a linear increase in this activity, suggesting an enzymatic reaction was occurring. From these results, LaMattina *et al.* concluded that the purpose of ChuY was to reduce a potentially toxic haem breakdown product.

Though LaMattina *et al.* were not able to test this hypothesis, it seems reasonable to assume that the role of ChuX *in vivo* in this context would be to shuttle

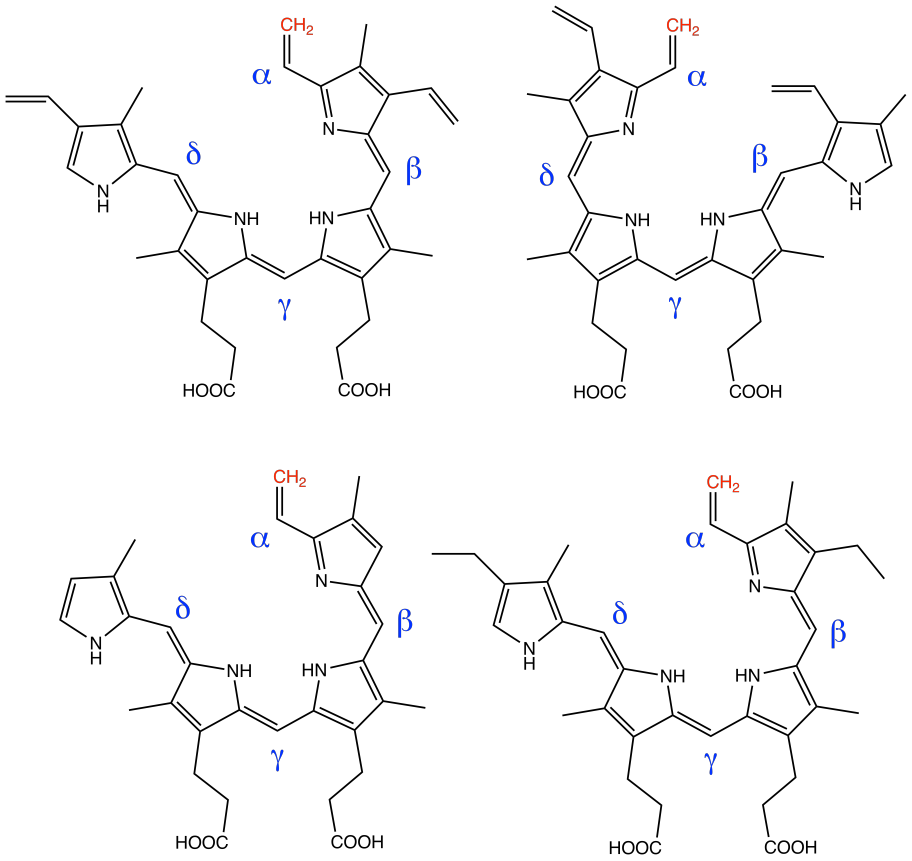


Figure 1.12: Proposed breakdown structures from LaMattina *et al.*[83] The methyl groups derived from SAM are shown in red. Top left and top right: proposed structures for anaerobilin. It is unknown which pyrrole the α-*meso*-carbon of haem remains attached to, or if both are produced, so the two possibilities are shown. Bottom left: deuteroanaerobilin (DAB) product. With respect to anaerobilin, DAB has two fewer vinyl groups, replacing them with H atoms. Bottom right: mesoanaerobilin (MAB) product. With respect to anaerobilin, MAB has two fewer vinyl groups, replacing them with ethyl groups. As with anaerobilin, it is unclear for both DAB and MAB which pyrrole the α-*meso*-carbon remains attached to. Only one possibility for each is shown.

anaerobilin from ChuW to ChuY to then be broken down further.

When discussing the role of ChuS and its homologues, many parallels between them and this ChuW,X,Y system shall become apparent.

1.3.5 The *Shu* Operon of *Shigella dysenteriae*

1.3.5.1 The Bacterium

Shigella is a genus of bacteria whose members, *S. dysenteriae*, *S. flexneri*, *S. boydii* and *S. sonnei*, it has been argued, would be better classified as strains of *E. coli*.[86;87] Infection typically spreads due to ingestion or handling of contaminated food and water. All members of the *Shigella* family have been implicated in shigellosis, an infection of the intestines, which typically causes diarrhoea and fever. *S. dysenteriae* is recognised as the most dangerous *Shigella* serogroup to human beings. It is the leading cause of dysentery epidemics worldwide, which often arise in refugee camps. It is estimated that *Shigella* accounts for over 1 million deaths across the world every year.[88]

1.3.5.2 The Operon

Considering that phylogenetic studies of *E. coli* O157:H7 suggest it should be reclassified under the *Shigella* subgenus (particularly as it contains the Shiga toxin) it is perhaps unsurprising that the *shu* and *chu* operons of *S. dysenteriae* and of *E. coli* O157:H7 are very similar. As with *chu*, the *shu* operon is split into four regions under different transcriptional controls: *shuR* (also known as *shuA*), *shuS*, *shuTWXY* and *shuUV*. Each of the genes within the operon show significant homologies to those in *chu*. There is a significant difference in the intergenic regions between *shu*/*chuR* and *shu*/*chuT* but Wyckoff *et al.* concluded that promoter elements and a Fur box were retained.[89] The Shu proteins have not been as well studied as Chu. Work that has been done would suggest close similarities. However, there are interesting differences between ShuS and ChuS, which shall be discussed primarily in Section 8.4.

1.3.6 The *Phu* Operon of *Pseudomonas aeruginosa*

The haem utilisation operon of *Pseudomonas aeruginosa*, *phu*, is not investigated in this work. However, it has intriguing similarities and differences from those operons which are studied, and so it is worth briefly considering.

1.3.6.1 The Bacterium

Though a Gram-negative Gammaproteobacteria like the other species considered thus far, *P. aeruginosa* is not a close relative to any of them. It is considered to be opportunistic, primarily attacking organisms that are immunocompromised or suffering from existing diseases, and can enter a human host *via* a variety of methods such as the urinary tract, burns or exposed wounds. As such, it is not as specific as the bacteria considered so far (*Y. pestis* excluded), which primarily target the gut, yet it still requires exogenous iron to survive and replicate. One haem-uptake strategy it uses is to produce siderophores, which capture and transport iron. This method can be wasteful, however, since the iron-siderophore complexes are not specific. This issue is exacerbated in populations of *P. aeruginosa* containing forms that do produce siderophores (known as cooperators) and mutated forms (known as cheaters) which do not. The cheaters can pick up siderophores, which they themselves did not produce, thus depriving the cooperators, which expend a significant amount of energy to produce these siderophores in the first place. In a mixed population, the cheaters therefore outcompete the cooperators, weakening the overall population over time and decreasing virulence.[90;91] Mitigating this effect are alternative haem-uptake strategies, such as that produced by the *phu* operon, as shall now be described.

1.3.6.2 The Operon

Unlike the operons discussed so far, the *phu* operon does not contain genes that code for the W, X and Y proteins. Those homologous proteins that are present are split into two regions under the control of different promoters and Fur boxes: *phuR* and *phuSTUV*. Following *phuV*, a further protein is coded for which was originally regarded as PhuW.[92] However, it bears little homology to other W proteins (with 11% identity to ChuW and HemW′ respectively). Instead, this molecule is now regarded as a ChaN lipoprotein, since it was found to have a 30% identity with ChaN from *Campylobacter jejuni*.[93] Rather than acting as a cytosolic haem breakdown enzyme, ChaN lipoproteins are thought to associate with the outer membrane, operating in some sort of partnership with ChaR (PhuR in *P. aeruginosa*).[94] Ochsner *et al.* showed that removal of this *chaN* gene limited, but did not stop, bacterial growth when haemin was the only exogenous source of iron. This arrest also proved to be the case when the *phuR* gene or the *phuSTUV* operon were removed.[92]

Whereas homologies between the proteins in the operons discussed previously tend to be relatively high (tending not to drop below 50% identity or 70% similarity) those proteins that are shared between them and the *phu* operon tend to be less

similar. This situation is not surprising, since *P. aeruginosa* is not of the same taxonomic order as the other bacteria considered. Despite this difference, the R, T, U and V proteins are thought to fulfil the same functions as those from the other operons. As far as the author is aware, none of these proteins have been extensively characterised (e.g. by crystallography or NMR studies). The exception is the S protein, PhuS, which has been shown to have interesting structural and functional differences compared to its homologues. This shall be discussed in Section 1.4.6.

1.4 HemS and its Homologues

1.4.1 HemS

HemS from *Yersinia enterocolitica* is the main focus of this thesis. The precise function of this protein has long been under debate. When they first examined the *hem* operon, Stojiljkovic & Hantke concluded that HemS 'could be either a cytoplasmic membrane permease that transfers hemin into the cytoplasm or a hemin-degrading enzyme.'[60] Though it has been shown since then that HemS is not a permease, it has remained unclear whether it is a haem transfer protein or a haem-degrading enzyme.

Despite not knowing its function, it was determined from the beginning that HemS was a key component of the *hem* operon. In their original study, Stojiljkovic & Hantke showed that HemR and HemS were both required for haem to be used as an iron source: gene knockouts led to cell death.[60]

The crystal structure of HemS was solved in 2006 by Schneider & Paoli, becoming the first cytosolic haemoprotein of its class to be resolved in both its *apo-* (without haem, PDB code 2J0R)[95] and its *holo-* (haem-bound, PDB code 2J0P)[96] forms.

Such studies showed that HemS is a large, 41 kDa protein consisting of two topologically homologous domains joined by an unstructured loop to give a pair of large, stacked β-sheets.[95] The fact that these domains are so similar has led to speculation that HemS is actually a fusion of two originally separate proteins. In the closely related *chu* operon (discussed in Section 1.3.4.2), for example, Suits *et al.* noted a close structural similarity between ChuS and the ChuX dimer, and hypothesised that both these proteins were fulfilling a similar function.[82] The equivalent protein in *hem*, HemX′, is poorly studied – it would be of interest to determine whether it bears the same relationship to HemS, as ChuX does to ChuS.

The unstructured loop connecting these domains is a poorly understood region, with its high conformational entropy making crystallographic resolution difficult. In the literature, the structure of this region is not properly known. In this thesis, a

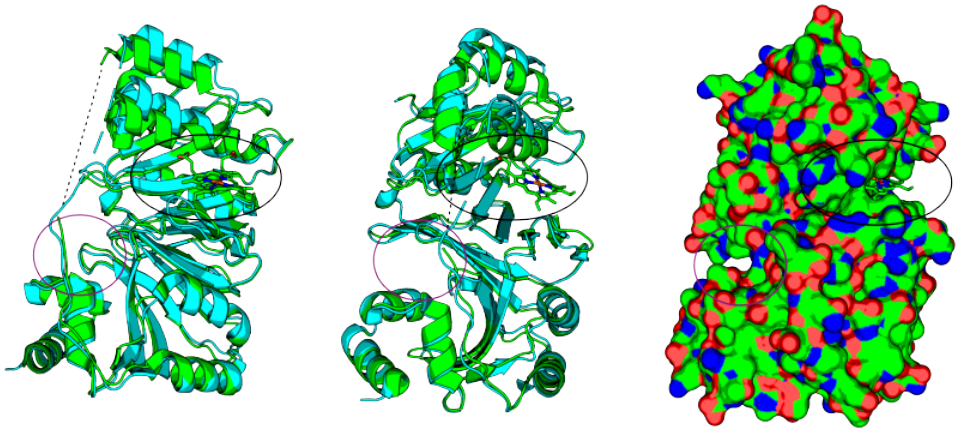


Figure 1.13: Representations of HemS. Left: *holo*-HemS (PDB 2J0P, green) superimposed on *apo*-HemS (PDB 2J0R, cyan). Middle: rotated version of left-hand representation. Right: Surface representation of *holo*-HemS. Dark blue and red colours represent nitrogen and oxygen atoms, respectively. In all of the representations, the black circle corresponds to the large cavity, and the purple circle to the small cavity. A dashed line is provided to show the missing loop region from the *holo*- structure. The *apo*- loop is also incomplete. The structural overlay clearly shows the large cavity 'clamping down' upon haem-binding when compared against the *apo*-form.

combination of further computational modelling and further X-ray crystallography shall shed more light on this region.

The stacked central β-sheets of HemS are twisted and capped by α-helices to form two distinct pockets. Haem can bind within the larger, deeper pocket. A histidine (H196) binds to one of the free iron axial ligands, anchoring the haem inside the pocket. It is less clear what is at the sixth coordination site, since previous crystallisation studies around this region could not be tightly refined, though it is strongly suspected that it is either a H_2O molecule or OH^- ion, clamped in place by an arginine (R102) residue. As well as binding to iron, further residues in HemS appear to have a role in binding to the propionate groups of haem. R209, K294, Q316, Y318 and R321 together form a polar, solvent-inaccessible region suited to these propionates. Whereas R321 coordinates to the more exposed propionate, R209, K294, Q316 and Y318 all form polar bonds with the propionate which is more deeply buried inside the pocket.

Inclusion of haem within this pocket leads to an induced fit conformational change. The C-domain moves towards the N-domain to facilitate H196-binding to the iron ion, clamping the haem molecule more tightly in place. This change causes the further burial of 350 $Å^2$ of solvent accessible surface area (SASA) on top of that occupied by the haem itself,[21] giving the docking event a considerable entropic drive to complement the favourable enthalpic drive caused by the new intermolecular in-

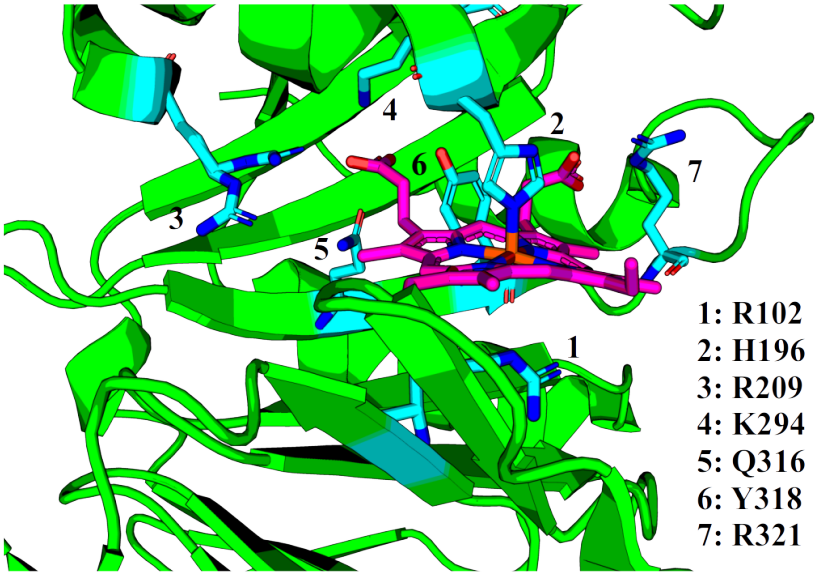


Figure 1.14: Close-up representation of the *holo*-HemS haem-binding pocket from 2J0P. The most important residues implicated in haem-binding, as detailed in the text, are shown in cyan.

teractions. The tight binding between H196 and the iron ion results in significant buckling and distortion to the porphyrin ring, suggesting that HemS is priming haem for degradation, rather than just transporting it within the cytoplasm.

The following sections show that there are indeed molecules that can react with *holo*-HemS to give novel, breakdown products. As these molecules, in particular NADH, are discussed, further residues shall be highlighted that are thought to be important to NADH-binding/regulation. First, however, a deeper discussion of the nature of haem-binding in HemS is required.

1.4.2 Haem-Binding in HemS

An in-depth study of haem-binding was conducted by Sawyer as part of her PhD at the University of Cambridge.[42] Though unpublished, these data are important for a proper understanding of this protein's relationship with haem.

Sawyer showed that the UV-Vis spectra of *holo*-HemS is dependent on pH. As pH was increased from 4 to 9, the Soret band shifted from 403 nm to 408 nm. This shift was accompanied by an increase in intensity, indicating stronger haem-binding at higher pH. It must be remembered that haem typically has lower solubility in more acidic solution, and that protonation of amino side chains around the haem-binding pocket are possible. However, this shift in intensity was regarded as too great to be

assigned just to these effects. Instead, this change of intensity was thought mainly to be due to changing affinities of the water/hydroxide acting as the sixth haem iron ligand, and of the pK_a of this group. Data fitting showed this pK_a to be 5.5, lower than for other reported proteins with hydroxide ligands.[97] Increasing the pH also led to a decrease in absorbance at ∼380 nm and at ∼650 nm, suggesting a transition to low-spin iron. Overall, therefore, it would seem that at high pH, haem is a low-spin, 6-coordinate, hydroxide complex. At low pH, the wavelength of the Soret peak would suggest a high-spin, 6-coordinate, water complex but the pK_a would instead indicate a 5-coordinate complex. Some sort of equilibrium is assumed, and because water is a weak ligand, it is recognised that haem iron may fluctuate between high- and low-spin states.

Over a series of difference spectra, in which a constant concentration of HemS was mixed with different concentrations of haem solution, Sawyer showed that the Soret peak reaches maximum absorbance once haem is equimolar to the protein. Thereafter, this peak remains approximately steady, but the overall spectra continues to change. This observation suggests that there is saturation of the known binding site, but then further protein-haem interactions can occur elsewhere. In other words, there is one favoured binding site but other (relatively spectroscopically 'silent') ones may be possible.

To determine whether HemS could accommodate more than one haem at a time, further studies were required. There was also a question as to whether these excess haem molecules were binding separately or as dimers. An analysis using non-denaturing mass spectrometry (ndMS) was therefore carried out. Here, it was important to be able to run MS on the entire samples injected, *and* on ions of particular m/z values selected in the quadrupole. In the first case, this experiment allowed for all the species to be separated at the end of the time of flight (ToF) tube and then be detected. This setup allows for an overall picture of protein-ligand binding stoichiometries to be developed. In the second case, only the daughter ions of the species selected in the quadrupole could be detected. This latter case is known as tandem MS. By explicitly selecting a particular protein-ligand stoichiometric combination, the ligand molecules stripped from the protein could be unambiguously identified with that parent ion.

Results from this experiment showed that HemS was indeed capable of binding more than one haem molecule at a time. Ratios up to *holo*-5-HemS were unambiguously observed; it is possible that higher stoichiometries were present, but the spectra become difficult to deconvolute in the regions where these species would be expected. Furthermore, haem could bind as monomers or dimers. There was a

particular preference for monomers, especially at low haem : protein ratios.[42]

Choy, as part of his PhD at the University of Cambridge, probed this hypothesis further by computation.[21] He showed that it might be possible both for haem to bind to the N-terminal pseudo-pocket in addition to the main pocket (termed 1,2-*bis*haem-HemS) or to bind in the main pocket as a dimer (1,2-*di*haem-HemS).

The pseudo-pocket has some properties that would appear to make it suitable for haem-binding. This situation was commented upon in the previous section when discussing the possible origins of HemS, and how it may have been a fusion of two smaller haem-binding proteins. This pseudo-pocket has an R-Q-Y-K-R line of residues in common with the main pocket, where both arginines and the lysine are implicated in propionate recognition, and tyrosine in iron coordination. Choy showed, by superimposing a second haem molecule onto his *holo*-HemS structure from experiment and then optimising, that haem could also bind inside this pocket and interact with these residues. However, the space available in this pocket was found to be very tight. It was further noted that there was a lack of conformational flexibility, and that there was an absence of some further propionate-recognising residues otherwise found in the main pocket. Altogether, this situation suggested that HemS would only use this pocket for haem-binding if the other was already occupied, in agreement with experiment.

Choy was also able to superimpose haem dimer into the main pocket and optimise to stable structures. He showed that these dimers were able to overlap in slipped-parallel mode with the propionates either perpendicular or antiparallel to one another. Having a second haem in this pocket precludes NADH from entering, thereby preventing the novel haem breakdown discussed in this thesis (to be introduced in Section 1.5).

1.4.3 HmuS

The structure of *Yersinia pestis* HmuS has not been resolved in either its *apo*- or *holo*-forms. It does, however, have 89.6% identity to HemS, retaining all of the important residues listed above (i.e. R102, H196, R209, K294, Q316, Y318 and R321). Of those residues which do differ, the changes tend to be cosmetic (% similarity between the two proteins is 94.8%). It is therefore thought, in the absence of detailed evidence, that the structures of HmuS and HemS are essentially identical.

1.4.4 ChuS

The structure and function of ChuS have both been studied in great detail. Its *apo*-crystal structure was first resolved by Suits *et al.* in 2005 (PDB Code 1U9T),[80] with the *holo*-form following the year after (PDB Code 4CDP).[98]

Due to the crystal structures for ChuS being solved only a matter of months after those for HemS, a detailed comparison between them was not made at the time of publication. However, the similarities are abundant. As with HemS, it was found that the N- and C-terminal halves of ChuS represented a structural duplication, giving a root mean square deviation of 2.1 Å between the repeats, despite them only having 19% sequence identity.[80] From this result, Suits *et al.* concluded that ChuS was fulfilling a similar role to ChuX, which could dimerise to give very similar pockets to those in ChuS, as noted in Section 1.4.1. Just like HemS, these two domains in ChuS are connected by an unstructured loop, which could not be entirely resolved in either the *apo*- or *holo*-forms. All of those residues important for haem docking in HemS noted in Section 1.4.1 (R102, H196, R209, K294, Q316, Y318, R321) are present in ChuS.

There are points of deviation, however, between HemS and ChuS. Firstly, their % Identity and % Similarity are 66.8% and 78.2%, respectively. A region with many differences, including the deletion of a glutamic acid found in HemS but not in ChuS, is the outermost α-helix of the α-loop-α-loop-α motif in the C-terminal domain. These differences result in this helix being shifted further from the central cavity in the *holo*-form (when compared against HemS). Due to the intimate connections between the α-helices of this motif, this change causes the innermost helix (that which directly forms the cavity) to also be shifted away, thus creating a larger cavity. However, when haem is not present, this effect is reversed; in other words, the HemS cavity is wider. Taken together, these observations suggest HemS 'clamps down' more than ChuS upon haem inclusion, as illustrated in Fig. 1.15. This difference between the two homologues would perhaps suggest that HemS is more effective at haem-binding than ChuS, as discussed in the Results.

The *holo*-ChuS crystal structure was of high resolution, allowing haem-binding to be closely investigated. Histidine, H193, coordinates to the iron ion at the proximal side, whereas an arginine, R100, is involved in coordination at the distal side, much as in HemS. Unlike HemS, however, this distal region was clearly resolved, showing two water molecules involved in coordination situated between R100 and the iron ion. Spectrally, haem-binding in this pocket results in a Soret maximum at ∼408 nm (plus further sets of peaks for the β-band at ∼545 nm and for the α-band at ∼580 nm).[80] Suits *et al.* showed that this Soret band depended on the H193 residue binding

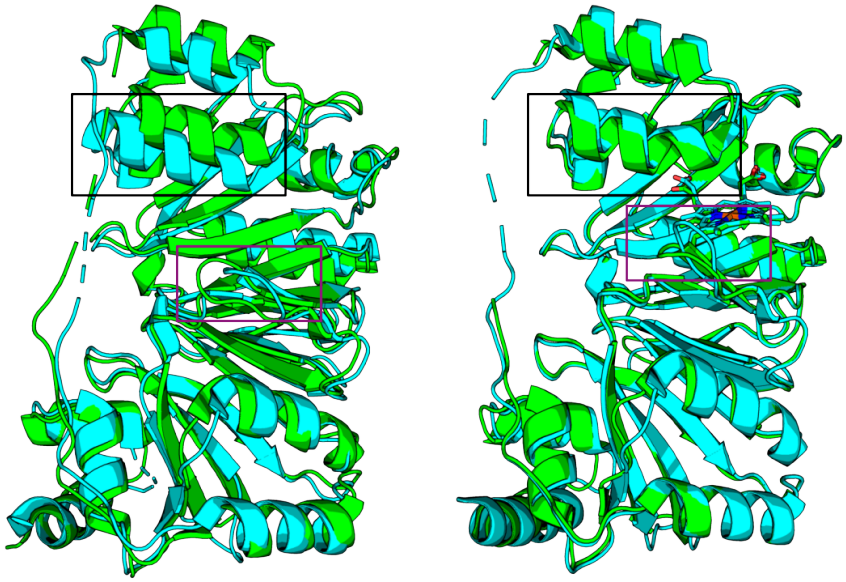


Figure 1.15: Comparison of HemS and ChuS structures. Left: overlay of *apo*-HemS (2J0R, green) and *apo*-ChuS (1U9T, cyan).[80] Here, the α-helix capping the central cavity is buried deeper in ChuS than in HemS (black square), thus restricting the pocket size. Right: overlay of *holo*-HemS (2J0P, green) and *holo*-ChuS (4CDP, cyan).[98] Here, the relative positions of the capping α-helices (black square) have now been reversed: the helix for HemS is now buried deeper in the pocket. It appears that inclusion of haem causes a more significant 'clamping' effect than it does for ChuS, thus explaining this reversal. The purple squares in each representation also highlight a conformational difference in a central loop between the HemS and ChuS structures. H196 is the preferred residue for haem iron-binding. However, this highlighted loop contains two further histidines – H85 and H89 in HemS, and H83 and H87 in ChuS. Their close proximity to H196 would suggest some sort of role in haem-binding. Therefore, such conformational differences in the loop could affect the relative abilities of HemS/ChuS to 'capture' and store haem.

to haem: mutations to the histidine, such as H193N, were shown to broaden the absorbance and shift the maximum to ∼390 nm.[98] In the wild type, ChuS was further shown to 'clamp down' on haem when the latter was introduced to the pocket, bringing the histidine closer to the iron ion. As mentioned above, however, this conformational change was less drastic than with HemS.

In their study of the structure of *holo*-ChuS, Suits *et al.* became the first to discover that this class of proteins can act as enzymes to promote haem degradation. Noting that typical bacterial HOs either use the Cytochrome P450 Reductase (CPR)-NADPH system, flavodoxin, ferredoxin or ascorbic acid as reducing partners in vitro,[80] they opted for ascorbic acid, introducing it to a *holo*-ChuS solution. Over time, the Soret band at 408 nm was shown to decrease, with a complementary increase then decrease in the near-IR (suggesting an intermediate) and a slower but steadier increase of absorbance at ∼560 nm (suggesting a final breakdown product). Further studies showed significant CO production when *holo*-ChuS was treated with ascorbic acid, an effect that could not be replicated when the protein was not present. They further showed that CPR-NADPH could be used as a reducing partner, although this produced less CO.

In 2016, Ouellet *et al.* continued this study to a detailed level. They proved that this reaction is aerobic and that hydrogen peroxide, H_2O_2, is a crucial component. More precisely, they demonstrated that the sodium ascorbate (c.f. ascorbic acid) was used to produce H_2O_2, which would then react with haem.[99] Adding catalase (a family of enzymes that decompose H_2O_2 to water and oxygen) to the reaction mixture slowed haem-degradation in a concentration-dependent manner. This observation proved the presence of H_2O_2 in solution, differentiating this reaction from those that follow a 'canonical' HO pathway, as described in Section 1.2.4 and in Fig. 1.5. It is not entirely clear how this H_2O_2 is produced. Ouellet *et al.* did, however, show that the synthesis of H_2O_2 in a solution of ascorbate increased upon the addition of *holo*-ChuS. As this increase did not correlate linearly with an increase in ChuS, it was concluded that this non-correlation must be due to the competing haem breakdown reaction (which results in H_2O_2 consumption), which ChuS then promotes. Ouellet *et al.* suggested that peroxide formation would be 'produced indirectly after the formation of superoxide.'[99] To demonstrate conclusively that H_2O_2 in solution was required for the reaction, Ouellet *et al.* set up an assay whereby peroxide was added to *holo*-ChuS rather than ascorbate. The same products were formed.

Ouellet *et al.* also monitored *holo*-ChuS and H_2O_2 under anaerobic conditions. Interestingly, the Soret maximum disappears over time, suggesting haem is either breaking down or being displaced, but there is no concomitant increase of absorbance

at ∼560 nm.[99] Unfortunately, these data have never been made publicly available as far as this author is aware, so a comparison with the novel reaction discussed in this thesis cannot be made.

Returning to the aerobic reaction, Ouellet *et al.* also closely studied the breakdown products. The peak at 565 nm did not indicate biliverdin formation, the expected product of a canonical HO reaction (which typically has peaks at ∼373 nm and ∼668 nm). However, upon complexation with pyridine of the intermediate species (i.e. pyridine was added to the reaction mixture when the near-IR species was most prevalent), the resulting spectrum was characteristic of a *bis*-pyridine Fe^{2+}-verdohaem complex,[99] with absorption maxima at 395, 499, 533 and 660 nm.[100] This structure was illustrated as part of the reaction pathway for a canonical HO shown in Fig. 1.5. As shown in that figure, verdohaem appears to be the penultimate structure before biliverdin, thus suggesting that the aerobic reaction of H_2O_2 with *holo*-ChuS perhaps overlaps at least to some degree with a typical HO reaction, but then diverges when it comes to the further breakdown of verdohaem. Pyridine was shown to have little effect on the spectrum of the final breakdown product, with the peak at ∼565 nm remaining, suggesting it was unable to coordinate.

Using a combination of electrospray ionisation mass spectrometry (ESI-MS), electron paramagnetic resonance (EPR) spectroscopy and nuclear magnetic resonance (NMR) spectroscopy, Ouellet *et al.* further characterised this breakdown product. From MS, they saw masses at 514.32 *m/z*, which disappeared as time progressed, and 437.19 *m/z* (for comparison, the mass of haem and PPIX are 616 Da and 563 Da respectively). The mass at 437.19 *m/z* corresponds to $C_{24}H_{26}N_3O_5$, suggesting a net loss of 10 C, 8 H, 1 N and a gain of 1 O with respect to PPIX ($C_{34}H_{34}N_4O_4$).

NMR studies clearly identified the formation of hematinic acid, a molecule with one pyrrole and formula $C_8H_9NO_4$. This formation would strongly suggest that PPIX loses one of its four pyrroles during the course of the reaction, and that the 437.19 *m/z* species is therefore a tripyrrole product. This reaction constitutes a significant difference from a canonical HO: even although the final product, biliverdin, is cleaved at one of its *meso*-positions, it retains all four of its pyrroles. Ouellet *et al.* were unable to determine the exact structure of this tripyrrole, but their working hypothesis for what this could be is shown in Fig. 1.16.

Recently, Mathew *et al.* proposed that ChuS could also act as a haem chaperone under anaerobic conditions.[101] They noted that, in the absence of molecular oxygen, ChuS can still bind, but not degrade, haem. Indeed, *holo*-ChuS can be isolated at substantial yields;[80;98] Mathew *et al.* quote 2 mM.[101] It was therefore concluded that

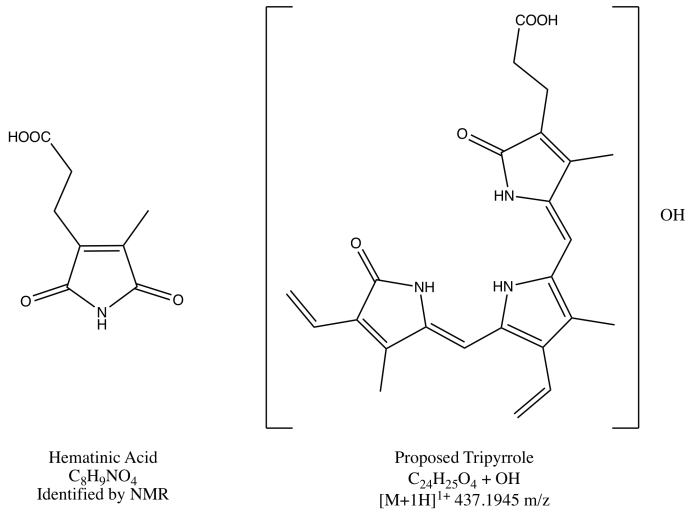


Figure 1.16: Confirmed hematinic acid fragment and proposed tripyrrole structure from Ouellet *et al.*[99] It would seem unlikely that the *m/z* fragment would retain a hydroxide counterion. Without the benefit of further evidence, this author would therefore suggest that the hydroxide should instead be incorporated covalently within the tripyrrole structure, though it is not clear how.

ChuS could store haem under anaerobic but iron-replete conditions (thus preventing toxicity caused by high cytosolic haem levels), and perhaps even deliver haem to ChuW for this latter enzyme to catalyse a SAM-mediated anaerobic breakdown (see Section 1.3.4.2 for details). Mathew *et al.* tested this hypothesis by running two ChuW-based assays where the only difference was in the substrate used: haem in one, and *holo*-ChuS in the other. The resulting UV-Vis spectra showed that anaerobilin (which has distinct peaks at 445 nm and 795 nm, and to a lesser extent at 570 nm) was produced in each case, strongly suggesting that *holo*-ChuS was transferring haem to ChuW for the latter to then break down. NADPH was used in both these assays which was of interest to the author because NAD(P)H is a reagent required for the anaerobic breakdown of haem in ChuS and its homologues, as shall be described in the Results. It is unclear whether the inclusion of other reagents in these assays precluded this reaction from occurring or whether it was outcompeted by the ChuW-mediated reaction, with the spectra for anaerobilin masking that for the haem breakdown product from ChuS (which has overlapping maxima at ∼590 nm and ∼810 nm).

1.4.5 ShuS

As with HmuS, the crystal structure for ShuS has not yet been determined. It does, however, retain 98.5% sequence identity with ChuS, and so it is predicted that the two would be structurally very similar. Paradoxically, however, investigations into

the properties of ShuS (undertaken by a different group from that which studied ChuS above) have led in a very different direction.

As far as this author is aware, the possibility of haem breakdown in ShuS has not been investigated in detail. The Wilks group at the University of Maryland tried to determine whether there was any HO activity in 'standard NADPH or ascorbate based assays of ShuS' and found there was none.[102] Unfortunately, the exact details of these assays and their results were not published. This study is particularly relevant to this thesis as it has been determined that *holo*-ShuS can indeed react with NAD(P)H, albeit not *via* a typical HO pathway (details of which shall be forthcoming in Chapter 9).

Instead, the binding of haem to ShuS in competition with DNA was the primary focus. Wilks was the first to show that DNA can bind competitively to *apo*-ShuS.[102] Such DNA-binding is non-sequence-specific. At first, it was thought that this DNA-binding could be analagous to Dps, a protein which can sequester and protect cellular DNA from oxidative stress or periods of low nutrient availability. However, a later study by Kaur & Wilks showed that the breakdown of DNA was only limited to a small degree by ShuS when exposed to oxidative stress.[103] Why DNA binds to ShuS is still a mystery, although a recent study, also from the Wilks group, on the homologue PhuS may be able to shed some light on this result, as discussed presently.

1.4.6 PhuS

PhuS is a more distant relative from HemS than any of the other homologues discussed thus far. Its % Identities with HemS, HmuS, ChuS and ShuS are quite low for them to be considered as homologues, at 42.6%, 43.8%, 41.1% and 40.8% respectively. This protein was not part of the present study, but some of its structural and binding properties are worth considering.

PhuS has been demonstrated to fulfil all of the proposed functions suggested for its homologues above: it has been shown to chaperone haem to other haem breakdown enzymes, degrade haem itself and bind to DNA.

Wilks *et al.* have, as with ShuS, taken the lead in the study of PhuS. In 2006, they discovered that haem in PhuS could be broken down by ascorbate or the CPR-NADPH system.[104] They noted that this was equivalent to experiments conducted on ChuS by the Jia group (mostly led by Suits) at Queen's University, Kingston,[80] as was discussed in Section 1.4.4. The Wilks group drew different conclusions from their results compared to the Jia group. Whereas Jia had taken this reaction to indicate HO activity, Wilks urged caution and argued that such a reaction did not

conform to the ordinary standards for a HO. Instead, it was taken to be a coupled oxidation reaction. This is a nonenzymatic class of reaction discovered nearly a century ago[105–107] which does not even necessarily require the presence of a protein; proposed mechanisms for this process, as gathered from the literature (especially Avila *et al.*)[108] are given in Fig. 1.17. Rather, haem in pyridine/water can be degraded aerobically by hydrazine or ascorbate. Though verdohaem is produced (and this verdohaem can then be hydrolysed to biliverdin upon addition of KOH and HCl), this reaction is unlike the one promoted by a canonical HO as the products are not regioselective. A canonical HO reaction almost exclusively attacks haem at its α-*meso*-carbon, whereas a coupled oxidation reaction can attack any of the four *meso*-carbons. In both a canonical HO reaction and a coupled oxidation, however, CO is produced. As the Jia group had claimed HO activity for ChuS based on their observation of the production of CO,[80] the Wilks group were therefore critical of this inference.[104] Wilks also found that adding catalase to the reducing mixture stopped the reaction with PhuS,[104] contradicting the claims of Jia that in ChuS, 'addition of catalase and superoxide dismutase did not affect the trends of the spectral change.'[80;98] Since catalase can break down H_2O_2, which is free in solution, and that a coupled oxidation reaction uses H_2O_2 whereas a canonical HO does not, Jia used their data to indicate that ChuS was indeed a HO. However, in a later paper, their position changed, and they too stated that addition of catalase inhibits the breakdown of haem in ChuS.[99] These findings all seemed to reinforce Wilks' position that the ascorbate-based assays conducted on ChuS and PhuS were both indicative of coupled oxidation. However, in this same later paper by Jia, the products were closely studied, as was described in Section 1.4.4. They found that the breakdown of haem in ChuS *was* regioselective, producing hematinic acid and a single tripyrrole. This outcome would suggest an enzymatic role for ChuS after all.

It is therefore still unclear whether ChuS/PhuS are simply able to produce H_2O_2 from ascorbate, which then attacks haem *via* a couple oxidation process, or whether they are indeed aerobic haem breakdown enzymes. Due to the regioselectivity of the reaction, it would seem to be the latter situation, although the reaction does not seem to bear as much resemblance to a canonical HO as originally thought.

Due to their scepticism surrounding the enzymatic role of PhuS, the Wilks group focussed more on its context within the cell, and whether it could be shuttling haem to other haem breakdown enzymes. They demonstrated that PhuS can transport haem to a *bona fide* haem oxygenase found in *Pseudomonas aeruginosa*, known in the literature as *pa*-HO or HemO. By observing the signature Soret bands (410 nm for *holo*-PhuS and 406 nm for *holo-pa*-HO), it was shown in the assays prepared that

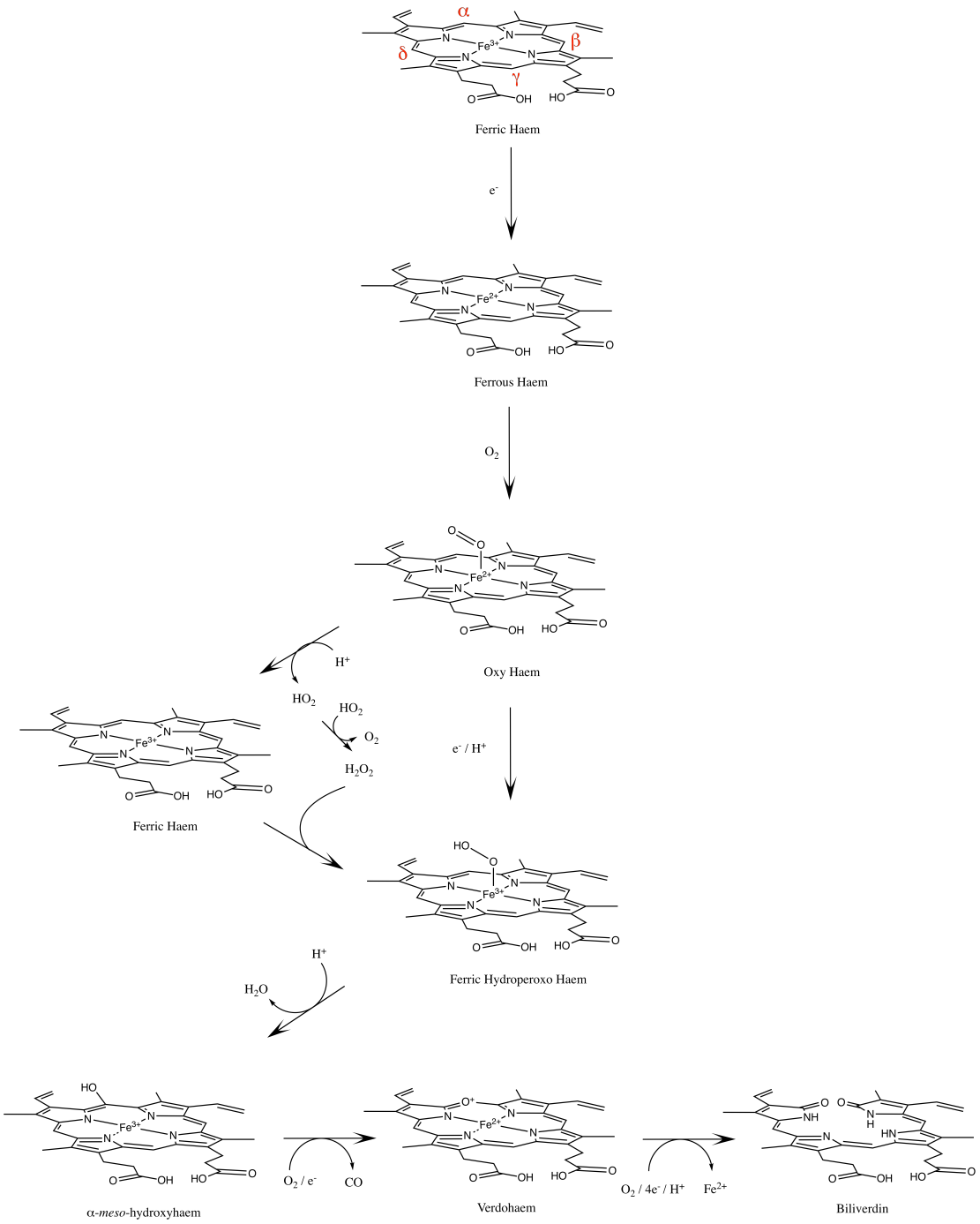


Figure 1.17: Proposed scheme for the coupled oxidation mechanism. Reaction at the α-*meso*-carbon atom is depicted, although the β, γ and δ positions can also be targetted. Oxy haem can be converted to ferric hydroperoxo haem directly. Alternatively, superoxide can be released to regenerate ferric haem; a second superoxide causes dismutation to peroxide, which can attack free ferric haem, thus generating ferric hydroperoxo haem. The hydroxyl can then attack one of the porphyrin *meso*-positions, producing a *meso*-hydroxyhaem. Conversion to verdohaem and biliverdin then proceeds *via* the same mechanism as in the canonical haem oxygenase case.

haem readily transfers from PhuS to *pa*-HO but that the reverse is not true. Surface Plasmon Resonance (SPR) studies revealed a K_D of 64 nM, a value considered consistent with PhuS associating to *pa*-HO in a specific and physiologically relevant manner.[104] It is also thought that a particular arrangement of histidine residues in PhuS could be conducive to the release of haem. Various studies by Wilks *et al.*[109;110] showed that PhuS contains an extra two histidines, H210 and H212, in addition to H209, the residue typically implicated in haem-binding (and the one shown to bind haem in the crystal structure, PDB 4MF9).[111] These three histidines are all in close proximity, and interestingly H210 and H212 do not have any equivalents in the aforementioned homologues of PhuS. Wilks *et al.* therefore suggested that haem, once PhuS was anchored to *pa*-HO, was transferred from His-209 to His-212, a more exposed histidine, to facilitate haem transfer. A H212A mutation confirmed this hypothesis. Though this is evidence from only one system and much more research into this area is needed, it suggests that haem transfer from pathogenic cytosolic haem chaperones requires multiple histidine residues at the haem-binding site in order to ultimately release the haem. Close relatives of PhuS, which lack these extra histidines, have not been implicated in haem transfer to HOs.

Within the last year, the Wilks group have demonstrated that PhuS is also capable of binding DNA.[112] Similar to their work with ShuS, they showed that such binding is mutually exclusive with respect to haem. However, in this case the binding was shown to be sequence-specific. Namely, PhuS was shown to target the *prrF1* promoter, an iron-responsive region directly downstream of the *phu* operon. PrrF bacterial small RNAs (sRNAs) are important during iron starvation as they cause mRNA degradation of nonessential iron-containing proteins,[113–115] thus directing iron to where it is needed most. In addition to this function, a haem-dependent read-through of the *prrF1* terminator yields a longer PrrH transcript, PrrH being a recently discovered sRNA with suspected importance for infection.[112] From these findings, Wilks therefore proposes that PhuS fulfils a dual function regulating haem flux through HemO, and in transcriptional modulation of PrrF/PrrH in conjunction with haem itself.

Research into HemS and its homologues (and their context within their operons) has been pulled in different directions by different groups. This research has led to some fascinating insights, and much debate and confusion surrounding the true roles of these interesting proteins. Table 1.2 and Fig. 1.18 are provided to summarise the data discussed above, highlighting some key points.

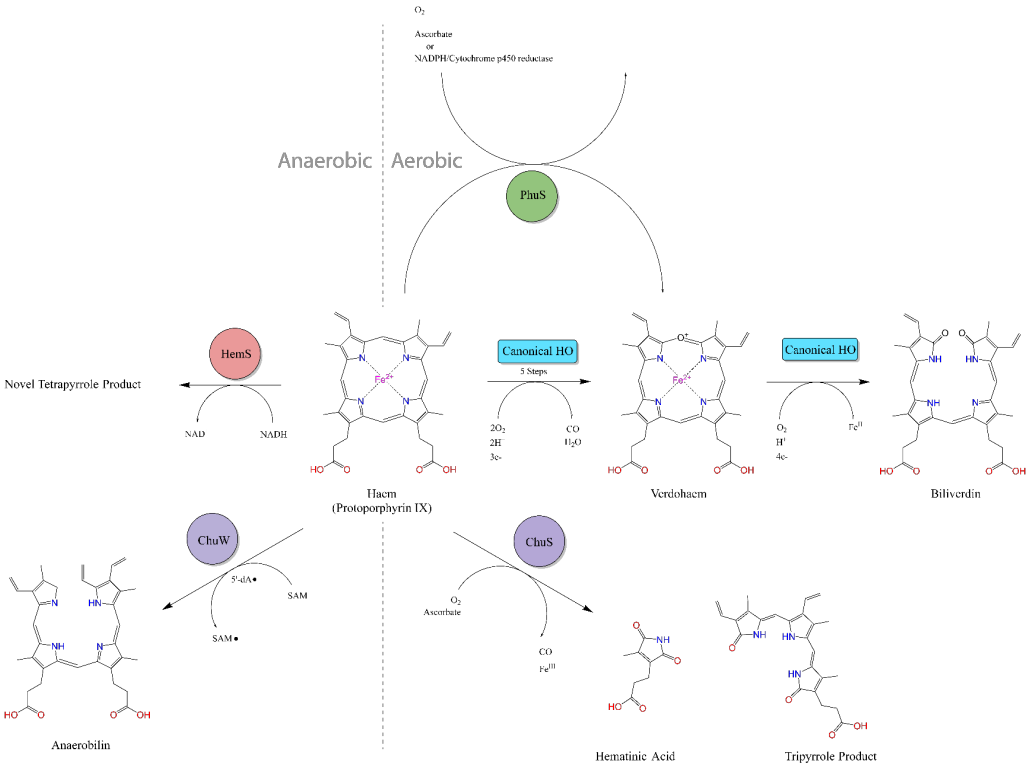


Figure 1.18: Summary of different known haem breakdown pathways. *Canonical HO*: using oxygen and reductants such as NADPH/Cytochrome P450 reductase, Fe^{2+} is extracted from haem, and biliverdin and CO are produced with ferrous verdohaem as the final intermediate. *ChuS*: non-canonical 'HO' reaction with ascorbate in oxygen releases CO, Fe^{3+}, hematinic acid and a tripyrrole product. Ferric verdohaem is the final intermediate. *PhuS*: oxidative haem degradation reaction that can use ascorbate or CPR-NADPH as reductants. Products of this reaction are not yet characterised but verdohaem is one of the intermediates and biliverdin is not produced. *ChuW*: anaerobic haem degradation using radical SAM mechanism to produce anaerobilin. *HemS*: anaerobic reductive haem degradation using NADH and producing a novel uncharacterised tetrapyrrole product, to be discussed below. Figure and caption reproduced from Xie.[81]

Species	Protein	Chaperone	DNA Binding	Haem Breakdown	*holo*/*apo* Crystal Form
Yersinia enterocolitica	HemS	–[a]	–	✓[21;42]	Monomer/Monomer
Yersinia pestis	HmuS	–	–	–	N/A
Escherichia coli O157:H7	ChuS	✓[101b]	–	✓[80;99c]	Dimer/Monomer
Shigella dysenteriae	ShuS	–	✓[102;103]	–[d]	N/A
Pseudomonas aeruginosa	PhuS	✓[104e]	✓[112]	✓[104f]	Dimer/Dimer

Table 1.2: Summary of homologues selected for study/consideration in this thesis. Together, they represent a wide range of homology with HemS. A ✓ indicates that the feature has been demonstrated using the protein in question, whereas a – indicates that it has not (although this could be because no group has tested for the activity under consideration). [a]HemS perhaps acts as a haem chaperone when the haem concentration is high.[21;42] [b]Demonstrated to transport haem to ChuW. [c]Reaction with CPR-NADH system demonstrated. Requires aerobic conditions. [d]Wilks attempted standard NADPH and ascorbate based assays but found no HO activity. [e]Demonstrated to transport haem to HemO. [f]Reaction with CPR-NADH system demonstrated, although it was suggested the observations may be due to a coupled oxidation process instead.

1.5 Novel, Anaerobic Haem Breakdown Discovered in HemS

1.5.1 NADH Structure and Properties

Sawyer tested a variety of small molecules to determine whether any of them disrupted haem-binding in HemS. It was found that glycerol, triethanolamine, maltose, trehalose, imidazole, ATP, galactose, NADH and NADPH could all do so. It is interesting to note that the latter four molecules all have an adenine group, since DNA-binding has been reported in ShuS and PhuS. In each case, these molecules did not appear to induce any other activity, except for NADH and NADPH. For reasons mainly associated with cost (NADPH tends to be ten times more expensive than NADH), NADH became the focus of further research, although where NADPH has been used instead of NADH, the reaction has behaved identically.

NADH is a common biological cofactor and can be used as a reducing agent and hydride source.[116;117] It consists of two ribose groups attached *via* phosphate groups at their 5′-positions. Both of these ribose units have substituents attached to their 1′-positions: an adenine at one end, and a nicotinamide at the other. NADH can exist as two different diastereoisomers, but the β-form is found almost exclusively in a biological context.

As has been demonstrated experimentally[116;117] and computationally,[21] NADH preferentially occupies a folded-up conformation when free in solution, with ribose-ribose stacking. However, this shape prohibits access to many protein pockets and so to fulfil its biological role, NADH is often required to unfold. Many crystal

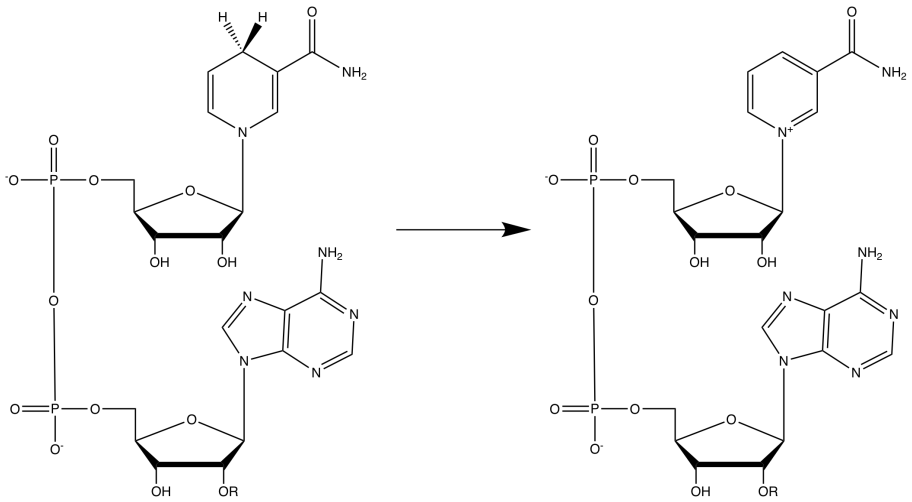


Figure 1.19: NAD(P)H/NAD(P)$^+$ structures. NAD(P)H (left) loses a hydride to produce NAD(P)$^+$ (right). R = H gives NADH/NAD$^+$. R = PO_3 gives NADPH/NADP$^+$.

structures of NADH bound inside proteins show it in a stretched conformation. Certain residues can assist with this unfolding process (particularly by providing hydrogen-bonds to the phosphate backbone), and identifying which ones can do so has become a recent area of interest. This process will be discussed with respect to HemS and its homologues more fully in the Results section.

Once inside the protein pocket, NADH functions by transferring its nicotinamide hydride, H^-, to its target, leaving NAD$^+$. The standard electrode potential of this NAD$^+$/NADH redox pair is $-0.32\,V$, making NADH a strong reducing agent.[118]

1.5.2 Biophysical Research into Haem Breakdown in HemS

Incubation of NADH with HemS and haem resulted in dramatic changes to the UV-Vis spectrum. Most noticeable is the large increase in absorbance at 591 nm, which is accompanied by the solution turning from light yellow to purple. Variations in the wavelength at ∼800 nm were noted but not studied – this effect shall be expanded on in the Results section. Unfortunately, NADH, which has a high extinction coefficient with a peak at 340 nm, was used in such high concentrations that the spectra of the HemS complex was effectively masked from 250-400 nm. Therefore, it was difficult to track the Soret band, and whether it was decreasing as the peak at 591 nm increased. A series of absolute spectra, showing the formation of this purple compound, is given in Fig. 1.20.

A variety of tests were undertaken to determine whether this peak could be an experimental artefact. It was shown that the *holo*-HemS complex was stable with no haem breakdown for a period of over two months; nor did any reaction occur

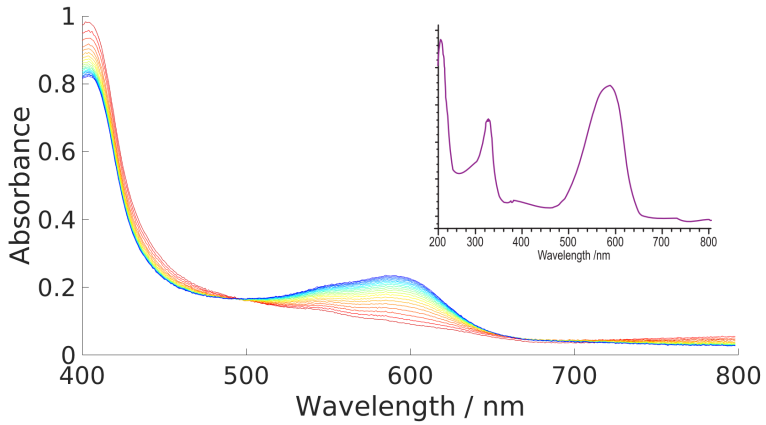


Figure 1.20: Representative UV-Vis absolute spectra showing haem breakdown by NADH in HemS. Evolution is charted in 1 minute intervals, from 1 min (red) to 20 min (blue), and reveals the formation of a purple haem breakdown product (591 nm). Reactants were mixed in the ratio 5 µM HemS : 20 µM Haem : 2000 µM NADH. The inset is from Sawyer,[42] and shows the absolute spectrum of the haem breakdown product following HPLC.

when free haem was incubated with NADH. An assay with cytochrome b_{562} showed that NADH could reduce haem but no enzymatic activity was exhibited. Tests were also conducted, by Gregory,[119] to determine whether the iron ion was an essential component of the reaction, or whether it could proceed with other transition metals, or even non-complexed PPIX. He showed that Co(III)-PPIX, Zn(II)-PPIX and PPIX were all incapable of reacting, suggesting that iron was essential.

Increasing the concentration of the protein or the NADH was demonstrated to increase the rate of formation of this 591 nm species. However, the concentration dependence on haem proved to be more complicated, as shown in Fig. 1.21. This figure shows that the initial rate of reaction increases with haem concentration when the haem concentration is low, but that the rate then decreases again when the haem concentration becomes too high. This behaviour would suggest therefore that at high concentrations, haem is inhibiting its own breakdown. Whether this inhibition is due to the haem : protein ratio being too high, or whether it is due to the absolute concentration of haem being too high, is unclear. Regardless, this behaviour suggested that HemS has a dual role: a haem breakdown enzyme at low haem concentrations to release iron; and a haem-storage protein at higher haem concentrations, protecting against toxicity and retaining the excess haem for future

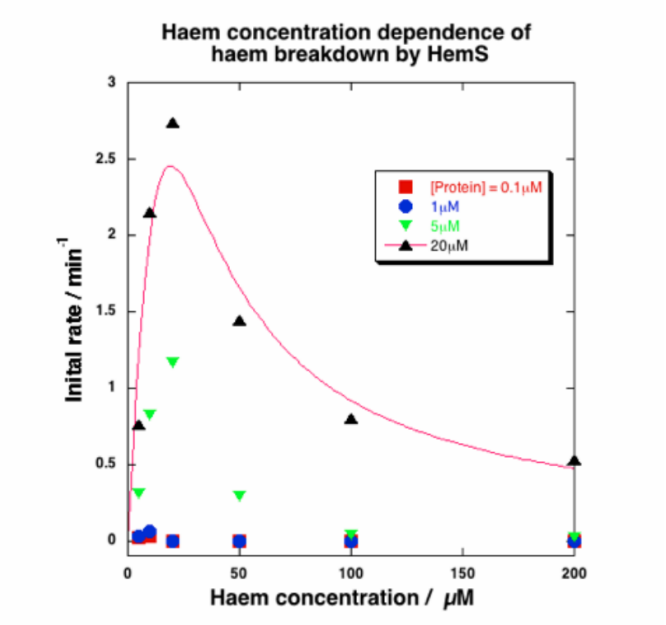


Figure 1.21: Plot of the initial rates of the NADH-dependent *holo*-HemS reaction *vs* haem concentration. This experiment was conducted at four different protein concentrations (red: 0.1 µM; blue: 1 µM; green: 5 µM; black: 20 µM). At low haem concentrations, the initial rate of reaction increases. A maximum is reached, however, after which the rate decreases with increasing haem concentration. Figure reproduced from Sawyer[42] and Choy.[21]

use.

As the properties of this 591 nm species do not match any known porphyrin structure in the literature, Sawyer attempted to isolate and characterise it. This investigation has proved to be a difficult process. It was suspected that the butanone extraction method used to separate this breakdown product from HemS was leading to its further breakdown, reflected by a further colour change from purple to yellow.[42] The extraction appeared to have removed iron completely, giving sharp resonances in a ^{1}H spectrum. Nevertheless, the resulting NMR analyses proved too difficult to interpret. Paradoxically, butanone extractions carried out by this author and George Biggs appear to have been more successful in retaining the integrity of the haem breakdown product (i.e. the purple colour remained) but that the iron still present is paramagnetic and so causes peak-broadening, making the NMR data uninterpretable. This problem shall be discussed further in Chapter 5.

Mass spectrometry has proved more fruitful. High performance liquid chromatography (HPLC) was used to separate the product from the rest of the HemS reaction mixture, so that it could then be characterised by liquid chromatography mass spectrometry (LCMS). These results are shown in Fig. 1.22A, taken from Sawyer.[42]

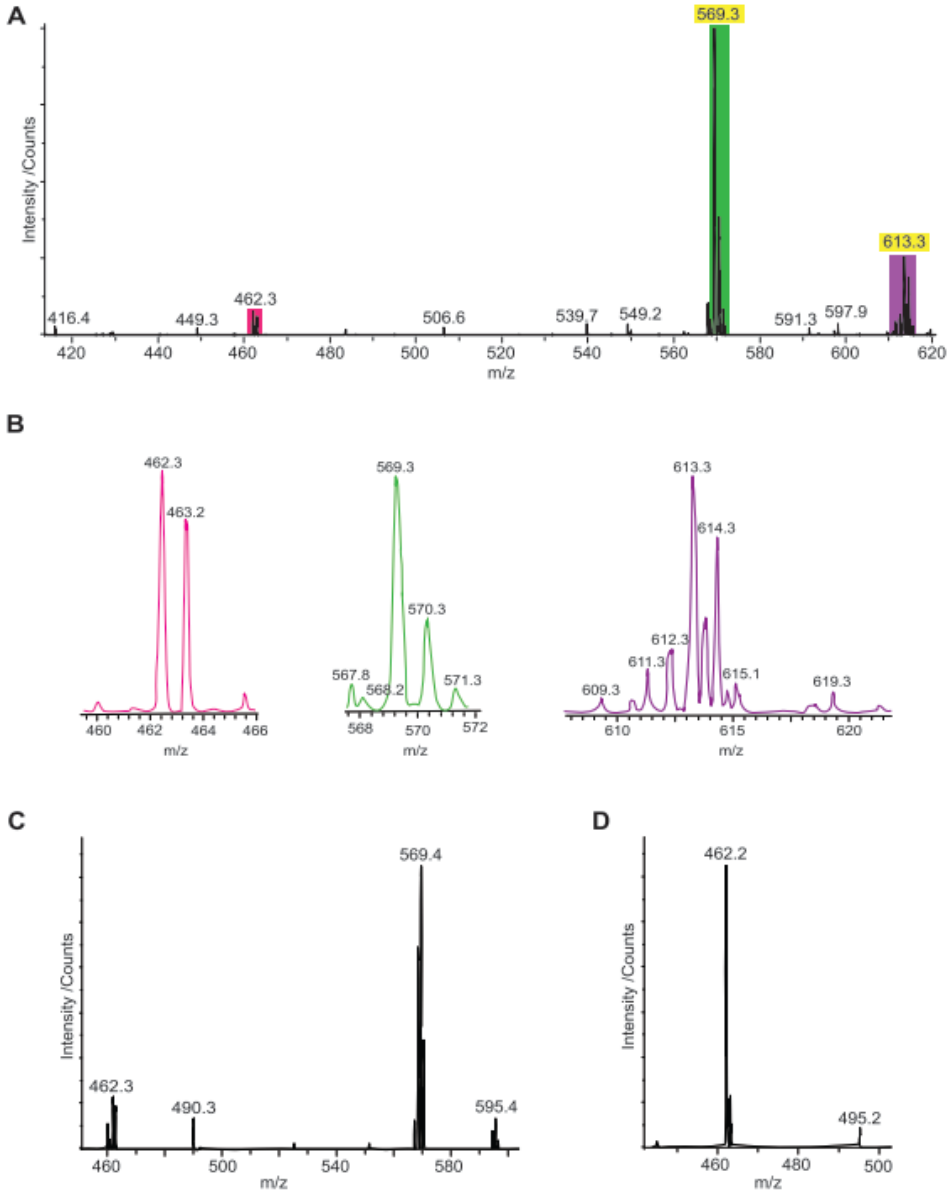


Figure 1.22: LCMS analysis of haem breakdown product. (**A**) Complete mass spectrum. (**B**) Magnified spectra of highlighted major components from (**A**), giving a clear picture of the isotope patterns for each. (**C**) MS^2 spectrum of the parent ion, m/z 613.3. (**D**) MS^3 spectrum of the base peak, m/z 569.3. Reproduced from Sawyer.[42]

This complete spectrum indicates that the parent ion (and so presumably the haem breakdown product) has a *m/z* ratio of 613.3. The base peak, meanwhile, has a *m/z* ratio of 569.3. This value corresponds to the loss of a carboxylate group, which was later confirmed by an accurate mass analysis. Another notable peak in the spectrum is at *m/z* 462.2.[b] MS^n analyses of *m/z* 613.3 and 569.3 showed that these ions can both break down to form this *m/z* 462.2 species, as shown in Fig. 1.22C+D. Accurate mass analysis shows clearly that, to give the *m/z* 569.3 or 462.2 species, iron is not lost in either case. The following scheme shows which fragments in principle may be lost to produce the *m/z* 462.2 species, and proposes what they may be:

$613.2643 - 462.2010 = 151.0633 = C_8H_9NO_2$, propionate-containing pyrrole
$569.2739 - 462.2010 = 107.0729 = C_7H_9N$, vinyl-containing pyrrole

This result is problematic because haem contains both of these potential fragments: it has two propionate-bearing pyrroles and two vinyl-bearing pyrroles. The question was therefore whether there was any way of determining whether it is one or the other, or a mixture of the two, which is lost during the fragmentation process.

Sawyer determined conclusively that it is a vinyl-containing pyrrole that is always lost,[42] implying that the *m/z* 462.2 ion is a tripyrrole with the haem molecule's two original propionate-bearing pyrroles and one vinyl-bearing pyrrole still present. This result suggests that either the β-*meso*-carbon or the δ-*meso*-carbon must be broken along with the α-*meso*-carbon to form this 462.2 ion. It is now known that it is the β-*meso*-carbon that is broken, as shall be described later in this Introduction. Returning to Sawyer's original experiment, the discovery that it is a vinyl-containing pyrrole that is always lost was made through comparative studies with deuterohaem and mesohaem. These molecules, chemically very similar to haem, can react in the same way with NADH in HemS. It was shown that the difference in masses that deuterohaem and mesohaem have with haem were replicated with those for the higher order products (i.e. those with *m/z* 613.3 and 569.3), but that this difference was halved when it came to *m/z* 462.2. This result is shown in Fig. 1.23 and Table 1.3.

The fact that there are no other ions of significant abundance would suggest that the *m/z* 613.3 and 569.3 ions are breaking down consistently. This situation would perhaps suggest that the haem breakdown product already has its ring cleaved at one of its *meso*-positions before ionisation in the spectrometer. Were this not the

[b] The *m/z* value of 462.3 quoted in Fig. 1.22 is a typo.

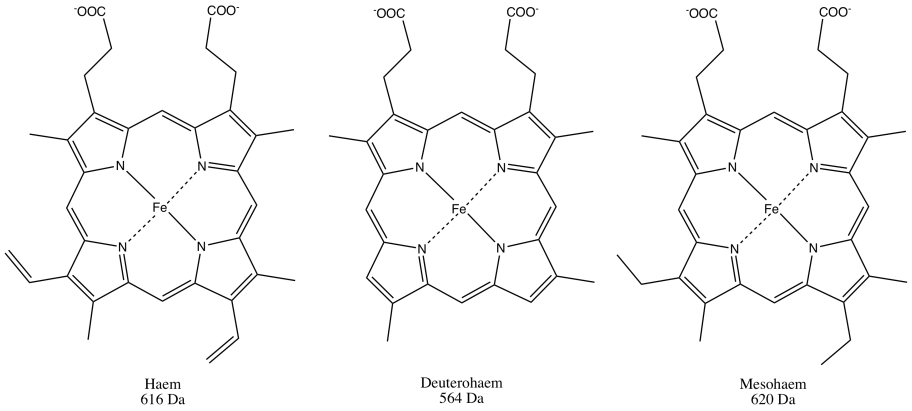


Figure 1.23: Haem with its deutero- and meso- alternatives.

case, and the cyclic tetrapyrrole was still intact, it would seem likely that there would not be a significant preference for which *meso*-carbon had one of its bonds broken first, resulting in a greater variety of fragments being lost. Such an inference is speculative, but it would seem sensible since haem breakdown proteins do typically cleave the PPIX scaffold of haem in order to reduce the effectiveness of iron chelation, and therefore to extract the iron.

LCMS Masses					
	Haem	**Deuterohaem**		**Mesohaem**	
	Mass	**Mass**	**Δ from Haem**	**Mass**	**Δ from Haem**
Before Reaction	616.2	564.2	−52	620.2	+4
LCMS Fragments	613.3	561.6	−52	617.2	+4
	569.3	517.2	−52	573.3	+4
	462.2	436.2	−26	464.2	+2

Table 1.3: Comparison between the masses, in Da, produced by haem from the NADH-induced reaction and some of its derivatives. The halved mass differences on the bottom line indicate that one of the pyrroles with a different substituent between the haem versions has been lost. Data from Sawyer.[42]

1.5.3 Limitations of the Biophysical Approach

Sawyer estimated that NADH-binding in HemS has a K_M of 1823 µM at pH 7.0, which was taken to be very weak.[42] The difficulties associated with crystallisation of proteins plus the suspected transitory nature of NADH-binding therefore discouraged an attempt at X-ray crystallography. Furthermore, it was realised that unless the reaction could be quickly arrested, it would be difficult to crystallise HemS with NADH and haem together in the pocket.

In order to probe this possible NADH-binding site and how NADH could interact with haem, bioinformatics and computational modelling were turned to. Sawyer, through informed choices of mutations to the two residues which coordinate with the iron of haem (e.g. H196A, R102W and H196W) had shown that the reaction with NADH could be severely limited.[42] These mutations were all with a view to disrupting haem-binding. It was wondered whether any residues could be identified which could disrupt NADH-binding, or the reaction between NADH and haem, and a combination of bioinformatics and modelling again seemed like a promising avenue of research to determine this.

1.6 Bioinformatics and Computational Biochemistry

1.6.1 Principles of Bioinformatics

Data science is the extraction of knowledge from data using a variety of scientific methods, algorithms and processes. This is to aid comprehension of data and identify trends. Within a generation, it has grown to become one of the major tools underpinning modern society, from recruitment[120] and manufacturing[121] to healthcare[122] and social networking.[123]

Bioinformatics is a subset of data science which focusses on developing software to better understand biological data. This field arose in order to handle the large quantities of data being generated when protein sequences first began to be determined in the 1950s. This took on further importance in the 1970s when first-generation sequencing technologies for DNA started to emerge. Today, bioinformatics plays a key role in biological research, having driven, for example, the discovery of several targetted anti-cancer drugs[124] and the completion of the human genome project.[125]

Large quantities of genomic (DNA-based), proteomic (protein-based) and structural sets of data are stored online in various databases. This allows for comparison between between sequences and structures. Therefore, newly-discovered molecules can be sequenced and/or their structures determined, and then compared against other molecules already in these databases. A range of algorithms have been developed to expedite these comparisons, allowing for tens of thousands of scans to be made in a few seconds.

One area of bioinformatics that is receiving increasing attention is protein-ligand/protein-cofactor docking. This approach is particularly useful if conventional experiments show that the ligand binds too transiently for full characterisation (as in

the case of NADH in HemS), or if a pocket appears capable of hosting a generic ligand whose properties are otherwise unknown. Using the databases available, models can be developed that 'score' existing, known protein-ligand interactions based on residue-ligand compatibilities and certain structural motifs. A comparison can then be made with the new protein-ligand combination(s). 'Hits' then catalogue areas of potential similarity between this new combination and those in the database, with higher scores typically indicating better alignment. Sophisticated methods to determine the relevant criteria for these hits and scoring methods have been developed, which are discussed in more detail later in this Introduction, and in the Methods section.

1.6.2 Principles of Computational Biochemistry

Computational biochemistry is often considered as synonymous to, or a subset of, bioinformatics. The terms are nebulous, but in this thesis the two fields shall be considered distinct. Whereas bioinformatics handles, processes, and then makes predictions based on deposited data (i.e. it is a 'data-based' approach), computational biochemistry uses *ab initio* or semi-empirical models to evaluate the behaviour of certain biological processes (i.e. it is a 'calculation-based' approach).

Computational biochemistry has also proven to be a useful new tool in the biochemist's repository. For example, it has been demonstrated to accurately predict the 3D structures of protein loops, which are usually difficult to determine experimentally, due to conformational flexibility.[126] This capability will be important when it comes to discussing the Results, since the published X-ray structures of the homologues under investigation all have at least a part of an important protein loop missing.

Computational biochemistry encompasses a wide range of methods. *Ab initio* techniques are the most accurate, but the computational resources required preclude them from the study of all but the smallest of molecules. Coarse-grained methods provide a way for studying very large molecules or biological systems, but do not provide atomic resolution. For the study of proteins, semi-empirical methods tend to be a sensible compromise, as they sacrifice some accuracy for computational expediency and yet maintain atomic resolution.

Semi-empirical methods consist of a force field[c] and a parameter set. A force field adheres to a carefully constructed functional form designed to accurately model covalent bonds, as well as electrostatic and van der Waals interactions. Though the

[c]Despite this term, a force field typically refers to an equation describing the potential energy of a system, from which forces can be derived.

overall form tends to be similar, details can vary between the functions on offer from different companies and research groups. The functional form for the AMBER family of force fields, used in the calculations throughout this thesis, is as follows:

$$E_{\text{AMBER}} = \sum_{\text{bonds}} K_r(r_b - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon r_{ij}} \right]. \tag{1.2}$$

This equation shows the potential split into components describing covalent bonds, angles and dihedrals, and non-bonded electrostatic and van der Waals interactions.

In such an equation, the variables are the relative atomic positions, whilst the parameters describe given atomic properties, such as force constants and charges. These latter properties can be derived from other quantum mechanical calculations or from experimental data. By comparing bond strengths, angles, torsions, etc. for selected atoms across a range of molecules, a picture of how such atoms behave when bonded to others or in certain environments can be developed. This informs the assignment of values to these parameters, which are then stored as a parameter set. The relevant parameters can then be selected when running a calculation.

Force fields are therefore effective at determining the potential energy of a particular arrangement of atoms, albeit with uncertainties introduced by the empirical approximations. The energy of a single conformation does not provide insight, though, into the molecular pathways of biological processes. A variety of methods can, however, be paired with force fields to propagate the system and therefore describe dynamical processes. The most commonly used is Molecular Dynamics (MD), which employs Newtonian equations of motion to determine trajectories and therefore the dynamic 'evolution' of a system. Such an approach has been used successfully to gain insight into a variety of biological problems, such as protein folding[127] and ligand-binding.[128]

There are drawbacks to the MD approach, however. One is that biological processes that require long timescales are computationally prohibitive. Trajectories are typically calculated using the Verlet integration algorithm.[129] A representation of the corresponding propagator is given below,

$$\mathbf{X}_\alpha(t + \delta t) = 2\mathbf{X}_\alpha(t) - \mathbf{X}_\alpha(t - \delta t) + \frac{F_\alpha(t)}{M_\alpha}\delta t^2. \tag{1.3}$$

Here, short time steps, δt, are required to maintain numerical stability and accuracy.

Using standard methods, this choice equates to $\delta t \sim 1$ fs. Biological processes tend to operate on timescales many orders of magnitude greater than this value: even a relatively quick process of 1 ns would require one million steps. This requirement can become computationally prohibitive, particularly when studying large biomolecules such as proteins. Unless advanced hardware solutions or massively parallel computers are used, the standard MD simulations tend to be limited to a few microseconds.[130] Even rapid protein folding events are commonly understood to occur over timescales from 10 µs to 1 ms,[131;132] and protein-ligand binding events can take even longer, so this restriction presents a major problem for MD. Some methods used to mitigate this problem, such as selecting every n^{th} trajectory, are often associated with a certain loss of accuracy. More sophisticated methods, under the umbrella term rare events sampling, including Transition Path Sampling (TPS) and Forward Flux Sampling (FFS), have been developed to address these issues.

Another disadvantage of standard MD is in the absence of true transition states to connect the frames underpinning the trajectory. Though post-processing or even *ad hoc* methods can be used to determine overall rate constants and kinetic properties, this is a significant omission. Typical protein folding, ligand-binding and enzymatic processes require thousands of conformational changes, each with an associated transition state. Knowledge of these structures and their energies can provide key insight into the finer details of the residue-residue, residue-solvent, and residue-ligand barriers that need to be overcome for a biological process to progress.

1.6.3 Energy Landscape Theory

The Energy Landscape provides an alternative framework for studying biomolecular pathways. This approach addresses some of the drawbacks of standard MD mentioned above, as real local transition states can be used, and post-processing methods provide access to long timescales using a fraction of the computational power typically required in standard MD.[133]

The potential energy surface (PES) is a fundamental entity that describes the energy of a system of atoms with respect to selected parameters (typically relative atomic positions). As such, the PES encodes all the information required to determine the ground state structural, thermodynamic and kinetic properties of any molecule.[133;134]

Accounting for bulk translations and rotations, a given collection of atoms, N, has $6N$–6 degrees of freedom. A PES is therefore a function of $6N$–6 variables, and therefore has high dimensionality for a protein. Fortunately, the PES can be well characterised by its stationary points (SPs). Of these, only minima (where all non-

zero normal mode frequencies are real) and transition states (TSs, where all but one of the non-zero normal mode frequencies are real) tend to be physically relevant. Therefore, a PES can be reduced to a collection of minima and the transition states that connect them. Strategies to find these minima and TSs, and determine how they connect, shall be discussed in the Methods section.

Once this process is complete, the user will have a large database of SPs (perhaps numbering in the tens of thousands) to draw from to determine the thermodynamics and kinetics of the system. Using the Dijkstra algorithm,[135] the shortest pathway can be found connecting two chosen endpoints (usually corresponding to the starting structure and end structure of the chemical process of interest).

Though theoretically informative, such a pathway may not be entirely representative of the overall biological process. To overcome this problem, PESs can be related to free energy surfaces (FESs); of course, at $T = 0\,\mathrm{K}$, the two surfaces coincide. In reality, complex biological processes such as protein folding or ligand-binding can be achieved *via* multiple competing pathways, unlike simpler chemical reactions that have defined series of intermediates and a single rate-determining step.[136] This complication renders the application of Dijkstra's algorithm to a PES inadequate as it considers just one such pathway.

To resolve this issue, the PES can be transformed into a FES and the graph transformation approach[137–139] used to sum over all of the discrete paths to give overall rate constants. The k distinct paths algorithm[140] can even be used to distinguish pathways that have different rate-determining steps.[141]

The PES can be transformed into a FES by averaging over a sensible interpolated range of order parameters, whilst providing scope for the more important ones to remain fixed. This averaging tends to be done on the implicit understanding that some degrees of freedom are more meaningful than others. In the case of a ligand entering a protein pocket, averaging would ideally be adjusted over minor variations in spectator residue conformations (typically separated by low barriers), whilst the long-range protein movements and residue-ligand interactions (relatively constrained processes) remain intact. It is important to note that this averaging process does not compromise any thermodynamic or kinetic analysis. The information is still retained; and indeed this averaging over an ensemble is a more realistic representation of an actual biological process, rather than focussing on a single pathway.

1.6.4 Some Successful Applications of ELT

The Wales group at Cambridge has successfully applied ELT-based methods to a variety of biological problems. The main focus has been on protein folding, a highly

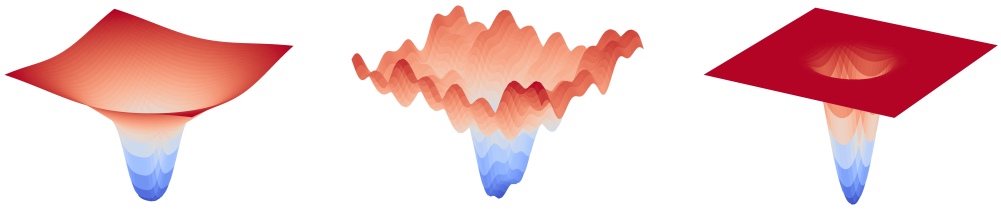


Figure 1.24: Representations of PESs, illustrating the principles of minimal frustration and folding funnels. In each case, high-energy, unfolded conformations would be found at the top of the red rim, and the dark blue region represents the native state at the bottom of the funnel. Left: Extreme case showing little-to-no frustration to protein folding. There are no barriers and multiple pathways are possible for folding. Middle: Moderate frustration. Many pathways for folding are possible but barriers and intermediates must be traversed. Frustration decreases the further down the funnel. This is the most realistic representation of protein folding. Right: Extreme case showing high levels of frustration. Few pathways are available for folding. Though conveying the idea of a folding funnel, these representations only hold when there are two degrees of freedom. To get to higher dimensionality (as required by any real protein) different ways to describe the PES are required. This is discussed in Section 4.9 in the Methods.

cooperative event, which can be difficult to capture experimentally.

Protein folding is the achievement of a native (lowest energy) state from an unfolded state. At first glance, it would appear that such a problem is so complicated, consisting of so many degrees of freedom, that it would not be possible to resolve. Levinthal, in 1969, introduced his famous paradox when he concluded that a random search to find this native state would take an unfeasible amount of time due to the sheer number of possible conformations involved.[142] Anfinsen, meanwhile, had demonstrated that proteins could fold and unfold on experimental time scales.[143]

This discrepancy between theory and experiment can be resolved by considering the energy landscape. This analysis reveals that the native state is not achieved through a random search of all of the possible conformations, but instead there is a bias towards the global minimum. The most popular description of this phenomenon is the 'folding funnel'.[144;145] This concept uses the principle of minimal frustration,[146] which states that 'frustration' is caused by competing low-energy minima separated by high barriers, and that proteins have therefore evolved to reduce this frustration to efficiently locate their native state. The 'folding funnel' extends this view by showing that frustration typically decreases the further along a folding pathway. This situation is illustrated in Fig. 1.24.

Provided effective global optimisation strategies are implemented (to be discussed further in the Methods), the fact that most proteins have a PES shaped like a funnel makes it possible for the native state structure to be identified within a

reasonable timeframe. This capability, in turn, makes it possible to identify folding pathways without the need to sample all frustrated, typically high energy, states. This approach has been used to determine the folding properties of a variety of biomolecules, from short peptides[147] and knotted proteins[148] to fold-switching[149] and designed proteins.[150]

Though the notion of a funnel was first developed to understand protein folding, this concept has a far wider range of application, from DNA base-pairing[151] to viral capsid assembly.[152] Of most importance to this thesis, is the application of the energy landscape approach to protein-ligand interactions. Due to the additional complexities of having multiple molecules within one simulation, this sort of problem is rarely tackled using an ELT-based approach. The principles, however, are largely the same. As mentioned above, protein folding from an evolutionary perspective requires the maximisation of bias towards stable conformational states. Protein-ligand interactions also require biasing, so that the ligands can be aligned appropriately and be primed for reaction. This organisation is usually again the product of evolutionary pressures, resulting in funnel-shaped landscapes.

1.7 Previous Work Using Computational Methods to Investigate HemS

Armed with a bioinformatic strategy and the ELT-based approach to computational modelling, it became possible to investigate the behaviour of HemS, haem and NADH in further detail. This work was begun by Desmond Choy,[21] and his key insights are summarised over the following sections.

1.7.1 Bioinformatics as Applied to the HemS-NADH Binding Problem

It was clear from sequence homology searches that the possible NADH-binding pocket bore little-to-no relation to any reported in the literature. Therefore, a bioinformatics package, Relibase$^+$,[153d] was used to scan the protein cavity and surface for structural comparisons against known NADH-binding pockets.[21] This package was specifically chosen because of its inclusion of CavBase,[154;155] a program explicitly designed to be independent of sequence and fold homology, yet still be able to

[d]This package was retired by the Cambridge Crystallographic Data Centre (CCDC) in 2018 as updated packages were unveiled, with most of its functions being embedded in its CSD Python API.

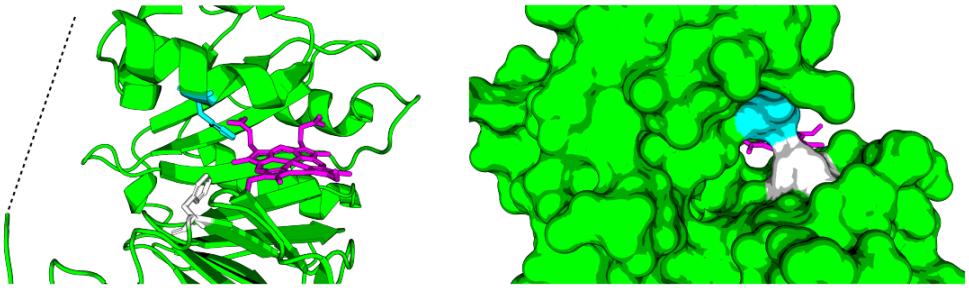


Figure 1.25: Large cavity of HemS. It is split into two pockets by a central pair of phenylalanine residues, F104 (white) and F199 (cyan). In one pocket, haem binds. In the other, Relibase$^+$ identified a possible NADH-binding site. As in previous figures, a dashed line has been included to represent the missing loop region from the crystal structure (2J0P).

identify unexpected similarities amongst protein cavities. CavBase worked first by defining the protein cavity within the protein of interest. This definition is made by placing the overall protein structure within a 3D spatial lattice. Grid points are then split into those representing protein atoms considered solvent inaccessible, and those representing atoms considered accessible, with the latter scored according to their degree of burial.[156] A higher score indicates a higher degree of burial, and if a cluster of grid points all exceed the score then a cavity is diagnosed. The residues flanking this cavity are then noted and generalised into (sometimes overlapping) groups called pseudocenters according to their hydrogen-bonding and/or aromatic properties. Relibase$^+$ then takes this information, picks a cavity of interest, selects all or a subset of its pseudocentres, and compares against the cavities of other proteins. Typically, this comparison would include all, or a subset, of the Protein Data Bank (PDB).[157]

When applied to HemS, Relibase$^+$ identified a portion of the large cavity already housing haem as a potential NADH-binding site. More specifically, this analysis revealed that the large cavity was actually a fusion of two smaller pockets separated by a double phenylalanine gate. This structure is illustrated in Fig. 1.25

One of these sub-pockets consisted of the known haem-binding site, but the other was a large void, which was assumed to be occupied by the loop missing from the crystal structures.[21] It was this void which, when tested against every cavity in the PDB, gave a hit with a known NADH-binding site. The hit was a mutant of the 2-trans enoyl-acyl carrier protein reductase enzyme InhA, from *Mycobacterium tuberculosis* (PDB Code 2AQI). Superimposing the coordinates of NADH from this structure onto *holo*-HemS produced a remarkable overlap. Not only did NADH fit favourably in the HemS pocket, but its conformation was uncommonly stretched

out in such a way that its hydride-bearing nicotinamide group was pointing directly at the β-meso carbon of haem. Curiously, the overlap score of HemS with 2AQI was only 21.1%, which was considered a 'weak hit.' This value was probably caused by a lack of relevant data in the PDB. Factors accounting for this lack of data most likely included the novel protein fold itself;[e] the unprecedented chemistry involved; the fact that the binding site is not a tight one (as established experimentally by Sawyer);[42] and the missing loop, suggesting that the search itself was missing relevant information from the start.

1.7.2 Energy Landscape Theory as Applied to the HemS-NADH Binding Problem

Though a promising confirmation that it can in theory enter the pocket, the superimposition of NADH into the *holo*-HemS pocket does not give the actual binding mode of NADH at an atomic level of detail. Moreover, it provides little information on how NADH enters and moves through the pocket.

Choy was able to address the first of these issues, but the second proved to be beyond the computational power of the day. Using the **GMIN** program,[158] he was able to identify the most important residues involved in the protein-ligand interactions. **GMIN**, which shall be discussed more fully in the Methods, includes basin-hopping (BH) algorithms, which provide an effective stochastic method for finding the global minimum of a system. For a system as complicated as HemS and its two ligands, it would be a tall order to find the precise global minimum. However, using **GMIN** even over a limited number of BH steps allows for local regions to be reasonably well explored. Using the superimposition of NADH from 2AQI into *holo*-HemS as a starting point, it became clear to Choy even after just 100 BH steps that certain residues (Q132, S171, K203, R250 and T312) were important for binding. This observation was based on the system's overall energetic responses to their respective changes in conformation, as catalogued after the BH steps. Fig. 1.26 shows their situations in the pocket. Whilst Q132, S171 and R250 all interact with the adenine base (the latter with the adenosine ribose oxygen and phosphate groups too), K203 was shown to always interact with the amide group of nicotinamide, and T312 with the hydroxyl groups of the adenosine ribose and the phosphate groups. Further BH runs (up to 500 steps) served to confirm this hypothesis, as well as identifying further residues of interest.

Two other residues comprising the aforementioned double phenylalanine (or dou-

[e] At the time of the analysis, ChuS was the only other protein shown to have a similar fold. Now, PhuS has been added to this list

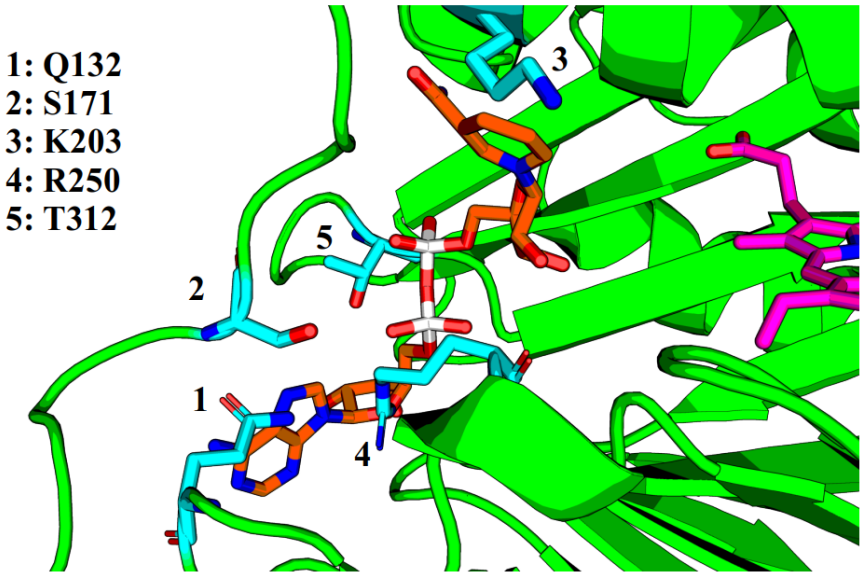


Figure 1.26: Close-up representation of the *holo*-HemS main pocket with NADH (computational reconstruction). NADH is shown in orange. The most important residues implicated in NADH-binding, as identified by Choy and detailed in the text, are shown in cyan.

ble phe-) gate, F104 and F199, are of particular interest. Situated between the known haem and putative NADH-binding pockets, Choy proposed that they could be there to perform some sort of regulatory role.[21] It was clear from space-filling models that NADH would only be able to easily access haem if these residues were in an 'open' conformation, as shown in Fig. 1.27. Choy further noted that this double phe-gate could be considered as closed under a few different scenarios, including 'the classic T-shaped or slipped parallel modes of π-π stacking.'[21]

Following Choy, Cheng Shang pushed this research further.[159] The aim of this work was to determine the mechanism of NADH movement along the pocket, and how this may influence the conformation of the double phe-gate.

To begin, Shang took the superimposed structure of NADH in HemS which had previously been identified with Relibase$^+$. He then conducted his own **GMIN** run, much as Choy had done, only this time with far larger basin-hopping steps. This approach increases the chances of finding unphysical, high energy structures, as well as the likelihood of cis-/trans-isomerism and chirality flipping. Such structures are not accessible in real-life biological processes, and so are rejected by **GMIN**.[158] These large BH steps therefore cause a large proportion of the structures found to be rejected. However, they allow for a greater area of the potential energy surface to be traversed.

This approach therefore allowed for [*holo*-HemS + NADH] structures to be iden-

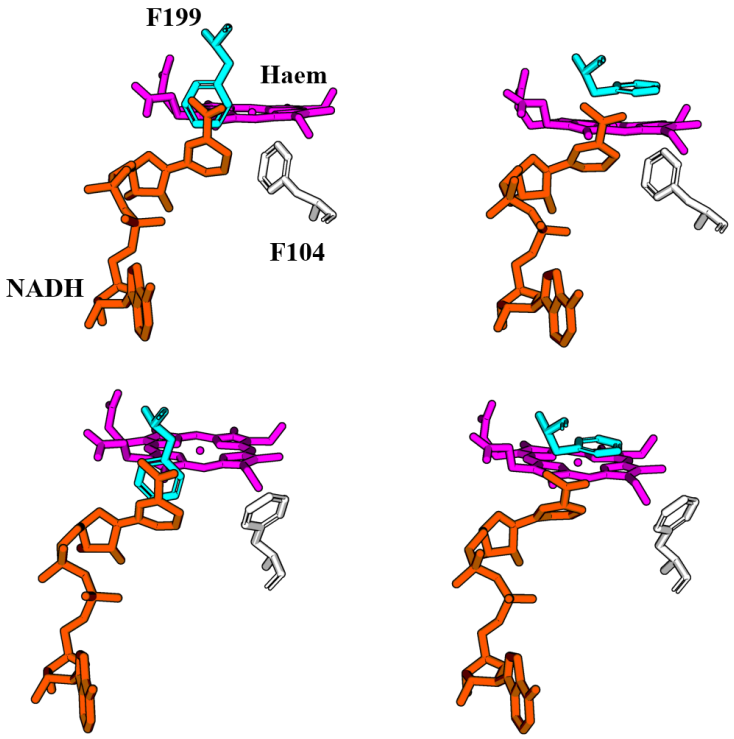


Figure 1.27: Representations of possible phe-gate conformations. All protein atoms other than F104 and F199 have been removed for clarity. Haem has been colour-coded magenta, NADH orange, F104 white and F199 cyan. Top left: both gates are closed, i.e. there is a closed-closed ($C_{F104}C_{F199}$ or CC) conformation. NADH therefore cannot slip past to access haem. Top right: Closed-open (CO) conformation. There is now a narrow window for NADH to access haem. Bottom left: Open-closed (OC) conformation. As with CO, there is a narrow opening for NADH to reach haem. Bottom-right: Open-open (OO) conformation. NADH has a clear route through to haem.

tified where the NADH conformation was dramatically different from the starting point, with the nicotinamide group much further away from haem. Shang selected three of these structures and conducted further **GMIN** runs on each (this time with smaller BH steps) in order to minimise the potential energies.

Interestingly, from these three selected structures, Shang showed that the further removed NADH was from haem, the more folded it became. This result implies that NADH approaches and enters the pocket in a roughly folded up conformation (which is not surprising given NADH in solution preferentially stacks its two ribose groups together) and is then induced by the protein to unfold in order to access haem.

Shang also investigated how each of these local conformations of NADH influenced the double phe-gate. For each one, he considered the possible rotamers of F104 and F199, and reminimised with these different rotamer combinations. This approach gave four structures for each of the local conformations chosen, labelled CC (i.e. $C_{F104}C_{F199}$), CO, OC and OO, respectively. For comparison, this calculation was performed on a representative structure of *holo*-HemS (i.e. without NADH) as well. Whereas for this *sans* NADH structure there was little difference in energies between the rotamer combinations, it was shown for the others, that as NADH approached haem, a biasing for the CO and OO conformations was instigated. This effect is shown in Fig. 1.28. Clearly, there is less steric crowding if the double phe-gate is opened up, thus allowing NADH to pass through more easily, and so it is interesting that there appears to be an energetic bias upon NADH approach to allow this opening to happen.

However, as can be seen from Fig. 1.28, this analysis was very limited, and indeed the results from 'Position 2' do not seem to fit the overall trend. This result was due to the computational limitations of the time, which also precluded any pathways for NADH moving through the pocket from being characterised. Fortunately, due to the speed-up afforded by implementing Wales group methods on GPUs, such an analysis has now become possible.

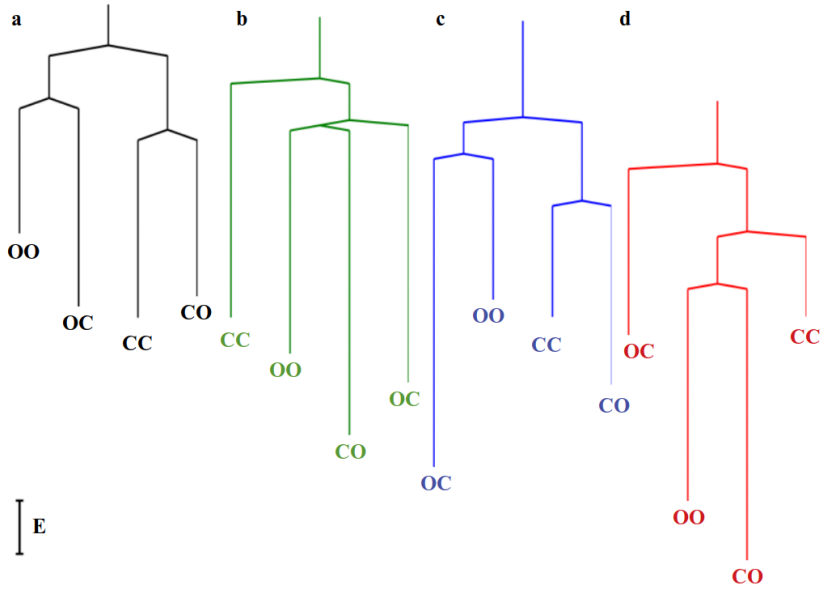


Figure 1.28: Some disconnectivity graphs, charting NADH approach to haem. The x-axis is arbitrary, and the y-axis corresponds to energy, **E**. Nodes correspond to superbasins, and line endpoints to minima. The construction and properties of disconnectivity graphs are explained more fully in Section 4.9 in the Methods. **a** *holo*-HemS, no NADH. **b** NADH in Position 1 (far from haem). **c** NADH in Position 2 (moderate distance from haem). **d** NADH in Position 3 (close to haem). To construct each graph, a representative structure with NADH at its given position (or not present, as in the case of **a**) was taken, and its phe-gates manually adjusted to the CC, CO, OC or OO positions. These structures were optimised to their nearest local minima, and then connected *via* transition states. As NADH is introduced and then moves along the pocket, the overall energy of these minima decrease. Furthermore, a biasing towards the CO and OO conformations is clear. Figure adapted from Shang.[159]

Chapter 2

Project Outline

The project was designed to answer a number of questions regarding the properties of HemS, the reductive haem breakdown process it promotes, and its wider phylogenetic context. Laboratory-based experiments, computational calculations and bioinformatic analyses were all used to help to find these answers. The original questions that were set, as well as those that arose during the course of the project, were as follows.

Does the haem breakdown process involve the direct transfer of a hydride from NADH to haem? This question was important for establishing the precise role of NADH in the reaction mechanism, as it was not known whether it was acting as a hydride transfer agent, or was simply engaged as an electron donor. Answers from an isotopic labelling study are provided in Chapter 5.

How does the reaction respond to various protein : haem ratios under anaerobic conditions? Studies by Sawyer had suggested that the reaction can proceed without oxygen. As it is unusual for haem breakdown to proceed anaerobically, this question was important for further understanding this novel process. Answers are provided in Chapter 5.

What is the structure of the haem breakdown product? Answering this question would provide clues regarding the details of the reaction mechanism and the overall purpose of the process. Partial answers, which appear to confirm that the porphyrin is cleaved, are provided in Chapter 5.

In addition to GPU acceleration, what strategies can be used to expedite the creation and expansion of stationary point databases describing

complicated systems? HemS is a large protein to treat atomistically, particularly when its interactions with two separate ligands are also being considered. To ensure that the creation and expansion of databases within a reasonable timeframe could be achieved, new subroutines were written, tested and applied. Their development is discussed in Chapter 6.

Can fully connected pathways showing the unfolding and approach of NADH towards haem be identified? These pathways are of interest because they would indicate which residues in the protein cavity facilitate NADH-binding and induce it to unfold. Pathways were indeed identified, and are discussed in Chapter 7.

Which residues should form the basis of further mutagenesis studies? Certain residues, based on their purported abilities to bind NADH or regulate its access to haem, were selected for further computational and experimental studies. Reasons for these selections, based on an analysis of the computationally-derived pathways, are given in Chapter 7.

How do potential energy landscapes representing mutated versions of HemS compare against the wild type (WT)? Using the WT database as a 'template', new databases incorporating various selected mutations were seeded, grown and refined. Selected features of the new energy landscapes arising from these databases were then compared and contrasted, to provide more information regarding the roles of these mutated residues. These landscapes are represented and discussed in Chapter 7.

How do potential energy landscapes representing selected homologues compare against HemS? The WT HemS database was also used as a 'template' to seed databases where HemS was replaced by HmuS, ChuS or ShuS. The intention was to compare and contrast the energy landscapes arising from these new databases with that for HemS. Although none of the homologue databases were fully refined, some conclusions could be drawn from the features of the new landscapes they described. These landscapes are represented and discussed in Chapter 7.

What is the phylogenetic context of HemS? How well are its residues conserved? HmuS, ChuS and ShuS are only a small snapshot of the total number of haemoproteins which are homologous to HemS. It is of interest to determine what

features, if any, this family of proteins share. Alongside an analysis into sequence conservation, it was anticipated that such information would provide clues as to the true function(s) of these proteins. A full discussion is provided in Chapter 8.

The residues identified by the computational mutagenesis study were also selected for a laboratory-based mutagenesis study. What effect do these mutations have on the ability of HemS to bind haem, and on its capacity to catalyse the NADH-dependent haem breakdown process? As noted above, these residues were selected due to their purported roles in binding NADH or regulating the access of NADH to haem. Results are provided in Chapter 9.

Can the selected homologues, HmuS, ChuS and ShuS, also catalyse the NADH-dependent reductive breakdown of haem? This question was crucial for determining whether this novel reaction is merely a unique (and perhaps even unintentional) artefact of the HemS protein, or if it is a genuine evolved feature of this family of proteins. Results are provided in Chapter 9.

What are the kinetic features of the enzymatic reaction, and how do they compare between WT HemS, its mutants and its homologues? Stopped-flow spectroscopy was used to track selected species during these reactions on short timescales. Results are discussed in Chapter 9.

Is Energy Landscape Theory a useful tool for predicting protein-ligand interactions? Computational methods had been used to reproduce the behaviour of WT HemS with haem and NADH, and then make predictions regarding certain single-point mutations, or by replacing HemS with a close homologue. Were these predictions reflected by subsequent experimentation? These questions are discussed in the Conclusions.

Why is there a family of haemoproteins that can break down haem anaerobically? Drawing upon evidence from experimentation, Energy Landscape Theory and bioinformatics, reasons as to why certain haemoproteins have this ability to break down haem under oxygen-limiting conditions are speculated upon in the Conclusions.

Chapter 3

Experimental Methods

3.1 General Conditions

Unless otherwise stated, all chemicals were purchased from Merck UK. Bovine haemin stocks were from Fluka Biochemika. All solutions were prepared with deionised water purified to a resistance of 18.2 MΩ cm^{-1} (PURELAB Chorus, Veolia Water Technologies). If required, the pH was adjusted using 5 M NaOH and 5 M HCl. Solution pHs were measured by an InLab Micro pH probe (Mettler Toledo) connected to a PHM240 digital pH meter (Radiometer Analytics), following three-point calibration with standard IUPAC buffers (pHs 4.005, 7.000 and 10.012). pH-adjusted solutions were then filtered through a 0.2 µm membrane (Sartorius Stedim Biotech).

Protein concentrations were determined by UV-Vis Spectroscopy using a Cary 60 spectrophotometer (Agilent Technologies) with the experimental Beer-Lambert Law:

$$A_\lambda = \epsilon_\lambda cl. \tag{3.1}$$

Here, A_λ and ϵ_λ are the absorbance and extinction coefficient at a given wavelength, λ, c is the concentration and l is the optical path length. ExPASY ProtParam[160] was used to theoretically predict ϵ_{280} values and formula masses for the *apo*-proteins being studied. The value for *apo*-HemS corresponded to that used in previous studies.[21;42;81;119;161] Mutants of HemS were assigned the same ϵ_{280} value as the wild type (WT). All relevant ϵ_{280} values and formula masses are given in Table 3.1.

Protein	ϵ_{280} / $M^{-1}cm^{-1}$	Formula Mass / Da
Wild Type HemS	44,070	39,360
Wild Type HmuS	44,920	39,104
Wild Type ChuS	52,940	38,845
Wild Type ShuS	52,940	38,831
F104A HemS	44,070	39,284
F104AF199A HemS	44,070	39,208
F104I HemS	44,070	39,326
F199A HemS	44,070	39,284
R209A HemS	44,070	39,275
R209K HemS	44,070	39,332
Q210A HemS	44,070	39,303

Table 3.1: Selected biophysical characteristics of the proteins studied in this work.

3.2 Buffers

Buffers used in this thesis were as follows:

Affinity Chromatography Buffer A (Aff A)
20 mM bis-tris propane (BTP), 10 mM imidazole (Acros Organics), 300 mM KCl (Fisher). Set to pH 6.5 using ∼7.8 mL 5 M HCl per 1 L solution.
Affinity Chromatography Buffer B (Aff B)
20 mM BTP, 500 mM imidazole, 300 mM KCl. Set to pH 6.5 using ∼72.5 mL 5 M HCl per 1 L solution.
Size Exclusion Chromatography Buffer (SEC)
20 mM BTP. Set to pH 6.5 using ∼5.5 mL 5 M HCl per 1 L solution.
Anion Exchange Chromatography Buffer A (AEC A)
Same as SEC.
Anion Exchange Chromatography Buffer B (AEC B)
20 mM BTP, 500 mM KCl. Set to pH 6.5 using ∼6.0 mL 5 M HCl per 1 L solution.
Thrombin His-Tag Cleavage Buffer (THTC)
20 mM BTP, 150 mM NaCl (Fisher), 1.5 mM $CaCl_2$ (VWR International). Set to pH 8.4 using 5 M NaOH.
ShuS Lysing Buffer
50 mM tris-HCl (USB Corporation), 1 mM EDTA, 1 mM $MgCl_2$. Set to pH 8.0 using 5 M NaOH.

Crystallisation Buffer A
50 mM HEPES (Sigma Aldrich), 150 mM NaCl. Set to pH 8.0 using 5 M NaOH.
Crystallisation Buffer B
100 mM tris-HCl (USB Corporation), 1.8 M $(NH_4)_2SO_4$ (Breckland Scientific Supplies), 2% (w/v) PEG 400. Set to pH 8.5 with 5 M NaOH.
NADD Buffer A
10 mM NH_4HCO_3 (BDH Laboratory Supplies). Set to pH 8.5 with 5 M NaOH.
NADD Buffer B
500 mM NH_4HCO_3. Set to pH 8.5 with 5 M NaOH.

3.3 Plasmid Preparation & Protein Expression / Purification

3.3.1 HemS

The *hemS* gene in a pGAT2 expression vector[162] was provided by Sabine Schneider and Max Paoli (University of Nottingham, UK). The gene sequence in this plasmid disagreed from the crystal structure sequence (2JOP in the Protein Data Bank)[96] at residue 333; replacing an aspartic acid with a glutamic acid. The full sequence plus a discussion of its relationship to genes quoted in the literature is provided in Appendix B. HemS was expressed with a His_6-tag fused at its N-terminal, connected *via* a thrombin cleavage site. The pGAT2 vector confers ampicillin resistance.

DNA manipulation was carried out using competent *E. coli* XL1-Blue cells (Agilent Technologies), and expression performed using electrocompetent *E. coli* BL21-Gold (DE3) cells (Agilent Technologies). Transformation was either by electroporation or heat shock. If electroporation was used, cells were incubated at 37 °C for 20 min in 500 µL lysogeny broth (LB) (Molecular Production and Characterisation Centre (MPACC) Service, Department of Chemistry, University of Cambridge), before being spread onto LB agar ampicillin plates (MPACC Service). If heat shock was used, cells were incubated at 37 °C with 200 rpm shaking for 1 h in 950 µL SOC medium (Thermo Fisher Scientific), before being spread onto LB agar ampicillin plates.

Plates were incubated at 37 °C overnight. From each plate, a colony was selected and added to a 50 mL starter culture of LB medium supplemented with 100 µg mL^{-1} ampicillin, which was then incubated overnight in a thermoshaker (200 rpm, 37 °C). 20 mL of overnight solution was then used to inoculate a 1 L flask of LB medium

supplemented with 100 µg mL^{-1} ampicillin. Cells were grown in a thermoshaker (200 rpm, 37 °C), until an optical density at 600 nm (OD_{600}) between 0.6-1.0 was reached. Protein expression was then induced by adding 0.4 mM isopropyl-β-D-thiogalactopyranoside (IPTG) (Melford Laboratories). Reducing the temperature to 25 °C at this point can give a five-fold increase in yield. The cells were incubated further (200 rpm, 25 °C) until the OD_{600} reached 2.25-2.50. These cells were then harvested by centrifugation (5000 rpm, 20 min, 4 °C) using a Sorvall SLC-4000 rotor (Thermo Scientific), and the pellets then resuspended in 20 mL Affinity Buffer A per 1 L LB medium used.

The cells were lysed by a high pressure homogeniser (EmulsiFlex-C5 – Biopharma Process Systems). Centrifugation (12,000 rpm, 50 min, 4 °C) using an F21S rotor (Thermo Scientific) was carried out, and the insoluble components removed.

The purification protocol used was based on that outlined by Schneider and Paoli[163] but had various modifications. The protein was loaded onto high density Ni resin (ABT – Agarose Bead Technologies) and eluted with Affinity Buffer B. HemS typically eluted at ∼250 mM imidazole. Using 10 kDa Amicon Ultra centrifugal filters (Merck Millipore), the eluate was concentrated and exchanged into ∼1.5 mL THTC Buffer. Incubation with 250 U bovine thrombin (MP Biomedicals) was then carried out for at least 36 hours at 4 °C to remove the His_6-tag. The protein was again loaded onto high density Ni resin which had been washed with Affinity Buffer A. The eluate which passed straight through the column was collected, separating cut from uncut protein. The cut protein was concentrated and exchanged into ∼2 mL SEC Buffer using Amicon Ultra centrifugal filters. The protein was then purified by size exclusion chromatography (SEC) using a Superdex 75 column (GE Healthcare).

Typically, this procedure yielded a highly purified protein sample. However, if there were still impurities following this size exclusion step, an anion exchange step was introduced. For this step, the HemS-containing fractions were pooled and concentrated down to ∼2 mL before being loaded onto a Mono-Q column (Pharmacia Biotech, now GE Healthcare). A linear gradient (0–500 mM KCl) was applied to elute HemS from the column.

Yield of protein was typically 20 mg per 1 L *E. coli* cells harvested.

3.3.2 HemS Mutants

Mutants were created from site-directed mutagenesis on the WT gene. Seven mutants were made: F104A, F104AF199A, F104I, F199A, R209A, R209K and Q210A. Primers (shown in Appendix B) were designed using SnapGene[164] and ordered from Merck UK Ltd. Mutagenesis was carried out using a QuikChange II XL Kit (Strata-

gene, Agilent Technologies). Transformation was also carried out as suggested in the kit (heat shock of ultracompetent *E. coli* XL10-Gold cells), culminating in the transformed cells being incubated overnight on LB agar ampicillin plates at 37 °C. Successfully grown cells were selected, and mutant plasmid extracted using a Miniprep Kit (Qiagen). Purity was determined using a Nanodrop 2000C Spectrophotometer (Thermo Scientific), and the sequences confirmed by standard Sanger Sequencing (DNA Sequencing Facility, Department of Biochemistry, University of Cambridge) of the plasmids using standard T7 primers. Mutagenesis for the F199A mutant was unsuccessful using the QuikChange II XL Kit, and so the QuikChange Lightning Kit (Stratagene, Agilent Technologies) was used instead. To make the double mutant, F104AF199A, point-mutations were done successively. First, the F104A mutation was introduced to the WT. The F199A mutation was then introduced to the purified F104A plasmid to create the double mutant.

Protein expression and purification for the mutants was the same as for the WT. Yields proved to be very similar.

3.3.3 HmuS, ChuS and ShuS

Synthetic genes for HmuS, ChuS and ShuS were designed and synthesised in commercial vectors by Thermo Fisher Invitrogen GeneArt. These inserts each included a region encoding a His_6-tag and thrombin-cleavage site, for expression at the N-terminal end of the protein. The plasmids were replicated in *E. coli* XL1-Blue cells. Restriction digests (RD) using NcoI (New England Biolabs, NEB) and BamHI (NEB) were then carried out on the plasmids and a further pET11d-containing plasmid (Barker Group stocks). Standard RD protocols (NEB) were followed.[165] Quick CIP was not added as it was found to limit ligation downstream. Fragments were separated on a 0.8% agarose gel, and the HmuS/ChuS/ShuS insert and pET11d backbone excised. These were then extracted using a QiaEXII Gel Extraction Kit (Qiagen). The fragments were purified using a Monarch PCR & DNA Cleanup Kit (NEB). Purified fragments were ligated using a Quick Ligation Kit (NEB).

The new plasmids with pET11d expression vector were transformed into competent *E. coli* XL1-Blue cells by heat shock, before being incubated at 37 °C with 200 rpm shaking for 1 h in 950 µL SOC medium (Thermo Fisher Scientific). These were then spread onto LB agar ampicillin plates, and incubated at 37 °C overnight. Plasmids were extracted and purified from selected colonies using a Miniprep Kit (Qiagen). Purity was determined using a Nanodrop, and sequences checked using the DNA Sequencing Facility, Department of Biochemistry, University of Cambridge.

Protein expression and purification for HmuS and ChuS was the same as for

HemS. Yields proved to be very similar.

Following protein expression and harvesting by centrifugation, ShuS was resuspended in ShuS Lysing Buffer rather than Affinity Chromatography Buffer A. 1 tablet of Roche Protease Inhibitor cocktail (Roche Diagnostics GmbH) was added per 40 mL buffer, in addition to 50 µL 250 U μL^{-1} benzonase nuclease (Sigma Aldrich). These were added to prevent DNA from binding to ShuS, causing it to aggregate. The cells were then lysed by a high pressure homogeniser (EmulsiFlex-C5 – Biopharma Process Systems) and the protein purification process continued as it is done for HemS. However, one further point of differentiation was that buffer pH levels were always maintained at 8.0 rather than 6.5, because ShuS has a tendency to precipitate at lower pHs. Yields proved to be very similar to those for the other homologues.

3.4 SDS-Polyacrylamide Gel Electrophoresis

SDS-PAGE analyses were carried out using NuPAGE 4–12% Bis-Tris gels (Thermo Fisher Invitrogen). 13 µL samples were pre-treated with 5 µL 4x NuPAGE LDS Sample Buffer (Thermo Fisher Invitrogen) and 2 µL 100 mM dithiothreitol (DTT) (Melford Laboratories). Gels were typically run for 40 minutes at 180 V in NuPAGE MES SDS Running Buffer (Thermo Fisher Invitrogen), before staining with 2.5% (w/v) coomassie blue in a methanol/acetic acid mixture.

3.5 Ultraviolet-Visible Spectroscopy

All UV-Vis spectra were collected on a Cary 60 (Agilent Technologies), a Cary 100 (Varian Ltd.) or a Cary 400 (Varian Ltd.) UV-Vis spectrophotometer. 10 mm pathlength standard 9/9/B quartz cuvettes (Starna Scientific) were used for recording all spectra. Scanning was set for 250–800 nm, with a scan rate of 300 nm min^{-1}.

3.5.1 Haem-Binding

These experiments ran over several days. Each day, a fresh haem stock was made up to approximately 500 µM, and its concentration then determined exactly. To achieve this, a baseline of buffer (20 mM BTP, 100 mM KCl, pH 6.5) was first taken. Cytochrome b_{562} (Barker Group stocks) was then added to ∼15 µM. Haem was then added to ∼5 µM. A crystal of sodium dithionite was added and the cuvette sealed with parafilm. Scans were then made every two minutes until the spectra stabilised

(ca. 10–20 minutes). The absorbance at 426.5 nm ($\varepsilon_{426.5}$ = 181,000 $M^{-1}\,cm^{-1}$) was then used to determine the concentration, thereby calibrating the haem stock solution.

Following calibration, a baseline of a known volume of buffer (20 mM BTP, 100 mM KCl, pH 6.5) was taken. The protein (WT HemS, or one of its mutants, or a homologue) was added to ∼10 µM (the exact concentration is not important as long as it is clearly in excess of haem). Using the calibrated stock, haem was then added to exactly 5 µM. Scans were taken every two minutes until the spectra stabilised (ca. 10–40 minutes). The absorbance readings quoted were an average of the last five scanned by the spectrophotometer. Using the known haem concentration and the absorbances, it was possible to convert to extinction coefficients for wavelengths above 300 nm.

Each protein was scanned twice and a further average taken. Another set of measurements were carried out at a pH value of 8.0.

3.5.2 Steady State Reaction of *holo*-HemS with NADH

Whenever an absolute spectrum was taken, a baseline of buffer solution was recorded, and this baseline subtracted from the sample data. Difference spectra were taken in a number of ways. Unless otherwise stated in the text, the baseline for these difference spectra consisted of the protein, haem and buffer, with spectra being recorded upon injection of NADH. It was important to allow the protein and haem to mix thoroughly before recording spectra to allow free and bound haem to equilibrate. A period of at least one hour is recommended to allow for equilibration. This interval was not always adhered to for experiments carried out early on in the project, where it was not appreciated how slow haem-binding could be; for experiments and spectra where this condition was the case, this is highlighted in the text.

Temperature was maintained at 25 °C unless otherwise specified by a Peltier heating block. Spectra were recorded in 1 nm intervals using a spectral bandwidth of 1 nm with full slit height.

3.6 Pre-Steady State Reaction Time-Course Using Stopped-Flow

An SX 20 stopped-flow spectrometer (Applied Photophysics) was used for all stopped-flow experiments. A Xe lamp was used as the light source, and the entrance and exit monochromator slit widths were both set to 2 mm. A photodiode array (PDA)

detector (Applied Photophysics), which had a resolution of 1.2 nm and a range of 250.0–722.9 nm, was used for most experiments, with a step size between wavelengths of 2.2 nm. The PDA was operated with an integration period of 1.260 ms, scan period of 1.097 ms, offset of 0 V, and gain of 127. For those experiments which required wavelengths longer than 722.9 nm (typically 806.0 nm), an absorbance photomultiplier (Applied Photophysics) with a 515 nm glass filter was used. All spectra were taken against a baseline of SEC Buffer. A water bath was used to keep the temperature at 25 °C.

Unless otherwise stated, all experiments consisted of a pre-incubated 1:1 sample of protein and haem being mixed with an excess of NADH.

It was important to take extra care in keeping consistency between experiments as stopped-flow is a sensitive method. Mixing between *apo*-protein and haem was always done the day before a set of experiments was to be recorded in order to allow for appropriate equilibration of the *holo*-protein structure. Washing of the 2.5 mL drive syringes before and between runs was thorough. Each was loaded with SEC buffer, which was flushed out by five consecutive drives. Following the fifth drive, a baseline was collected. The left-hand drive syringe was then loaded with the *holo*-protein mixture, and the right-hand drive syringe with NADH; this order was always maintained to avoid cross-contamination of the syringes or pistons. Two consecutive drives were performed, and the drive syringes then fully reloaded. A further two drives were then performed. Following this, data collection was started. Every experiment was done in triplicate, and so three further drives were required, thus emptying the drive syringes. The reason for doing two drives before data collection after fully reloading the drive syringes was to ensure there were no bubbles in the sample when it came to recording absorbance values. The reason why an additional two drives had been performed prior to this reloading was to ensure that the mixing chamber had been thoroughly flushed through with the solutions under study before data acquisition.

Initial analysis of the collected data was conducted using ProK-IV software (Applied Photophysics). Detailed model building was then performed using KinTek.[166]

3.7 Anaerobic Reaction

Both as a proof of principle and when high-purity products were required (see Results) it was necessary to sometimes perform the haem breakdown reaction with NADH under anaerobic conditions. To achieve this condition, vessels containing appropriate stocks of both *holo*-HemS and NADH were fitted with Suba seals and

thoroughly degassed (for at least 30 minutes) with N_2. A syringe was used to transfer the appropriate volume of NADH to the *holo*-HemS solution, and the reaction allowed to run for at least 75 minutes, whilst maintaining a N_2 overflow throughout. A deep purple colour developed over time.

3.8 Extraction and Purification of the NADH-Dependent Haem Breakdown Product

Two methods were attempted to extract the haem breakdown product from HemS.

Method 1 employed a 75 mL scale. WT HemS, haem and NADH were reacted in a 5 µM : 20 µM : 2 mM ratio for 30 minutes in a shaker-incubator (200 rpm, 25 °C). The solution turned a murky, dark purple colour in this time. 1-butanol was used to extract this product from the aqueous mixture. The organic layer was then dried using $MgSO_4$ and filtered under vacuum. Using a Schlenk line, 1-butanol was then drawn off by vacuum, leaving a dark purple residue.

So as not to overload the solid phase extraction cartridge, Method 2 employed a 10 mL scale. HemS and haem were mixed stoichiometrically to minimise residual unbound haem; therefore, WT HemS, haem and NADH were mixed together in a 10 µM : 10 µM : 2 mM ratio. After reaction for 75 minutes in a shaker-incubator (200 rpm, 25 °C), the sample was loaded onto an activated (with 20 mL SEC Buffer, then 20 mL acetonitrile, then a further 20 mL SEC Buffer) reverse-phase Hypersep C18 solid phase extraction column (Thermo Scientific). The sample was washed with deionised water (5×20 mL), removing the protein and NADH. The haem breakdown product was then eluted as a deep purple solution using 3 mL acetonitrile.

3.9 Nuclear Magnetic Resonance Spectroscopy

The extracted haem breakdown product was dissolved in either D_2O or d^6-DMSO. All NMR data was collected at 298 K using a 400 MHz Bruker Avance NMR spectrometer with a cryoprobe. Experiments were kindly performed by George Biggs.

3.10 Mass Spectrometry

Data on small molecules were collected using a Xevo G2-S ToF mass spectrometer (capillary voltage 2 kV, cone voltage 40 kV; desolvation temperature 350 °C), with error limits of ±5 ppm mass units. Injection was automatic, and nitrogen was used

as the desolvation gas, with a total flow of 850 L h^{-1}.

Data on proteins were collected by liquid chromatography mass spectrometry (LCMS) using a Xevo G2-S ToF mass spectrometer (capillary voltage 2.0 kV, cone voltage 40 kV; desolvation temperature 350 °C) coupled to an Acquity UPLC BEH300 C4 column (1.7 µm, 2.1×50 mm). The mobile phase consisted of H_2O with 0.1% formic acid (Solvent A) and 95% : 5% acetonitrile : H_2O with 0.1% formic acid (Solvent B), which was run at a flow rate of 0.2 mL min^{-1}. The gradient was run as follows: 95% A for 0.93 min; gradient to 100% B for 4.28 min; 100% B for 1.04 min; gradient to 95% A for 1.04 min. Nitrogen was used as the desolvation gas, with a total flow of 850 L h^{-1}. The maximum entropy (MaxEnt) algorithm[167] pre-installed on the MassLynx software (Waters, v4.1) was used to reconstruct total mass spectra from the ion series. Deconvoluted spectra were consistently split into 0.25 Da channels, and used a width at half-peak height of 0.75 Da.

Most experiments were kindly performed by George Biggs or Jamie Klein.

3.11 X-ray Crystallography

3.11.1 Crystallisation

An anaerobic reaction, as described in Section 3.7, was carried out to give as pure a product as possible with limited verdohaem/biliverdin formation. 10 µM protein, 10 µM haem and 2 mM NADH were reacted in a total volume of 17.6 mL. The product mixture was then exchanged into Crystallisation Buffer A and concentrated to 30 mg mL^{-1} by centrifugation (4000g, 4 °C) using 10 kDa Amicon Ultra centrifugal filters. 5 µL samples were applied to individual glass coverslips and mixed with 5 µL Crystallisation Buffer B. These coverslips were then applied to individual wells for crystallisation by the hanging drop method, with 700 µL Crystallisation Buffer B used as the precipitant solution in each well. Crystals were incubated at 277 K, typically for 2-4 weeks.

3.11.2 Data Collection and Analysis

The following was conducted by Dr Paul Brear. Cryoprotection of the crystals was performed using a solution of 0.1 M Tris-HCl pH 8.5, 1.8 M ammonium sulfate, 2% PEG 400, and 1.2 M sodium malonate. Crystals were then cryo-cooled in liquid nitrogen for data collection. X-ray diffraction data were collected at the Diamond Light Source (Didcot, UK) and data derived from automated data processing using autoProc[168] were utilised for the structure determination. Structures were solved by

using programs in the CCP4 package.[169] Models were iteratively refined and rebuilt using the Refmac[170] and Coot[171] programs.

3.12 Synthesis and Characterisation of (*R*)/(*S*)-NADD

The two NADD stereoisomers were synthesised and purified using methods adapted from Northrop & Duggleby,[172] Basran *et al.*[173] and Pudney *et al.*[174]

(*R*)-[4-^{2}H]-NADD. 11.4 mL 10 mM NH_4HCO_3 set to pH 8.5 and 0.6 g 1-[$^2H_6$]-ethanol (Eurisotop) were mixed. The pH was monitored to ensure it remained at 8.5, requiring occasional adjustment with 1 M NaOH. 120 U alcohol dehydrogenase (from *Saccharomyces cerevisiae*, EMD Millipore) and 60 U aldehyde dehydrogenase (from *Saccharomyces cerevisiae*, EMD Millipore) were added, keeping the pH at 8.5 throughout. 300 mg NAD^+ (Roche Diagnostics GmbH) was then added **slowly**, again ensuring the pH was kept steady at 8.5. After all the NAD^+ was added, the reaction was monitored and further NaOH added to ensure the pH did not drop far below 8.5. This was done until the pH stopped decreasing, indicating reaction completion. Anion exchange chromatography was then used to separate any unreacted NAD^+ from (*R*)-NADD. To achieve this separation, the sample was applied to a Mono-Q HR 10/10 column (Pharmacia Biotech, now GE Healthcare) pre-equilibrated in NADD Buffer A (10 mM NH_4HCO_3, pH 8.5). The column was washed with further NADD Buffer A, before introducing a gradient with NADD Buffer B (500 mM NH_4HCO_3, pH 8.5). (*R*)-NADD eluted at approximately 300 mM NH_4HCO_3. Relevant fractions were pooled and freeze-dried to concentrate the samples and exchange into SEC buffer. Purity was then checked by UV-Vis spectroscopy – a ratio of $A_{260}/A_{340} \leq 2.3$ was deemed pure.[173] Stocks were frozen at −20 °C.

(*S*)-[4-^{2}H]-NADD. 15.2 mL 10 mM NH_4HCO_3 set to pH 8.5 and 0.8 g 1-[^{2}H]-glucose were mixed. The pH was monitored to ensure it remained at 8.5, requiring occasional adjustment with 1 M NaOH. 200 U glucose dehydrogenase (from *Pseudomonas* sp.) was added, keeping the pH at 8.5 throughout. 400 mg NAD^+ was then added **slowly**, again ensuring the pH was kept steady at 8.5. After all the NAD^+ was added, the reaction was monitored and further NaOH added to ensure the pH did not drop far below 8.5. This was done until the pH stopped decreasing, indicating reaction completion. Anion exchange, fraction pooling and freeze-drying were then carried out as described for (*R*)-NADD.

To ensure that the correct stereoisomers had been made, these samples were characterised by ^{1}H NMR spectroscopy using a 400 MHz Bruker Avance NMR spectrometer with a cryoprobe. Spectra were kindly taken by Jamie Klein. These spectra, confirming the correct stereoisomers were made, are given in Section 5.3.

Chapter 4

Computational Methods

4.1 AMBER Potential and Force Field

Before embarking on a semi-empirical computational study, it is essential the correct potential and parameter set are selected. The AMBER force field of Case *et al.*[175] fulfils many of the criteria required. In providing an all-atom representation, it gives higher resolution for protein-ligand interactions than are possible with united atom and coarse-grained representations. As outlined in Eq. (1.2), the AMBER potential has terms accounting for covalent bonds with respect to their lengths, angles and dihedrals, as well as further terms for van der Waals and electrostatic interactions. A further strength is that AMBER can be GPU-accelerated.[176;177]

The AMBER12[178] package was used throughout this work as later packages, such as AMBER16,[179] had not been linked to any of the Wales group programs when the project started. The ff99SB force field[180] was the sole force field used. This was the force field used by Choy[21] and Shang,[159] and so using it provided continuity. It is based on the original ff99[181] parameterisation, which has been demonstrated to be effective at simulating short peptides and DNA helices.[181] However, this force field proved to be weak at reproducing energy differences between secondary structures arising from protein backbone atoms.[134;182;183] Backbone torsions were therefore reparameterised, giving the ff99SB force field. Since 2006 (when ff99SB was developed), new force fields have been published, which are considered to give more accurate representations for protein structure. However, it was judged that such improvements would not offset the inconsistency of using a different force field from that used by Choy and Shang. Indeed, ff99SB is still a widely used force field, as demonstrated by many recent studies.[184–186] As well as having good protein secondary structure balance, ff99SB has been demonstrated within the Wales group to perform well with a multitude of Generalised Born implicit solvent models,[21]

making it a suitable force field to simulate protein pockets.

4.2 Small Molecule Parameterisation

AMBER force fields tend to provide parameters for the building blocks essential to large biomolecules only, which are sufficient for investigating protein folding or DNA twisting problems. However, for a study of protein-ligand interactions, further parameters not included in the standard AMBER package are required in the input topology files. Fortunately, AMBER provides methods both to prepare (antechamber)[187] and read in (LEaP) more unusual molecules. Care must be taken with antechamber to ensure that parameters generated truly reflect the properties of the molecule being investigated. There is also a problem with standardisation – for example, different groups investigating the same molecule may generate different and inconsistent parameter sets. One solution is to use parameters generated and curated by established groups. This strategy was followed in the current thesis. Haem and NADH (as well as NADPH and NAD^+) are all common biomolecules for which AMBER parameters have already been generated. These parameters are curated by the Bryce Group at the University of Manchester,[188] and so the parameters were taken from there. Specifically, those parameters used were the all-atom haem representation adapted from Giammona,[189] the NADH and NAD^+ representations from Walker[190;191] and the NADPH representation from Ryde.[192]

4.3 Implicit Solvent Model

To approximate an aqueous, ionic environment, two strategies have been developed. The first is the inclusion of explicit solvent molecules. Although this is the most accurate method, it requires many water molecules and various ions, and so excessive computational effort becomes necessary to determine their respective conformations in addition to those for the molecule(s) of interest. A far less computationally expensive strategy is to include a solvent implicitly. This is typically a mean-field approach which uses the system configuration to approximate solute-solvent interactions. In AMBER, the total solvation free energy, ΔG_{solv}, is made up of two distinct, non-polar and electrostatic, contributions,

$$\Delta G_{solv} = \Delta G_{np} + \Delta G_{el}. \tag{4.1}$$

The non-polar contribution is derived empirically, and is designed to reflect surface tension arising from the solvent-exposed surface area. The electrostatic contribution,

meanwhile, is required to chart the distribution of electric potential perpendicular to a charged surface in solution. A numerical solution can be derived from the Poisson-Boltzmann equation, but is computationally prohibitive. An analytical approach, known as the Generalised Born (GB) model, can be used instead to reproduce an approximated Poisson-Boltzmann result at a fraction of the computational cost. This approach builds on the Born formula, the analytical solution for a spherical particle with radius R and charge q:

$$\Delta G_{el} \;=\; -\frac{\tau q^2}{2R}\,,\ \text{where}\,\tau \;=\; \frac{1}{\epsilon_{protein}} - \frac{1}{\epsilon_{solv}}. \tag{4.2}$$

The GB formula is simply a polyatomic extension of this equation.[193] The most common formula used, from Still *et al.*,[194] is,

$$\Delta G_{el} \;=\; -\frac{\tau}{2}\sum_{i,j}\frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)}\left(1 - \frac{e^{-\kappa f_{GB}}}{\epsilon}\right), \tag{4.3}$$

where r_{ij} is the distance between atoms i and j. These atoms have respective charges (q_i,q_j) and effective Born radii (R_i,R_j). κ is the Debye-Hückel screening parameter. To make the simulation more realistic, it is necessary to include the electrostatic screening effects of monovalent salt, which is estimated by $\kappa/\text{Å} \approx 0.316\sqrt{[salt]}$. A salt concentration of 0.1 M is consistently used as an estimation of the salt content of a typical bacterial cell throughout this thesis. Furthermore, for the GB model in the current work, the choice for f_{GB}, consistent with Still *et al.*,[194] is set as,

$$f_{GB} \;=\; \sqrt{r_{ij}^2 + R_i R_j e^{\left(-\frac{r_{ij}^2}{4R_i R_j}\right)}}. \tag{4.4}$$

Distinctions between GB models mainly arise due to different methods of calculating the effective Born radii.[134] Each Born radius describes a solute atom within the overall configuration of the total solute; a way of expressing this is that it approximates atomic burial from the molecular surface. As these values depend on the Coulomb field approximation and differ with each changing molecular conformation, their calculation can become prohibitive too. Therefore, further approximations are introduced. Different strategies for approximation are implemented by the various *igb* models in AMBER. The simplest is *igb*1,[195] which uses,

$$R_i^{-1} \;=\; \rho_i^{-1} - \frac{1}{4\pi} I, \tag{4.5}$$

to estimate the Born radii, where ρ_i approximates the van der Waals radius of atom i and I is an integral which sums over the volume surrounding i. Choy briefly con-

sidered using this model when beginning computational work on HemS.[21] However, as it treats all spaces between van der Waals radii as being filled with water, rather than vacuum,[196] it dramatically underestimates the effective radii of buried atoms in macromolecules.[21;197]

An alternative to the model employed by *igb*1 is,

$$R_i^{-1} = \tilde{\rho}_i^{-1} - \rho_i^{-1} \tanh\left(\alpha\tilde{R}_i I - \beta(\tilde{R}_i I)^2 - \gamma(\tilde{R}_i I)^3\right), \tag{4.6}$$

where $\tilde{\rho}_i^{-1}$ is the 'intrinsic radius' (i.e. ρ_i^{-1} offset by a uniform value, 0.09 Å). α, β and γ are dimensionless parameters determined experimentally from pH titrations. Some versions of *igb* assign different values to these parameters. Based on benchmarking exercises, Choy found that *igb*2,[197] for which $\{\alpha,\beta,\gamma\} = \{0.8,0.0,2.909125\}$, was most suitable for the [HemS + Haem + NADH] system,[21] and so this implicit solvent model is used throughout this thesis.

4.4 Generating AMBER Input Files

AMBER requires input coordinate and topology files. For this project, input files for the following systems were required:

- WT HemS + Haem
- WT HemS + Haem + NADH
- WT HemS + Haem + NAD^+
- WT HemS + Haem + NADPH
- F104A HemS + Haem + NADH
- F104AF199A HemS + Haem + NADH
- F104I HemS + Haem + NADH
- F199A HemS + Haem + NADH
- R209A HemS + Haem + NADH
- R209K HemS + Haem + NADH
- Q210A HemS + Haem + NADH
- WT HmuS + Haem + NADH

- WT ChuS + Haem + NADH
- WT ShuS + Haem + NADH
- WT PhuS + Haem + NADH

Input files for the first two of these listed systems already existed from Choy.[21] Attempts were made to keep new input files as consistent with Choy's original as possible.

The initial coordinate files for these two systems had been taken from the *holo*-HemS PDB structure, 2J0P.[96] Solvent and crystallant molecules were removed and the missing loop (see 1.4.1) reconstructed using Swiss-PdbViewer.[198] For the [WT HemS + Haem + NADH] system, NADH coordinates were appended to the coordinate file, its position within the protein cavity having been determined by Relibase$^+$.

To generate a topology file, the LEaP program had been used. Additional parameter files for the ligands, haem and NADH, were necessary, and so they were uploaded to LEaP. Hydrogen atoms were also added. None of the non-histidine charged residues were neutralised since they were either exposed to the surface or part of a salt bridge. All histidine residues were neutralised, however, to better reflect the pH of 6.5 HemS-based experiments are typically conducted at. Since it is ordinarily the atom most capable of being protonated, all but one histidine therefore included this extra proton at the Nε position. The exception was H196, which was protonated at the δ position, to allow it to coordinate to haem, as in the crystal structure. Indeed, an explicit bond was placed between haem FE and H196 Nε.

This choice proved to be a significant point of deviation when it came to the generation of input files for all of the other systems in this thesis. The present author is of the opinion that this explicit bond between haem FE and H196 Nε is not fully justified. Although experiment has indeed shown that haem binds strongly *via* its iron ion to H196, the inclusion of a covalent bond here has the potential to unduly bias optimisations towards structures where haem is explicitly bound. It is perhaps possible instead for haem to move within the pocket with slightly more freedom. For the present study, this bond was therefore removed for all of the other systems. It was, however, retained for the [WT HemS + Haem] and [WT HemS + Haem + NADH] systems to keep consistency between the results generated in this thesis and those generated previously by Choy[21] and Shang.[159]

Otherwise, generating topology files for these other systems was effectively the same. When preparing the LEaP files, it was ensured that the relevant mutations were included, or alterative sequences (for HmuS, ChuS, ShuS and PhuS) were used.

If the ligands changed, then the appropriate files enumerated in Section 4.2 were downloaded from the Bryce Group database.[188]

Starting coordinate files for these alternative systems were generated using **CHECKSPMUTATE**. This routine was written during the course of the project, and it shall be discussed in more detail in Chapter 6.

4.5 Basin-Hopping and Minima

Section 1.6.3 showed that a PES can be reduced to a collection of minima and transition states, and suggested there are various methods capable of finding these stationary points. **GMIN**[158] is a program encoding one such method, developed by the Wales group, which can be used to find relevant minima efficiently.

GMIN was designed to find the global minimum of a system stochastically. It has been successfully applied to protein folding problems. It would be unrealistic, however, to expect it to find the global minimum for a system composed of three biomolecules, since intermolecular conformations would need to be optimised in addition to intramolecular ones. Instead, **GMIN** in this thesis was used to sample local conformations of NADH at various stages along the pocket. This approach allowed for optimal NADH conformations to be determined at various stages of its progress through the protein cavity in an efficient manner, thus making it easier for a fully connected pathway involving both minima and transition states to be identified using discrete path sampling (see Sections 4.6 and 4.7).

GMIN includes the basin-hopping (BH) algorithm, and transforms the PES to simplify its structure. The aim is to find the global minimum on this simplified surface. Local minimisation is applied to achieve this landscape transformation:[133]

$$V^*(\mathbf{X}) = \min\{V(\mathbf{X})\}. \tag{4.7}$$

Here, $V^*(\mathbf{X})$ is the transformed energy after a local energy minimisation is carried out starting from $\mathbf{X}$. Therefore, all points on the PES are transformed by moving downhill – crucially, without crossing any barriers – to a local minimum. The PES is therefore converted into a configuration space that is partitioned into catchment basins.[133] The plateaux of this new representation correspond to the energies of the local minima for the original PES, as illustrated in Fig. 4.1. Having transformed the landscape, a search strategy must be implemented. The strategy used in this report is similar to Li and Scheraga's 'Monte Carlo plus energy minimisation' (MCM) procedure.[199;200] In general, Monte Carlo-type sampling techniques work by perturbing the coordinates of a minimum with energy V_{old} before minimising to generate a

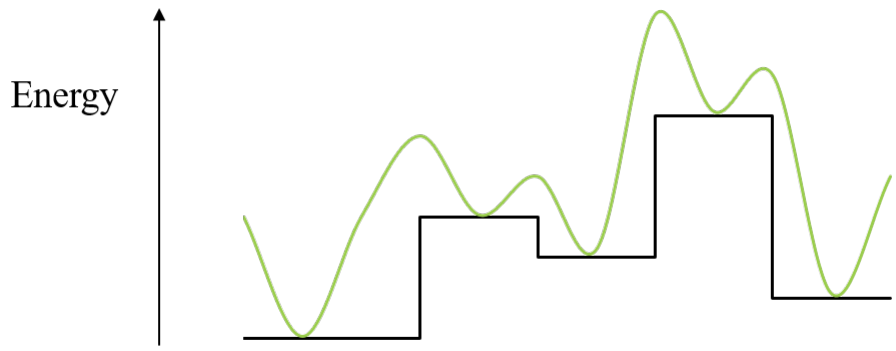


Figure 4.1: Schematic of the landscape transformation applied in BH. The curved green line represents the original PES, whereas the stepped black line represents the transformed surface.

new minimum of energy V_{new}. The step is accepted or rejected according to:

$$Step = \begin{cases} Accept & (V_{new} < V_{old}) \text{ or} \\ & (V_{new} > V_{old}) \text{ and } \exp[(V_{old} - V_{new})/k_B T] > \text{Ran}[0,1] \\ Reject & (V_{new} > V_{old}) \text{ and } \exp[(V_{old} - V_{new})/k_B T] < \text{Ran}[0,1] \end{cases} \quad (4.8)$$

In other words, the step is always accepted if the energy of the new minimum is less than the one preceding it. Otherwise, it is only accepted if a function of the energy increase is greater than a threshold value, which is generated randomly for every step taken.[201–203] The structure can then either be reset to V_{new} or V_{old} (depending upon whether the step was accepted or rejected) as is the case in the MCM procedure, or allowed to vary continuously.[133] A significant advantage of this scheme is that the effective temperature is the only variable when a fixed average acceptance ratio is achieved *via* dynamic step size adjustment.[204] This scheme can therefore be transferred between different, unrelated systems. Perturbations are also local and so the transformation expressed by Eq. (4.7) is not applied to the surface as a whole for each step.

This search strategy, however, still depends on an effective minimisation procedure. It is particularly advantageous to use the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)[205–210] algorithm for larger systems. This is a quasi-Newton optimisation algorithm where an approximation of $\mathbf{H}^{-1}$, the inverse Hessian,

is stored implicitly. This representation allows for the required storage of data to scale with respect to N^2 as opposed to N^3, an important consideration when the number of atoms, N, is large.[133] A line search is typically used to dictate a suitable step size. However, a different algorithm has been implemented in GMIN,[158] which has been determined to be more efficient.[211] Rather than having to calculate a specific step size along the descent direction found, this algorithm is less stringent, instead ensuring that the step along the descent direction does not produce an excessive energy rise or exceed a maximum step size. The step size is scaled down if either of these conditions is breached.[211]

4.6 Transition States

For the PESs being investigated in this thesis, and to give accurate kinetic information, transition states connecting the minima need to be identified.

A minimum on the PES is a stationary point with positive curvature in all directions aside from degrees of freedom corresponding to overall translations and rotations. A transition state has one negative eigenvalue and hence negative curvature in one direction (the reaction coordinate) but retains positive or zero curvature in all other directions.[212] Therefore, to find a transition state, the energy must be minimised along all modes apart from along the reaction coordinate, where the energy should be maximised. The system is balanced in one degree of freedom, thus requiring a precise search strategy.

The strategy implemented in the program OPTIM[213] is to use the doubly-nudged elastic band (DNEB)[214;215] algorithm to generate transition state candidates, and then use hybrid eigenvector-following (HEF)[216–218] to refine them.

DNEB transition state searches begin by attempting to connect two minima. A direct path may not necessarily connect these two minima, in which case more (intermediate) minima and transition states may need to be considered (for more details, see Section 4.7). In the simple case where a direct path is possible, the DNEB search operates as follows.[214;215]

After receiving the two minima geometries, $\mathbf{X}_0$ and $\mathbf{X}_{n_{spr}+1}$, as input, a series of interpolated geometries known as images, $\{\mathbf{X}_1,\mathbf{X}_2...\mathbf{X}_{n_{spr}}\}$, are generated. $\mathbf{X}_i$ is simply a vector used to represent the coordinates of an endpoint or image, i.

An attractive spring interaction is placed between the images to ensure that they do not simply collapse to the endpoint minima. This spring is what gives rise to the elastic band appellation: it can be imagined that an elastic band has been stretched across the PES and terminates at the endpoint structures. These springs

connect atoms of adjacent images, and so the spring potential, $\tilde{V}$, is distinct from the 'true' potential, V_t, which binds the atoms within each image together. In order to connect all of the images interpolated between the endpoints, n_{spr}+1 springs are required. The overall 'spring' potential introduced is therefore:

$$\tilde{V} = \frac{1}{2} k_{spr} \sum_{i=1}^{n_{spr}+1} |\mathbf{X}_i - \mathbf{X}_{i-1}|^2. \tag{4.9}$$

Optimisation is carried out to minimise the forces acting on the images. In practice, this is difficult because the 'spring' potential interferes with the 'true' potential, the magnitude of such interference being system dependent.[219] To overcome this problem, both the 'true' and 'spring' gradients, $\mathbf{g}$ and $\tilde{\mathbf{g}}$, are split into components parallel and perpendicular to the path: $\mathbf{g}^{\parallel}$, $\mathbf{g}^{\perp}$, $\tilde{\mathbf{g}}^{\parallel}$ and $\tilde{\mathbf{g}}^{\perp}$. 'Nudging' the elastic band results in $\mathbf{g}^{\parallel}$ and $\tilde{\mathbf{g}}^{\perp}$ being projected out, flattening the PES and pulling the images down, respectively.[134;220–222] Taken together, this procedure results in a more stable band, and the new gradient is simplified to,

$$\mathbf{g}^{NEB} = \mathbf{g}^{\perp} + \tilde{\mathbf{g}}^{\parallel}. \tag{4.10}$$

Following tests, Trygubenko & Wales[214] found that giving the band a 'second nudge' (encapsulated by $\tilde{\mathbf{g}}^{*}$ in Eq. (4.11)) causes the divergence of images from the path to be reduced. This new 'doubly-nudged' gradient, which reincorporates an element of the perpendicular component of the spring gradient, can therefore be expressed as:

$$\mathbf{g}^{DNEB} = \mathbf{g}^{\perp} + \tilde{\mathbf{g}}^{\parallel} + \tilde{\mathbf{g}}^{*}, \tag{4.11}$$

where

$$\tilde{\mathbf{g}}^{*} = \tilde{\mathbf{g}}^{\perp} - (\tilde{\mathbf{g}}^{\perp} \cdot \hat{\mathbf{g}}^{\perp})\hat{\mathbf{g}}^{\perp}. \tag{4.12}$$

The L-BFGS algorithm is then used to minimise $\mathbf{g}^{DNEB}$. The evolution of the band during this DNEB process is illustrated in Fig. 4.2.

Once the images have been optimised to below a specified root mean square (RMS) force, HEF is then used to refine candidates which appear to be maxima.

HEF,[216–218] in contrast to DNEB, is a single-ended method for determining transition state structures.

EF in general is a geometry optimisation method, which uses first and, optionally, second derivatives of the potential. Specifically, a Taylor expansion around the

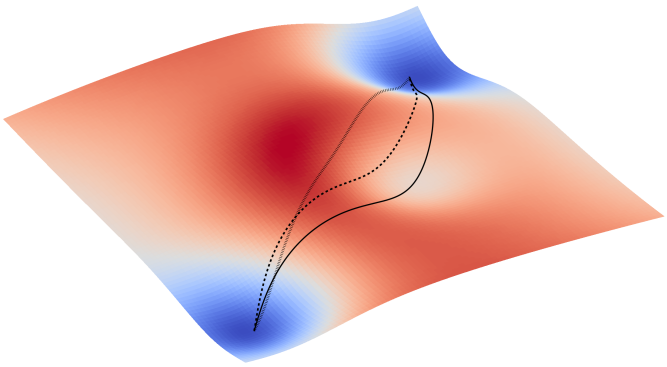


Figure 4.2: Representation of the evolution of an interpolation band as the DNEB algorithm progresses. Blue areas corresponds regions of low potential energy (the endpoint minima being connected) and red to areas of high potential energy. The line with fine dashes represents the original linear interpolation. The dashed and solid lines show progressively more optimised bands. Local maxima are taken as transition state candidates for refinement by hybrid eigenvector-following (HEF). Figure reproduced from Röder.[134]

present geometry, $\mathbf{X}$, can be implemented:

$$V(\mathbf{X}+\mathbf{x}) = V(\mathbf{X}) + \mathbf{G}(\mathbf{X})^{\hat{T}}\mathbf{x} + \frac{1}{2}\mathbf{x}^{\hat{T}}\mathbf{H}(\mathbf{X})\mathbf{x}. \tag{4.13}$$

Here, $\mathbf{G}(\mathbf{X})$ and $\mathbf{H}(\mathbf{X})$ are the gradient and Hessian at $\mathbf{X}$ respectively, with $\mathbf{x}$ being a small displacement. If the following condition is applied,

$$\frac{\mathrm{d}V(\mathbf{X}+\mathbf{x})}{\mathrm{d}\mathbf{x}} = \mathbf{0}, \tag{4.14}$$

then the standard Newton-Raphson formula can be derived:[133]

$$\mathbf{x}_{\mathrm{NR}} = -\mathbf{H}^{-1}\mathbf{G}. \tag{4.15}$$

Diagonalisation of the Hessian,

$$\mathbf{H}\boldsymbol{\nu}_i = \tilde{\lambda}_i\boldsymbol{\nu}_i, \tag{4.16}$$

can then be carried out, giving a Newton-Raphson step of

$$\mathbf{x}_{\mathrm{NR}} = \sum_{i}^{3N} \frac{-F_i}{\tilde{\lambda}_i}\boldsymbol{\nu}_i, \tag{4.17}$$

where $\tilde{\lambda}_i$ and $\boldsymbol{\nu}_i$ are eigenvalue and eigenvector i, N is the number of atoms, and F_i is the component of the gradient along each eigenvector. Eq. (4.17) decreases in energy when a positive eigenvalue is being followed and increases when a negative eigenvalue is being followed.

This method has an important drawback, however, in that the Hessian scales quantitatively with the number of atoms. At very large numbers of atoms, therefore, the size of the Hessian becomes very substantial, rendering its calculation and diagonalisation (which scales as N^3) computationally expensive.

A significant advantage of HEF is that it does not require the Hessian matrix of second derivatives to be calculated at any point during its operation. Instead, a variational approach can be used to find the smallest eigenvalue along with its associated eigenvector.[133;223] First, a Rayleigh-Ritz ratio for a displacement $\mathbf{x}$ from the present geometry is defined:

$$\tilde{\lambda}(\mathbf{x}) = \frac{\mathbf{x}^{\hat{T}}\mathbf{H}\mathbf{x}}{\mathbf{x}^2}. \tag{4.18}$$

Though it may appear from this equation that the Hessian, $\mathbf{H}$, is still essential to solving this problem, it is now possible to use a numerical second derivative to calculate $\tilde{\lambda}(\mathbf{x})$, where $\xi \ll 1$:[211]

$$\tilde{\lambda}(\mathbf{x}) \approx \frac{\{\nabla V(\mathbf{X} + \xi\mathbf{x}) - \nabla V(\mathbf{X} - \xi\mathbf{x})\} \cdot \mathbf{x}}{2\xi\mathbf{x}^2}. \tag{4.19}$$

By minimising this ratio iteratively, the smallest non-zero eigenvalue can be determined. This, in turn, allows its associated eigenvector to be calculated. Uphill steps can therefore be taken in this direction whilst the orthogonal subspace is minimised. It has been found that the L-BFGS algorithm is also effective in carrying out these calculations, and so it is used to carry out the HEF steps.[224]

To summarise, the DNEB & HEF approach is relatively inexpensive computationally as it is capable of determining the eigenvalue and eigenvector of a transition state without needing to calculate or diagonalise the Hessian. After the eigenvalue-eigenvector pair has been found, an uphill step is taken in the direction of the eigenvector, followed by partial minimisation orthogonally. This procedure is repeated iteratively until a user-defined RMS force convergence criterion is achieved.

4.7 Discrete Path Sampling

The previous section showed how a transition state can be found by interpolating in a region I between two endpoint minima, A and B. Thus far, however, no method has been provided to determine whether this transition state is indeed directly connected to these starting minima. For long, complicated pathways, this situation is unlikely as it is probable that there will be intervening local minima in I. Therefore, a

methodology is required to determine how individual stationary points are connected within the pathway. Discrete path sampling (DPS) achieves this objective, first by determining the two local minima a transition state is directly connected to. This connection is known as a minimum-TS-minimum 'triple'. DPS can then connect these 'triples' to give an overall chain between the two original endpoint minima. This chain is known as a discrete path.

The minima directly connected to transition states are determined using steepest-descent algorithms. The steepest-descent paths, defined in Eq. (4.20), can be followed in both the parallel and antiparallel directions to give two connected local minima (with any such minima belonging to the intervening I set):

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}s} = -\frac{\mathbf{G}(\mathbf{X})}{G(\mathbf{X})}. \tag{4.20}$$

In this steepest-descent equation, s is the integrated path length between two points on the curve, $\mathbf{G}(\mathbf{X})$, as in previous definitions, is the gradient vector, and $G(\mathbf{X})$ is the magnitude of the gradient ($G(\mathbf{X}) = \mid \mathbf{G}(\mathbf{X}) \mid$).[133] When launched from a true transition state, there are only two solutions to Eq. (4.20).[225]

One of the minima the steepest-descent path reaches will sometimes be that from which the transition state search was first instigated. A decision then has to be made whether to remain at this original minimum or to move to the newly connected one in preparation for the next interpolation and transition state search. It is also possible that the transition state is not connected to either of the original minima at all, in which case these new minima can be saved separately and then checked periodically to determine whether they have been connected to the database in any subsequent search.[133]

When the two endpoints chosen are far apart in configuration space, the possible number of connections that could be made becomes combinatorial. An efficient selection process for searches therefore must be utilised. DPS uses the Dijkstra shortest path algorithm,[135] a greedy algorithm, which describes the total set of minima (whether they belong to A, B or I) as a weighted, directed graph, $\zeta(M_s,E_s)$.[226] Here, M_s is the set of minima and E_s the set of edges connecting them. In other words, a complete graph is defined from a set of edges, with an edge assigned to each pair of (i,j) minima. An edge weight, $w(i,j)$, defines the state of connection of these pairs:[226;227]

$$w(i,j) = \begin{cases} 0, & \text{if a known path exists between } i \text{ and } j \\ \infty, & \text{if } n_u(i,j) = n_{max} \\ f(D(i,j)), & \text{otherwise.} \end{cases} \tag{4.21}$$

The rationale for this algorithm is to identify the shortest connected path and, if that is not possible due to gaps in the pathway, to identify the most likely pairs of minima to bridge these gaps. An edge weight of 0 indicates that i and j are already connected. The shorter a gap is, the more likely it is to be selected. This behaviour becomes apparent when $f(D(i,j))$ is broken down. $D(i,j)$ is the Euclidean distance between the locally permutationally aligned minima i and j,[228] and $f(D(i,j))$ is a weighting function based on the distance. The edge weight is set to ∞ for a pair of minima if the number of connection attempts, n_u, reaches a user defined maximum, n_{max}; this condition is applied to prevent the same pair of minima from being attempted repetitively.

The Dijkstra algorithm is considered complete either once a path between the endpoints can be extracted or if a user-defined limit of searches is exhausted. The first scenario occurs when the total of the edge weights between the two selected endpoints reaches zero, which indicates that there are no gaps in the path. If the pathway between the endpoints still includes at least one edge weight which is non-zero, then more DNEB connection attempts will have to be made between those i and j minima that have non-zero weightings, in a new attempt to find transition states between them. This process cycles until all $w(i,j)$ values required to connect the endpoints become 0 (i.e. the path is completely connected).

The entire discrete path sampling process is outlined in Fig. 4.3.

4.8 From Initial Pathways to Complete Representations

Discrete Path Sampling is a useful tool for identifying possible pathways for chemical processes. However, as discussed in Section 1.6.3, protein-ligand interactions are typically complicated processes, involving many steps and multiple possible pathways. Two problems arise if DPS is used simply as described in the previous section:

1. The landscape is almost certainly undersampled. DPS has indeed found and identified a fastest pathway between the two starting endpoints, but it is unlikely to be as efficient as possible (in other words, the fastest path identified is not necessarily the true fastest path). There may be artificially high barriers and/or superfluous intermediate structures present. Further sampling is therefore required.

2. DPS gives the fastest single pathway. However, as discussed in Section 1.6.3, in order to derive relevant thermodynamic and kinetic information, a consid-

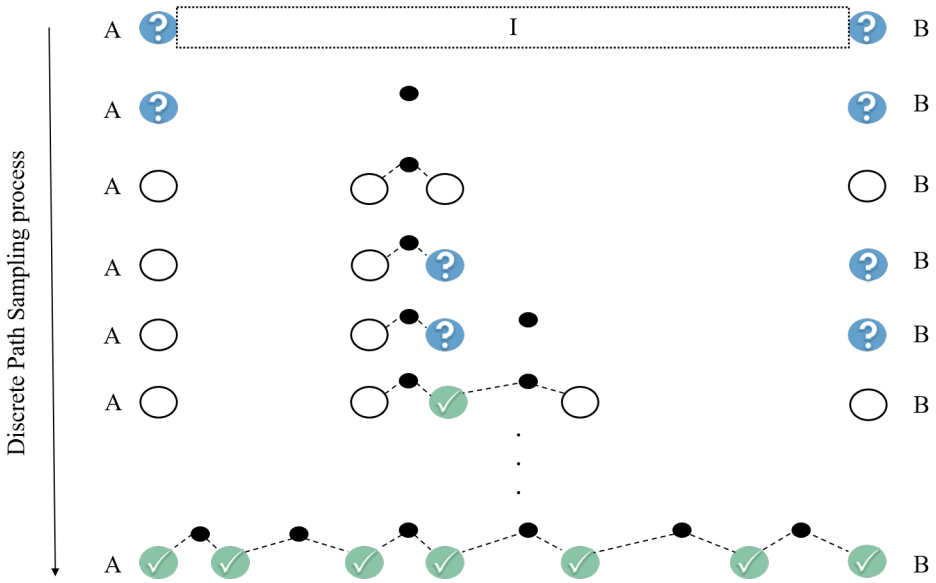


Figure 4.3: Discrete path sampling process to find a pathway. Small filled circles represent transition states. Large circles represent minima. Dashed lines between minima and transition states indicate a direct connection. If a minimum is filled in blue with a question mark, this means it has been selected as a candidate for an attempted connection. If filled in green with a tick, this indicates that the minimum is connected to two transition states in the pathway. Progression flows from top to bottom. The first step illustrates the selection of two endpoint minima A and B. A search is made in the intervening space, I, typically through the use of DNEB and HEF. Steepest-descent algorithms are then used to find the minima directly connected to this transition state candidate. These new minima are added to a stationary point database. A Dijkstra analysis is then used to select two points in the expanding database. Again, these two minima are represented by blue circles containing question marks, and an attempt is made to connect them in the same way that a connection between A and B was attempted before. This process continues until a fully connected path is found, or some user-defined number of connection attempts, n_{max}, is exhausted. This figure is inspired but adapted significantly from Whittleston.[229]

eration of the FES (which constitutes multiple pathways connecting basins rather than individual minima) rather than the PES is required.

To tackle the first of these issues, four main strategies are typically employed. These are listed and described below and in Fig. 4.4.

1. **Shortening Pathways.** Most often, DPS will find a fully connected pathway which is unnecessarily long. Due to the bias for connection attempts between minima that are conformationally similar, such pathways tend to consist of many intermediate, largely identical structures. This path is not necessarily the most efficient method of getting from A to B. Therefore, the **SHORTCUT** scheme,[226;230] employed in the program **PATHSAMPLE**,[231] selects pairs of minima on this 'fastest' path with a user-defined number of intervening transition states between them, and tries to connect them directly. If the connection is successful, many superfluous intermediates are therefore removed from the pathway.

2. **Removing Large Barriers.** A fully connected pathway found by DPS may also have artificially high barriers. During interpolation between minima, it is possible that the transition state(s) identified are high in energy. Unless there are specific conditions in place to ignore high-energy TSs, these are incorporated into the pathway. Therefore, the final fully connected pathway contains these high-energy TSs, significantly slowing down the fastest path. The **SHORTCUT BARRIER** scheme[230] selects the pairs of minima at either side of these high barriers, and tries to reconnect them *via* lower energy alternatives.

3. **Removing Kinetic Traps.** It is also possible that kinetic traps are present, here defined as low-lying minima separated by large barriers from A and/or B.[134] The **UNTRAP** scheme[230] selects minima that have high ratios of the energy barrier separating them from A or B versus the energy difference to those states. This is therefore similar to the **SHORTCUT BARRIER** scheme but tends to select minima further apart in conformational space. As such, it is not included in Fig. 4.4.

4. **Connecting Stationary Points to the *AB* Set.** As a database of stationary points expands, many transition states are located, which are not connected to either the A or B endpoints. Come the end of the DPS scheme and the identification of a fully connected pathway, many of these transition states and their steepest-descent-derived minima are still not connected in any way to either A or B. In other words, they are not part of the AB connected

set. Sometimes these AB-unconnected stationary points are themselves part of large, connected clusters. The database therefore contains a lot of information on relevant transition states and minima, which are nevertheless not included in the determination of the fastest fully connected path. The **CONNECTUNC** scheme[232] has therefore been developed in order to connect these stationary points to the AB set. This scheme first works by splitting the minima in the database into two sets, AB and $\overline{AB}$. A minimum i within AB is then selected, either explicitly by the user, or **CONNECTUNC** can identify the lowest energy minimum and choose that, and the n closest minima in $\overline{AB}$ are then identified. Connections are then attempted. This is an efficient method for expanding the AB set, using stationary points already in the database. In Chapter 6, a new feature added to **CONNECTUNC** shall be discussed.

The second issue, namely the consideration of multiple pathways to derive thermodynamic and kinetic properties, rather than from a single pathway, is dealt with by considering free energies. Unfortunately, a full consideration of free energies proved beyond the scope of the current work. However, as a derivation of these values would most likely be the next stage of this project, the principles behind the computation of free energies are given in Appendix C.

4.9 Disconnectivity Graphs

Due to the sheer number of degrees of freedom possible in a large biomolecular system, the number of possible minima and TSs, which scale exponentially with system size,[233] is immense. The representation of these data therefore becomes problematic. Textbook-style depictions of PESs, as in Fig. 1.24, are simply not suitable as these are merely 3D representations of landscapes, whereas the problems being studied in this work have much higher dimensionalities to consider. As it is not possible for human beings to visualise high-dimensional objects, an alternative strategy to represent these stationary points faithfully is therefore required.

One such strategy is to use disconnectivity graphs.[133;234;235] Such graphs, also known as 'trees', consist of nodes which correspond to superbasins and line endpoints which correspond to minima. The y-axis captures the energy of the system, whereas the x-axis is a free variable which can be altered by the user to give the clearest horizontal representation of the data. Superbasins arise by running analyses at fixed energy intervals, E_n, from which minima are classified into disjoint sets. Minima are considered as being in the same set (or superbasin) if a discrete path, consisting of intermediate minima and TSs, is possible between them for which no energy exceeds

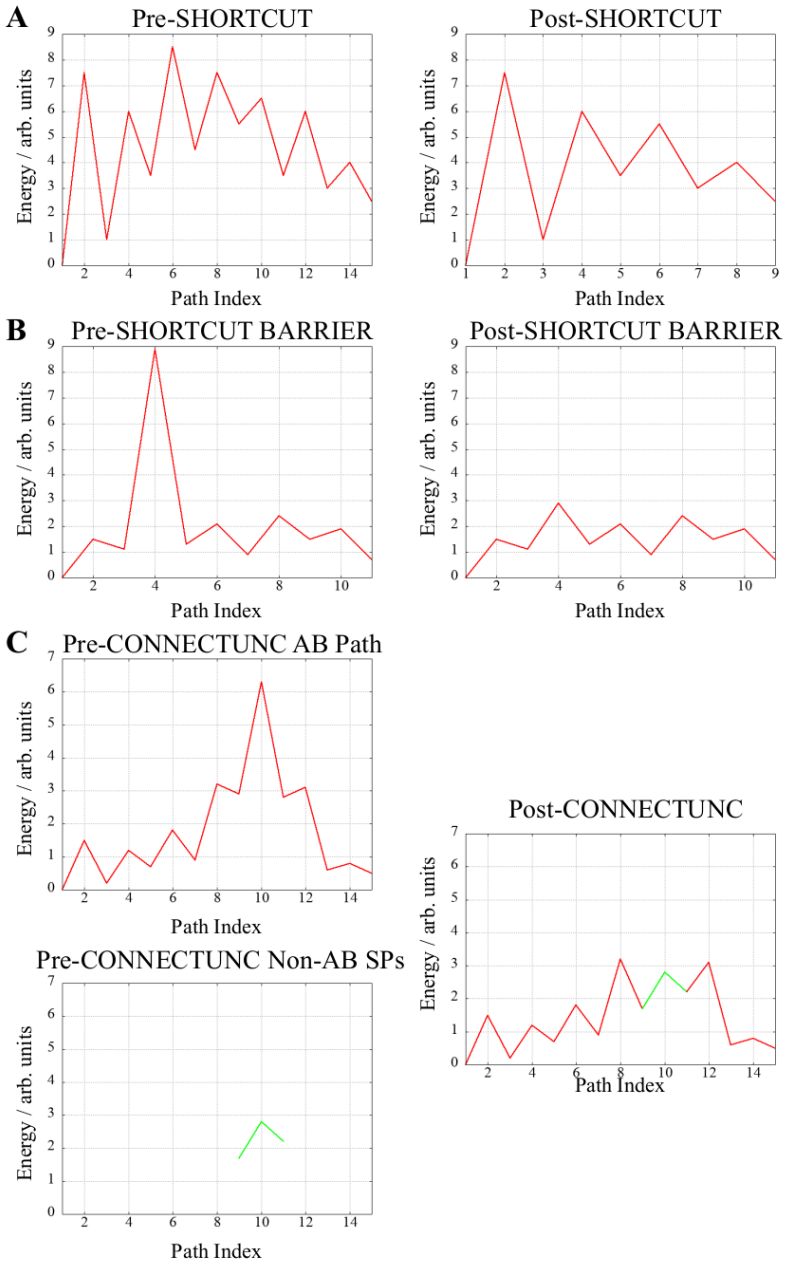


Figure 4.4: Illustration of refinement schemes. The red line in each chart shows a connected pathway between endpoints A (the locus at the far left) and B (the locus at the far right). Loci corresponding to an odd number in the integrated pathway are minima, and loci with even numbers are TSs. **A SHORTCUT**. A shorter path between minima 5 and 13 *via* a new TS was found, shortening the overall pathway. **B SHORTCUT BARRIER**. An alternative TS with a lower barrier was found between minima 3 and 5. **C CONNECTUNC**. A min-TS-min triple unconnected to the main AB set was connected and incorporated into the pathway, giving an alternative route with lower barriers.

E_n. Each superbasin is marked on the graph by a node at this corresponding energy, E_n. Therefore, minima in separate superbasins need to traverse this node (i.e. one or more stationary points in the pathway separating them is above E_n) in order to interchange. By sampling at various energies, E_n, the disconnectivity graph takes on a structure, which can provide a useful pictorial guide for the basic properties of the system being investigated. This construction is illustrated in Fig. 4.5.

4.10 Implementation of Wales Group Methods on GPUs

As emphasised in Section 1.7.2, previous computational work on the HemS system by Choy[21] and Shang[159] was severely limited by computational speed. This problem was largely due to these calculations necessarily having to be run on central processing units (CPUs). Since then, however, the basin-hopping, DNEB and HEF routines (including interfaces with AMBER) have been made compatible with graphics processing units (GPUs). GPUs require less resources to be allocated to data-caching and flow control compared to CPUs, freeing up more transistors for data-processing. This design increases efficiency for high intensity arithmetic calculations. Benchmarking has shown that Wales group methods typically run two orders of magnitude faster on GPUs than on CPUs.[211] This speed-up was exploited in the work described in this thesis: all calculations were run on GPUs.

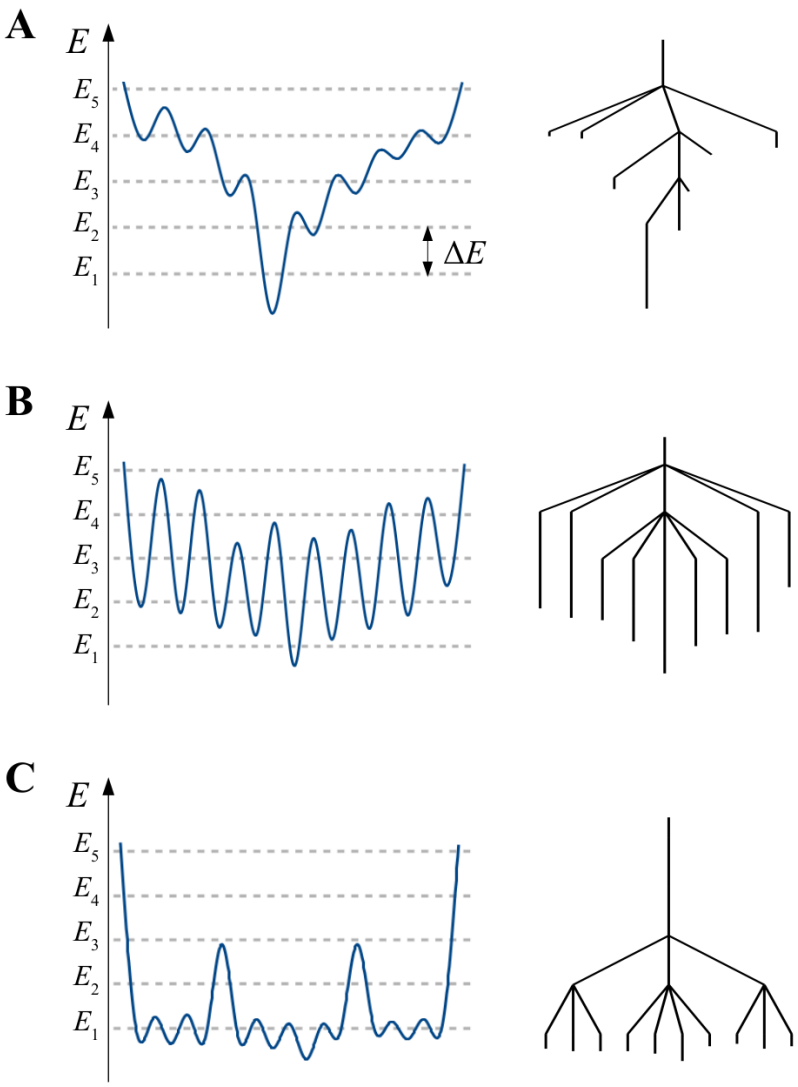


Figure 4.5: Pictorial correspondence between ordinary 2D-representations of a PES and disconnectivity graphs for three different energy landscapes, inspired by Wales.[133] E corresponds to energy, and E_n to energies at which a superbasin analysis was performed. **A** 'Palm tree' motif, which arises when the landscape corresponds to a steep funnel with low barriers. Protein folding and protein-ligand-based landscapes tend to have landscapes like this. **B** 'Weeping willow' motif, which arises when there is a shallow funnel with large barriers. **C** 'Banyan tree' motif, which arises from a 'rough' landscape and many competing low-energy minima.

Chapter 5

Further Experimental Insight into Haem Breakdown by HemS

5.1 Aims

There were some outstanding issues from the experimental work on HemS previously done in the Barker group. Sawyer had clearly demonstrated that addition of NADH leads to the breakdown of haem in the HemS pocket to produce a novel product.[42] The structure of this product, however, remained unknown, despite a large volume of biophysical data having been gathered on it. The mechanism of the reaction was also unknown. Indeed, it was not even certain if NADH was directly involved in haem breakdown. It was proposed, from the modelling done by Choy,[21] that NADH transferred a hydride directly over to haem, but this had not been proven experimentally. There were also unanswered questions concerning the low turnover rate of the reaction. Was it, for example, due to inhibition by either of the products, and, if so, which one? Furthermore, though Sawyer had noted that the reaction was capable of proceeding anaerobically, these were preliminary results. Therefore, the aims of further research into this breakdown reaction were:

1. To establish whether the reaction can indeed proceed anaerobically, and more generally to note any points of differentiation from the aerobic reaction.
2. To investigate possible product inhibition in greater detail.
3. To conduct deuterium labelling experiments to determine whether a hydride is transferred from NADH or not.
4. To determine the structure of the final haem breakdown product.

Outcomes from experiments inspired by these aims are described in the following sections.

5.2 Proof of Anaerobic Reaction

The NADH-dependent breakdown of haem in HemS was discovered under aerobic conditions,[42] and these are the conditions under which the reaction has typically been carried out in the laboratory.

Sawyer showed that the reaction was possible under anaerobic conditions too, with a HemS : Haem : NADH ratio of 5 µM : 20 µM : 2000 µ M, and that it proceeded at approximately the same rate as under aerobic conditions.[42] These were preliminary results, and so the experiment was repeated by the author using the procedure described in Section 3.7. This repeat confirmed the discovery by Sawyer that the reaction could proceed without oxygen.

This analysis was extended to investigate anaerobic reactivity under different haem : protein ratios. Sawyer had demonstrated in UV-Vis studies that increasing this ratio under aerobic conditions led to a greater likelihood of a further peak at ∼700 nm developing over time, in addition to the standard haem breakdown product peak at 591 nm. Increasing this ratio still further then led to the total shutdown of the reaction. From this result, it had been concluded that haem simply inhibited the reaction once its concentration was too high, and that this was most likely due to its propensity to dimerise at high concentrations.

Anaerobic studies with higher haem : protein ratios than 4:1 shed light on the identity of this ∼700 nm species. At a ratio of 20:1, it simply did not appear, as shown in Fig. 5.1. This figure shows two sets of difference spectra. Baselines consisted of SEC buffer, as well as the appropriate concentrations of HemS and haem. Haem was given time (ca. 30 minutes) to bind to HemS before the reaction was initiated by the addition of NADH.

The fact that this extra peak at ∼700 nm occurs in the aerobic case but not the anaerobic one would strongly suggest that there is a competing side-reaction that requires oxygen. The greater loss of haem in the the aerobic case, indicated by a greater loss of absorbance at the Soret band (408 nm), also suggests this conclusion. Furthermore, it is apparent that this side-reaction only becomes competitive when the haem concentration appreciably exceeds the HemS one.

Verdohaem and biliverdin under similar conditions are known to display peaks in the 700 nm range.[236] It was proposed, therefore, that this side-reaction was a coupled oxidation reaction, and was operating non-regiospecifically and without the

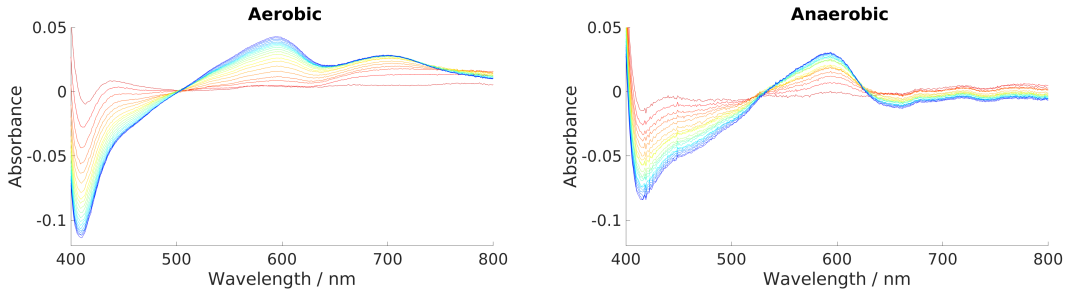


Figure 5.1: UV-Vis difference spectra charting the progress of NADH-dependent breakdown of haem in HemS over time. Experiments were run in SEC buffer, and the pH was set to 6.5. Spectra were recorded for 20 minutes in 1 minute increments, indicated by the colour scheme which runs from red (1 minute) to blue/purple (20 minutes). Exact stoichiometries were 1 µM HemS : 20 µM Haem : 2000 µM NADH. Left: Aerobic conditions. A peak at ∼700 nm develops over time. Right: Anaerobic conditions. No such peak develops.

need for any enzyme. This type of reaction was discussed in Section 1.4.6, where the Wilks group had pointed out that such reactions are often confused with genuine regiospecific haem oxygenase ones. NADH is a mild reductant, just like ascorbate, and is capable of reacting with oxygen to generate hydrogen peroxide, thus initiating non-enzymatic haem breakdown. As this is a non-catalytic process, this would also explain why it only becomes competitive when haem (and NADH) vastly exceed the protein concentration; otherwise, the catalytic haem breakdown process to form the 591 nm species dominates.

To test this theory, LCMS was performed on both the aerobic and anaerobic product mixtures, following a reaction in the ratio 1 µM HemS : 40 µM haem : 2000 µM NADH. As displayed in Fig. 5.2, the aerobic sample gives a large, clear m/z peak at 583.3 in addition to the signature haem breakdown product peaks at 569.3 and 462.2. This value corresponds to the mass of biliverdin. In the anaerobic LCMS spectrum, this 583.3 peak was negligible.

This result was tangible proof, therefore, that a coupled oxidation reaction was capable of competing as a side-reaction at high enough haem concentrations, but only under aerobic conditions. To minimise this side-reaction, most experiments following this discovery were therefore run anaerobically, particularly if high product purity was sought.

5.3 Deuterium Labelling to Determine Hydride Transfer

To better understand the role of NADH in the breakdown of haem, deuterium labelling experiments were conducted.

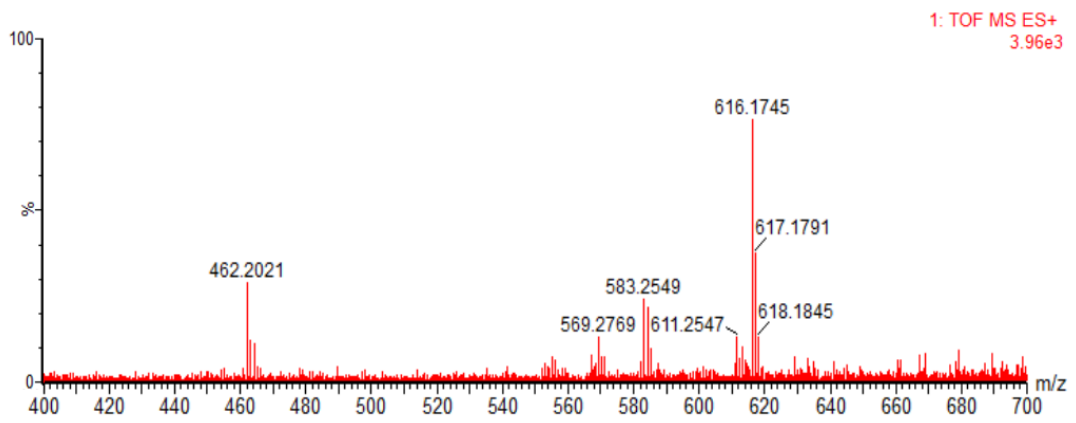


Figure 5.2: LCMS data for the aerobic reaction of *holo*-HemS with NADH, at a HemS : haem ratio of 1 : 40. The m/z peak at 583.3 corresponds to biliverdin.

NADH is known both as a hydride donor and as a reducing agent. As discussed in Sections 1.4.4, 1.4.5 and 1.4.6, NAD(P)H can constitute one part of a group of reactants which produces hydrogen peroxide, and thus initiate coupled oxidation, which leads to the breakdown of haem and formation of biliverdin. However, direct transfer of hydride from NADH to haem has never, as far as the author is aware, been demonstrated in HemS or any of its homologues.

One of the most common isotopes used in labelling studies is deuterium, which contains one neutron, whereas its isotope, hydrogen, has none. This makes deuterium almost twice as heavy as hydrogen, which leads to some different physical properties, such as a thicker viscosity and decreased quantum tunnelling efficiency, but its chemical properties remain largely the same. These properties make deuterium ideal for isotopic labelling experiments, especially when a particular hydrogen atom is of interest.

Choy's calculations suggested that the nicotinamide head of NADH points towards the haem once they are in close proximity. This nicotinamide head contains two hydride atoms, and their proximity to haem would suggest at least one of them is involved in haem breakdown. Tracking one or both of these hydrides would therefore reveal if such an involvement is indirect or direct; the latter scenario would be indicated by the labelled atom becoming part of the protoporphyrin structure.

Two enantiomers of deuterated NADH, (*R*)-NADD and (*S*)-NADD, were therefore synthesised using the methods described in Section 3.12. NMR studies confirmed that these were the correct (*R*)- and (*S*)-forms respectively, and that they were of high purity, as shown in Fig. 5.3.

Three reactions were performed in parallel. Conditions were all the same, except that NADH was used in the first reaction, (*R*)-NADD in the second, and (*S*)-NADD

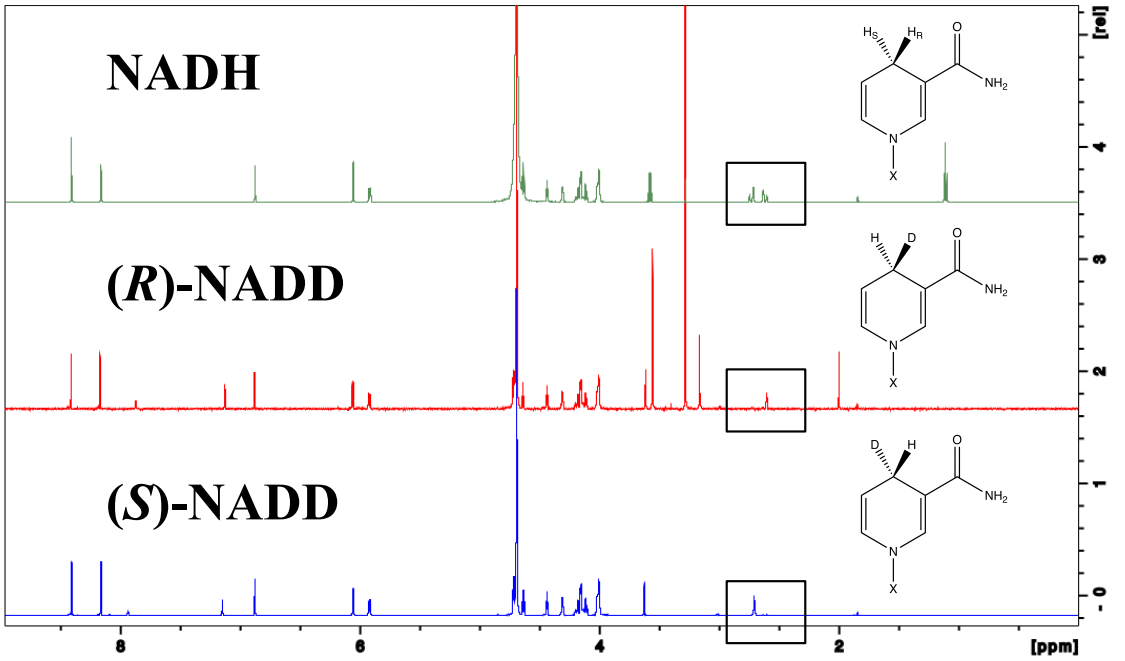


Figure 5.3: Overlay of NMR spectra taken for NADH, (*R*)-NADD and (*S*)-NADD. Black boxes highlight the region corresponding to the hydride signals. The X label corresponds to the rest of the NADH molecule.

in the third. Following the reactions, LCMS was performed on the product samples to determine whether deuterium transferred to the product fragments in any of the reactions. These spectra are shown in Fig. 5.4.

The signals around the *m/z* 613.3 region were difficult to deconvolute, but the 569.3 and 462.2 regions showed clear variations in the isotope patterns in the two deuterated cases with respect to the NADH sample. In both the deuterated samples, the signals for 570.3 and 463.2 were greater than for those at 569.3 and 462.2, which is a reversal of the NADH-based sample. This difference suggests that deuterium can be transferred over to the porphyrin ring and, because each of the experiments were conducted in non-deuterated solvents, that the deuterium must have originated from NADD. Therefore, haem breakdown does occur *via* deuteride (and therefore, by implication, hydride) transfer. The fact that both the (*R*)-NADD and (*S*)-NADD experiments showed significant deuterium transfer suggests that the reaction is not stereospecific, and that hydride can therefore be released from either side of the nicotinamide head. This suggestion is reinforced by the fact that in each case, a mixture of deuterated and non-deuterated products were found at the fragment regions centred round 569.3 and 462.2, rather than the deuterated or non-deuterated products being found exclusively.

These reactions had all been monitored using UV-Vis spectroscopy at a HemS : Haem : NADH/NADD ratio of 10 µM : 10 µM : 1000 µM, which provided some low-

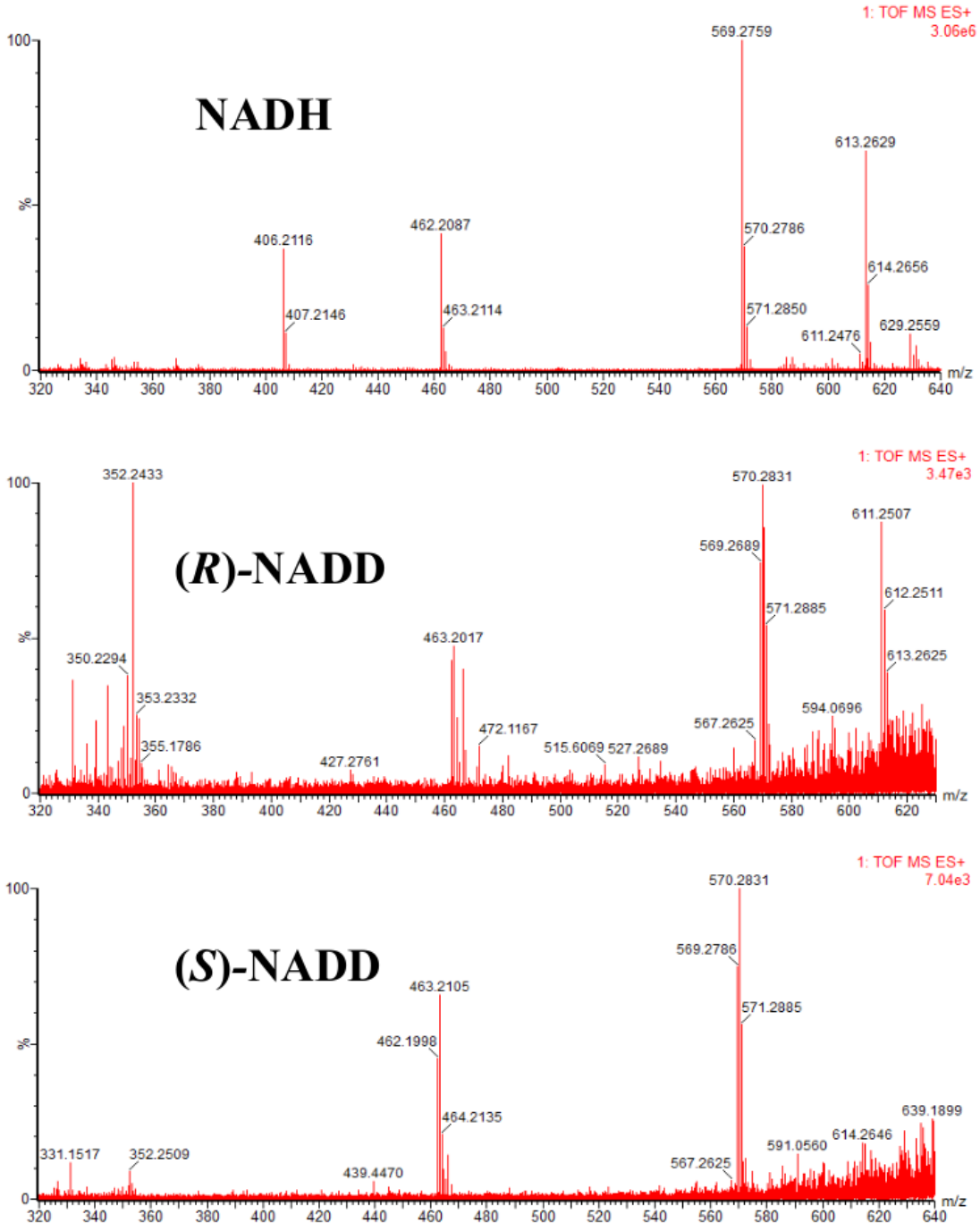


Figure 5.4: LCMS data for the product mixtures following reactions with NADH, (*R*)-NADD and (*S*)-NADD, respectively. Data for the molecular ion (*m/z* of 613.3 in the NADH-based sample) is inconclusive, but the peaks based at 569.3 and 462.2 have different isotope patterns in the (*R*)-NADD and (*S*)-NADD spectra compared to NADH.

resolution preliminary kinetic data, as shown in Fig. 5.5.

The first thing to note about these spectra is their different evolution at the 408 nm Soret peak. Rather than pre-equilibrate HemS and haem for these reactions (unlike the typical procedure used throughout this thesis), HemS and NADH/NADD were pre-equilibrated first, a baseline taken, and the reactions were then initiated by the addition of haem. This procedure was to prevent NADH/NADD distorting the spectra around the 400 nm region. NADH/NADD has a broad, strong absorbance at 340 nm, which extends above 400 nm at high concentrations. Typically, experiments were run from the same NADH stock, and so these absorbances could be easily accounted for. However, for these reactions, three different NADH/NADD stocks were required and so, in case of slight concentration variations when added, it was decided that a baseline should be taken to correct for this possibility before initiating the reaction. This procedure has the benefit that the Soret peak seen at 408 nm for these experiments is purely due to HemS-haem binding, with no interference from NADH/NADD absorbance.

These Soret peaks grew significantly up to absorbances of $\sim$0.8 in the two deuterated experiments, whereas in the non-deuterated case it did not even reach half that value. The conclusion reached was that in all three cases, haem was entering the pocket and coordinating to HemS quickly. However, in the case of NADH, conversion to the HBP was fast (thus preventing a build-up of HemS coordinated to haem, as indicated by the peak at 408 nm) whereas for (*R*)-NADD and (*S*)-NADD it was slow (where such a build-up did occur). This effect is reflected by the growth of absorbance at 591 nm, indicating the growth in concentration of the HBP. In the NADH case, this increase occurs quickly so that it has effectively reached a maximum at 0.25 after 40 minutes, as shown by the inset in Fig. 5.5. For (*R*)-NADD, it appears that the absorbance is beginning to level off after 160 minutes, again at an absorbance of around 0.25. (*S*)-NADD appears to be slower still: it is still rising, and has only reached an absorbance of 0.2, after 160 minutes. These relative rates between the two deuterated forms of NADD are reflected by the data at the Soret peak. For (*R*)-NADD, this peak has decreased almost to zero after 160 minutes, suggesting that the haem added is almost used up. However, after the same amount of time in the (*S*)-NADD case, the Soret peak is still far from zero, indicating that a significant amount of haem is still to be converted.

It is therefore clear that both deuterated forms of NADH slow down the reaction significantly, but one more so than the other; it is not entirely apparent why. The LCMS data in Fig. 5.4 unambiguously show that each form of NADD leads to a mixture of deuterated and non-deuterated product fragments. This mixture would

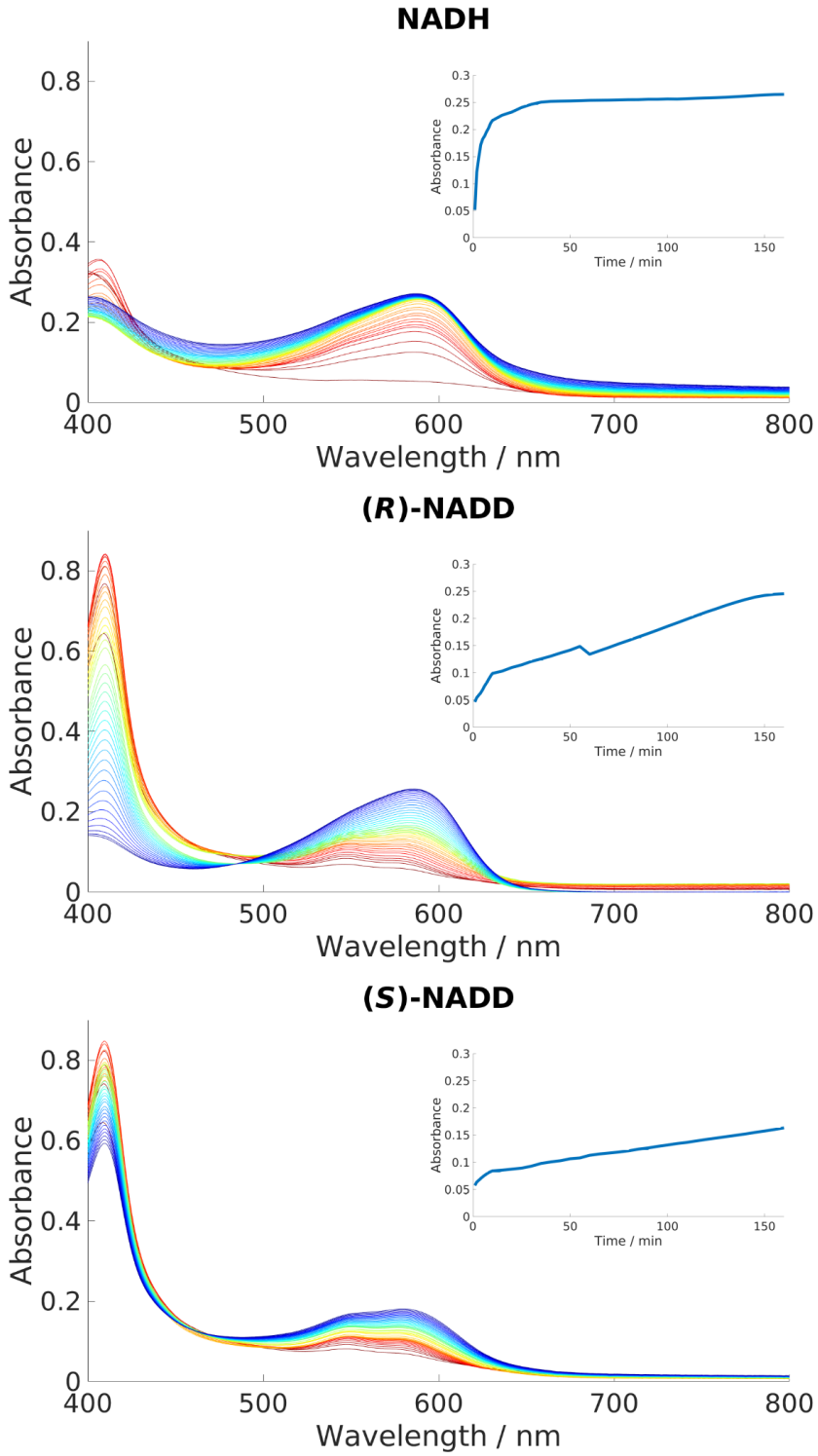


Figure 5.5: UV-Visible spectra comparing the rates of reaction between NADH, (*R*)-NADD and (*S*)-NADD. Reactants were mixed in the ratio 10 µM WT HemS : 10 µM Haem : 1000 µM NADH/NADD. Spectra were collected every minute for 10 minutes, then every 5 for 150 minutes, indicated by the colour change from red (1 minute) to blue/purple (160 minutes). Insets chart the changing absorbance at 591 nm over time.

suggest that both the hydride and the deuteride from each stereoisomer are capable of being transferred over to haem. Presumably, deuteride transfer is slower than hydride transfer. The kinetic isotope effect is a well-known phenomenon in chemistry, where a heavier isotope, courtesy of a lower vibrational frequency, requires more energy to reach a given transition state. This difference would explain why both forms of NADD are slower than NADH, but does not explain the discrepancy between the (*R*)- and (*S*)-NADD forms.

In the LCMS spectrum for NADH, the isotope patterns surrounding the 569.3 and 462.2 *m/z* peaks show smaller but significant amounts of species with one additional mass unit. This result is to be expected considering that isotopes, most commonly ^{13}C, occur naturally. Meanwhile, Fig. 5.4 shows that, when either (*R*)-NADD or (*S*)-NADD is used, these heavier 570.3 and 463.2 *m/z* fragments are more abundant than their 569.3 and 462.2 non-deuterated alternatives. Therefore, accounting for naturally occurring isotopes, it appears that the proportion of deuterated and non-deuterated fragments when either (*R*)-NADD or (*S*)-NADD is used is approximately 1:1. This ratio suggests there is no, or very little, preference for which hydride/deuteride transfers over to haem. This result, at first glance, appears to contradict the slower rate of reaction (*S*)-NADD displays compared to (*R*)-NADD. A possible explanation is that there are, broadly speaking, two different conformations NADH/NADD can adopt when transferring hydride/deuteride over to haem, and that these different conformations are each exclusive in terms of which hydride (either (*R*)- or (*S*)-) it can transfer. Furthermore, it is surmised that both conformations are almost equally likely but one of them results in a slower rate of transfer of the hydride to haem (perhaps because a greater distance has to be traversed between the molecules). Since hydrides transfer faster than deuterides, hydride transfer will govern the overall rate. In (*S*)-NADD, the deuteride is at the (*S*)-position and the hydride is therefore at the (*R*)-position. As the rate of reaction is so much slower when the hydride is in this position (as opposed to (*R*)-NADD, when the hydride is at the (*S*)-position) we conclude that the rate of transfer is more difficult from the (*R*)-position. In other words, transfer of a hydride from the (*S*)-position is faster. The deuterides do not enter into this consideration because of their relatively slow rate of transfer with respect to hydrides.

Because these different rates result in no noticeable difference between the proportion of the deuterated and non-deuterated forms in the LCMS spectra for either the (*R*)-NADD or (*S*)-NADD cases, it is further proposed that NADH/NADD must 'commit' to a conformation before transferring over one of its hydrides. In other words, even though a hydride at the (*R*)-position transfers more slowly to haem

than it does from the (*S*)-position, this process is still faster than the conformational changes that would be required for the nicotinamide head to switch round and present its (*S*)-hydride instead. Possible conformations of the nicotinamide head of NADH shall be discussed further in Chapter 7. Efforts were made to characterise these deuterated haem breakdown products but, as with the non-deuterated products, this task proved to be difficult.

5.4 Identification of an Intermediate

Sawyer had shown that the formation of the 591 nm species from UV-Vis spectroscopy was directly dependent on haem breakdown. However, it was unclear whether this process was a straightforward conversion without any intermediates.

It was noticed by the author after UV-Vis experiments under a variety of conditions that a further 'peak' seemed to arise, but then quickly disappear, at the edge of the spectra. It was realised that this peak must actually be in the near-IR (NIR) region, as only its shoulder could be observed due to the UV-Vis spectrophotometer cutoff at 800 nm. Often, this shoulder was only fleeting, particularly when the rate of formation of the 591 nm species was fast. This observation suggested that the shoulder represented an intermediate species on the pathway between haem and the HBP.

To investigate this problem further, a stopped-flow spectrometer was used, which has the double benefit of being able to access the NIR region, and give high resolution over short timescales.

Using an absorbance detector coupled with appropriate glass filters, a series of experiments charting the reaction of NADH with pre-incubated *holo*-HemS over a wide range of wavelengths (515–875 nm) were performed. The precise ratio used for these experiments was 5 µM HemS : 20 µM Haem : 2000 µM NADH, and the wavelengths were measured in increments of 40 nm. Results are shown in Fig. 5.6.

These experiments showed that the absorbances at the selected wavelengths in the 515–635 nm range grew steadily and then began to level off over the 1000 s time period measured. The absorbance was greatest at 595 nm, which is consistent with the peak for the HBP identified in other UV-Vis studies. The selected wavelengths between 675–875 nm, however, revealed that this change was not simply a conversion from haem to the HBP. These spectra showed a short burst in absorbance over the first 20 s, followed by a steadier decrease. This decreasing absorbance then began to level off around the 1000 s mark. The wavelength where this short-timescale 'burst' was shown to give the greatest absorbance was at 795 nm. As other UV-

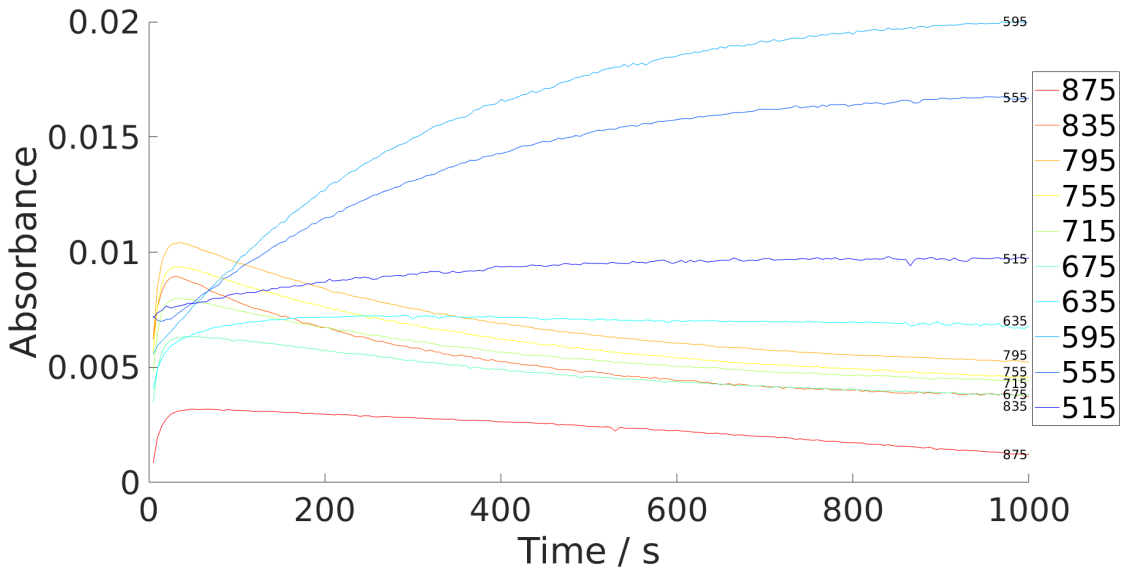


Figure 5.6: Stopped-flow spectra showing 5 µM HemS : 20 µM Haem : 2000 µM NADH reaction progress at selected wavelengths. 200 time points were recorded over 1000 s. Spectra from 515–635 nm show a growth in absorbance over time, which begins to level off after ∼1000 s. This absorbance growth is greatest at 595 nm, consistent with the HBP peak seen in other UV-Vis studies. The behaviour is different from 675–875 nm. At these wavelengths, a short burst in absorbance (0–20 s), then a longer decrease which eventually levels off (20–1000 s), is observed. This result suggests intermediate formation, which then converts to the HBP. The absorbance growth is greatest at 795 nm for this intermediate within the wavelengths sampled.

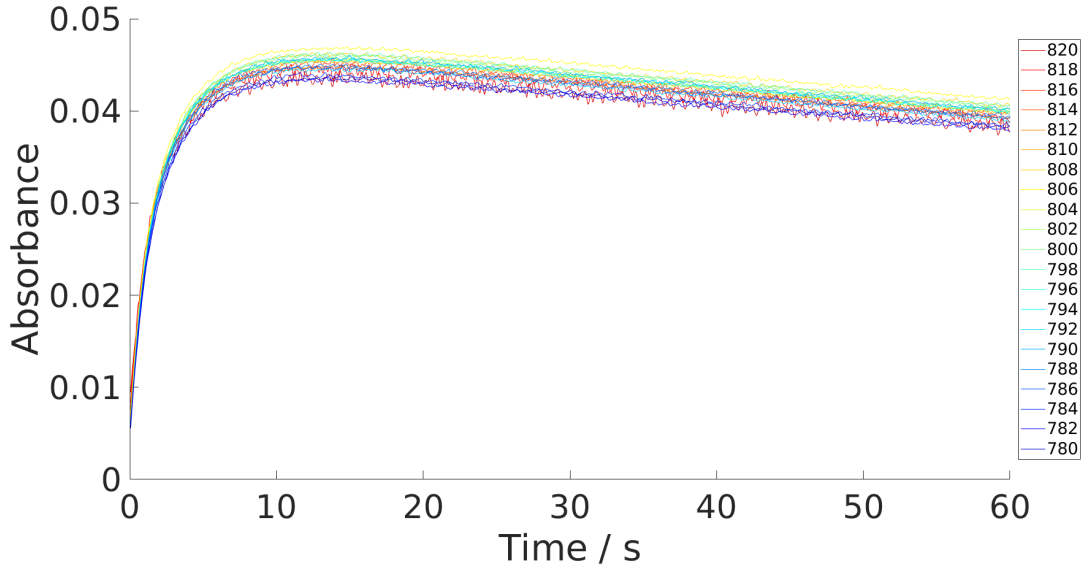


Figure 5.7: Stopped-flow spectra showing 8 µM HemS : 8 µM Haem : 2000 µM NADH reaction progress at selected wavelengths and high resolution. 1000 time point were recorded over 60 s. There is a greatest growth in absorbance at 806 nm, indicating this is the peak for the intermediate species.

Vis studies using the Cary spectrophotometers clearly showed that this peak must have a higher wavelength than 800 nm, it must therefore be located somewhere between 800 nm and the next increment tested by stopped-flow, 835 nm. These results suggest that an intermediate, with a signature peak at approximately 800 nm, is therefore formed first from haem breakdown, and then converts to the HBP (which has its signature peak at 591 nm). Having an intermediate with a peak at such a long wavelength could suggest that initial haem breakdown involves a disruption of aromaticity and/or change in the iron oxidation state. Given the porphyrin is eventually cleaved open to produce the HBP, it may be the case that some sort of sigmatropic rearrangement then occurs.

Further stopped-flow experiments were then conducted to determine the exact wavelength of this intermediate peak using a shorter range of wavelengths (780–820 nm) with shorter increments (2 nm). The location for this peak turned out to be at 806 nm, as shown in Fig. 5.7.

Due to the fleeting nature of this intermediate, it was recognised that it would be difficult to extract from the protein pocket, isolate and characterise. Focus, instead, turned back to the final HBP as well as the precise role of NADH in the breakdown of haem. This intermediate shall be discussed in more detail in Chapter 9.

5.5 Attempting to Determine the Haem Breakdown Product Structure

The exact structure of the HBP was still unknown, despite advances in understanding the reaction mechanism. Attempts were therefore made to determine this structure. This task was difficult, though, as the yield of product was consistently low. One reason for this low yield is product inhibition; it is suspected that, in its biological context, HemS has an accompanying protein which it can transfer the HBP to. Another reason is that the reaction effectively shuts down at high haem concentrations.

5.5.1 NMR

Sawyer[42] and Choy[21] had both reported difficulties in preparing HBP samples for NMR spectroscopy. They each tried different approaches: Sawyer tried a butanone extraction, and Choy used a C18 solid phase extraction column.

Sawyer had shown that it was possible to extract the HBP from the protein on a large scale using 2-butanone. However, this procedure resulted in an immediate colour change of the overall solution from purple to light yellow, suggesting that iron had been removed from the product. This iron seemed to be important to the structural integrity of the multipyrrole, as a subsequent NMR analysis by Sawyer suggested that multiple species were present due to the sheer number of different resonances observed. It is interesting that iron was removed, as evidence from other experiments – albeit in the gas phase – would suggest that it binds quite strongly to the product multipyrrole. All three of the main LCMS fragments (613.3, 569.3 and 462.2 m/z) are known from accurate mass analysis to still have iron bound. Furthermore, samples of the product mixture have been seen by the author to retain their purple colour (and so presumably the iron ion is still bound to the protein) when kept at 4 °C over a matter of months.

An attempt was therefore made to repeat Sawyer's experiment, but with an effort to keep iron inside the HBP. Butanol was used instead of butanone since it was reckoned that this solvent would give a better separation from the aqueous layer. The reaction was performed under anaerobic conditions, as described in Section 3.8, and then kept under vacuum for the extraction process and subsequent removal of butanol by rotary evaporation. This procedure gave a deep purple residue, suggesting that iron was still coordinated to the HBP.

This sample was redissolved in d^6-DMSO and submitted for ^{1}H NMR analysis. It was hoped that, since the structure of the HBP seemed to have been kept intact,

that the spectrum would be cleaner than that achieved by Sawyer. This point was rendered moot, however, because of the paramagnetism of the iron still left in the sample, which resulted in extreme peak broadening in the NMR spectrum, to such an extent that none could be resolved.

Though this problem meant that the structure of the HBP still could not be resolved by NMR, it at least revealed some details concerning the oxidation state of iron. Fe^{3+} is necessarily paramagnetic due to its odd number of electrons. Fe^{2+}, meanwhile, can be diamagnetic or paramagnetic depending on the splitting energies of the ligands it is coordinated to. In the samples submitted for NMR, therefore, it is likely that iron is either in its Fe^{3+} or Fe^{2+} high-spin state.

5.5.2 Crystallisation

NMR analysis therefore presented a paradox. If extraction successfully retained the iron ion, paramagnetism caused peak broadening to such an extent that the spectra became uninterpretable. On the other hand, if iron was released during extraction, it seemed that the main HBP disintegrated, giving a mixture of compounds that could not be deconvoluted. Alternative methods of characterisation were therefore sought. One such method was X-ray crystallography.

Despite the fact that it had been observed to be capable of retaining its purple colour (implying stability) at low temperatures over a long period of time, it was not clear how crystallisation of the isolated HBP should be approached. Even though crystallisation of haem (in its haemin form at least) has been known for a long time, with Teichmann crystals discovered in 1853,[237] it was thought unlikely that the HBP would crystallise under the same conditions, given that it is most likely cleaved at one of the *meso*-carbons, resulting in a more flexible structure that is perhaps more of a linear than a cyclic tetrapyrrole. It was suspected that, even if crystallisation was successful, it would only occur after further breakdown of the product.

Therefore, crystallisation with the HBP still bound inside the protein was attempted instead. As detailed in Section 1.4.1, suitable conditions have been developed for HemS to crystallise in both its *apo*- and *holo*-forms. The conditions from Schneider & Paoli[163] were largely followed with minor adaptations, which are detailed in Section 3.11.1. Samples of *holo*-HemS (i.e. HemS with unreacted haem bound) were crystallised as a standard in parallel with samples of HemS containing the HBP.

Before getting to this stage, a method of concentrating the HBP-containing HemS samples and exchanging into the appropriate crystallisation buffer was required. It was found that concentration could be achieved using centrifugal filters,

with the HBP remaining in the protein pocket.

This approach solved an important paradox which could have prevented crystallisation altogether. As noted in Section 1.5.2 (particarly in Fig. 1.21), the NADH-dependent breakdown of haem does not occur at high haem concentrations (above ∼20 µM it begins to shut down). Meanwhile, crystallisation of HemS requires concentrated samples (Schneider & Paoli quote 30 mg mL^{-1},[163] which works out as approximately 750 µM). If as many HemS molecules are to contain the HBP as possible, then the starting concentration of haem would need to be approximately 750 µM as well if it were not for this post-reaction concentration step. In short, the ability to concentrate the HemS and HBP samples *via* a method unlikely to significantly disrupt the structure of the protein was required in order for crystallisation to be possible, and those conditions were fulfilled by the ultrafiltration step.

Crystal growth proved to be inconsistent between samples. Those stored at 4 °C tended to grow better than those stored at 25 °C. For those samples that did show growth, the crystals tended to be small. Most were long, thin and blue, and none appeared to exceed 0.16×0.04 mm.

With the help of Dr Paul Brear, Facility Manager of the Crystallographic X-ray Facility at the Department of Biochemistry, University of Cambridge, the most suitable crystals were treated with cryoprotectant and sent to Diamond Light Source, as described in Section 3.11.2. The resolution of the protein itself in these samples typically proved to be of good quality (at around 1.67 Å). In many cases, there was clearly electron density in the pocket corresponding to the region where haem binds. This species was therefore taken to be the HBP. However, the resolution of this electron density was too low to determine a structure exactly. One side appeared more 'open' than the others, suggesting that the porphyrin ring had been cleaved, as expected. There was no obvious electron density corresponding to iron, suggesting that it had been removed, although it was unknown whether this loss occurred during crystallisation or cryoprotection. Without iron to bind to, this result would suggest the multipyrrole had a degree of conformational flexibility, thus giving rise to a degree of disorder and therefore potentially explaining its low resolution within the protein pocket. Alternatively, these densities in the protein pocket could be showing multiple stages of the degradation of the HBP at once; inconsistencies between unit cells would explain the low resolution. A representation of the electron density of the HBP from one of the crystals is given in Fig. 5.8.

The unstructured loop of HemS discussed in detail in Sections 1.4.1 and 1.7.1 was also of too low a resolution to accurately resolve the structure in any of the wild type (WT) crystal structures considered. However, crystallisation was also performed on

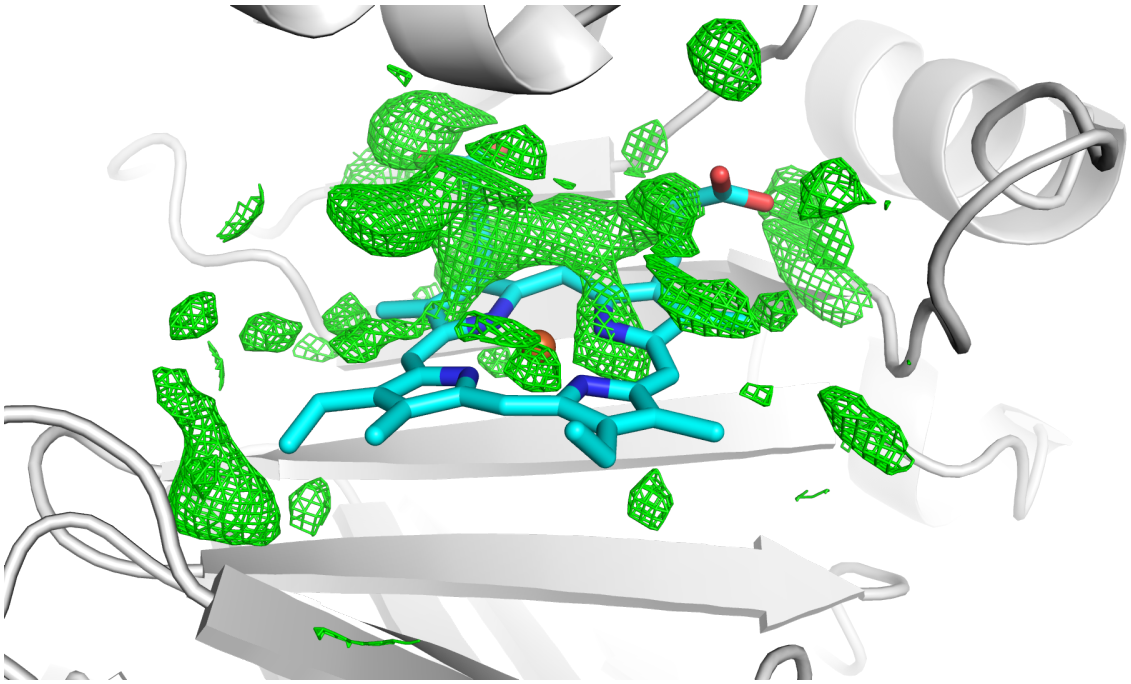


Figure 5.8: Electron density from a crystal of WT HemS after haem breakdown. Unassigned electron density is shown as a green mesh. The haem molecule, in cyan, is overlaid from a crystal of *holo*-HemS to provide a reference point. The shape of the unassigned electron density suggests the tetrapyrrole is still intact but has been cleaved. The absence of a region of large electron density, indicative of iron, suggests that iron has been extracted.

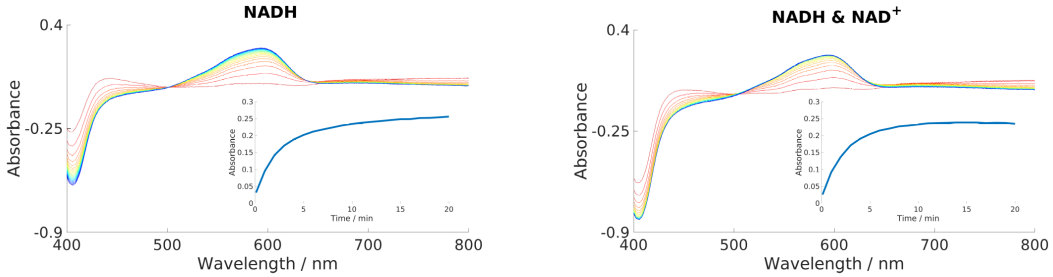


Figure 5.9: UV-Visible difference spectra comparing a standard 5 µM WT HemS : 20 µM Haem : 2000 µM NADH reaction with one which also has 2000 µM NAD^+ present. Spectra were collected every minute for 20 minutes, indicated by the colour change from red (1 minute) to blue/purple (20 minutes). Insets chart the changing absorbance at 591 nm over time. Results show that the presence of NAD^+ has little effect on the reaction.

the double mutant, F104AF199A, where the resolution was good enough to construct the loop. This experiment shall be discussed further in Section 9.6.

5.6 Product Inhibition and NAD^+

It was unclear why the turnover for the NADH-dependent haem breakdown reaction is so low. Experiments by the author exhibited a diminishing rate of conversion upon further haem addition, even when it was assumed that all of the original stock of haem had been converted. This result suggested that one or both of the reaction products, the HBP and/or NAD^+, were still occupying the pocket and involved in product inhibition.

Therefore, two reactions were run in parallel and tracked by UV-Vis spectroscopy. In both cases, 5 µM WT HemS and 20 µM haem were pre-incubated for 30 minutes, but in one of them 2000 µM NAD^+ was also present. Baselines were then taken. NADH was then added to 2000 µM in each, and the reactions monitored for 20 minutes. The results are shown in Fig. 5.9.

It had been hypothesised that pre-soaking with NAD^+ would result in NADH being blocked from accessing the protein pocket. However, the results in Fig. 5.9 show little difference between the reactions run with and without NAD^+. The peak at 591 nm grows to an absorbance of ∼0.25 in each case over a similar time period. After 20 minutes, the reaction without NAD^+ still appears to be rising slightly whereas the reaction with NAD^+ appears to have reached a maximum and levelled off. However, this difference is only slight, so no meaningful conclusion can be drawn from it. A slight discrepancy is in the absolute values of the Soret band at

408 nm between the two experiments. In the case without NAD^+, the absorbance starts at approximately −0.25, and drops to roughly −0.60 over 20 minutes. In the case with NAD^+, meanwhile, the absorbance starts at approximately −0.55 and decreases to roughly −0.80 over 20 minutes. The greater decrease in absorbance in the non-NAD^+ case suggests that haem is being used up faster in a given amount of time. This difference is perhaps because NADH does not have to displace any NAD^+ molecules from the protein pocket, so can access haem to break it down more readily. The reason why the absolute absorbance values at 408 nm are higher in the case without NAD^+ is not clear, but is perhaps due to an inadequate amount of time being allowed for HemS-haem equilibration before the baseline was taken.

These experiments show that NAD^+ does not inhibit the haem breakdown reaction to any significant degree. The NAD^+/NADH ratio is therefore presumably not of great importance to the reaction. If product inhibition is a factor limiting the turnover, it is perhaps due to the HBP or some sort of [HBP-NAD^+] complex instead.

5.7 Discussion and Summary

Greater insight had been attained from these further investigations into the NADH-dependent haem breakdown reaction in HemS.

Firstly, the possibility of the reaction taking place anaerobically had been confirmed. Furthermore, performing this reaction anaerobically also demonstrated minimal side-product formation from a non-regiospecific, non-enzymatic coupled oxidation process, unlike under aerobic conditions, where this process competes. This side-reaction had not been noticed in previous studies but it can produce a significant quantity of biliverdin, particularly when the haem concentration is high. These facts taken together – that NADH-dependent haem breakdown can occur aerobically or anaerobically, but that competition only occurs aerobically – would suggest that HemS is a haem breakdown enzyme specifically designed to work under anaerobic conditions. As discussed in Section 1.3.2.1, the bacterium that produces HemS, *Yersinia enterocolitica*, generally attacks the gastrointestinal tract of its host organism. It then typically travels to the terminal ileum to replicate.[238] This is a region of low oxygen content[239] and so *Y. enterocolitica*, an aerobic but facultative (i.e. discretionary) anaerobic bacterium, has to adapt to its new conditions. Whatever the conditions, the bacterium still needs to obtain iron. The *hem* operon has been demonstrated to be able to obtain such iron in the form of haem, with HemS purportedly the end user. Either the *hem* itself adapts to the oxygen-limiting con-

ditions, or it may be a backup iron-uptake system, which only switches on when oxygen levels are low. Whatever the case, it provides an anaerobic method of haem uptake and, now that HemS has been demonstrated to be capable of breaking down haem without oxygen, utilisation.

The reaction mechanism was also investigated in more detail. It is now clear, from stopped-flow experiments, that an intermediate with a peak at 806 nm is produced before the final HBP, which has its signature at 591 nm. This intermediate is transient, typically lasting for only 20 s under the conditions investigated, and its loss corresponds to the production of the final HBP. It was unclear what structure this intermediate could have but the long wavelength associated with it implied that the conjugation of the porphyrin had been broken and/or the iron oxidation state had changed.

Further insight into the reaction mechanism was gained from deuterium labelling experiments, which proved that NADH was acting as a direct hydride donor to haem. These experiments further showed that the reaction was not stereospecific in terms of which face of the nicotinamide head the hydride was delivered from. Indeed, it appeared that there was an approximately 50 : 50 chance of delivery from either face. Despite this ratio, the rate of transfer of the hydride was clearly faster from the (*S*)-position than it was from the (*R*)-position. At first sight, this result seemed to be a contradiction: if transfer from one of the faces was markedly faster than transfer from the other, then how did the deuterated : non-deuterated product ratio end up near to 1 : 1 no matter whether (*R*)-NADD or (*S*)-NADD was used? The current hypothesis is that NADH unfolds in the pocket in such a manner that either its (*S*)- or (*R*)-face is presented towards haem, and that the energy that would be required to change this conformation so that the alternative hydride is presented exceeds the energy required for the hydride to transfer, even if it is the less favoured face that is directed towards haem. Thus, the more difficult hydride transfer is still achieved, albeit at a slower rate.

Further attempts were also made to determine the structure of the HBP. A method of HBP extraction from the protein using butanol was found to successfully retain the integrity of the HBP. However, ^{1}H NMR spectroscopy on these samples was unsuccessful due to extreme peak broadening, which made the spectra difficult to interpret. This problem was most likely due to paramagnetic effects caused by the retention of the iron ion inside the HBP. One conclusion from this NMR analysis was possible though: the very fact that paramagnetism was occurring suggested that this iron ion was either in a Fe^{3+} or a high-spin Fe^{2+} state.

Crystallisation and X-ray crystallography of the HBP inside HemS was also at-

tempted. To achieve a low enough haem concentration for the reaction itself and then arrive at high enough concentrations for crystallisation, samples were concentrated by ultrafiltration. Crystal growth between samples was inconsistent. Nevertheless, the resolution of the protein molecules inside the crystals selected for analysis at Diamond proved to be good. The resolution of the HBP, on the other hand, was never high enough for an accurate structural determination. This could have been due to a number of factors. The HBP may have been degraded by the centrifugation step required to concentrate the samples, although this seems unlikely given the samples each retained their purple colour. It may also have been degraded during the crystallisation process itself as the crystallants used, such as PEG 400, may have reacted with it to cause further breakdown. Furthermore, the cryoprotectants added by Dr Paul Brear may have caused breakdown. Despite the poor resolution, it was clear from all of the crystals that the multipyrrole was not a full ring, with one of the sides longer than the others, and with density missing from its midpoint. This structure strongly suggested that it had been cleaved at one of the *meso*-carbon positions.

Due to the low turnover of this reaction, the possibility of product inhibition (specifically from NAD^+) was investigated. The results showed that NAD^+ caused no such inhibition. Either product inhibition is caused by the HBP or by a [HBP-NAD^+] complex, or the low turnover rate of this reaction is not due to the products at all.

Further experimental results are discussed in Chapter 9.

Chapter 6

Computational Method Development

A glossary and some diagrams are provided in Appendix D to clarify the terms and concepts discussed in this method development section. It is therefore recommended to refer to this resource whilst reading this chapter.

6.1 Aims

Due to the weak and transient nature of NADH-binding inside the HemS pocket, the types of laboratory-based experiments that could be used to further elucidate NADH-residue interactions were limited. The overall aim of computational research into the [WT HemS + Haem + NADH] system[a] was therefore to find pathways for NADH approaching haem within the pocket (and thereby build up a thermodynamic and kinetic profile). As outlined in the Methods, this analysis is most easily achieved if a single pathway using discrete path sampling (DPS) is searched for first. Furthermore, such an approach requires good starting points. For a pathway as long and involved as the one under study, it was decided it should be split into smaller segments, with NADH at various stages of progression along the pocket. These starting structures (five in all) were sampled extensively before attempts were made to connect all of them together using DPS. This approach resulted in five large 'sub-databases' of stationary points, and so the question became one of how to connect them to uncover a complete pathway.

It was mainly due to the implementation of Wales group software to GPUs since

[a]Please note that for the rest of this chapter, whenever the WT HemS system (or any of its mutants or homologues) are referred to, this implies inclusion of haem and NADH unless explicitly stated otherwise.

the studies of Choy[21] and Shang[159] that far larger databases of stationary points describing the WT HemS system could be generated. However, the system under study contains a large number of atoms, as well as multiple biomolecules, implying that many intermolecular forces in addition to the intramolecular forces of the protein must be included. Coupled with the fact that the conformational gaps between the sub-databases were large (even despite the attempt to split the pathway into more manageable segments), it was clear that efficient selection methods would be required to ensure sensible connection attempts were made. Therefore, the **CONNECTUNC** keyword (which was explained in Section 4.8) was expanded to reflect this particular issue of selecting efficiently for large sub-database connection attempts.

Despite this development, it still took a long time to find the pathway for the WT HemS system. It was the intention of the author to study pathways from selected mutants and homologues too, but it was apparent that building these other databases from scratch using standard methods and the computational resources available would take too long. Therefore, a method was developed to take the stationary points in the WT HemS database, use them as templates, apply mutations to the relevant residues, reoptimise, and then fill in any gaps in the pathway where reoptimisation failed. This strategy led to **CHECKSPMUTATE**, a subroutine now implemented in the Wales group code.

Therefore, method development projects pertinent to this thesis were:

1. To develop an efficient procedure for selecting minima far apart in conformational space and part of different sub-databases for connection attempts.

2. To develop a routine that can 'mutate' selected residues for all of the stationary points on a given pathway, and reoptimise, reconnect, and refine them to therefore give a pathway for a new, mutated system.

Details of these methods, and how well they perform, are provided in the following sections.

6.2 Expanding CONNECTUNC

As detailed in Section 1.7.2, Shang had identified, using **GMIN**, three possible binding modes for NADH in the HemS pocket in addition to the one that had been identified by Choy using Relibase$^+$.[21;159] Further studies by the author identified another site.

Therefore, five possible binding sites for NADH in the *holo*-protein pocket had been identified, all at different stages of unfolding and proximity to haem. These sites are numbered from 1 to 5, with site 1 representing the case where NADH is

furthest from haem and at its most folded, and site 5 representing the case where NADH is closest to haem and is most unfolded. It was recognised that connecting these sites would therefore give a pathway showing NADH unfolding and travelling along the pocket towards haem.

Despite there being three intermediate stages of NADH progression along this pathway to start calculations with, rather than just the start and end points, it was realised that these structures would still be difficult to connect *via* DPS. The sheer size of the system (5501 atoms in total) plus the requirement to model intermolecular as well as intramolecular interactions, promised that such connection attempts would be computationally expensive. Furthermore, these connections between binding sites often looked as if they would require significant conformational changes, both from the protein residues associated with the pocket and from the NADH molecule itself.

To better understand the local energy landscapes around these different sites, basin-hopping runs using short step sizes were first run for each. This approach allowed for many minima close in conformational space to the site being studied to be discovered. These minima were then connected using DPS, as described in Sections 4.6 and 4.7. Minima derived from the same site were far easier to connect to each other than minima derived from different binding sites, because they were far closer in conformational space.

This approach results in five 'sub-databases', each comprising the original binding site minimum structure discovered by Choy, Shang or the author, and a large number of minima and connecting TSs close in conformational space. It can therefore be said that this conformational space around each site has been well sampled, increasing the possibility that there are minima in these sites that can connect with less difficulty to minima in other sites. Because all of the minima in the sub-databases are connected to one another, just one successful connection between two minima in two different sub-databases would mean that all of the minima in these two sub-databases become connected. To save computational resources, an efficient scheme to choose which two minima to connect based on the likelihood of success was therefore developed. The criterion was Euclidean distance, as it was assumed that minima closer together in conformational space would be more likely to successfully connect in a given number of cycles. In other words, the shorter the Euclidean distance between a pair of minima, the more likely they would be selected for a connection attempt.

The **CONNECTUNC** keyword in the **PATHSAMPLE** program (see a discussion of **CONNECTUNC** in Section 4.8) was therefore extended to achieve this selection. It was first ensured that the database being studied contained only the two sub-databases

to be connected.[b] **CONNECTUNC** then splits the database into the AB set (one of the sub-databases) and the $\overline{AB}$ set (the other sub-database).

If the **LOWEST** option is applied to **CONNECTUNC**, the lowest energy minimum from AB is identified, and the n closest minima in $\overline{AB}$ are chosen to try to connect to this AB minimum. For this work, **LOWEST** was adapted to give another option, designated **LOWESTTEST**. Instead of submitting a job after the closest minima in $\overline{AB}$ to the lowest minimum in AB were identified, **LOWESTTEST** stores this information and then cycles to the next-lowest minimum in AB. The closest minima of $\overline{AB}$ to this minimum are then identified too, and this information is also stored. This continues until the closest minima from $\overline{AB}$ have been identified for all of the minima from AB. In other words, the Euclidean distance for each and every combination of minima between AB and $\overline{AB}$ are identified. A connection is then attempted between the two minima with the shortest distance between them. The fact that it is the shortest distance to connect means that, in general, it should be the most likely to succeed.

As sites (and therefore sub-databases) 1 through to 5 approximately charted the progression of NADH through the pocket (with 1 showing NADH just as it enters the pocket, and 5 showing NADH in close proximity to haem) it was therefore clear which sub-databases to try to connect. It would make far more sense to try to connect sub-databases 2 and 3 together as well as 3 and 4, rather than trying to connect 2 and 4 directly, as the Euclidean distance between any two minima in the latter case is likely to be very high. However, the author could envisage a scenario (totally removed from the WT HemS system) where far more than five sub-databases are brought together for connection attempts, and there would be no obvious relationship between them. Therefore, the **CONNECTUNC LOWESTTEST** algorithm was generalised still further, so that, as well as choosing which minima between two sub-databases are most suitable to select for connection attempts, the best sub-databases to select for connection attempts in the first place are considered as well.

6.3 CHECKSPMUTATE

Despite the use of **CONNECTUNC LOWESTTEST** to connect the five sub-databases, plus the benefit of GPU-acceleration, it still took over 12 months to find a fully connected pathway showing NADH unfolding and moving from a position at the

[b] If there were further minima and/or TSs in the database not belonging to either of the sub-databases of interest then these were removed, after being copied and saved elsewhere, to be added back in later. There are subroutines, such as **RETAINSP** and **REMOVESP**, implemented in **PATHSAMPLE** which can do this.

edge of the pocket to one where its nicotinamide head is in close contact with haem. When we consider the computational efforts expended by Choy and Shang to generate the original binding sites in the first place, whether from Relibase$^+$ or large step size basin-hopping, it is clear that building such a pathway from scratch constitutes an 18-24 month project.

Analysis of this pathway had suggested seven mutants of HemS to study (to be discussed in the next chapter), and the literature indicated at least four homologues (HmuS, ChuS, ShuS and PhuS) of interest. Pathways were therefore sought for these other ten systems (PhuS was discontinued after failing the reoptimisation process, as shall be discussed in Section 6.3.3) to determine whether NADH interacts with the protein in a similar manner in each. Clearly, this task would not have been possible if the methods employed to find the WT HemS pathway had been followed, even if it was possible to run some of the systems in parallel.

Therefore, it was decided that these other systems for the mutants and homologues should not be derived from scratch. Instead, the WT HemS system was used as a 'template' for the others. Specifically, all of the stationary points (minima and TSs) on the Dijkstra shortest pathway for WT HemS were reoptimised after having applied the necessary mutations.

The question, therefore, was how to apply these mutations before reoptimisation. Furthermore, how were the homologues to be handled? Not only did they require changes to multiple residues throughout the sequence, but they also required residues to be added/removed from certain points in the sequences.

6.3.1 Test Tripeptide System

In developing the **CHECKSPMUTATE** subroutine, a small tripeptide, RQQ, was used as a test system, which also provides a useful conceptual framework to explain the operation of the algorithm.

The simplest mutation to model is a single-point mutation which reduces the overall number of atoms. Increasing the size of the residue is more likely to lead to reoptimisation issues as it increases the steric bulk of the molecule under consideration. Therefore, the first mutation attempted was a change of the central glutamine to a glycine, i.e. RQQ $\rightarrow$ RGQ.

The first thing to consider was therefore using the LEaP program in AMBER to generate a new topology file for RGQ (see Section 4.4 for further details on topology file generation). This is a simple, well-documented process which provides the necessary information on atomic type, bond strengths/angles etc. required for any system to run on AMBER.

However, generating the new set of input coordinates for the RGQ peptide presents more difficulties. As the overarching intention of the 'template-based' approach was to preserve the conformations of the reoptimised stationary points as far as possible, a method of ensuring the mutated residues were as close conformationally to what they had been changed from was required. Fortunately, this problem had been addressed by Röder when he developed the Mutational Basin-Hopping (MBH) subroutine for the Wales group.[134]

MBH considers both the original and mutated residues as individual graphs, each consisting of nodes (which represent atoms) and edges (which represent covalent bonds). Each node stores information on the atom name, type and hybridisation. All nodes that are consistent between the original and mutated residues (i.e. have the same atomic properties) and are at the same position within their respective graphs are considered as preserved, and so the coordinates they had in the original residue are carried over to those for the mutated residue. Backbone atoms are therefore always preserved, and usually a given number of side-chain atoms are too. This process therefore preserves the overall orientation of the residue.

For those mutated atoms not assigned coordinates at this stage, a subgraph is created to allow for the coordinates to be created one-by-one, thus 'building up' the remainder of the mutated residue. This construction is achieved first by locating all of the atoms in the subgraph that have a connection to a node already assigned coordinates (known as the host atom). Unit vectors from the host atom, which take into account its hybridisation state, are constructed to generate coordinates for those atoms connected to the host. These unit vectors are then adapted (i.e. scaled) to account for the properties (atom type and hybridisation state) of the new atoms to give a final set of coordinates. This procedure is then repeated iteratively until all the mutated atoms have been assigned a position.

There are caveats for certain residue types, particularly ring systems. Also, methods have been put in place to deal with potential chirality problems. For more details, please refer to the PhD thesis of Röder.[134]

By interfacing CHECKSPMUTATE with this method to generate starting coordinates for mutated residues, a new overall structure with mutations applied to the user-specified residues could be created, ready for optimisation. If the starting stationary point had been a minimum, reoptimisation used the L-BFGS routine as described in Section 4.5. Otherwise, if the stationary point had been a TS, reoptimisation was carried out using hybrid eigenvector-following (HEF), as described in Section 4.6. This process was cycled until all stationary points in the pathway had been mutated and an attempt made to reoptimise them.

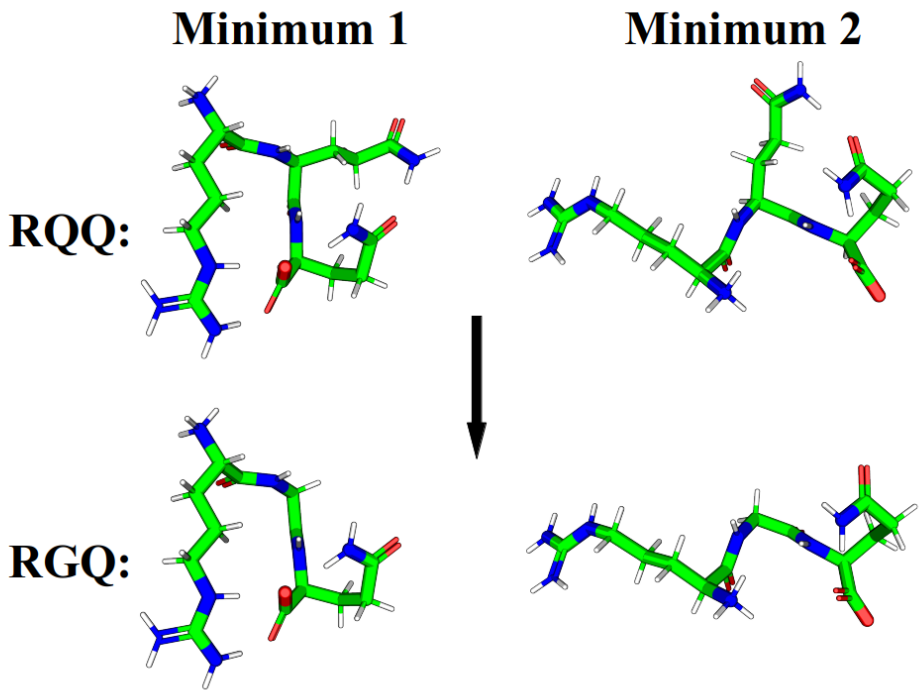


Figure 6.1: RQQ to RGQ mutations and reoptimisations. The top structures show two different minima (labelled 1 and 2) from a database describing the RQQ tripeptide. A Q → G mutation was applied to the central residue in each case, and the resulting set of coordinates reoptimised. The general conformational properties of the RQQ minima are retained in the new RGQ minima in each case.

To illustrate this approach, Fig. 6.1 shows two different stationary points of the RQQ tripeptide and their equivalent stationary points in the RGQ tripeptide after mutation and reoptimisation.

However, mutations that increased the overall number of atoms in the residue being mutated also had to be considered. Furthermore, it was recognised that for homologues several residues would need to be mutated, plus some may need to be deleted from or added to the sequence. Therefore, **CHECKSPMUTATE** was developed further to account for each of these possibilities. In all, four types of transformation were attempted on the test system, RGQ. These were:

RGQ → RFR. Two mutations attempted at once. Both constituted an increase in the number of atoms.

RGQ → RQ. Deletion of the central G residue.

RGQ → RGQFF. Insertion of two residues at the C terminal (which is located at Q).

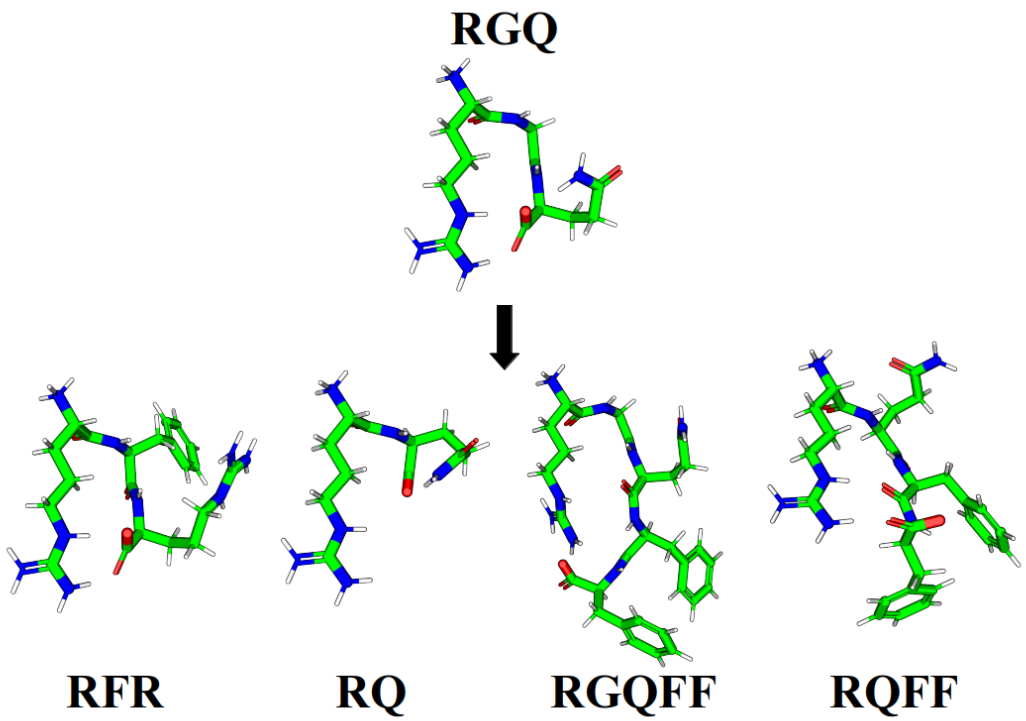


Figure 6.2: RGQ mutations and reoptimisations. Each mutation was from the same RGQ minimum. All reoptimisations converged to stable minima, giving mutated minima of similar conformational properties to the original RGQ tripeptide minimum.

RGQ → RQFF. Deletion of the central G residue and insertion of two residues at the C terminal.

The results from each of these mutations, all applied to the same RGQ minimum, are shown in Fig. 6.2. Reoptimisations following mutations were generally successful. However, particularly for cases where the mutations resulted in steric clashes, certain minima and TSs did not converge to within a user-specified RMS force criterion.

6.3.2 Point Mutations

CHECKSPMUTATE had been demonstrated to successfully mutate and reoptimise the majority of stationary points for the tripeptides studied. This subroutine was therefore applied to the WT HemS system. The Dijkstra shortest pathway was identified for this system and extracted to give a smaller, more manageable database. This database was then copied eleven times, the combined total of HemS with point mutations applied, or homologues of HemS, to be studied. **CHECKSPMUTATE** was then applied to each one.

	No. Minima	% Successfully Converted	No. Transition States	% Successfully Converted
Wild Type	**1235**	**N/A**	**1234**	**N/A**
F104A	1105	89.5%	803	65.1%
F104AF199A	1059	85.7%	557	45.1%
F104I	1110	89.9%	979	79.3%
F199A	1068	86.5%	838	67.9%
R209A	1050	85.0%	642	52.0%
R209K	1090	88.3%	960	77.8%
Q210A	1079	87.4%	660	53.5%

Table 6.1: Success rate of point mutations and reoptimisations from the minima and transition states of the WT HemS Dijkstra shortest pathway.

Analysis of the WT HemS pathway (see the next chapter) had inspired further study of six different single-point mutations (F104A, F104I, F199A, R209A, R209K and Q210A) and one double-point mutation (F104AF199A). As these each constituted only one or two mutations in the protein out of a possible 338, and all such mutations led to a decrease in the number of atoms in the residue(s) being changed, the success rate of reoptimisations proved to be high. They are listed in Table 6.1.

With so many stationary points successfully mutated and reoptimised in each case, the pathways for the mutants were mainly complete, and the gaps to fill in were generally small. Strategies to fill in these gaps will be explained in Section 6.3.4.

6.3.3 Homologues

Conversion to homologues is a far more involved process than single- or double-point mutations. Even HmuS, which has an 89.6% sequence identity with HemS, required 34 residues to be converted. ChuS, meanwhile, required the conversion of 111 residues as well as the deletion of one midway through the sequence and another at the C terminus. ShuS required a change of 115 residues and the deletion of one near the middle of the sequence. Conversion to PhuS was also attempted, a transformation requiring 199 residues to change, 5 to be added at various points in the sequence, and one to be removed. The sequences themselves are given for each homologue in Appendix B.

It was therefore not surprising that the rate of reoptimisation success was not as high for these homologues as it had been for the point mutations to HemS. Nevertheless, other than for PhuS, these rates remained good enough to give fairly complete pathways, as shown in Table 6.2.

It is unsurprising that, as the % Identity of the homologue with respect to HemS decreases, the number of minima and TSs successfully reoptimised also decreases. Not a single minimum or TS was successfully mutated and reoptimised to PhuS.

This failure was probably due to certain mutations causing steric clashes within the structure common to all of the stationary points along the pathway. The crystallographic structures of PhuS and HemS (PDB codes 4mf9[111] and 2J0P,[96] respectively) reveal that haem in PhuS is less tightly clamped, probably due to extra steric bulk in that region. Trying to fit these bulkier residues in a narrower pocket, as **CHECKSPMUTATE** using HemS as a template does, may have resulted in overlapping residues, precluding successful structural reoptimisation. This result is a reminder that **CHECKSPMUTATE** is only feasible where the structure of the new system is sufficiently similar to the template system. Results will vary between different protein-ligand systems, but it would seem, from these data on HemS, that an advisable rule-of-thumb would be to avoid trying to mutate and reoptimise to homologues where the sequence identity falls below 65%.

	% Identity to HemS	No. Minima	% Successfully Converted	No. Transition States	% Successfully Converted
HemS	**100%**	**1235**	**N/A**	**1234**	**N/A**
HmuS	89.6%	674	54.6%	706	57.2%
ChuS	66.8%	208	16.8%	255	20.7%
ShuS	66.2%	131	10.6%	184	14.9%
PhuS	42.6%	0	0%	0	0%

Table 6.2: Success rate of homologue conversions and reoptimisations from the minima and transition states of the WT HemS Dijkstra shortest pathway. % Identities are taken from experimental homology calculations using the sequences from the accession numbers listed in Appendix A.

6.3.4 Post-Processing

Though PhuS failed, all of the other point mutations and homologues studied gave reasonably complete pathways. It should be stressed that, even for ShuS, where only 131 minima and 184 TSs were successfully reoptimised (% conversions of 10.6% and 14.9%, respectively), this still provides far more to start with than had been the case with WT HemS, where just five minima (the original binding sites identified by Choy, Shang and the author) were used to connect the entire pathway using DPS. Even accounting for the fact that certain regions of the pathway would be poorly optimised (if one stationary point failed to reoptimise, it was probable that its neighbouring stationary points, most likely having similar conformational characteristics, would also not reoptimise successfully), this result suggested that the gaps to fill in would be far shorter than those which had had to be tackled for HemS.[c]

[c]Even though each of these five sites were sampled locally before connections were attempted, thus giving more minima to choose from, it must be remembered that they were all concentrated in particular regions of the landscape, whereas the mutation and reoptimisation scheme of **CHECKSPMUTATE** provides for a wider spread of minima, reducing the likelihood that any large gaps will need to be connected.

To fill in these gaps for each system, steepest-descent paths were followed for all of the TSs to find the minima they were directly connected to, which were often the reoptimised minima already identified. However, a number of new minima were also found.

Attempts were then made to connect the gaps in the pathway. As the gaps arose from stationary points that failed to converge when reoptimised, to bridge these gaps, those minima that had been successfully reoptimised and were closest according to the template pathway to those that had failed were selected for connection attempts. This approach generally worked well because the gaps tended to be small. Therefore, in each mutated system (whether point mutation of HemS, or homologue) the majority of the gaps were filled. However, large gaps (which can arise, for example, if a long chain of stationary points from the original pathway is not reoptimised successfully) proved to be harder to bridge.

When taken together, all of the DPS runs, whether they had succeeded in connecting gaps or not, generated a large number of new minima and TSs. Therefore, each of the sub-databases within the overall database had grown large.[d] **CONNECTUNC LOWESTTEST**, discussed in the previous section, was therefore used to choose which sub-databases would be best to try to connect, and which minima within these respective sub-databases should be selected to attempt these connections.

Overall, it took approximately six months to find a fully connected pathway for all ten mutated systems (seven point mutations and three homologues). These runs were done in parallel using limited computational resources, and so if each had been performed separately they could each have taken approximately three weeks. This timescale is a significant speed-up from the original WT HemS system, where it took over twelve months to find a fully connected pathway from the five starting structures that had been identified by Choy, Shang and the author.

Care must be taken, however, not to assume that these new mutated pathways are optimal. These new pathways are liable to have artificially high barriers, and so refinement using the methods outlined in Section 4.8 (i.e. **SHORTCUT**, **SHORTCUT BARRIER**, **UNTRAP** and standard **CONNECTUNC**) is required. This refinement is particularly important, as an inherent bias was introduced by using the WT HemS pathway as a template. Because of this bias, the new pathways for these mutated systems are likely to follow a route more akin to the original WT HemS pathway than possible alternative, lower energy pathways unique to the new system. Suffi-

[d]Recall that a sub-database in this context is a collection of minima and TSs all connected to each other, whether directly or indirectly. Therefore, a certain segment of a pathway containing no gaps can be considered as a sub-database, and it is separated from other sub-databases by gaps in the pathway.

cient refinement and sampling is expected to iron out these differences, provided the point mutations administered or homologues considered do not significantly alter the protein structure. Experimental evidence would suggest both of these assumptions are justified for the systems under study in this work. The *apo*-structures of the mutants studied, following certain biophysical investigations (see Chapter 9), showed little evidence of misfolding and the homologues, where crystal structures are available (e.g. for HemS and ChuS), show similar overall structures.

6.4 Discussion and Summary

This chapter described the development of two different strategies to assist in the overall connection of difficult pathways.

The first scheme uses **CONNECTUNC LOWESTTEST**, a simple algorithm that determines, from a set of sub-databases, the fewest and narrowest set of conformational gaps that need to be filled in order to connect all of the sub-databases (and therefore, by implication, all of the stationary points contained within each). It does this both by considering which sub-databases would be best to try to connect, as well as which minima to select within those sub-databases.

CONNECTUNC LOWESTTEST was first applied to the WT HemS system in order to best select the minima from each of the five original sites identified by Choy, Shang and the author for connection attempts. Due to the extensive sampling of the landscapes surrounding these sites before **CONNECTUNC LOWESTTEST** was used, these regions were well-characterised. This situation contrasts with the regions of the landscape between sites, which were not characterised at all until **CONNECTUNC LOWESTTEST** was used.

There is therefore an issue with this overall **CONNECTUNC LOWESTTEST** strategy in that it leads to some regions of conformational space being well sampled (those regions surrounding the original sites) whereas other regions are less well sampled (those between the original sites). This imbalance can be solved by pathway refinement, where the refinement schemes described in Section 4.8 can be set to focus more on the stationary points of these less sampled regions.

It is difficult to determine how much computational time, if any, the strategy using **CONNECTUNC LOWESTTEST** saved with respect to finding a fully connected pathway for the WT HemS system, as a parallel study using more conventional selection methods to fill gaps was not conducted. As the strategy was able to identify the closest two minima between two separate sub-databases, and it was later demonstrated by the work with **CHECKSPMUTATE** just how important it is to

try to minimise the gap size between attempted connections, it is likely that this strategy saved a significant amount of computational time and resources.

The second strategy developed could only be applied to systems where a pre-existing system with similar characteristics (in terms of sequence *and* suspected structure) was already available. This strategy was a 'template-based' approach, and depended upon the subroutine, **CHECKSPMUTATE**. In this strategy, all of the stationary points from the original system (in this study, describing the pathway showing movement of NADH towards haem in WT HemS) were mutated at specified residues and the resulting coordinates reoptimised. In addition to this procedure, provisions were made so that residues could be inserted or deleted if required. The success rate for reoptimisations was typically good, but generally became harder when more residues had to be mutated. This trend was confirmed when HemS was transformed to PhuS (a homologue with 42.6% homology) and not a single minimum or transition state was successfully reoptimised.

The success rate for reoptimisation was never 100%, so gaps appeared in the pathways for the new systems. Most of these gaps were short and could easily be bridged by considering where these gaps would have appeared in the original WT HemS system, and connecting the closest successfully reoptimised minima either side of this gap.

For larger gaps that could not be easily connected, **CONNECTUNC LOWESTTEST** was used to select the closest pairs of minima for connection attempts in what were by now quite extensive sub-databases, as well as which sub-databases would be best to try to connect in the first place.

Fully connected pathways were found for all of the new systems (ten in all) within six months. Had each system been allowed to run individually, rather than computational resources being pooled between all ten, it is estimated that each one would have completed within three weeks. This timescale is a significant speed-up with respect to the original approach used to find a fully connected pathway for the WT HemS system, which took over twelve months.

Caution must be exercised, however, when it comes to using this 'template-based', **CHECKSPMUTATE** approach. It is only suitable when similar pockets within a protein are being investigated, without large-scale differences between the pathways of the different systems. For example, this method would not be suitable for investigating the majority of protein-folding problems, where the conformational changes a protein undergoes are often very significant, unless it is apparent that the two proteins fold *via* very similar pathways. One of just a few scenarios where this exemption would apply would perhaps be if a single-point mutation was applied to

a protein that was expected to only affect protein folding at a local level.

Instead, this 'template-based' approach is more suited to studying protein-ligand interactions (as in this thesis) as such interactions tend to be localised in a particular section of the protein. As long as the mutations do not significantly affect the overall shape of the binding pocket, there is justification for using the pathway from one system to derive the other. Although this approach biases the mutated pathway to be like the template, the fact that the mutated coordinates are reoptimised allows for new energies to be obtained, showing how the landscape transforms with respect to the new mutations. Sufficient refinement of the pathway should also identify any possible alternative pathways, which are likely to be similar since the binding pockets between the template and the mutated systems are related.

Provided these refinement conditions are met, this 'template-based' approach can be an effective method for deriving pathways for new, yet related, systems at a fraction of the computational cost these would otherwise require. The 'template-based' approach therefore allows for multiple systems of mutated proteins or homologues to be constructed quickly, in preparation for thermodynamic and kinetic comparisons.

Chapter 7

Computational Comparison of HemS with its Mutants and Homologues

7.1 Aims

Using GPUs and the novel methods described in the previous chapter, it became possible to construct a large stationary point database for the [WT HemS + Haem + NADH] system for the first time. From this database, it was possible to find fully connected pathways showing NADH moving along the main cavity towards haem. It was thought that these pathways would shed light on the residues involved in NADH-binding, unfolding and movement. The double phenylalanine gate was of particular interest due to its suspected role in regulating NADH access to haem.

Once a fully connected pathway had been found and the WT HemS system analysed, further aims were introduced. Certain residues had been identified as important for NADH-binding, and so these were mutated to ascertain what the effects would be. A system using NADPH in place of NADH was also briefly investigated to determine whether it behaved in the same manner. Finally, three HemS homologues, HmuS, ChuS and ShuS, were investigated to determine whether they too could facilitate NADH unfolding and movement inside their pockets to promote reaction with haem. These aims were all attempted with the aid of the **CHECKSPMUTATE** subroutine. To summarise, the aims of the computational investigation were:

1. To construct a large stationary point database for the [WT HemS + Haem + NADH] system.

2. To find a fully connected pathway showing NADH moving towards haem

within the WT HemS pocket.

3. To investigate which residues are of importance to NADH movement and unfolding within the WT HemS pocket, with a special emphasis on the double phenylalanine gate.

4. To use the [WT HemS + Haem + NADH] pathway as a template to find pathways for other systems, including selected single-point mutants, homologues and alternatives to NADH.

5. To compare the structural and energetic properties of these systems.

Outcomes from calculations inspired by these aims are described in the following sections.

7.2 Expansion of the Wild Type HemS Database, and Further Analysis of the Double Phe-Gate

The author inherited the investigation into the relationship between the double phe-gate and NADH distance from haem started by Shang.[159] This work was discussed towards the end of Section 1.7.2, particularly in Fig. 1.28.

Due to the enormous increase in computational efficiency afforded by GPUs, this work could be continued on a much larger scale. For certain calculations, this speed-up amounted to two orders of magnitude. Therefore, where the number of basin-hopping steps had typically been 100 in previous studies for an individual run, this value was increased to 10,000, allowing for larger regions of local landscapes to be sampled.

To recap, Shang had identified three binding sites (optimised minima structures) for NADH at various stages of its progression towards haem in the protein pocket, in additon to the original site identified by Choy using Relibase^{+}.[21] These sites were numbered from 1-4, with 1 indicating the site where NADH was furthest from haem, and 4 indicating the site where NADH was closest. Through manual manipulation with PyMOL[240] and subsequent reoptimisation, the author identified a further site (labelled 5), where NADH was even closer to haem.

Basin-hopping was therefore carried out starting at each of these sites (and also from a *holo*-HemS system, i.e. without NADH) using short step sizes to better sample the local landscapes surrounding each of these sites. Ten separate basin-hopping runs of 10,000 steps were conducted. The lowest minimum structure from each of these runs was then selected for further study.

The overall aim of this set of calculations was to determine the relative stabilities of the phe-gate conformations at various stages of the progression of NADH along the pocket (and also when NADH was not present at all). It was therefore necessary to define exactly what constituted open and closed gates. As the double gate consisted of two residues, F104 and F199, it could be described as having one of four conformations – closed-closed (CC, i.e. $C_{F104}C_{F199}$), closed-open (CO), open-closed (OC) or open-open (OO). Four representative examples of these conformations were provided in Fig. 1.27 in the Introduction. These labels describing these conformations were introduced by Shang. However, a systematic way of determining what constituted a closed or an open state had not been necessary at that time; it was possible for Shang to look at the relatively few minima available on a visualisation package, and determine what conformation the two phenylalanine residues were in by sight. However, it was anticipated that many thousands of stationary points would be generated using the GPUs, and so a more rigorous, mathematical method for classifying which conformations the phenylalanines were in without the need to visualise them was developed. Due to the desire to distinguish between OC and CO states, it was realised that some sort of metric for the two phenylalanines with respect to each other would not work as the distances and angles between these two residues would be very similar for these OC and CO conformations. Instead, each phenylalanine was defined with reference to the haem molecule, whose location between stationary points was very consistent, not least because of the FE–H196 Nε bond introduced by Choy.[21] Specifically, for each phenylalanine, a dihedral angle, θ, between it and the haem was defined according to three of its carbon atoms (α, γ and ζ) and the β-*meso*-carbon of haem, in the order Cα–Cγ–Cζ–Cβ-*meso*. Through observation and testing, the four conformational states were classified as follows:

- **CC:** $-90^\circ \leq \theta_{F104} < 80^\circ$ and $0^\circ \leq \theta_{F199} < 175^\circ$
- **CO:** $-90^\circ \leq \theta_{F104} < 80^\circ$ and ($\theta_{F199} < 0^\circ$ or $\theta_{F199} \geq 175^\circ$)
- **OC:** ($\theta_{F104} < -90^\circ$ or $\theta_{F104} \geq 80^\circ$) and $0^\circ \leq \theta_{F199} < 175^\circ$
- **OO:** ($\theta_{F104} < -90^\circ$ or $\theta_{F104} \geq 80^\circ$) and ($\theta_{F199} < 0^\circ$ or $\theta_{F199} \geq 175^\circ$)

The conformational state of the double phe-gate for the lowest minimum from each of the BH runs was determined. The conformations of these two phenylalanines were then changed (by substitution, using phenylalanine coordinates from other stationary points) to reflect the other three possibilities, and these resulting structures reoptimised. Therefore, the four different double phe-gate conformations could be

compared for minima which were otherwise identical.[a] As such, a direct comparison between the relative stabilities of these double phe-gate conformations could be made. These four minima were then connected to one another in order to identify possible barriers and to allow for them to be plotted as disconnectivity graphs. Typically, pathways between these four minima involved intermediates, and so the disconnectivity graphs tended to grow to more than four minima.

As stated above, ten different BH runs were performed on each site (plus for *holo*-HemS), giving sixty BH runs in all. The lowest minimum from each of these was extracted, and the phe-gate substitutions/subsequent connections applied to each. This procedure gave sixty disconnectivity graphs.

As a result, at each of the six sites, there were ten separate sets of minima describing how the energy of the system varied when the F104 and F199 gates were in different (i.e. CC, CO, OC or OO) conformations. These ten sets were averaged and plotted, as shown in Fig. 7.1, allowing for general trends between the sites (i.e. as NADH moved through the pocket towards haem) to be identified. This analysis is discussed further in the caption to Fig. 7.1, but the main findings were that, as NADH moved further into the pocket, the general stability of the system increased, and that the OO state became more favoured over the other conformations.

The ten outputs from each site were then connected to each other using Discrete Path Sampling (DPS). The sites themselves were then connected to one another using the **CONNECTUNC LOWESTTEST** strategy described in the previous chapter, producing a large database of all of the stationary points identified for the [WT HemS + Haem + NADH] system. Refinement of the Dijkstra shortest pathway resulted in further minima and TSs, giving a total of 21,852 connected minima in the disconnectivity graphs shown in Figures 7.2 and 7.3.

The two disconnectivity graphs in Figures 7.2 and 7.3 represent the same database but are colour-coded differently. One represents the double phe-gate conformations within the respective minima, whereas the other is coded according to the NADH-haem distance. Specifically, this NADH-haem distance was determined using the β-*meso*-carbon of haem and the hydride-bearing carbon of NADH, as it was these two carbon atoms which were thought to be most involved in the hydride transfer process.

Fig. 7.2 shows that the CO conformation is preferred by the vast majority of minima in the database. This result is most likely due to the stabilising effect of a T-shaped π-π bonding interaction that this conformation affords. However, as

[a]Identical save for very smaller differences introduced when the substituted structures were reoptimised.

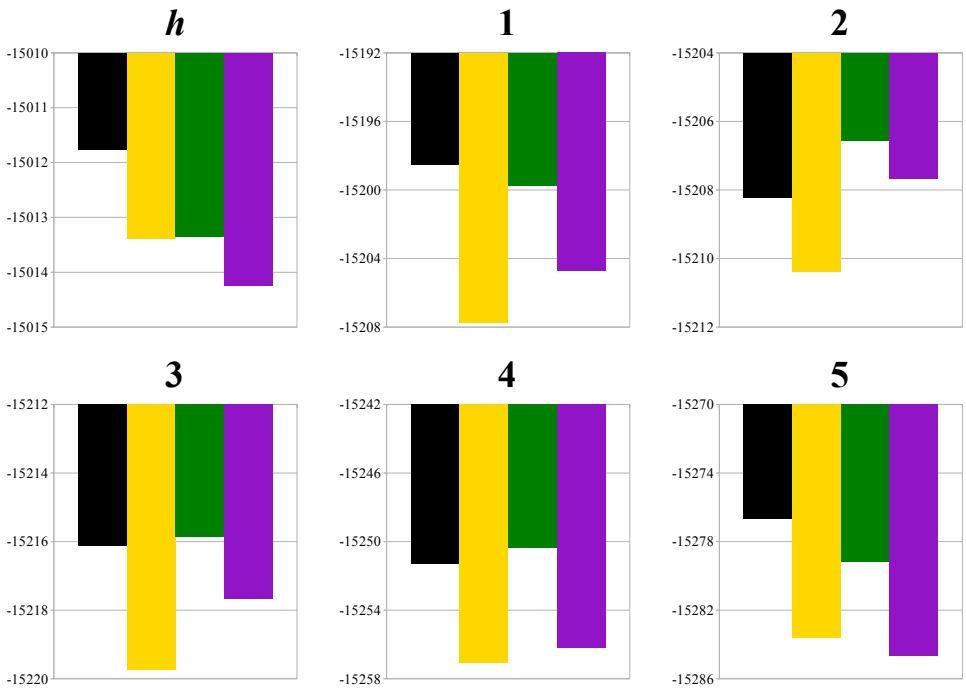


Figure 7.1: Charts showing the relative stabilities of the double phe-gate conformations when NADH is located in different parts of the pocket. ***h*** represents the *holo*-HemS system (i.e. without NADH), whereas **1**–**5** represent an increasing degree of NADH unfolding and progression towards haem within the pocket. Bars are colour-coded according to the double phe-gate conformation: CC, black; CO, yellow; OC, green; OO, purple. The y-axis gives a measure of the potential energy of the system, in kcal mol^{-1}. The scale varies between the systems. In moving from **1**–**5**, the energies for each of the conformations drop significantly, indicating that the movement of NADH through the pocket is enthalpically favourable. In the *holo*-system, the energy difference between conformations is low, suggesting that the OO conformation is marginally favoured against the others, and CC marginally disfavoured. Upon inclusion of NADH, at site 1, the CO conformation becomes significantly more favoured than the others. At **2**, CO remains the most favoured conformation. OC is particularly disfavoured, being even higher in energy than the CC conformation. These two features also prove to be the case at **3** and **4**, although the OO conformation is becoming relatively more stable. At **5**, a switch occurs, whereupon the OC conformation becomes more stable than CC, and the OO conformation becomes more stable than CO, and thereby the most stable conformation of them all.

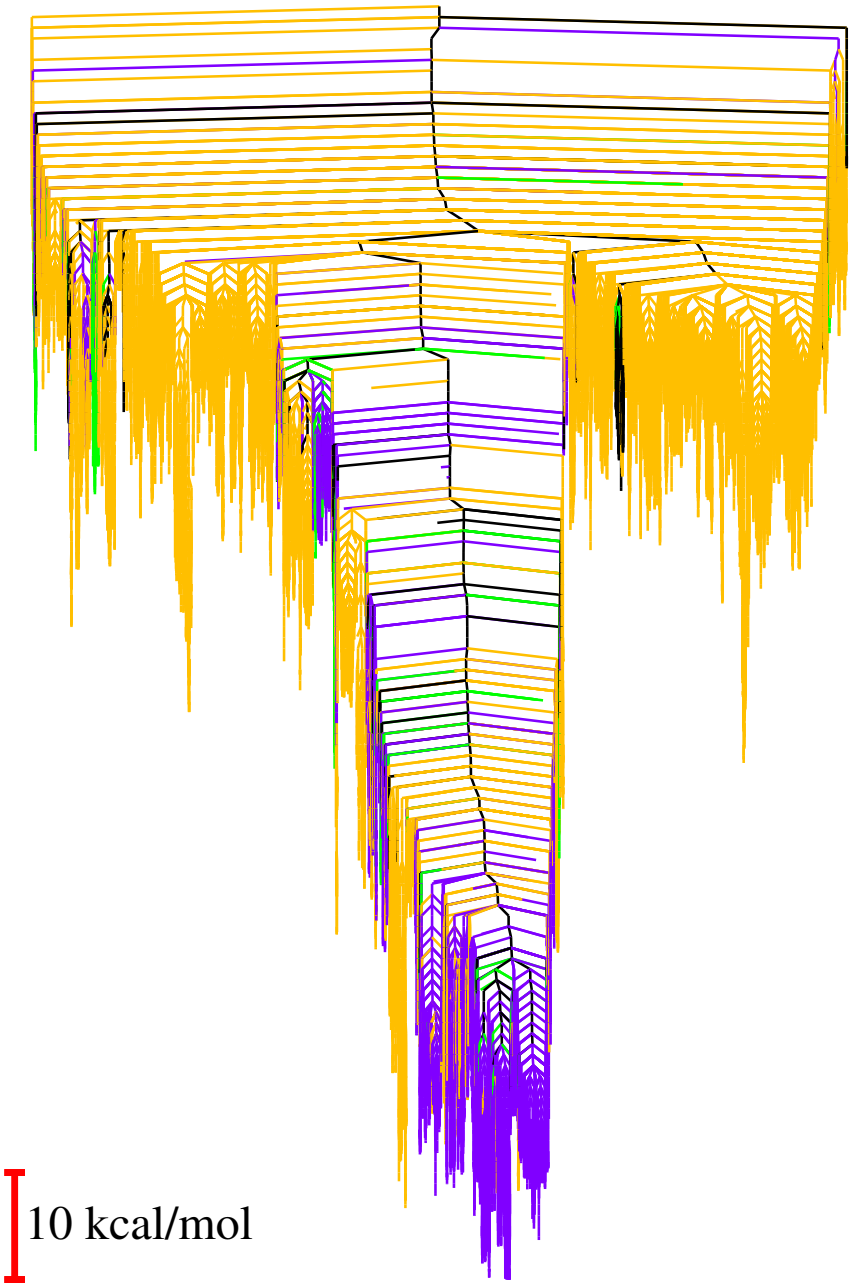


Figure 7.2: Disconnectivity graph of [WT HemS + Haem + NADH], colour-coded according to double phe-gate conformation. Minima with a CC conformation are coloured black, CO yellow, OC green and OO purple. The vast majority of minima in the database displayed a CO conformation, suggesting most minima have a preference for this T-shaped π-π bonding interaction. The majority of the lowest energy minima, however, have an OO conformation, suggesting that the most stable minima prefer an arrangement where the two phenylalanine residues point away from each other, and away from the centre of the pocket.

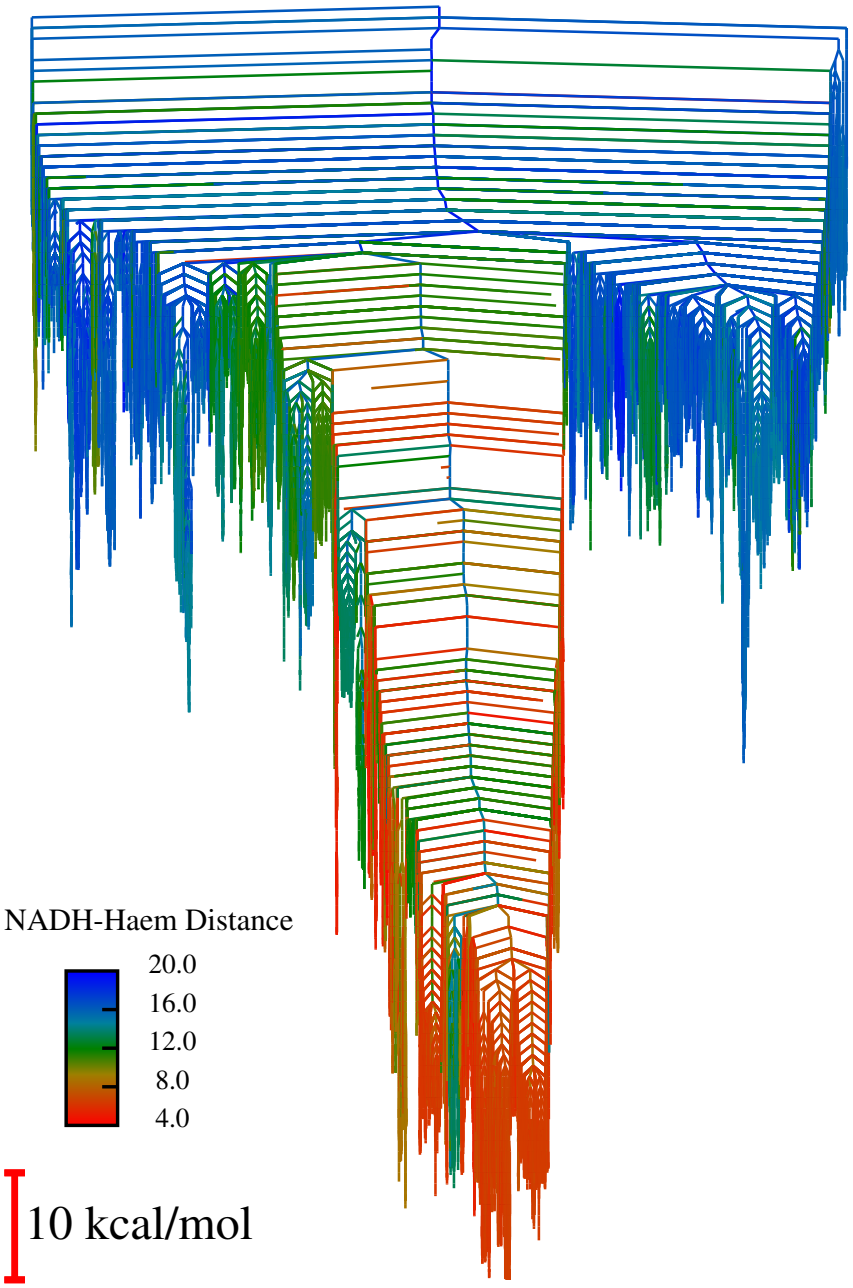


Figure 7.3: Disconnectivity graph of [WT HemS + Haem + NADH], colour-coded according to NADH-haem distance, as shown by the colour-bar, which is demarcated in Angstroms. The main funnel is dominated by minima with low NADH-haem distances, suggesting that the approach of NADH to haem has a stabilising effect.

NADH approaches, F104 is required to open to allow passage of the ligand through to haem. Therefore, when NADH is close to haem, the OO conformation becomes the preferred option. This conclusion can be established by comparing Figures 7.2 and 7.3, where the minima with low NADH-haem distances are seen to overlap almost exactly with the minima possessing an OO conformation. Furthermore, these minima appear at the bottom of the main funnel, suggesting that the NADH approach to haem is an optimised process (i.e. that the protein has evolved to perform this function or one very much like it).

The appearance of the disconnectivity graph in Fig. 7.2 therefore led to a revision of the double phe-gate hypothesis. Rather than occupying one of four conformations, it is clear (once NADH has been introduced to the system, at least) that two conformations (CO and OO) are far more prevalent than the other two. Furthermore, the CO conformation predominates in all cases other than when NADH is close to haem. This result suggests that the CO conformation is the most stable option, save for those instances when NADH occupies the space between the two residues. Even though this CO conformation is possible when NADH comes between the two phenylalanines, it is clear from a steric consideration that the protein retaining this CO conformation would lead to a very crowded region in the cavity. It is therefore not surprising that F104 flips round so that it points away from NADH, giving the nicotinamide head more space to orient itself towards haem. More properly, therefore, the [WT HemS + Haem + NADH] system should perhaps be considered as only having two possible conformations at the 'double' phe-gate: the CO conformation, which should be considered simply as the closed state, and the OO conformation, which should be considered as the open state. Meanwhile, the CC and OC states, though possible, are so rare that they should be removed from consideration as meaningful competitors.

This result is especially interesting because the *holo*-HemS (i.e. without NADH) system shows no such bias towards the CO (closed) state, as illustrated in Fig. 7.4. Though this *holo*-HemS system was not as thoroughly sampled, the disconnectivity graph shows a clear preference for the OO conformation, followed by OC. This observation would suggest that CO preference is dependent upon the inclusion of NADH, whereupon F104 switches from an open to a closed state. It was unclear why this preference would be the case; perhaps the inclusion of NADH, even at the very edge of the pocket, leads to some sort of structural change within the protein, which has a knock-on effect on residues deeper within the pocket, culminating in the change of F104 conformation.

Further investigation of the stationary points upon NADH inclusion revealed the

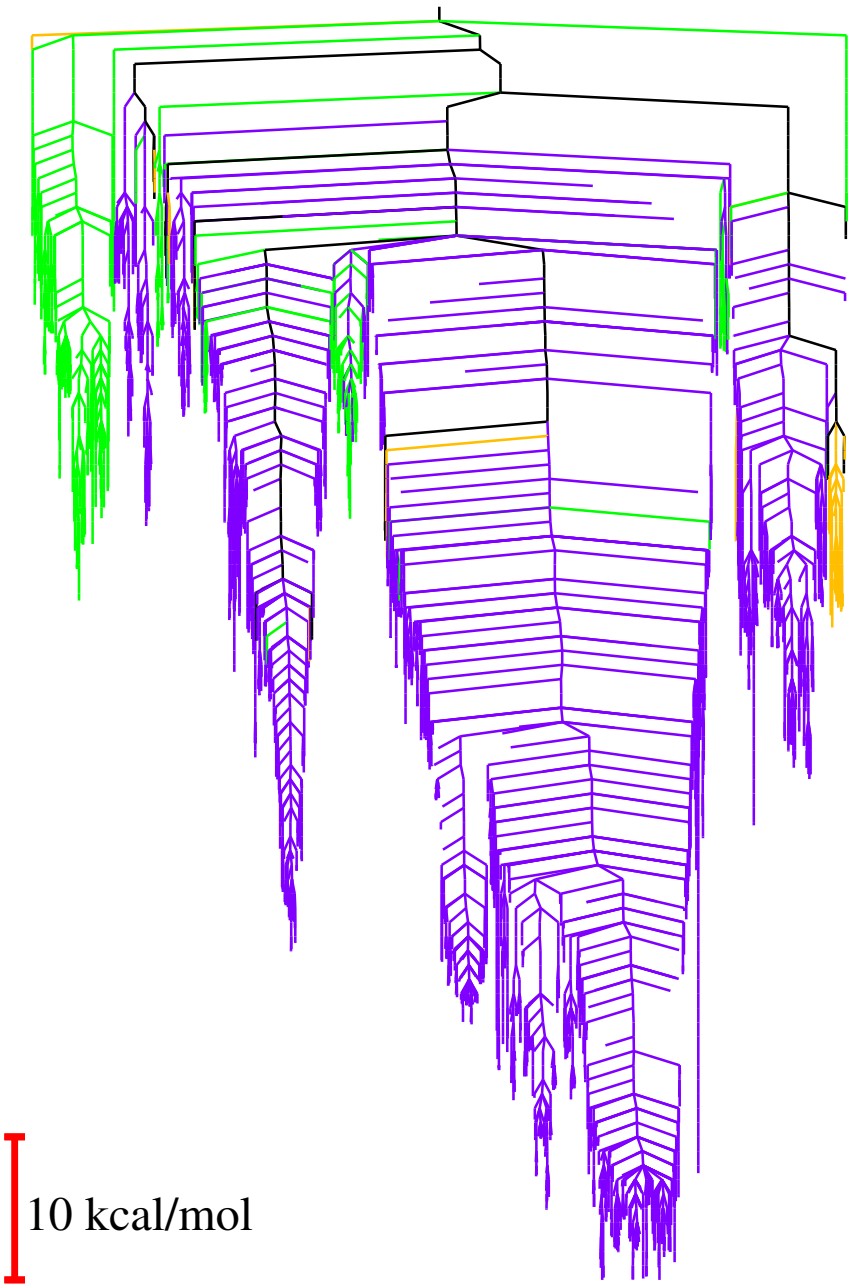


Figure 7.4: Disconnectivity graph of *holo*-HemS, colour-coded according to double phegate conformation. The OO conformation is preferred, and there are few instances of CO.

precise reasons behind this effect.[b] When NADH is introduced to the pocket, its adenine nucleobase positions itself such that it binds to two residues at opposite sides of the cavity (namely, the backbone of P169 and the carboxylate group of D310). This binding serves to narrow the entrance to the cavity significantly, holding the NADH molecule in place. This alteration to the cavity affects the unstructured loop further down its length, bringing a glutamate residue (E174) into much closer vicinity to an asparagine (N106). The glutamate hydrogen-bonds to the asparagine, preventing it from bonding to the π system of an open F104. Without this stabilising binding interaction to asparagine, F104 moves to a closed position. As NADH unfolds and its nicotinamide head approaches the double phe-gate, E174 is 'pushed' upwards again, causing the E174-N106 hydrogen-bond to break. This change leaves N106 free to bind to F104 again, thus causing F104 to open, and thereby let NADH slip past. Three representative structures – two from the database with NADH, and the other from the database without – are shown in Fig. 7.5, with the relevant residues and bond lengths highlighted, to illustrate the interactions involved.

7.3 Identifying Residues to Mutate

As the double phe-gate appears to be key to the regulation of NADH access to haem, the F104 and F199 residues were chosen to form the basis of a mutagenesis study. In all, four mutants involving this double phe-gate were chosen: F104A, F199A, F104I and F104AF199A. The two single-point mutations to alanine were to determine if a reduction in the steric bulk at either of the gates led to NADH being able to access haem more easily. F104AF199A was chosen for similar reasons, only this time the entire double gate would be removed. F104I was chosen as isoleucine is of a similar size to phenylalanine but is non-polar and aliphatic rather than polar and aromatic. This choice was to probe whether a change in electronics (specifically, the disruption of the T-shaped π-π interaction between F104 and F199) would have an effect on the gate, rather than simply a change of sterics.

In addition to the double phe-gate, two other residues were selected for mutagenesis, namely R209 and Q210. The first of these, R209, was of interest because it can bind to both haem and NADH simultaneously. R209 is one of the residues (detailed in Section 1.4.1) that binds to the buried propionate group of haem, helping to anchor the haem in the pocket.

NADH at the edge of the pocket is too far away from R209 to interact. However,

[b]These findings regarding the change in F104 conformation when NADH is introduced to the pocket were only discovered towards the end of the project, hence the residues concerned were not included in the mutagenesis study.

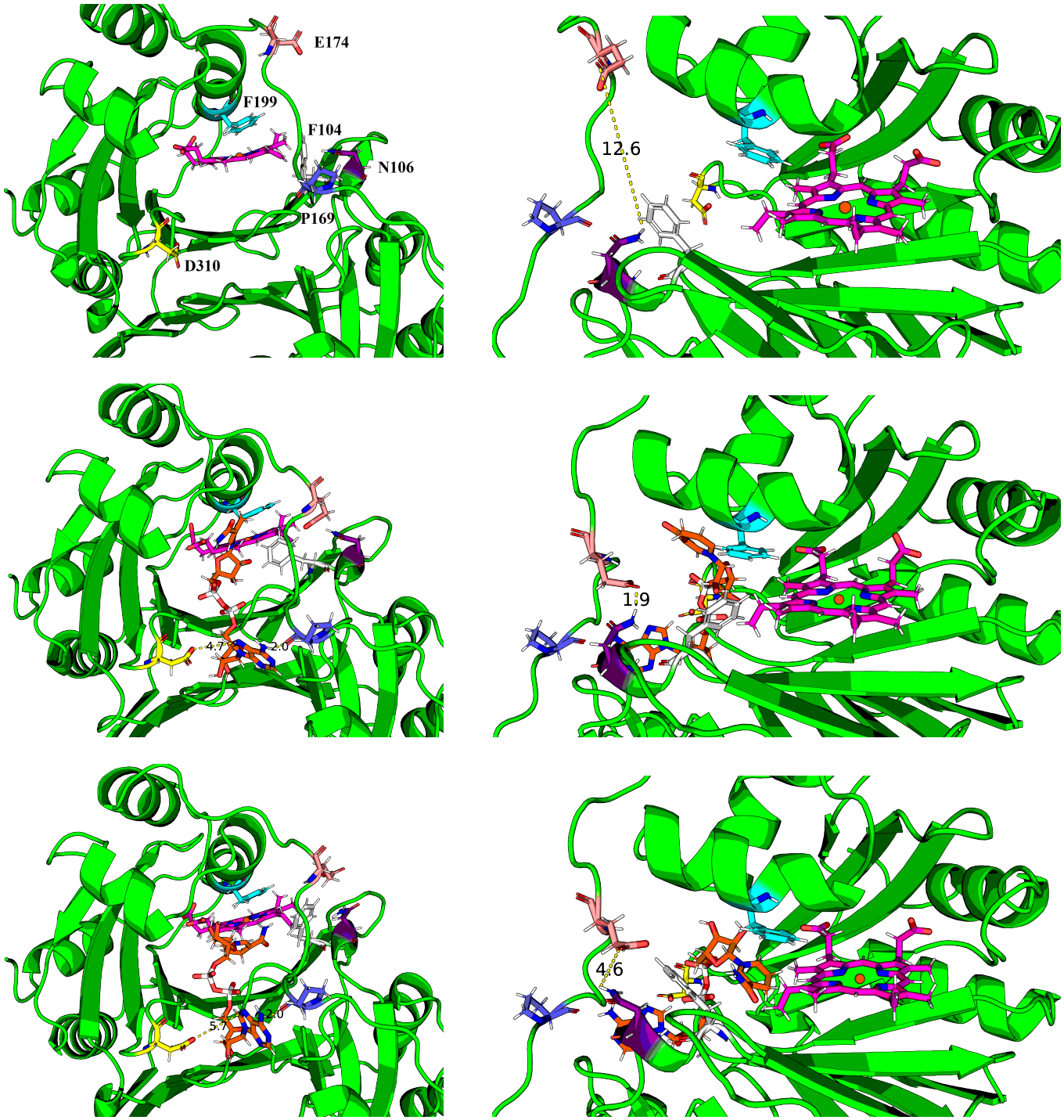


Figure 7.5: Front and side views illustrating the long-scale effect of NADH-binding on F104 conformation. Haem is represented with a magenta skeleton, NADH (where present) in orange, F104 in white, N106 in purple, P169 in blue, E174 in salmon, F199 in cyan and D310 in yellow. Top left & right: front and side views of a selected minimum from the *holo*-HemS database. The cavity is open wide, with P169 and D310 far apart. E174 is also far from N106 (12.6 Å), which frees up N106 to form a polar contact with F104, keeping the latter in an open conformation. Middle left & right: front and side views of a minimum from the [WT HemS + Haem + NADH] database, where NADH has not unfolded enough to reach the double phe-gate. P169 and D310 hydrogen-bond to the adenine nucleobase of NADH from different sides (2.0 Å and 4.7 Å respectively), narrowing the cavity entrance.

Figure 7.5: (continued) This movement brings E174 in close contact (1.9 Å) with N106. N106 therefore shifts in conformation to make this bond, leaving F104 free. F104 changes from an open to closed conformation in order to form a T-shaped π-π bond with F199. This movement blocks NADH from easy access to haem. Bottom left & right: front and side views of a minimum from the [WT HemS + Haem + NADH] database, where NADH has extended fully and slipped past the double phe-gate. P169 and D310 still dock the adenine nucleobase of NADH in place (2.0 Å and 5.7 Å respectively). However, the approach of the nicotinamide head of NADH to the double phe-gate has caused E174 to shift back upwards. To remain bonded, N106 therefore changes its conformation considerably, ultimately yielding a bond 4.6 Å long. This movement of N106 frees up the space required for F104 to flip back to an open conformation, which in turn provides more space for NADH to slip between the two phenylalanine residues and access the haem molecule.

as it approaches haem, the ribose adjoining the nicotinamide head can interact with the free $-NH_2$ of R209. Depending on the orientation of the ribose, this interaction is through either the C2′-hydroxyl or the ring-based oxygen. If the former case, then as NADH slips past the double phe-gate, the hydrogen-bond formed from the free $-NH_2$ of R209 shifts from the C2′-hydroxyl of the ribose to the C3′-hydroxyl. Relevant structures are shown in Fig. 7.6.

Analysis of the [WT HemS + Haem + NADH] database suggested that NADH did not readily slip through the double phe-gate when the ring-based oxygen of its ribose was hydrogen-bonded to R209, yet slipped through more readily when the ribose was oriented so that its C2′/C3′-hydroxyls could bond to R209. It is important to note that the two sites identified by Shang[159] and the author, which were in closest proximity to haem (sites 4 and 5), both had the NADH oriented so that its C2′/C3′-hydroxyls would bond to R209, rather than the ring-based oxygen. On the one hand, the very fact that these were identified by BH suggests that this is the preferred orientation. However, on the other hand, the fact that these geometries were used as starting structures in connecting the pathway, and that structures with the ring-based oxygen bonded to R209 were therefore precluded, suggests that there may have been a bias against identifying these latter structures. With that caveat in place, it is nevertheless notable, that no structures were identified that had the nicotinamide head of NADH past the double phe-gate and the ring-based oxygen of the ribose bound to R209 simultaneously.

This apparent selection for a particular orientation of the NADH molecule as it approaches haem suggests that R209 not only helps to bring NADH and haem into close proximity by hydrogen-bonding to both simultaneously, but also plays a role in ensuring that NADH obeys some aspect of stereospecificity. On the face of it, this result is surprising because the deuterium labelling experiments detailed in Section

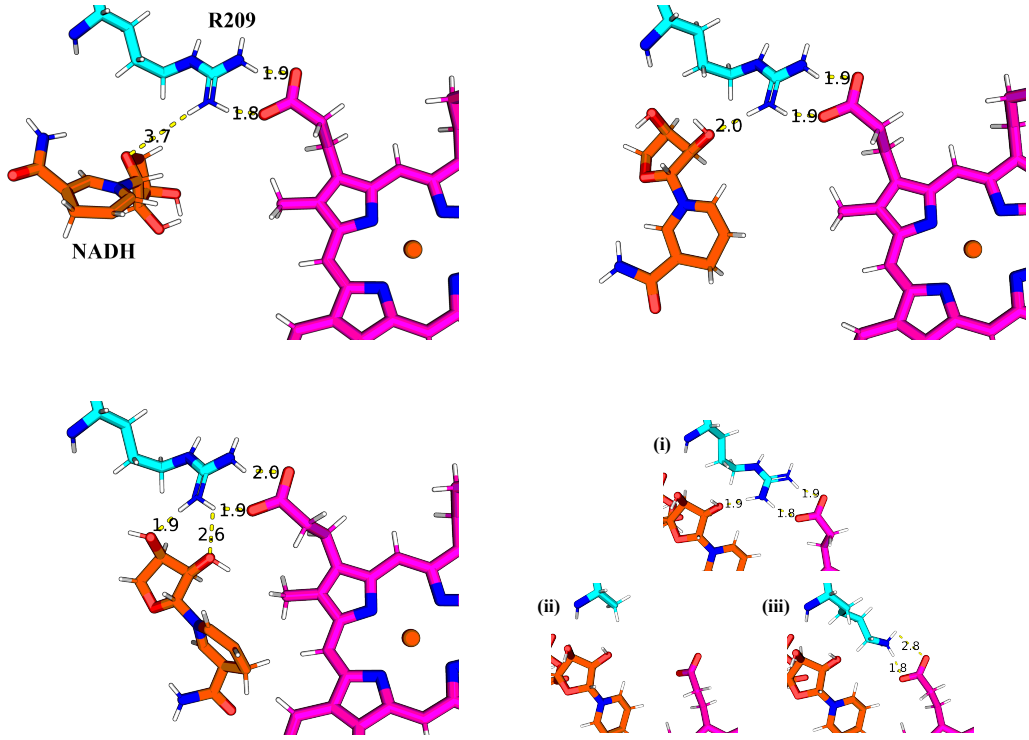


Figure 7.6: Selected minima highlighting the different bonding capabilities of R209. Haem is represented with a magenta skeleton, NADH with an orange skeleton, and R209 in cyan. In every case, one hydrogen from each of R209's two $-NH_2$ groups form tight hydrogen-bonds with the two oxygens of the innermost haem propionate. Top left: The free (i.e. non-haem bound) hydrogen of the $-NH_2$ group of R209 more deeply embedded in the pocket hydrogen-bonds to the ring-based oxygen of the nicotinamide-bearing ribose of NADH. The double phe-gate (not shown) blocks NADH access to haem. Top right: The free hydrogen of R209 hydrogen-bonds to the C2′-hydroxyl oxygen of the NADH ribose. The double phe-gate has opened, allowing NADH to slip past. Bottom left: The free hydrogen of R209 hydrogen-bonds to the C3′-hydroxyl oxygen of the NADH ribose. NADH has slipped further along the pocket, with the nicotinamide head approaching the β-*meso*-position of haem. Bottom right (i): Segment of a selected minimum from the WT HemS database, showing the hydrogen-bonds formed between R209, haem and NADH. (ii): The same minimum following **CHECKSPMUTATE** mutation and reoptimisation to R209A. No hydrogen-bonds are formed between the alanine and haem or NADH. (iii): The same minimum following **CHECKSPMUTATE** mutation and reoptimisation to R209K. The lysine is hydrogen-bonded to the haem propionate. Other minima studied showed lysine was also capable of hydrogen-bonding to NADH, but it could not bond to both haem and NADH simultaneously.

5.3 showed that both the (*R*)- and the (*S*)-hydrides of NADH could be transferred over to haem. However, it was also shown in these deuterium labelling experiments that transfer from the (*R*)-position was far slower than that from the (*S*)-position. It is perhaps the case that R209 plays some role in this difference, although it is not apparent how or for what reason.

Two mutations were applied to R209. One was to mutate to alanine, which was of interest because it reduced the hydrogen-bonding capabilities to zero, and significantly reduced the steric bulk. In case this mutation was to prove too disruptive when applied in the laboratory, a less drastic mutation to lysine was also investigated. This residue retains the ability to hydrogen-bond, but only *via* one amino side chain group, rather than *via* the three present in the guanidino group of arginine. It would therefore not be possible for this lysine group to anchor both NADH and haem in place at the same time.

The final residue of interest was the one adjacent to R209, Q210. This residue was shown to change conformation considerably as NADH progressed through the pocket. Generally, for minima in the *holo*-HemS (i.e. without NADH) database, Q210 hydrogen-bonds to E288, and thus points away from the main cavity. Such a conformation has been termed the 'resting position' for Q210 by the author. The top left image of Fig. 7.7 shows that Q210 remains in this resting position when NADH is at the edge of the pocket. However, as NADH begins to unfold, Q210 swings round to hydrogen-bond first to the G311 backbone. As NADH extends along the pocket still further, Q210 flips over again, so that it can hydrogen-bond to both G311 and the NADH diphosphate backbone. This change most likely stabilises NADH enough for it to overcome its final barriers to full extension. Once NADH is fully extended, Q210 swings back to its original resting position, hydrogen-bonded to E288. This configuration perhaps encourages NAD^+ to fold back up after the transfer of hydride to haem.

Since Q210 appears to have a key role in stabilising extended conformations of NADH, it was hypothesised that its removal may disrupt, or even stop, the transfer of hydride from NADH to haem. Q210A was therefore chosen as a mutant to study as a change to alanine would remove this residue's hydrogen-bonding capabilities.

Analysis of the stationary points and pathways generated for the WT HemS system had therefore helped to develop a deeper understanding of residue-ligand bonding within the protein pocket. Furthermore, it is unlikely that the behaviour of Q210 would have been identified without a fully connected pathway, as such a pathway allowed for the residue (and its changing conformations) to be monitored continuously as NADH progressed along the pocket. In all, this analysis had identi-

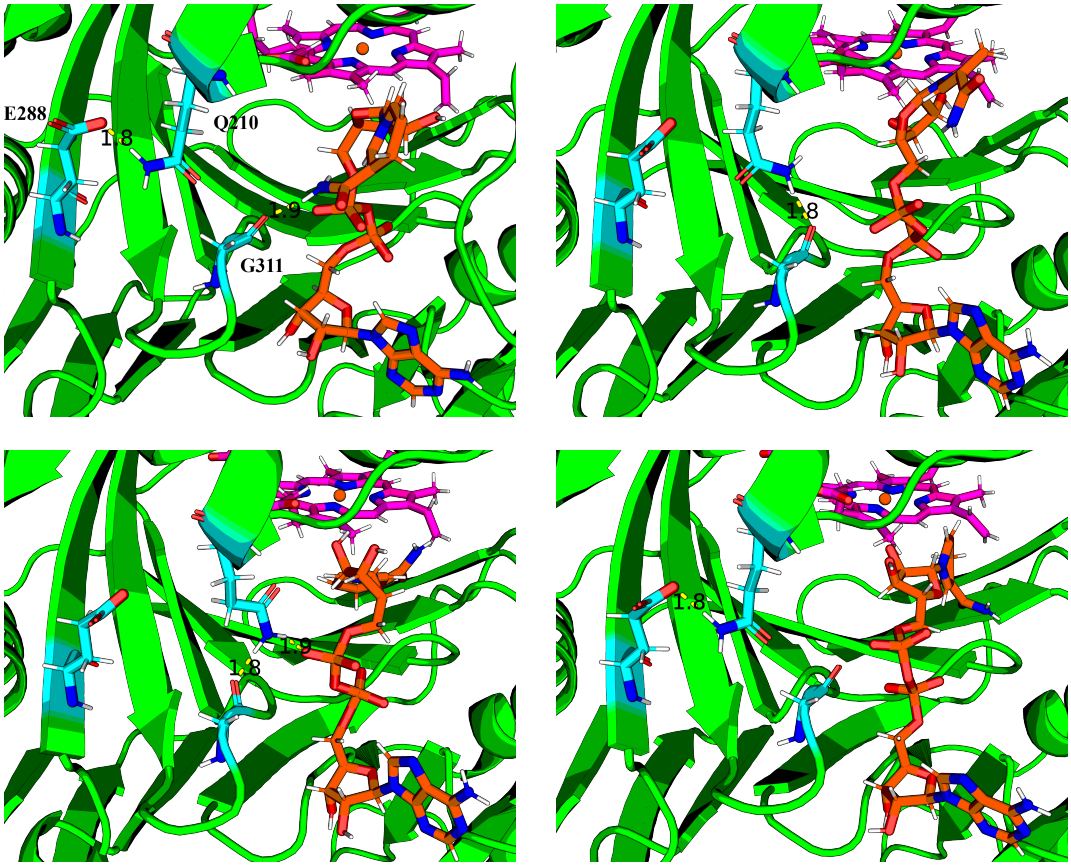


Figure 7.7: Selected minima highlighting the conformational changes of Q210 as NADH progresses through the pocket. Haem is represented with a magenta skeleton and NADH with an orange skeleton. Residues Q210 (centre, cyan), E288 (left, cyan) and G311 (right, cyan) are represented explicitly. Top left: NADH is folded and at the edge of the pocket. A hydrogen from the $-NH_2$ group of the nicotinamide hydrogen-bonds with the G311 backbone. Meanwhile, Q210 is hydrogen-bonded to E288. Top right: NADH has partially unfolded and progressed further along the pocket. Its nicotinamide is no longer hydrogen-bonded to G311. However, Q210 has swung around to hydrogen-bond with G311, thus abandoning E288. Bottom left: NADH has extended further into the pocket. Q210 has flipped over to another conformation, so that the opposite hydrogen of its $-NH_2$ group is now hydrogen-bonded to G311 from that which was bound previously. The hydrogen atom that had originally been hydrogen-bonded to G311 is therefore freed, and bonds to the diphosphate backbone of NADH. This presumably helps to stabilise NADH as it occupies a high-energy extended conformation, and perhaps to help it overcome some further barriers to achieve full extension. Bottom right: NADH is now fully extended within the pocket. Q210 has swung back round to hydrogen-bond to E288.

fied a set of residues (F104, F199, R209 and Q210) with interesting properties, which were thought to be suitable for parallel computational and experimental mutagenesis studies.

7.4 NADPH

Before examining any of the mutated (or homologue) systems, a scenario where NADH was replaced with NADPH was briefly considered.

Experimental studies had revealed that NADPH can react with haem to give the same haem breakdown product as NADH. In essence, this result is unsurprising given that both molecules have similar structures and are capable of donating hydrides. However, these two molecules tend to have different roles in the cell, with NADH primarily being used for catabolic purposes, and NADPH for anabolic purposes. Both NADH and NADPH are found almost ubiquitously in cells, but there are many cases in biology of an enzyme being specific for only one of these molecules. Typically, this specificity occurs when the extra phosphate group of NADPH (see Fig. 1.19) interferes within the binding pocket. Computational studies of the WT HemS system with NADH suggest that this extra phosphate group would point away from the pocket and is therefore unlikely to destabilise NADPH docking. Though CHECKSPMUTATE is capable of 'mutating' ligands (as long as the AMBER parameters for the new ligand are uploaded) as well as protein residues, the decision was made to not do a full study of the system with NADPH. However, a selection of stationary points were studied, and showed that the extra phosphate group does indeed point away from the pocket. In fact, this phosphate group was shown to be capable of forming a stabilising hydrogen-bond with the hydroxyl group of the T312 residue, perhaps indicating that NADPH is the natural ligand after all, and not NADH. This bond is shown in Fig. 7.8.

7.5 Mutant and Homologue Systems

7.5.1 Analysing the Databases

Using the WT HemS system as a template, databases were generated and grown for the F104A, F104AF199A, F104I, F199A, R209A, R209K and Q210A mutants, and HmuS, ChuS and ShuS homologues. The databases are at various stages of completion. Those databases describing the mutants can all be considered 'complete', as their disconnectivity graphs (shown in Figures 7.9 and 7.10) display clearly the gen-

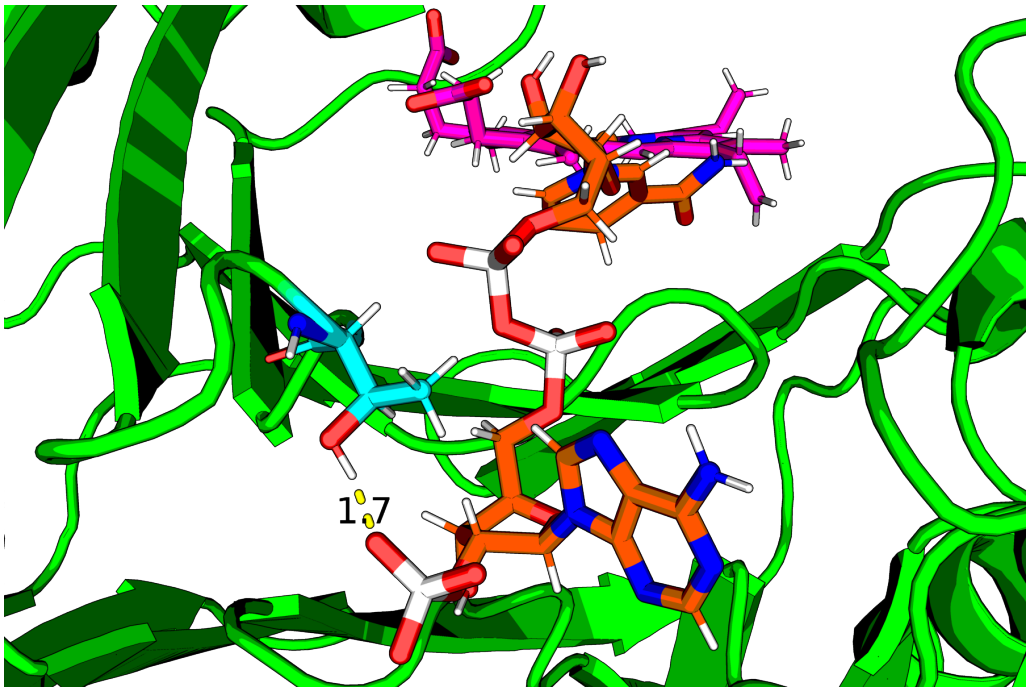


Figure 7.8: Selected minimum showing NADPH extended along the WT HemS pocket. The extra phosphate group NADPH has with respect to NADH is shown hydrogen-bonding to the T312 residue of HemS. This phenomenon suggests that NADPH is perhaps even more suited to docking inside the HemS pocket than NADH, suggesting that it is the natural ligand instead.

eral features of the system, plus the rate constants for the pathways investigated have converged.[c] The databases describing the homologues cannot be considered complete. In each case, the disconnectivity graphs, though showing many mis-assigned minima, are still able to reveal interesting general features of the homologues. However, successive attempts to refine the pathways of interest for these homologues were still giving large changes to the features of these pathways and their rate constants. Further sampling (beyond the scope of this project) will therefore be required to complete these databases.

The disconnectivity graphs for the mutants and homologues were each colour-coded according to the haem-NADH distance, as defined in Fig. 7.3 for the WT HemS system. Two minima are highlighted on each graph, labelled S for start, and F for finish, respectively. For the original, template WT HemS system (shown again in Fig. 7.9), S represents a manually selected minimum which is folded up and at the edge of the pocket. F, meanwhile, represents the minimum with the lowest NADH-

[c]However, it must be noted that the databases describing these systems cannot be considered as being truly complete, due to the very high degrees of freedom these systems each contain. Though this observation implies that there are many possible unsampled minima and transition states in each of these systems, the assumption is made that the refinement schemes were robust enough to identify those structures most relevant to the lowest energy pathways.

haem distance in the entire database, as defined by the β-*meso*-carbon of haem and the hydride-bearing carbon of NADH (consistent with the colouring scheme for the disconnectivity graphs). These two minima, S and F, were those that were selected for the fastest pathway analysis, as it was thought that they were suitable representative structures for the endpoints of the movement of NADH within the protein pocket.

The disconnectivity graphs for the mutants and homologues were also labelled S and F. The minimum representing S in each case was the reoptimised S minimum from the WT HemS system. It was not possible in every case for the minimum representing F from the WT HemS system to be successfully reoptimised following mutation. Therefore, F for each mutated or homologue system was the one with the lowest NADH-haem distance, as was used to define F in the WT HemS case. For some of these mutated systems, the minimum with the lowest NADH-haem distance coincided with that from the WT HemS system. As in the WT HemS case, these two S and F minima were those used to define the endpoints for any subsequent fastest pathway analysis.

Figures 7.12 and 7.13 show the fastest pathways calculated for the WT HemS and mutant systems studied. Due to their databases not being sufficiently refined, which can give rise to artificial barriers and kinetic traps, pathways for the homologues are not provided. For those pathways that are provided, the starting minima (point 0 along the integrated path length) and the finishing minima (the last point along the integrated path length for that system) correspond to the S and F minima in the disconnectivity graphs for their respective systems.

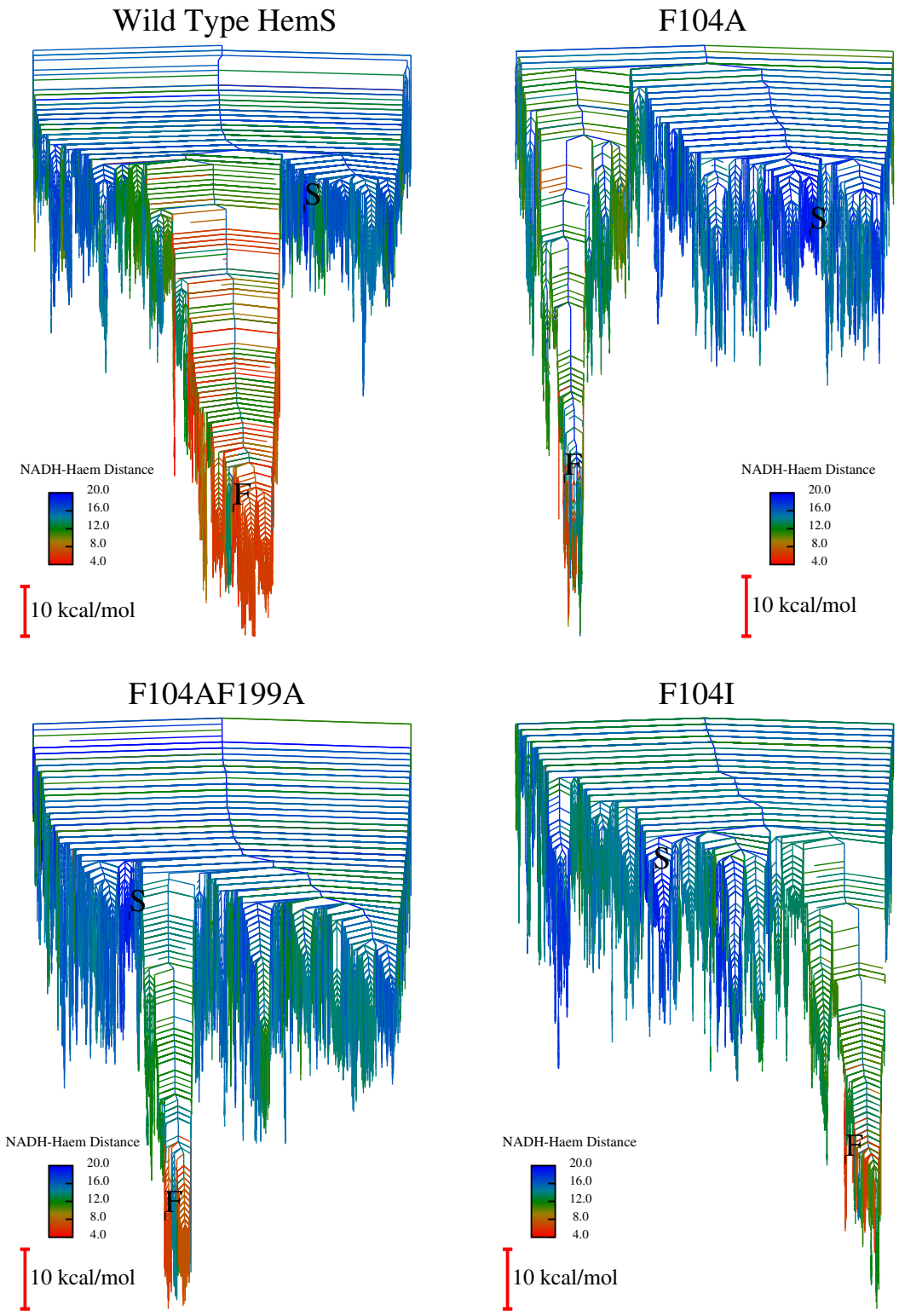


Figure 7.9: WT, F104A, F104AF199A and F104I HemS Disconnectivity Graphs.

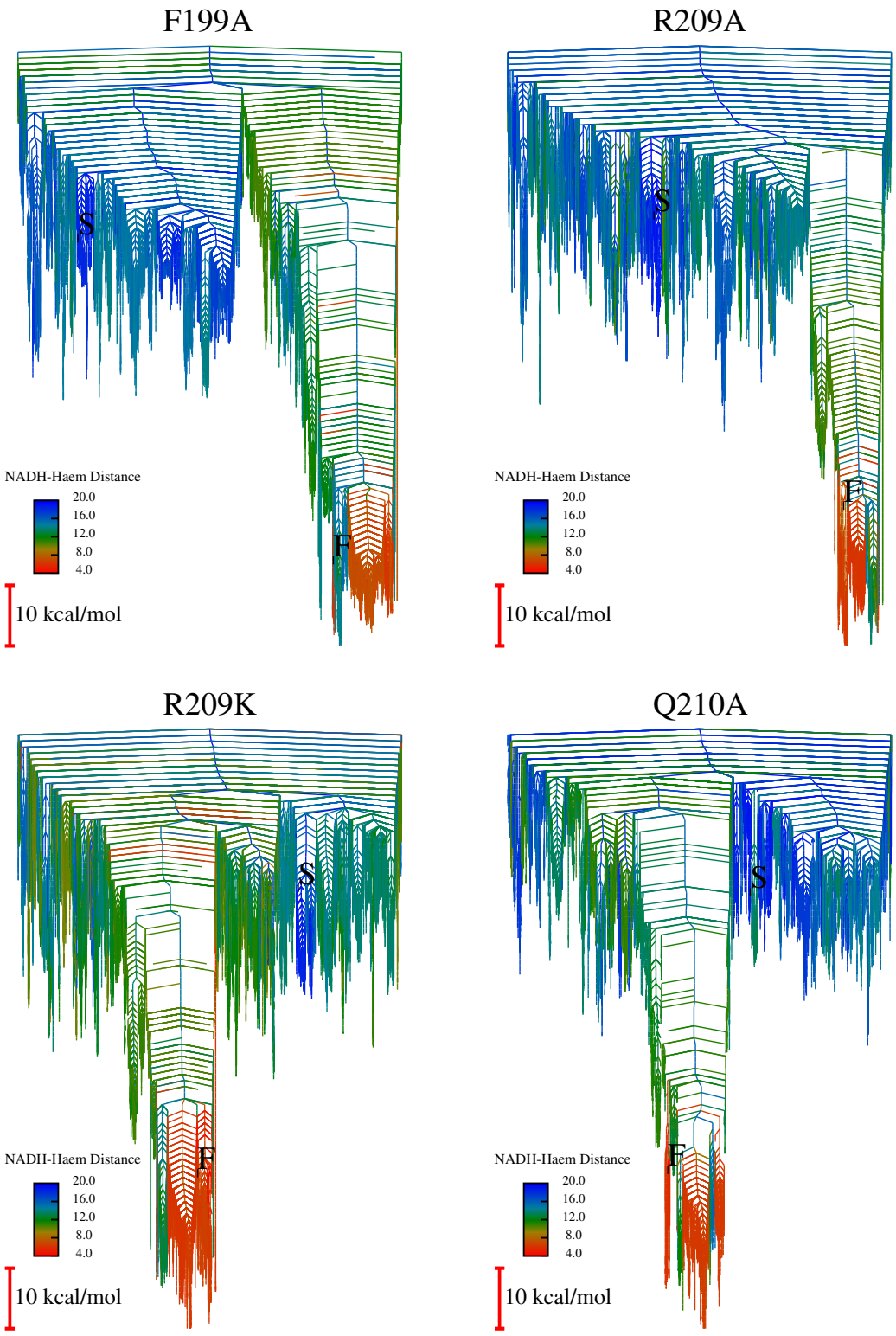


Figure 7.10: F199A, R209A, R209K and Q210A HemS Disconnectivity Graphs.

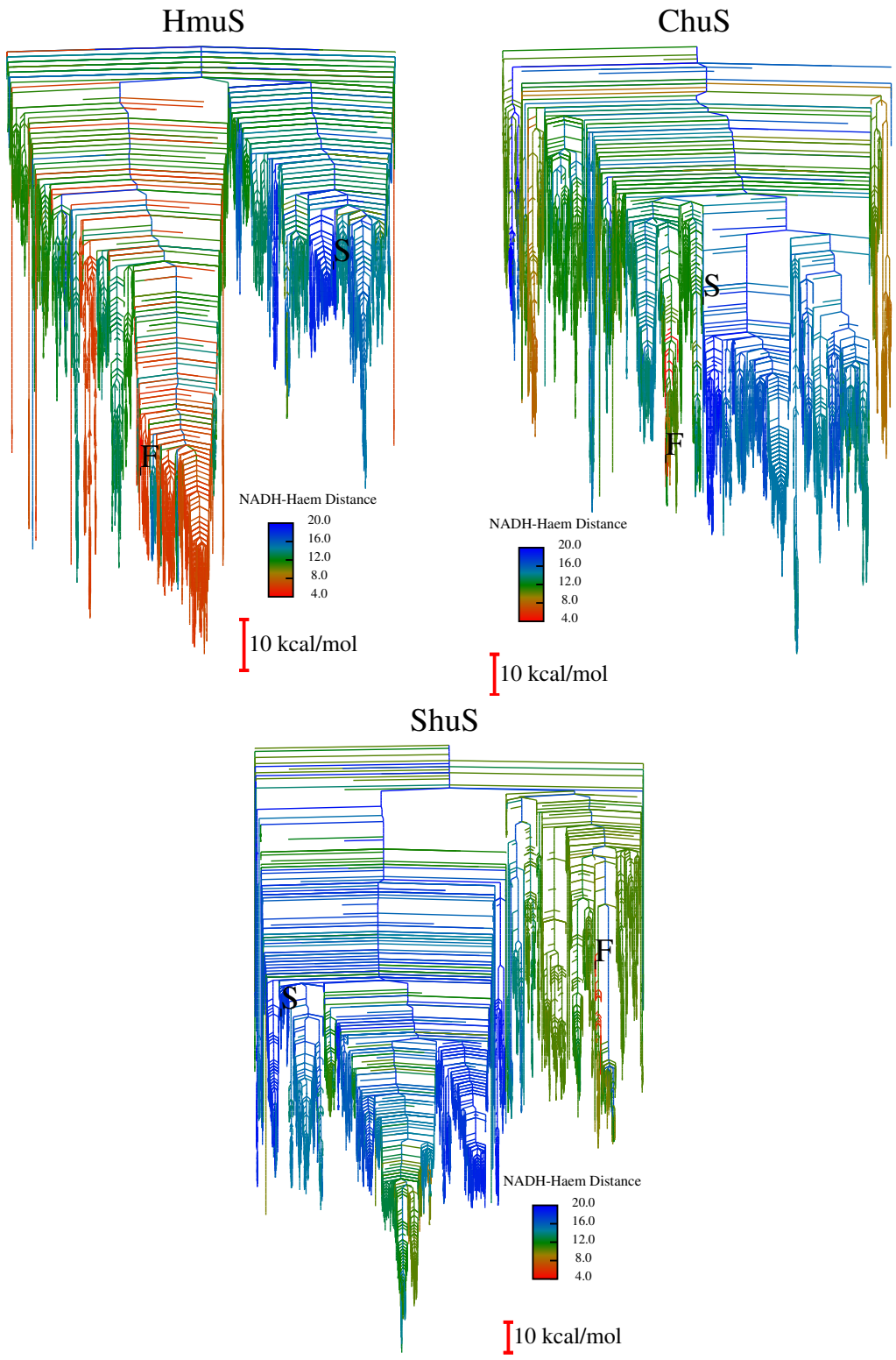


Figure 7.11: HmuS, ChuS and ShuS Disconnectivity Graphs.

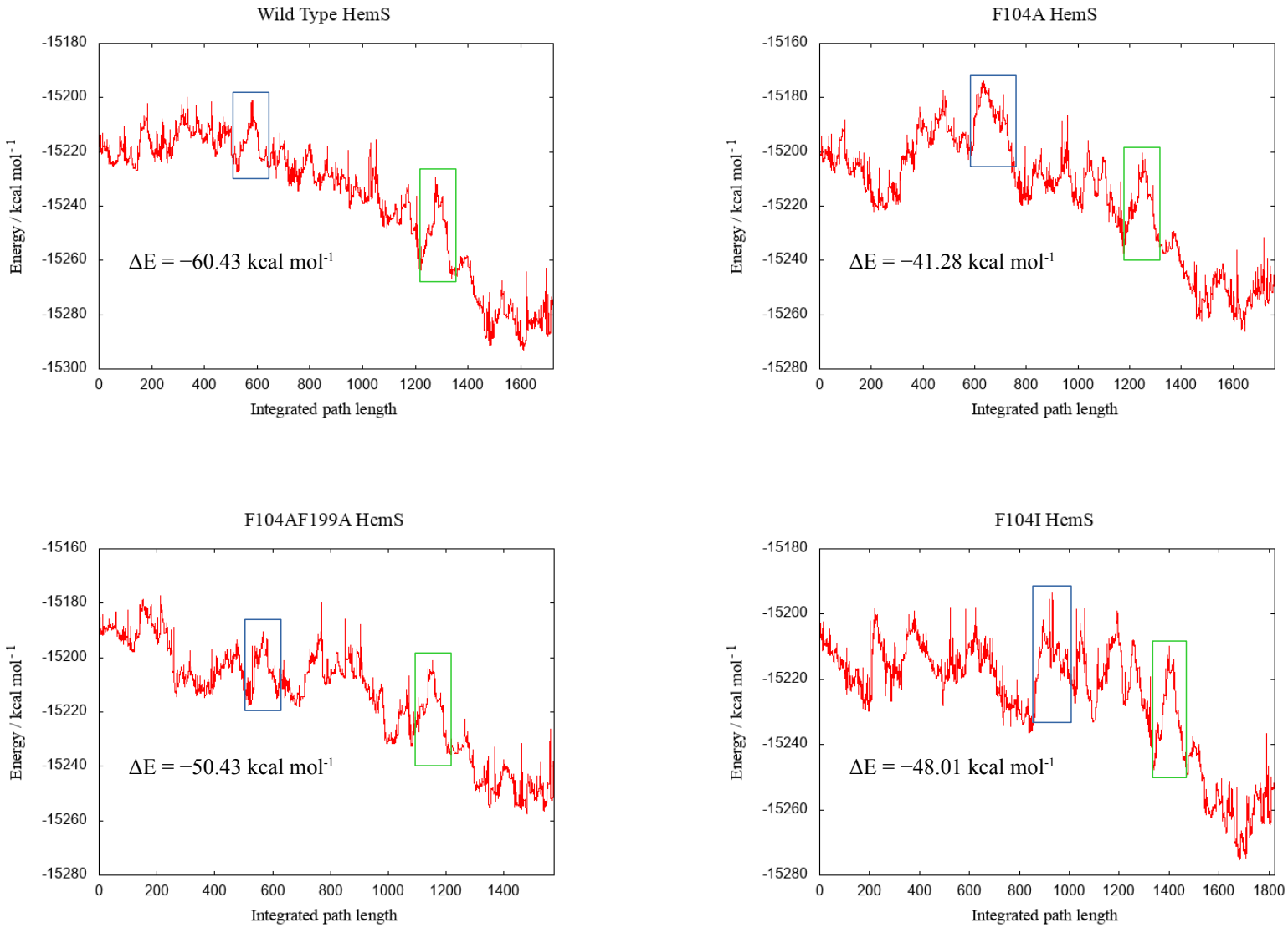


Figure 7.12: Dijkstra fastest pathways between minima S and F for WT, F104A, F104AF199A and F104I HemS. ΔE represents the overall potential energy change. Blue boxes highlight a region of NADH phosphate backbone twisting common to all systems, and green boxes highlight a region of NADH ribose twisting, also common to all systems.

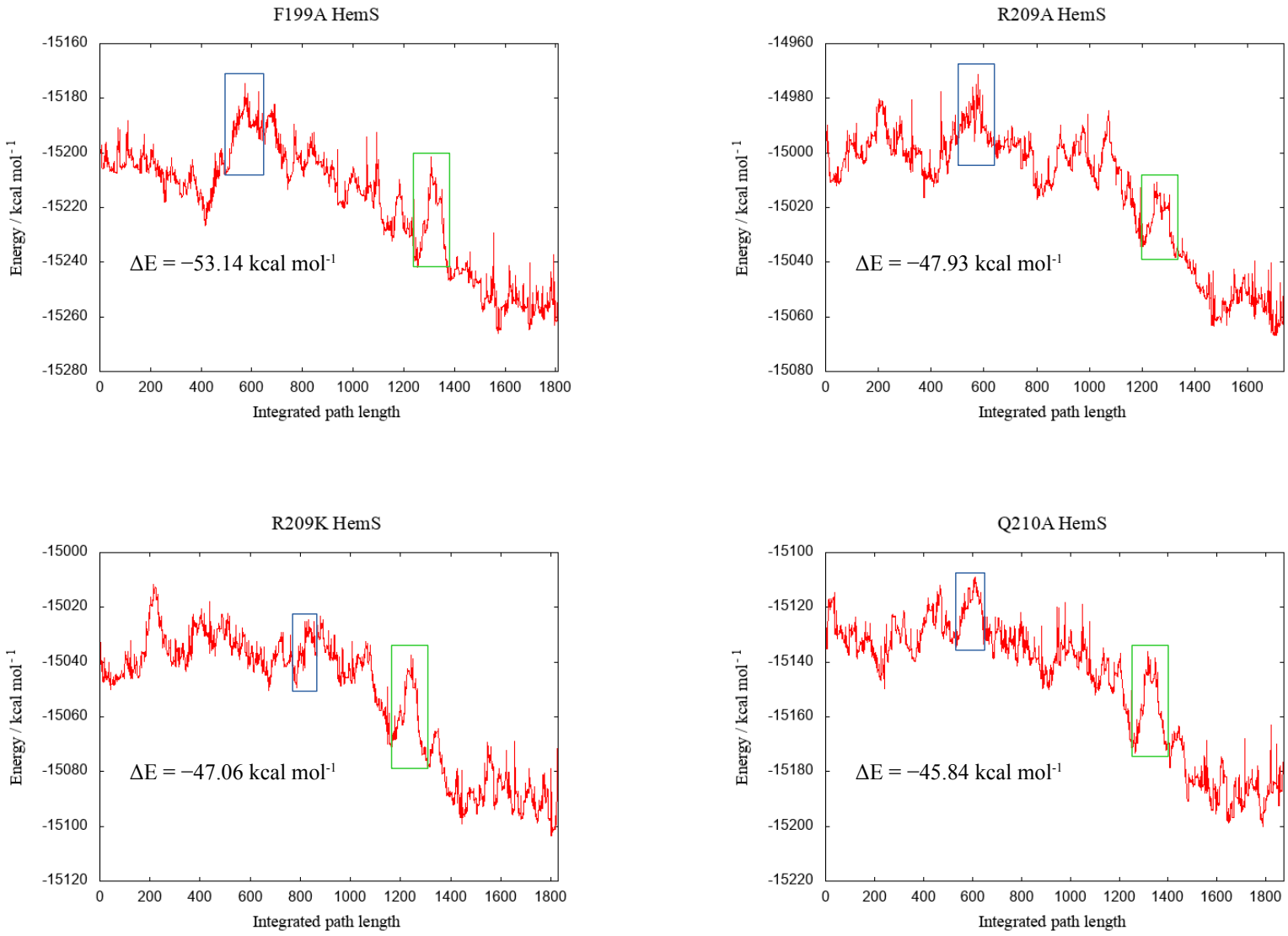


Figure 7.13: Dijkstra fastest pathways between minima S and F for F199A, R209A, R209K and Q210A HemS. ΔE represents the overall potential energy change. Blue boxes highlight a region of NADH phosphate backbone twisting common to all systems, and green boxes highlight a region of NADH ribose twisting, also common to all systems.

When its disconnectivity graph is compared against the other systems, it is apparent that the WT HemS system contains many more low-lying minima with short NADH-haem distances. Though it is not complete, it appears that the HmuS graph is the only other one that comes close to WT HemS in this regard. This result would suggest that the WT forms for both of these proteins are efficiently optimised to bring NADH deeper into the pocket, and that any single-point mutations can disrupt this fine balance. Analysis of the fastest pathways between WT HemS and its mutants supports this conclusion: the movement of NADH towards haem in the WT HemS pocket results in an overall potential energy change of $-60.43\,\mathrm{kcal\,mol^{-1}}$, which is a steeper decrease than for any of the mutants studied.

Though not complete, certain conclusions can be drawn from the ChuS and ShuS disconnectivity graphs. It is clear from each that the approach of NADH towards haem does not give a potential energy decrease to the same extent as any of the HemS systems, nor of HmuS. Indeed, in the case of ShuS, minimum F is actually higher in energy than minimum S. In both the ChuS and ShuS databases, there are few minima where the NADH-haem distance is small, suggesting that there is a relatively small drive for NADH to approach and break down haem in these proteins. The disconnectivity graph for ShuS is complete enough to show a funnel developing. It is noteworthy that this funnel is totally removed from the basin that contains minimum F, as it suggests that there is a different function this protein may be engaged in. Furthermore, the barrier to move from S to F (i.e. the conventional unfolding of NADH and its approach to haem) is very large, whereas to move from S to the most stable structure, it appears to be almost entirely downhill. The fastest pathway to move from S to this lowest energy minimum was therefore calculated. Rather than charting movement of NADH through the protein pocket, this pathway instead showed significant movement of the N–terminal α-helix, which opens up the smaller cavity. Snapshots of this pathway are shown in Fig. 7.14. The stabilisation brought about by this α-helix movement suggests that ShuS may be priming itself to bind ligands within this smaller cavity. Analysis with bioinformatics (see Section 8.4 in the following chapter) strongly suggests that the preferred ligand is double-stranded DNA.

The plotted pathways in Figures 7.12 and 7.13 show that the fastest pathway is roughly the same length between the WT and the mutants (ranging from approximately 1600–1900 stationary points). The shortest pathway is with F104AF199A. It would seem that the removal of the entire double phe-gate opens up the cavity such that NADH can move through it and approach haem more directly.

There are two highlighted regions in each of these plotted pathways, indicated by

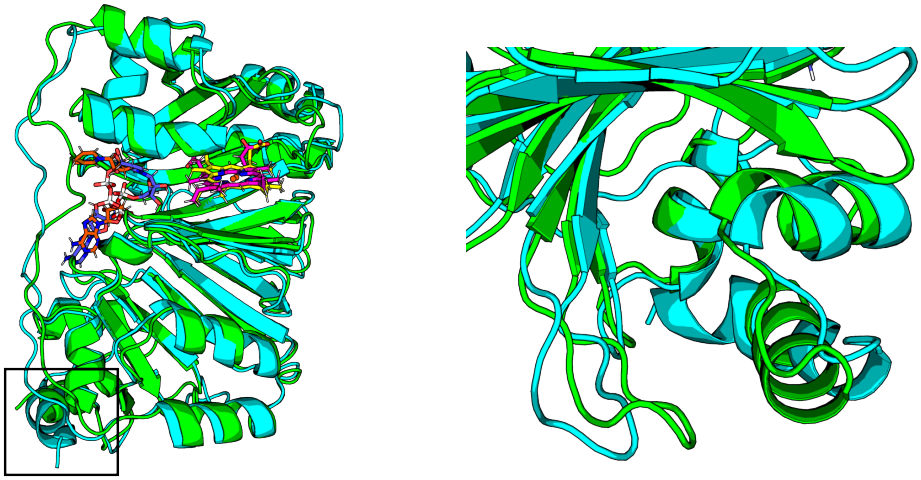
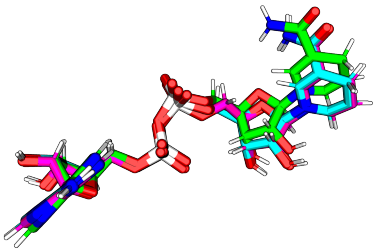


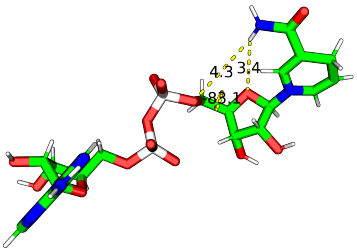
Figure 7.14: Overlay of the starting, S (green cartoon), and lowest energy, L (cyan cartoon), minima from the ShuS database. Left: The haem and NADH conformations forming part of the S minimum are colour-coded in magenta and orange, respectively. For the L minimum, haem and NADH are colour-coded in yellow and dark blue, respectively. In the L minimum, NADH is further inside the pocket, but the nicotinamide head is still folded back towards the phosphate backbone. The most significant region of structural change between the S and L minima is in the N-terminal α-helix, highlighted by a black box. Right: Reverse angle, magnified image of the entrance to the small cavity. This image shows that the movement of the N-terminal α-helix in the transition from the S minimum to the L minimum results in a significant enlargement of the entrance to the small cavity.

blue and green boxes, respectively. The blue boxes cover the region of the pathway in which a certain twisting motion of one of the NADH phosphate-ribose linking bonds takes place, and is highlighted because it is typically a high-energy process, which often results in it being the highest energy region across the entire pathway. All mutants retain this region, but the energy required to traverse it varies. For the F104A and F104I mutants, the overall barrier is particularly high, suggesting that the F104 residue is important for stabilising this high-energy backbone twist. The key stationary points involved in this twisting process are shown in Fig. 7.15, as demonstrated with the WT HemS system. This figure shows that, as the amide of the nicotinamide unfolds and thus moves away from the phosphate backbone, where it was engaged in a stabilising hydrogen-bond, the energy increases, giving the maximum along the region of the pathway encapsulated by the blue box. The system is then stabilised when the C11–C12 dihedral bond shifts from $-85.9°$ to $-177.8°$. None of the residues that were selected for mutation appear to be directly involved in this process.

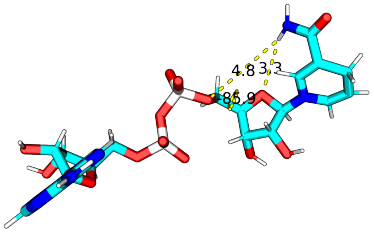
The other region common to all of the pathways, and highlighted inside a green box in Figures 7.12 and 7.13, is concerned with a change of conformation of the nicotinamide-bearing ribose. It occurs late on in the pathway in each case, where



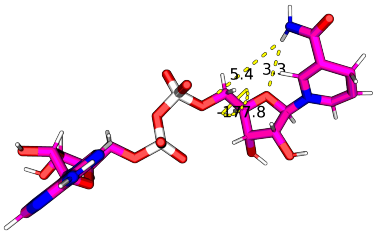
Overlaid Representations



Minimum 529: −15227.52 kcal mol^{-1}



TS 584: −15201.10 kcal mol^{-1}



Minimum 609: −15222.45 kcal mol^{-1}

Figure 7.15: Illustration of the high-energy NADH conformational change that occurs along the fastest pathway, as encapsulated by the blue box for the WT system in Fig. 7.12. The green representation is for the low-lying minimum at position 529 along the pathway before the barrier, the cyan representation for the highest transition state at position 584, and the magenta representation for the low-lying minimum at position 609 on the other side of the barrier. Top left: Overlay of these three stationary points. Top right: Minimum 529. One of the hydrogen atoms of the nicotinamide amide hydrogen-bonds to both the cyclic oxygen of the ribose and to the closest backbone phosphate oxygen. The bond lengths are 3.4 and 4.3 Å, respectively. The dihedral angle around the C11–C12 covalent bond (i.e. the bond linking the phosphate backbone to the nicotinamide-bearing ribose) is −83.1°. Bottom left: Transition state 584. The nicotinamide amide group has moved away from the backbone, increasing the bond length from the nicotinamide amide hydrogen to the closest backbone phosphate oxygen from 4.3 to 4.8 Å. This hydrogen-bond therefore becomes less stabilising. The C11–C12 dihedral angle has remained approximately the same (shifting from −83.1° to −85.9°) giving a high energy conformation. Bottom right: Minimum 609. Twisting this C11–C12 bond (changing the dihedral angle from −85.9° to −177.8°) stabilises the structure again.

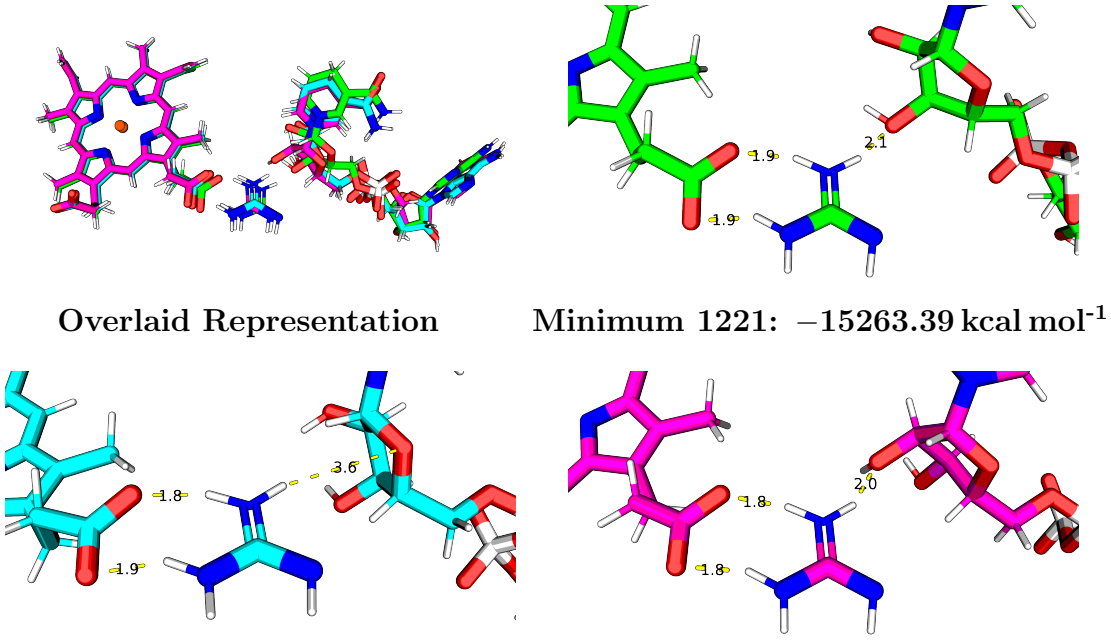
the overall energy of the system tends to be lower, and yet it constitutes a significant barrier. Residue R209 is closely involved with the process, binding alternatively to the C2′-hydroxyl oxygen and C3′-hydroxyl oxygen of the ribose as the NADH conformation changes. Mutation of this residue to a lysine (R209K) does not significantly change the shape or size of this barrier, but a mutation to alanine (R209A) was shown to reduce it significantly. R209 is a very well conserved residue (see Section 8.3 for details), and so despite the alanine residue providing a smoother alternative for this particular region of the fastest pathway, there must be other reasons for HemS and its homologues to retain it so consistently. Some possible reasons, such as stereospecific considerations, were suggested in Section 7.3. The conformational changes of NADH involved in this highlighted green region, along with the involvement of R209, are shown in Fig. 7.16, as demonstrated with the WT HemS system.

7.5.2 Lowest Energy Minima from Each Database

The disconnectivity graphs in Figures 7.9 and 7.10 show that the lowest energy minimum does not always have a short NADH-haem distance. Furthermore, this feature is typically true for those mutants where the double phe-gate has been mutated in some way. The lowest energy minima from each of these databases were extracted, and are shown in a reduced form in Fig. 7.17.

For those cases where the double phe-gate has not been mutated (WT, R209A, R209K and Q210A), atoms C5 (a methyl carbon) of haem and (*R*)-hydride of NADH are in close proximity, with the hydride pointing directly at this methyl carbon. Given these are the lowest energy minima in their respective databases, this result would suggest that these minima are the true endpoints of the movement of NADH through the HemS pocket, and that hydride transfer is to this methyl carbon of haem, rather than the β-*meso*-carbon, as previously assumed. There is no experimental evidence to definitively show which haem carbon is attacked. Were it to be *via* this methyl atom, then some sort of rearrangement would be required to bring about the cleaving of the ring at the β-*meso*-carbon position, which experiment has shown does occur. Furthermore, given the reaction can proceed with either the (*R*)- or the (*S*)-hydride of NADH (see Section 5.3) then an alternative mechanism whereby the (*S*)-hydride is brought into close proximity to this haem methyl carbon (of which there is little evidence in the databases studied) would be required.

When the double phe-gate is disrupted, the lowest energy structure can change dramatically. For F104A, F104I and F199A, these structures occur before NADH has passed through the gate. This information suggests that the double phe-gate does



Overlaid Representation **Minimum 1221: −15263.39 kcal mol^{-1}**

TS 1276: −15229.24 kcal mol^{-1} **Minimum 1329: −15262.52 kcal mol^{-1}**

Figure 7.16: Illustration of the high-barrier nicotinamide-bearing ribose conformational change which occurs along the fastest pathway, as encapsulated by the green box for the WT system in Fig. 7.12. Segments of haem, NADH and R209 (functional group only) are shown. The green representation is for the low-lying minimum at position 1221 along the pathway before the barrier, the cyan representation for the highest transition state at position 1276, and the magenta representation for the low-lying minimum at position 1329 on the other side of the barrier. Top left: Overlay of these three stationary points. Top right: Minimum 1221. Residue R209 is hydrogen-bonded to both of the oxygen atoms of the deeply buried propionate of haem. It is also hydrogen-bonded to the C3′-hydroxyl oxygen of the ribose. Bottom left: Transition state 1276. NADH has twisted away from R209, leaving the cyclic oxygen of the ribose closest. Bottom right: NADH has twisted further, resulting in the C2′-hydroxyl oxygen of the ribose hydrogen-bonding with residue R209 instead. This conformation is primed to unfold, with few large barriers to traverse along the remainder of the pathway. As part of the NADH unfolding process, R209 bonds again to the C3′-hydroxyl oxygen, as shown in Fig. 7.6.

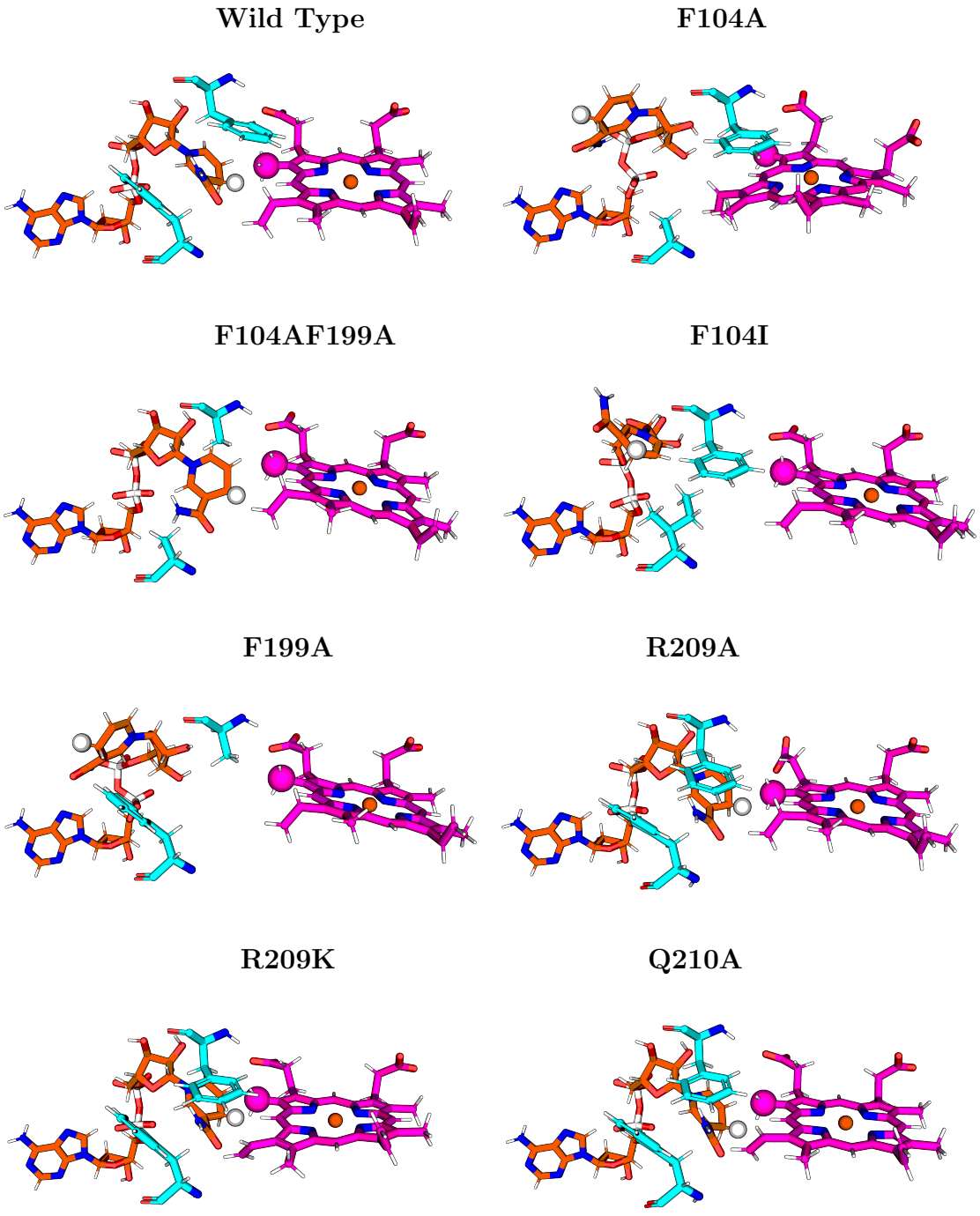


Figure 7.17: Lowest energy structures for the respective WT and mutated HemS systems, represented in a reduced form. Haem is represented in magenta, NADH in orange, and residues 104 and 199 in cyan. Methyl carbon C5 of haem (see Fig. 1.1 for a labelled diagram of the haem structure) and the (*R*)-hydride of NADH are shown as expanded spheres. In the WT, R209A, R209K and Q210A cases, these two atoms are in close proximity. Given that these are the lowest energy minima, this result suggests this location may be where the hydride transfer occurs. For F104A, F104I and F199A, the lowest energy structure occurs before NADH has passed through the gate, suggesting that these mutations give proteins that are not as suitable for facilitating hydride transfer.

indeed organise the energetics of the pocket to facilitate the controlled movement of NADH towards haem. Interestingly, the lowest energy structure of the F104AF199A double mutant shows NADH unfolded to the same extent as the cases where the double phe-gate is not disrupted at all, but that its hydride is not directly pointing towards the C5 methyl carbon of haem. This finding therefore suggests that the double phe-gate not only facilitates favourable energetics for NADH to unfold and approach haem, but that it also plays a role in orienting the nicotinamide head once in that close proximity.

7.6 Discussion and Summary

Computational work helped to deepen an understanding of the WT HemS system, as well as of selected mutants and homologues.

Taking advantage of GPU implementation, as well as development of the algorithms discussed in the previous chapter, the [WT HemS + Haem + NADH] database was grown to a much greater extent (approximately one thousandfold) than had previously been possible. Plotting this database as a disconnectivity graph revealed that the system exhibits funnel-like behaviour. Furthermore, Fig. 7.3 revealed that descent down the funnel correlates with a reduction of the NADH-haem distance. This structure implies that the approach of NADH towards haem within the pocket is a favourable process, strengthening the case that HemS is indeed an enzyme that has evolved to bring NADH and haem together in order to react. From this database, a fastest pathway could be identified, which clearly showed a favourable route with few large barriers for NADH to unfold and approach haem.

Having a large database also allowed for the behaviour of the double phe-gate to be investigated more fully. This analysis showed that, in the absence of NADH, the gate is typically fully open. However, once NADH is introduced at the edge of the pocket, F104 flips to a closed conformation, resulting in a T-shaped π-π bonding interaction between the two gates. Only once NADH is in close proximity to these gates does F104 open up again in order to allow it to pass through to haem. This behaviour is due to a series of intramolecular residue interactions instigated by the presence of NADH, as described in Fig. 7.5. It seems that the overarching purpose of these interactions is to allow NADH to be held in place effectively throughout its time inside the pocket.

The double phe-gate was clearly of importance to the reaction and so residues F104 and F199 were chosen for mutagenesis studies. Two other residues were also selected. The first, R209, was chosen because of its ability to bind to both NADH

and haem when they are in close proximity. It was therefore thought that this residue may be important in bringing the two ligands close to one another, or orienting them in a specific way. The second was Q210, which was selected due to the significant conformational changes it engages in, seemingly to stabilise the phosphate backbone of NADH at certain stages along the pathway.

Before engaging in mutagenesis, a short study on NADPH was performed, with the aim being to explain the experimental finding that both NADH and NADPH can engage in the breakdown of haem. This study showed that the extra phosphate group of NADPH points away from the centre of the pocket, thus explaining why both of these ligands can access haem and react in the same way. It was further shown that this extra phosphate group can hydrogen-bond to residue T312, suggesting that NADPH association with HemS is more stabilising, and it is therefore more likely to be the natural ligand, compared to NADH.

Using **CHECKSPMUTATE** and the WT HemS system as a template, databases were created and refined for the F104A, F104AF199A, F104I, F199A, R209A, R209K and Q210A mutants, and HmuS, ChuS and ShuS homologues. The databases describing the mutants can all be considered as being complete. The disconnectivity graphs arising from these databases (see Figures 7.9 and 7.10) revealed that each had a far smaller proportion of minima with low NADH-haem distances compared to the WT, suggesting they were less capable of directing NADH towards haem. The mutations to F104 and F199 (other than F104AF199A) were shown to be particularly disruptive, where the lowest energy minima showed NADH still partially folded up, with the nicotinamide head pointing away from haem.

The Dijkstra fastest pathway between two minima specially selected to demonstrate NADH unfolding and approach to haem was derived from each database describing a mutant, and compared against the WT pathway. These pathways showed that the change in potential energy, ΔE, was negative for each mutant, but not by as much as for the WT. The two most prominent barriers in the WT were highlighted, and the effect of each mutation on these barriers investigated. The barrier earlier in the pathway pertained to a twisting motion of the NADH phosphate-ribose linking bond. This barrier grew significantly when F104 was mutated, demonstrating that this phe-gate is important for stabilising the dihedral changes required of NADH to allow it to unfold. The second highlighted barrier was concerned with the conformation of the nicotinamide-bearing ribose, a process which a mutation to the R209 residue disrupts.

Though they were not entirely complete, certain conclusions could be drawn from the databases describing HmuS, ChuS and ShuS. The disconnectivity graph

for HmuS looked very similar to the one for WT HemS, containing a wide funnel with many minima where the two ligands are in close proximity, suggesting that these systems behave in a similar manner. The disconnectivity graphs for ChuS and ShuS, however, were entirely different. Neither had an obvious funnel, and very few minima where NADH and haem were close. The appearance of these graphs strongly suggests that these two proteins are not optimised to facilitate the reaction of NADH with haem. This structure does not preclude their ability to perform this reaction (as shall be demonstrated in Chapter 9), but it does suggest these proteins have different possible functions. Analysis of the lowest energy structures of the ShuS database gave an indication as to what this alternative function could be. These structures showed the N-terminal α-helix in a markedly different position from where it is typically located, with this movement corresponding to an opening up of the smaller cavity. The following chapter shall show, *via* bioinformatics, why this is likely to be part of an alternative DNA-binding function.

Overall, computational analysis therefore provided a number of predictions to be tested in the laboratory. All of the mutants were demonstrated to lessen, but most likely not entirely prevent, the ability of NADH to access haem. It also showed that, of the homologues, HmuS was most likely to engage in this NADH-dependent haem breakdown reaction but that ChuS and ShuS might be more limited by competing alternative functions.

Chapter 8

Bioinformatic Study of HemS Homologues

The bioinformatics work in this chapter was undertaken by Yuhang Xie (King's College, Cambridge, 2020-21) as part of his Master's project, under the day-to-day supervision of the author. The following discussion is original to the author apart from the figures and passages highlighted.

8.1 Aims

HmuS, ChuS and ShuS had been chosen as alternative proteins to HemS for computational study, as they were the best characterised in the literature. However, these homologues together only provide a small snapshot of the overall phylogenetic context of HemS. Given the advances in the field of bioinformatics since Choy's original study,[21] it was reasoned that a fresh study could shed new light. With proteomic databases having expanded significantly, it was possible new proteins had been identified with NADH-binding homology to HemS. A further aim was to identify a consensus sequence within the HemS family and determine whether the residues selected for mutation are conserved. Due to the reported DNA-binding properties of ShuS and PhuS (see Sections 1.4.5 and 1.4.6 in the Introduction, respectively) an investigation of possible DNA-binding sites was also conducted. To summarise, the aims of the bioinformatic investigation were:

1. To investigate the phylogenetic context of HemS.
2. To search for newly-discovered proteins in the literature with possible links to NADH metabolism or NADH binding modes homologous to those in HemS.

3. To identify a consensus sequence for HemS and use this information to determine whether the residues predicted by computation to be important for NADH-binding are conserved.

4. To investigate possible regions for DNA-binding in HemS and its homologues.

Outcomes from this research are described in the following sections.

8.2 Phylogenetics

Xie began by placing HemS within its phylogenetic context. He amassed homologous sequences of HemS using the default settings in pBLAST,[241] except that the maximum number of sequences was increased to 5000. The sequences identified came from a total of 218 different genera. One sequence was manually selected from each of these genera, and a maximum-likelihood phylogenetic tree created in MEGA-X[242] using default settings, except that the number of bootstrap replications was increased to 2000. iTOL[243] was then used to visualise, manipulate and export the tree. This tree is provided in Fig. 8.1.

Fig. 8.1 clearly shows that HemS and HmuS are closely related, as are ChuS and ShuS. This result is not surprising, considering that HemS/HmuS both belong to the *Yersinia* genus, and ChuS/ShuS respectively belong to *Escherichia* and *Shigella*, which are, as noted in Section 1.3.5.1, arguably the same genus. Interestingly, though, these HemS/HmuS and ChuS/ShuS pairs are shown to be only distantly related to one another, with many intervening genera between them. The same is true for PhuS, which is shown to be only distantly related to both the HemS/HmuS and ChuS/ShuS pairs. This situation serves to emphasise the sheer number of bacterial species which contain a version of this protein. Those that appear in genera intervening between the HemS/HmuS and ChuS/ShuS pairs can be reasonably assumed to engage in anaerobic haem breakdown, since all four of these proteins have been shown to catalyse this reaction with NADH (see the next chapter for details).

Xie further noted that these homologues originated from both pathogenic and non-pathogenic bacteria, suggesting these proteins (and the operons they form part of) do not necessarily adversely affect the host. Xie also noted that all 5000 of the homologues were listed as either haem transport or hemin degradation factors. It is interesting that so many of these homologues have been listed as haem degrading enzymes, without any apparent experimental evidence. This classification was probably the result of Stojiljkovic & Hantke's original investigations[60] into the *hem*

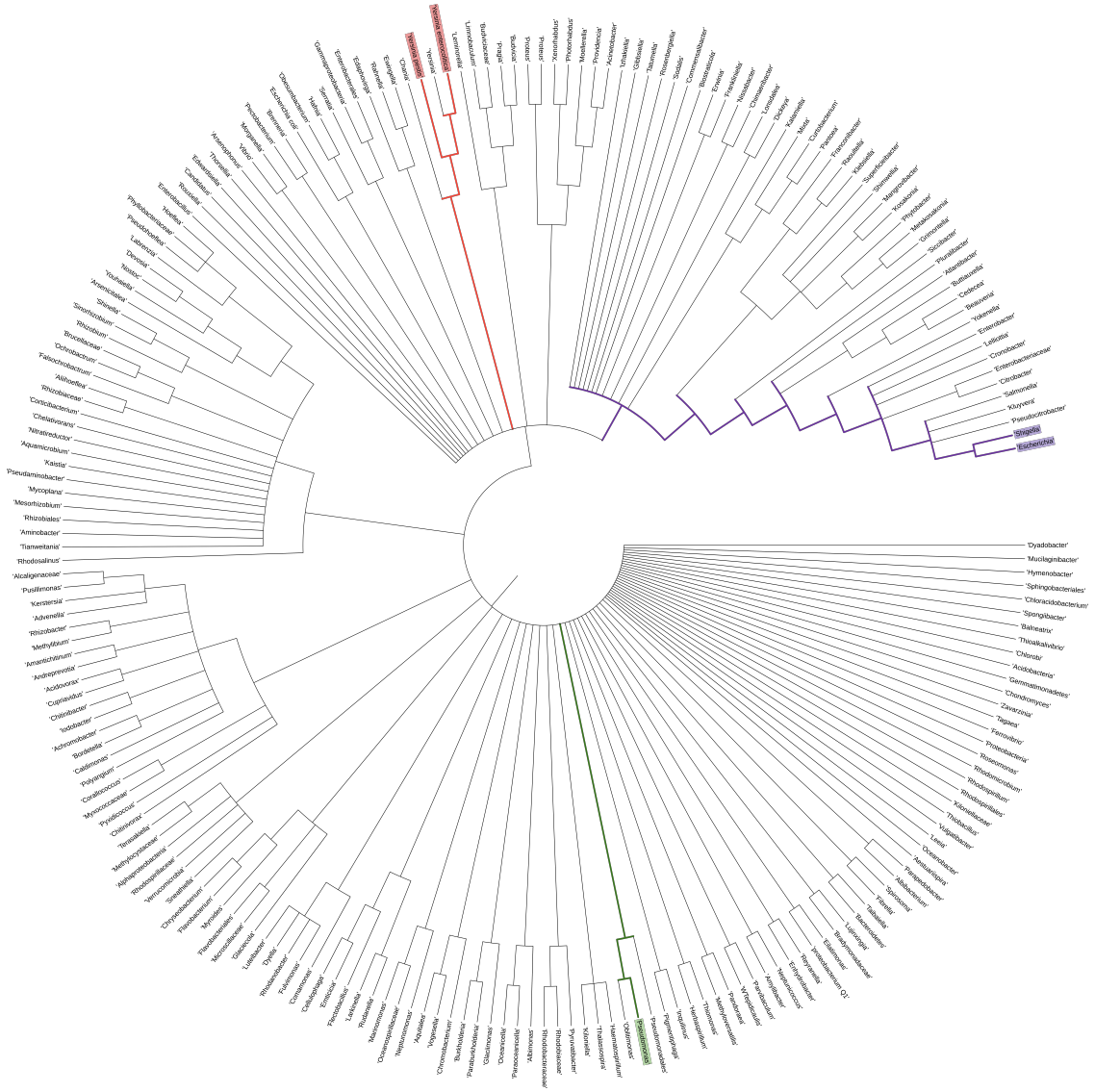


Figure 8.1: Maximum likelihood phylogenetic tree for HemS and its homologues. 218 different genera are represented, including both pathogenic and non-pathogenic bacteria. Selected species are highlighted. Red: *Yersinia enterocolitica* (HemS), *Yersinia pestis* (HmuS). Purple: *Escherichia coli* (ChuS), *Shigella dysenteriae* (ShuS). Green: *Pseudomonas aeruginosa* (PhuS). A fuller version, which includes the Accession Numbers for each sequence and the bootstrap values for the nodes, is included in Appendix E. Figure reproduced from Xie,[81] with minor adaptations.

operon, where they speculated that HemS itself could be a haem degrading enzyme (see Section 1.4.1). It is therefore satisfying that such a haem breakdown process has now been identified in at least some of these homologues (i.e. HemS, HmuS, ChuS and ShuS), using NADH as a hydride donor.

This result is remarkable because none of the homologues gathered by Xie are annotated as being able to bind NAD(P)H. This observation implies that NADH binds in the HemS pocket in a manner not seen before, as Choy had inferred when his Relibase$^+$ searches only gave 'weak' hits. None of the homologues were known or speculated to bind to nucleic acids either (apart from ShuS and PhuS), a feature which shall be of interest in Section 8.4.

8.3 Sequence Analysis and Conservation

The 218 sequences selected to construct the phylogenetic tree were aligned, and the consensus sequence taken to be 345 residues long. This alignment revealed that 177 of the 345 residues were conserved where the consensus threshold was set to 50%, that 105 were conserved where the threshold was set to 70%, and that 59 were conserved where the threshold was set to 90%. Of those very highly conserved residues, 44.1% were involved in the large cavity and 27.1% in the small cavity. Fig. 8.2 shows the residues which form the large and small cavities respectively, as assigned by MetaPocket.[244].

Those residues that were suspected to be involved in haem-binding (see Fig. 1.14) showed remarkable conservation. R102, H196 and R209 all showed 100% conservation[a] across the 218 homologues, and K294 and R321 98.1% and 99.1%, respectively. The residue at 318 showed 69.7% conservation for phenylalanine, although HemS has a tyrosine at this position. Tyrosine and phenylalanine are both aromatic residues, and so presumably this change does not significantly alter the haem-binding properties. Together, they account for 97.2% of residues at this position. Q316, meanwhile, is not as strictly conserved across the homologues, at 71.1%.

Those residues identified by Choy[21] as possibly being involved in NADH-binding (see Fig. 1.26) did not show as high levels of conservation. These, respectively, were: Q132, 29.8%; S171, 4.1%; K203, 43.1%; R250, 61.5%; and T312, 30.7%. The fact that conservation of haem-binding residues is so much higher than the conservation of the purported NADH-binding residues suggests that all of the homologues are involved in haem-binding, but not all of them are haem degrading enzymes, or at

[a]Unless otherwise stated, all % conservation values listed are % identities rather than % similarities.

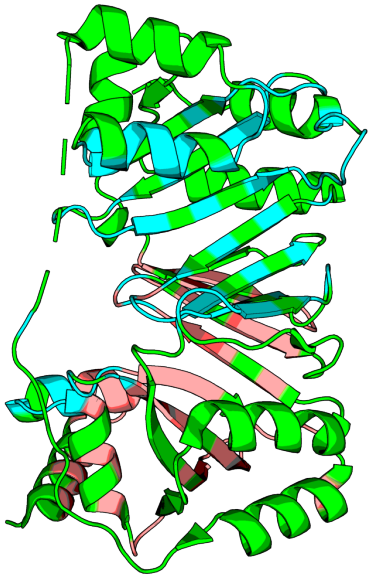


Figure 8.2: PyMOL representation of HemS, using structure 2J0R[95] from the Protein Data Bank. Residues were assigned to the large or small cavities using the default values in MetaPocket.[244] Those residues determined to be part of the large cavity are highlighted in cyan, and those determined to be part of the small cavity are highlighted in salmon. Where MetaPocket assigned a residue to both cavities, the residue was manually assigned to the cavity where its ranking was highest.

least not all of them have haem degradation as their primary function.

It was also important to note the extent of conservation of the residues, which had become the focus of both the computational and experimental investigations in this thesis. The first, R209, which was already mentioned above as a haem-binding residue, is 100% conserved, suggesting that it is important at least for binding haem, if not for enzymatic activity. Q210 is only conserved in 42.2% of the homologues, suggesting it is not as important for protein function. This conclusion assumes, however, that all of the homologues are designed to bind to NADH. If this is not the case, then it is not particularly surprising that Q210, a residue predicted to bind to NADH, and not to haem, is not as strictly conserved. Importantly, this residue is conserved in HmuS, ChuS and ShuS, which were all shown to be capable of utilising NADH to break down haem (see Section 9.4 in the next chapter for details).

Considering that they are not suspected to be directly involved in binding to haem or NADH, the two phenylalanine residues that make up the double-phe gate are remarkably well conserved. F104 is conserved across 98.6% of the homologues, and F199 across 82.6% of them (rising to 90.8% when the other possible aromatic residues, histidine and tyrosine, are included). This conservation lends weight to the notion that these two residues are performing some sort of regulatory function

in controlling access of other ligands to haem. It is thought that the large cavity of HemS may have been the result of two smaller cavities merging in an ancestor, and that the two phenylalanines are situated where those cavities merged. Their presence at the centre of the pocket may therefore have originated as some sort of evolutionary 'accident'. However, their remarkably high rates of conservation would suggest that these two residues then became essential to protein function.

8.4 DNA-Binding Exhibited by Some Homologues

Sections 1.4.5 and 1.4.6 described the DNA-binding properties of ShuS and PhuS, and their link to iron and haem regulation, as discovered by the Wilks group. As far as the author is aware, no DNA-binding studies have been carried out on any of the other homologues (including HemS itself, HmuS and ChuS). No such studies were carried out in this work either. However, it was found that the expression and purification of ShuS (to be described in the next chapter) required DNase/RNase treatment while lysing the cells. The presumption is that this requirement was because DNA was binding to ShuS and causing it to aggregate, a phenomenon not found in any of the other homologues.

It was remarkable that ShuS required this treatment with benzonase nuclease, and yet ChuS did not. The two homologues have 98.5% sequence identity, varying in only five of their residues. Xie therefore used bioinformatics to investigate why two such similar proteins have such apparently different DNA-binding characteristics.

Xie first showed that ChuS and ShuS have very similar predicted isoelectric points (5.5 and 5.7, respectively) and somewhat similar net charges at pH 7.4 (-12.7 and -10.7, respectively).[81]

The study was therefore extended to determine whether there were any specific areas of the respective homologue surfaces that would be suitable for DNA-binding. Coulombic surface charge calculations were carried out on the *apo*-forms of ChuS, ShuS, HemS and PhuS.[b] These calculations showed that both the large and the small cavity (as shown in Fig. 1.13) are cationic in contrast to the rest of the surface, which was either anionic or neutral, making them prime candidates for binding to the highly anionic DNA. A charged surface for ShuS is shown in Fig. 8.3.

The DNAbind package[246] was then used to predict the number of DNA-binding

[b]Surfaces could readily be derived for ChuS, HemS and PhuS from their published crystal structures. HmuS was not included in this study because no such crystal structure was readily available. The same was true for ShuS, but because a comparison between the DNA-binding properties of ChuS and ShuS was sought after, a ShuS structure was modelled upon the ChuS structure using SwissModel.[245]

residues for each homologue, estimating ten apiece for ShuS and ChuS, eight for HemS and seventeen for PhuS. For HemS, these residues were spaced far apart at seemingly random positions of the surface, suggesting that HemS did not have a contiguous DNA-binding surface. For PhuS, the binding residues clustered around the large cavity (suggesting that is where DNA would bind), whereas for ShuS and ChuS the residues clustered around the small cavity. This clustering perhaps explains the laboratory-based findings by the Wilks group, where they determined that PhuS is sequence specific, but ShuS is not.[102;112] In addition to this specificity, PhuS can readily dimerise, and it is apparent from structural studies that the small cavity of PhuS is involved in this.[111] DNA-binding at this smaller cavity would interfere with this dimerisation process.

None of the predicted binding residues (Y18, R20, R29, R67, T205, R206, F243, G245, N246 and R247) overlap with the five residues that differ between ChuS and ShuS (ChuS/ShuS: A/V19, H/R46, T/A78, S/R123 and T/M151). However, two of those five residues are located close together in the small cavity. One of those residues is at position 123 and is a serine in ChuS but an arginine in ShuS. As serine is a polar, neutral residue but arginine is cationic, this mutation may explain why the ChuS pocket does not bind to anionic DNA but ShuS does. The other residue close by in this cavity, which is different between the two homologues, is at position 19. In ChuS, it is an alanine, but in ShuS it is a valine. Both are hydrophobic residues, but valine has significantly more steric bulk, which allows the valine in ShuS to interact with R123, whereas the alanine in ChuS is too far away to do the same with S123, as shown in Fig. 8.3. Having R123 anchored to the centre of the cavity may assist ShuS to bind to DNA, although further studies would be required to determine if this is indeed the case.

Wilks had noted that both ShuS and PhuS can bind to DNA when in their *apo*-forms but not when in their *holo*-forms.[103;112] Assuming it is true that DNA binds to the large cavity of PhuS, this behaviour would therefore seem to be a simple case of direct competition between DNA and haem. However, assuming that DNA binds to the small cavity of ShuS, there would be no such direct competition to explain this phenomenon. Rather, for ShuS, it may be that certain conformational changes induced by haem-binding could preclude the small cavity from binding to DNA. To date, there are no crystal structures published for ShuS. Wilks was able to infer from circular dichroism that the secondary structure of ShuS does not change significantly with haem-binding.[102] Crystal structures for ChuS are available,[80;98] which suggest minimal structural changes once haem binds. Haem docking to ChuS does, however, result in a subtle movement of the very first α-helix at the N-terminus, which serves

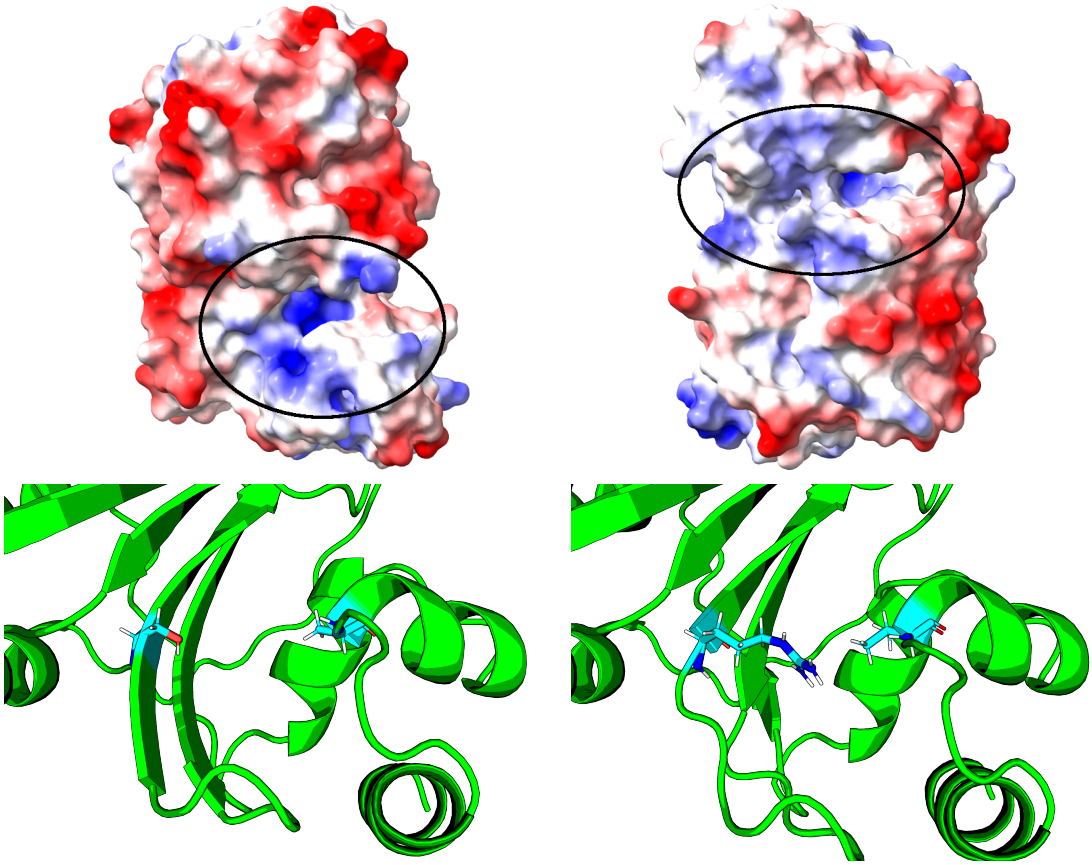


Figure 8.3: DNA-binding in ChuS and ShuS. Top left: Charged surface representation of ShuS, calculated using Chimera.[247] Red represents regions of negative charge, blue positive regions and white neutral regions. The small cavity is circled. Top right: Same structure, rotated 180°. The large cavity is circled. Bottom left: Small cavity of ChuS. Residues A19 and S123 are shown in cyan. Bottom right: Small cavity of ShuS. Residues V19 and R123 are highlighted in cyan. Both residues are larger than their equivalents in ChuS. It is perhaps the case that V19 anchors R123 in position, in a suitable position to bind to DNA.

to narrow the small cavity. This movement of the N-terminal α-helix reflects certain computational findings regarding ShuS, which were expressed in Section 7.5.1 of the previous chapter. Experiments are needed to investigate this effect further, but this conformational change caused by haem-binding may be enough to prevent DNA from binding to ShuS.

The fact that DNA can bind to the *apo*-forms of PhuS and ShuS, but not the *holo*-forms, points to another aspect of binding, which DNAbind unfortunately cannot capture. Namely, these predictions assume that DNA-binding itself does not lead to significant changes in the protein structure. However, this assumption is not reflected by the experiments carried out by the Wilks group, which found that DNA-binding to ShuS led to significant aggregation.[103] The author experienced this aggregation as well when trying to lyse cells containing ShuS without DNase. It is perhaps the case that dimerisation/oligomerisation of ShuS gives a larger contiguous cationic surface capable of binding DNA.[81] Presumably, if haem is already bound to ShuS, the protein is afforded a certain structural integrity, which would prevent this collapse to a protein-DNA aggregate. This is perhaps, therefore, an alternative explanation for why *apo*-ShuS can bind to DNA, but *holo*-ShuS cannot.

The Wilks group have provided strong evidence that PhuS binds to the *prrF1* promoter, which regulates haem flux.[112] However, ShuS has only been shown to bind DNA non-specifically. Furthermore, to reiterate, the research by Xie has suggested that PhuS and ShuS bind to DNA at different cavities. It is even possible that ShuS is intended to bind ssDNA or RNA rather than dsDNA.

Both experimental research by the author and bioinformatic studies by Xie suggest there is no obvious binding site for DNA in HemS or ChuS. Further research into this interesting area is required to determine further the DNA-binding capabilities of these homologues.

8.5 Discussion and Summary

This chapter used recent advances in bioinformatics to study the phylogenetic context of HemS.

Using pBLAST, the 5000 closest sequences to HemS were identified. These sequences came from a total of 218 different genera, including both pathogenic and non-pathogenic varieties. This diversity suggests that these proteins are not necessarily, or solely, for the extraction and breakdown of haem from host species.

Of the 5000 sequences investigated, none had been reported as capable of catalysing the reductive degradation of haem. The Jia group had identified CPR-NADPH

as a possible oxidative agent for haem breakdown in ChuS, but the Wilks group, in similar studies on ShuS and PhuS, suggested this observation may just be a coupled oxidation reaction (see Section 1.4.6 for details). Other than these studies, it does not appear that NAD(P)H interactions with any of the homologues have been investigated. Coupled with the weak hits Choy received when using Relibase$^+$ to investigate NADH-binding in the HemS pocket, this result implies that the method of NAD(P)H-binding inside HemS, and within a certain number of its homologues, is novel.

Analysis of the 218 sequences selected from the different genera revealed remarkable conservation of the haem-binding residues, therefore indicating that all of the homologues bind haem. However, the conservation of the NADH-binding residues was not as high, suggesting that not all of the homologues are haem breakdown enzymes. The conservation of F104 and F199, the two phenylalanine residues involved in the double phe-gate, was very high. It is interesting that these residues, which were thought to be important for controlling NADH access to haem, are conserved to a far higher extent than the supposed NADH-binding residues themselves. This conservation suggests that this double phe-gate has a role in controlling not just the access of NADH to haem, but of alternative ligands that some of the homologues are perhaps capable of binding to. Whatever the case, it is noteworthy that these two residues, which are not directly involved in haem-binding, should be so well conserved, suggesting that regulation of ligand access to haem is of foremost importance.

A phylogenetic tree was constructed, with HemS, HmuS, ChuS, ShuS and PhuS highlighted. This tree emphasised how closely related HemS/HmuS and ChuS/ShuS are to one another. Indeed, ChuS and ShuS only vary in five residues, and yet their DNA-binding and reductive haem breakdown properties have been found to be very different. The next chapter will describe the alterations to the protein expression and purification protocols that had to be made for ShuS because of its propensity to aggregate and precipitate out of solution upon cell lysis, which is thought to be a result at least in part of DNA-binding. Such treatment was not required for ChuS.

The bioinformatic package, DNAbind, identified ten residues apiece in ChuS and ShuS which could potentially bind to DNA. The majority of these residues were clustered around the small cavity but none of them coincided with the five residues which differ between the homologues. It is speculated that having R123 in ShuS rather than S123 found in ChuS could increase the positive charge of the small cavity such that it becomes more attractive to anionic DNA.

DNAbind was also used to study HemS and PhuS. This analysis found no obvious

region for DNA-binding in HemS, whilst in PhuS the most likely region for binding was determined to be the large rather than the small cavity. This difference may explain why experiments by other groups have suggested that ShuS binds to DNA non-specifically whereas PhuS binds selectively to the *prrF1* promoter.

Overall, incorporating a bioinformatic element into this study has allowed for analysis of HemS within its broader context. Given the wide range (and often seemingly contradictory) roles assigned to HemS, HmuS, ChuS, ShuS and PhuS from the literature, phylogenetic and residue conservation studies have been very useful in providing possible reasons for this. Clearly, these proteins have evolved from a common ancestor and retained their unique haem-binding cavity. Yet, as evolution has proceeded, it seems these homologues have diverged to exploit different niches depending upon the requirement of the species producing them, the two obvious functions being to control haem flux *via* promoter suppression and to degrade haem reductively.

Chapter 9

Experimental Characterisation of Mutants and Homologues

9.1 Aims

Computation predicted clear mutants and homologues to test experimentally. First, therefore, efforts were made to express and isolate these proteins. This process proved successful, with any deviations from the standard WT HemS growth and purification protocol noted. Haem-binding characteristics of these proteins were investigated, and whether they could catalyse the NADH-dependent haem breakdown reaction. Stopped-flow spectroscopy was used to investigate the kinetics of these reactions and to test for the presence of the intermediate identified in the WT reaction. Crystallisation of selected mutants was also attempted. Therefore, to summarise, the aims of the characterisation of HemS mutants and homologues were:

1. To determine whether the mutant and homologous proteins could be grown successfully in *E. coli* cells using the same protocol that was developed for WT HemS.

2. To compare haem-binding characteristics between WT HemS and these mutants and homologues.

3. To determine whether these mutants and homologues could also catalyse the NADH-dependent breakdown of haem.

4. To use stopped-flow spectroscopy to compare the kinetics of reaction between NADH and haem in the mutants and homologues, with respect to WT HemS.

5. To attempt crystallisation and X-ray crystallography of selected mutants to determine the effect the mutations have on the protein structure, and to attempt

crystallisation and X-ray crystallography of the homologues, particularly those that do not have published structures.

Outcomes from experiments inspired by these aims are described in the following sections.

9.2 Expression and Purification

The method of expression and purification used for HemS proved to be suitable for all of its mutants and homologues, with all giving similar yields. The one exception was ShuS.

The first attempt to lyse *E. coli* cells containing over-expressed ShuS resulted in the protein aggregating and precipitating out of solution. ShuS is known to be capable of binding to DNA,[102;103] and so a fresh batch of cells were lysed in the presence of benzonase nuclease and protease inhibitor, which prevented aggregation, and the purification process proceeded as normal. However, over time, it was found that ShuS was still precipitating out of solution when stored in high concentrations. Increasing the pH to 8.0 from 6.5 prevented this aggregation.[a]

It was considered noteworthy that none of the other homologues required this treatment, which implied that ShuS was the only one capable of binding to DNA. As ChuS and ShuS share 98.5% sequence identity, a further study into why these two homologues behave so differently was instigated, which was carried out by Xie[81] (see Section 8.4).

An SDS-PAGE gel of all of the mutants and homologues studied is provided in Fig. 9.1, highlighting their respective purities and variations in mass. The exact masses, confirmed by mass spectrometry, are given in Table 9.1.

9.3 Haem-Binding Properties

The haem-binding capabilities of the mutants and homologues were investigated to determine whether there were any differences between them and WT HemS. It

[a]It is therefore not entirely clear what causes ShuS to precipitate out of solution, making this effect worthy of further study. Solutions treated with benzonase nuclease then stored at pH 6.5 in low concentrations ($< 100\,\mu M$) do not appear to precipitate out of solution faster than any of the other homologues, thus suggesting that the presence of DNA during cell lysis is indeed a problem. However, at higher concentrations, ShuS precipitates faster than any of its homologues, even if it was treated with benzonase nuclease. This problem can be minimised either by reducing the concentration or increasing the pH to 8.0.

Protein	Expected Mass / Da	Actual Mass / Da	Difference / Da
Wild Type HemS	39,360	39,360	0
F104A	39,284	39,284	0
F104AF199A	39,208	39,208	0
F104I	39,326	39,326	0
F199A	39,284	39,284	0
R209A	39,275	39,275	0
R209K	39,332	39,337	+5
Q210A	39,303	39,303	0
HmuS	39,104	39,104	0
ChuS	38,845	38,845	0
ShuS	38,831	38,833	+2

Table 9.1: Accurate masses of the proteins. Mass spectrometry was used to determine the actual masses. The masses for the R209K HemS mutant and ShuS were both slightly different from their respective expected values. Care must therefore be taken when interpreting the data involving these proteins.

was not expected that there would be any significant differences, since the homologues were already known haemoproteins, and none of the mutated residues studied (except R209) were thought to be directly involved in haem-binding.

UV-Visible spectroscopy experiments were carried out as described in Section 3.5.1. The intention had been to run a set of experiments for all mutants and homologues at pH values of 8.0, 6.5 and 5.0. However, it was found that the protein samples were not stable enough at pH 5.0, typically precipitating out of solution within minutes, and so measurements at this pH value were abandoned. Furthermore, the F104A and R209K samples had aggregated over time in the 4 °C refrigerator and so were not suitable for study at all. Spectra are shown in Fig. 9.2, and numerical results in Table 9.2.

The results show, consistent with the research carried out by Sawyer,[42] that the wavelength and the intensity of the Soret band of WT HemS both increase as the pH is increased. The increase in the wavelength was attributed to the fact that high-spin water complexes, which are more likely to arise at lower pH, tend to have shorter wavelength Soret peaks (∼406 nm) than low-spin hydroxide complexes (∼408 nm). The increase in intensity could be attributed to haem-binding being more efficient at higher pH, a change in the electronic conditions of the porphyrin, or a change in pK_a due to deprotonation.

This increase in intensity with respect to increasing pH is a property shared by all of the mutants and homologues apart from F104AF199A (where it decreased) and F199A (where it effectively stayed the same).

R209A stood out, as the wavelengths for its Soret maxima at both pH values

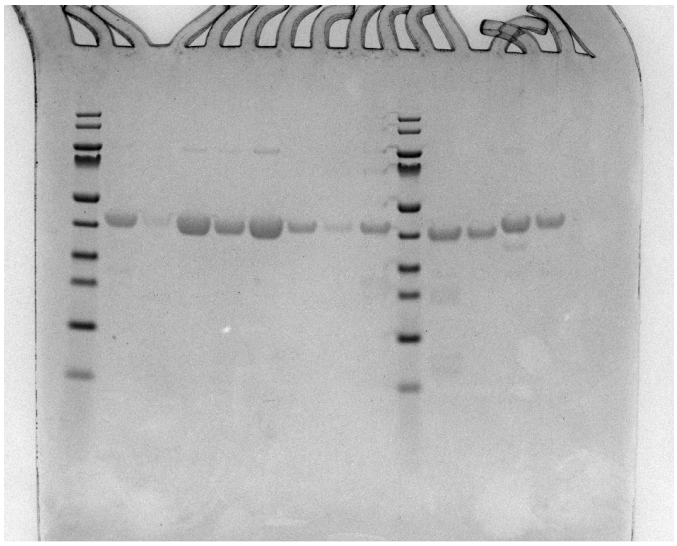
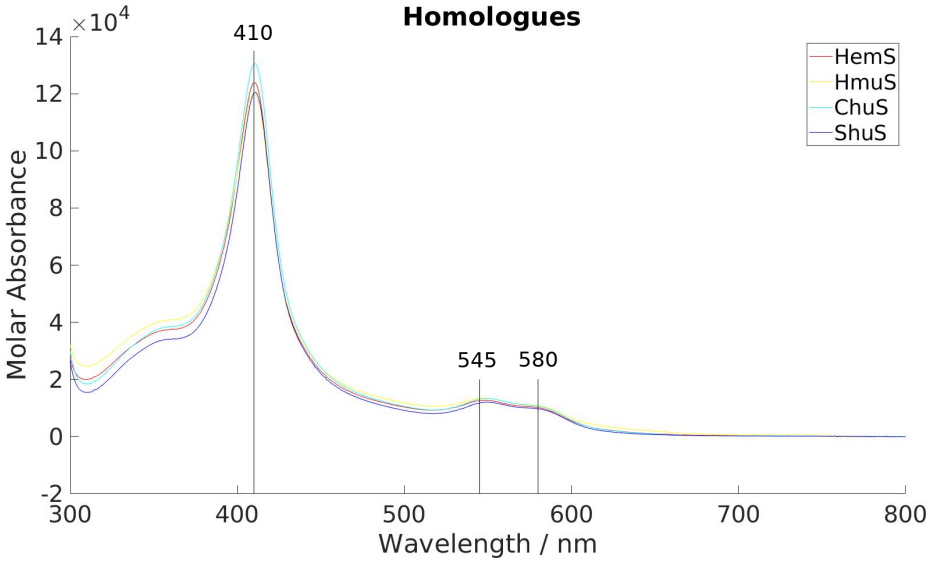


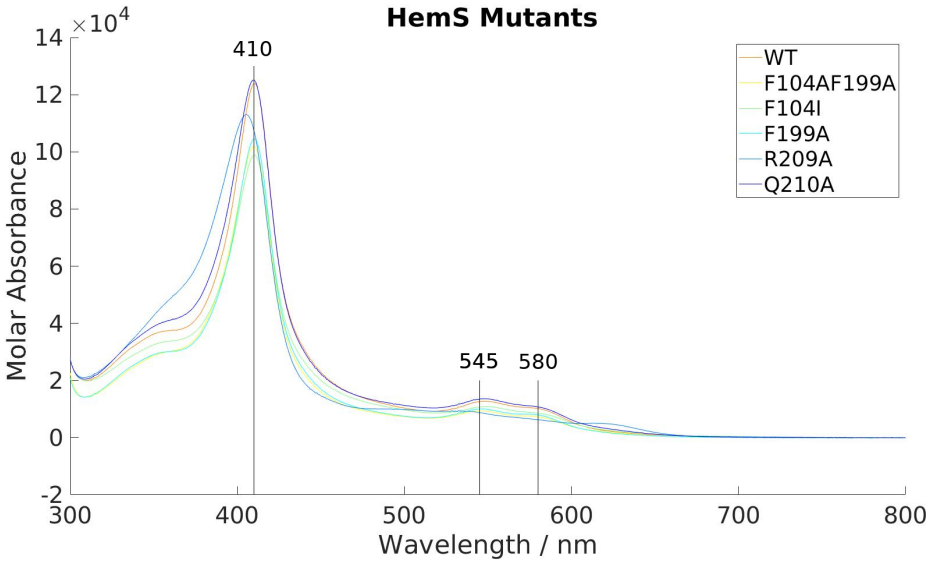
Figure 9.1: SDS-PAGE gel of the proteins studied in this work. All have had their His-tags cleaved. Lanes are numbered from left-to-right. Lanes 1 and 10 are markers, using PageRuler Protein Ladder, where the bands range from 10 kDa to 180 kDa. The samples are all pure, and the proteins are all situated close to the fifth band, which corresponds to 40 kDa. Lane 2: WT HemS; 3: F104A; 4: F104AF199A; 5: F104I; 6: F199A; 7: R209A; 8: R209K; 9: Q210A; 11: ChuS; 12: ShuS; 13: HmuS; 14: WT HemS (repeated).

were noticeably lower than for any of the other proteins. Furthermore, it did not have the signature set of peaks for the β-band at ∼545 nm and for the α-band at ∼580 nm. It was also unique, in that the wavelength decreased in going from pH

	pH 6.5		**pH 8.0**	
	Soret Max.	$\boldsymbol{\varepsilon_{SM}}$	**Soret Max.**	$\boldsymbol{\varepsilon_{SM}}$
WT HemS	409.00	103,000	410.25	123,900
F104AF199A	406.50	104,300	409.75	101,800
F104I	408.50	93,900	410.50	98,800
F199A	408.00	104,400	410.50	104,500
R209A	406.50	93,600	405.25	113,000
Q210A	408.50	99,400	409.75	125,100
HmuS	409.00	109,500	410.25	120,300
ChuS	409.50	112,300	410.50	130,500
ShuS	410.00	101,600	411.00	120,500

Table 9.2: Haem-binding properties of WT HemS and selected mutants and homologues. The Soret maximum is quoted in nm, and ε_{SM} is the extinction coefficient in $M^{-1}\,cm^{-1}$ at that Soret maximum. The ShuS values are lower than those reported previously (Wilks reported an ε_{410} value of $159\,mM^{-1}\,cm^{-1}$ at pH 7.4),[102] but appear to be consistent with the other homologues investigated in this study.


Homologues
×10^4
410
545
580
HemS
HmuS
ChuS
ShuS
Molar Absorbance
Wavelength / nm


HemS Mutants
×10^4
410
545
580
WT
F104AF199A
F104I
F199A
R209A
Q210A
Molar Absorbance
Wavelength / nm

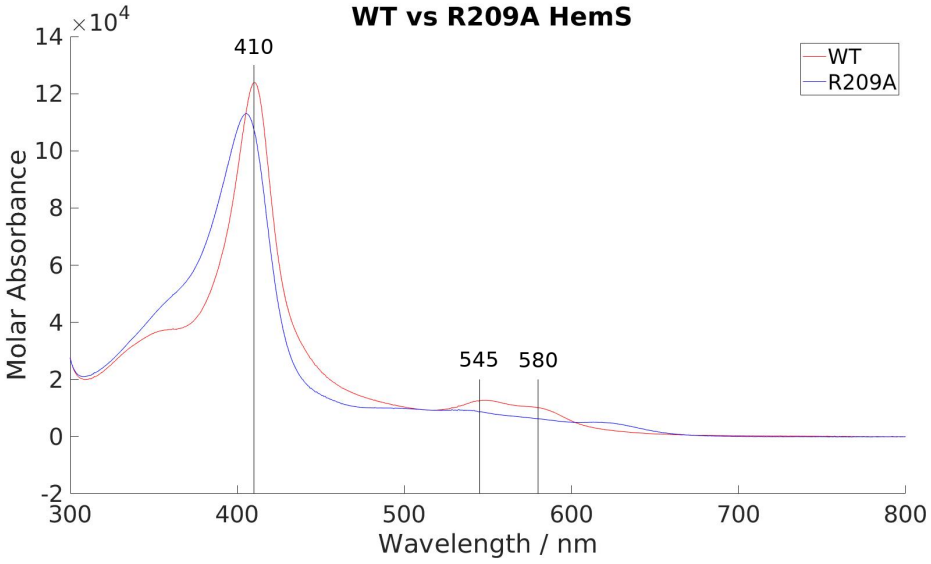

WT vs R209A HemS
×10^4
410
545
580
WT
R209A
Molar Absorbance
Wavelength / nm

Figure 9.2: Selection of UV-Visible haem-binding spectra at pH 8.0. All spectra were normalised by setting the absorbance at 800 nm to 0, thus giving a consistent baseline. The absorbance at each Soret peak was then set to the appropriate ε_{SM} values given in Table 9.2, therefore yielding molar absorbances. Top: Homologues. All of the proteins show similar haem-binding properties. Middle: Most of the HemS mutants show similar haem-binding characteristics with respect to the WT. Bottom: WT vs R209A HemS. R209A was least similar to the WT of all of the mutants. The removal of bonding to one of the haem propionates was most likely responsible, and possible implications of this removal on the cavity environment are given in the text.

6.5 to pH 8.0. The shorter wavelengths perhaps suggest that R209A causes the pK_{a} of the distal water/hydroxide ligand to change, shifting the equilibrium in favour of water. This shift in turn could have been the result of changed electronic properties of the haem (as a result of R209 no longer binding to one of its propionates) or of a change in the environment in which water/hydroxide was situated.[161] Furthermore, this mutation will alter the pK_{a} values of the propionates themselves, relative to the WT protein. Therefore, the removal of the interaction between the R209 residue and the more deeply buried of these propionates could significantly influence the pH dependence of the spectra as well.

9.4 Reaction with NADH as Monitored by UV-Visible Spectroscopy

Once all of the mutants and homologues had been expressed and purified, their reactivity with haem and NADH was tested using UV-Visible spectroscopy. The spectra are given in Fig. 9.3, and these all show growth in absorbance at 591 nm, implying that every mutant and homologue studied is capable of producing the HBP.

The fact that all of the mutants are capable of reactivity suggest that the structural integrity of the proteins have been retained and that none of those residues that were mutated were essential for the reaction to proceed. It is not a surprise that the F104 and F199 mutations did not stop reactivity, as they are mainly suspected to be involved in the regulation of NADH access to haem, and not its complete allowance/prevention of access. It is interesting that neither of the R209 mutations, nor Q210A, appear to significantly reduce the rate of reaction, although this shall be discussed further when considering the stopped-flow data.

All three homologues, HmuS, ChuS and ShuS, were also shown to be able to catalyse this haem breakdown process, which is proof, for the first time, that proteins other than HemS can produce this novel 591 nm HBP. This finding strongly suggests

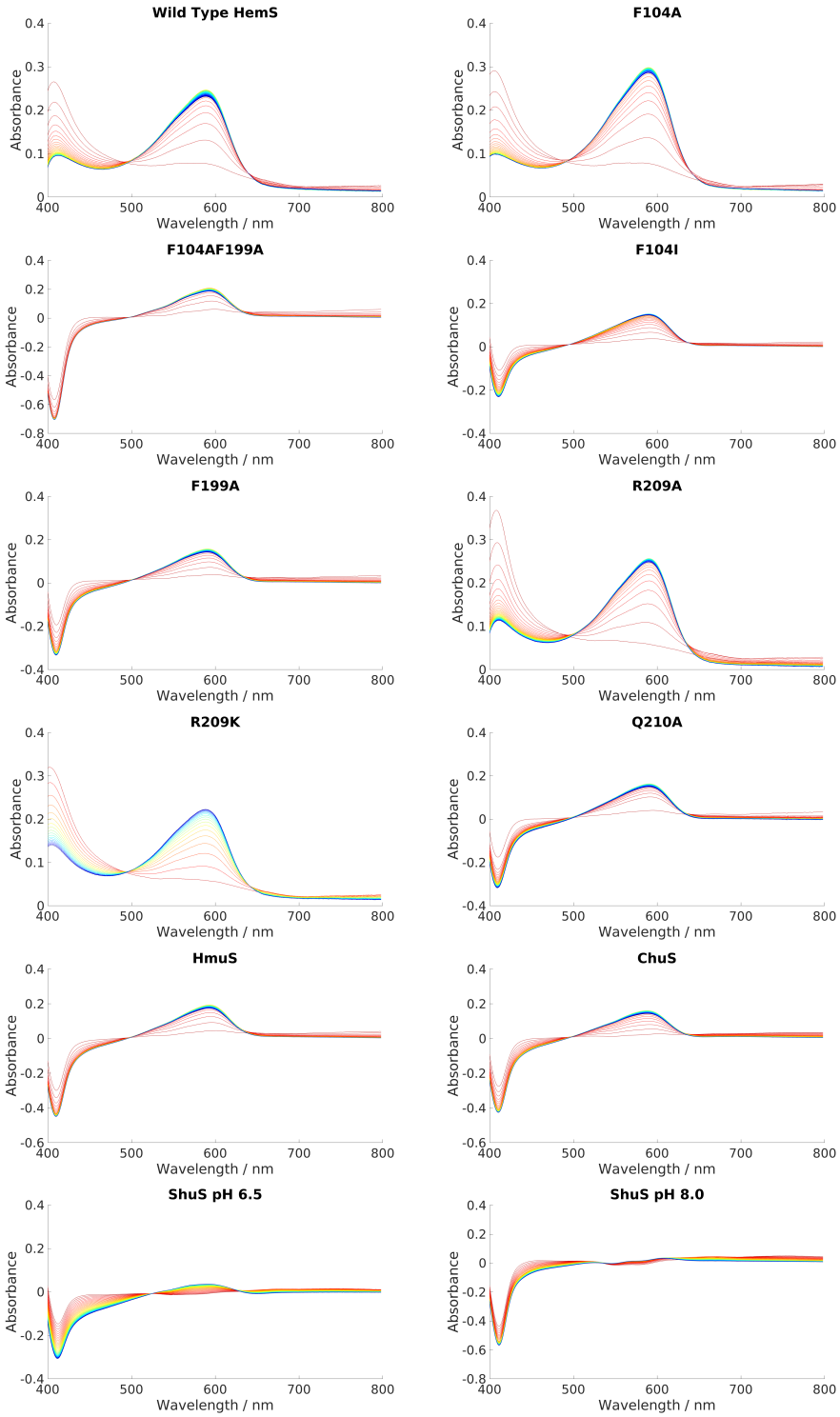


Figure 9.3: UV-Visible difference spectra of the mutants and homologues studied. Reaction stoichiometry was 8 µM protein : 8 µM haem : 2000 µM NADH. Scans were run every minute for 50 minutes, indicated by the colour scheme change from red (1 minute) to blue/purple (50 minutes), except for R209K, which only ran for 20 minutes. The positive absorbance values between 400-450 nm for the WT, F104A, R209A and R209K HemS samples indicate that haem had not properly equilibrated before the baseline was taken.

that proteins have evolved specifically to carry out this reaction, and that it was not merely an incidental artefact unique to HemS that it could catalyse haem breakdown *via* this reductive route.

However, the ability of ShuS to catalyse the reaction differed markedly from the other homologues. On the face of it, this is surprising since ShuS only differs from ChuS by five residues. However, as discussed in Section 8.4, ShuS can bind DNA yet ChuS seemingly cannot, suggesting a different role for ShuS. Nevertheless, ShuS can still catalyse the breakdown of haem with NADH, albeit at a much reduced rate. A second experiment was run where the pH was set to 8.0 rather than 6.5, as ShuS is less prone to precipitate out of solution at this higher pH. Though loss of haem was observed (indicated by the decreasing absorbance of the Soret band at ∼408 nm), there appeared to be only little growth at ∼591 nm, suggesting that haem was perhaps being broken down by some other mechanism.

9.5 Stopped-Flow Spectroscopy

As explained in Chapter 5, stopped-flow spectroscopy is a technique with high temporal resolution, making it suitable for further kinetic studies into the haem breakdown process in HemS and its homologues, and on the intermediate formation and consumption in particular. This intermediate had been identified in studies on HemS in Chapter 5. Were it to be identified in studies on the homologues, this confirmation would provide further evidence that they are all promoting the same reaction. First, however, the correlation of the intermediate formation with the loss of haem was examined.

9.5.1 Deconvolution of the Stopped-Flow Spectra

Experiments were run using a PDA detector over the range 400-722 nm, the latter being the long-wavelength limit for the detector. As the peak corresponding to the intermediate at 806 nm is broad, readings at 722 nm can capture the evolution of this intermediate qualitatively. Therefore, experiments run with a detector operating between 400-722 nm can chart the progress of haem loss (at 408 nm), HBP gain (at 591 nm) and intermediate evolution (at 722 nm) over time. In order to find out more about these individual species (namely, protein-bound haem, the HBP and the intermediate), an effective means of deconvoluting the overall spectra was sought. Singular value decomposition (SVD) provides a ready means for doing this analysis. The time course for the reaction of pre-incubated 5 µM WT HemS and 5 µM haem with 1000 µM NADH is shown in Fig. 9.4, along with decomposed spectra. SVD sug-

gested that the 3D spectrum consisted of three primary components. The first two corresponded closely to the Soret band at 408 nm and to the HBP peak at 591 nm, respectively. The third revealed a shoulder of a peak at the upper wavelength limit of 722 nm which is consistent with the intermediate species, which is known to have a broad peak centred around 806 nm. SVD also suggested that this component has a further, sharper peak at ∼433 nm. Such a peak is always masked by the Soret band, explaining why it has not been observed before. Much of the spectrum of the intermediate between 400-650 nm is obscured by haem and the HBP, but these data suggest that the intermediate species has two signature peaks in the 400-850 nm region, one around 430-440 nm and the 806 nm absorbance itself. These are significantly red-shifted compared with haem and suggest that the conjugated bonding in the porphyrin has moved to lower energy, but that the tetrapyrrole remains cyclic.

The SVD amplitudes, shown in the bottom right of Fig. 9.4, strongly indicate a direct inverse correlation between haem loss and intermediate formation. At the beginning of the reaction, the component representing haem steeply declines, which is matched by a sharp increase in the component representing the intermediate. Then, as the intermediate reaches its maximum value and starts to decline, this change is mirrored by the time at which the decline of haem starts to level off. The rates at which both the haem and intermediate level off then appear to match up until the reaction cutoff of 1020 s. The component representing the HBP, however, behaves differently. At a very short timescale (within the first 10 s), there is a slight lag, before the HBP concentration starts to increase. This increase is not as rapid as that for the intermediate, but it lasts longer, with the formation of the HBP beginning to level off after ∼400 s.

9.5.2 Dependence of the Intermediate on NADH Concentration

The rates of formation and consumption of the intermediate at 806 nm were investigated over a range of NADH concentrations. Parallel investigations were run with the WT and F104AF199A HemS proteins. The time courses are shown in Fig. 9.5.

Increasing the initial concentration of NADH leads to both an increase in the rate of formation and in the rate of consumption of the intermediate. Attempts were made to fit curves to these two sets of data using KinTek.[166] Even with the introduction of normalisation constants, this fitting proved too difficult, perhaps due to issues with the data itself. In the case of the WT, it does appear as if the absolute absorbances at an NADH concentration of 400 µM are consistently too high, plus the readings at 150 µM and 250 µM appear to be too close to one another

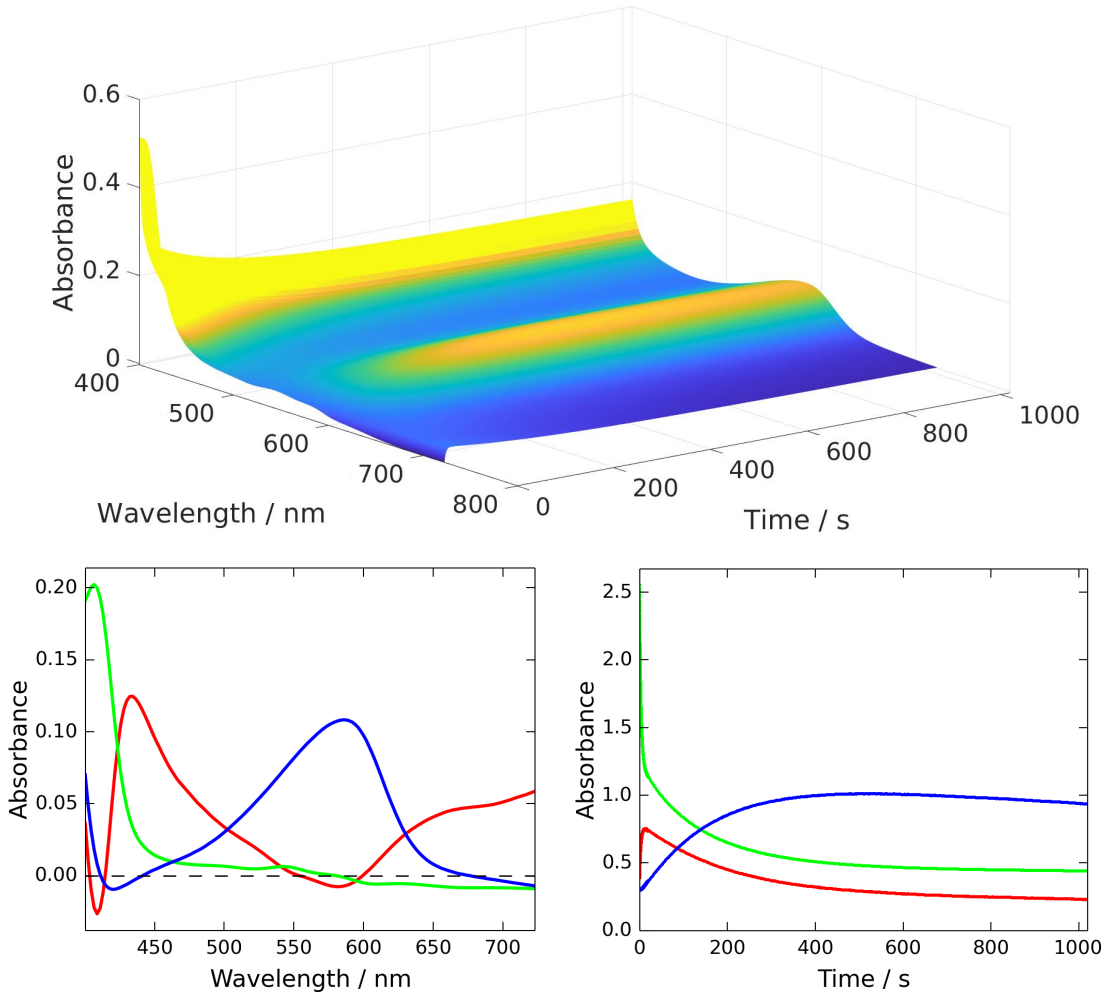


Figure 9.4: SVD analysis of time course for pre-incubated 5 µM WT HemS + 5 µM haem with 1000 µM NADH, recorded from 400-722 nm. Top: The 3D spectrum. The Soret band at 408 nm decreases over time, indicating loss of haem. A peak centred around 591 nm increases over time, indicating product formation. At 722 nm, there is a small but fast initial increase in absorbance, followed by a steady decline, which indicates intermediate formation then consumption. Bottom left: SVD component spectra. The green curve corresponds closely to the Soret band at 408 nm, and so gives an indication of haem concentration. The blue curve peaks at 591 nm, therefore reflecting HBP concentration. The red curve appears to contain a broad peak, with the wavelength cutoff at 722 nm capturing its shoulder. This is to be expected for the intermediate species, and it is assumed that this peak will be centred at 806 nm. This red curve also reveals a further peak, located at ∼433 nm. This peak is masked by the Soret band in convoluted spectra. Bottom right: SVD amplitudes. The colour-coding corresponds to that found in the SVD component spectra. These amplitudes show haem being consumed over time, an initial burst of intermediate formation before a steady decline, and a slower rate of product formation, which begins to level off and then slowly decline after ∼400 s.

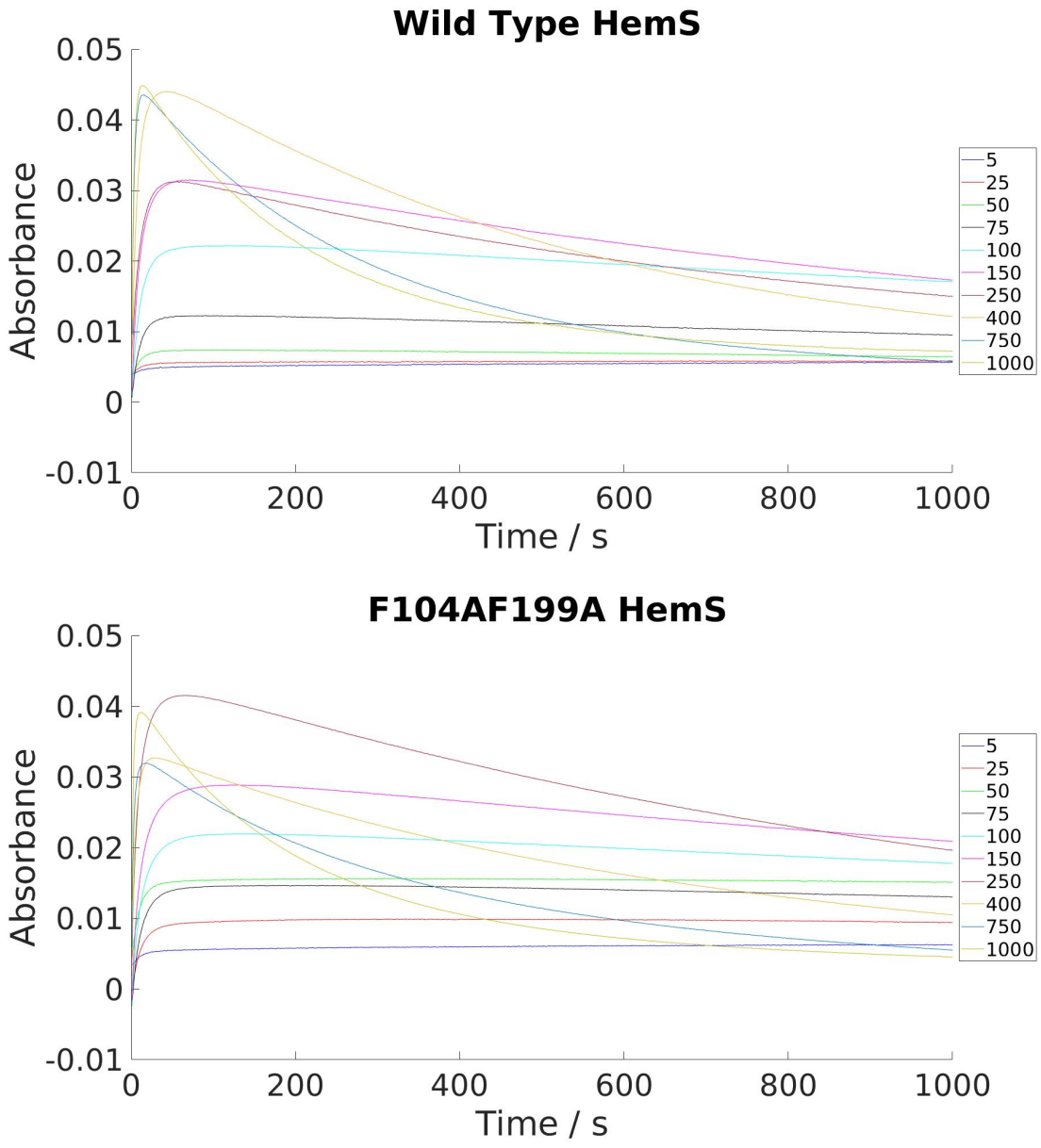
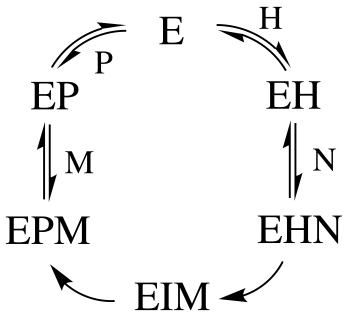


Figure 9.5: 806 nm time courses for WT HemS (top) and F104AF199A HemS (bottom), with varying starting NADH concentrations, described in the legend in units of µM. In each experiment, NADH was added to a pre-incubated mixture of 5 µM HemS and 5 µM haem. The rates of formation and consumption of the intermediate both increase with increasing NADH concentration.

to fit the overall trend. In the case of F104AF199A, meanwhile, some of the time courses start at negative absorbances. Another explanation for the failure to derive accurate curves from the data could be that the proposed reaction mechanism, given in Scheme 2 and described below, is wrong.



Scheme 2: Proposed reaction scheme for the overall haem breakdown process. E corresponds to the enzyme, H to haem, N to NADH, M to NAD^+, I to the intermediate with a characteristic peak at 806 nm and P to the haem breakdown product. The hydride transfer and subsequent step are considered as being irreversible.

Evidence from Sections 5.5.2 and 5.6 suggests that NAD^+, M, leaves the pocket faster than the haem breakdown product, P, thus explaining the order in which they leave in the scheme. Furthermore, the experiments were all performed on samples where haem, H, and the enzyme, E, were pre-incubated, and so the scheme reflects this by showing H being introduced to the enzyme before NADH, N. The assumption was made that only one intermediate, I, was formed throughout the scheme.

As simulations run using this scheme did not yield any fully converged curves, alternative mechanisms were attempted, including ones that considered allostery, but none of these proved successful either.

In the absence of a more detailed analysis, conclusions on the dependence of the intermediate on NADH concentration can only be tentative. One observation is that the rates of both the formation and the consumption of the intermediate increases with increasing NADH concentration, suggesting that the second phase of the reaction depends on NADH concentration in addition to the first phase. As the NADH concentration increases, the time at which the absorbance reaches a maximum decreases, although this maximum rises in absolute value.

Though a global fit had failed, fitting of the individual kinetic traces was achieved. This fitting used a simple two-step model, where the first step consisted of the irreversible conversion of *holo*-HemS and non-bound NADH to the intermediate species, and the latter step comprised the irreversible conversion of this intermediate to the final product, as shown in Scheme 3.

The coefficients and associated errors derived from these fits are given in Table

$$\mathrm{EH} + \mathrm{N} \xrightarrow{k_1} \mathrm{EIM}$$
$$\mathrm{EIM} \xrightarrow{k_2} \mathrm{EPM}$$
Therefore, equation to fit:
$$\mathrm{a} * \mathrm{EIM} + \mathrm{b}$$

Scheme 3: Simple two-step mechanism for haem breakdown, which was used to derive fits for the individual kinetic traces gathered by stopped-flow spectroscopy. The symbols follow the same logic as in Scheme 2.

9.3, and plots of the fits overlaid on top of the data are shown in Appendix G. For the Wild Type case, fits to the data gathered at NADH concentrations of 5 µM and 25 µM could not be achieved, and for the F104AF199A case, could not be generated when [NADH] was 5 µM, 25 µM and 50 µM, respectively. These fits could not be accurately derived as the absorbance at 806 nm (i.e. the concentration of the intermediate) was still rising after 1000 s, thus complicating the estimation of the rate of the conversion of the intermediate to the product.

Protein	Coefficient	50 µM	75 µM	100 µM	150 µM	250 µM	400 µM	750 µM	1000 µM
WT	$\mathbf{k_1}$	0.001392(1068)	0.001037(747)	0.000732(90)	0.000548(33)	0.000497(63)	0.000278(42)	0.000303(146)	0.000266(70)
WT	$\mathbf{k_2}$	0.000316(40)	0.000301(27)	0.000296(30)	0.000665(16)	0.000797(27)	0.001710(551)	0.004112(1151)	0.004791(500)
WT	**a**	0.000806(6)	0.002396(12)	0.004353(24)	0.006549(9)	0.006368(13)	0.008727(46)	0.008295(89)	0.008038(37)
WT	**b**	0.003531(31)	0.000727(64)	0.001159(132)	0.000336(45)	0.000299(63)	0.003926(19)	0.006054(27)	0.007168(11)
DM	$\mathbf{k_1}$	–	0.000810(256)	0.000558(113)	0.000419(103)	0.000385(50)	0.000378(73)	0.000236(62)	0.000375(118)
DM	$\mathbf{k_2}$	–	0.000111(45)	0.000241(36)	0.000333(40)	0.000809(24)	0.001483(736)	0.002123(568)	0.004836(636)
DM	**a**	–	0.002855(57)	0.004337(33)	0.005894(37)	0.008768(16)	0.006160(43)	0.007763(42)	0.007280(43)
DM	**b**	–	0.000576(306)	0.000980(186)	0.000303(208)	0.000103(76)	0.003358(23)	0.005638(18)	0.004890(14)

Table 9.3: Coefficients and associated errors derived from fitting to individual kinetic traces, using Scheme 3 as the model. WT stands for Wild Type HemS and DM for the double mutant of HemS, F104AF199A. k_1, representing irreversible conversion of the haem to the intermediate, is shown to decrease with increasing NADH concentration. This is due to the faster consumption of this intermediate, as represented by the increasing value of k_2 with increasing NADH concentration.

From these data, plots of k_1 and k_2 versus NADH concentration were made, as shown in Fig. 9.6. These plots show that, whereas k_1 decreases with increasing NADH, k_2 increases. Such a phenomenon was concluded to be due to the faster conversion of the intermediate species to the HBP masking any possible increase in the rate of the conversion of the reactants to the intermediate (which is possible since the rates are both ultimately derived from the absorbance levels at the signature peak for the intermediate at 806 nm).

Fig. 9.7 gives the variation in the initial formation rates of the intermediate species with respect to NADH concentration. This figure shows that the initial rate of intermediate formation increases with increasing NADH concentration, but then begins to level off as NADH reaches millimolar quantities. This data would therefore suggest that saturation of the enzyme occurs when NADH is approximately in 100-fold excess of the HemS : haem mixture.

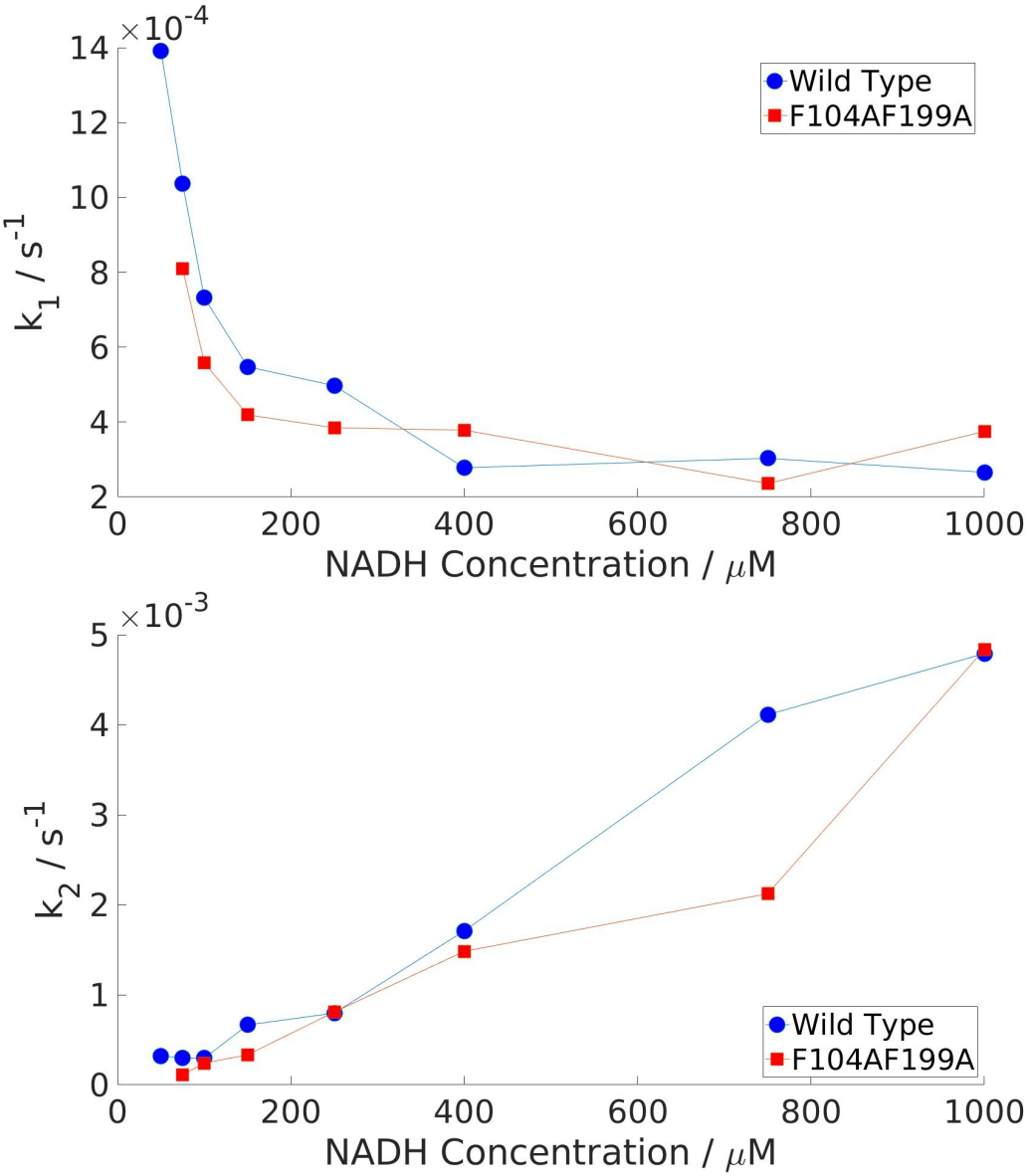


Figure 9.6: k_1 and k_2 values with respect to initial NADH concentration. In each experiment, NADH was added to a pre-incubated mixture of 5 µM HemS and 5 µM haem. Due to the extinction coefficient at 806 nm being unknown, these values are simply for the change in absorbance over the given time interval, rather than a direct description of the change in intermediate concentration.

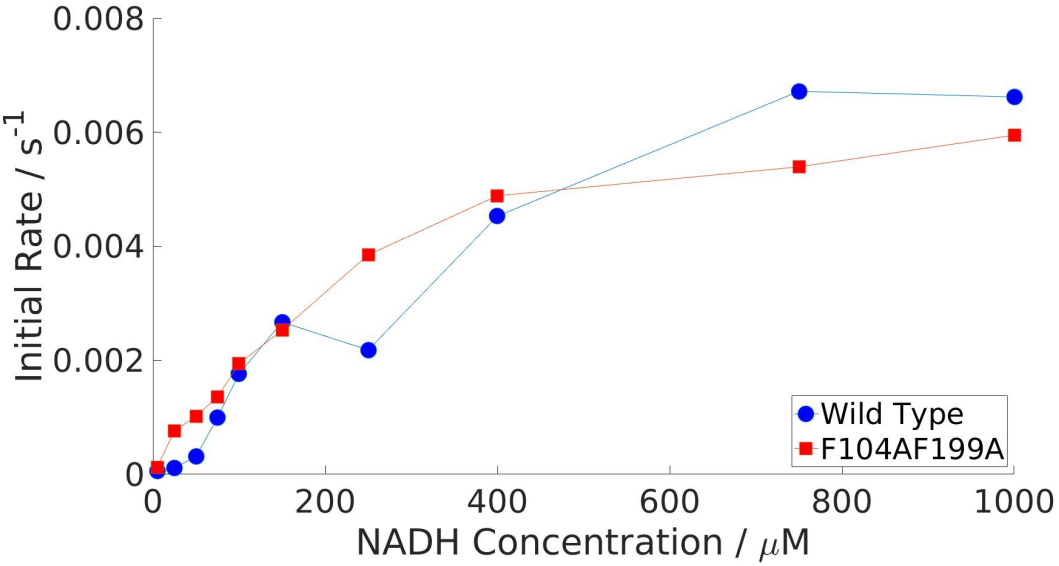


Figure 9.7: Initial rates of formation of the intermediate species with respect to initial NADH concentration. In each experiment, NADH was added to a pre-incubated mixture of 5 µM HemS and 5 µM haem. These rates were calculated from the first five seconds of the respective reactions, where the rises in absorbances are linear. Due to the extinction coefficient at 806 nm being unknown, these rates are simply for the rise in absorbance over a given time period, rather than a direct description of the change in intermediate concentration.

9.5.3 Effect of Mutants and Homologues on Intermediate Formation and Consumption

Time courses for each of the homologues at 806 nm are provided in Fig. 9.8. These time courses show that, in the presence of each homologue, there is a quick initial rise in absorbance at 806 nm, followed by a more gradual decrease, when NADH is added in excess to pre-incubated *holo*-protein. These changes in absorbance, as with HemS, suggest the formation and consumption of an intermediate species. The curves charting HemS and HmuS fall away more rapidly than for the other homologues (with the ShuS samples being especially slow), suggesting that the intermediate is being used up more rapidly in the presence of these two homologues. These data therefore reinforce the conclusion drawn from the previous spectroscopic data that HemS and HmuS are more effective at breaking down haem to produce the HBP than ShuS. It appears that build-up of the intermediate is greatest with HmuS. However, care should be taken when analysing absolute absorbances. Despite the best efforts of the author, including repeating all experiments in triplicate (see Section 3.6 in the Methods for further details), it is recognised that the Xe lamps used in typical stopped-flow setups can fluctuate. This problem is especially relevant to the current

	$A_{0.125\,s}$	t_{peak} / s	A_{peak}	$A_{250\,s}$	$R_{form.}$ / μs^{-1}	$R_{cons.}$ / μs^{-1}
WT HemS	0.0049	14.000	0.0436	0.0194	2787	102
F104AF199A	0.0044	15.625	0.0459	0.0245	2676	91
F104I	0.0042	25.250	0.0509	0.0298	1859	94
F199A	0.0047	16.125	0.0448	0.0252	2507	84
R209A	0.0043	43.375	0.0284	0.0177	556	52
Q210A	0.0226	23.375	0.0720	0.0487	2128	103
HmuS	0.0103	18.125	0.0611	0.0358	2822	109
ChuS	0.0059	19.500	0.0462	0.0309	2078	66
ShuS pH 6.5	0.0057	24.625	0.0377	0.0300	1308	34
ShuS pH 8.0	0.0030	64.125	0.0396	0.0338	572	31

Table 9.4: Rates of formation and consumption of the intermediate species. t_{peak} corresponds to the time at which the intermediate absorbance at 806 nm reaches a maximum, A_{peak} is the absorbance at this time, $A_{0.125\,s}$ is the absorbance after 0.125 s, and $A_{250\,s}$ is the absorbance after 250 s. $R_{form.}$ and $R_{cons.}$ are simple approximations for the rates of formation and consumption of the intermediate species, respectively. $R_{form.}$ is calculated by taking the difference between A_{peak} and $A_{0.125\,s}$, and dividing by the difference in time between those two readings. $R_{cons.}$ is the same but for A_{peak} and $A_{250\,s}$ instead. Extinction coefficients at 806 nm are unknown, and so absorbance values are compared directly, rather than concentrations. In interpreting results, an assumption is made that the extinction coefficient will be the same for all mutants and homologues. The data show that all mutations to HemS give a slower rate of formation, and that all mutations other than Q210A result in a slower rate of consumption of the intermediate. Furthermore, HemS and HmuS promote similar rates of formation and consumption, whereas ShuS gives the slowest rates of any homologue. F104A and R209K are not included in this study because these samples had degraded before the assays could be prepared.

experiments, where the absorbance values being detected were often low and the experiments were conducted over several days.

The time courses at 806 nm with the mutants of HemS were also studied, and are provided in Fig. 9.8. Here again, care must be taken when interpreting the absolute absorbance values. It is apparent, though, that the rate of consumption of the intermediate is greater for the WT than for most of the mutants. A numerical comparison of these rates of consumption, along with those from the homologues, is provided in Table 9.4.

9.6 Crystallography

Attempts were made to crystallise a selection of the mutants and homologues. A limited number of crystals did form for the majority of them. However, these crystals proved to either not be of a high enough quality for diffraction, or were destroyed by the cryoprotectants added to them.

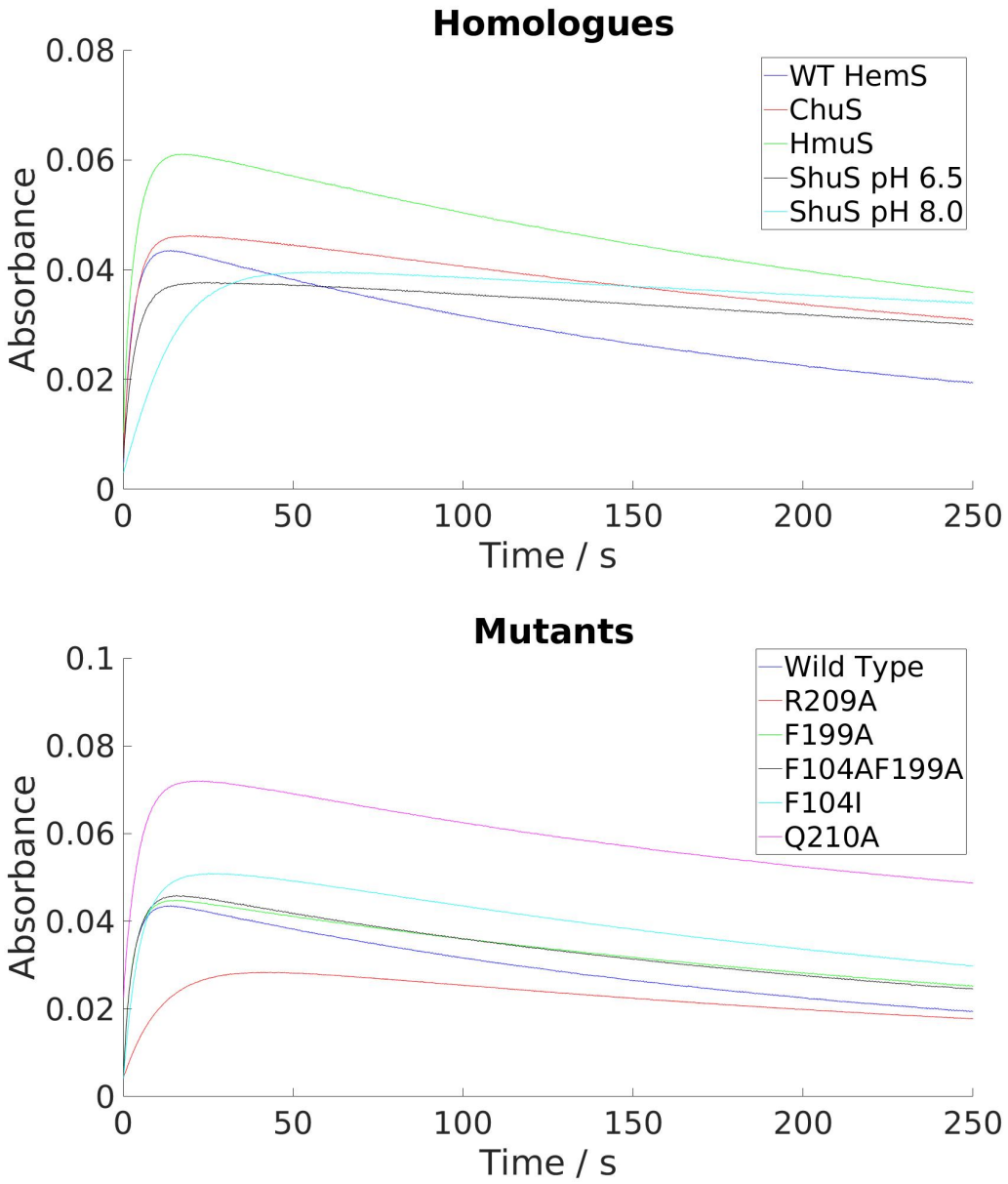


Figure 9.8: Time courses of the intermediate concentration, as reported by the absorbance at 806 nm. In each experiment, 1000 µM NADH was added to a pre-incubated mixture of 5 µM protein and 5 µM haem. Top: Homologues. All plots show an initial fast rise in absorbance, followed by a gradual decrease, suggesting that an intermediate is formed in the presence of each homologue. The consumption of the intermediate is more rapid for HemS and HmuS than it is for the other homologues. Bottom: Mutants. All plots also show an initial fast rise in absorbance, followed by a gradual decrease, suggesting that an intermediate is formed in the presence of each mutant. Apart from Q210A, the consumption of the intermediate is more rapid for the WT than it is for the mutants, suggesting each of the mutations made deleteriously affect the conversion of the intermediate to the product.

As it is central to the main pocket, separating the haem-binding and NADH-binding regions of the protein, it was hypothesised that removing the double phe-gate may lead to significant structural changes. Therefore, additional assays were run to crystallise F104AF199A HemS. For reference, the results for WT HemS can be found in Section 5.5.2. That section discussed how the resolution of the unstructured loop region, consistent with all previous crystallographic studies on HemS and its homologues other than the PhuS dimer,[111] was too low to be accurately resolved. However, the resolution with F104AF199A samples proved to be higher, so that this 'missing loop' could be constructed for the first time. This is therefore the first X-ray structure of HemS or one of its homologues (other than the PhuS dimer)[111] where all residues are resolved. The structure is shown in Fig. 9.9, superimposed upon the WT structure.

Superimposing resolved crystal structures for the WT and for F104AF199A shows that the main pocket changes in subtle ways upon removal of the double phe-gate. Indeed, the main pocket is more open, with the capping α-helix less deeply buried. However, simulations suggest the reverse is true, with this α-helix more tightly clamped down in the F104AF199A case than in the WT case. As the simulations contain haem and NADH, and yet the crystal structures did not, it may be the case that these simulations are not inaccurate, but that they are instead showing that F104AF199A clamps down on haem more significantly than the WT does upon the inclusion of haem. Since the removal of the double phe-gate affords more space in the central cavity, this conclusion would seem to be consistent with basic steric considerations. Looking from one perspective, it is curious that HemS has therefore evolved to contain this double phe-gate, since its absence results in the protein being better able to clamp down on haem, and therefore presumably better hold it in place. However, the fact that the mutated protein shows greater conformational flexibility between its *apo*- and *holo*-forms perhaps points to a greater inherent instability in its structure. When left at refrigerated temperatures, it was found that samples of F104AF199A did tend to aggregate quicker than the WT.

9.7 Discussion and Summary

The research described in this chapter used experimental techniques to better understand the properties of the selected mutants and homologues of HemS, as well as to deepen an understanding of the anaerobic breakdown of haem.

Firstly, it was shown that all of the mutants and homologues were able to bind to haem, with each giving a clear Soret peak at ∼408 nm. The most significant

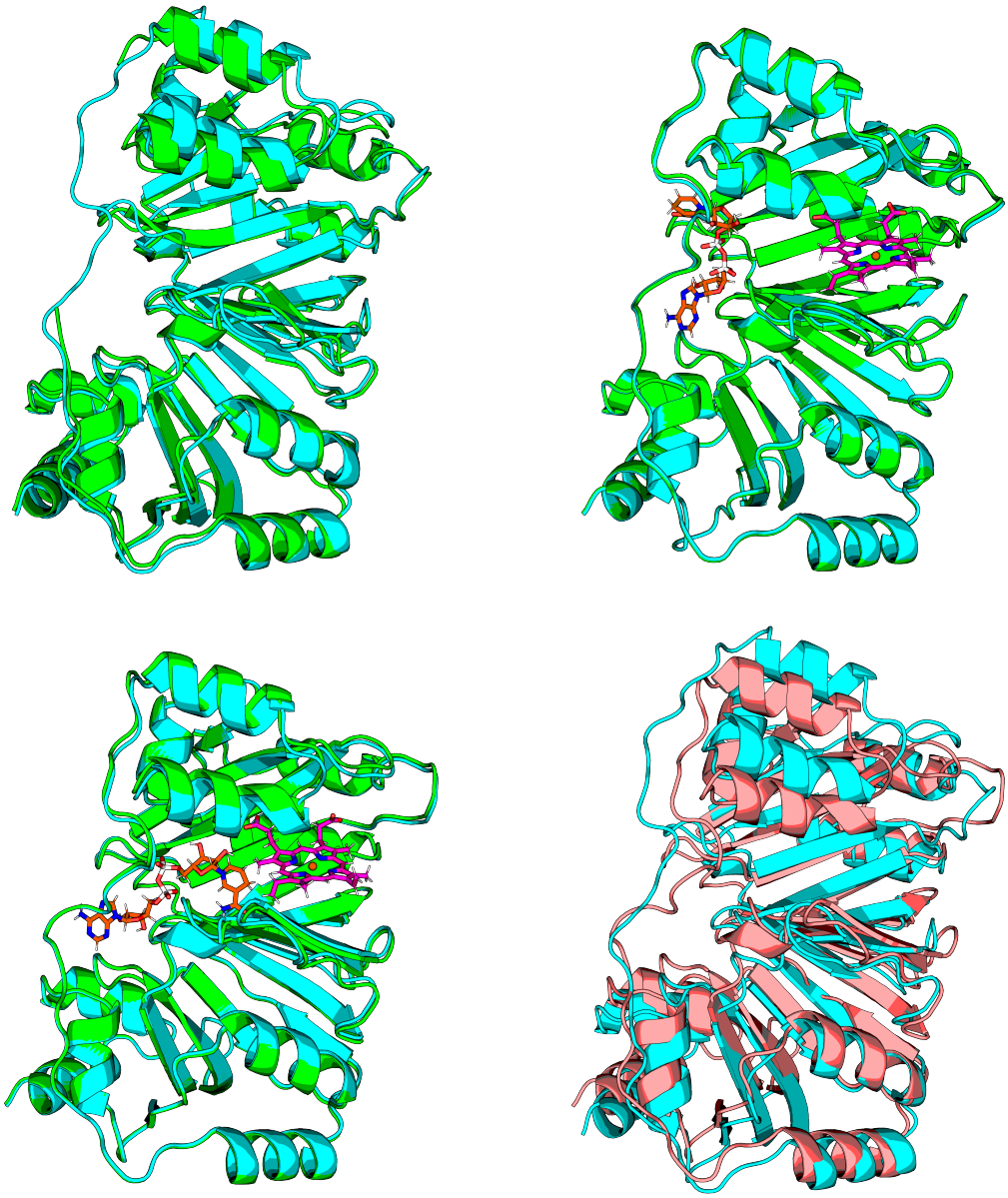


Figure 9.9: WT and F104AF199A HemS structures. The WT is represented in green and F104AF199A in cyan. Top left: The structures resolved from X-ray crystallography. The (cyan) capping α-helix of F104AF199A is shown to be further from the central cavity compared to the WT capping α-helix. Top right: The starting structures in the Dijkstra fastest pathways, as derived from computation. Bottom left: The final structures in the Dijkstra fastest pathways, as derived from computation. For clarity, only NADH and haem from the WT structure are shown in each case. Bottom right: Two representations of F104AF199A superimposed on one another, to demonstrate the movement of the unstructured loop and the pocket-capping α-helix upon NADH- and haem-binding. The cyan representation is the structure resolved from X-ray crystallography and the salmon representation is the final structure in the Dijkstra fastest pathway.

conclusion from this chapter is that this haem breakdown reaction is not unique to HemS, but can also occur in the homologues, HmuS, ChuS and ShuS. These results would therefore suggest that the anaerobic breakdown of haem is indeed an important strategy engaged in by pathogenic bacteria to extract iron from their host organisms. As was predicted from both computational and bioinformatic studies, these homologues had different propensities to catalyse the reaction, with ShuS being the least effective enzyme. Studies with the bioinformatic package, DNAbind, had suggested that ShuS may have competing DNA-binding properties in addition to its ability to bind to haem. Though no studies in this current work were specifically designed to test this hypothesis, its veracity was inferred from the fact that the expression of ShuS required the inclusion of benzonase nuclease to prevent DNA-protein aggregates from forming.

All of the mutants were able to catalyse the reaction as well, showing that none of the residues selected for mutation – which were all predicted to influence the NADH unfolding and its access to haem – are essential for the reaction to proceed, and so HemS can therefore tolerate a wide range of mutations to the NADH-binding region.

Stopped-flow spectroscopy was used to further probe how the reaction varies between the mutants and homologues. The original intention had been to engage in a full kinetic study with global analysis. A combination of the mechanism's complexity, plus suspected inconsistencies in absorbances arising from lamp fluctuations between runs, precluded this analysis.

Stopped-flow experiments were run where the detector was set to the wavelength range 400-722 nm. This setting allowed for the concentrations of haem and the HBP to be tracked, plus the intermediate imperfectly (since its peak is at 806 nm, but its shoulder can be observed at 722 nm). Analysis by SVD gave three primary components whose spectra matched these three species (i.e. haem, the intermediate and the HBP) closely. The component corresponding to the intermediate revealed that this species has a peak at ∼433 nm, which tends to be masked by the haem Soret peak, in addition to the one at 806 nm. Tracking these components over the time course showed that there is a direct inverse correlation between haem and the intermediate, demonstrating that haem breakdown is indeed directly correlated to intermediate formation. This result strengthens the hypothesis that the intermediate is the immediate product resulting from hydride transfer from NADH to haem.

The dependence of intermediate formation and consumption on NADH concentration was then investigated. It was shown that both rates increase with increasing NADH concentration, suggesting that, in addition to its formation step, the inter-

mediate breakdown step also depends on NADH. Furthermore, the time at which the intermediate peaks in concentration shortens, plus its concentration at this time increases. The increase in intermediate concentration may be due to faster rates of NADH association with the protein, unfolding and hydride transfer, before the proposed sigmatropic rearrangement to form the final HBP takes meaningful effect. However, it is thought that the shift to shorter time is due to the rate for this proposed sigmatropic rearrangement increasing faster than for any of these preceding steps in the mechanism.

It was shown that all of the selected mutations to HemS lead to a reduced rate of intermediate formation, and that all mutations other than Q210A result in a reduced rate of consumption. The fact that this Q210A mutant behaves this way is not surprising. Unlike the other residues that were mutated, Q210 is not located near to the region of the pocket where hydride transfer is expected. Instead, it is located further towards the edge of the pocket, where it was hypothesised from the computational data (see Section 7.3) that it helps NADH to unfold by hydrogen-bonding to the phosphate backbone. Its mutation to alanine would explain the decreased rate of formation of the intermediate as this residue cannot form hydrogen-bonds, and so, under this hypothesis, it would become more difficult for NADH to unfold and therefore it would take longer to reach haem. However, once in close proximity to haem, NADH would presumably be positioned and oriented similarly to the WT case, given that the residues in that region are all retained. Assuming NADH is involved in the breakdown of haem following the transfer of hydride, perhaps just by influencing the electronic environment, the position and orientation of NADH (or, more correctly, NAD^+) could have a significant influence on the rate. The WT and Q210A cases should provide similar electronic environments, which would explain why they give similar rates of intermediate consumption whereas all of the other mutants give reduced values.

The R209A mutation reduces the rates of formation and consumption of the intermediate the most. If R209 binds both the NADH and haem molecules, that such a mutation should be so disruptive is also not surprising.

Qualitatively, it was shown that consumption of the intermediate species for HemS and HmuS was significantly higher than for the other homologues, reinforcing the conclusions drawn from the standard UV-Visible spectroscopic data, as well as from computational and bioinformatic predictions, that ChuS and ShuS are less effective enzymes.

Attempts to crystallise the majority of the mutants and homologues either did not work, were subsequently destroyed by cryoprotectants, or produced too low a

resolution for accurate structural determination. Crystals of F104AF199A, however, were successfully grown and their structures resolved, using the procedures outlined in Section 3.11. These structures showed that, in the absence of haem or NADH, the main cavity of HemS is more open than for the WT. As the computational work consistently shows that, with haem and NADH present, the reverse is true, two conclusions can be drawn. The first is that the simulations do not accurately represent the WT and/or F104AF199A proteins. The second is that the removal of the double phe-gate allows for greater conformational flexibility, meaning the double mutant can 'clamp down' more strongly on haem. The second conclusion implies that F104AF199A may not be as stable in its *apo*-protein form, since it seems unable to maintain its cavity size and shape as consistently, perhaps therefore explaining the observation that it can aggregate more readily than the WT.

From these F104AF199A crystals, the unstructured loop spanning the entrance to one side of the main cavity was resolved fully for the first time for HemS or any of its close homologues. This structure, therefore, is the first one that has been fully resolved for this class of haemoprotein other than the PhuS dimer.

Chapter 10

Conclusions and Future Work

10.1 Conclusions

The most significant conclusion to be drawn from this work is that the novel anaerobic haem breakdown reaction discovered by Sawyer is not unique to HemS. Instead, such breakdown has been demonstrated to occur, to various extents, in the HemS homologues, HmuS, ChuS and ShuS. These proteins are a small subset of a wider class, which are typically notated as 'haem transport' or 'haem-degrading' factors. The research by Sawyer, Choy and the current author is the first demonstration that this class of protein can indeed promote the breakdown of haem.

Novel features of this reaction were investigated in more detail. The hypothesis that haem breakdown was initiated by reductive hydride transfer from NADH was confirmed through a series of deuterium labelling experiments. Both stereoisomers of NADD produced deuterated and non-deuterated HBPs, suggesting that the reaction is not stereospecific. Such reactions are unusual but not unheard of in enzymology. Transfer of a hydride from the (*S*)-position of the nicotinamide head of NADH was shown to be significantly faster than from the (*R*)-position. It is perhaps the case that, once NADH has committed to unfolding in a certain manner and presenting one of its hydrides to haem, the barrier to changing conformation is higher than the transfer of the less-favoured hydride. As NADPH hydrogen-bonds to an additional (T312) residue when it associates with HemS, this structure may induce the stereospecificity not found with NADH. Studies on deuterated NADPH would be required to determine whether this is the case.

The reaction mechanism was elucidated further through the discovery of a short-lived intermediate. This intermediate has signature UV-Visible absorption maxima at 806 nm and 433 nm, the latter only becoming apparent after SVD analysis of the time evolution of the UV-Visible spectra. Together, these peaks suggest a molecule

that is still largely similar to haem structurally, but which has had its conjugated system moved to lower energy. It was shown that the formation of the intermediate coincides with the loss of the haem absorbance in the Soret band. The working hypothesis is therefore that this intermediate is the immediate result of hydride transfer, and that a subsequent sigmatropic rearrangement causes the cyclic porphyrin to open to produce the final HBP, a linear tetrapyrrole.

The ability of the reaction to proceed anaerobically was demonstrated definitively. Under aerobic conditions, there is an indirect competition for haem breakdown, which arises *via* coupled oxidation and produces non-regioselective biliverdin. Under certain protein : haem ratios, this competition significantly reduces formation of the HBP. All of the proteins studied in this work are from pathogenic bacteria, which are known to thrive in areas of the human gut where oxygen levels are low. It would therefore not be surprising for these bacteria to have developed alternative strategies for the controlled breakdown of haem, which are not dependent on oxygen. Further to this conclusion, there was no evidence of 'canonical haem oxygenase' activity in any of the proteins studied.

Attempts to characterise the HBP by NMR proved unsuccessful. Other than trying paramagnetic NMR, it would seem that the electronic properties of this species, as well as its instability outside the protein, would preclude further study by this method. X-ray crystallography has proven to be more fruitful. Post-reaction structures of HemS have been successfully crystallised and resolved. The structure of the HBP in the pocket has not proved to be as straightforward to resolve. This problem may be due to an inherent instability in the structure caused by loss of iron. It is thought that the cryoprotectants used may have been causing iron to escape from the cleaved porphyrin prematurely. If this is the case, then alternative cryoprotectants may prove capable of retaining the HBP for accurate resolution.

Previous studies had shown little homology between HemS and other proteins known to bind to NADH, suggesting that the details of NADH-binding in HemS and its homologues are unprecedented. As NADH binds transiently in the pocket, computational experiments were required to probe this novel interaction more closely. All-atom models using AMBER and the programs developed by the Wales group at the University of Cambridge provided the means for this analysis. Previous studies had been limited by the computational methods and resources available at the time. However, for this study, calculations could be run on GPUs, giving speed-ups of up to two orders of magnitude.

Even with this increase in efficiency, further strategies were sought to expedite these calculations. To accelerate growth of the WT HemS database, a method for

choosing unconnected minima in separate sub-databases for connection attempts was developed, based both on minimising the number of attempts required to fully connect all of the sub-databases, and on choosing the minima closest in conformational space to make those connections. Once this WT HemS database was complete, a new subroutine was written, which could use the stationary points (or a selection thereof) in this database as a template to seed new databases where the protein has either been subject to single-point mutation(s), transformed into a homologue, or the ligand(s) changed. This subroutine removes the need to build up new databases from scratch, and thus allowed fresh databases describing seven HemS mutants and three homologues to be constructed and connected in a fraction of the time it would have required by alternative strategies.

The construction of large databases afforded an opportunity for the general properties of these systems to emerge, particularly when displayed as disconnectivity graphs. For those databases which could be considered as being 'complete' (i.e. all of those involving HemS), each showed one distinct low-energy funnel in their respective graphs. Further analysis showed that the bottoms of these funnels generally corresponded to the minima with the shortest NADH-haem distances. This structure was especially true for WT HemS, and it was noted that some of the single-point mutations disrupted this feature of the landscape. A steep funnel in an energy landscape with few competing kinetic traps is indicative in a biological context of a system which is optimised (i.e. has evolved) to achieve a particular structure. Typically, this principle has been used to determine the folding characteristics of proteins and nucleic acids. In this context, the steep funnels show that HemS is optimised to bring haem and NADH into close proximity. Though none of their databases are fully complete, of the homologues it would appear that HmuS has a similar funnel, but ChuS and ShuS do not.

Another advantage of being able to build up the WT HemS database quickly was that a pathway showing NADH unfolding and approaching haem could be fully connected for the first time. Further refinement of the database then allowed for alternative pathways to be identified, giving a detailed picture of the routes by which NADH can most effectively reach haem. The Dijkstra fastest pathway showed that this process was energetically favourable, with few large barriers to be traversed.

These pathways identified an interesting conformational change in residue Q210. This residue was shown to swing towards the pocket as NADH moved further inside, with other residues anchoring it in a convenient location to hydrogen-bond to the phosphate backbone of the ligand. This structure appeared to facilitate a number of the unfolding steps required of NADH to reach haem and, once these processes were

completed, Q210 then swung back to its original position. Such behaviour would be difficult to identify by experiment or using computational methods which only consider snapshots in the pathway or rely on coarse-graining the protein. R209 was also identified as a residue of interest due to its ability to hydrogen-bond to both haem and NADH when the two ligands are in close proximity.

The double phe-gate, first identified by Choy and Shang, was also studied in greater detail. It was shown that association of NADH with the edge of the main cavity caused the preferred conformation of the double phe-gate to change from open-open (OO) to closed-open (CO). This change was due to NADH hydrogen-bonding to a pair of residues, N106 and P169, which brought about a series of conformational changes, culminating in F104 closing and forming a T-shaped π-π interaction with F199. As NADH approaches haem, F104 typically opens up again, which alleviates steric crowding at that region of the pocket. It is possible, therefore, that this double phe-gate is engaging in a sophisticated regulatory mechanism to control the access of NADH to haem, which only comes into effect once NADH associates with the protein.

Bioinformatics was also utilised to further understand the phylogenetic context of HemS. This work showed that HemS is a member of a family of at least 5000 related haemoproteins derived from 218 different genera. The majority of these genera are pathogenic bacteria although some are non-pathogenic. The working hypothesis developed as a result of this project is that HemS and its homologues are used to break down haem under anaerobic conditions as a means of allowing a bacterium to extract iron from its host organism. Non-pathogenicity would suggest that for some of these bacteria, this process is either benign to the host or does not occur at all.

Analysing the sequences of these homologues revealed remarkable rates of conservation among some of the residues. Those directly involved with haem-binding (such as R102, H196 and R209) showed 100% rates of conservation, suggesting that all of the homologues are at least involved in binding to haem. Those residues thought to be involved in NADH-binding were not nearly as faithfully conserved, suggesting either that this region of the pocket is only required to bind to NADH weakly (as implied by the weak Michaelis constant for NADH determined by Sawyer) and therefore tolerates a wider range of possible residues, or the breakdown of haem by NADH is not the primary function of all of the homologues. The double phe-gate, which occupies a section of the cavity between the haem-binding and NADH-binding regions, has a much higher degree of conservation than any of the residues comprising the NADH-binding area. One possible reason for this difference is that the

double phe-gate helps to maintain the structural integrity of the main cavity, as determined from crystallographic studies and computational simulations, which together showed that F104AF199A HemS cannot maintain the pocket size as well as the WT. Another possible reason is that the double phe-gate is not just involved in regulating the approach of NADH to haem, but of alternative ligands as well.

One possible alternative function for certain haemoproteins in this class, which has already been demonstrated in the literature, is DNA-binding. It has been shown that PhuS, when not bound to haem, is capable of binding to the *prrF1* promoter, which is known to regulate the expression of certain nonessential iron-containing proteins. This homologue, at least, is therefore potentially capable of controlling its own concentration in the cell. ShuS has also been demonstrated to bind to DNA, but non-specifically. Studies with the analysis package, DNAbind, in the present work showed that PhuS and ShuS are likely to bind to DNA at different cavities, suggesting that the ability of ShuS to bind to DNA is for reasons other than to control its own expression *via prrF1*. DNAbind also provided good reasons as to why ChuS has not been shown to bind DNA, yet ShuS has, despite the two proteins being different in only five residues.

All but one of the mutants and homologues selected for the joint computational and experimental analyses were successfully synthesised using the protocol designed for WT HemS. In order to extract ShuS from the *E. coli* cells it was grown in, the protocol had to be adapted to account for possible aggregation with DNA.

All of the mutants and homologues were demonstrated to bind haem similarly to WT HemS, and to catalyse its reductive breakdown by NADH to give the same intermediate species and final HBP. All of the mutants were shown to deleteriously affect the rate of the breakdown of haem, although not by as much as expected. It therefore seems that HemS is optimised to break down haem, but that it can also tolerate a wide range of mutations to its NADH-binding residues and double phe-gate. Previous studies by Sawyer had shown that mutations to the residues that bind haem at its iron centre, R102 and H196, effectively prevent haem from binding and therefore, by extension, severely limit the breakdown process. Together, these findings reinforce the conclusion derived from the bioinformatic study that these homologues have primarily evolved to bind to haem, and that they may be engaging in several functions, one of which is anaerobic, reductive degradation of the haem by NADH.

10.2 Future Work

Despite the investment of much effort and the use of a wide range of characterisation techniques, the exact structure of the HBP remains unsolved. The most promising avenue for definitively resolving this structure appears to be X-ray crystallography. Post-reaction crystals with the intact HBP still inside (as determined by their bright purple colour) can reliably be produced, and crystallographic studies have shown that the protein structures can be clarified to high resolution. The next stage of this research is to find a method of protecting the integrity of the HBP as cryoprotectant is added. A series of assays using different combinations of cryoprotectants at different concentrations may alight on a formulation that does not lead to HBP degradation. Were this approach to work, further studies using crystallography to determine the intermediate structure may be possible as well. By soaking *holo*-HemS crystals with NADH and freezing collections of them rapidly through a series of time-steps, it may be possible to track the evolution of haem to the HBP over time. Depending on how much the reaction can be slowed down and how high the temporal resolution could be increased to, such experiments could show exactly which atom of haem the hydride from NADH is transferred over to. As stocks have been retained of all of the homologues made, further assays could be attempted to crystallise and resolve these proteins too. This information would be particularly helpful for HmuS and ShuS, which have never been crystallised before; comparing these structures against HemS and ChuS, respectively, could shed new light on the subtle differences between these HemS/HmuS and ChuS/ShuS homologue pairs.

Marked progress has been made in developing an understanding of the kinetics of the reaction, and how it varies between the mutants and homologues. Unfortunately, neither the computational, nor the experimental approaches produced full quantitative kinetic analyses. To expedite the expansion of databases, the frequencies of new stationary points were not calculated from a full consideration of the Hessian. These frequencies are not required when considering potential energies, but are important when considering free energies or actual rates. There are routines used by the Wales group that can readily calculate and add in these frequencies, although such calculations would most likely take 3-4 months to run. In terms of the experimental kinetic analysis, it is suspected that limitations to the stopped-flow technique prevented its full realisation. Although stopped-flow spectroscopy is a powerful method for tracking reactions over short timescales, the lamp can fluctuate and is difficult to normalise. For many reactions, where strong absorbances are involved, this fluctuation does not prove to be a significant issue. However, for the breakdown of haem by NADH, absorbances tend to be low, particularly for the

intermediate species, making these fluctuations more prominent. A possible way to resolve this problem would be to change the light source from a xenon arc-lamp to monochromatic LEDs. Such LEDs have been shown to be very intense and highly stable.[248] Experiments using LEDs would be required to run at specific wavelengths rather than over a range, as is possible with the xenon lamp. For experiments where the relationship between haem and the HBP is of interest, these LEDs would therefore not be suitable. They would be suitable, however, if the object of the experiment was to track the behaviour of a single compound (e.g. the intermediate at 806 nm). Another possible reason that curve fitting did not prove possible is that the assumed mechanism is not actually correct. There could, for example, be further, silent intermediate species and/or alternative, competing pathways. A deeper understanding of the mechanism, which could possibly be achieved by the crystallography experiments described above, could lead to a model being developed which would allow for more accurate simulation and fitting to the data.

To further understand the NADH-binding site, alternative mutations and combinations thereof could be applied. Mutating R250 and Q313 could prove to be interesting, as these residues are both implicated in binding to the NADH phosphate backbone, and are therefore possibly important in stabilising the molecule as it unfolds.

The comparison between homologues could also be expanded. Those chosen (HmuS, ChuS and ShuS) all had at least 66% identity with HemS. By extending this study to more distant homologues, the juncture at which haem breakdown ceases (if indeed that does happen) could be pinpointed, and the combination of residues required to bring this change about could perhaps even be identified. PhuS would be a sensible homologue to start with, considering that this protein has already been studied in detail by other groups. It would also be interesting to study homologues from non-pathogenic bacteria to determine whether any of these haemoproteins can catalyse the haem breakdown reaction.

PhuS, together with ShuS, could be subjected to DNA-binding assays to test the hypotheses suggested by the bioinformatic study. HemS, HmuS and ChuS should also be subjected to these assays to determine definitively whether these proteins are indeed not able to bind DNA. Attempts could be made to crystallise PhuS bound to the *prrF1* promoter. Provided it does not aggregate, crystallisation could also be attempted on ShuS bound to certain DNA sequences. The structures obtained could then be compared against *apo*-forms or those with haem bound instead, in order to determine what kind of changes, if any, DNA-binding causes.

To further understand the role of these S proteins within their biological con-

texts, protein-protein interactions could be investigated by pull-down assays. By running these assays under both haem-rich and haem-limiting conditions, it may even be possible to determine if such interactions with the S proteins are dependent upon the presence of haem or not. A series of gene knockouts could also be applied to the *hem* operon before attempting to grow the parent organism under iron-replete and iron-limiting conditions. The working hypothesis of this thesis is that the catalytic constant of HemS is low due to the breakdown reaction being limited by product inhibition, and that this inhibition may be due to the absence of a haemoprotein which can capture the HBP from HemS and transport it to another location. Pull-down assays and gene knockout studies provide a promising avenue for confirming/disproving this hypothesis.

The most immediate concern of future computational work on this project should be to complete the refinement of the HmuS, ChuS and ShuS databases. Refinement of the HmuS database would allow for a more detailed comparison with HemS to be made, whereas refinement of the ChuS and ShuS databases may cause alternative funnels to emerge and therefore reveal more about possible alternative functions. These databases could then be subjected to detailed kinetic studies with a full consideration of the frequencies too.

If **CHECKSPMUTATE** could be made more robust to comprehensive changes to the protein sequence, then a wider range of homologues could be studied. Attempts to transform HemS to PhuS could be revisited. These two proteins have 42.6% homology, and none of the transformed stationary points could be successfully re-optimised during the course of the project. It is thought that this rate of failure was due to three related reasons: firstly, the large number of residues along the sequence being mutated increased the likelihood that at least one would lead to steric clashes; secondly, five residues, including two pairs, had to be added at various points in the sequence, which significantly disrupts those local environments and also leads to increased likelihood of steric clashes; and thirdly, the main cavity of PhuS has a larger volume than that of HemS, and so the cavity as found in the HemS stationary points may not be compatible when these stationary points are set to use PhuS residues. A possible solution to these issues would be to expand the main cavity to give a buried volume akin to that found experimentally with PhuS. This expansion would be implemented after the transformation of the sequence, but before any attempt to optimise the new structures. The expansion would also be required to preserve the conformations of the residues in the chains that define the cavity. It could be achieved by artificially extending the bond lengths making up the protein backbone that surround the cavity. This change would need to be done in

very small increments. On the assumption that such an expansion would ultimately yield a structure with no steric clashes between the residues in the main cavity, and that none had been introduced to other parts of the protein in the meantime, optimisation could then be performed, which would relax these backbone bonds, bringing them back to their original lengths. Such a procedure would increase the likelihood that individual residue conformations would change significantly during the optimisation and so the stationary point of the transformed PhuS protein would not necessarily be a faithful replication of its HemS equivalent. This sacrifice may just be necessary to achieve a reasonable number of successfully reoptimised stationary points for proteins with low % sequence homology. Thorough refinement of the new database would be essential.

Rather than changing the protein, extra focus could also be given to transforming the ligands. This feature is already possible with **CHECKSPMUTATE**, with tests having been carried out on changing NADH to NADPH and NAD^+, respectively. However, care must to taken to ensure that the new ligand shares some features with the ligand found in the template. Otherwise, the template becomes largely redundant, as the new ligand is more likely to bind to the protein *via* different residues and in different conformational arrangements.

The computational approach could also be expanded to investigate the haem breakdown mechanism itself. Calculations in this thesis were concerned primarily with the approach of NADH to haem, and the way these two ligands orientate themselves prior to hydride transfer. However, using QM/MM, proposed intermediate and product structures could be tested. There is still uncertainty over which atom in haem the transferred hydride bonds to. A reasonable assumption is that it attaches to the β-*meso*-carbon of haem, as one of the bonds this carbon is engaged in must break to open up the tetrapyrrole. However, the lowest energy minima in most of the databases studied featured a hydride within 3 Å of the haem C5 methyl. Comparing the stabilities of molecules where hydride has attached to either of these positions could therefore provide clues as to where the hydride transfer occurs. QM/MM calculations are typically computationally expensive and so starting parameters for intermediate and product candidate structures should be chosen carefully, ideally only after accurate resolution of the HBP and, if possible, the 806 nm intermediate by X-ray crystallography. These calculations could then shed light on the molecular orbitals involved in hydride transfer and possible subsequent sigmatropic rearrangements, as well as giving an indication of the kinetic barriers involved in this haem breakdown process. Detailed QM/MM and QM-cluster calculations have already been carried out by a group at Shahid Beheshti University together with Ulf Ryde

at Lund University, which show the breakdown of haem *via* the 'canonical HO mechanism' to verdohaem and biliverdin.[249–252] Their approach has proved to be in good agreement with experiment, and could perhaps be adapted to investigate this novel reductive haem breakdown process as well.

10.3 A Broader View

In July of this year, two ground-breaking papers were published in Nature by the developers of AlphaFold, an artificial intelligence program managed by Google's DeepMind which uses deep learning to predict protein structure.[253;254] The same month, a further paper was published in Science by the developers of an alternative protein-prediction tool, known as RoseTTaFold.[255]

Together, these papers showed that protein structures can regularly be predicted to atomic levels of accuracy, even when there are no homologous structures available from experiment. Structure prediction is also scalable, and this feature was exploited using AlphaFold to predict 98.5% of the entire human proteome. This achievement is remarkable, especially when it is considered that it has taken many decades of diligent work for experimentalists to structurally determine 17% of the total number of residues in human protein sequences.[256]

The ability to predict protein structures so quickly and to such consistently high levels of accuracy constitutes a significant advancement in the field of proteomics. It is not difficult to see how an accurate knowledge of the protein structures making up the human proteome will provide valuable insight for pharmaceutical and medical researchers.

However, to gain a true understanding of the human proteome, or of the proteomes of other species, a deeper comprehension of protein function is also required. As shown by a combination of experimental, computational and bioinformatic research in this thesis, homologues with very similar structures can engage in a variety of different functions. Care must therefore be taken by the scientific community going forward not to assign protein function based on structural homologies alone.

Thorough experimentation in the laboratory is without a doubt the gold standard approach to elucidating protein function. However, given constraints on both time and cost, computational research should not be discounted as an increasingly capable alternative. The methods used by the Wales group have been proven to accurately predict protein folding pathways for a wide variety of protein types and in a broad range of simulated environments. Whilst there has not been much atomistic protein-ligand research done in the group, it is hoped that this research on HemS and

its homologues demonstrates that this avenue is both feasible, and faithfully replicates/predicts features identified by experiment. The **CHECKSPMUTATE** subroutine, in particular, could hopefully prove to be useful in studying other protein-ligand problems, where a knowledge of the effects of mutations or a comparison between closely-related homologues is required. The 'template-based' approach for seeding and growing new databases was very much a proof-of-principle in this thesis to determine how well this method reflected findings in the laboratory. In the future, it is hoped that it may be possible to use such a method to effectively screen certain mutations to determine whether they give feasible proteins with interesting properties, before committing to full experimental studies.

Recent advances in artificial intelligence, such as those achieved by AlphaFold and RoseTTafold, are helping to rapidly expand the collective knowledge of the genomic and proteomic sciences. With such knowledge, it is important to develop a detailed understanding of the accumulated data, one aspect of which is being able to relate protein structure to function accurately. Bench-top experimentation, computational simulation and bioinformatic analysis have all been demonstrated to be important means of effecting such development, particularly when used together. The 21st century has indeed proved to be the 'century of biology' thus far, and Nature still has many secrets to share.

Appendix A

% Homologies of Operon Proteins

There are both similarities and differences between the five operons of interest in this report. A simple, but informative, approach to determining the degree of similarity of proteins is to compare their sequences.

To make these comparisons, particular genomes from each of the five types of bacteria were considered. The Accession Numbers for these genomes are listed in Table A.1. The annotations which came with these were often incomplete or inconsistent with the naming conventions used in this report. Therefore, RefSeq (NCBI Reference Sequences) are provided in Table A.2. ChuW, ChuX and ChuY were discovered to be an integral part of the *chu* operon. In contrast, the roles of HemW, HemX and HemY in *hem* are less clear. Complicating matters is the fact that there are other proteins coded for in the *Y. enterocolitica*, *Y. pestis*, *E. coli*, *S. dysenteriae* and *P. aeruginosa* genomes that are commonly given the appellations HemW, HemX and HemY as well. These genes are far away from the respective *hem*/*hmu*/*chu*/*shu*/*phu* operons, suggesting they are part of alternative haem-uptake mechanisms available to these bacteria. Using these reference sequences, percentage homologies between the proteins could be determined. These results are split into individual Tables A.3 to A.13.

Bacterium	Accession Number
Yersinia enterocolitica	NC_017564.1
Yersinia pestis	NC_017168.1
Escherichia coli O157:H7	NC_002695.2
Shigella dysenteriae	CP000034.1
Pseudomonas aeruginosa	NC_002516.2

Table A.1: Accession numbers which provide the bacterial proteomes used to select the proteins and their sequences for the derivations of % homologies.

XxxX	*Y. enterocolitica*	*Y. pestis*	*E. coli*	*S. dysenteriae*	*P. aeruginosa*
XxxP	**HemP**	**HmuP**	**ChuP**	**ShuP**	**PhuP**
	WP_005181153.1	No RefSeq	N/A	N/A	N/A
XxxR	**HemR**	**HmuR**	**ChuR**	**ShuR**	**PhuR**
	WP_005181150.1	WP_002209062.1	NP_312407.1	WP_000089574.1	NP_253398.1
XxxS	**HemS**	**HmuS**	**ChuS**	**ShuS**	**PhuS**
	WP_005156541.1	WP_002209061.1	NP_312406.1	WP_001017208.1	NP_253397.1
XxxT	**HemT**	**HmuT**	**ChuT**	**ShuT**	**PhuT**
	WP_005156544.1	WP_002209060.1	NP_312409.3	WP_001081846.1	NP_253396.1
XxxU	**HemU**	**HmuU**	**ChuU**	**ShuU**	**PhuU**
	WP_005156547.1	WP_002209059.1	NP_312413.1	WP_005019013.1	NP_253395.1
XxxV	**HemV**	**HmuV**	**ChuV**	**ShuV**	**PhuV**
	WP_005156550.1	WP_002209058.1	NP_312414.1	WP_001626196.1	NP_253394.1
XxxW	**HemW′**	**HmuW**	**ChuW**	**ShuW**	**PhuW**
(Operon)	WP_005156531.1	WP_002209066.1	NP_312410.1	No RefSeq	N/A
XxxX	**HemX′**	**HmuX**	**ChuX**	**ShuX**	**PhuX**
(Operon)	WP_005156534.1	WP_002209065.1	NP_312411.1	WP_000020038.1	N/A
XxxY	**HemY′**	**HmuY**	**ChuY**	**ShuY**	**PhuY**
(Operon)	WP_013649113.1	WP_002209064.1	NP_312412.1	WP_000189360.1	N/A
XxxW	**HemW**	**HemW**	**HemW**	**HemW**	**HemW**
(Other)	WP_013649113.1	WP_002209064.1	NP_312412.1	WP_000239935.1	NP_249077.1
XxxX	**HemX**	**HemX**	**HemX**	**HemX**	**HemX**
(Other)	WP_005166100.1	WP_002211463.1	NP_312760.1	WP_000138987.1	NP_253945.1
XxxY	**HemY**	**HemY**	**HemY**	**HemY**	**HemY**
(Other)	WP_005166099.1	WP_002211462.1	NP_312759.1	WP_000921781.1	NP_253944.1

Table A.2: Accession numbers for each of the protein sequences used in the derivation of % homologies. The W, X and Y proteins corresponding to those found in the *hem* operon are labelled with a dash, ′. The W, X and Y proteins annotated (Other) correspond to those found outside of the *hem*, *hmu*, *chu*, *shu* or *phu* operons.

	HemR	HmuR	ChuR	ShuR	PhuR
HemR	–	84.7	66.3	66.1	24.8
HmuR	91.7	–	68.0	67.7	24.3
ChuR	79.4	81.0	–	99.5	25.7
ShuR	79.4	81.0	99.8	–	25.4
PhuR	38.0	38.3	39.9	39.8	–

% Identity

% Similarity

Table A.3: XxxR Homologies.

	HemS	HmuS	ChuS	ShuS	PhuS
HemS	–	89.6	66.8	66.2	42.6
HmuS	94.8	–	67.1	66.5	43.8
ChuS	78.2	78.8	–	98.5	41.1
ShuS	77.9	78.5	98.5	–	40.8
PhuS	56.5	58.3	56.8	56.5	–

% Identity

% Similarity

Table A.4: XxxS Homologies.

	HemT	HmuT	ChuT	ShuT	PhuT
HemT	–	90.7	36.2	36.6	29.1
HmuT	94.3	–	34.8	35.2	29.2
ChuT	55.9	57.0	–	97.4	36.1
ShuT	55.5	56.6	99.3	–	36.5
PhuT	33.1	32.2	58.0	58.0	–

% Identity

% Similarity

Table A.5: XxxT Homologies.

	HemU	HmuU	ChuU	ShuU	PhuU
	% Identity				
HemU	–	93.1	66.9	66.6	49.2
HmuU	97.0	–	67.5	67.8	48.6
ChuU	81.1	80.4	–	99.4	48.8
ShuU	80.4	80.4	99.4	–	48.8
PhuU	65.2	64.4	67.8	67.8	–
	% Similarity				

Table A.6: XxxU Homologies.

	HemV	HmuV	ChuV	ShuV	PhuV
	% Identity				
HemV	–	90.3	57.9	58.7	45.5
HmuV	95.7	–	59.2	60.0	44.0
ChuV	72.0	73.3	–	98.4	41.2
ShuV	71.7	72.9	99.6	–	41.2
PhuV	58.4	57.6	57.6	57.6	–
	% Similarity				

Table A.7: XxxV Homologies.

	HemW′	HmuW	ChuW	ShuW	PhuW
	% Identity				
HemW′	–	93.1	58.5	x	x
HmuW	97.5	–	57.8	x	x
ChuW	72.0	71.8	–	x	x
ShuW	x	x	x	–	x
PhuW	x	x	x	x	–
	% Similarity				

Table A.8: XxxW (Operon) Homologies.

% Identity

	HemX′	**HmuX**	**ChuX**	**ShuX**	**PhuX**
HemX′	–	87.6	60.4	59.1	x
HmuX	91.7	–	59.4	58.1	x
ChuX	77.4	76.9	–	98.2	x
ShuX	76.1	75.6	98.2	–	x
PhuX	x	x	x	x	–

% Similarity

Table A.9: XxxX (Operon) Homologies.

% Identity

	HemY′	**HmuY**	**ChuY**	**ShuY**	**PhuY**
HemY′	–	85.2	55.5	55.9	x
HmuY	92.6	–	55.9	55.9	x
ChuY	67.8	68.2	–	97.1	x
ShuY	68.2	68.7	99.0	–	x
PhuY	x	x	x	x	–

% Similarity

Table A.10: XxxY (Operon) Homologies.

% Identity

	Y. ent.	***Y. pest.***	***E. coli***	***S. dys.***	***P. aer.***
Y. ent.	–	95.2	83.2	83.2	55.1
Y. pest.	97.1	–	81.4	81.4	54.0
E. coli	89.4	88.8	–	98.9	55.3
S. dys.	89.4	88.8	99.2	–	55.3
P. aer.	71.0	69.9	69.7	69.4	–

% Similarity

Table A.11: HemW (Non-Operon) Homologies.

	% Identity				
	Y. ent.	***Y. pest.***	***E. coli***	***S. dys.***	***P. aer.***
Y. ent.	–	90.9	68.0	68.3	27.1
Y. pest.	94.7	–	68.1	68.1	26.8
E. coli	81.6	81.2	–	97.7	26.5
S. dys.	81.6	80.9	98.0	–	26.2
P. aer.	44.5	44.5	47.0	46.2	–
	% Similarity				

Table A.12: HemX (Non-Operon) Homologies.

	% Identity				
	Y. ent.	***Y. pest.***	***E. coli***	***S. dys.***	***P. aer.***
Y. ent.	–	92.0	70.9	70.6	26.9
Y. pest.	95.5	–	70.6	70.4	26.9
E. coli	86.3	84.6	–	99.7	25.1
S. dys.	86.1	84.3	99.7	–	25.1
P. aer.	49.6	49.1	48.0	47.7	–
	% Similarity				

Table A.13: HemY (Non-Operon) Homologies.

Appendix B

Gene / Protein Sequences

DNA Sequences

Each line of data spans 50 bases. Red boxes denote areas of differentiation between the gene used in this thesis and that quoted in the literature – for HemS, this is European Nucleotide Archive (ENA) entry CAA54865.1, from Stojijlkovic & Hantke.[37] Table B.1 catalogues these differences. Formatting for sequences is from the dnaseq package.[257]

HmuS, ChuS and ShuS inserts were bought in commerical plasmids from Thermofisher Invitrogen GeneArt. Following reconstitution into an appropriate plasmid (pET11d), each insert was sequenced. These sequences were then compared against those quoted in the literature (specifically, ENA entry AAC64867.1.[78] for HmuS, BAB37802.1.[258] for ChuS and AAC27810.1.[259] for ShuS), with any differences also highlighted in red.

Codon	**Lit. → Thesis**	**Δ to Residue**
241-243	tac → tat	TYR → TYR
484-486	att → aat	ILE → ASN
493-495	ttg → tta	LEU → LEU
694-696	ttg → tta	LEU → LEU
997-999	gac → gag	ASP → GLU
1000-1002	gaa → caa	GLU → GLN

Table B.1: Codon differences between *apo*-HemS as quoted in the literature *versus* those used in this study.

WT HemS – DNA Sequence

```
   1 atgagcaaat caatatacga gcagtatcta caagctaaag cagataatcc
  51 gggcaaatat gcgcgcgatt tggccacgct gatggggatt tcagaagcgg
 101 aactgaccca tagccgcgtt agtcatgatg ccaaacgtct gaaaggtgat
 151 gcccgcgcac tactggccgc attggaagct gtcggtgagg tcaaagctat
 201 cacccgcaac acctatgccg tacatgagca aatgggccgt tatgaaaatc
 251 aacatctgaa tggccatgct ggtttgatcc tcaatccacg caatttagat
 301 ttacgcctgt tcctcaacca gtgggccagc gcattcacgc tgacagaaga
 351 aactcgccac ggtgtacgcc atagcatcca gtttttcgac catcaaggcg
 401 atgctctgca taaagtgtat gtcactgaac aaactgacat gccagcctgg
 451 gaagcgctac tggcgcagtt tatcaccaca gaaaatccag agttacagct
 501 agagccactg agcgcacctg aagtcactga accgacagcc accgatgaag
 551 ctgtcgatgc tgaatggcgt gctatgactg acgtgcatca gttcttccag
 601 ttgctcaaac gcaataattt gacccgtcag caagccttcc gtgccgtggg
 651 taatgatctg gcttatcagg ttgataacag ttctctgacc cagttactga
 701 acattgctca gcaagaacag aatgaaatca tgatttttgt gggtaaccgt
 751 ggctgtgtac aaatattcac cggcatgatt gaaaaggtta caccacatca
 801 agattggatt aatgttttca accagcgctt cacgctgcat ctgattgaaa
 851 caacgattgc tgaaagctgg attacccgca agccaacaaa agacggtttc
 901 gtgaccagtt tggaactgtt tgctgctgat ggcacccaaa ttgcacaact
 951 ttacggtcag cgcaccgaag gccagccaga acaaacgcaa tggcgtgagc
1001 aaattgctcg cctcaataat aaggatatcg ccgcatga
```

HmuS – DNA Sequence

```
   1 ---aacgcat cattatacca acaatatgta caggctaaag cagagcaccc
  51 tggcaaatat gcccgtgatt tagccaccct gatggggatt tcagaagcag
 101 agctgaccca tagccgcgtc gggcatgatg caaaacgttt acaaagtgat
 151 gctcgtgcat tattggccgc attggaatcc gtcggcgaag tcaaagccat
 201 tacccgcaac acctatgcag ttcatgagca agtgggccgc tatgagaacc
 251 aacacttaaa tggtcatgca gggttaatcc tcaatccacg cgccttggac
 301 ctccggttat tcctgaatca gtgggcaagc gcctttacac tgaccgaaga
 351 gacccgccac ggcgtgcgcc atagcatcca atttttcgac catcagggcg
 401 atgcattaca caaagtgtat gtgacagaac agacagatat gtctgcctgg
 451 gaagccttgc tggcacaatt tatcatcccg gaaaacccgg cattgcagtt
 501 agaacctttg agcaccccag aagcggtaga acctacagcc gatgatgcaa
 551 ccgtggatag cgaatggcgt gccatgaccg atgtacacca gttcttccaa
 601 ctgcttaaac gcaataatct gacccgtcag caggcgttcc gcgctgttgg
 651 tgatgatctg gcttaccagg tcgataacaa ctcactgact cagctgttgc
 701 acatcgccca gcaagatcag aacgagatca tgatttttgt cggcaaccgc
 751 ggctgtgtac aaattttcac cggcctgatt gaaaaagtca caccacacaa
 801 cgaatggatt aatgtcttca atcagcgctt tacactgcat ctgatcgaaa
 851 cggccattgc cgaaagctgg atcacccgca aaccaacaaa agacggtttt
 901 gtcaccagcc tagaactgtt tgctgctgat ggtactcaac ttgcccaact
 951 ctacggccag cgcaccgaag ggcagccaga acaaaaccaa tggcgtgaac
1001 agattgcccg cctaatcaac aaggatatcg ccgcatga
```

ChuS – DNA Sequence

```
   1 atgaaccact acacacgctg gcttgagtta aaagaacaaa atcccggaaa
  51 gtacgcgcgt gacatcgcag ggttaatgaa tattagagaa gcagaactgg
 101 catttgcacg cgtcacgcac gatgcgtggc ggatgcacgg cgatatccgt
 151 gaaattctgg cggcgctcga aagtgttggc gaaaccaaat gtatttgtcg
 201 taatgaatat gcagtccatg agcaagttgg tacgttcaca aaccagcatt
 251 tgaacggaca tgccggattg atcctcaatc cgcgcgcgct ggatttacgt
 301 ctgtttctca atcaatgggc cagtgttttc cacatcaaag aaaacacggc
 351 tcgtggcgaa cgccagagta ttcagttctt tgatcatcag ggcgatgcat
 401 tactaaaagt ttatgccacc gacaataccg atatggcggc atggagtgag
 451 cttctggcac ggtttatcac cgatgagaat acgccgcttg agttaaaagc
 501 cgttgatgcg ccagttgttc aaacgcgagc cgatgccact gtggtcgagc
 551 aagagtggcg ggcgatgacc gacgttcatc agtttttac gttgctcaag
 601 cgccacaacc tgacgcgcca acaggcgttc aatctggtgg cagacgattt
 651 ggcctgcaaa gtatccaaca gtgcgttggc gcaaattctt gaatctgcac
 701 agcaggatgg taatgaaatc atggtgtttg ttggcaaccg tggctgcgta
 751 cagattttca ccggtgtggt agaaaaagtg gtgccaatga aaggttggct
 801 gaatattttc aacccgacgt ttactcttca tctattagaa gagagcattg
 851 ctgaagcctg ggttacccgt aaaccgacca gcgatggcta cgtaaccagt
 901 ctggaattgt ttgcccatga tggtacgcag atagcgcaac tttatggtca
 951 acgtacagaa ggcgaacagg agcaagcgca atggcgtaag caaattgctt
1001 cgctgatacc ggaaggcgtt gctgcataa
```

ShuS – DNA Sequence

```
   1 ---aaccact acacacgctg gcttgagtta aaagaacaaa atcccggaaa
  51 gtacgtgcgt gacatcgcag ggttaatgaa tattagagaa gcagaactgg
 101 catttgcacg agtcacgcac gatgcgtggc ggatgcgcgg cgatatccgt
 151 gaaattctgg cggcgctcga aagtgttggc gagaccaaat gtatttgccg
 201 taatgaatat gcagtccatg agcaagttgg tgcgttcaca aaccagcatt
 251 tgaatggaca tgccggattg atcctcaatc cacgcgcgct ggatttacgt
 301 ctgtttctca atcaatgggc cagtgttttc cacatcaaag aaaacacggc
 351 tcgtggcgaa cgccagagga ttcagttctt tgatcatcag ggcgatgcat
 401 tactaaaagt ttatgccacc gacaataccg atatggcggc atggagtgag
 451 cttctggcac ggtttatcac cgatgagaat atgccgcttg agttaaaagc
 501 cgttgatgcg ccagttgttc aaacgcgagc cgatgccact gtggtcgagc
 551 aagagtggcg agcgatgacc gacgttcatc agtttttac gttgctcaag
 601 cgccacaacc tgacgcgcca acaggcgttc aatctggtgg cagacgattt
 651 ggcctgcaaa gtatccaaca gtgcgttggc gcaaattctt gaatctgcac
 701 agcaggatgg taatgaaatc atggtgtttg ttggcaaccg tggctgcgta
 751 cagattttca ccggtgtggt agaaaaagtg gtgccaatga aaggttggct
 801 gaatattttc aacccgacgt ttactcttca tctattagaa gagagcattg
 851 ctgaagcctg ggttacccgt aaaccgacca gcgatggcta cgtaaccagt
 901 ctggaattgt ttgcccatga tggtacgcag atagcgcaac tttatggtca
 951 acgtacagaa ggcgaacagg agcaagcgca atggcgtaaa caaattgctt
1001 cgctgatacc ggaaggcgtt gctgcataa
```

Protein Sequences (Experimental)

For HemS, of the six different codons between the literature and the sample used in the present study, three lead to different residues being coded for. All three of these residues are situated on solvent-exposed loops, which are nowhere near the main binding pocket. It was therefore concluded that these mutations would not affect functionality, and so were not corrected. In any case, the crystal structure for HemS (2J0R)[95] corresponds with the present study rather than the ENA sequence at positions 162 (c.f. codon 484-486) and 334 (c.f. codon 1000-1002), respectively. In other words, there is only one deviation in residue type from the published crystal structure for *apo*-HemS and the WT structure used in the present study (a conversion from ASP to GLU at position 333).

In reality, there is a His_6-tag encoded in the expression vector along with a thrombin cleavage site. Once cleaved, a glycine and a serine residue are retained at the N-terminus, attached to the first methionine residue. As they are situated at a solvent-exposed region far from the main cavity, these two are also thought not to significantly affect the functionality of HemS. They are not included in the sequences below in order to keep the residue indices consistent with those quoted throughout the thesis.

For each experimental protein sequence, residues highlighted in red denote differences between those used in this work, and those quoted in the literature. For HemS, the published crystal structure of *apo*-HemS (2J0R) is used as the literature reference, whereas for HmuS, ChuS and ShuS the reference sequences are from ENA codes AAC64867.1.,[78] BAB37802.1.[258] and AAC27810.1.[259] respectively. Residues highlighted in blue and detailed in Table B.2 denote differences between the experimental and computational *apo*-HemS sequences. Where a residue differs both from the literature and the computational sequence, it is highlighted in maroon. These are also detailed in Table B.2.

Reside Location	Experimental	Computational
1-3	MSK	- - -
197	Q	E
333	E	D
342-346	DIAA*	- - - - -

Table B.2: Residue differences between the experimental and computational sequences used for *apo*-HemS.

WT HemS – Protein Sequence (Experimental)

```
  1 MSKSIYEQYL QAKADNPGKY ARDLATLMGI SEAELTHSRV SHDAKRLKGD
 51 ARALLAALEA VGEVKAITRN TYAVHEQMGR YENQHLNGHA GLILNPRNLD
101 LRLFLNQWAS AFTLTEETRH GVRHSIQFFD HQGDALHKVY VTEQTDMPAW
151 EALLAQFITT ENPELQLEPL SAPEVTEPTA TDEAVDAEWR AMTDVHQFFQ
201 LLKRNNLTRQ QAFRAVGNDL AYQVDNSSLT QLLNIAQQEQ NEIMIFVGNR
251 GCVQIFTGMI EKVTPHQDWI NVFNQRFTLH LIETTIAESW ITRKPTKDGF
301 VTSLELFAAD GTQIAQLYGQ RTEGQPEQTQ WREQIARLNN KDIAA*
```

HmuS – Protein Sequence (Experimental)

```
  1 -NASLYQQYV QAKAEHPGKY ARDLATLMGI SEAELTHSRV GHDAKRLQSD
 51 ARALLAALES VGEVKAITRN TYAVHEQVGR YENQHLNGHA GLILNPRALD
101 LRLFLNQWAS AFTLTEETRH GVRHSIQFFD HQGDALHKVY VTEQTDMSAW
151 EALLAQFIIP ENPALQLEPL STPEAVEPTA DDATVDSEWR AMTDVHQFFQ
201 LLKRNNLTRQ QAFRAVGDDL AYQVDNNSLT QLLHIAQQDQ NEIMIFVGNR
251 GCVQIFTGLI EKVTPHNEWI NVFNQRFTLH LIETAIAESW ITRKPTKDGF
301 VTSLELFAAD GTQLAQLYGQ RTEGQPEQNQ WREQIARLIN KDIAA*
```

ChuS – Protein Sequence (Experimental)

```
  1 MNHYTRWLEL KEQNPGKYAR DIAGLMNIRE AELAFARVTH DAWRMHGDIR
 51 EILAALESVG ETKCICRNEY AVHEQVGTFT NQHLNGHAGL ILNPRALDLR
101 LFLNQWASVF HIKENTARGE RQSIQFFDHQ GDALLKVYAT DNTDMAAWSE
151 LLARFITDEN TPLELKAVDA PVVQTRADAT VVEQEWRAMT DVHQFFTLLK
201 RHNLTRQQAF NLVADDLACK VSNSALAQIL ESAQQDGNEI MVFVGNRGCV
251 QIFTGVVEKV VPMKGWLNIF NPTFTLHLLE ESIAEAWVTR KPTSDGYVTS
301 LELFAHDGTQ IAQLYGQRTE GEQEQAQWRK QIASLIPEGV AA*
```

ShuS – Protein Sequence (Experimental)

```
  1 -NHYTRWLEL KEQNPGKYVR DIAGLMNIRE AELAFARVTH DAWRMRGDIR
 51 EILAALESVG ETKCICRNEY AVHEQVGAFT NQHLNGHAGL ILNPRALDLR
101 LFLNQWASVF HIKENTARGE RQRIQFFDHQ GDALLKVYAT DNTDMAAWSE
151 LLARFITDEN MPLELKAVDA PVVQTRADAT VVEQEWRAMT DVHQFFTLLK
201 RHNLTRQQAF NLVADDLACK VSNSALAQIL ESAQQDGNEI MVFVGNRGCV
251 QIFTGVVEKV VPMKGWLNIF NPTFTLHLLE ESIAEAWVTR KPTSDGYVTS
301 LELFAHDGTQ IAQLYGQRTE GEQEQAQWRK QIASLIPEGV AA*
```

Experimental/Computational Sequence Differences

HemS The residues removed at the N-terminal (MSK) and C-terminal (DIAA*) regions should not unduly affect protein functionality as they are far from the main cavity. Furthermore, these residues do not appear in any crystal structures anyway, whether in the literature or in the work described in this thesis. Conversion of residue 333 from glutamic acid to aspartic acid, being in a solvent-exposed loop, should not have much effect. Indeed, this change is simply a reversal of the mutation at residue 333, making the computational sequence at this region more faithful to the literature 2J0P structure than the experimental sequence. Another difference is at residue 197, where the glutamine from the experimental sequence has been changed to a glutamic acid. This seems to have been a mistake inherited from Choy,[21] and was not noticed until well into the project. Though this residue is situated very close to the main cavity, it is hoped that it will not significantly affect the protein functionality, due to the fact that it is pointing away from the cavity. Though glutamine is a polar, neutral residue, and glutamic acid polar, acidic, it is also hoped that their similar steric bulks will help to preserve the protein function as well.

HmuS Due to the computational sequence for HmuS being derived from that for HemS, some of the 'errors' from HemS were carried over during the **CHECKSPMUTATE** routine. Namely, the computational sequence for HmuS has some residues removed from the N-terminal region (NA) and the C-terminal region (DIAA*) with respect to the experimental sequence. Being part of outer loops, these differences should have little bearing on protein functionality. Also, E197 replaces Q197. This is equivalent to HemS.

ChuS The computational sequence for ChuS has some residues removed from the N-terminal region (M) and the C-terminal region (EGVAA*) with respect to the experimental sequence. Being part of outer loops, these differences should have little bearing on protein functionality.

ShuS The computational sequence for ShuS has some residues removed from the C-terminal region (GVAA*) with respect to the experimental sequence. Being part of an outer loop, this difference should have little bearing on protein functionality. Furthermore, there is a difference at residue 327, where a glutamic acid replaces a glutamine in the computational sequence, which was most likely due to an error carried over from HemS during the **CHECKSPMUTATE** routine. Being situated on an outer loop, this difference should also be of little consequence.

Sequence Transformations by CHECKSPMUTATE

CHECKSPMUTATE used the [WT HemS + Haem + NADH] sequence as a template. Those residues which were mutated to give one of the HemS mutants studied in this work (F104A, F104AF199A, F104I, F199A, R209A, R209K and Q210A) are highlighted in brown in the WT sequence.

For the HmuS, ChuS, ShuS and PhuS sequences, the residues which had to be mutated from the HemS template are highlighted in red. When a residue was removed, this is highlighted in blue, with a dash used as a placeholder for the missing residue. When a residue was inserted, this is highlight in green; a dash was also used as a placeholder where these residues were missing from the other homologues.

WT HemS Sequence (Computational)

```
  1 SIYEQYLQAK ADNPGKYARD LATLMGISEA ELTHSRVSHD AKRLKGDARA
 51 LLAALEAVGE VKAITRNTYA VHEQMGRYEN QHL--NGHAG LILNPRNLDL
101 RLFLNQWASA FTLTEETRHG VRHSIQFFDH QGDALHKVYV TEQTDMPAWE
151 ALLAQFITTE NPELQ-LEP- -LSAPEVTEP TATDEAVDAE WRAMTDVHEF
201 FQLLKRNNLT RQQAFRAVGN DLAYQVDNSS LTQLLNIAQQ EQNEIMIFVG
251 NRGCVQIFTG MIEKVTPHQD WINVFNQRFT LHLIETTIAE SWITRKPTKD
301 GFVTSLELFA ADGTQIAQLY GQRTEGQPEQ TQWRDQIARL NNK
```

HmuS Sequence (Computational)

```
  1 SLYQQYVQAK AEHPGKYARD LATLMGISEA ELTHSRVGHD AKRLQSDARA
 51 LLAALESVGE VKAITRNTYA VHEQVGRYEN QHL--NGHAG LILNPRALDL
101 RLFLNQWASA FTLTEETRHG VRHSIQFFDH QGDALHKVYV TEQTDMSAWE
151 ALLAQFITPE NPALQ-LEP- -LSTPEAVEP TADDATVDSE WRAMTDVHEF
201 FQLLKRNNLT RQQAFRAVGD DLAYQVDNNS LTQLLHIAQQ DQNEIMIFVG
251 NRGCVQIFTG LIEKVTPHNE WINVFNQRFT LHLIETAIAE SWITRKPTKD
301 GFVTSLELFA ADGTQLAQLY GQRTEGQPEQ NQWREQIARL INK
```

ChuS Sequence (Computational)

```
  1 NHYTRWLELK EQNPGKYARD IAGLMNIREA ELAFARVTHD AWRMHGDIRE
 51 ILAALESVGE TKCICRNEYA VHEQVGTFTN QHL--NGHAG LILNPRALDL
101 RLFLNQWASV FHIKENTARG ERQSIQFFDH QGDALIKVYA TDNTDMAAWS
151 FLLARFITDE NTPLE-LKA- -VDAPVVQT- RADATVVEQE WRAMTDVHQF
201 FTLLKRHNLT RQQAFNLVAD DLACKVSNSA LAQILESAQQ DGNEIMVFVG
251 NRGCVQIFTG VVEKVVPMKG WLNIFNPTFT LHLLEESIAE AWVTRKPTSD
301 GYVTSLELFA HDGTQIAQLY GQRTEGEQEQ AQWRKQIASL IPF-
```

ShuS Sequence (Computational)

```
  1 NHYTRWLELK EQNPGKYVRD IAGLMNIREA ELAFARVTHD AWRMRGDIRE
 51 ILAALESVGE TKCICRNEYA VHEQVGAFTN QHL--NGHAG LILNPRALDL
101 RLFLNQWASV FHIKENTARG ERQRIQFFDH QGDALIKVYA TDNTDMAAWS
151 FLLARFITDE NMPLE-LKA- -VDAPVVQT- RADATVVEQE WRAMTDVHQF
201 FTLLKRHNLT RQQAFNLVAD DLACKVSNSA LAQILESAQQ DGNEIMVFVG
251 NRGCVQIFTG VVEKVVPMKG WLNIFNPTFT LHLLEESIAE AWVTRKPTSD
301 GYVTSLELFA HDGTQIAQLY GQRTEGEQEQ AEWRKQIASL IPF
```

PhuS Sequence (Computational)

```
  1 ELYRAWQDLR AERPQLRARD AAALLQVSEG ELVASRVGID AVRLRPDWAA
 51 LLPALGELGP IMALTRNEHC VHERKGPYRF VTVSANGQMG LVVSP-DIDL
101 RLFLGGWNAV FAIAEETARG TQRSIQVFDQ QGVAVHKVFL AEASDVRAWE
151 PLVERLRAAE QDAVLALHFP RAPAAALVDA QIDAAALREG WAALKDTHHF
201 HALLKKHGAQ RTQALRLAGG EWAERLDNGD LAKLFEAAAE SGLPIMVFVG
251 NAHCIQIHTG PVCNLKWLDD WFNVLDPEFN LHLKTTGIAE LWRVRKPSTD
301 GIVTSWEAFD PDGELIVQLF GARKPGEPER DDWRELAESF KAIL
```

HemS Mutants

The primers used for point-mutations were as follows. Sequences are all shown from 5′ to 3′, and mutagenic bases are capitalised.

F104A

Forward Sequence: agatttacgcctgGCcctcaaccagtggg
Reverse Complement: cccactggttgaggGCcaggcgtaaatct

F104I

Forward Sequence: agatttacgcctgAtcctcaaccagtgggc
Reverse Complement: gcccactggttgaggaTcaggcgtaaatct

F199A

Forward Sequence: gactgacgtgcatcagttcGCccagttgctcaaacgc
Reverse Complement: gcgtttgagcaactggGCgaactgatgcacgtcagtc

R209A

Forward Sequence: cgcaataatttgaccGCtcagcaagccttccg
Reverse Complement: cggaaggcttgctgaGCggtcaaattattgcg

R209K

Forward Sequence: cgcaataatttgaccAAtcagcaagccttccg
Reverse Complement: cggaaggcttgctgaTTggtcaaattattgcg

Q210A

Forward Sequence: cgcaataatttgacccgtGCgcaagccttccg
Reverse Complement: cggaaggcttgcGCacgggtcaaattattgcg

Appendix C

Computing Free Energies

As discussed in Section 1.6.3, free energies can be derived from the underlying PES. This analysis is achieved by dividing the PES into catchment basins and applying the superposition approach,[260] which defines the canonical partition function, $Z(T)$, and density of states, $\Omega(E)$, as the summation of individual contributions from each minimum, i, thus giving,[133;134;261;262]

$$Z(T) = \sum_i Z_i(T) \quad \text{and} \quad \Omega(E) = \sum_i \Omega_i(E). \tag{C.1}$$

Meanwhile, the vibrational partition function of each local minimum can be estimated using the harmonic approximation (which should be valid as protein folding and protein-ligand problems do not typically involve covalent bond breaking/making events). Therefore, the partition function for each local minimum can be expressed as,

$$Z_i(T) = \frac{n_i e^{-\beta V_i}}{(\beta h \overline{\nu_i})^{\chi}} \quad , \tag{C.2}$$

where n_i is the number of distinct permutational isomers in minimum i, V_i is the potential energy for minimum i, $\overline{\nu_i}$ its geometric mean vibrational frequency and χ the number of vibrational degrees of freedom. β is simply $1/k_B T$, where k_B is the Boltzmann constant and T a user-defined temperature in K.

The free energy of each minimum i, $F_i^E(T)$, can then be calculated using,

$$F_i^E(T) = -k_B T \ln Z_i(T). \tag{C.3}$$

Eq. (C.1) can then be used to determine the total canonical partition function. Overall, therefore, this formulation encapsulates the equilibrium occupation probability for each minimum as a function of temperature, and is explicitly ergodic.[133;263]

From this framework, many other interesting properties can be derived. For ex-

ample, heat capacities can be generated by differentiating the total partition function as follows,

$$C^H(T) \;=\; \frac{1}{k_B T^2}\frac{\partial^2 \ln Z(T)}{\partial \beta^2}, \tag{C.4}$$

giving,

$$C^H(T) \;=\; \chi k_B \;+\; k_B T^2 \sum_i g_i(T) \left(\frac{\partial \ln P_i(T)}{\partial T} \right), \tag{C.5}$$

where $P_i(T)$ is the occupation probability of minimum i at temperature T, expressed as,

$$P_i(T) \;=\; \frac{n_i e^{-\beta V_i}/\overline{\nu_i}\chi}{\sum_\gamma n_\gamma e^{-\beta V_\gamma}/\overline{\nu_\gamma}\chi}, \tag{C.6}$$

and $g_i(T)$ is the change in occupation probability given by,

$$g_i(T) \;=\; \frac{\partial P_i(T)}{\partial T}. \tag{C.7}$$

A maximum in the sum in Eq. (C.5) corresponds to a maximum in the heat capacity. $g_i(T)$ and $\partial \ln P_i(T)/\partial T$ both necessarily have the same sign and so such peaks can be interpreted in terms of contributions from local minima with positive and negative temperature derivatives. Therefore, for a decomposition according to the normal mode approximation, this analysis can reveal which local minima are responsible for any given heat capacity feature.[263]

As well as heat capacity curves, it is possible to use vibrational densities of states to investigate kinetic properties. This approach is based on the principle that the coupling between all possible individual transitions will yield a description of the overall transition rates between particular states in a system.[133;134] Therefore, overall kinetic properties can be derived from a combination of unimolecular rate theory and master equation dynamics.

Transition state theory (TST)[264–266] is the unimolecular rate theory implemented in the Wales group program, **PATHSAMPLE**.[231] From this theory, the unimolecular rate constant through a transition state † from minimum i at temperature T is given by,

$$k_i^\dagger(T) \;=\; \frac{k_B T}{h}\frac{Z^\dagger}{Z_i} e^{-\Delta\beta V_i^\dagger}, \tag{C.8}$$

where $\Delta V_i^\dagger$ is the difference in energy between the minimum i and TS †, and $Z^\dagger$ is the partition function of the TS without the degree of freedom associated with the negative eigenvalue. This formulation shows that an equilibrium between the start and end minima *via* the TS is not required.[267;268] Also, Rice-Ramsperger-Kassel-Marcus (RRKM) theory[269–272] yields the same expression but from the viewpoint

of reactive flux flowing through a dividing transition surface.[134] Therefore, the dynamics of the overall system can be described using a master equation if they are Markovian.[273]

The master equation approach assumes that the system occupies individual minima i for periods long enough so that its motion to i+1 can be effectively decorrelated from the previous motion describing i–1 $\rightarrow i$.[134] Such a system is colloquially described as having no 'memory' between transitions, and is the signature of a Markovian system.

If $\mathbf{P}(t)$ is a vector describing the occupation probability for all states at time t, the change in the occupation probabilities can be expressed by a master equation,

$$\frac{dP_a(t)}{dt} = \sum_{b \neq a} [k_{ab} P_b(t) - k_{ba} P_a(t)]. \tag{C.9}$$

Here, k_{ab} is the rate constant for the transition to minimum a from minimum b as defined from TST, and P_a is the occupation probability of minimum a.

To solve the master equation analytically, the transition matrix, $\mathbf{W}$ is introduced, and Eq. (C.9) can be rewritten as,

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{W}\mathbf{P}(t). \tag{C.10}$$

In a collection of minima that are all connected, a zero eigenvalue arises in $\mathbf{W}$, which corresponds to the equilibrium occupation probability vector, $\mathbf{P}^{eq}$. Detailed balance is therefore obeyed, i.e.,

$$\sum_{b \neq a} [k_{ab} P_b^{eq} - k_{ba} P_a^{eq}] = 0. \tag{C.11}$$

Combining this result with the discrete path sampling approach (where transitions between two states, A and B are considered) requires all of the local minima to be sorted into the A, B or intervening I sets. Considering just the A and B sets, this approach gives rise to phenomenological rate constants, which are weighted sums of the contributions from all transition states from each discrete path joining the A and B regions. These phenomenological rate constants, k_{AB} and k_{BA}, can therefore be expressed as,

$$k_{AB} = \frac{1}{P_B^{eq}} \sum_{b \in B} \sum_{a \in A} k_{ab} P_b^{eq} \quad \text{and} \quad k_{BA} = \frac{1}{P_A^{eq}} \sum_{a \in A} \sum_{b \in B} k_{ba} P_a^{eq}. \tag{C.12}$$

For the complicated protein-ligand interactions being investigated in this thesis,

however, it would be unusual for the A and B regions to be directly connected. Therefore, the I region must be included too. A steady state approximation is applied to these I minima, suggesting that the rate of change in the probability of occupation for these minima tends to zero. This result can then be applied to the master equation. However, it is also possible to relax this criterion and find non-steady state rates, using committor probabilities, C_b^A and C_a^B, giving[134;137;274;275]

$$k_{AB}^{NSS} = \frac{1}{P_B^{eq}} \sum_{a \leftarrow b} \frac{C_b^A P_b^{eq}}{t_b} \quad \text{and} \quad k_{BA}^{NSS} = \frac{1}{P_A^{eq}} \sum_{b \leftarrow a} \frac{C_a^B P_a^{eq}}{t_a}. \tag{C.13}$$

Here, C_b^A denotes the probability of encountering a minimum in A before encountering one from B when following a random walk starting from minimum b in B. Moreover, t_b denotes the mean waiting time for a transition to any minimum in A from minimum b. Whereas such waiting times are automatically zero in the steady state approximation, an estimate for them must be made in the non-steady state condition. The new graph transformation (NGT) method, as it is applied in **PATHSAMPLE**, is one such method that can be used.[138]

Should the above analysis prove too complicated for the system under study, it is also possible to apply regrouping schemes to simplify the data.[147] This regrouping is achieved by lumping minima separated by barriers beneath a user-defined threshold.

Appendix D

Specifics of Method Development

Glossary of Terms for Method Development

CHECKSPMUTATE. Subroutine that mutates and reoptimises any user-specified residues for all of the stationary points from a template system.

CHECKSPMUTATE Strategy. Overall strategy, which uses **CHECKSPMUTATE** to mutate and reoptimise all of the stationary points of a template system, and then fills in the gaps using a combination of steepest-descent algorithms, DPS and, if required, **CONNECTUNC LOWESTTEST**.

CONNECTUNC LOWESTTEST. Subset of the larger **CONNECTUNC** subroutine that provides an efficient way to connect sub-databases, both by selecting which sub-databases to connect, and by selecting appropriate minima for connection attempts between these sub-databases

CONNECTUNC LOWESTTEST Strategy. Overall strategy, which uses **CONNECTUNC LOWESTTEST** to connect sub-databases efficiently. It is intended for databases that contain many sub-databases far apart in conformational space.

Converge. The act of optimising/reoptimising a given stationary point to within a given RMS force threshold value. Achievement of this value indicates a stable stationary point.

Database. The overall collection of stationary points for a particular system.

Fully connected pathway. A chain of minima and TSs between two minima of interest with no intervening gaps.

Gap. Any break in a pathway between two selected minima. This could be due to missing TSs and/or minima.

Mutation. A change of the atomic makeup of a selected residue, resulting in a list of new properties and of new coordinates.

Region. A particular part of the landscape, usually used to describe the stationary points surrounding a site.

Reoptimisation. Attempt to converge a mutated system to a stable stationary point.

Site. One of the original minima identified to be of interest for a particular system. In the case of the [WT HemS + Haem + NADH] system, there were five such sites, based on the work by Choy, Shang and the author.

Sub-database (SD). A subset of a database where all of the minima and TSs are interconnected. In this work, sub-databases typically arose by sampling around a site, which provided a network of stationary points in the landscape surrounding that site. These stationary points tended to be close to one another in conformational space. As long as all of these stationary points are connected to one another, they constitute a sub-database.

System. The collection of atoms being studied, which could comprise just one molecule (e.g. a protein) or many molecules (e.g. a protein with ligands). If any of the residues in the protein are mutated, or any of the ligands changed, the resulting collection of atoms should be considered as a new system. Please note that, for easier readability, the [WT HemS + Haem + NADH] system is typically referred to in the text simply as the WT HemS system, with the inclusion of haem and NADH being implied. This shorthand is also true for the mutants and homologues of HemS.

Template. The original system studied. Its database is usually derived from more traditional sampling methods. To study a mutated system, the stationary points from this original system can be used as a basis for determining the likely coordinates of the respective stationary points of the new system.

Template-based Strategy. A strategy that uses a template system to derive others. It is synonymous with the **CHECKSPMUTATE** strategy.

Figures Describing the Methods Developed

Fig. D.1 illustrates the operation of the overall strategy involving **CONNECTUNCLOWESTTEST** to find connections between widely separated sub-databases.

Fig. D.2 illustrates the operation of the overall **CHECKSPMUTATE** strategy to find a fully connected pathway of a mutated system, using the fully connected pathway from a similar system as a template.

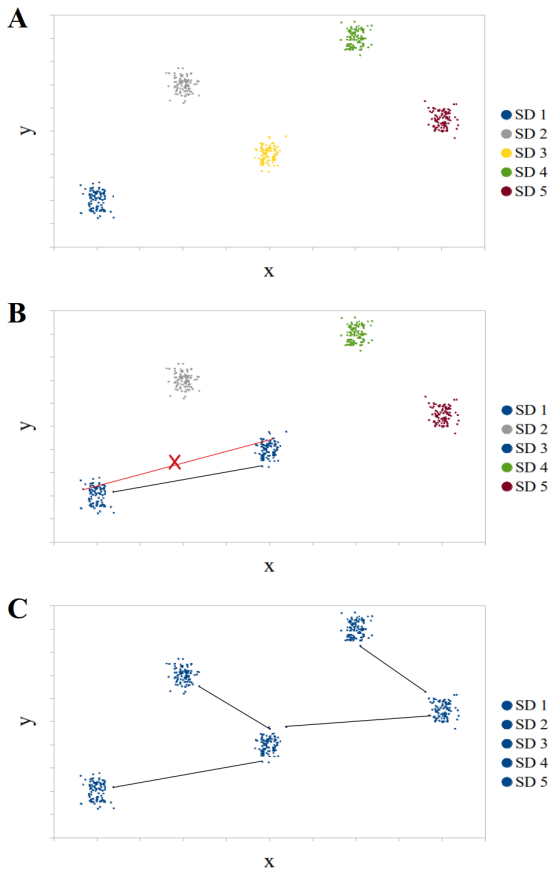


Figure D.1: Cartoon representing the overall CONNECTUNC LOWESTTEST strategy. Each point represents a minimum in an energy landscape, colour-coded according to which of the five original sub-databases (SD) they belong to. x and y are arbitrary parameters to give a sense of how close these minima are in conformational space. (**A**) None of the five sub-databases are connected to one another. (**B**) CONNECTUNC LOWESTTEST is used to identify which two minima from two separate sub-databases are closest in conformational space, and select them for a connection attempt. Alternative pairs of minima, such as that represented by the red line with the cross through it, are not considered for connection attempts unless the closer pairs fail. Once a connection has successfully been established between two pairs in two separate sub-databases, all of the minima comprising these sub-databases become connected. This situated is indicated by the colour change of SD 3 from yellow to blue. (**C**) CONNECTUNC LOWESTTEST also identifies which sub-databases to try to connect. The figure shows two minima between SD 2 and SD 3 in closer proximity to one another than any minima between SD 1 and SD 2. There is therefore no need to make an attempted connection between SD 1 and SD 2 directly – provided the connection attempt between SD 2 and SD 3 is successful, SD 1 and SD 2 are also automatically connected to one another. CONNECTUNC LOWESTTEST therefore considers the fewest number of connection attempts required to connect all of the sub-databases within a database, and the closest minima within these respective sub-databases to select for these connection attempts.

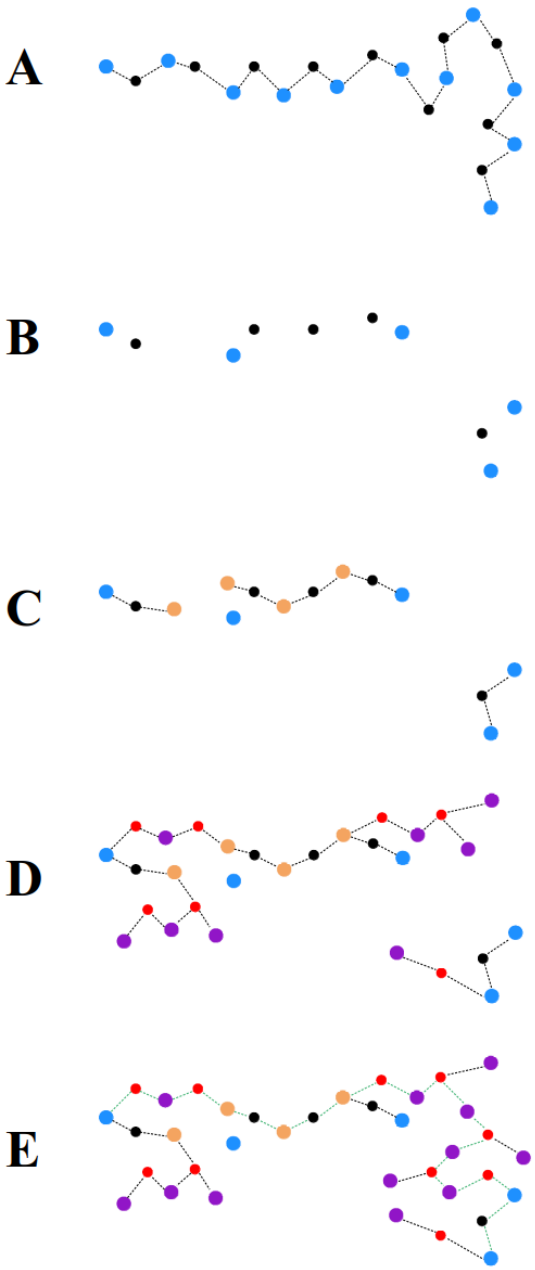


Figure D.2: Cartoon representing the overall CHECKSPMUTATE strategy. (**A**) Fully connected pathway for an original, template system. Blue circles represent minima and smaller black cirlces represent TSs. Black dashed lines indicate direct connections between the minima and TSs. (**B**) Following mutations and reoptimisations to the stationary points from the template system, a new set of stationary points describing the mutated system emerges. These structures should closely resemble the original stationary points they were based on from the template, and so their relative positions are generally retained (in the figure, they are kept in the exact same positions for conceptual convenience, plus their colour-coding is kept consistent). However, it is unlikely that all stationary points will have converged successfully upon reoptimisation, and so gaps arise in the new pathway. Furthermore, information on connectivities between the minima and TSs is lost. (**C**) Steepest-descent pathways from the successfully reoptimised TSs are calculated to

Figure D.2: (continued) determine the minima each TS directly connects. Where these minima coincide with successfully reoptimised minima already present in the database, this newly-found connection is indicated by a dashed line from the TS to that (blue) minimum. Where the steepest-descent pathway culminates in a minimum not already found in the database, this new minimum is therefore added, indicated by an orange circle, and its connection to the TS also represented by a dashed line. (**D**) Attempts are made to connect any gaps still present in the pathway. Minima to connect are selected according to the relative positions their equivalents had in the original pathway. New minima identified are represented by purple circles and new TSs by smaller red circles. (**E**) The process in **D** typically bridges most of the gaps in the pathway. However, there may be some longer gaps that prove difficult to connect. **CONNECTUNC LOWESTTEST** is therefore used to try to connect the remaining parts of the pathway (which can be considered as sub-databases) based on an efficient selection of which minima from each part to consider. The fully connected pathway that is equivalent to the one seen in the original system is indicated by a series of green dashed lines.

Appendix E

Full Phylogenetic Tree

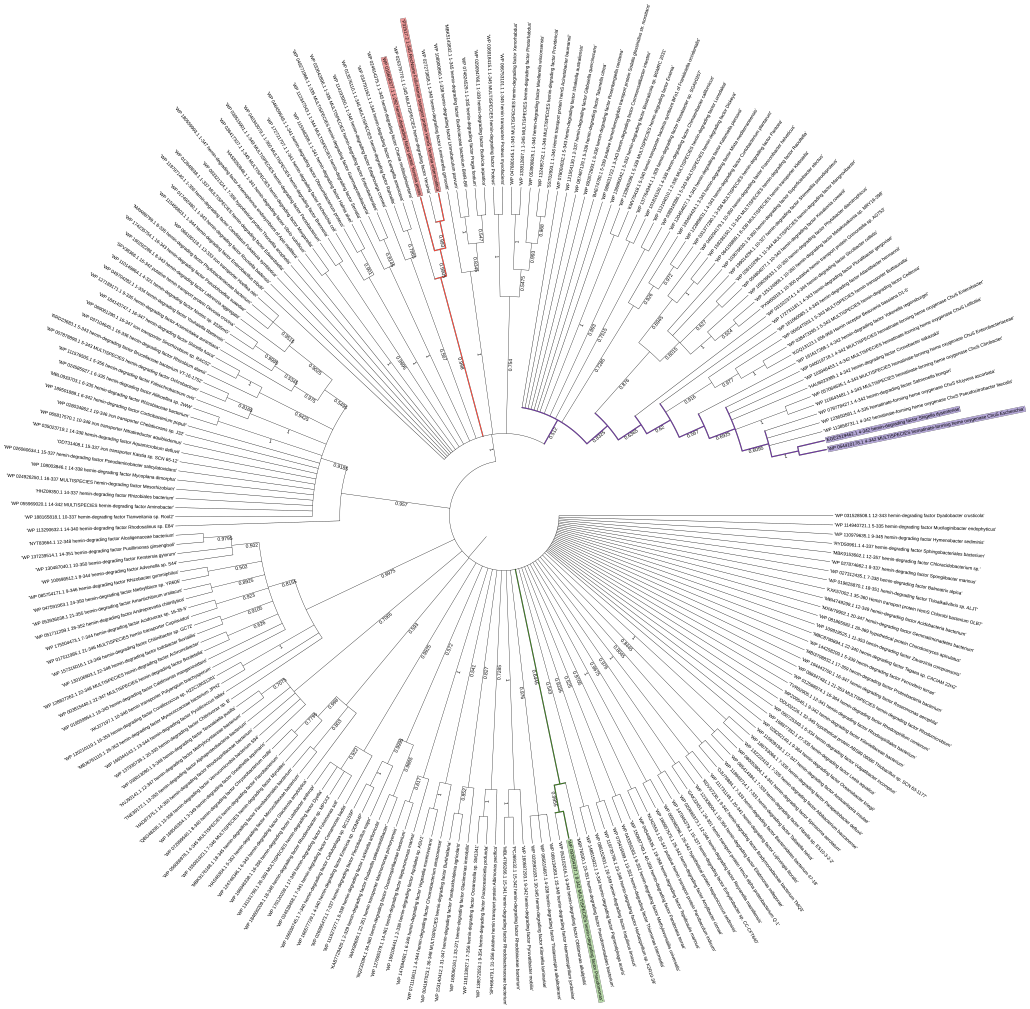


Figure E.1: Maximum likelihood phylogenetic tree for HemS and its homologues. Accession numbers and bootstrap values are given. Figure reproduced from Xie,[81] with small adaptations.

Appendix F

Sequence Conservation

The following sequence is of WT HemS according to the experimental sequence used throughout this thesis (see Appendix B). Residues are highlighted in bold if they showed >90% conservation with respect to the other 218 homologues in the bioinformatic study. A background cyan colour indicates that the residue forms part of the large cavity, and a salmon colour indicates it forms part of the small cavity. The assignment of residues to these cavities was achieved using the default values in MetaPocket.[244]

```
  1 MSKSIYEQYL QAKADNPGKY ARDLATLMGI SEAELTHSRV SHDAKRLKGD
 51 ARALLAALEA VGEVKAITRN TYAVHEQMGR YENQHLNGHA GLILNPRNLD
101 LRLFLNQWAS AFTLTEETRH GVRHSIQFFD HQGDALHKVY VTEQTDMPAW
151 EALLAQFITT ENPELQLEPL SAPEVTEPTA TDEAVDAEWR AMTDVHQFFQ
201 LLKRNNLTRQ QAFRAVGNDL AYQVDNSSLT QLLNIAQQEQ NEIMIFVGNR
251 GCVQIFTGMI EKVTPHQDWI NVFNQRFTLH LIETTIAESW ITRKPTKDGF
301 VTSLELFAAD GTQIAQLYGQ RTEGQPEQTQ WREQIARLNN KDIAA*
```

Appendix G

Stopped-Flow Curve Fitting

Please see the following page.

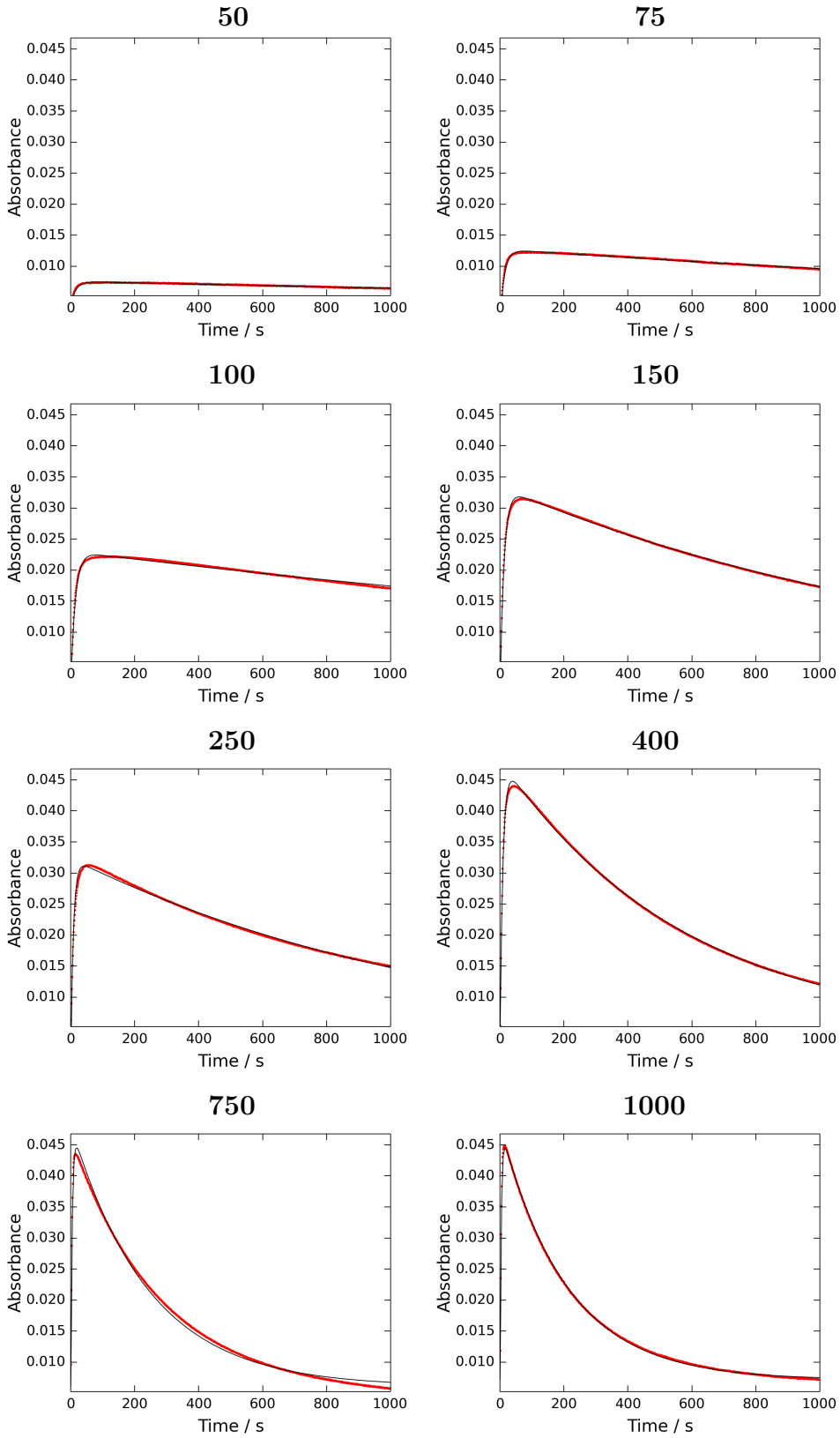


Figure G.1: Overlay of curves fitted to Wild Type HemS stopped-flow data. Graph titles correspond to the NADH concentration.

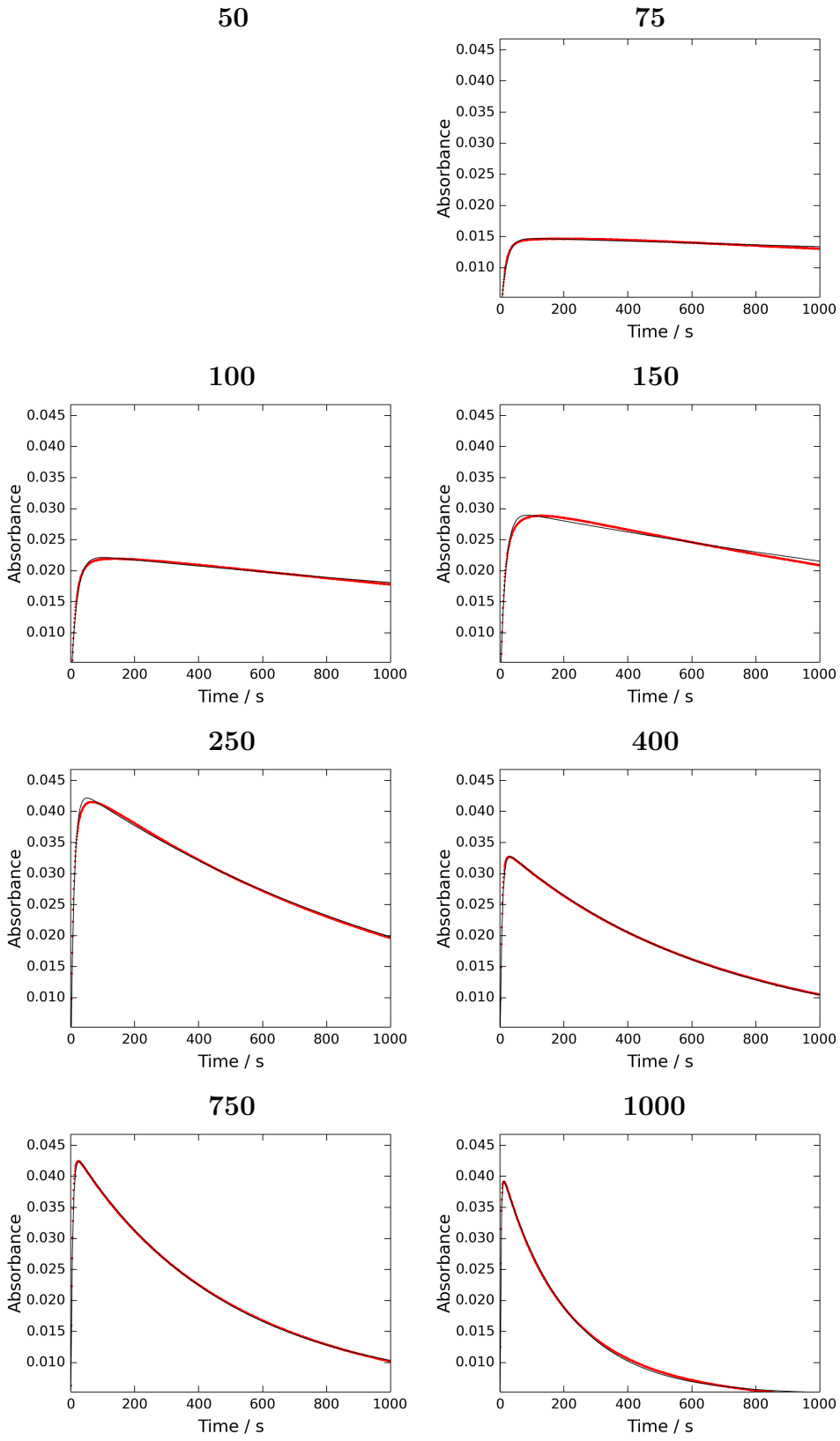


Figure G.2: Overlay of curves fitted to F104AF199A HemS stopped-flow data. Graph titles correspond to the NADH concentration. The 50 µM NADH case could not be accurately fitted to the data, and so has been left blank.

References

[1] Venter, C.; Cohen, D. The 21st century: the century of Biology. *New Perspect. Q.* **1997**, *14*, 26–31.

[2] Herman, J. G.; Graff, J. R.; Myohanen, S.; Nelkin, B. D.; Baylin, S. B. Methylation-specific PCR: A novel PCR assay for methylation status of CpG islands. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 9821–9826.

[3] Barski, A.; Cuddapah, S.; Cui, K.; Roh, T. Y.; Schones, D. E.; Wang, Z.; Wei, G.; Chepelev, I.; Zhao, K. High-resolution profiling of histone methylations in the human genome. *Cell* **2007**, *129*, 823–837.

[4] Booth, M. J.; Branco, M. R.; Ficz, G.; Oxley, D.; Krueger, F.; Reik, W.; Balasubramanian, S. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **2012**, *336*, 934–938.

[5] Arnold, F. H. Design by Directed Evolution. *Acc. Chem. Res.* **1998**, *31*, 125–131.

[6] Romero, P. A.; Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866–876.

[7] Woolf, T. M. Therapeutic repair of mutated nucleic acid sequences. *Nat. Biotechnol.* **1998**, *16*, 341–344.

[8] Cermak, T.; Doyle, E. L.; Christian, M.; Wang, L.; Zhang, Y.; Schmidt, C.; Baller, J. A.; Somia, N. V.; Bogdanove, A. J.; Voytas, D. F. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **2011**, *39*, 1–11.

[9] Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A.; Charpentier, E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **2012**, *337*, 816–822.

[10] Dubochet, J.; McDowall, A. W. Vitrification of pure water for electron microscopy. *J. Microsc.* **1981**, *124*, 3–4.

[11] Frank, J.; Radermacher, M.; Penczek, P.; Zhu, J.; Li, Y.; Ladjadj, M.; Leith, A. SPIDER and WEB: Processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **1996**, *116*, 190–199.

[12] Henderson, R.; Baldwin, J.; Ceska, T.; Zemlin, F.; Beckmann, E.; Downing, K. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **1990**, *213*, 899–929.

[13] Wu, X. *et al.* Rational Design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* **2010**, *329*, 856–861.

[14] Gatenby, R. A.; Smallbone, K.; Maini, P. K.; Rose, F.; Averill, J.; Nagle, R. B.; Worrall, L.; Gillies, R. J. Cellular adaptations to hypoxia and acidosis during somatic evolution of breast cancer. *Br. J. Cancer* **2007**, *97*, 646–653.

[15] Smallbone, K.; Gatenby, R. A.; Gillies, R. J.; Maini, P. K.; Gavaghan, D. J. Metabolic changes during carcinogenesis: potential impact on invasiveness. *J. Theor. Biol.* **2007**, *244*, 703–713.

[16] Metzcar, J.; Wang, Y.; Heiland, R.; Macklin, P. A review of cell-based computational modeling in Cancer Biology. *JCO Clin. Cancer Inform.* **2019**, *3*, 1–13.

[17] Ramaprasad, A.; Pain, A.; Ravasi, T. Defining the protein interaction network of human malaria parasite *Plasmodium falciparum*. *Genomics* **2012**, *99*, 69–75.

[18] De Laeter, J. R.; Böhlke, J. K.; De Bièvre, P.; Hidaka, H.; Peiser, H. S.; Rosman, K. J.; Taylor, P. D. Atomic weights of the elements: Review 2000 (IUPAC Technical Report). *Pure Appl. Chem.* **2003**, *75*, 683–800.

[19] Haynes, W. M.; Lide, D. R.; Bruno, T. J. *Abundance of Elements in the Earth's Crust and in the Sea* in *CRC Handbook of Chemistry and Physics*, 94th ed., CRC Press: Boca Raton, FL, **2016**, 14–17.

[20] Andrews, S. C.; Robinson, A. K.; Rodríguez-Quiñones, F. Bacterial iron homeostasis. *FEMS Microbiol. Rev.* **2003**, *27*, 215–237.

[21] Choy, D. C. Y. Haemoproteins and the study of protein-ligand interactions. PhD Thesis, University of Cambridge, **2015**.

[22] Fenton, H. J. H. Oxidation of tartaric acid in presence of iron. *J. Chem. Soc. Trans.* **1894**, *65*, 899–910.

[23] Theil, E. C. Ferritin: structure, gene regulation, and cellular function in animals, plants, and microorganisms. *Ann. Rev. Biochem.* **1987**, *56*, 289–315.

[24] Pistorius, E. K.; Axelrod, B. Iron, an essential component of lipoxygenase. *J. Biol. Chem.* **1974**, *249*, 3183–3186.

[25] Hayashi, T.; Matsuo, T.; Hitomi, Y.; Okawa, K.; Suzuki, A.; Shiro, Y.; Iizuka, T.; Hisaeda, Y.; Ogoshi, H. Contribution of heme-propionate side chains to structure and function of myoglobin: chemical approach by artificially created prosthetic groups. *J. Inorg. Biochem.* **2002**, *91*, 94–100.

[26] Barker, P. D.; Ferguson, S. J. Still a puzzle: why is haem covalently attached in c-type cytochromes? *Structure* **1999**, *7*, 281–290.

[27] Kumar, S.; Bandyopadhyay, U. Free heme toxicity and its detoxification systems in human. *Toxicol. Lett.* **2005**, *157*, 175–188.

[28] Chiabrando, D.; Vinchi, F.; Fiorito, V.; Mercurio, S.; Tolosano, E. Heme in pathophysiology: a matter of scavenging, metabolism and trafficking across cell membranes. *Front. Pharmacol.* **2014**, *5*, 1–24.

[29] Neidhardt, F. C.; Umbarger, H. E. *Chemical composition of Escherichia coli* in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, Am. Soc. Microbiol. (ASM) Press: Washington DC, **1996**, Chapter 3.

[30] Alberts, B.; Johnson, A.; Lewis, J. *The shape and structure of proteins* in *Molecular Biology of the Cell*, 4th ed., Garland Science: New York, **2002**, Chapter 3.

[31] Lodish, H.; Berk, A.; Zipursky, S. L. *Molecular Cell Biology*, 4th ed., W. H. Freeman: New York, **2000**.

[32] Hutchins, D. A.; Rueter, J. G.; Fish, W. Siderophore production and nitrogen fixation are mutually exclusive strategies in *Anabaena* 7120. *Limnol. and Oceanogr.* **1991**, *36*, 1–12.

[33] Wilks, A.; Burkhard, K. A. Heme and virulence: how bacterial pathogens regulate, transport and utilize heme. *Nat. Prod. Rep.* **2007**, *24*, 511–522.

[34] Runyen-Janecky, L. J. Role and regulation of heme iron acquisition in gram-negative pathogens. *Front. Cell. Infect. Microbiol.* **2013**, *3*, 1–11.

[35] Ahn, S. H.; Han, J. H.; Lee, J. H.; Park, K. J.; Kong, I. S. Identification of an iron-regulated hemin-binding outer membrane protein, HupO, in *Vibrio fluvialis*: effects on hemolytic activity and the oxidative stress response. *Infect. Immun.* **2005**, *73*, 722–729.

[36] Cobessi, D.; Meksem, A.; Brillet, K. Structure of the heme/hemoglobin outer membrane receptor ShuA from *Shigella dysenteriae*: heme binding by an induced fit mechanism. *Proteins* **2010**, *78*, 286–294.

[37] Stojiljkovic, I.; Hantke, K. Transport of haemin across the cytoplasmic membrane through a haemin-specific periplasmic binding-protein dependent transport system in *Yersinia enterocolitica*. *Mol. Microbiol.* **1994**, *13*, 719–732.

[38] Allen, W. J.; Phan, G.; Waksman, G. *Structural biology of periplasmic chaperones* in *Advances in Protein Chemistry and Structural Biology*, 1st ed., Elsevier, **2009**, 51–97.

[39] Mattle, D.; Zeltina, A.; Woo, J. S.; Goetz, B. A.; Locher, K. P. Two stacked heme molecules in the binding pocket of the periplasmic heme-binding protein HmuT from *Yersinia pestis*. *J. Mol. Biol.* **2010**, *404*, 220–231.

[40] Burkhard, K. A.; Wilks, A. Functional characterization of the *Shigella dysenteriae* heme ABC transporter. *Biochemistry* **2008**, *47*, 7977–7979.

[41] Wegiel, B.; Otterbein, L. E. Go green: the anti-inflammatory effects of biliverdin reductase. *Front. Pharmacol.* **2012**, *3*, 1–8.

[42] Sawyer, E. B. Biophysical analysis of haem-protein interactions in bacterial haem transfer systems. PhD Thesis, University of Cambridge, **2009**.

[43] Maines, M. D.; Kappas, A. Cobalt induction of hepatic heme oxygenase; with evidence that cytochrome P-450 is not essential for this enzyme activity. *Proc. Natl. Acad. Sci. USA* **1974**, *71*, 4293–4297.

[44] Unno, M.; Matsui, T.; Ikeda-Saito, M. Structure and catalytic mechanism of heme oxygenase. *Nat. Prod. Rep.* **2007**, *24*, 553–570.

[45] Schuller, D. J.; Wilks, A.; Ortiz De Montellano, P. R.; Poulos, T. L. Crystal structure of human heme oxygenase-1. *Nat. Struct. Biol.* **1999**, *6*, 860–867.

[46] Wilks, A. Heme oxygenase: evolution, structure and mechanism. *Antioxid. Redox Signal.* **2002**, *4*, 603–614.

[47] Maharshak, N.; Ryu, H. S.; Fan, T.-J.; Onyiah, J. C.; Otterbein, S. L.; Wong, R.; Hansen, J.; Otterbein, L. E.; Plevy, S. E. *Escherichia coli* heme oxygenase modulates host innate immune reponses. *Microbiol. Immunol.* **2015**, *59*, 452–465.

[48] Lehmann, E.; El-Tantawy, W. H.; Ocker, M.; Bartenschlager, R.; Lohmann, V.; Hashemolhosseini, S.; Tiegs, G.; Sass, G. The heme oxygenase 1 product biliverdin interferes with hepatitis C virus replication by increasing antiviral interferon response. *Hepatology* **2010**, *51*, 398–404.

[49] Zhu, Z.; Wilson, A. T.; Luxon, B. A.; Brown, K. E.; Mathahs, M. M.; Bandyopadhyay, S.; McCaffrey, A. P.; Schmidt, W. N. Biliverdin inhibits hepatitis C virus nonstructural 3/4A protease activity: mechanism for the antiviral effects of heme oxygenase? *Hepatology* **2010**, *52*, 1897–1905.

[50] Stocker, R.; Glazer, A. N.; Ames, B. N. Antioxidant activity of albumin-bound bilirubin. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 5918–5922.

[51] Stocker, R.; Yamamoto, Y.; McDonagh, A. F.; Glazer, A. N.; Ames, B. N. Bilirubin is an antioxidant of possible physiological importance. *Science* **1987**, *235*, 1043–1046.

[52] Ohrui, T.; Yasuda, H.; Yamaya, M.; Matsui, T.; Sasaki, H. Transient relief of asthma symptoms during jaundice: a possible beneficial role of bilirubin. *J. Exp. Med.* **2003**, *199*, 193–196.

[53] Barañano, D. E.; Rao, M.; Ferris, C. D.; Snyder, S. H. Biliverdin reductase: a major physiologic cytoprotectant. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 16093–16098.

[54] Rivera, M.; Rodríguez, J. C. The dual role of heme as cofactor and substrate in the biosynthesis of carbon monoxide. *Met. Ions Life Sci.* **2009**, *6*, 241–293.

[55] Saebø, A.; Lassen, J. Acute and chronic gastrointestinal manifestations associated with *Yersinia enterocolitica* infection: a Norwegian 10-year follow-up study on 458 hospitalized patients. *Ann. Surg.* **1992**, *215*, 250–255.

[56] Karachalios, G.; Bablekos, G.; Karachaliou, G.; Charalabopoulos, A. K.; Charalabopoulos, G. Infectious endocarditis due to *Yersinia enterocolitica*. *Chemotherapy* **2002**, *48*, 158–159.

[57] Zińczuk, J.; Wojskowicz, P.; Kiśluk, J.; Fil, D.; Kemona, A.; Dadan, J. Mesenteric lymphadenitis caused by *Yersinia enterocolitica*. *Prz. Gastroenterol.* **2015**, *10*, 118–121.

[58] Reinicke, V.; Korner, B. Fulminant septicemia caused by *Yersinia enterocolitica*. *Scand. J. Infect. Dis.* **1977**, *9*, 249–251.

[59] *Centers for Disease Control and Prevention (CDC) Yellow Book 2020: Health Information for International Travel*, Oxford University Press: New York, **2020**.

[60] Stojiljkovic, I.; Hantke, K. Hemin uptake system of *Yersinia enterocolitica*: similarities with other TonB-dependent systems in Gram-negative bacteria. *EMBO J.* **1992**, *11*, 4359–4367.

[61] Amarelle, V.; Koziol, U.; Rosconi, F.; Noya, F.; O'Brian, M. R.; Fabiano, E. A new small regulatory protein, HmuP, modulates haemin acquisition in *Sinorhizobium meliloti*. *Microbiol.* **2010**, *156*, 1873–1882.

[62] Escamilla-Hernandez, R.; O'Brian, M. R. HmuP is a coactivator of Irr-dependent expression of heme utilization genes in *Bradyrhizobium japonicum*. *J. Bacteriol.* **2012**, *194*, 3137–3143.

[63] Sato, T.; Nonoyama, S.; Kimura, A.; Nagata, Y.; Ohtsubo, Y.; Tsuda, M. The small protein HemP is a transcriptional activator for the hemin uptake operon in *Burkholderia multivorans* ATCC 17616. *Appl. Environ. Microbiol.* **2017**, *83*, 1–14.

[64] Troxell, B.; Hassan, H. M. Transcriptional regulation by Ferric Uptake Regulator (Fur) in pathogenic bacteria. *Front. Cell. Infect. Microbiol.* **2013**, *3*, 1–13.

[65] Jacobi, C. A.; Gregor, S.; Rakin, A.; Heesemann, J. Expression analysis of the yersiniabactin receptor gene *fyuA* and the heme receptor *hemR* of *Yersinia enterocolitica* in vitro and in vivo using the reporter genes for green fluorescent protein and luciferase. *Infect. Immun.* **2001**, *69*, 7772–7782.

[66] LaCross, N. C.; Marrs, C. F.; Gilsdorf, J. R. Otitis media associated polymorphisms in the hemin receptor HemR of nontypeable *Haemophilus influenzae*. *Infect. Genet. Evol.* **2014**, *26*, 47–57.

[67] Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010**, *5*, 725–738.

[68] Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **2008**, *9*, 1–8.

[69] Zhang, Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **2007**, *69*, 108–117.

[70] Braun, V.; Hantke, K. *Genetics of bacterial iron transport* in *Handbook of Microbial Iron Chelates*, CRC Press: Boca Raton, FL, **1991**, 107–138.

[71] Shine, J.; Dalgarno, L. The 3′ terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. USA* **1974**, *71*, 1342–1346.

[72] Woo, J. S.; Zeltina, A.; Goetz, B. A.; Locher, K. P. X-ray structure of the *Yersinia pestis* heme transporter HmuUV. *Nat. Struct. Mol. Biol.* **2012**, *19*, 1310–1315.

[73] Walker, J. E.; Saraste, M.; Runswick, M. J.; Gay, N. J. Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1982**, *1*, 945–951.

[74] Hanson, P. I.; Whiteheart, S. W. AAA+ proteins: have engine, will work. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 519–529.

[75] Davis, K. M. All *Yersinia* are not created equal: phenotypic adaptation to distinct niches within mammalian tissues. *Front. Cell. Infect. Microbiol.* **2018**, *8*, 1–8.

[76] Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, 480–489.

[77] Hornung, J. M.; Jones, H. A.; Perry, R. D. The hmu locus of *Yersinia pestis* is essential for utilization of free haemin and haem-protein complexes as iron sources. *Mol. Microbiol.* **1996**, *20*, 725–739.

[78] Thompson, J. M.; Jones, H. A.; Perry, R. D. Molecular characterization of the hemin uptake locus (*hmu*) from *Yersinia pestis* and analysis of *hmu* mutants for hemin and hemoprotein utilization. *Infect. Immun.* **1999**, *67*, 3879–3892.

[79] Torres, A. G.; Payne, S. M. Haem iron-transport system in enterohaemorrhagic *Escherichia coli* O157:H7. *Mol. Microbiol.* **1997**, *23*, 825–833.

[80] Suits, M. D.; Pal, G. P.; Nakatsu, K.; Matte, A.; Cygler, M.; Jia, Z. Identification of an *Escherichia coli* O157:H7 heme oxygenase with tandem functional repeats. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16955–16960.

[81] Xie, Y. Predicting function from structure: haem degradation and DNA binding in the HemS family. Master's Thesis, University of Cambridge, **2021**.

[82] Suits, M. D.; Lang, J.; Pal, G. P.; Couture, M.; Jia, Z. Structure and heme binding properties of *Escherichia coli* O157:H7 ChuX. *Protein Sci.* **2009**, *18*, 825–838.

[83] LaMattina, J. W.; Nix, D. B.; Lanzilotta, W. N. Radical new paradigm for heme degradation in *Escherichia coli* O157:H7. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 12138–12143.

[84] Zhang, Q.; van der Donk, W. A.; Liu, W. Radical-mediated enzymatic methylation: a tale of two SAMS. *Acc. Chem. Res.* **2012**, *45*, 555–564.

[85] Huang, W.; Xu, H.; Li, Y.; Zhang, F.; Chen, X. Y.; He, Q.-L.; Igarashi, Y.; Tang, G.-L. Characterization of yatakemycin gene cluster revealing a radical S-adenosylmethionine dependent methyltransferase and highlighting spirocyclopropane biosynthesis. *J. Am. Chem. Soc.* **2012**, *134*, 8831–8840.

[86] Lan, R.; Reeves, P. R. *Escherichia* coli in disguise: molecular origins of Shigella. *Microb. Infect.* **2002**, *4*, 1125–1132.

[87] Jin, Q. *et al.* Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.* **2002**, *30*, 4432–4441.

[88] World Health Organisation (WHO): State of the art of vaccine research and development. **2005**, 1–99.

[89] Wyckoff, E. E.; Duncan, D.; Torres, A. G.; Mills, M.; Maase, K.; Payne, S. M. Structure of the *Shigella dysenteriae* haem transport locus and its phylogenetic distribution in enteric bacteria. *Mol. Microbiol.* **1998**, *28*, 1139–1152.

[90] Griffin, A. S.; West, S. A.; Buckling, A. Cooperation and competition in pathogen bacteria. *Nature* **2004**, *430*, 1024–1027.

[91] Harrison, F.; Browning, L. E.; Vos, M.; Buckling, A. Cooperation and virulence in acute *Pseudomonas aeruginosa* infections. *BMC Biol.* **2006**, *4*, 1–5.

[92] Ochsner, U. A.; Johnson, Z.; Vasil, M. L. Genetics and regulation of two distinct haem-uptake systems, *phu* and *has*, in *Pseudomonas aeruginosa*. *Microbiology* **2000**, *146*, 185–198.

[93] Chan, A. C.; Lelj-Garolla, B.; Rosell, F. I.; Pedersen, K. A.; Mauk, A. G.; Murphy, M. E. Cofacial Heme Binding is Linked to Dimerization by a Bacterial Heme Transport Protein. *J. Mol. Biol.* **2006**, *362*, 1108–1119.

[94] van Vliet, A. H.; Ketley, J. M.; Park, S. F.; Penn, C. W. The role of iron in *Campylobacter* gene regulation, metabolism and oxidative stress defense. *FEMS Microbiol. Rev.* **2002**, *26*, 173–186.

[95] Schneider, S.; Sharp, K. H.; Barker, P. D.; Paoli, M. An induced fit conformational change underlies the binding mechanism of the heme transport proteobacteria-protein HemS. *J. Biol. Chem.* **2006**, *281*, 32606–32610.

[96] Schneider, S.; Paoli, M. Haem-binding properties and crystallisation of the bacterial protein HemS. *Acta Cryst. A* **2005**, *61*, 343.

[97] Song, Y.; Mao, J.; Gunner, M. R. Electrostatic environment of hemes in proteins: pK_as of hydroxyl ligands. *Biochemistry* **2006**, *45*, 7949–7958.

[98] Suits, M. D.; Jaffer, N.; Jia, Z. Structure of the *Escherichia coli* O157:H7 heme oxygenase ChuS in complex with heme and enzymatic inactivation by mutation of the heme coordinating residue His-193. *J. Biol. Chem.* **2006**, *281*, 36776–36782.

[99] Ouellet, Y. H.; Ndiaye, C. T.; Gagné, S. M.; Sebilo, A.; Suits, M. D.; Jubinville, É.; Jia, Z.; Ivancich, A.; Couture, M. An alternative reaction for heme degradation catalyzed by the *Escherichia coli* O157:H7 ChuS protein: Release of hematinic acid, tripyrrole and Fe(III). *J. Inorg. Biochem.* **2016**, *154*, 103–113.

[100] Sakamoto, H.; Omata, Y.; Adachi, Y.; Palmer, G.; Noguchi, M. Separation and identification of the regioisomers of verdoheme by reversed-phase ion-pair high-performance liquid chromatography, and characterization of their complexes with heme oxygenase. *J. Inorg. Biochem.* **2000**, *82*, 113–121.

[101] Mathew, L. G.; Beattie, N. R.; Pritchett, C.; Lanzilotta, W. N. New insight into the mechanism of anaerobic heme degradation. *Biochemistry* **2019**, *58*, 4641–4654.

[102] Wilks, A. The ShuS protein of *Shigella dysenteriae* is a heme-sequestering protein that also binds DNA. *Arch. Biochem. Biophys.* **2001**, *387*, 137–142.

[103] Kaur, A. P.; Wilks, A. Heme inhibits the DNA binding properties of the cytoplasmic heme binding protein of *Shigella dysenteriae* (ShuS). *Biochemistry* **2007**, *46*, 2994–3000.

[104] Lansky, I. B.; Lukat-Rodgers, G. S.; Block, D.; Rodgers, K. R.; Ratliff, M.; Wilks, A. The cytoplasmic heme-binding protein (PhuS) from the heme uptake system of *Pseudomonas aeruginosa* is an intracellular heme-trafficking protein to the δ-regioselective heme oxygenase. *J. Biol. Chem.* **2006**, *281*, 13652–13662.

[105] Warburg, O.; Negelein, E. Grunes haemin aus blast-haemin. *Chem. Ber.* **1930**, *63*, 1816–1819.

[106] Lemberg, R. Transformation of haemins into bile pigments. *Biochem. J.* **1935**, *29*, 1322–1336.

[107] Wilks, A.; Ikeda-Saito, M. Heme utilization by pathogenic bacteria: not all pathways lead to biliverdin. *Acc. Chem. Res.* **2014**, *47*, 2291–2298.

[108] Avila, L.; Huang, H.-w.; Damaso, C. O.; Lu, S.; Moënne-Loccoz, P.; Rivera, M. Coupled oxidation vs heme oxygenation: insights from axial ligand mutants of mitochondrial cytochrome b_5. *J. Am. Chem. Soc.* **2003**, *125*, 4103–4110.

[109] O'Neill, M. J.; Bhakta, M. N.; Fleming, K. G.; Wilks, A. Induced fit on heme binding to the *Pseudomonas aeruginosa* cytoplasmic protein (PhuS) drives interaction with heme oxygenase (HemO). *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 5639–5644.

[110] Tripathi, S.; O'Neill, M. J.; Wilks, A.; Poulos, T. L. Crystal Structure of the *Pseudomonas aeruginosa* cytoplasmic heme binding protein, *apo*-PhuS. *J. Inorg. Biochem.* **2013**, *128*, 131–136.

[111] Lee, M. J.; Schep, D.; McLaughlin, B.; Kaufmann, M.; Jia, Z. Structural analysis and identification of PhuS as a heme-degrading enzyme from *Pseudomonas aeruginosa*. *J. Mol. Biol.* **2014**, *426*, 1936–1946.

[112] Wilson, T.; Mouriño, S.; Wilks, A. The heme binding protein PhuS transcriptionally regulates the *Pseudomonas aeruginosa* tandem sRNA *prrF1,F2* locus. *J. Biol. Chem.* **2021**, *296*, 100275–100285.

[113] Wilderman, P. J.; Sowa, N. A.; FitzGerald, D. J.; FitzGerald, P. C.; Gottesman, S.; Ochsner, U. A.; Vasil, M. L. Identification of tandem duplicate regulatory small RNAs in *Pseudomonas aeruginosa* involved in iron homeostasis. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 9792–9797.

[114] Oglesby, A. G.; Farrow, J. M.; Lee, J. H.; Tomaras, A. P.; Greenberg, E. P.; Pesci, E. C.; Vasil, M. L. The influence of iron on *Pseudomonas aeruginosa* physiology: a regulatory link between iron and quorum sensing. *J. Biol. Chem.* **2008**, *283*, 15558–15567.

[115] Reinhart, A. A.; Powell, D. A.; Nguyen, A. T.; O'Neill, M.; Djapgne, L.; Wilks, A.; Ernst, R. K.; Oglesby-Sherrouse, A. G. The *prrF*-encoded small regulatory RNAs are required for iron homeostasis and virulence of *Pseudomonas aeruginosa*. *Infect. Immun.* **2015**, *83*, 863–875.

[116] Weber, G. Intramolecular transfer of electronic energy in dihydro diphosphopyridine nucleotide. *Nature* **1957**, *180*, 1409.

[117] Patel, D. J. 220 MHz proton nuclear magnetic resonance spectra of retinals. *Nature* **1969**, *221*, 825–828.

[118] Unden, G.; Bongaerts, J. Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors. *BBA Bioenerg.* **1997**, *1320*, 217–234.

[119] Gregory, J. Investigating the NADH dependent reaction of the *Yersinia enterocolitica* haem chaperone HemS. Master's Thesis, University of Cambridge, **2012**.

[120] Ikram, A.; Su, Q.; Fiaz, M.; Khadim, S. Big data in enterprise management: transformation of traditional recruitment strategy. *IEEE ICBDA* **2017**, 414–419.

[121] Lee, C.-Y.; Chien, C.-F. Pitfalls and protocols of data science in manufacturing practice. *J. Intell. Manuf.* **2020**, 1–19.

[122] Dash, S.; Shakyawar, S. K.; Sharma, M.; Kaushik, S. Big data in healthcare: management, analysis and future prospects. *J. Big Data* **2019**, *6*, 1–25.

[123] Esfahani, H. J.; Tavasoli, K.; Jabbarzadeh, A. Big data and social media: a scientometrics analysis. *Int. J. Data Netw. Sci.* **2019**, *3*, 145–164.

[124] Li, K.; Du, Y.; Li, L.; Wei, D.-Q. Bioinformatics Approaches for anti-cancer drug discovery. *Curr. Drug Targets* **2020**, *21*, 3–17.

[125] Collins, F. S. *et al.* Finishing the euchromatic sequence of the human genome. *Nature* **2004**, *431*, 931–945.

[126] Feng, J.-J.; Chen, J.-N.; Kang, W.; Wu, Y.-D. Accurate structure prediction for protein loops based on molecular dynamics simulations with RSFF2C. *J. Chem. Theory Comput.* **2021**, *17*, 4614–4628.

[127] Cho, S. S.; Weinkam, P.; Wolynes, P. G. Origins of barriers and barrierless folding in BBL. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 118–123.

[128] Salsbury, F. R.; Crowley, M. F.; Brooks III, C. L. Modeling of the metallo-β-lactamase from *B. Fragilis*: structural and dynamic effects of inhibitor binding. *Proteins* **2001**, *44*, 448–459.

[129] Verlet, L. Computer "Experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **1967**, *159*, 98–103.

[130] Elber, R. Perspective: Computer simulations of long time dynamics. *J. Chem. Phys.* **2016**, *144*, 1–12.

[131] Eaton, W. A.; Munõz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. Fast kinetic and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 327–359.

[132] Prigozhin, M. B.; Gruebele, M. Microsecond folding experiments and simulations: a match is made. *Phys. Chem. Chem. Phys.* **2013**, *15*, 3372–3388.

[133] Wales, D. J. *Energy Landscapes*, Cambridge University Press: Cambridge, **2003**.

[134] Röder, K. Energy landscaping - on the relationship between functionality and sequence mutations for multifunctional biomolecules. PhD Thesis, University of Cambridge, **2018**.

[135] Dijkstra, E. A note on two problems in connexion with graphs. *Numer. Math.* **1959**, *1*, 269–271.

[136] Baldwin, R. L. Matching speed and stability. *Nature* **1994**, *369*, 183–184.

[137] Wales, D. J. Calculating rate constants and committor probabilities for transition networks by graph transformation. *J. Chem. Phys.* **2009**, *130*, 204111–204118.

[138] Stevenson, J. D.; Wales, D. J. Communication: Analysing kinetic transition networks for rare events. *J. Chem. Phys.* **2014**, *141*, 041104–041108.

[139] Trygubenko, S. A.; Wales, D. J. Graph transformation method for calculating waiting times in Markov chains. *J. Chem. Phys.* **2006**, *124*, 234110–234126.

[140] Sharpe, D. J.; Wales, D. J. Identifying mechanistically distinct pathways in kinetic transition networks. *J. Chem. Phys.* **2019**, *151*, 124101–124114.

[141] Burke, D. F.; Mantell, R. G.; Pitt, C. E.; Wales, D. J. Energy landscape for the membrane fusion pathway in influenza A hemagglutinin from discrete path sampling. *Front. Chem.* **2020**, *8*, 1–11.

[142] Levinthal, C. How to fold graciously. Mossbauer spectroscopy in biological systems: proceedings of a meeting held at Allerton House, Monticello, Illinois. **1969**, 22–24.

[143] Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **1973**, *181*, 223.

[144] Dill, K. A. *The stabilities of globular proteins* in *Protein Engineering*, Alan R. Liss, Inc.: New York, **1987**, 187–192.

[145] Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 8721–8725.

[146] Bryngelson, J. D.; Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 7524–7528.

[147] Carr, J. M.; Wales, D. J. Folding pathways and rates for the three-stranded β-sheet peptide Beta3s using discrete path sampling. *J. Phys. Chem. B* **2008**, *112*, 8760–8769.

[148] Prentiss, M. C.; Wales, D. J.; Wolynes, P. G. The energy landscape, folding pathways and the kinetics of a knotted protein. *PLoS Comput. Biol.* **2010**, *6*, 1–12.

[149] Joseph, J. A.; Chakraborty, D.; Wales, D. J. Energy Landscape for fold-switching in regulatory protein RfaH. *J. Chem. Theory Comput.* **2019**, *15*, 731–742.

[150] Neelamraju, S.; Gosavi, S.; Wales, D. J. Energy landscape of the designed protein Top7. *J. Phys. Chem. B* **2018**, *122*, 12282–12291.

[151] Chakraborty, D.; Wales, D. J. Energy landscape and pathways for transitions between Watson-Crick and Hoogsteen base pairing in DNA. *J. Phys. Chem. Lett.* **2018**, *9*, 229–241.

[152] Fejer, S. N.; James, T. R.; Hernández-Rojas, J.; Wales, D. J. Energy landscapes for shells assembled from pentagonal and hexagonal pyramids. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2098–2104.

[153] Hendlich, M. Databases for protein-ligand complexes. *Acta Cryst. D* **1998**, *54*, 1178–1182.

[154] Schmitt, S.; Hendlich, M.; Klebe, G. From structure to function: a new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angew. Chem. Int. Ed.* **2001**, *40*, 3141–3144.

[155] Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406.

[156] Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **1997**, *15*, 359–363.

[157] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

[158] Wales, D. J. GMIN: A program for basin-hopping global optimisation. **2021**, `http://www-wales.ch.cam.ac.uk/GMIN`.

[159] Shang, C.; Choy, D.; Barker, P. D.; Wales, D. J. Energy landscape for an automated enzymatic double gate: analysis of phenylalanine flipping in HemS [Unpublished]. **2015**, 1–12.

[160] Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A. *Protein identification and analysis tools on the ExPASy server* in *The Proteomics Protocols Handbook*, Humana Press: Totowa, New Jersey, **2005**, 571–607.

[161] Cole, J. Experimental investigations into the mechanism of the HemS-catalysed reaction between NADH and haem from computational studies. Master's Thesis, University of Cambridge, **2020**.

[162] Peränen, J.; Rikkonen, M.; Hyvönen, M.; Kääriäinen, L. T7 vectors with a modified T7 *lac* promoter for expression of proteins in *Escherichia coli*. *Anal. Biochem.* **1996**, *236*, 371–373.

[163] Schneider, S.; Paoli, M. Crystallization and preliminary X-ray diffraction analysis of the haem-binding protein HemS from *Yersinia enterocolitica*. *Acta Cryst. F* **2005**, *61*, 802–805.

[164] SnapGene software (from Insightful Science). **2021**, `https://www.snapgene.com/`.

[165] NEB Protocol for Restriction Endonuclease Reaction. **2012**, `https://international.neb.com/protocols/2012/12/07/optimizing-restriction-endonuclease-reactions`.

[166] Johnson, K. A.; Simpson, Z. B.; Blom, T. Global Kinetic Explorer: a new computer program for dynamic simulation and fitting of kinetic data. *Anal. Biochem.* **2009**, *387*, 20–29.

[167] Merow, C.; Smith, M. J.; Silander, J. A. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* **2013**, *36*, 1058–1069.

[168] Vonrhein, C.; Flensburg, C.; Keller, P.; Sharff, A.; Smart, O.; Paciorek, W.; Womack, T.; Bricogne, G. Data processing and analysis with the *autoPROC* toolbox. *Acta Cryst. D* **2011**, *67*, 293–302.

[169] Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Cryst. D* **2011**, *67*, 235–242.

[170] Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst. D* **1997**, *53*, 240–255.

[171] Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Cryst. D* **2010**, *66*, 486–501.

[172] Northrop, D. B.; Duggleby, R. G. Preparation and storage of isotopically labeled reduced nicotinamide adenine dinucleotide. *Anal. Biochem.* **1987**, *165*, 362–364.

[173] Basran, J.; Harris, R. J.; Sutcliffe, M. J.; Scrutton, N. S. H-tunneling in the multiple H-transfers of the catalytic cycle of morphinone reductase and in the reductive half-reaction of the homologous pentaerythritol tetranitrate reductase. *J. Biol. Chem.* **2003**, *278*, 43973–43982.

[174] Pudney, C. R.; Hay, S.; Sutcliffe, M. J.; Scrutton, N. S. α-Secondary isotope effects as probes of "tunneling-ready" configurations in enzymatic H-tunneling: insight from environmentally coupled tunneling models. *J. Am. Chem. Soc.* **2006**, *128*, 14053–14058.

[175] Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham III, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

[176] Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542–1555.

[177] Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory and Comput.* **2013**, *9*, 3878–3888.

[178] Case, D. A. *et al.* **AMBER12**, University of California, San Francisco. **2012**.

[179] Case, D. A. *et al.* **AMBER16**, University of California, San Francisco. **2016**.

[180] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple AMBER force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712–725.

[181] Wang, J.; Cieplak, P.; Kollman, P. A. How well does a Restrained Electrostatic Potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21*, 1049–1074.

[182] Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *J. Chem. Theory Comput.* **2006**, *2*, 420–433.

[183] Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. Secondary structure bias in Generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.

[184] Rieloff, E.; Skepö, M. Phosphorylation of a disordered peptide – structural effects and force field inconsistencies. *J. Chem. Theory Comput.* **2020**, *16*, 1924–1935.

[185] Gopal, S. M.; Wingbermühle, S.; Schnatwinkel, J.; Juber, S.; Herrmann, C.; Schäfer, L. V. Conformational preferences of an intrinsically disordered protein domain: a case study for modern force fields. *J. Phys. Chem. B* **2021**, *125*, 24–35.

[186] Mustafa, G.; Nandekar, P. P.; Mukherjee, G.; Bruce, N. J.; Wade, R. C. The effect of force-field parameters on cytochrome P450-membrane interactions: structure and dynamics. *Sci. Rep.* **2020**, *10*, 1–11.

[187] Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.

[188] Bryce, R. Bryce Group - AMBER Parameter Database, University of Manchester. **2021**, `http://amber.manchester.ac.uk/`.

[189] Giammona, D. A. An examination of conformational flexibility in porphyrins and bulky-ligand binding in myoglobin. PhD Thesis, University of California, Davis, **1984**.

[190] Pavelites, J. J.; Gao, J.; Bash, P. A. A molecular mechanics force field for NAD^+, NADH and the pyrophosphate groups of nucleotides. *J. Comput. Chem.* **1997**, *18*, 221–239.

[191] Walker, R. C.; De Souza, M. M.; Mercer, I. P.; Gould, I. R.; Klug, D. R. Large and fast relaxations inside a protein: calculation and measurement of reorganization energies in alcohol dehydrogenase. *J. Phys. Chem. B* **2002**, *106*, 11658–11665.

[192] Holmberg, N.; Ryde, U.; Bülow, L. Redesign of the coenzyme specificity in L-Lactate dehydrogenase from *Bacillus stearothermophilus* using site-directed mutagenesis and media engineering. *Protein Eng.* **1999**, *12*, 851–856.

[193] Bashford, D.; Case, D. A. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.

[194] Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

[195] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

[196] Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins* **2004**, *55*, 383–394.

[197] Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born model suitable for macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

[198] Guex, N.; Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **1997**, *18*, 2714–2723.

[199] Li, Z.; Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6611–6615.

[200] Li, Z.; Scheraga, H. A. Structure and free energy of complex thermodynamic systems. *J. Mol. Struct.* **1988**, *179*, 333–352.

[201] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

[202] Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109.

[203] Hoffmann, K. H.; Franz, A.; Salamon, P. The structure of best possible strategies for finding ground states. *Phys. Rev. E.* **2002**, *66*, 046706–046714.

[204] Wales, D. J.; Doye, J. P. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

[205] Broyden, C. G. The convergence of a class of double-rank minimization algorithms. *IMA J. Appl. Maths.* **1970**, *6*, 76–90.

[206] Fletcher, R. A new approach to variable metric algorithms. *Comput. J.* **1970**, *13*, 317–322.

[207] Goldfarb, D. A family of variable-metric methods derived by variational means. *Math. Comput.* **1970**, *24*, 23–26.

[208] Shanno, D. F. Conditioning of quasi-Newton methods for function minimization. *Math. Comput.* **1970**, *24*, 647–656.

[209] Nocedal, J. Updating quasi-Newton matrices with limited storage. *Math. Comput.* **1980**, *35*, 773–782.

[210] Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Prog.* **1989**, *45*, 503–528.

[211] Mantell, R. G.; Pitt, C. E.; Wales, D. J. GPU-Accelerated exploration of biomolecular energy landscapes. *J. Chem. Theory Comput.* **2016**, *12*, 6182–6191.

[212] Murrell, J.; Laidler, K. Symmetries of activated complexes. *Trans. Faraday Soc.* **1968**, *64*, 371–377.

[213] Wales, D. J. **OPTIM**: A program for optimising geometries and calculating pathways. **2021**, `http://www-wales.ch.cam.ac.uk/OPTIM`.

[214] Trygubenko, S. A.; Wales, D. J. A doubly nudged elastic band method for finding transition states. *J. Chem. Phys.* **2004**, *120*, 2082–2094.

[215] Trygubenko, S. A.; Wales, D. J. Erratum: A doubly nudged elastic band method for finding transition states (Journal of Chemical Physics (2004) 120 (2082)). *J. Chem. Phys.* **2004**, *120*, 7820.

[216] Cerjan, C. J.; Miller, W. H. On finding transition states. *J. Chem. Phys.* **1981**, *75*, 2800–2806.

[217] Pancíř, J. Calculation of the least energy path on the energy hypersurface. *Coll. Czech Chem. Commun.* **1975**, *40*, 1112–1118.

[218] Munro, L. J.; Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B* **1999**, *59*, 3969–3980.

[219] Jónsson, H.; Mills, G.; Jacobsen, K. W. *Nudged elastic band method for finding minimum energy paths of transitions* in *Classical and Quantum Dynamics in Condensed Phase Simulations*, World Scientific, **1998**, 385–404.

[220] Henkelman, G.; Jónsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.* **1999**, *111*, 7010–7022.

[221] Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

[222] Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

[223] Hildebrand, F. *Methods of Applied Mathematics*, Dover Publications: New York, **1992**.

[224] Kumeda, Y.; Wales, D. J.; Munro, L. J. Transition states and rearrangment mechanisms from hybrid eigenvector-following and density functional theory. Application to $C_{10}H_{10}$ and defect migration in crystalline silicon. *Chem. Phys. Lett.* **2001**, *341*, 185–194.

[225] Pechukas, P. On simple saddle points of a potential surface, the conservation of nuclear symmetry along paths of steepest descent, and the symmetry of transition states. *J. Chem. Phys.* **1976**, *64*, 1516.

[226] Carr, J. M.; Trygubenko, S. A.; Wales, D. J. Finding pathways between distant local minima. *J. Chem. Phys.* **2005**, *122*, 234903–234909.

[227] Rhee, Y. M. Construction of an accurate potential energy surface by interpolation with Cartesian weighting coordinates. *J. Chem. Phys.* **2000**, *113*, 6021–6024.

[228] Wales, D. J.; Carr, J. M.; Khalili, M.; de Souza, V. K.; Strodel, B.; Whittleston, C. S. *Pathways and rates for structural transformations of peptides and proteins* in *Proteins: Energy, Heat and Signal Flow*, Taylor and Francis/CRC Press: Oxford, **2009**, 315–340.

[229] Whittleston, C. S. Energy landscapes of biological systems. PhD Thesis, University of Cambridge, **2011**.

[230] Strodel, B.; Whittleston, C. S.; Wales, D. J. Thermodynamics and kinetics of aggregation for the GNNQQNY peptide. *J. Am. Chem. Soc.* **2007**, *129*, 16005–16014.

[231] Wales, D. J. PATHSAMPLE: A driver for OPTIM to create stationary point databases using discrete path sampling and perform kinetic analysis. **2021**, http://www-wales.ch.cam.ac.uk/PATHSAMPLE.

[232] Röder, K.; Wales, D. J. Energy landscapes for the aggregation of Aβ_{17-42}. *J. Am. Chem. Soc.* **2018**, *140*, 4018–4027.

[233] Stillinger, F. H.; Weber, T. A. Hidden structure in liquids. *Phys. Rev. A* **1982**, *25*, 978–989.

[234] Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495–1517.

[235] Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758–760.

[236] Liu, Y.; Ortiz De Montellano, P. R. Reaction intermediates and single turnover rate constants for the oxidation of heme by human heme oxygenase-1. *J. Biol. Chem.* **2000**, *275*, 5297–5307.

[237] Teichmann, L. K. Uber die krystallisation der organischen bestandteile des bluts. *Zeitschrift fur Kationelle Medicin* **1853**, *3*, 375–388.

[238] Le Baut, G.; O'Brien, C.; Pavli, P.; Roy, M.; Seksik, P.; Tréton, X.; Nancey, S.; Barnich, N.; Bezault, M.; Auzolle, C.; Cazals-Hatem, D.; Viala, J.; Allez, M.; Hugot, J. P.; Dumay, A. Prevalence of *Yersinia* species in the ileum of Crohn's disease patients and controls. *Front. Cell. Infect. Microbiol.* **2018**, *8*, 1–9.

[239] Friedman, E. S. *et al.* Microbes vs. chemistry in the origin of the anaerobic gut lumen. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4170–4175.

[240] Schrödinger, L. L. C.; DeLano, W. PYMOL. **2020**, http://www.pymol.org/pymol.

[241] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

[242] Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549.

[243] Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, 256–259.

[244] Huang, B. Metapocket: a meta approach to improve protein ligand binding site prediction. *OMICS J. Integr. Biol.* **2009**, *13*, 325–330.

[245] Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; De Beer, T. A.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*, 296–303.

[246] Liu, R.; Hu, J. DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches. *Proteins* **2013**, *81*, 1885–1899.

[247] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera – A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.

[248] Law, J.; Cole, L. Applied PhotoPhysics SX Series Application Note: LED Light Sources for Stopped-Flow Spectroscopy. **2011**.

[249] Alavi, F. S.; Zahedi, M.; Safari, N.; Ryde, U. QM/MM study of the conversion of oxophlorin into verdoheme by heme oxygenase. *J. Phys. Chem. B* **2017**, *121*, 11427–11436.

[250] Gheidi, M.; Safari, N.; Zahedi, M. Density functional theory studies on the conversion of hydroxyheme to iron-verdoheme in the presence of dioxygen. *Dalton Trans.* **2017**, *46*, 2146–2158.

[251] Alavi, F. S.; Gheidi, M.; Zahedi, M.; Safari, N.; Ryde, U. A novel mechanism of heme degradation to biliverdin studied by QM/MM and QM calculations. *Dalton Trans.* **2018**, *47*, 8283–8291.

[252] Alavi, F. S.; Zahedi, M.; Safari, N.; Ryde, U. QM/MM study of the conversion of biliverdin into verdoheme by heme oxygenase. *Theor. Chem. Acc.* **2019**, *138*, 1–8.

[253] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

[254] Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596.

[255] Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876.

[256] SWISS-MODEL. Homo sapiens (human). **2021**, `https://swissmodel.expasy.org/repository/species/9606`.

[257] Pedersen, B. The dnaseq package. **2002**.

[258] Yokoyama, K. *et al.* Complete nucleotide sequence of the prophage VT1-Sakai carrying the Shiga toxin 1 genes of the enterohemorrhagic *Escherichia coli* O157:H7 strain derived from the Sakai outbreak. *Gene* **2000**, *258*, 127–139.

[259] Mills, M.; Payne, S. M. Identification of *shuA*, the gene encoding the heme receptor of *Shigella dysenteriae*, and analysis of invasion and intracellular multiplication of a *shuA* mutant. *Infect. Immun.* **1997**, *65*, 5358–5363.

[260] Strodel, B.; Wales, D. J. Free energy surfaces from an extended harmonic superposition approach and kinetics for alanine dipeptide. *Chem. Phys. Lett.* **2008**, *466*, 105–115.

[261] Hoare, M. R.; McInnes, J. A. Morphology and statistical statics of simple microclusters. *Adv. Phys.* **1983**, *32*, 791–821.

[262] Joseph, J. A.; Röder, K.; Chakraborty, D.; Mantell, R. G.; Wales, D. J. Exploring biomolecular energy landscapes. *Chem. Commun.* **2017**, *53*, 6974–6988.

[263] Wales, D. J. Decoding heat capacity features from the energy landscape. *Phys. Rev. E* **2017**, *95*, 1–6.

[264] Marcelin, R. Expression of speeds of transformations of physico-chemical systems in an affinity function. *Cr. Hebd. Acad. Sci.* **1913**, *157*, 1419–1422.

[265] Eyring, H. The activated complex in chemical reactions. *J. Chem. Phys.* **1935**, *3*, 107–115.

[266] Evans, M. G.; Polanyi, M. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Trans. Faraday Soc.* **1935**, *31*, 875–893.

[267] Miller, W. H. Importance of nonseparability in quantum mechanical transition-state theory. *Acc. Chem. Res.* **1976**, *9*, 306–312.

[268] Miller, W. H. Unified statistical model for "complex" and "direct" reaction mechanisms. *J. Chem. Phys.* **1976**, *65*, 2216–2223.

[269] Rice, O. K.; Ramsperger, H. C. Theories of unimolecular gas reactions at low pressures. I. *J. Am. Chem. Soc.* **1927**, *49*, 1617–1629.

[270] Rice, O. K.; Ramsperger, H. C. Theories of unimolecular gas reactions at low pressures. II. *J. Am. Chem. Soc.* **1928**, *50*, 617–620.

[271] Kassel, L. S. Studies in homogeneous gas reactions. *J. Phys. Chem.* **1928**, *32*, 225–242.

[272] Marcus, R. A. Unimolecular dissociations and free radical recombination reactions. *J. Chem. Phys.* **1952**, *20*, 359–364.

[273] Fain, B. Theory of rate constants: master equation approach. *J. Stat. Phys.* **1981**, *25*, 475–489.

[274] Onsager, L. Initial recombination of ions. *Phys. Rev.* **1938**, *54*, 554–557.

[275] Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. On the transition coordinate for protein folding. *J. Chem. Phys.* **1998**, *108*, 334–350.