Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Combining Experiment and Energy Landscapes to Explore Anaerobic Heme Breakdown in Multifunctional Hemoproteins

Alasdair D. Keith,*,† Elizabeth B. Sawyer,† Desmond C.Y. Choy,† Yuhang Xie,† George S. Biggs,† Oskar James Klein,† Paul D. Brear,‡ David J. Wales,*,† and Paul D. Barker*,†

†*Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom*

‡*Department of Biochemistry, Sanger Building, University of Cambridge, Cambridge CB2 1GA, United Kingdom*

E-mail: adk44@cam.ac.uk; dw34@cam.ac.uk; pdb30@cam.ac.uk

Contents

A Methods **4**

B Reviewed Operons **16**

C Canonical Heme Oxygenase Reaction Mechanism **17**

D The Fur box and the P Protein as Controlling Elements for Heme Uptake **18**

E Phylogenetic Tree **22**

F Selected Protein Homologies **23**

G Genetic Sequences **24**

H Protein Sequences **28**

I Sequence Conservation **33**

J E197 *vs* Q197 Comparative Analysis **36**

K Homologue Properties **40**

L Heme Structure **42**

M Experimental Heme-Binding Properties **43**

N Calculated Heme-Binding Properties **47**

N.1 1,2-*bis*heme-HemS 48

N.2 1,1-*di*heme-HemS 48

N.3 Conclusion 49

O Heme-Binding in a 1 : 8 HemS : Heme Mixture **50**

P Small Molecules Used to Probe Disruption to Heme-Binding **52**

Q NADH-Induced Heme Breakdown with HemS – Raw Data **53**

R HPLC and MS Comparing Heme, Deuteroheme and Mesoheme – Raw Data **58**

S LCMS of Aerobic Reaction **65**

T NADPH **66**

U X-ray Crystallography Structure of the HBP in HemS **68**

V NADD NMR **69**

W DNA-Binding in ChuS and ShuS **70**

A Methods

The methods given here are extracted from the doctoral theses of Sawyer,[1] Choy[2] and Keith.[3] For full details, please refer to those resources.

General Materials & Conditions. Unless otherwise stated, all chemicals were purchased from Merck UK. Bovine hemin stocks were from Fluka Biochemika. All solutions were prepared with deionized 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. pH-adjusted solutions were filtered through 0.2 µm membranes (Sartorius Stedim Biotech). Protein concentrations were determined by UV-Vis Spectroscopy using a Cary 60 spectrophotometer (Agilent Technologies) and applying the experimental Beer-Lambert Law. ExPASY ProtParam[4] 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.[1,2,5–7] All relevant ϵ_{280} values and formula masses are given in Table S5.

Buffers.

–*Affinity Chromatography Buffer A.* 20 mM bis-tris propane (BTP), 10 mM imidazole (Acros Organics), 300 mM KCl (Fisher), pH 6.5.

–*Affinity Chromatography Buffer B.* 20 mM BTP, 500 mM imidazole, 300 mM KCl, pH 6.5.

–*Size Exclusion Chromatography Buffer (SEC).* 20 mM BTP, pH 6.5.

–*Anion Exchange Chromatography Buffer A (AEC A).* Same as SEC.

–*Anion Exchange Chromatography Buffer B (AEC B).* 20 mM BTP, 500 mM KCl, pH 6.5.

–*Thrombin His-Tag Cleavage Buffer (THTC).* 20 mM BTP, 150 mM NaCl (Fisher), 1.5 mM $CaCl_2$ (VWR International), pH 8.4.

–*ShuS Lysing Buffer.* 50 mM tris-HCl (USB Corporation), 1 mM EDTA, 1 mM $MgCl_2$, pH 8.0.

–*Crystallization Buffer A.* 50 mM HEPES (Sigma Aldrich), 150 mM NaCl, pH 8.0.

–*Crystallization Buffer B.* 100 mM tris-HCl (USB Corporation), 1.8 M $(NH_4)_2SO_4$ (Breckland Sci-

entific Supplies), 2% (w/v) PEG 400, pH 8.5.

–*NADD Buffer A.* 10 mM NH_4HCO_3 (BDH Laboratory Supplies), pH 8.5.

–*NADD Buffer B.* 500 mM NH_4HCO_3, pH 8.5.

Heme Stocks. Hemin chloride stock solutions were made by initially dissolving the solid in 1.0 M NaOH followed by dilution with buffer solutions to suitable concentrations of heme (1.0 mM or 0.1 mM final concentration).

Plasmid Preparation. The *hemS* gene in a pGAT2 expression vector[8] was provided by Sabine Schneider and Max Paoli (University of Nottingham, UK). The sequence in this plasmid disagreed with the crystal structure sequence (2JOP in the Protein Data Bank)[9] at residue 333, where an aspartic acid had been replaced with a glutamic acid. The full sequence is provided in Section H. 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.

Synthetic genes for HmuS, ChuS and ShuS were designed and synthesized 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 (own group stocks). Standard RD protocols (NEB) were followed.[10] Quick CIP was not added as it was found to limit ligation downstream. Fragments were separated on a 0.8% agarose gel, and the inserts/backbone excised. These were then extracted using a QiaEXII Gel Extraction Kit (Qiagen) and 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 lysogeny broth (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 / Purification. Protein was expressed 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 LB, before being spread onto LB agar ampicillin plates. 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 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 homogenizer (EmulsiFlex-C5 – Biopharma Process Systems) and the insoluble components removed by centrifugation (12,000 rpm, 50 min, 4 °C) using an F21S rotor (Thermo Scientific).

The purification protocol used was based on that outlined by Schneider and Paoli[11] 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 (GE Healthcare). A linear gradient (0–500 mM KCl) was applied to elute HemS from the column. The yield of protein was typically 20 mg per 1 L *E. coli* cells harvested.

For ShuS, the protocol was adapted slightly to prevent DNA-protein aggregation. 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). The cells were then lysed by a high pressure homogenizer (EmulsiFlex-C5 – Biopharma Process Systems) and the protein purification process continued as described above. Yields proved to be very similar to those for the other homologues once these measures were taken. However, pH levels were always maintained at 8.0 rather than 6.5 during storage, because ShuS has a tendency to precipitate at lower pHs.

Small molecule screening experiment with HemS. The assay was conducted in a quartz 96-well plate (Starna Scientific) using a FLASHScan 550 plate reader (Analytik Jena). Experiments were all conducted in 20 mM BTP, pH 7 buffer. The order of addition of ligands to HemS (i.e. small molecule or heme) was also investigated.

Stopped-flow Spectroscopy. All experiments were run using an SX 20 stopped-flow spectrometer (Applied Photophysics), with a Xe lamp light source and entrance/exit slit widths both set to 2 mm. Detection was achieved using an absorbance photomultiplier (Applied Photophysics) with appropriate glass filters. All spectra were taken against a baseline of SEC buffer, and temperature was maintained at 25 °C using a water bath. ProK-IV software (Applied Photophysics) was used in the analysis.

UV-Visible Spectroscopy. Spectra were collected either on a Cary 60 (Agilent Technologies), Cary 100 (Varian Ltd) or Cary 400 (Varian Ltd) spectrophotometer. 10 mm pathlength standard 9/9/B quartz cuvettes (Starna Scientific) were used throughout. Scanning was set for 250–800 nm,

with a scan rate of 300 nm min^{-1}. **Heme-binding.** Fresh heme stocks were always made on the day, to approximately 500 μM. Calibration was achieved first by taking a baseline of buffer (20 mM BTP, 100 mM KCl, pH 6.5). Cytochrome b_{562} was then added to approximately 15 μM. Heme was then added to approximately 5 μM. A crystal of sodium dithionite was added and parafilm used to seal the cuvette. Scans were made every 2 minutes until the spectra stabilized (typically 10–20 minutes). The absorbance at 426.5 nm ($\epsilon_{426.5}$ = 181,000 M^{-1} cm^{-1}) was then tracked in order to determine the concentration, thereby calibrating the heme stock solution. Following this calibration, a baseline of known volume of buffer (20 mM BTP, 100 mM KCl, pH 6.5 or pH 8.0) was taken. The protein (HemS, HmuS, ChuS or ShuS) was then added to approximately 10 μM. Using the calibrated stock, heme was then added to exactly 5 μM, and scans taken every two minutes until the spectra stabilized (ca. 10–40 minutes). For each protein-pH combination, the experiment was duplicated. The last five readings from these duplicated runs were pooled together and averaged. Normalisation so that the reading at 800 nm corresponded to zero was carried out across the entire wavelength range sampled. Molar absorbances were then achieved by converting the reading at the Soret maximum to its ϵ_{SM} value, and applying this correction across the entire wavelength range. **Steady State Reaction of *holo*-HemS with NADH.** For the spectra in Fig. 4, a baseline of buffer solution was recorded, and this baseline subtracted from the sample data. Unless otherwise stated in the text, the baseline for all other spectra consisted of the protein, heme and buffer, with data collected upon injection of NADH. It was important to allow the protein and heme to mix thoroughly before recording spectra to allow free and bound heme to equilibrate (>60 min). 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.

Mass Spectrometry. Small molecules. Data 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}. **Proteins** Data 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[12] 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. **Non-denaturing MS (ndMS).** Experiments were performed using a Q-ToF II instrument (Waters) modified by MS Horizons with the addition of a high-pressure inlet sleeve (to allow gentler desolvation) and tuning of the quadrupole (to allow for high m/z ions to be detected). Following standard preparation, samples were exchanged using Bio-Spin columns (Bio-Rad) into 200 mM ammonium acetate pH 7.0 buffer. 1-2 μL samples were placed in a gold-coated nanoflow capillary and delivered to the mass analyzer by nano-ESI. Cone and extractor voltages and gas flow rates were adapted for the respective experiments, with the cone voltages and extractor voltages always kept between 100–150 V and 0–20 V, respectively.

NADD Synthesis. The two NADD stereoisomers were synthesized and purified using methods adapted from Northrop & Duggleby,[13] Basran *et al.*[14] and Pudney *et al.*[15] **(*R*)-[4-^{2}H]-NADD.** 11.4 mL NADD Buffer A 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 (GE Healthcare) pre-equilibrated in NADD Buffer A. The column was washed with further NADD Buffer A. A gradient of NADD Buffer B was then introduced. (*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.[14] Stocks were frozen at −20 °C. **(*S*)-[4-^{2}H]-NADD.** 15.2 mL NADD Buffer A 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. Anion exchange, fraction pooling and freeze-drying were then carried out as described for (*R*)-NADD. **Characterization.** To ensure that the correct stereoisomers had been made, these samples were characterized by ^{1}H NMR spectroscopy using a 400 MHz Bruker Avance NMR spectrometer with a cryoprobe. These spectra, confirming the correct stereoisomers were made, are given in Fig. S30.

***holo*-HemS Reaction with NADD.** UV-Visible spectroscopy and MS were carried out as for standard protein + heme + NADH experiments described above, except that baselines were taken with NADH/NADD, and heme then added to initiate the reaction. This protocol was to prevent distortion of the spectra at the Soret maximum, allowing for heme-binding to be tracked more effectively. Initial rates of reaction were calculated by subtracting the absorbance value at 1 minute from the absorbance value at 10 minutes, and dividing by the time difference between these two readings.

Maintaining anaerobic conditions. When required, anaerobic conditions were achieved by fitting vessels containing appropriate stocks of both *holo*-HemS and NADH with Suba seals and thoroughly degassing them (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. As in the case when anaerobic conditions were not maintained (i.e. *holo*-HemS and NADH were simply mixed on the benchtop), a deep purple color developed over time.

HPLC Analysis of HBP. Analyses were carried out using an Agilent 1050 series HPLC instrument, fitted with an online de-gasser, binary pump, autosampler and PDA detector (Agilent Technologies) set to measure absorbance at 254 nm and 591 nm. Separation of reaction components was achieved on a 150×3.2 mm Kromasil-C8 column (Phenomenex) using a linear gradient of wa-

ter/acetonitrile with 0.1% formic acid. The flow rate was set to 1 ml min^{-1}. 5 μL sample was always injected onto a 95 : 5 water : acetonitrile + 0.1% formic acid-equilibrated column.

Non-HPLC Extraction / Purification of HBP. 75 mL Scale. Following reaction of pre-mixed 5 µM HemS : 20 µM heme with 2 mM NADH, 1-butanol was added to extract the HBP from the aqueous mixture. The organic layer was then dried using $MgSO_4$ and filtered under vacuum. 1-butanol was then drawn off by vacuum, leaving a dark purple residue. **10 mL Scale.** HemS and heme were mixed stoichiometrically to minimize residual unbound heme. Therefore, following reaction of pre-mixed 10 µM HemS : 10 µM heme with 2 mM NADH, 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 deionized water (5×20 mL), removing the protein and NADH. The heme breakdown product was then eluted as a deep purple solution using 3 mL acetonitrile.

Nuclear Magnetic Resonance Spectroscopy. The extracted heme 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.

X-ray Crystallography. Crystallization. Pre-mixed 10 µM protein : 10 µM heme were reacted with 2 mM NADH anaerobically in a total volume of 17.6 mL. The product mixture was then exchanged into Crystallization Buffer A and concentrated to 30 mg mL^{-1} by centrifugation (5700 rpm, 4 °C) using 10 kDa Amicon Ultra centrifugal filters. 5 µL samples were applied to individual glass coverslips and mixed with 5 µL Crystallization Buffer B. These coverslips were then applied to individual wells for crystallization by the hanging drop method, with 700 µL Crystallization Buffer B used as the precipitant solution in each well. Crystals were incubated at 277 K for 2-4 weeks. **Data Collection and Analysis.** 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[16] were utilized for the structure determination. Structures were solved by using programs in the CCP4 package.[17] Models were iteratively refined and rebuilt using the Refmac[18] and Coot[19] programs.

Relibase$^+$. The Relibase$^+$ package[20] (now retired) was used according to the developers' instructions.[21] This package was used to identify likely binding sites for NADH in *holo*-HemS. It was also used to identify likely secondary heme sites to give 1,2-*bis*heme-HemS and 1,2-*di*heme-HemS (see Section N). Specific settings followed those described in Choy.[2]

GPU Implementation. All calculations were run on GPUs using interfaces of the Cambridge energy landscapes software to AMBER.[22]

Starting Configurations and Force Field Settings. Initial coordinate files were taken from the *holo*-HemS structure, 2J0P.[9] Solvent and crystallant molecules were removed and the missing loop reconstructed using Swiss-PdbViewer.[23] NADH was added to the cavity using coordinates determined by Relibase$^+$.[20] AMBER LEaP[24] was used to generate topology files and add hydrogen atoms. None of the non-histidine charged residues were neutralised since they were either exposed to the surface or part of a salt bridge. However, all histidine residues were neutralised, 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 heme, as in the crystal structure. An explicit bond was placed between heme FE and H196 Nϵ in the HemS system to aid heme-anchoring studies and was retained in later investigations to maintain consistency. This bond was not included for HmuS, ChuS or ShuS, where such heme-anchoring investigations were not conducted. Antechamber[25] was used to prepare the parameters for non-protein molecules before uploading to LEaP. All parameters were taken from the database curated by the Bryce group at the University of Manchester.[26] Specifically, those parameters used were the all-atom heme representation adapted from Giammona,[27] the NADH representation from Walker[28,29] and the NADPH representation from Ryde.[30] Starting coordinate files for the HmuS, ChuS and ShuS systems were generated using **CHECKSPMUTATE** (discussed below). All simulations were run using the ff99SB force field[31] from the AMBER12 package,[32] following benchmarking exercises (see Choy[2] and Keith[3] for details). The implicit generalised Born, *igb2*,[33] solvent model with no cutoff and an effective monovalent ionic concentration of 0.1 M using the Debye-Hückel approximation[34] was selected for all calculations following further benchmarking.

Landscape Exploration – System Sampling. Sampling of the *holo*-HemS + NADH sys-

tem was achieved by basin-hopping (BH) global optimization,[35–39] as implemented in the GMIN program[40] interfaced with AMBER12.

Landscape Exploration – Creating Kinetic Transition Networks (KTNs). KTNs were created using the discrete path sampling (DPS) strategy,[41–43] as implemented in the OPTIM program[44] interfaced to AMBER12. Transition state candidates were located using the doubly-nudged elastic band (DNEB) algorithm,[45,46] and then converged using hybrid eigenvector-following (HEF).[47–49] The local minima directly connected to these transition states were then identified using an L-BFGS-derived steepest-descent algorithm.[50,51] Selected minima were then connected to one another iteratively using the Dijkstra shortest path algorithm,[52] which describes the total set of minima as a weighted, directed graph.[53] This approach allows for the construction of a priority list to efficiently attempt connections based on edge weights. To further improve connectivity and refine the KTN, the PATHSAMPLE program[54] program (which acts as a driver for OPTIM) was used. The PATHSAMPLE strategies employed were the SHORTCUT scheme,[53,55,56] which shortens artificially long pathways; the SHORTCUT BARRIER scheme,[55,56] which removes artificially large barriers; the UNTRAP scheme,[55,56] which removes artificial kinetic traps; and the CONNECTUNC scheme,[57] which connects unconnected minima to the main set, until rate constants are converged. A new feature of the CONNECTUNC scheme was developed specifically for this work. From a consideration of all of the sets of interconnected stationary points (known as sub-databases) within an overall database, the new algorithm determines the fewest and narrowest 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) within the database. This calculation is achieved through a consideration of which sub-databases would be best to try to connect directly, as well as which minima to select for connection within those sub-databases. This approach allowed for the efficient connection of clusters of stationary points with NADH at various points of progress along the *holo*-HemS pocket. For further details, see Keith.[3]

Templating Strategy. A new subroutine, CHECKSPMUTATE, was developed to grow the *holo*-HmuS + NADH, *holo*-ChuS + NADH and *holo*-ShuS + NADH systems using the *holo*-HemS + NADH system as a template, rather than growing these databases from scratch, using the methods described for the *holo*-HemS + NADH system detailed in the sections above. Briefly, CHECKSPMUTATE took

all of the stationary points from the 'template' *holo*-HemS + NADH system database (or a subset thereof, in this work the Dijkstra fastest pathway[52,58] for NADH motion towards heme) and mutated, inserted or deleted all of the relevant residues to transform HemS to HmuS, ChuS or ShuS, respectively. Following these mutations, these new stationary points were reoptimized. If any gaps arose, these were filled by the standard methods in **PATHSAMPLE** and **OPTIM** described above. Multiple rounds of **PATHSAMPLE**-based refinement calculations were also used to ensure that the pathways for these new systems were optimal, and not unduly influenced by the original *holo*-HemS + NADH system. We estimate that templating gave a 10-20 fold speed-up with respect to the approach used to generate the original database described above. For full details, see Keith.[3]

Visualization and Analysis of Landscapes. Energy landscapes were represented using disconnectivity graphs.[59,60] These graphs consist of nodes that correspond to superbasins and branches corresponding to local minima. The vertical axis represents the potential energy of the system, whereas the branch positions on the horizontal axis are chosen to give the clearest horizontal representation of the data. Superbasins are calculated at fixed energy intervals, E_n, where the minima are classified into disjoint sets. These graphs therefore provide a useful pictorial guide to the structure of an energy landscape. Further properties can be inferred by color-coding the graphs, for example, using NADH-heme distances within each minimum. Specifically, these NADH-heme distances were calculated using the Cartesian distance between the β-*meso*-carbon in heme and the hydride-bearing carbon in NADH within each minimum, since we expect that these are the two carbon atoms involved in the hydride transfer process. PyMol[61] and VMD[62] were used to visualize structures and pathways.

Phylogenetic Analysis. Homologous sequences of HemS were collected using the default settings of pBLAST,[63] except for an increase in the maximum number of sequences to 5000. The sequences identified issued from a total of 218 different genera. One sequence from each of these 218 genera was manually selected, and a maximum-likelihood phylogenetic tree was generated in MEGA-X[64] using default settings, apart from the bootstrap setting being extended to 2000. Visualization, manipulation and export of the tree was achieved using iTOL,[65] and is available to view in Section E.

DNA-Binding Analysis. Published structures for *apo*-HemS (PDB: 2J0R)[66] and *apo*-ChuS

(PDB: 1U9T)[67] were selected for analysis. Since there was no published structure for ShuS, an *apo*-ShuS structure was modelled from the *apo*-ChuS structure using SwissModel.[68] The DNAbind package was used to predict DNA-binding residues, as described in Liu.[69]

B Reviewed Operons

Table S1: List of selected operons, and their respective sources, studied for heme breakdown capabilities.

Operon	**Source**
hem	*Yersinia enterocolitica*
hmu	*Yersinia pestis*
chu	*Escherichia coli* O157:H7
shu	*Shigella dysenteriae*
phu	*Pseudomonas aeruginosa*

C Canonical Heme Oxygenase Reaction Mechanism

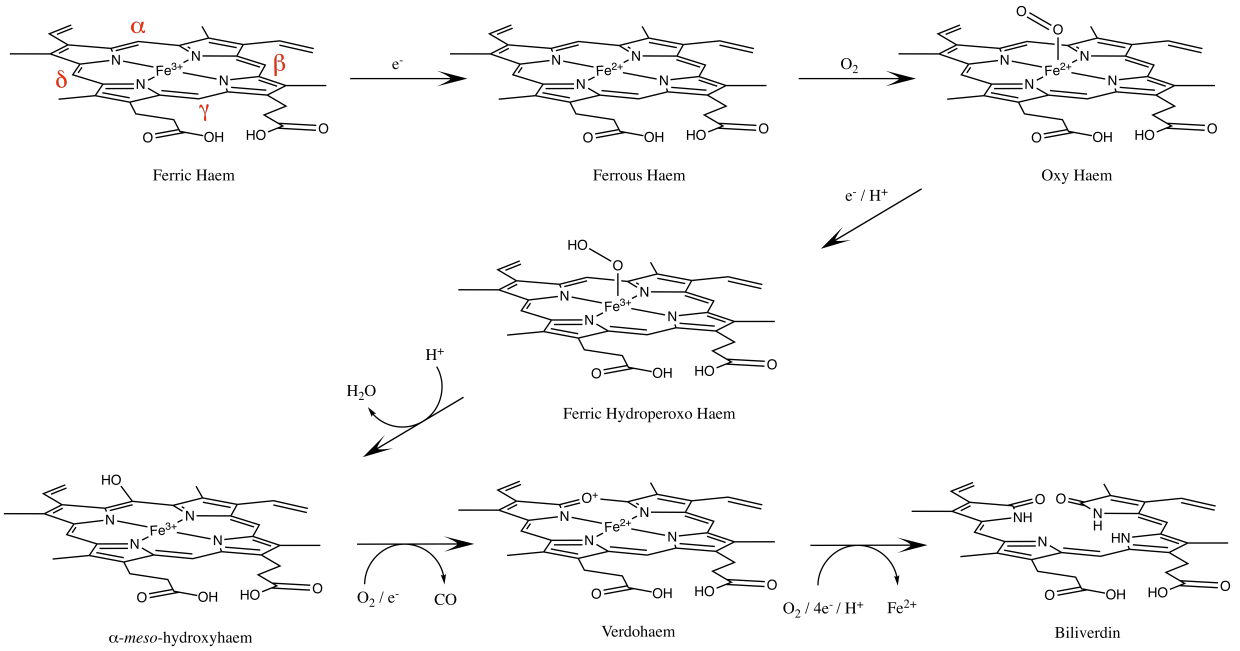


Figure S1: Schematic of the canonical HO mechanism. Ferric heme is first reduced to ferrous heme so that it can coordinate O_2. Further reduction gives a ferric hydroxoperoxo heme. The hydroxyl group is then transferred to the porphyrin ring itself at the α-*meso*-position. Oxidation at this position yields verdoheme and expels CO. Verdoheme can then be cleaved, releasing iron and producing biliverdin. Figure inspired by Unno *et al.*[70]

D The Fur box and the P Protein as Controlling Elements for Heme Uptake

HemP, derived from the *hem* operon of *Yersinia enterocolitica*, is a poorly studied protein. Original studies by Stojiljkovic & Hantke determined that this protein was 81 amino acids long and approximately 8.5 kDa in weight but were unable to isolate it.[71] However, it was shown that this protein was required (along with HemR and HemS) for *Y. enterocolitica* to grow under iron-deficient conditions.[71] Furthermore, studies on other pathogenic heme uptake systems have suggested that their corresponding versions of HemP are involved in transcriptional activation of the genes encoding the outer membrane heme receptors.[72–74]

In the *hem* operon, the *hemP* gene (which has an open reading frame spanning nucleotides 373-615)[71] overlaps an overall Ferric Uptake Regular (Fur) box. A model for Fur repression is given in Fig. S2.

Though they never considered the role of HemP in their studies, data from Jacobi *et al.*[75] may provide insight into the role of this protein, and its possible role working in conjunction with the standard Fur box. Their research showed three distinct levels of HemR expression when *Y. enterocolitica* was placed in different mouse tissues.[75] In tissues where the iron content is high, such as those from the liver and intestinal lumen, HemR expression was weak. This reflects the impulse *Y. enterocolitica* has not to become overloaded with iron. Meanwhile, in tissues where there is a moderate iron content, such as those from the spleen, there is moderate expression of HemR. Interestingly, however, in tissues where iron content is severely limited, such as those from the peritoneum, HemR was 'hyperexpressed'. Jacobi *et al.* concluded that a third level of HemR expression must be due to additional activators besides the Fur box. At the time of their study, it was not known that HemP could be involved in transcriptional activation. Combining these findings, it is speculated that HemP could help to control HemR expression in the manner illustrated in Fig. S3. Studies will be required to determine whether this idea is correct or not.

Of the other operons studied in detail in the main paper, only *hmu* codes for an obvious P protein, HmuP. This protein was found to be half the size of HemP (41 amino acids long, as opposed to

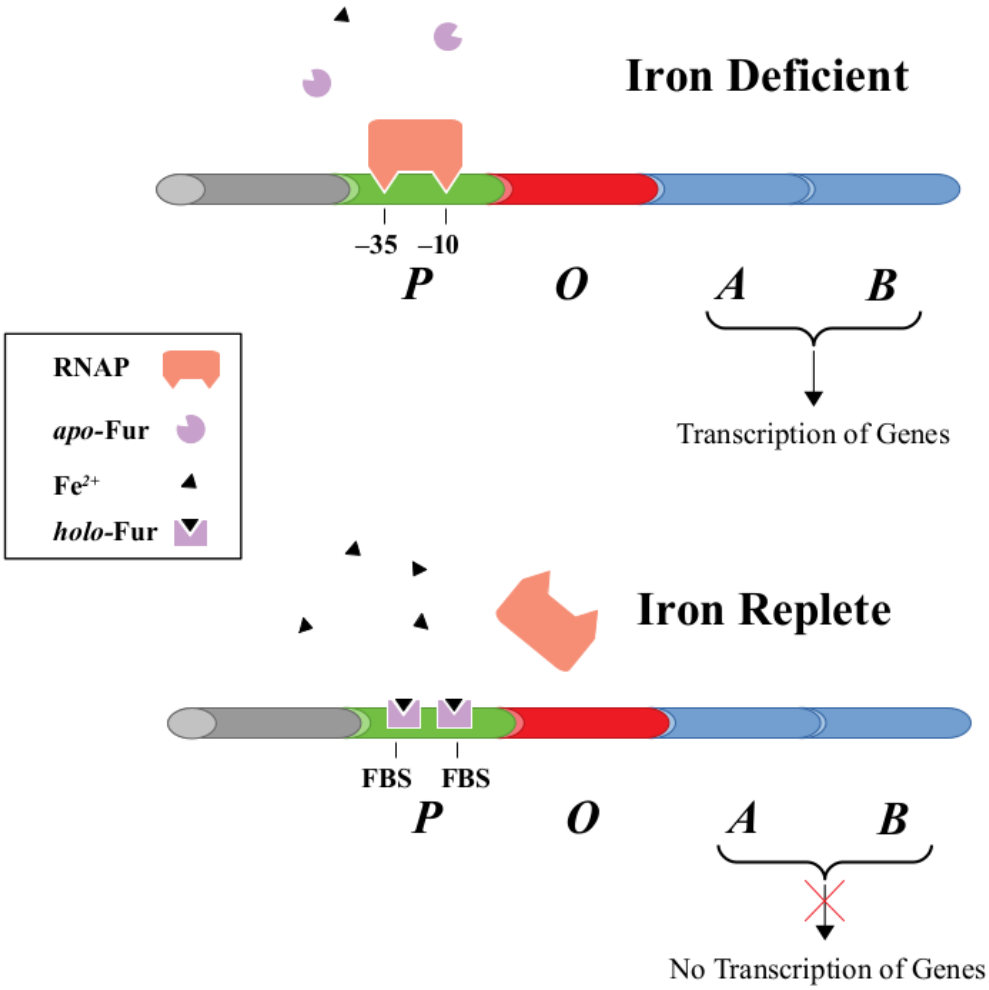


Figure S2: Model of Fur repression. The promoter, P (green), and operator, O (red), control the expression of the structural genes, A, B, etc. (blue). Under iron limiting conditions, transcription of genes occurs as the RNA polymerase (RNAP) is free to bind to the -35 and -10 sites of the promoter. Under iron replete conditions, the coordination of Fe^{2+} to Fur significantly increases its affinity for the Fur-binding sites (FBSs) as a result of conformational changes. As these FBSs overlap with the -35 and -10 sites, this effect prevents the coordination of RNAP to the promoter, and hence switches off the transcription of downstream genes.

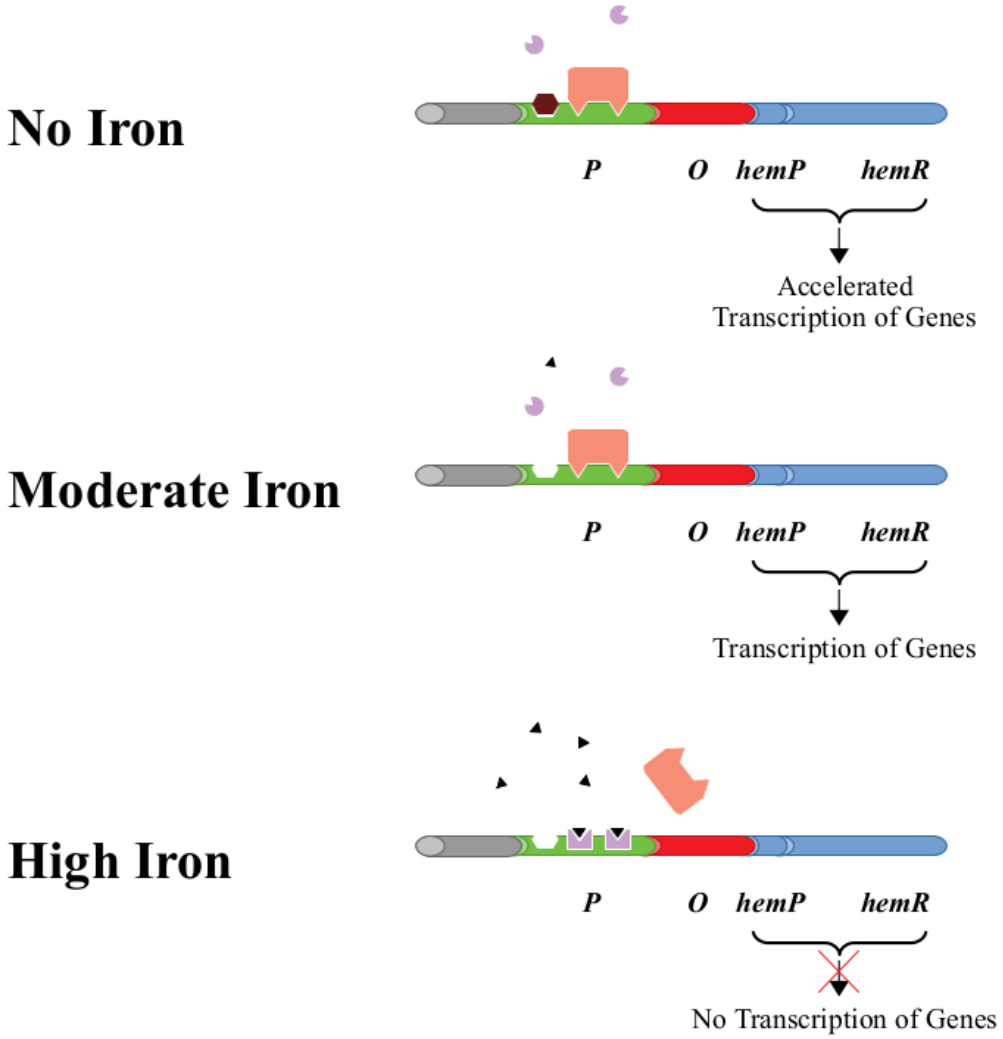


Figure S3: Proposed model for HemR expression based on iron availability in *Y. enterocolitica*. When iron levels are moderate or high, expression behaves in a similar manner to a standard Fur model, as described in Fig. S2. However, at very low iron concentrations, HemR is 'hyperexpressed'. This effect could be due to transcriptional activation by HemP (brown hexagon), although how such activation can only be induced at very low iron concentrations remains unknown, and further studies are required.

81). This variation in length gives credence to the idea that these are transcriptional regulators[72] as opposed to structural proteins, as protein length tends to be less important for function within this class of proteins.

E Phylogenetic Tree

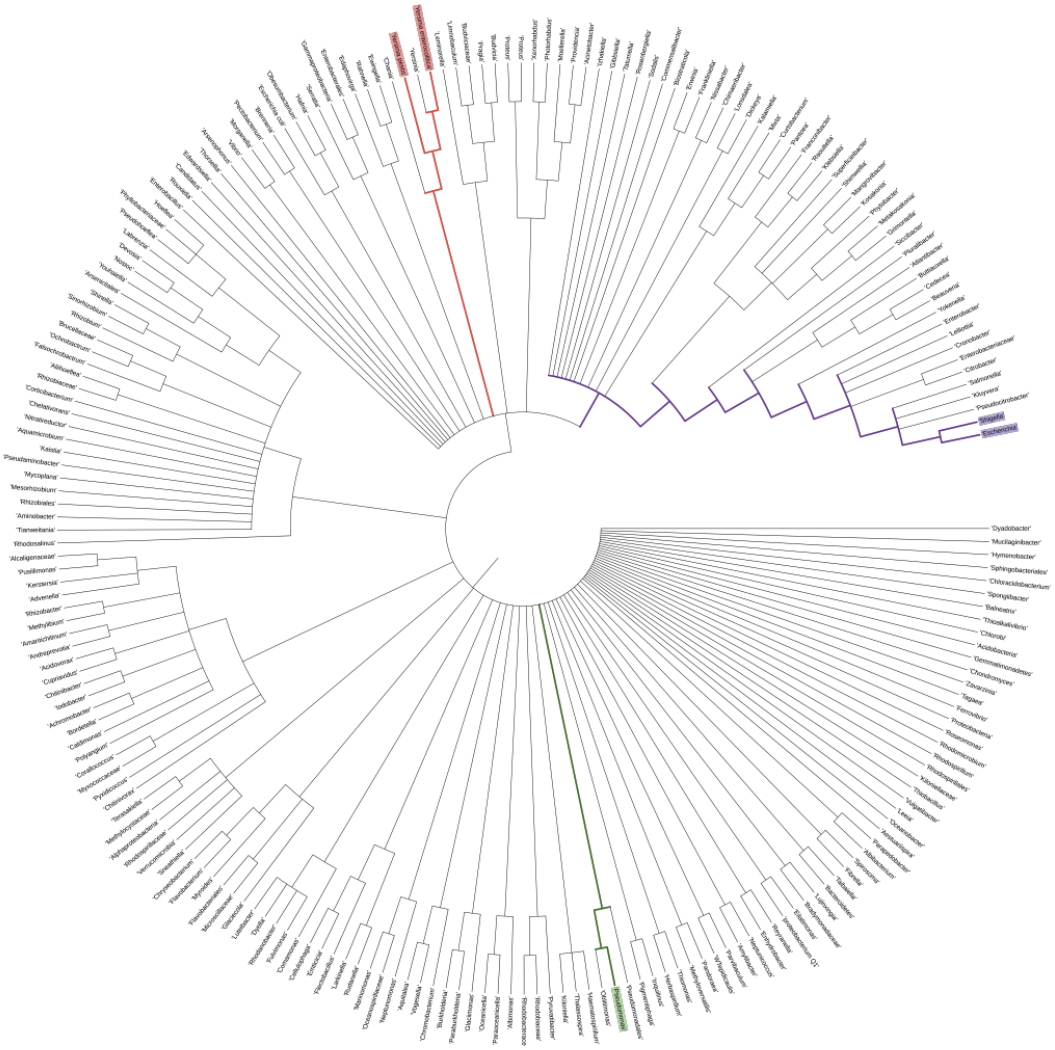


Figure S4: Maximum likelihood phylogenetic tree for HemS and its homologues. 218 different genera are represented, including both pathogenic and non-pathogenic bacteria. Selected species are highlighted. Red: *Yersinia enterocolitica* (HemS), *Yersinia pestis* (HmuS). Purple: *Escherichia coli* (ChuS), *Shigella dysenteriae* (ShuS). Green: *Pseudomonas aeruginosa* (PhuS).

F Selected Protein Homologies

Table S2: Accession numbers of the protein sequences used to calculate % homologies.

Protein	**Bacterium**	**Accession Number**
HemS	*Yersinia enterocolitica*	WP_005156541.1
HmuS	*Yersinia pestis*	WP_002209061.1
ChuS	*Escherichia coli* O157:H7	NP_312406.1
ShuS	*Shigella dysenteriae*	WP_001017208.1
PhuS	*Pseudomonas aeruginosa*	NP_253397.1

Table S3: XxxS Homologies.

% Identity

	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	–

% Similarity

G Genetic Sequences

The dnaseq package[76] was used to represent the data given below. For full details of the genetic data in this work, see the appendices of Keith.[3]

HemS – DNA Sequence

```
   1 atgagcaaat caatatacga gcagtatcta caagctaaag cagataatcc gggcaaatat
  61 gcgcgcgatt tggccacgct gatggggatt tcagaagcgg aactgaccca tagccgcgtt
 121 agtcatgatg ccaaacgtct gaaaggtgat gcccgcgcac tactggccgc attggaagct
 181 gtcggtgagg tcaaagctat cacccgcaac acctatgccg tacatgagca aatgggccgt
 241 tatgaaaatc aacatctgaa tggccatgct ggtttgatcc tcaatccacg caatttagat
 301 ttacgcctgt tcctcaacca gtgggccagc gcattcacgc tgacagaaga aactcgccac
 361 ggtgtacgcc atagcatcca gtttttcgac catcaaggcg atgctctgca taaagtgtat
 421 gtcactgaac aaactgacat gccagcctgg gaagcgctac tggcgcagtt tatcaccaca
 481 gaaaatccag agttacagct agagccactg agcgcacctg aagtcactga accgacagcc
 541 accgatgaag ctgtcgatgc tgaatggcgt gctatgactg acgtgcatca gttcttccag
 601 ttgctcaaac gcaataattt gacccgtcag caagccttcc gtgccgtggg taatgatctg
 661 gcttatcagg ttgataacag ttctctgacc cagttactga acattgctca gcaagaacag
 721 aatgaaatca tgatttttgt gggtaaccgt ggctgtgtac aaatattcac cggcatgatt
 781 gaaaaggtta caccacatca agattggatt aatgttttca accagcgctt cacgctgcat
 841 ctgattgaaa caacgattgc tgaaagctgg attacccgca agccaacaaa agacggtttc
 901 gtgaccagtt tggaactgtt tgctgctgat ggcacccaaa ttgcacaact ttacggtcag
 961 cgcaccgaag gccagccaga acaaacgcaa tggcgtgagc aaattgctcg cctcaataat
1021 aaggatatcg ccgcatga
```

HmuS – DNA Sequence

```
   1 aacgcatcat tataccaaca atatgtacag gctaaagcag agcaccctgg caaatatgcc
  61 cgtgatttag ccaccctgat ggggatttca gaagcagagc tgacccatag ccgcgtcggg
 121 catgatgcaa aacgtttaca aagtgatgct cgtgcattat tggccgcatt ggaatccgtc
 181 ggcgaagtca aagccattac ccgcaacacc tatgcagttc atgagcaagt gggccgctat
 241 gagaaccaac acttaaatgg tcatgcaggg ttaatcctca atccacgcgc cttggacctc
 301 cggttattcc tgaatcagtg ggcaagcgcc tttacactga ccgaagagac ccgccacggc
 361 gtgcgccata gcatccaatt tttcgaccat cagggcgatg cattacacaa agtgtatgtg
 421 acagaacaga cagatatgtc tgcctgggaa gccttgctgg cacaatttat catcccggaa
 481 aacccggcat tgcagttaga acctttgagc accccagaag cggtagaacc tacagccgat
 541 gatgcaaccg tggatagcga atggcgtgcc atgaccgatg tacaccagtt cttccaactg
 601 cttaaacgca ataatctgac ccgtcagcag gcgttccgcg ctgttggtga tgatctggct
 661 taccaggtcg ataacaactc actgactcag ctgttgcaca tcgcccagca agatcagaac
 721 gagatcatga tttttgtcgg caaccgcggc tgtgtacaaa ttttcaccgg cctgattgaa
 781 aaagtcacac cacacaacga atggattaat gtcttcaatc agcgctttac actgcatctg
 841 atcgaaacgg ccattgccga aagctggatc acccgcaaac caacaaaaga cggttttgtc
 901 accagcctag aactgtttgc tgctgatggt actcaacttg cccaactcta cggccagcgc
 961 accgaagggc agccagaaca aaaccaatgg cgtgaacaga ttgcccgcct aatcaacaag
1021 gatatcgccg catga
```

ChuS – DNA Sequence

```
   1 atgaaccact acacacgctg gcttgagtta aaagaacaaa atcccggaaa gtacgcgcgt
  61 gacatcgcag ggttaatgaa tattagagaa gcagaactgg catttgcacg cgtcacgcac
 121 gatgcgtggc ggatgcacgg cgatatccgt gaaattctgg cggcgctcga aagtgttggc
 181 gaaaccaaat gtatttgtcg taatgaatat gcagtccatg agcaagttgg tacgttcaca
 241 aaccagcatt tgaacggaca tgccggattg atcctcaatc cgcgcgcgct ggatttacgt
 301 ctgtttctca atcaatgggc cagtgttttc cacatcaaag aaaacacggc tcgtggcgaa
 361 cgccagagta ttcagttctt tgatcatcag ggcgatgcat tactaaaagt ttatgccacc
 421 gacaataccg atatggcggc atggagtgag cttctggcac ggtttatcac cgatgagaat
 481 acgccgcttg agttaaaagc cgttgatgcg ccagttgttc aaacgcgagc cgatgccact
 541 gtggtcgagc aagagtggcg ggcgatgacc gacgttcatc agttttttac gttgctcaag
 601 cgccacaacc tgacgcgcca acaggcgttc aatctggtgg cagacgattt ggcctgcaaa
 661 gtatccaaca gtgcgttggc gcaaattctt gaatctgcac agcaggatgg taatgaaatc
 721 atggtgtttg ttggcaaccg tggctgcgta cagattttca ccggtgtggt agaaaaagtg
 781 gtgccaatga aaggttggct gaatattttc aacccgacgt ttactcttca tctattagaa
 841 gagagcattg ctgaagcctg ggttacccgt aaaccgacca gcgatggcta cgtaaccagt
 901 ctggaattgt ttgcccatga tggtacgcag atagcgcaac tttatggtca acgtacagaa
 961 ggcgaacagg agcaagcgca atggcgtaag caaattgctt cgctgatacc ggaaggcgtt
1021 gctgcataa
```

ShuS – DNA Sequence

```
   1 aaccactaca cacgctggct tgagttaaaa gaacaaaatc ccggaaagta cgtgcgtgac
  61 atcgcagggt taatgaatat tagagaagca gaactggcat ttgcacgagt cacgcacgat
 121 gcgtggcgga tgcgcggcga tatccgtgaa attctggcgg cgctcgaaag tgttggcgag
 181 accaaatgta tttgccgtaa tgaatatgca gtccatgagc aagttggtgc gttcacaaac
 241 cagcatttga atggacatgc cggattgatc ctcaatccac gcgcgctgga tttacgtctg
 301 tttctcaatc aatgggccag tgttttccac atcaaagaaa acacggctcg tggcgaacgc
 361 cagaggattc agttctttga tcatcagggc gatgcattac taaaagttta tgccaccgac
 421 aataccgata tggcggcatg gagtgagctt ctggcacggt ttatcaccga tgagaatatg
 481 ccgcttgagt taaaagccgt tgatgcgcca gttgttcaaa cgcgagccga tgccactgtg
 541 gtcgagcaag agtggcgagc gatgaccgac gttcatcagt tttttacgtt gctcaagcgc
 601 cacaacctga cgcgccaaca ggcgttcaat ctggtggcag acgatttggc ctgcaaagta
 661 tccaacagtg cgttggcgca aattcttgaa tctgcacagc aggatggtaa tgaaatcatg
 721 gtgtttgttg gcaaccgtgg ctgcgtacag attttcaccg gtgtggtaga aaaagtggtg
 781 ccaatgaaag gttggctgaa tattttcaac ccgacgttta ctcttcatct attagaagag
 841 agcattgctg aagcctgggt tacccgtaaa ccgaccagcg atggctacgt aaccagtctg
 901 gaattgtttg cccatgatgg tacgcagata gcgcaacttt atggtcaacg tacagaaggc
 961 gaacaggagc aagcgcaatg gcgtaaacaa attgcttcgc tgataccgga aggcgttgct
1021 gcataa
```

H Protein Sequences

There were some subtle differences between the experimental and computational sequences used. Sequences listed below were those used in laboratory-based experiments, with those residues highlighted in red denoting points of differentiation in the computational sequences used. Reasons for these differences are provided in the text below.

For full details, including a discussion of subtle variations these sequences have with respect to certain sequences found elsewhere in the literature, see the appendices in the PhD thesis of Keith.[3]

HemS – Protein Sequence (Experimental)

In the computational sequence, the N-terminal (MSK) and C-terminal (DIAA) residues were removed. This was to simplify computations since these residues, being far from the protein cavity plus not appearing in any experimental crystal structures, whether in the literature or in this work, were not thought to be important to protein structure or function. Residue 333 was converted from glutamic acid to aspartic acid; this mutation actually causes the computational sequence to be more in line with the published structure for HemS (PDB Code: 2J0P).[9] *It is thought that the gene gifted from Schneider and Paoli had a mutation at this position, causing all expressed HemS in this work to bear glutamic acid. As this residue is located in a solvent exposed loop, it was determined that this mutation should have little effect on function. The final point of differentiation is at residue 197, where the glutamine from the experimental sequence has been changed to a glutamic acid for computations. This mutation is due to a mistaken assignment in structure 2J0P deposited to the PDB.*[66] *Since we used this structure as a starting-point for our calculations, we therefore inherited this error. Though this residue is situated close to the main cavity, it consistently points away from the pocket and so its effect on protein function is assumed to be minimal. Furthermore, test calculations have been run to determine the effect of this different residue on heme-NADH interactions, with conclusions provided in Section J of the SI.*

```
  1 MSKSIYEQYL QAKADNPGKY ARDLATLMGI SEAELTHSRV SHDAKRLKGD ARALLAALEA
 61 VGEVKAITRN TYAVHEQMGR YENQHLNGHA GLILNPRNLD LRLFLNQWAS AFTLTEETRH
121 GVRHSIQFFD HQGDALHKVY VTEQTDMPAW EALLAQFITT ENPELQLEPL SAPEVTEPTA
181 TDEAVDAEWR AMTDVHQFFQ LLKRNNLTRQ QAFRAVGNDL AYQVDNSSLT QLLNIAQQEQ
241 NEIMIFVGNR GCVQIFTGMI EKVTPHQDWI NVFNQRFTLH LIETTIAESW ITRKPTKDGF
301 VTSLELFAAD GTQIAQLYGQ RTEGQPEQTQ WREQIARLNN KDIAA
```

HmuS – Protein Sequence (Experimental)

As with HemS (reasons explained there), the N-terminal (-NA) and C-terminal (DIAA) residues were removed for computations. Due to the computational sequence for HemS having acted as a template for the HmuS computational sequence, the error at residue 197 (glutamic acid, rather than glutamine) was inherited.

```
  1 -NASLYQQYV QAKAEHPGKY ARDLATLMGI SEAELTHSRV GHDAKRLQSD ARALLAALES
 61 VGEVKAITRN TYAVHEQVGR YENQHLNGHA GLILNPRALD LRLFLNQWAS AFTLTEETRH
121 GVRHSIQFFD HQGDALHKVY VTEQTDMSAW EALLAQFIIP ENPALQLEPL STPEAVEPTA
181 DDATVDSEWR AMTDVHQFFQ LLKRNNLTRQ QAFRAVGDDL AYQVDNNSLT QLLHIAQQDQ
241 NEIMIFVGNR GCVQIFTGLI EKVTPHNEWI NVFNQRFTLH LIETAIAESW ITRKPTKDGF
301 VTSLELFAAD GTQLAQLYGQ RTEGQPEQNQ WREQIARLIN KDIAA
```

ChuS – Protein Sequence (Experimental)

As with HemS (reasons explained there), the N-terminal (M) and C-terminal (EGVAA) residues were removed for computations.

```
  1 MNHYTRWLEL KEQNPGKYAR DIAGLMNIRE AELAFARVTH DAWRMHGDIR EILAALESVG
 61 ETKCICRNEY AVHEQVGTFT NQHLNGHAGL ILNPRALDLR LFLNQWASVF HIKENTARGE
121 RQSIQFFDHQ GDALLKVYAT DNTDMAAWSE LLARFITDEN TPLELKAVDA PVVQTRADAT
181 VVEQEWRAMT DVHQFFTLLK RHNLTRQQAF NLVADDLACK VSNSALAQIL ESAQQDGNEI
241 MVFVGNRGCV QIFTGVVEKV VPMKGWLNIF NPTFTLHLLE ESIAEAWVTR KPTSDGYVTS
301 LELFAHDGTQ IAQLYGQRTE GEQEQAQWRK QIASLIPEGV AA
```

ShuS – Protein Sequence (Experimental)

As with HemS (reasons explained there), the C-terminal (GVAA) residues were removed for computations. Residue 327 also differs (glutamic acid, rather than glutamine). As this residue is situated on a solvent-exposed loop, it should have little effect on protein function.

```
  1 -NHYTRWLEL KEQNPGKYVR DIAGLMNIRE AELAFARVTH DAWRMRGDIR EILAALESVG
 61 ETKCICRNEY AVHEQVGAFT NQHLNGHAGL ILNPRALDLR LFLNQWASVF HIKENTARGE
121 RQRIQFFDHQ GDALLKVYAT DNTDMAAWSE LLARFITDEN MPLELKAVDA PVVQTRADAT
181 VVEQEWRAMT DVHQFFTLLK RHNLTRQQAF NLVADDLACK VSNSALAQIL ESAQQDGNEI
241 MVFVGNRGCV QIFTGVVEKV VPMKGWLNIF NPTFTLHLLE ESIAEAWVTR KPTSDGYVTS
301 LELFAHDGTQ IAQLYGQRTE GEQEQAQWRK QIASLIPEGV AA
```

I Sequence Conservation

The following is the HemS experimental sequence, with residues highlighted in bold if they showed ≥90 % conservation with respect to the 218 homologues identified by the phylogenetic study. Residues forming the large, heme- and NADH-binding cavity are highlighted in cyan, and those forming the small, possibly DNA-binding cavity in salmon. This corresponds with the color-coding found in Fig. S5. % conservations were calculated using SnapGene,[77] and cavity-forming residues identified using the default settings in MetaPocket.[78] When a residue was assigned to both cavities, it was manually assigned to the cavity where its ranking was highest.

HemS Sequence – % Conservation & Cavity-Forming Residues

```
  1 MSKSIYEQYL QAKADNPGKY ARDLATLMGI SEAELTHSRV SHDAKRLKGD ARALLAALEA
 61 VGEVKAITRN TYAVHEQMGR YENQHLNGHA GLILNPRNLD LRLFLNQWAS AFTLTEETRH
121 GVRHSIQFFD HQGDALHKVY VTEQTDMPAW EALLAQFITT ENPELQLEPL SAPEVTEPTA
181 TDEAVDAEWR AMTDVHQFFQ LLKRNNLTRQ QAFRAVGNDL AYQVDNSSLT QLLNIAQQEQ
241 NEIMIFVGNR GCVQIFTGMI EKVTPHQDWI NVFNQRFTLH LIETTIAESW ITRKPTKDGF
301 VTSLELFAAD GTQIAQLYGQ RTEGQPEQTQ WREQIARLNN KDIAA
```

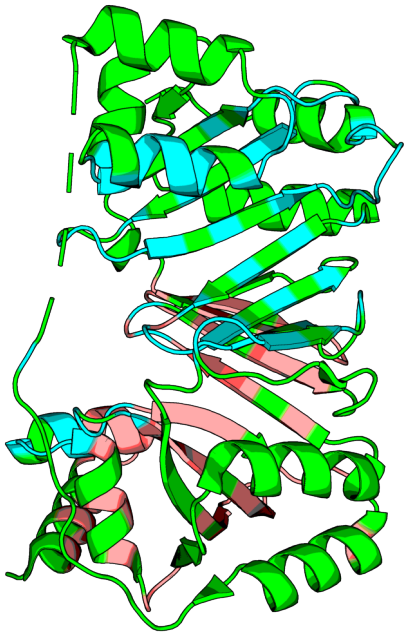


Figure S5: PyMOL representation of *apo*-HemS (PDB: 2J0R).[66] Residues are not explicitly shown, but their positions on the main chain are highlighted if they are involved in forming the main pocket (cyan) or secondary pocket (salmon).

Table S4: Sequence conservation of selected residues. Indexing of residues is according to their positions in HemS. Though residue 318 is a tyrosine in HemS, it is a phenylalanine across 69.7% of the homologues.

Function	Residue	% Conservation
Heme-binding	**R102**	**100%**
Heme-binding	**H196**	**100%**
Heme-binding	**R209**	**100%**
Heme-binding	**K294**	**98.1%**
Heme-binding	**F318**	**69.7%**
Heme-binding	**R321**	**99.1%**
NADH-binding	**Q132**	**29.8%**
NADH-binding	**K203**	**43.1%**
NADH-binding	**R250**	**61.5%**
NADH-binding	**T312**	**30.7%**
Phe-gate	**F104**	**98.6%**
Phe-gate	**F199**	**82.6%**

J E197 *vs* Q197 Comparative Analysis

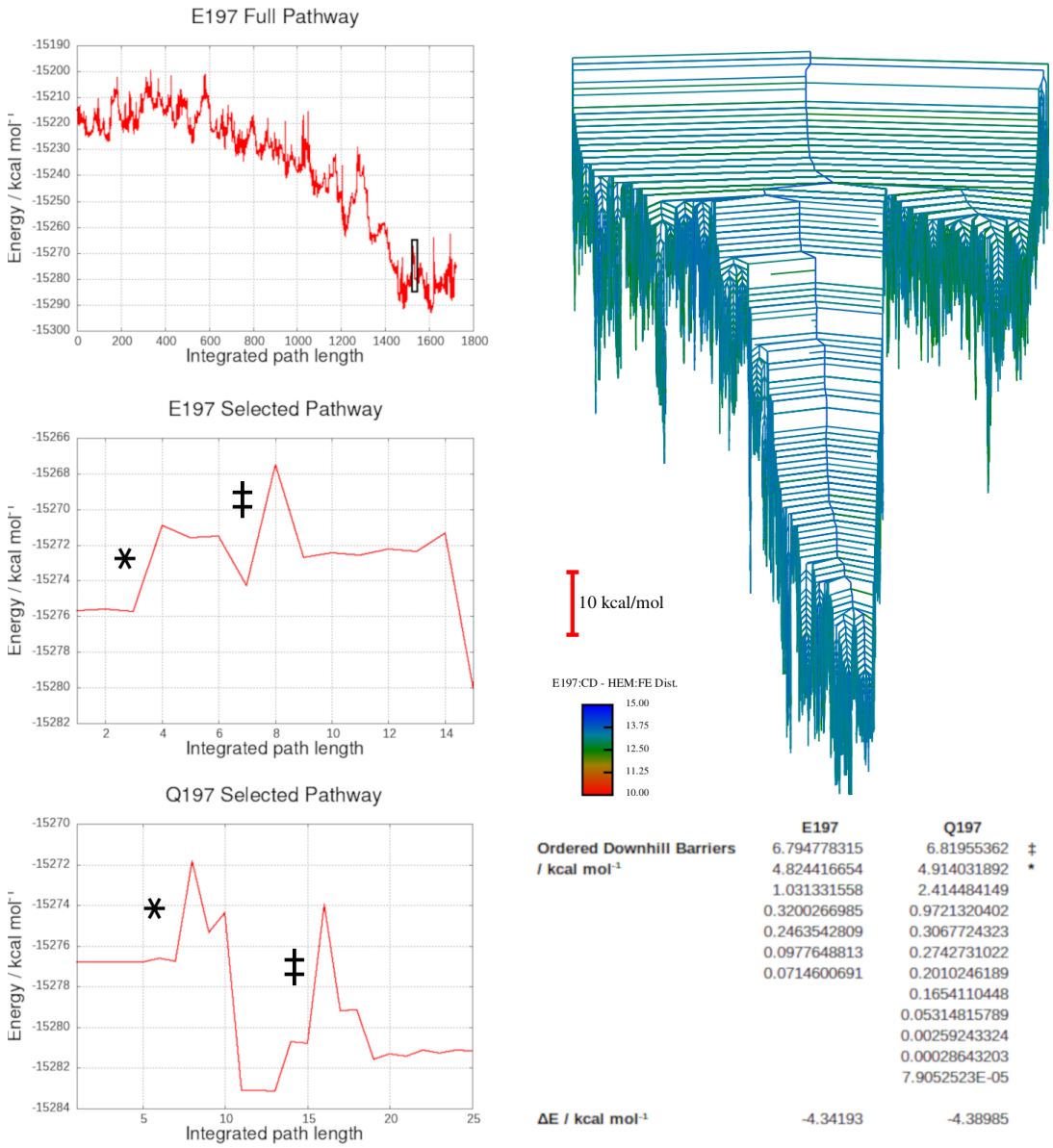
Section G explains how an error in the PDB-deposited *holo*-HemS structure, 2J0P,[66] propagated into the starting structure for our calculations. Hence the HemS, HmuS and ShuS structures erroneously have residue 197 assigned as a glutamic acid (E) rather than a glutamine (Q). In lieu of repeating all these time-consuming calculations, a comparative analysis was undertaken to determine whether the misassignment would have a significant effect on our conclusions.

Firstly, an analysis of the entire HemS database was performed to determine whether any of the 43,233 minima had E197 pointing towards the pocket, and therefore interfering directly with heme or NADH. Fortunately, none of the local minima have E197 pointing towards the pocket, as shown by the disconnectivity graph in Fig. S6.

To determine whether E197 significantly changes the interaction of heme with NADH in the pocket, a small subset of the Dijkstra fastest pathway for NADH approach towards heme was selected for a repeat calculation, with glutamine replacing glutamic acid at position 197. This subset consisted of eight minima separated by seven transition states. The minima were reoptimized with Q197 in place, and attempts were made to connect the two endpoint minima (i.e. reoptimized minima 1 and 8), and the database then refined, all according to standard routines in PATHSAMPLE[54] and OPTIM.[44] This subpath was chosen based on a number of criteria. The barriers between the two endpoint minima are relatively high and NADH undergoes a significant conformational change, suggesting that this is a critical stage of the overall pathway. Furthermore, this path segment occurs when NADH is in close proximity to both heme and the backbone of residue 197. Therefore, we expect any changes to residue 197 at this stage of the pathway would highlight whether a problem exists. Structures for these endpoint minima, along with the largest transition state barriers in the downhill direction, are shown in Fig. S7, from both the E197 and Q197 databases.

Following reoptimization of the two endpoint minima, structural visualization (compare the top left, top right, middle left and middle right images in Fig. S7) revealed that these minima had not changed other than for the change in residue 197 itself. Therefore, a direct comparison between the pathways between these minima for the E197 and Q197 systems could be made.

The E197 and Q197 selected pathways are depicted in Fig. S6. Together they show that the



	E197	Q197	
Ordered Downhill Barriers / kcal mol^{-1}	6.794778315	6.81955362	‡
	4.824416654	4.914031892	*
	1.031331558	2.414484149	
	0.3200266985	0.9721320402	
	0.2463542809	0.3067724323	
	0.0977648813	0.2742731022	
	0.0714600691	0.2010246189	
		0.1654110448	
		0.05314815789	
		0.00259243324	
		0.00028643203	
		7.9052523E-05	
ΔE / kcal mol^{-1}	-4.34193	-4.38985	

Figure S6: E197 *vs* Q197 pathway analysis. Top right: Disconnectivity graph for the WT HemS + Heme + NADH system, with residue 197 assigned as a glutamic acid. Color-coding is according to E197 : CD – Heme : Fe distance. The graph shows that E197 never comes into close proximity to heme (i.e. E197 never points towards the pocket). Top left: Dijkstra fastest pathway for the full approach of NADH towards heme. Middle left: Subset of the Dijkstra pathway chosen for the comparative analysis. This corresponds to the boxed section of the full pathway depicted above it. Bottom left: From the E197 Selected Pathway, the minima were reoptimized with Q197 in place, and these minima were reconnected. This Q197 Selected Pathway is the fastest resulting pathway according to Dijkstra fastest pathway analysis. Bottom right: List of the ordered downhill barriers for the E197 and Q197 Selected Pathways, in addition to the overall energy change, ΔE. Barriers * and ‡ correspond to those highlighted on the E197 and Q197 Selected Pathways.

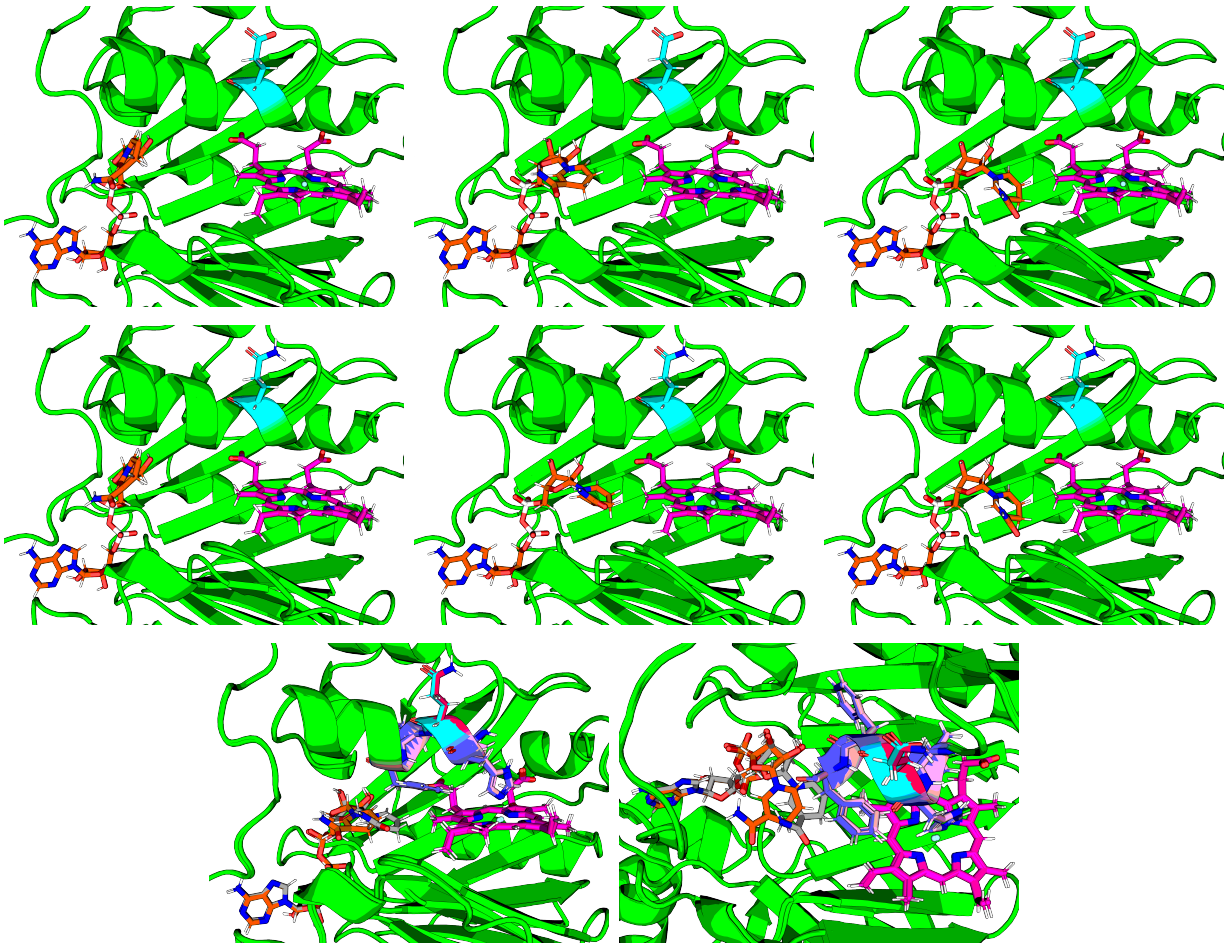


Figure S7: Selected stationary points for E197 *vs* Q197 structural comparison. Heme is shown in magenta, NADH in orange and E197/Q197 in cyan. Top left: Starting minimum for the selected pathway with E197. Middle left: Starting minimum for the selected pathway with Q197. Top central: Maximum transition state (corresponding to ‡ in Fig. S6) with E197. Middle central: Maximum transition state (corresponding to ‡ in Fig. S6) with Q197. Top right: Last minimum for the selected pathway with E197. Middle right: Last minimum for the selected pathway with Q197. Bottom left: Overlay of the ‡ transition states described above. In this case, NADH in the system for Q197 is depicted in gray, Q197 itself in red, and the surrounding residues (V195, H196, F198 and F199) in pink. For the system where E197 is set, V195, H196, F198 and F199 are depicted in blue. Bottom right: Alternative view of the overlaid structures. Residues Q200–N205 have been removed for clearer visualization.

pathway between the two selected minima for the Q197 case is longer than for the E197 case. However, closer inspection reveals that these extra intermediate structures and the transition states connecting them do not make a significant difference to the overall thermodynamics and kinetics of the pathway. The two main transition states separating the endpoint minima (depicted by a ‡ and a * in Fig. S6) in the E197 pathway are replicated in the Q197 pathway, and the barrier sizes change only minimally. The largest of these barriers, ‡, is shown explicitly in Fig. S7. Overlaying these structures reveals that NADH penetrates the pocket marginally further in the Q197 case compared to the E197 case. Apart from this minor difference, the effect of changing E197 to Q197 appears to be minimal: this residue still points away from the pocket, plus its surrounding residues (V195, H196, F198 and F199) almost exactly overlay on one another.

Due to the absence of significant structural or thermodynamic changes upon the E197Q mutation at a critical juncture in the overall pathway, we conclude that having the incorrect residue at this position should not affect our principal conclusions.

K Homologue Properties

Table S5: Accurate masses of the proteins, given in Da, and ϵ_{280} values. Mass spectrometry was used to determine the observed masses, and ExPASY ProtParam[4] used to theoretically predict ϵ_{280} values and expected masses. The mass for ShuS differed by 2 Da from its expected value, which was perhaps due to unexpected protonation behaviour. Care must therefore be taken when interpreting the data involving this homologue.

Protein	**Expected Mass**	**Observed Mass**	**Difference**	ϵ_{280}
HemS	39,360	39,360	0	44,070
HmuS	39,104	39,104	0	44,920
ChuS	38,845	38,845	0	52,940
ShuS	38,831	38,833	+2	52,940

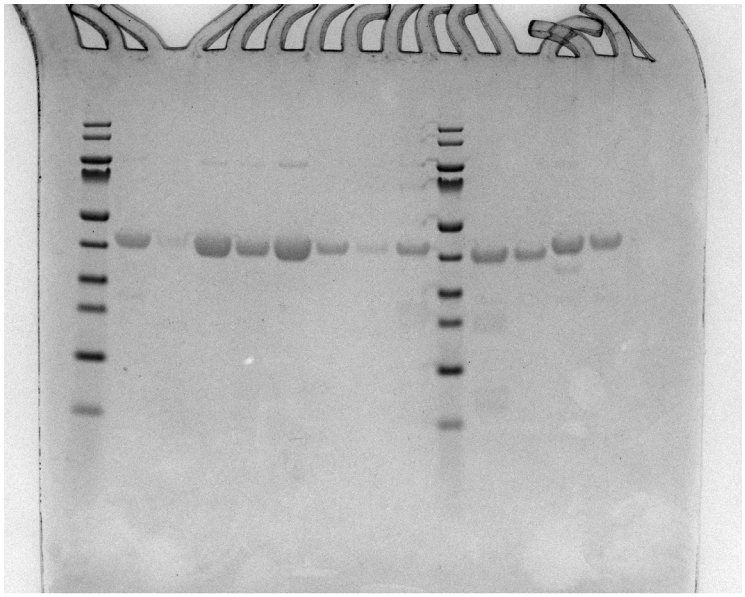


Figure S8: SDS-PAGE gel of the proteins studied in this work and in our associated paper. All have had their His-tags cleaved. Lanes are numbered from left-to-right. Lanes 1 and 10 are markers, using PageRuler Protein Ladder, where the bands range from 10 kDa to 180 kDa. The samples are all pure, and the proteins are all situated close to the fifth band, which corresponds to 40 kDa. Lane 2: WT HemS; 3: F104A HemS; 4: F104AF199A HemS; 5: F104I HemS; 6: F199A HemS; 7: R209A HemS; 8: R209K HemS; 9: Q210A HemS; 11: WT ChuS; 12: WT ShuS; 13: WT HmuS; 14: WT HemS (repeated).

L Heme Structure

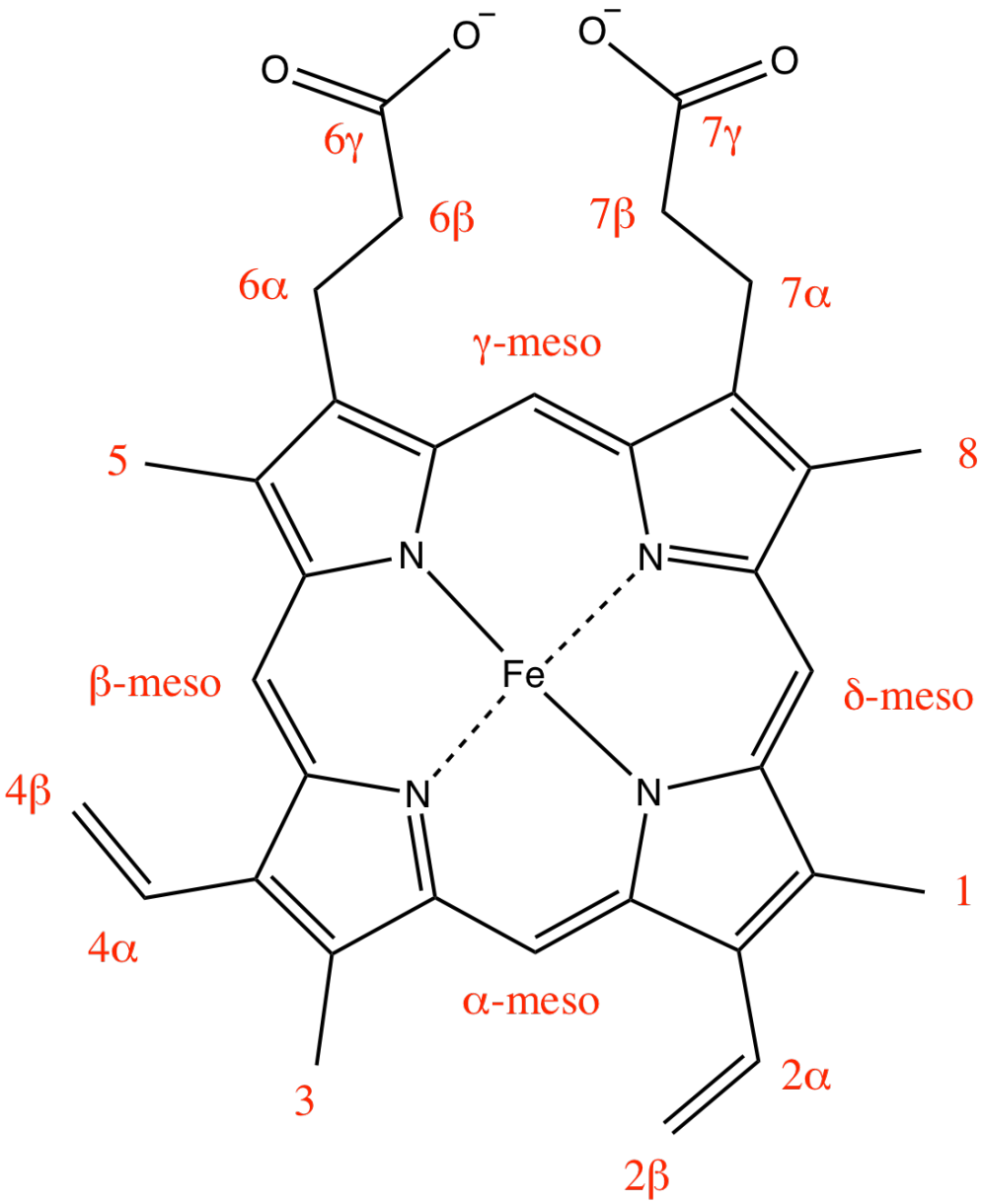


Figure S9: Labeled representation of heme b.

M Experimental Heme-Binding Properties

Table S6: Heme-binding properties of the 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),[79] but appear to be consistent with the other homologues investigated in this study.

	pH 6.5		**pH 8.0**	
	Soret Max.	ϵ_{SM}	**Soret Max.**	ϵ_{SM}
HemS	409.00	103,000	410.25	123,900
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

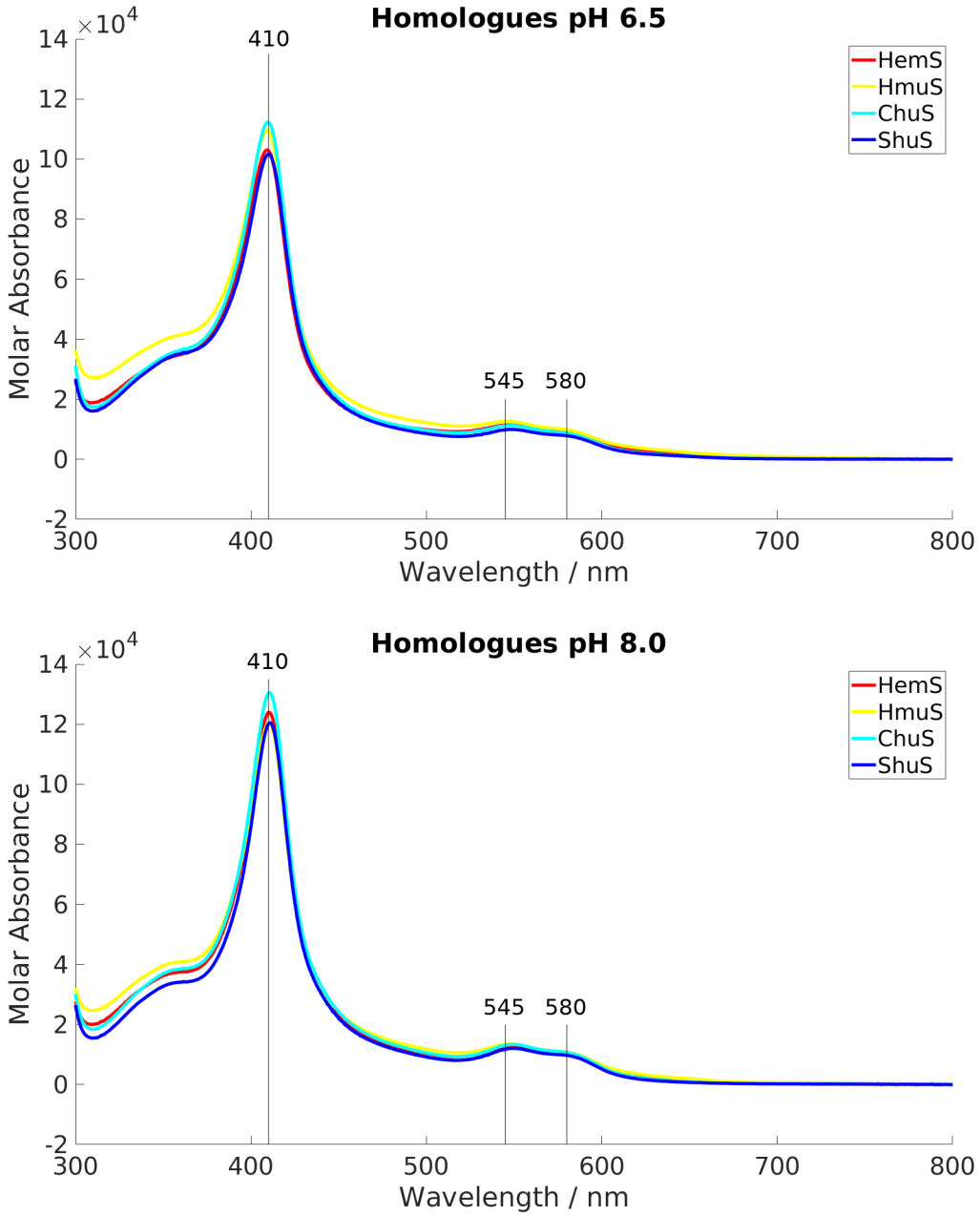


Figure S10: Homologue UV-Visible heme-binding spectra at pH 6.5 and 8.0. All spectra were normalized by setting the absorbance at 800 nm to 0, thus giving a consistent baseline. The absorbance at each Soret peak was then set to the appropriate ϵ_{SM} values given in Table S6, therefore yielding molar absorbances. All of the proteins show similar heme-binding properties. Heme-binding is stronger at pH 8.0 than at pH 6.5.

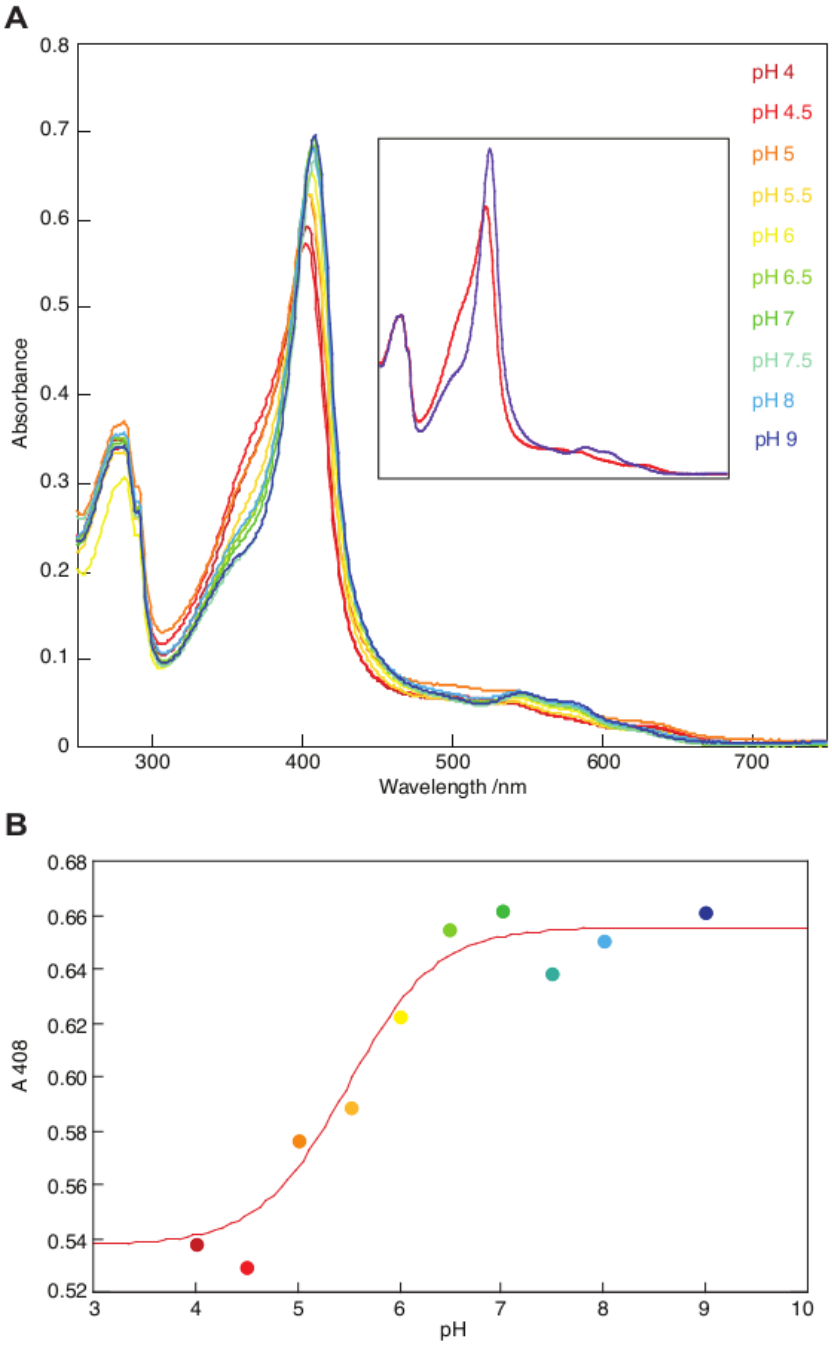


Figure S11: A: HemS UV-Visible heme-binding spectra from pH 4–9. Inset: Two extreme pH values of 4.5 and 9.0. B: Plot of Soret (408 nm) absorbance against pH. The thin red line represents the best fit to an equation describing a single titratable group. The pH at which the Soret intensity maximally changes appears to be halfway between the low pH (water) and high pH (hydroxide) complexes. This feature therefore provides an estimate for the pK_a of the axial heme ligand of ∼5.5.

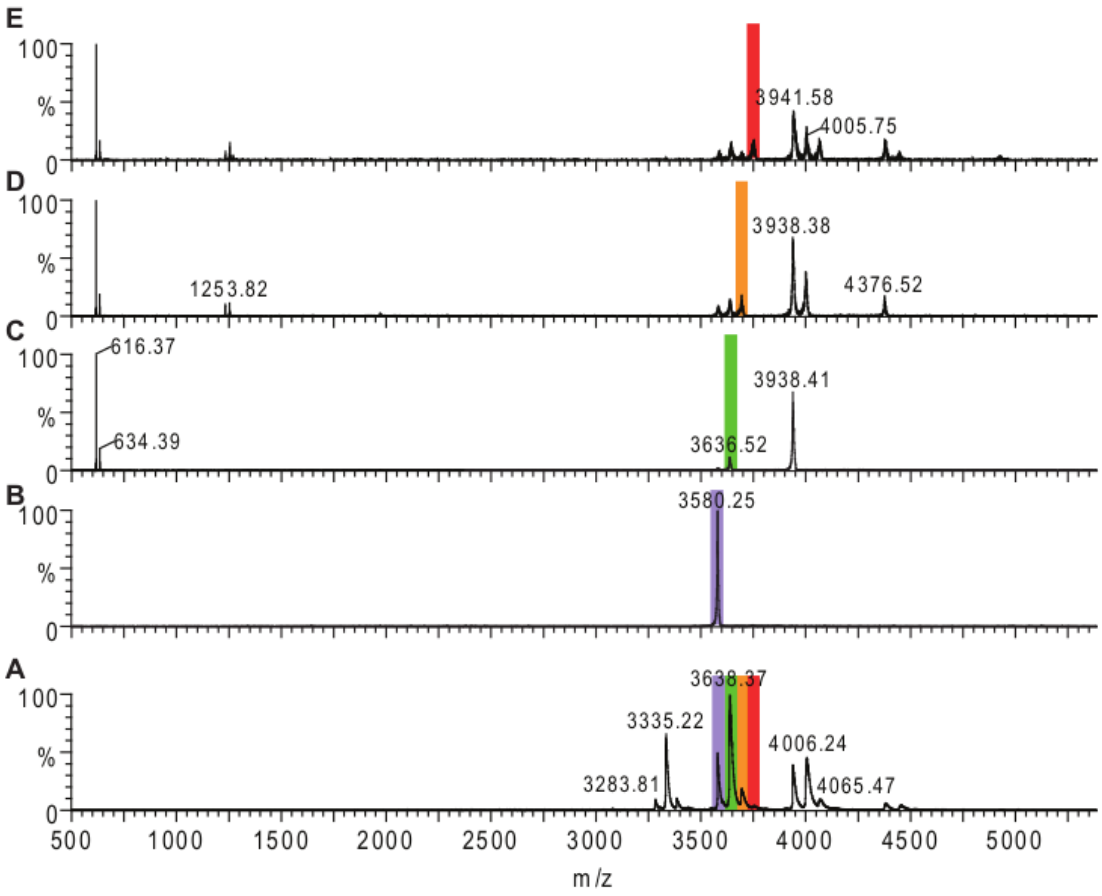


Figure S12: A: Non-denaturing mass spectrometry (ndMS) spectrum detected in a 1:1 mixture of HemS and heme at charge state 11+: purple *apo*; green *holo*-1; orange *holo*-2; red *holo*-3. B–E: MS^2 spectra following parent ion (highlighted) quadrupole selection and subjection to 20 V collision energy. Heme can be dissociated as a cation, giving a m/z 616.4 peak and resulting in the protein being observed at higher m/z, or as a neutral molecule, producing a protein daughter ion of the same charge state but with a lower mass than the parent ion.

N Calculated Heme-Binding Properties

Non-denaturing mass spectrometry (ndMS) of a 1 : 1 HemS : heme mixture had detected traces of HemS with two, or even three, molecules of heme bound. This suggested that heme could bind *via* alternative low-affinity heme-binding sites, or as a dimer at the main binding site.

Considering the case of two heme molecules only, simulations of 1,2-*bis*heme-HemS (where 1 represents the main heme-binding pocket, and 2 represents the alternative heme-binding pocket, as represented in Fig. S13) and 1,1-*di*heme-HemS were compared and contrasted.

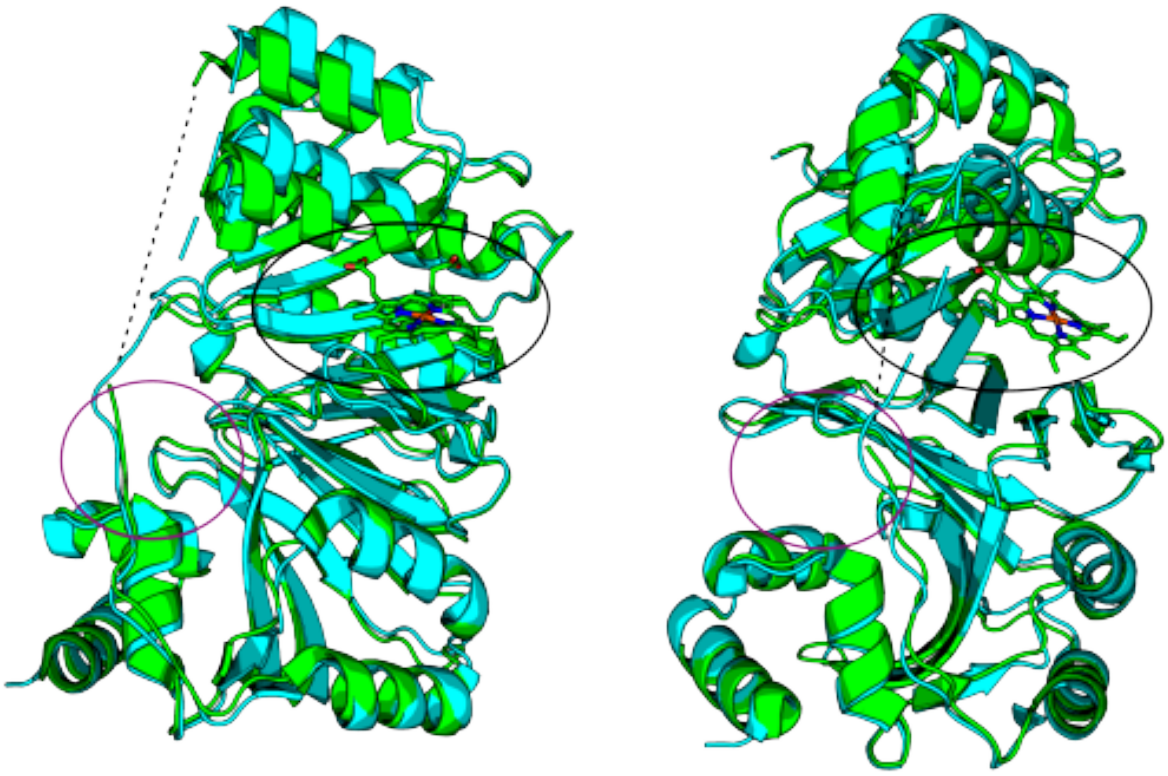


Figure S13: Representations of HemS, showing a likely alternative heme-binding pocket in addition to the main one. *holo*-HemS (PDB 2J0P, green) superimposed on *apo*-HemS (PDB 2J0R, cyan). A black circle shows the large, main cavity, and the purple circle the small, alternative cavity. A dashed line is provided to show the missing loop region from the *holo*-structure. The *apo*- loop is also incomplete.

N.1 1,2-*bis*heme-HemS

A heme molecule was superimposed onto the *holo*-HemS (PDB: 2J0P) structure following suggestions from Relibase^{+}. 36 100-step basin-hopping parallel calculations were then used to refine the protein-ligand binding interactions, with common trends in local recognition emerging. Y20 and H280 flanked opposite faces of the Fe ion to give an extended π-π stacking configuration. It was further shown that the propionate groups of heme are always in polar contacts with R69, R123 and K138 and inconsistently with Q127 and Y140, the former perhaps explained by Q127 being comparatively buried. For the raw data from these calculations, see the PhD thesis of Choy.[2]

The trends from these calculations reveal that this secondary, alternative pocket is certainly capable of binding heme, both *via* iron coordination and propionate recognition. However, the pocket is narrower, and appears to be less flexible, than the main one. These issues, together with the fact that there are less propionate-recognising residues in this pocket, suggest why this pocket binds heme weakly in comparison.

N.2 1,1-*di*heme-HemS

A heme molecule was superimposed onto the *holo*-HemS (PDB: 2J0P) structure, in close proximity to the existing heme molecule. As this addition resulted in steric clashes in a crowded pocket, various residues were manually altered (using the Mutagenesis Wizard in PyMOL)[61] to provide the space. Two starting positions were considered. In both cases the two heme molecules were stacked in slipped-parallel mode. In the first case, the propionate groups of one heme molecule pointed perpendicularly to those in the other heme molecule (such an arrangement inspired by crystal structures of MhuD, a hemoprotein known to bind heme dimers in this manner),[80] whereas in the other case they pointed anti-parallel to one another (inspired by HmuT, which binds stacked heme with this orientation).[81] For each case, 36 100-step basin-hopping parallel calculations were run, and general trends identified. See the PhD thesis of Choy for further details.[2]

These simulations collectively revealed that the main HemS pocket widens to accommodate this extra heme molecule, and that various residues help to anchor this extra heme in place by coordinating to its propionate groups. These interactions are strongest when the heme molecules

are stacked perpendicularly to one another, as in MhuD.

N.3 Conclusion

Therefore, the calculations demonstrate that the extra heme molecules identified by ndMS could be accommodated both as a dimer or through binding to an additional smaller cavity.

O Heme-Binding in a 1 : 8 HemS : Heme Mixture

Table S7: Comparison of HemS-Heme complex expected vs observed ndMS masses, given in Da. 1 : 1 and 1 : 8 protein : heme mixtures are both considered. At the higher heme ratio, it is possible for species up to *holo*-9-HemS to be unambiguously identified. Such high capacity for heme-binding at higher heme ratios suggests that under conditions where heme is excessive, HemS may be acting as a heme storage protein. The presence of *holo*-2-HemS and *holo*-3-HemS species under the more physiologically relevant 1 : 1 mixture of HemS and heme reinforces the validity of this hypothesis.

	Expected Mass	**Observed Mass**	
		1 : 1	**1 : 8**
***apo*-HemS**	39,360	39,373	39,374
***holo*-1-HemS**	39,976	39,989	39,992
***holo*-2-HemS**	40,592	40,603	40,623
***holo*-3-HemS**	41,209	41,239	41,239
***holo*-4-HemS**	41,825		41,879
***holo*-5-HemS**	42,441		42,508
***holo*-6-HemS**	43,057		43,128
***holo*-7-HemS**	43,673		43,744
***holo*-8-HemS**	44,289		44,361
***holo*-9-HemS**	44,906		44,977

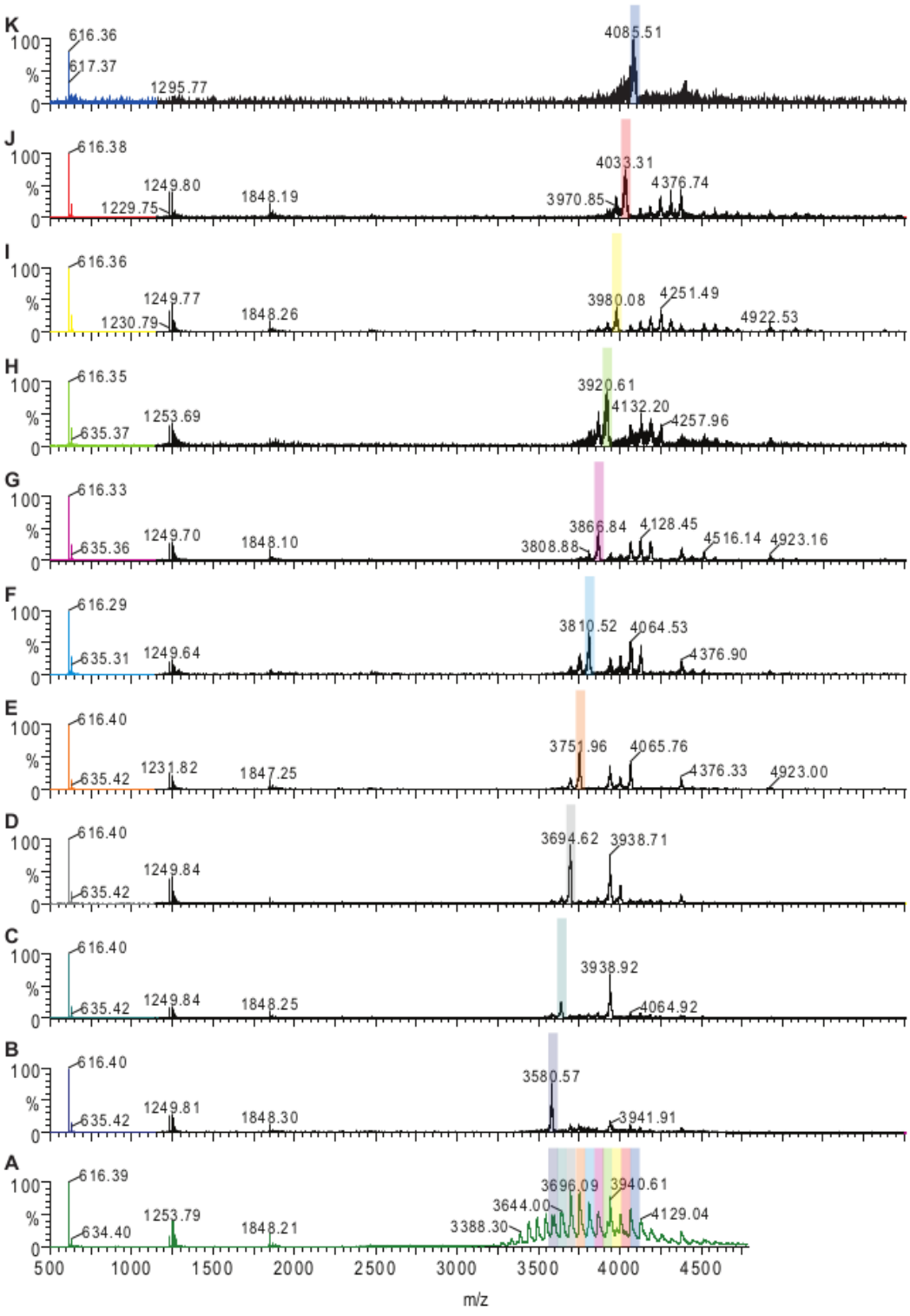


Figure S14: A: Non-denaturing mass spectrometry (ndMS) spectrum detected in a 1:8 mixture of HemS and heme, color-coded according to the MS^2 spectra above. B–K: MS^2 spectra following parent ion (highlighted) quadrupole selection and subjection to 20 V collision energy.

P Small Molecules Used to Probe Disruption to Heme-Binding

Table S8: Effect of various compounds on HemS heme-binding. Compounds were selected to reflect those likely to be encountered by HemS *in vivo*, to probe for heme oxygenase activity, and to introduce a diverse range of functional groups. They were incubated with stoichiometric *apo*-HemS and heme. Of those compounds determined to reduce heme-binding affinity, ATP, galactose, NADH and NADPH all have an adenine group, which is of interest as DNA-binding has been reported in ShuS and PhuS. NADH and NADPH were the only molecules demonstrated to induce activity.

Molecule	**Effect on Heme-Binding**
ATP	Reduction in affinity
Ascorbic acid	No effect
dsDNA (pRSET plasmid)	No effect
Galactose	Reduction in affinity
Glucose	No effect
Glycerol	Reduction in affinity
Imidazole	Reduction in affinity
Maltose	Reduction in affinity
MOPS (buffer)	No effect
NADH	Induces reaction
NADPH	Induces reaction
Phosphate (buffer)	No effect
Ribose	No effect
Sucrose	No effect
Trehalose	Reduction in affinity
Triethanolamine	Reduction in affinity

Q NADH-Induced Heme Breakdown with HemS – Raw Data

The following four figures show the respective dependence of the reaction on heme, HemS and NADH concentrations, and on pH. In each case, spectra were collected in 2 minute intervals, ranging from 2 minutes (red) to 40 minutes (purple). The insets chart the change of absorbance at 591 nm (corresponding to the signature peak of the HBP). These insets are intended to provide approximate trends only, and so for clarity the scale bars have been removed. When not being varied, the concentrations of HemS, heme and NADH were 5 μM, 20 μM and 2000 mM, respectively, and the pH set to 6.5.

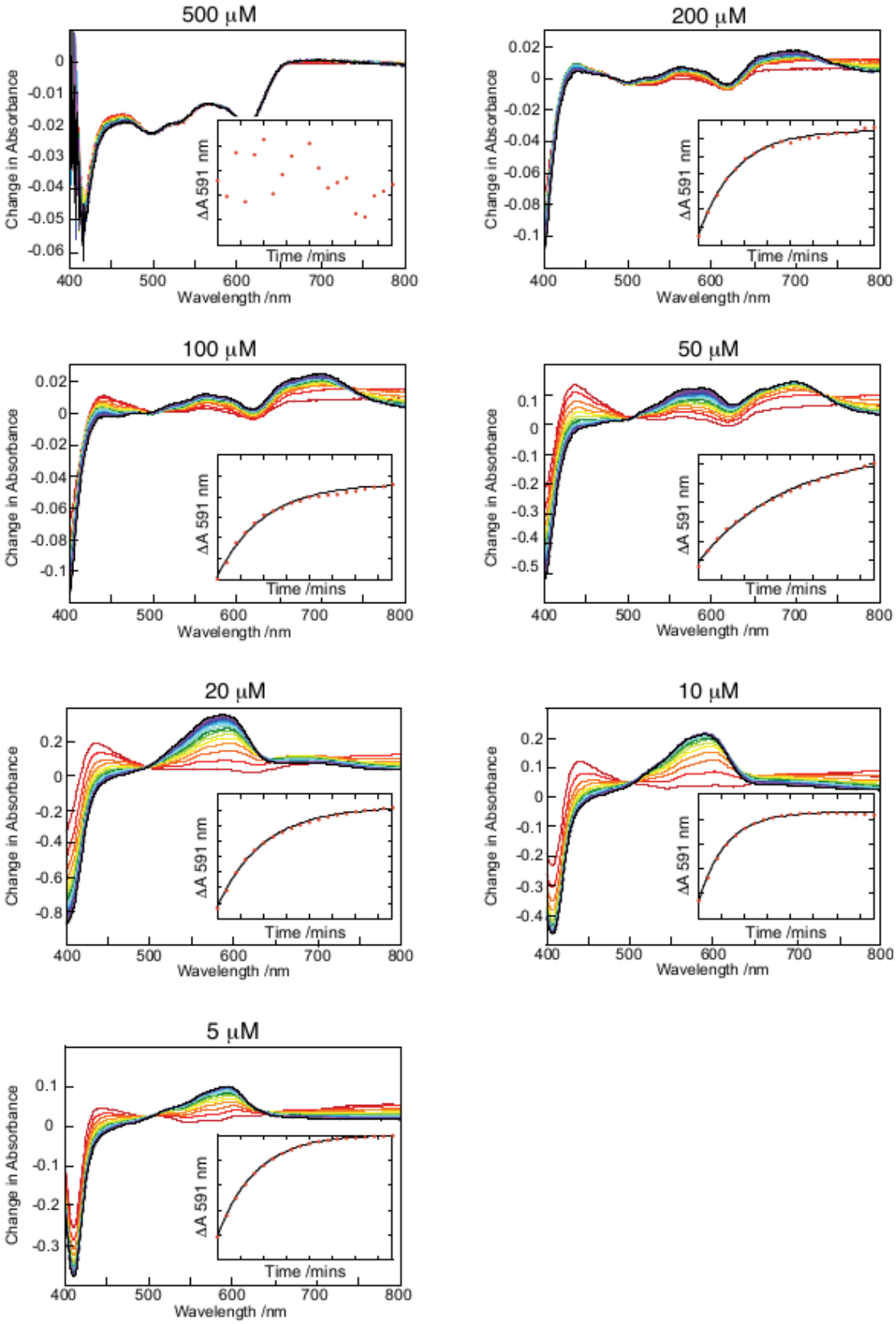


Figure S15: Heme concentration dependence series.

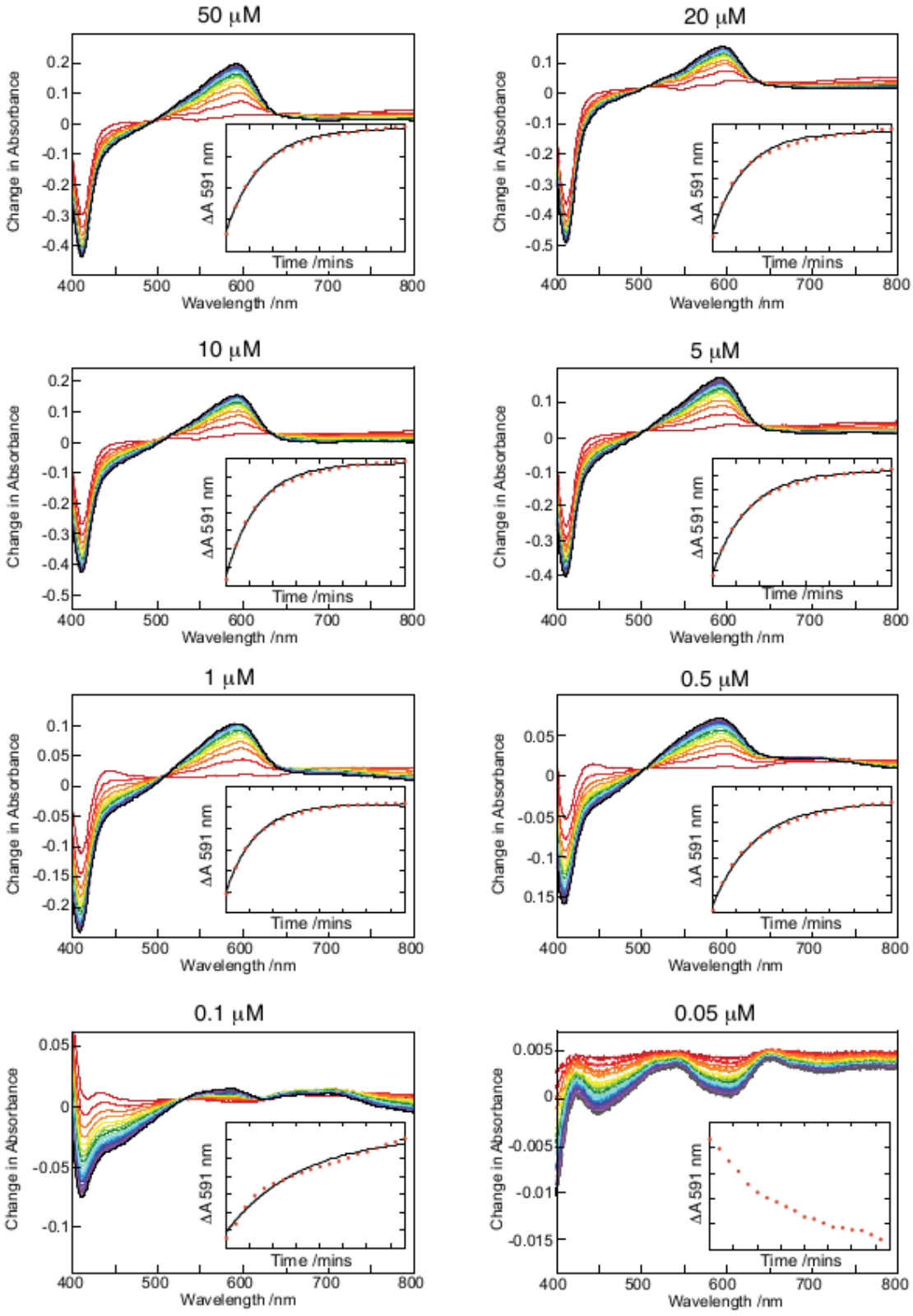


Figure S16: HemS concentration dependence series.

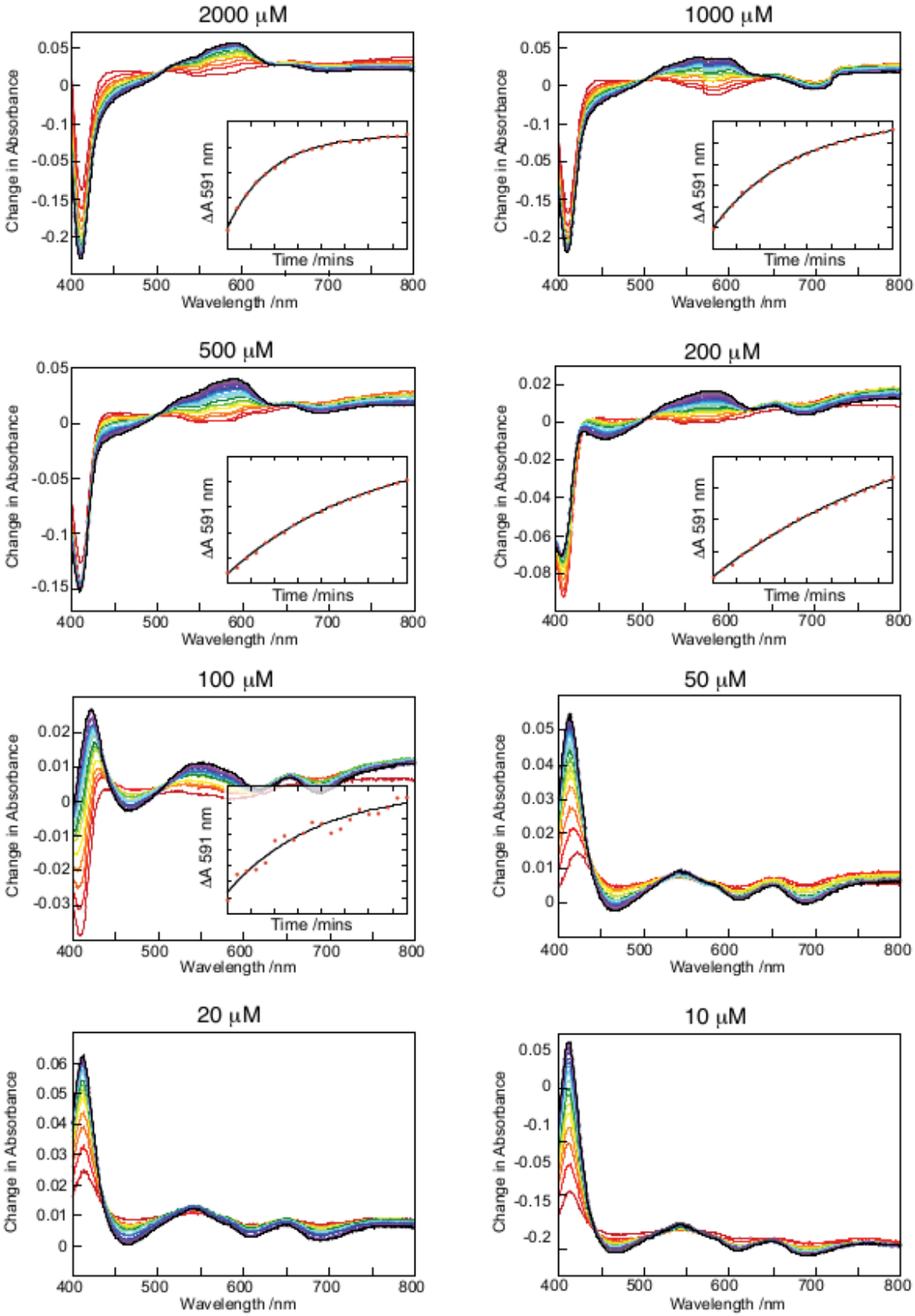


Figure S17: NADH concentration dependence series.

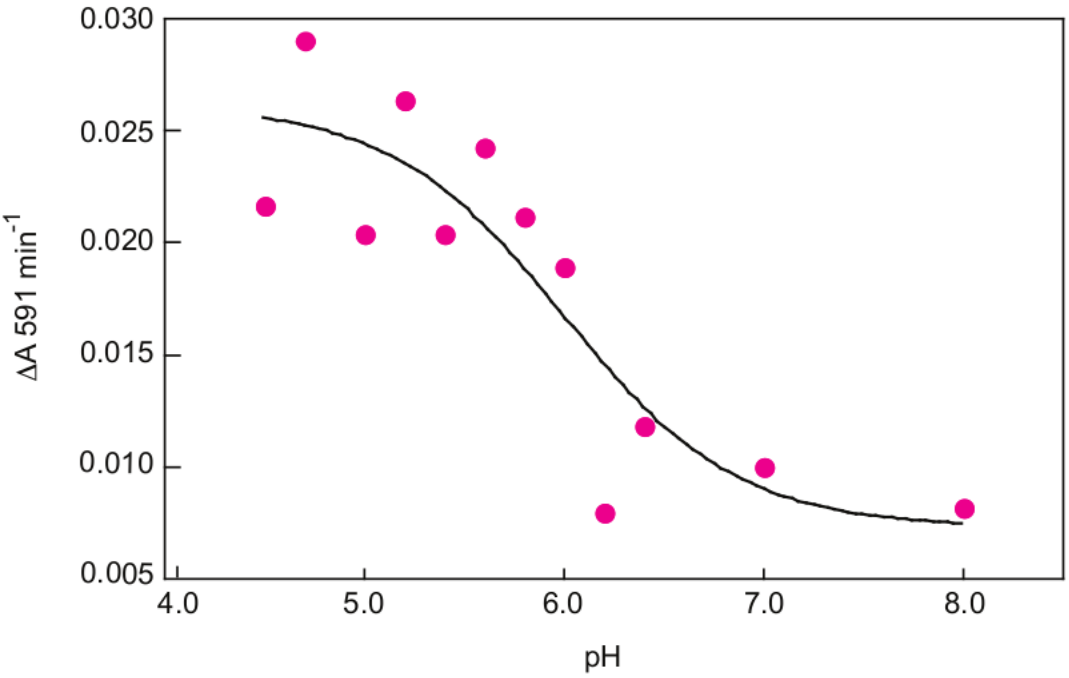


Figure S18: pH dependence series. The temperature was kept constant at 20 °C for all experiments.

R HPLC and MS Comparing Heme, Deuteroheme and Mesoheme – Raw Data

The HBP was separated from the protein/ligand mixture by HPLC for analysis by LCMS. This procedure revealed the characteristic fragmentation pattern for this product. In separate experiments, deuteroheme and mesoheme were used in place of heme, and the respective reaction products also separated by HPLC and analysed by LCMS. The differences in the fragmentation patterns allowed for hypotheses to be reached concerning the structure of the HBP, as discussed in the main paper.

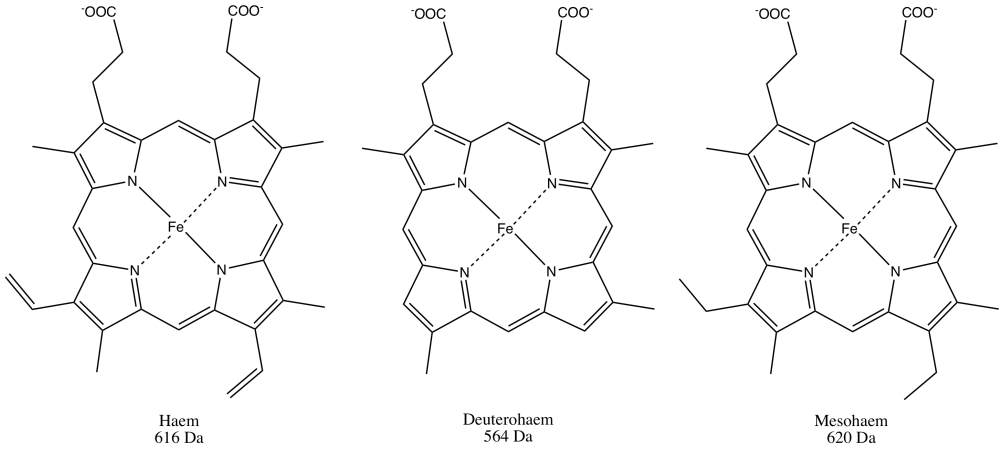


Figure S19: Heme with its deutero- and meso- alternatives.

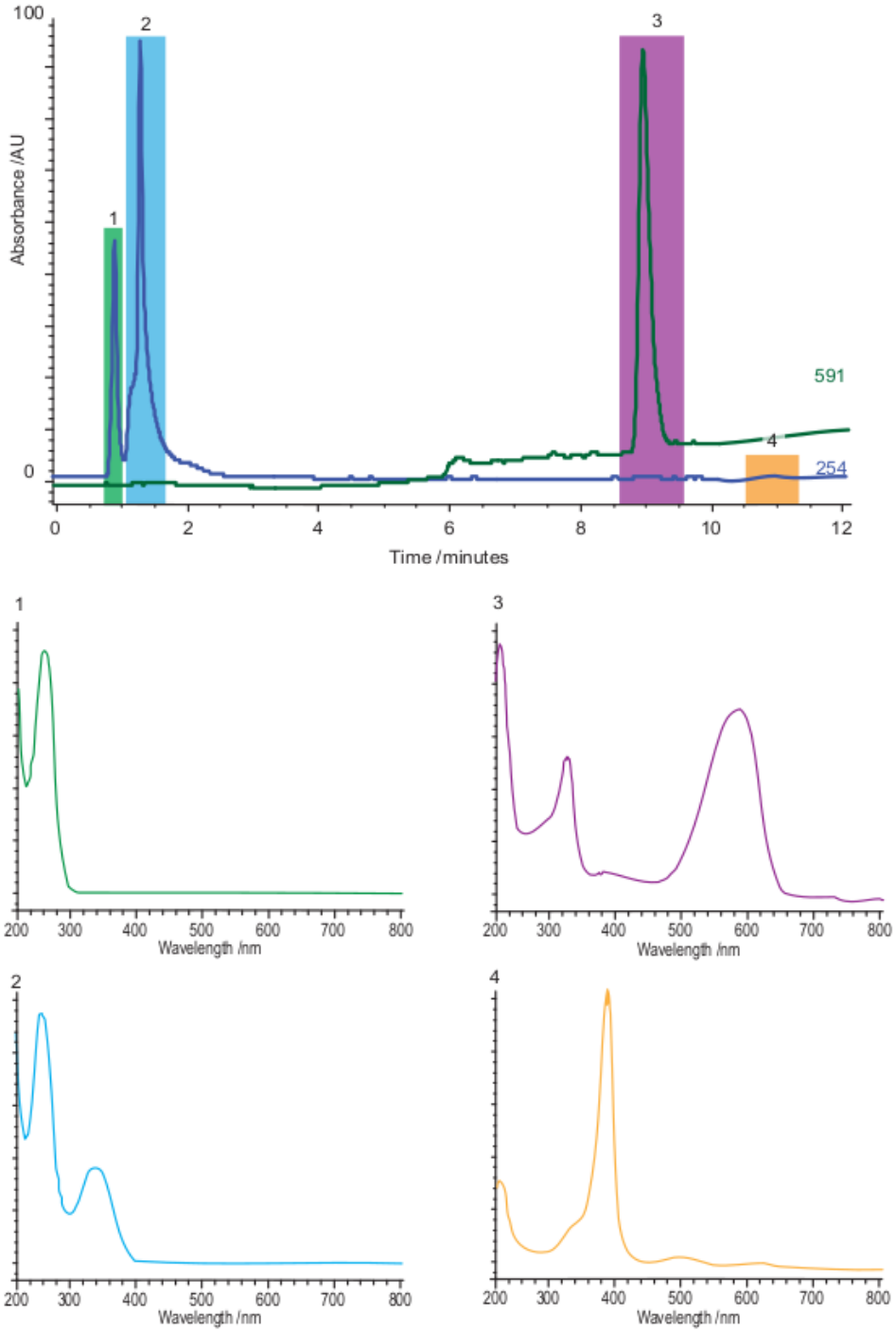


Figure S20: HPLC separation of the HBP. Top: chromatogram charting absorbances at 254 nm and 591 nm, with peaks highlighted. Bottom: Spectra for the main species detected in the chromatogram (indicated by corresponding numbers and colors). Spectra 1, 2, 3 and 4 are characteristic of the HemS protein, NADH/NAD^+, the HBP and unreacted heme, respectively.

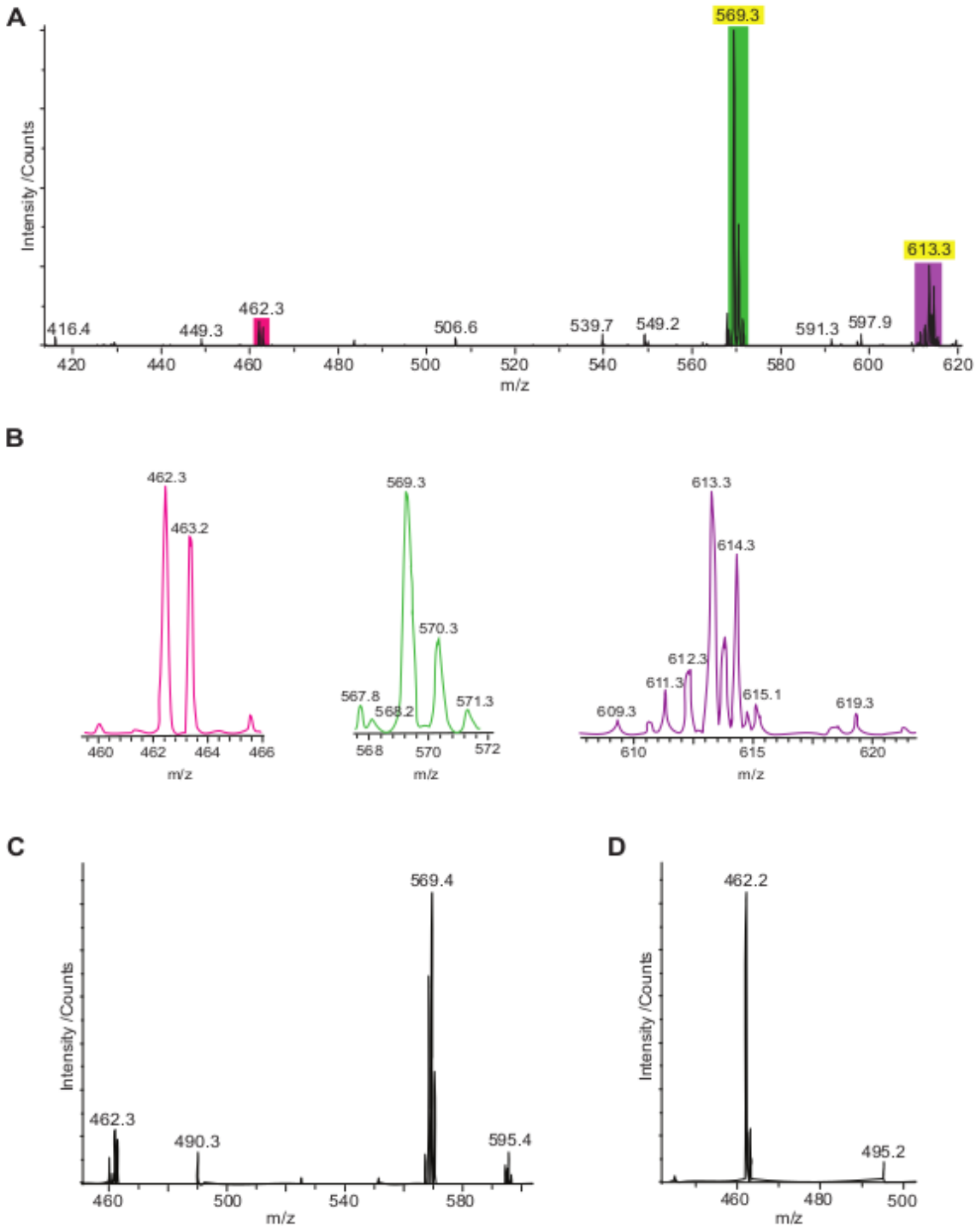


Figure S21: HBP LCMS analysis. A: HBP mass spectrum. The three most abundant ions are highlighted. Yellow highlights indicate ions which were then selected for MS^2 analysis. B: Magnified fragmentation patterns for the three main ions, showing their isotopic patterns. C: MS^2 spectrum of m/z 613.3 ion. D: MS^2 spectrum of m/z 569.3 ion.

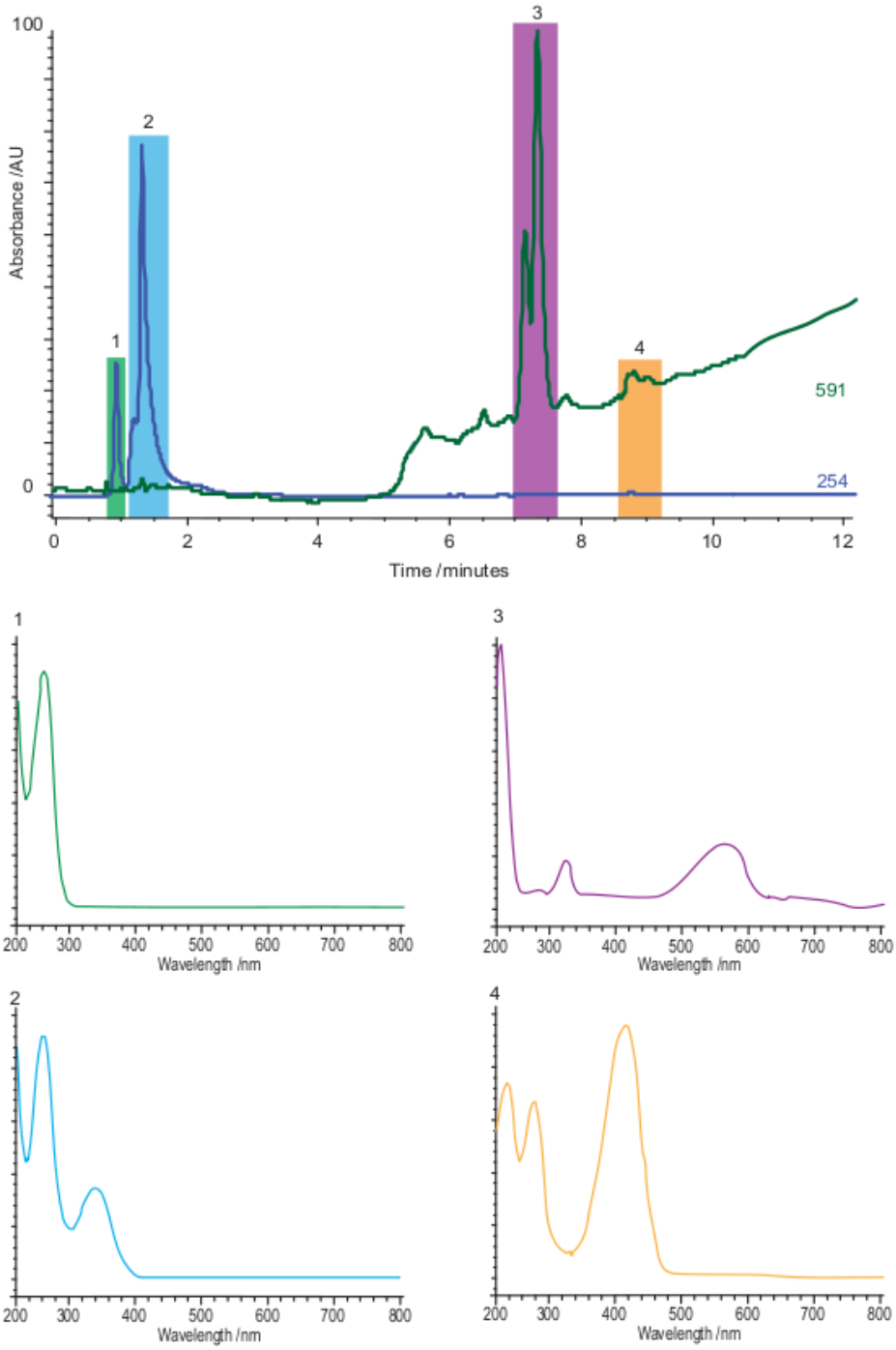


Figure S22: HPLC separation of the deuteroheme breakdown product. Top: chromatogram charting absorbances at 254 nm and 591 nm, with peaks highlighted. Bottom: Spectra for the main species detected in the chromatogram (indicated by corresponding numbers and colors). Spectra 1, 2, 3 and 4 are characteristic of the HemS protein, NADH/NAD^+, the deuteroheme breakdown product and unreacted deuteroheme, respectively.

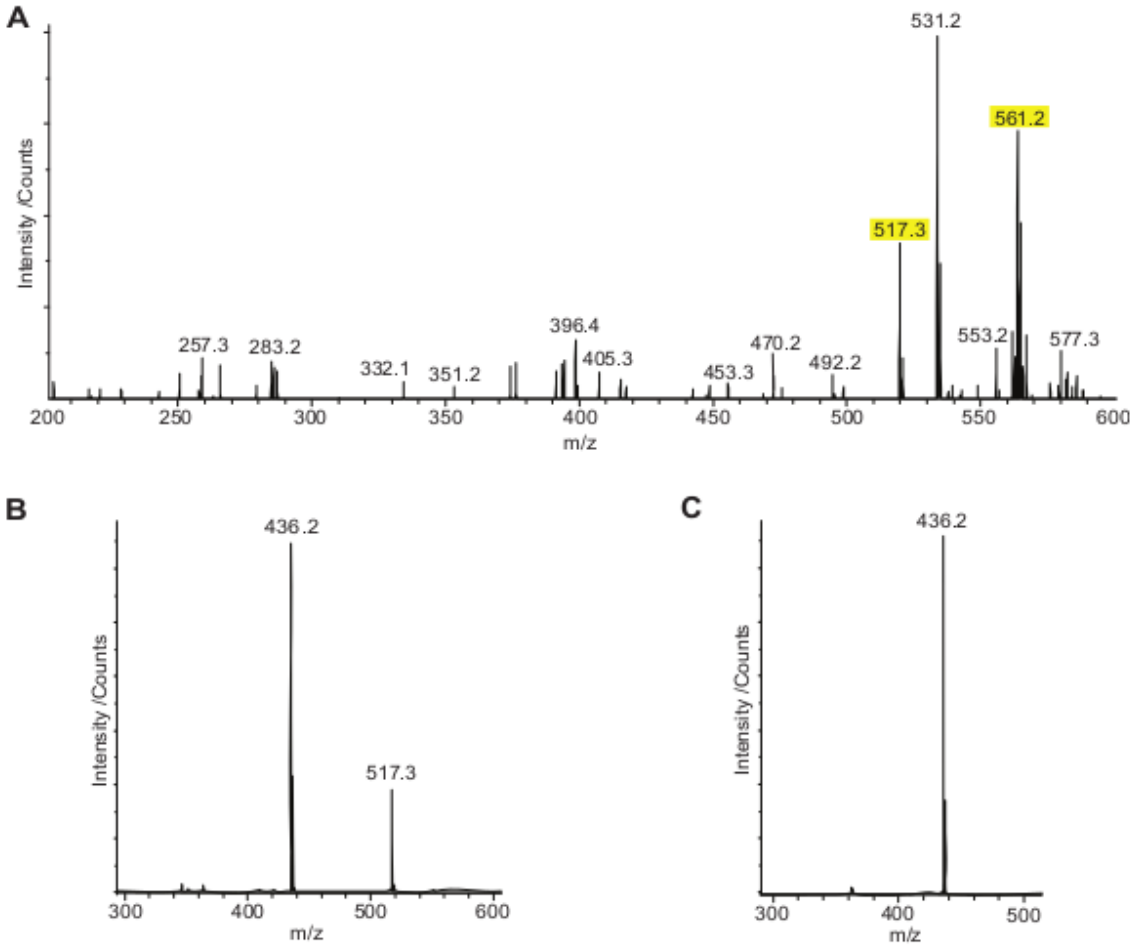


Figure S23: Deuteroheme breakdown product LCMS analysis. A: Deuteroheme breakdown product mass spectrum. Yellow highlights indicate ions which were then selected for MS^2 analysis. B: MS^2 spectrum of m/z 561.2 ion. C: MS^3 spectrum of m/z 517.3 ion.

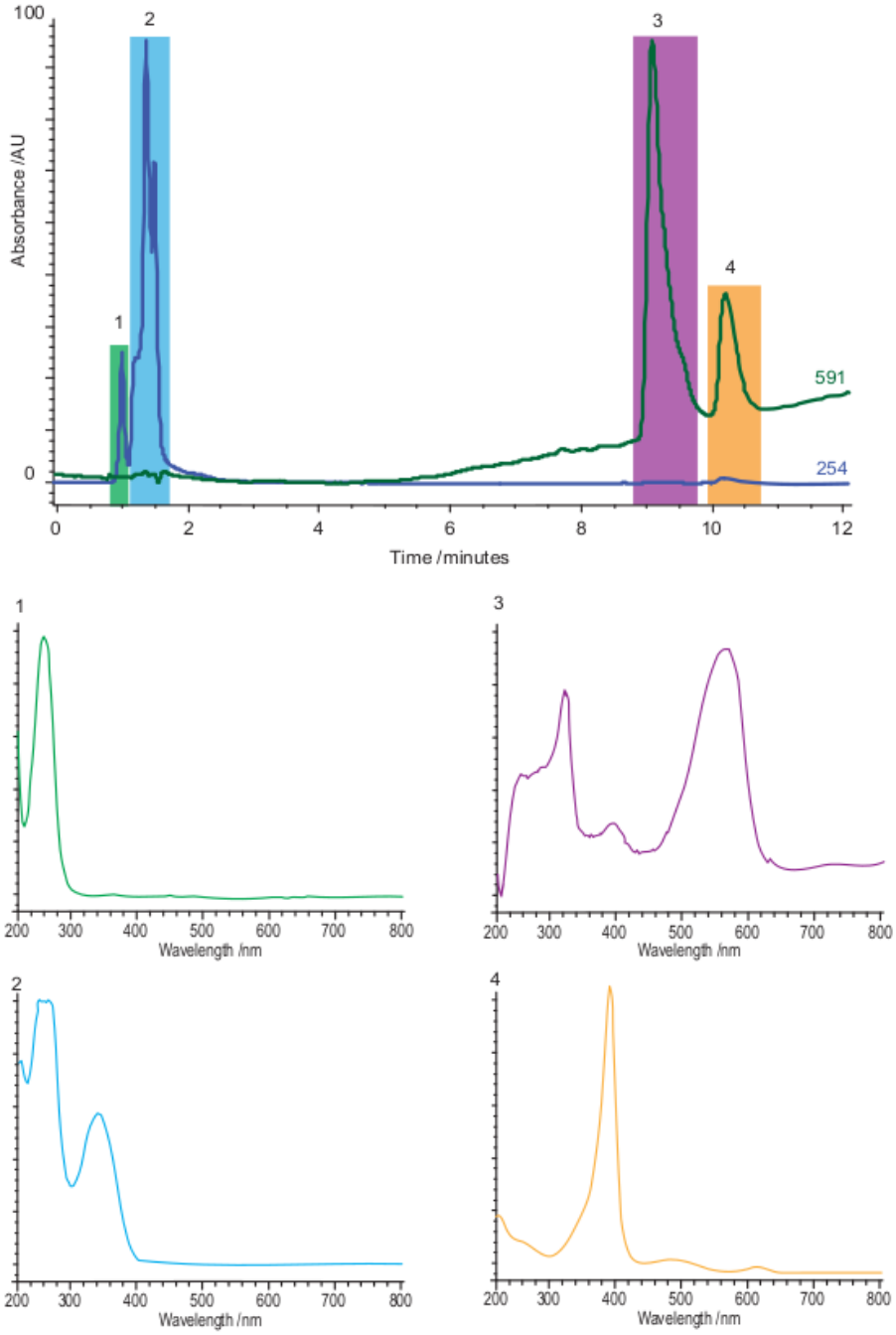


Figure S24: HPLC separation of the mesoheme breakdown product. Top: chromatogram charting absorbances at 254 nm and 591 nm, with peaks highlighted. Bottom: Spectra for the main species detected in the chromatogram (indicated by corresponding numbers and colors). Spectra 1, 2, 3 and 4 are characteristic of the HemS protein, NADH/NAD^+, the mesoheme breakdown product and unreacted mesoheme, respectively.

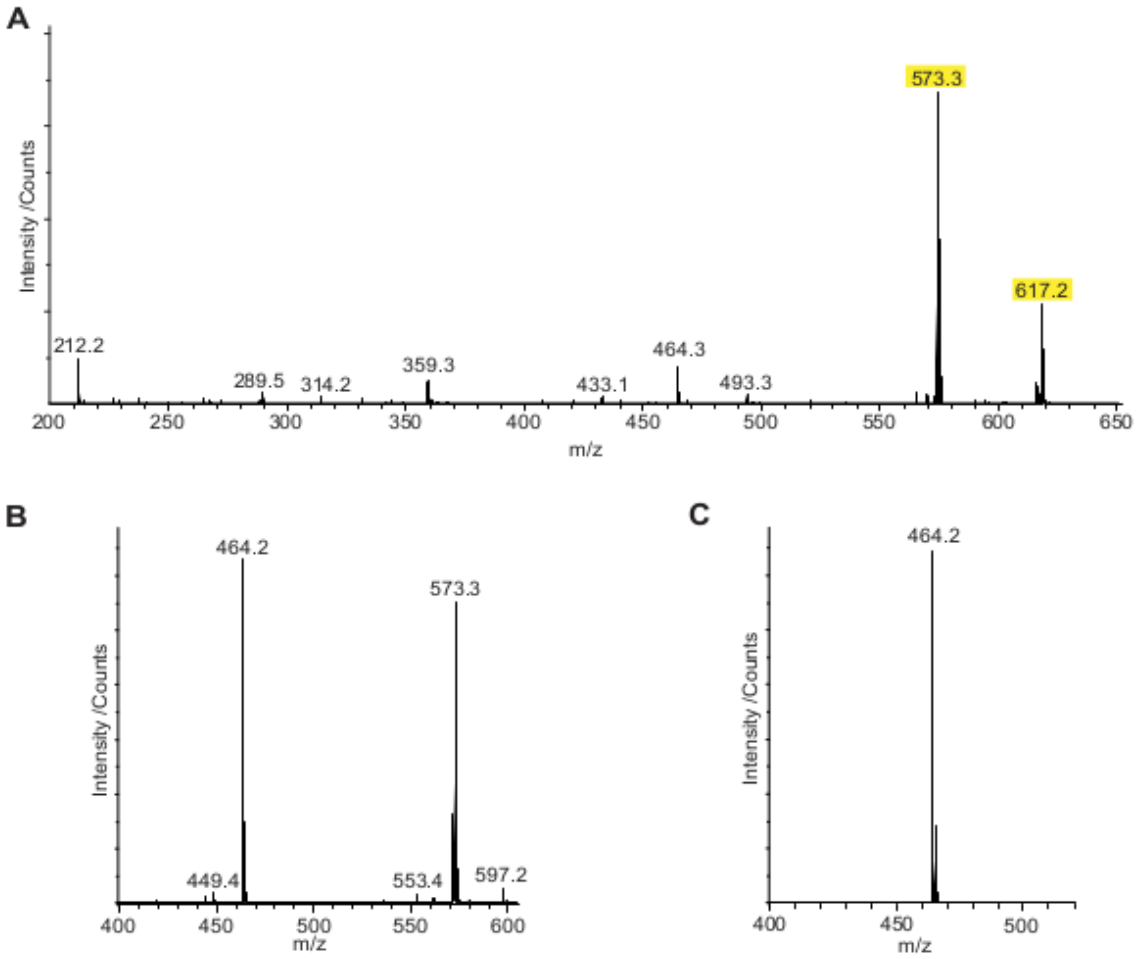


Figure S25: Mesoheme breakdown product LCMS analysis. A: Mesoheme breakdown product mass spectrum. Yellow highlights indicate ions which were then selected for MS^2 analysis. B: MS^3 spectrum of m/z 617.2 ion. C: MS^3 spectrum of m/z 573.3 ion.

S LCMS of Aerobic Reaction

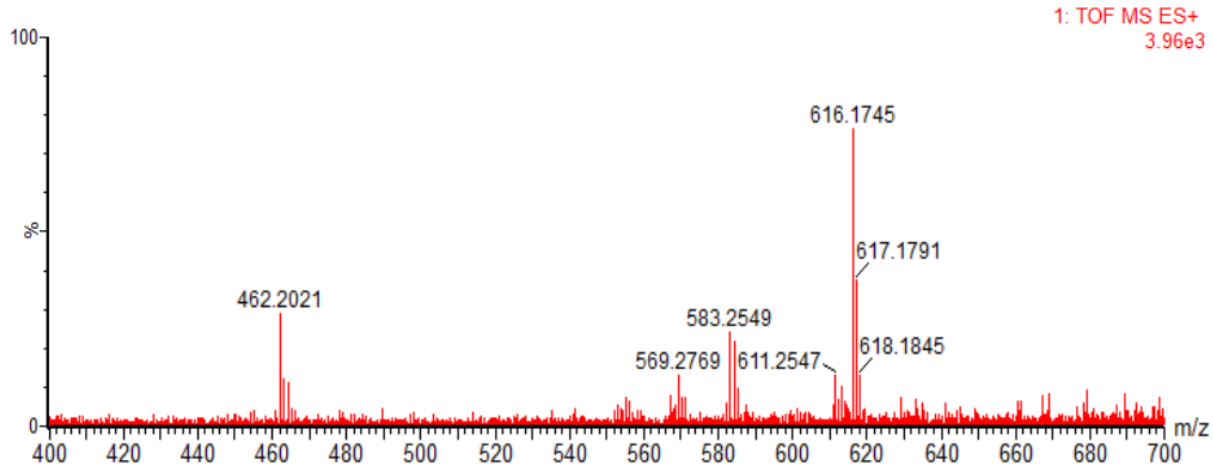


Figure S26: LCMS data for the aerobic reaction between 1 μM HemS, 40 μM heme and 2000 μM NADH. The peak at m/z 616.2 corresponds to unreacted heme and those at 569.3 and 462.2 to the signature peaks of the HBP. The peak at 583.3 corresponds to biliverdin, a by-product from a competing coupled oxidation process.

T NADPH

Preliminary experimental (UV-Visible spectroscopy) data suggested that NADH and NADPH can both be used as the ligand to initiate heme breakdown within HemS. A full investigation to determine whether one or the other, or both, is the correct natural ligand was not pursued. However, computational calculations were used to determine how NADPH could also fit inside the cavity, and some hypotheses were formulated suggesting that this ligand may indeed be the natural one.

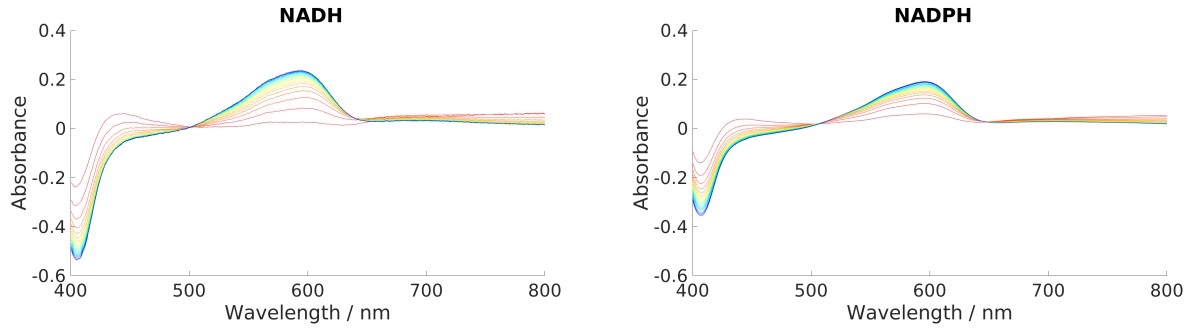


Figure S27: UV-Visible difference spectra comparing NADH vs NADPH. Reaction stoichiometries were 5 μM HemS : 20 μM heme : 2000 μM NAD(P)H, and pH set to 6.5. Scans were run every minute for 20 minutes, reflected by the color scheme change from red (1 minute) to blue/purple (20 minutes). Breakdown of heme appears marginally slower in the NADPH case. This is most likely due to a lack of proper equilibration between HemS and heme before NADPH was added to initiate the reaction. These results should therefore not be taken quantitatively. They do, however, indicate that reaction with NADPH produces the same product as with NADH.

A selection of stationary points from the *holo*-HemS database were taken, and NADH 'mutated' to NADPH using **CHECKSPMUTATE**, and these points then reoptimised. These calculations showed that the extra phosphate group possessed by NADPH tended to point away from the cavity, and so is therefore unlikely to disrupt docking, thus showing why this protein, unusually for a biomolecule, can utilize either NADH or NADPH to effect its function. Closer inspection of this extra phosphate group revealed that it is capable of forming a stabilising hydrogen bond with the hydroxyl group of the T312 residue, suggesting that NADPH may be the natural ligand after all, whilst the protein is able to utilize NADH as well. Though not tested, it is assumed that NADPH would be a suitable (and perhaps the natural) ligand for the family of homologues related to HemS too.

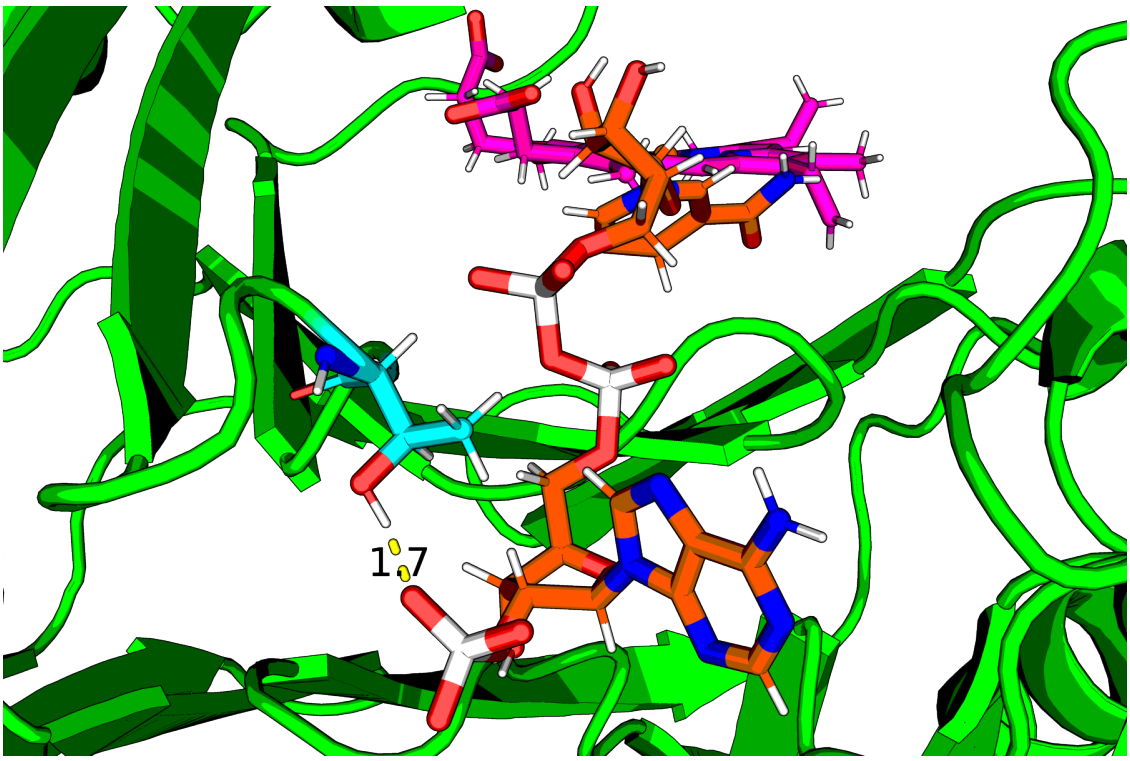


Figure S28: Magnified image of the cavity from a selected minimum structure generated computationally, showing NADPH extended along the HemS pocket. The extra phosphate group of NADPH is shown hydrogen-bonding to the T312 residue of HemS, suggesting that NADPH is more suited to docking inside the pocket than NADH. This would indicate that, though capable of utilizing NADH or NADPH to effect heme breakdown, the natural ligand of HemS is NADPH. Here, the protein backbone is represented in green, whilst its T312 residue is depicted explicitly in cyan. Heme is represented in magenta, and NADH in orange. The distance, 1.7, is in Angstroms.

U X-ray Crystallography Structure of the HBP in HemS

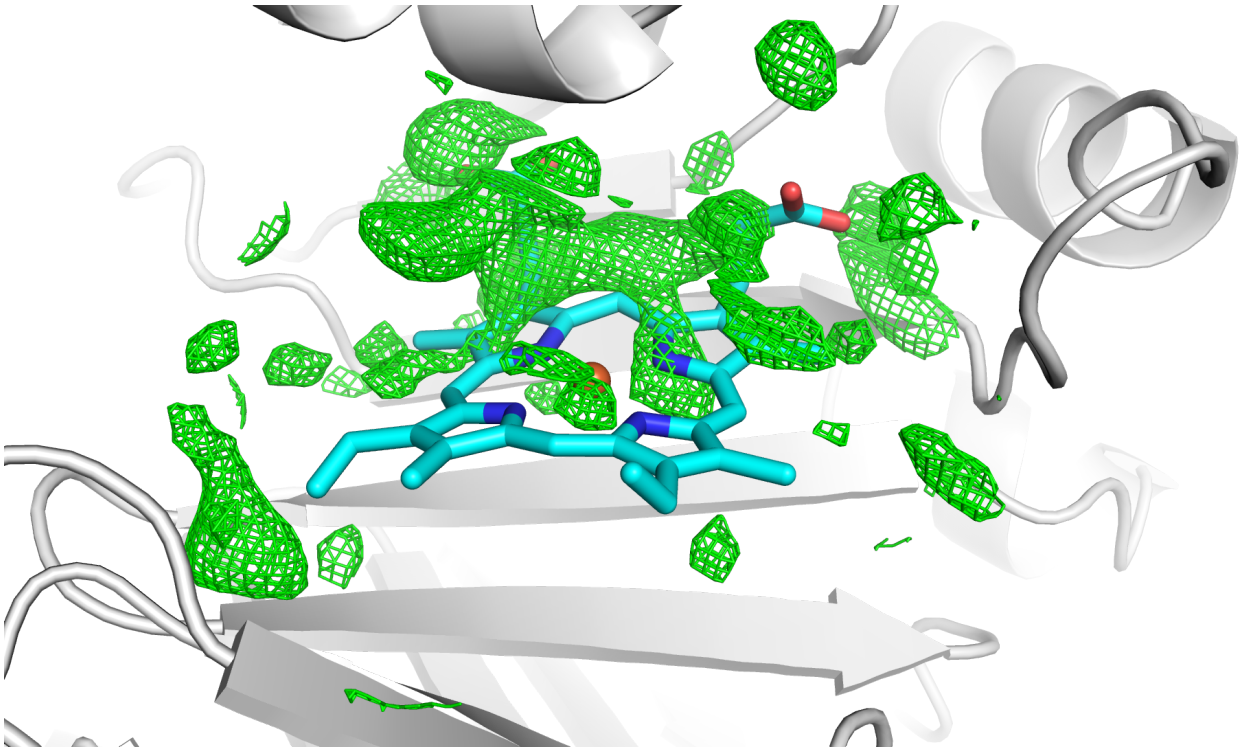


Figure S29: Electron density from a crystal of HemS after heme breakdown. The green mesh corresponds to unassigned electron density. Accurate refinement and assignment of atoms was not possible due to the partial disintegration of the HBP. The heme molecule, in cyan, is for reference only, having been overlaid on the original density from a crystal of *holo*-HemS. The shape of the unassigned electron density suggests cleavage of the tetrapyrrole. That iron has been extracted is indicated by an absence of any region of large electron density.

V NADD NMR

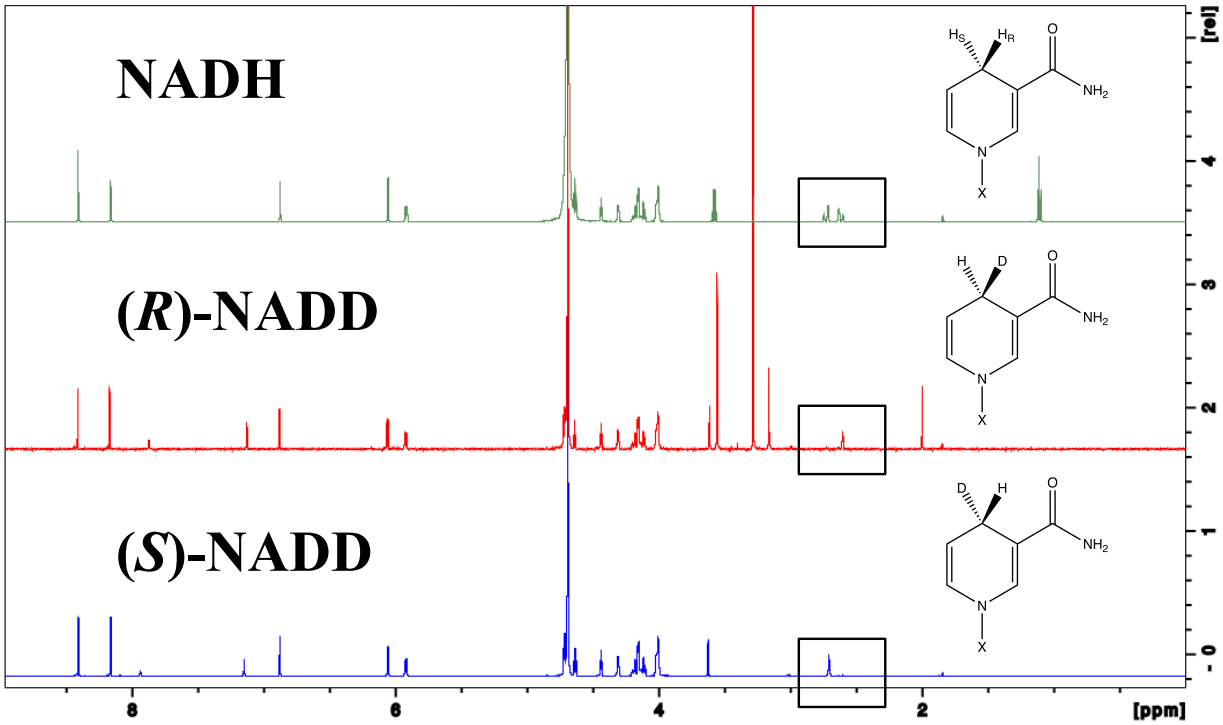


Figure S30: Overlay of NMR spectra taken for NADH, (*R*)-NADD and (*S*)-NADD. Black boxes highlight the region corresponding to the hydride signals. The X label corresponds to the rest of the NADH molecule.

W DNA-Binding in ChuS and ShuS

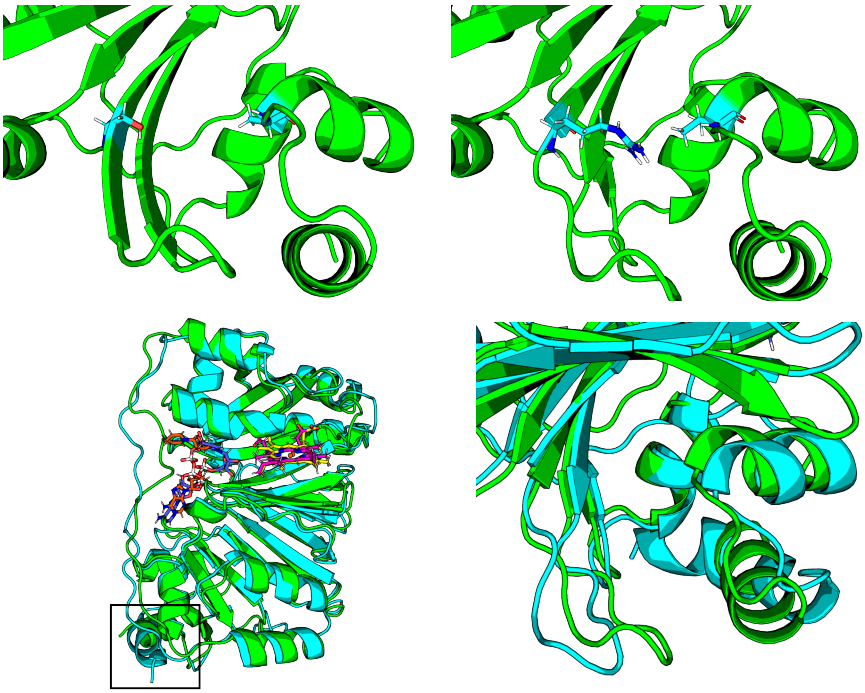


Figure S31: Top: Possible DNA-binding site in ChuS and ShuS. Left: Small cavity of ChuS, with residues A19 and S123 highlighted in cyan. Right: Small cavity of ShuS, with residues V19 and R123 highlighted in cyan. Both highlighted residues in ShuS are larger than their equivalents in ChuS. It appears that V19 anchors R123 in position, in a suitable position to bind to DNA. Bottom: Overlay of starting (green) and lowest energy (cyan) structural minima from the ShuS database. Left: The heme and NADH conformations forming part of the starting minimum are color-coded in magenta and orange, respectively. For the lowest energy minimum, heme and NADH are color-coded in yellow and dark blue, respectively. In the lowest energy 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 starting and lowest energy 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. Movement of the N-terminal α-helix in the transition from the starting minimum to the lowest energy minimum results in a significant enlargement of the entrance to the small cavity.

References

(1) Sawyer, E. B. Biophysical analysis of haem-protein interactions in bacterial haem transfer systems. PhD Thesis, University of Cambridge, 2009.

(2) Choy, D. C. Y. Haemoproteins and the study of protein-ligand interactions. PhD Thesis, University of Cambridge, 2015.

(3) Keith, A. D. Anaerobic, NADH-dependent haem breakdown in a family of haemoproteins: an experimental and computational study. Ph.D. thesis, University of Cambridge, 2021.

(4) 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.

(5) Gregory, J. Investigating the NADH dependent reaction of the *Yersinia enterocolitica* haem chaperone HemS. Master's Thesis, University of Cambridge, 2012.

(6) 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.

(7) Xie, Y. Predicting function from structure: haem degradation and DNA binding in the HemS family. Master's Thesis, University of Cambridge, 2021.

(8) 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.

(9) Schneider, S.; Paoli, M. Haem-binding properties and crystallisation of the bacterial protein HemS. *Acta Cryst. A* **2005**, *61*, 343.

(10) NEB Protocol for Restriction Endonuclease Reaction. 2012; `https://international.neb.com/protocols/2012/12/07/optimizing-restriction-endonuclease-reactions`.

(11) 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.

(12) 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.

(13) Northrop, D. B.; Duggleby, R. G. Preparation and storage of isotopically labeled reduced nicotinamide adenine dinucleotide. *Anal. Biochem.* **1987**, *165*, 362–364.

(14) 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.

(15) 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.

(16) 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.

(17) Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Cryst. D* **2011**, *67*, 235–242.

(18) 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.

(19) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Cryst. D* **2010**, *66*, 486–501.

(20) Hendlich, M. Databases for protein-ligand complexes. *Acta Cryst. D* **1998**, *54*, 1178–1182.

(21) 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.

(22) Mantell, R. G.; Pitt, C. E.; Wales, D. J. GPU-Accelerated exploration of biomolecular energy landscapes. *J. Chem. Theory Comput.* **2016**, *12*, 6182–6191.

(23) Guex, N.; Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **1997**, *18*, 2714–2723.

(24) Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(25) 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.

(26) Bryce, R. Bryce Group - AMBER Parameter Database, University of Manchester. 2021; `http://amber.manchester.ac.uk/`.

(27) Giammona, D. A. An examination of conformational flexibility in porphyrins and bulky-ligand binding in myoglobin. PhD Thesis, University of California, Davis, 1984.

(28) 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.

(29) 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.

(30) 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.

(31) 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.

(32) Case, D. A. et al. **AMBER12**, University of California, San Francisco. 2012.

(33) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born model suitable for macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(34) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor. Chem. Acc.* **1999**, *101*, 426–434.

(35) 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.

(36) Li, Z.; Scheraga, H. A. Structure and free energy of complex thermodynamic systems. *J. Mol. Struct.* **1988**, *179*, 333–352.

(37) 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.

(38) Doye, J. P. K.; Wales, D. J. Thermodynamics of global optimization. *Phys. Rev. Lett.* **1998**, *80*, 1357–1360.

(39) Doye, J. P. K.; Wales, D. J.; Miller, M. A. Thermodynamics and the global optimization of Lennard-Jones clusters. *J. Chem. Phys.* **1998**, *109*, 8143–8153.

(40) Wales, D. J. GMIN: A program for basin-hopping global optimisation. 2021; `http://www-wales.ch.cam.ac.uk/GMIN`.

(41) Wales, D. J. Discrete path sampling. *Mol. Phys.* **2002**, *100*, 3285–3305.

(42) Wales, D. J. *Energy Landscapes*; Cambridge University Press: Cambridge, 2003.

(43) Wales, D. J. Some further applications of discrete path sampling to cluster isomerization. *Mol. Phys.* **2004**, *102*, 891–908.

(44) Wales, D. J. OPTIM: A program for optimising geometries and calculating pathways. 2021; `http://www-wales.ch.cam.ac.uk/OPTIM`.

(45) Trygubenko, S. A.; Wales, D. J. A doubly nudged elastic band method for finding transition states. *J. Chem. Phys.* **2004**, *120*, 2082–2094.

(46) 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.

(47) Cerjan, C. J.; Miller, W. H. On finding transition states. *J. Chem. Phys.* **1981**, *75*, 2800–2806.

(48) Pancíř, J. Calculation of the least energy path on the energy hypersurface. *Coll. Czech Chem. Commun.* **1975**, *40*, 1112–1118.

(49) Munro, L. J.; Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B* **1999**, *59*, 3969–3980.

(50) Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Prog.* **1989**, *45*, 503–528.

(51) Nocedal, J.; Wright, S. J. *Numerical Optimization* in *Physical Review*; Springer-Verlag: Berlin, Germany, 2006.

(52) Dijkstra, E. A note on two problems in connexion with graphs. *Numer. Math.* **1959**, *1*, 269–271.

(53) Carr, J. M.; Trygubenko, S. A.; Wales, D. J. Finding pathways between distant local minima. *J. Chem. Phys.* **2005**, *122*, 234903–234909.

(54) 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`.

(55) 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.

(56) Joseph, J. A.; Röder, K.; Chakraborty, D.; Mantell, R. G.; Wales, D. J. Exploring biomolecular energy landscapes. *Chem. Commun.* **2017**, *53*, 6974–6988.

(57) Röder, K.; Wales, D. J. Energy landscapes for the aggregation of Aβ_{17-42}. *J. Am. Chem. Soc.* **2018**, *140*, 4018–4027.

(58) Evans, D. A.; Wales, D. J. Folding of the GB1 hairpin peptide from discrete path sampling. *J. Chem. Phys.* **2004**, *121*, 1080–1090.

(59) 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.

(60) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758–760.

(61) Schrödinger, L. L. C.; DeLano, W. **PYMOL**. 2020; `http://www.pymol.org/pymol`.

(62) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

(63) 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.

(64) 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.

(65) Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, 256–259.

(66) 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.

(67) 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.

(68) 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.

(69) 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.

(70) Unno, M.; Matsui, T.; Ikeda-Saito, M. Structure and catalytic mechanism of heme oxygenase. *Nat. Prod. Rep.* **2007**, *24*, 553–570.

(71) 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.

(72) 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*. *Microbiology* **2010**, *156*, 1873–1882.

(73) 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.

(74) 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.

(75) 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.

(76) Pedersen, B. The dnaseq package. 2002.

(77) SnapGene software (from Insightful Science). 2021; `https://www.snapgene.com/`.

(78) Huang, B. Metapocket: a meta approach to improve protein ligand binding site prediction. *OMICS J. Integr. Biol.* **2009**, *13*, 325–330.

(79) Wilks, A. The ShuS protein of *Shigella dysenteriae* is a heme-sequestering protein that also binds DNA. *Arch. Biochem. Biophys.* **2001**, *387*, 137–142.

(80) Chim, N.; Iniguez, A.; Nguyen, T. Q.; Goulding, C. W. Unusual diheme conformation of the heme-degrading protein from *Mycobacterium tuberculosis*. *J. Mol. Biol.* **2010**, *395*, 595–608.

(81) 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.