



**This electronic thesis or dissertation has been
downloaded from Explore Bristol Research,
<http://research-information.bristol.ac.uk>**

Author:

Marson, Noa A

Title:

**Exploration into the use of bioinformatics and genetically encoded sensors to
understand the binding and dynamics of heme in the cell**

General rights

Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode> This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy

Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

- Your contact details
- Bibliographic details for the item, including a URL
- An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

Exploration into the Use of Bioinformatics and Genetically Encoded Sensors to Understand the Binding and Dynamics of Heme in the Cell

Noa Marson

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Chemistry MSc by Research in the School of Chemistry, Faculty of Science.

Supervisor: Prof. Emma Raven

September 2022

Word count: 15,472

Abstract

Heme, or iron protoporphyrin IX, has been traditionally understood to bind exclusively as a cofactor to proteins. However, regulatory roles have since been uncovered, for example, in circadian rhythms, gas sensing, gene expression, and immunity. As a result, the need to uncover more regulatory heme proteins has become apparent. Many methods for studying heme binding to proteins are available, including traditional techniques (such as mutational analysis and measurements of heme binding affinity) and novel approaches (machine learning and proteomic analyses). However, little is known about how heme is made available to bind to these proteins. Knowledge of the bioavailability and the oxidation state of heme in the cell is required to understand the physiological context in which these proteins bind heme.

To this aim, we have adopted a dual approach, the first of which is an investigation into the use of bioinformatics to predict heme binding. To do this, we have employed the ligand binding prediction tool, ProFunc, in combination with AlphaFold. The predicted binding sites were assessed and compared to existing crystal structures (with or without heme bound) and with suspected binding residues from the literature. This work is the start of an exploration into the techniques available for predicting heme binding and demonstrates their practical uses and limitations. The second part of the approach is the creation of a genetically encoded redox-sensitive heme sensor for deployment in live cells. The sensor is a recombinant fusion protein of the mKate2 fluorescent protein and a variant of the heme-binding protein, myoglobin, and the first steps of its development are described in this thesis. This sensor will provide much-needed understanding into the significance of redox state in heme binding and cellular regulation and be an important step forward for the heme community.

Acknowledgements

I would first like to thank Emma Raven for welcoming me into the Raven group, and all the wisdom and guidance she has given me in the past year. I am also very grateful for all the help I have received from Andrea Gallio, for he has given me invaluable knowledge and skills and supported me every step of the way. I would also like to thank Roman Laskowski, Andrew Hudson, Samuel Freeman and Anna Bailey for their help and advice during this project. Finally, I am eternally grateful for the unwavering support from my family who never stop believing in me.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed:

Date: 17/09/2022

Table of contents

List of figures, tables and equations	vii
Abbreviations.....	xi
List of amino acids and their abbreviations	xiv
Mathematical terms.....	xv
1. Introduction	1
1.1. The importance of heme in the cell.....	1
1.1.1. Chemical and structural properties of heme	1
1.1.2. Traditional heme proteins.....	1
1.1.3. Emerging roles of cellular heme.....	2
1.2. Using bioinformatics to search for heme binding proteins	3
1.2.1. Why do we want to predict heme binding proteins?	3
1.2.2. Characterisation of permanent and transient heme binding sites	3
1.2.3. Using bioinformatics to predict heme binding	5
1.2.4. Combining bioinformatics and experimental approaches	7
1.3. Detecting heme in the cell.....	7
1.3.1. Resonance energy transfer.....	7
1.3.2. RET-based heme sensors.....	8
1.3.3. Lifetime-based measurements of fluorescence.....	10
1.3.4. The need for a redox-sensitive sensor	10
1.4. Thesis objectives	11
1.4.1. Exploration of using bioinformatics methods to predict heme binding in cells.....	11
1.4.2. Development of a novel, redox-sensitive sensor for detecting heme in cells.....	11
2. Materials and Methods.....	13
2.1. Bioinformatics.....	13
2.1.1. Prediction of heme binding sites	13
2.1.2. Visualisation and analysis of binding sites.....	13

2.2. Preparation of competent <i>Escherichia coli</i> (<i>E. coli</i>) cells	13
2.3. Transformation into competent <i>E. coli</i> cells	14
2.4. Preparation of sperm whale myoglobin variant (H64Y/V68F).....	14
2.4.1. Protein expression.....	14
2.4.2. Protein purification	14
2.5. Extraction of heme from <i>holo</i> -myoglobin.....	15
2.6. Ultraviolet-Visible absorption spectroscopy	16
2.7. Spectrophotometric titration of hemin into <i>apo</i> -myoglobin	16
2.8. Estimations of the affinity of heme binding.....	17
2.9. Agarose gel electrophoresis	17
2.10. SDS-PAGE gel electrophoresis.....	17
2.11. Cloning.....	18
2.11.1. PCR amplification.....	18
2.11.2. Restriction Digestion Cloning.....	18
3. Identification and Analysis of Heme Binding Sites.....	20
3.1. Summary of results.....	20
3.2. Predictions for proteins for which there are heme-bound crystal structures available: STEAP1, Rev-erb β , PGRMC1 and HO-2.....	21
3.3. Predictions for proteins with available crystal structures, without heme bound: GAPDH, BACH1, BACH2, p53, Rev-erb α , PER2, CLOCK, and IRP2.....	24
3.3.1. GAPDH	24
3.3.2. BACH1	24
3.3.3. BACH2.....	26
3.3.4. p53	26
3.3.5. Rev-erb α	26
3.3.6. PER2.....	27
3.3.7. CLOCK.....	27
3.3.8. IRP2	27

3.3.9. Discussion of the results	27
3.4. Predictions for proteins without available crystal structures: ALAS1 and NPAS2	29
3.4.1. ALAS1	29
3.4.2. NPAS2	30
3.5. Conclusions	31
4. Redox-Sensitive Sensor – Design and Development of the Construct.....	33
4.1. Theoretical basis for the redox sensor	33
4.2. Sperm whale myoglobin mutant (H64Y/V68F).....	35
4.2.1. Protein expression and purification	35
4.2.2. Protein characterisation	37
4.2.3. Determination of binding affinity using heme titrations	39
4.3. Red fluorescent protein - mKate2	41
4.4. Constructing the redox sensor	42
4.5. Conclusions	44
5. Perspectives and Future Work	46
References.....	47
Appendix I: NPAS2 sequence alignment between two species (human and mouse).....	55
Appendix II: Additional information on the preliminary calculations for redox sensor.	56
Appendix III: Plasmid and sequence information for sperm whale Mb(H64Y/V68F).	57
Appendix IV: The full set of data collected from heme titration into gMb.....	58
Appendix V: UV-Visible spectrum of hemin stock used for heme titrations.....	59
Appendix VI: Plasmid map for TOPO mKate2 plasmid, used for cloning of sensor.....	60
Appendix VII: Primers used during PCR amplification of gMb gene.....	61
Appendix VIII: Sequence of gMbmKate2 construct.	62

List of figures

Figure 1. The structure of heme	1
Figure 2. Summary of the dynamic roles of heme in the cell	2
Figure 3. Conventional description of heme binding sites	4
Figure 4. The first example of RET-based sensors.....	9
Figure 5. Modified RET-based sensors to include an internal standard for ratiometric measurements.....	9
Figure 6. FRET-based sensor where RET occurs between two fluorophores	10
Figure 7. The workflow used to obtain the results shown in Table 3.....	20
Figure 8. A comparison of the ProFunc model and crystal structure for four proteins for which the heme-bound structure is available in the PDB	23
Figure 9. Predicted heme binding sites in the AlphaFold models for proteins without a heme-bound structure in the PDB	28
Figure 10. Predicted binding sites from ProFunc in the AlphaFold models of ALAS1 and NPAS2	30
Figure 11. NPAS2 alignment.....	32
Figure 12. Spectra illustrating the different overlap between the fluorescence emission spectrum of mKate2 and gMb and the mKate2gMb sensor	34
Figure 13. UV-Visible spectrum of pEMBL19-Mb(H64Y/V68F).....	35
Figure 14. Elution profiles from the size exclusion chromatography during the purification of gMb	36
Figure 15. SDS-PAGE during gMb purification	37
Figure 16. UV-Visible spectrum of gMb following purification.....	38
Figure 17. Extraction of heme from gMb to form the <i>apo</i> -protein.....	38
Figure 18. UV-Visible spectra showing hemin titration into gMb and deconvolution of the spectra	40
Figure 19. TOPO mKate2 restriction digestions.....	42
Figure 20. Cloning protocol for the insertion of the gMb gene into TOPO mKate2 for the formation of the gMbmKate2 sensor	42
Figure 21. PCR amplification of gMb	43
Figure 22. Plasmid map of gMbmKate2.....	44
Figure 23. Alignment of the full sequence of human and mouse NPAS2	55
Figure 24. Plasmid map for pEMBL19-Mb(H64Y/V68F).....	57

Figure 25. UV-Visible spectrum of hemin stock	59
Figure 26. Plasmid map for TOPO mKate2 plasmid	60
Figure 27. gMbmKate2 sequence data.....	63

List of tables

Table 1. Summary of the heme binding prediction tools available in the literature	5
Table 2. PCR protocol for the amplification of the gMb gene from pEMBL19-Mb(H64Y/V68F)	18
Table 3. Summary of the AlphaFold models, obtained using the relevant UniProt IDs, submitted to ProFunc and an estimated score for heme binding (as determined by ProFunc) for a number of heme proteins, along with a brief description of protein function.....	22
Table 4. Summary of predicted binding sites in AlphaFold models of human proteins for which there is a crystal structure for the <i>apo</i> -protein available only, and relevant comparisons to binding residues in the literature.....	25
Table 5. Preliminary calculations of fluorescence lifetimes for gMbmKate2 sensor.....	56
Table 6. The full set of titration data collected to produce the results in Section 4.2.3.....	58

List of equations

Equation 1. The expression for Förster distance.....	8
Equation 2. RET efficiency in terms of Förster distance and distance between chromophores	8
Equation 3: RET efficiency in terms of fluorescent lifetimes for the <i>holo</i> - and <i>apo</i> -proteins	10
Equation 4. The relationship between absorbance with the molar extinction coefficient and path length (Beer-Lambert Law)	16
Equation 5. The equation used to determine the concentration of DNA.....	16
Equation 6. RET efficiency in terms of the distance between fluorescent protein and heme and the Forster distances for ferric and ferrous heme.....	34
Equation 7. RET efficiency in terms of the fluorescent lifetimes.....	34
Equation 8. The equation used to calculate the spectral overlap from the molar absorptivity spectrum of the acceptor, gMb, and the relative emission spectrum of the donor, mKate2.....	56

Abbreviations

ALAS1	nonspecific 5-aminolevulinate synthase
APX	ascorbate peroxidase
BACH1	BTB domain and CNC homolog 1
BACH2	BTB domain and CNC homolog 1
CLOCK	circadian locomotor output cycles kaput
CP motif	cys-pro motif
cyt <i>b</i> ₅₆₂	cytochrome <i>b</i> ₅₆₂
DEAE	diethylaminoethanol
DNA	deoxyribonucleic acid
<i>E. coli</i>	<i>Escherichia coli</i>
EDTA	ethylenediaminetetraacetic acid
EGFP	enhanced green fluorescent protein
Fe	iron (element)
FLIM	fluorescence lifetime imaging microscopy
FP	fluorescent protein
FRET	Förster resonance energy transfer
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
gMb	green myoglobin
heme	iron protoporphyrin IX
HO-2	heme oxygenase 2
HRM	heme regulatory motif
hsp90	heat shock protein 90
IDR	intrinsically disordered region

iNOS	inducible nitric oxide synthase
IRP2	iron-responsive element-binding protein 2
kb	kilobase
kDa	kilodalton
LB	Lysogeny broth
Mb	myoglobin
MCR-ALS	Multivariate Curve Resolution – Alternating Least Squares
mKate2	monomeric Katushka 2
MEK	methyl ethyl ketone
NPAS2	neuronal PAS domain-containing protein 2
P53	tumour protein 53
PAS	Per-Arnt-Sim
PCR	polymerase chain reaction
PER2	period 2
PGRMC1	progesterone receptor membrane component 1
PPIX	protoporphyrin IX
RET	resonance energy transfer
Rev-erb α	Nuclear receptor, NR1D1 (nuclear receptor subfamily 1, group D, member 1)
Rev-erb β	Nuclear receptor, NR1D2 (nuclear receptor subfamily 1, group D, member 2)
SapA	saposin A
SDS-PAGE	sodium dodecyl sulfate–polyacrylamide gel electrophoresis
sGC	soluble guanylate cyclase
SpdH	spermidine dehydrogenase
STEAP1	Six-Transmembrane Epithelial Antigen of Prostate 1

STEAP3	Six-Transmembrane Epithelial Antigen Of Prostate 3
T _a	annealing temperature
UV-Visible	ultraviolet/visible
WESA	weighted ensemble solvent accessibility

List of amino acids and their abbreviations

Ala	A	alanine
Arg	R	arginine
Asn	N	asparagine
Asp	D	aspartic acid
Cys	C	cysteine
Gln	Q	glutamine
Glu	E	glutamic acid
Gly	G	glycine
His	H	histidine
Ile	I	isoleucine
Leu	L	leucine
Lys	K	lysine
Met	M	methionine
Phe	F	phenylalanine
Pro	P	proline
Ser	S	serine
Thr	T	threonine
Trp	W	tryptophan
Val	V	valine

Mathematical terms

R_0	Förster distance in resonance energy transfer
κ	orientation factor
Φ	fluorescence quantum yield
n	refractive index
J	integral overlap between donor and acceptor
r	distance
τ	fluorescent lifetime
E	efficiency of energy transfer
A	absorbance
ε	extinction coefficient
c	concentration
l	path length
K_D	dissociation constant

1. Introduction

1.1. The importance of heme in the cell

1.1.1. Chemical and structural properties of heme

The iron protoporphyrin IX complex, commonly referred to as heme (Fig. 1), is a hydrophobic and cytotoxic molecule and thus is presumed to be under tight regulatory control in the cell.¹ Despite this, heme has many essential and dynamic roles and is most well-known for its role as a cofactor, in addition to its emerging regulatory role, in cells.² These roles are summarised in Sections 1.1.2 and 1.1.3.

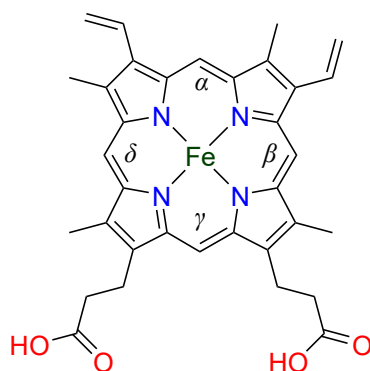


Figure 1. The structure of heme (ChEBI ID: 26355). The central iron (ferric, Fe^{3+} or ferrous, Fe^{2+}) coordinates to a tetrapyrrole ring: protoporphyrin IX (PPIX).³ There is polarity across the porphyrin ring as a result of the propionate groups. This is important for the functionality of the molecule, for example in the binding of heme to proteins.¹ The edges of heme can be denoted using the letters α , β , γ , and δ , and these are shown in the Figure.⁴

1.1.2. Traditional heme proteins

Early hints to the involvement of heme in biology go as far back as 1893, when hemoglobin was first understood to contain iron.⁵⁻⁷ Since then, heme has been understood to act as a cofactor, facilitating the functions of countless proteins.⁸ These functions include oxygen storage and transport (*e.g.* hemoglobin and myoglobin),⁹⁻¹⁰ electron transfer (*e.g.* cytochrome *c*),¹¹ oxygen activation (cytochrome P450),^{8, 11} catalysis (*e.g.* indoleamine 2,3-dioxygenase, IDO),⁸ synthesis of gases such as nitric oxide (nitric oxide synthase, iNOS),¹² heme-based gas sensor functions (soluble guanylate cyclase, sGC)¹³⁻¹⁴ and many others. These processes are vital to the workings of the cell.

Heme binds to these ‘traditional’ heme proteins with high affinity and, by all practical means, irreversibly. However, this is not always the case and transient binding is required when heme is acting in regulatory roles.¹⁵

1.1.3. Emerging roles of cellular heme

Recent discoveries point at an ever-growing number of heme-regulated cellular functions where transient, low affinity heme binding is key for the reversible activation/inactivation of cellular pathways.¹⁶⁻¹⁷ Research into understanding the role of heme as a regulator has uncovered novel binding modes which enable heme to be readily exchanged in the cell.^{2, 18} This has revealed that we know very little about the diverse roles of heme and is encouraging us to reconsider the significance of heme in the cell.

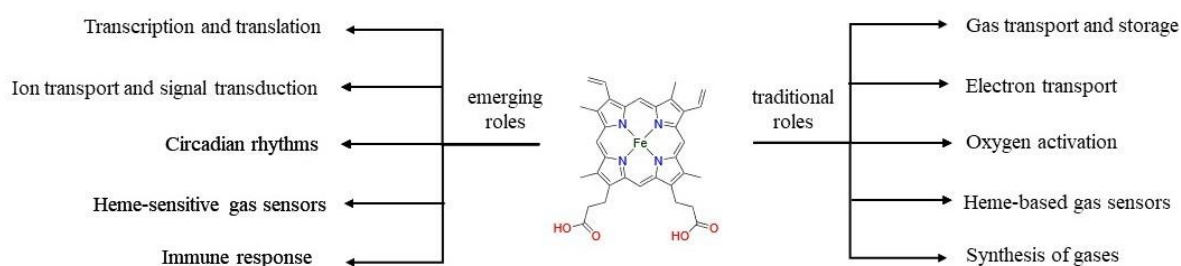


Figure 2. Summary of the dynamic roles of heme in the cell. This Figure considers both the emerging, regulatory roles of heme as well as the prototypical roles of heme where it is acting as a cofactor. However, how heme is made available to act as a cofactor or in its emerging regulatory roles remains unknown.

As an example, novel “heme-responsive sensor” functions, whereby heme binding acts as a first signal in signal transduction, are being discovered.¹⁹ This is important for physiological functions such as gene expression,^{18, 20-21} ion channel regulation,²²⁻²⁵ regulation of the circadian clock,²²⁻²⁵ and protein phosphorylation and degradation.²⁶ Heme can also act as a gas sensor by switching on and off enzymatic activities, including that of phosphodiesterase and histidine kinase.²⁷

More recently, the presence of an exchangeable supply of heme (also known as ‘labile’ heme)²⁸ has become apparent. Exchangeable heme is bound reversibly to proteins and is a fraction of total heme content in the cell. It is available to meet instantaneous demands as well as carrying out regulatory functions and its thought to bind transiently to proteins.²⁹ When proteins are transiently bound to heme, they may be moonlighting. Moonlighting is a common strategy for enzymes and unsuspected proteins may be involved in heme trafficking. For example, the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH)³⁰ has recently been established as a heme chaperone by Stuehr *et al.*^{12, 14, 31-33} Given that moonlighting roles can be very different from the primary function of the enzyme, the heme binding sites are likely to

be unconventional. This makes it difficult to identify enzymes exerting moonlighting roles relevant in heme biology, especially when using techniques that are designed to target and predict ‘permanent’ heme binding.³⁴ There are several bioinformatics techniques which can be employed to facilitate the discovery of other heme-dependent, regulatory functions, and are discussed in Section 1.2.

1.2. Using bioinformatics to search for heme binding proteins

1.2.1. Why do we want to predict heme binding proteins?

As the list of the diverse, regulatory roles of heme expands, it has become increasingly desirable to identify novel heme binding proteins. This will contribute to the understanding of the distribution of heme in the cell, as well as its regulatory role.

There are numerous suspected but unconfirmed regulatory heme proteins (including GAPDH,^{12, 14} heat shock proteins such as Hsp70,³⁵ SapA,³⁶ and spermidine dehydrogenase, SpdH³⁷). The use of bioinformatics may enable researchers to narrow the number of putative heme-binders and focus in-depth studies to the most probable binding partners. This would pave the way for the discovery of novel aspects of heme biology, that might have been hitherto concealed. A standout example is the glycolytic enzyme, GAPDH, which was shown, unexpectedly, to be responsible for the incorporation of heme into heme proteins (such as iNOS, sGC).^{12, 14, 33} The duties of GAPDH as a heme chaperone strengthen the idea that proteins carry out different functions that are context specific in the cellular milieu (*i.e.* they moonlight). These secondary functions can be quite different from their primary role, which makes the identification of such moonlighting proteins a particularly challenging endeavour.

Before discussing the existing tools available for the prediction of heme binding, it is necessary to discuss common motifs in heme binding, and how the binding sites will differ depending on the role of the heme-protein interaction.

1.2.2. Characterisation of high affinity and transient heme binding sites

There are many common features that can be identified across heme binding sites in proteins. For example, heme binding sites tend to be rich in alpha-helical structure, however not exclusively.⁶ Binding sites generally consist of two axial ligands on each side of the porphyrin ring. The most common binding residue is histidine, although other ligands are possible, including methionine, cysteine and tyrosine.⁶ Axial ligands may be ‘distal’ or ‘proximal’ and are represented either above or below the heme plane respectively (Fig. 3A).³⁸ Proximal

ligands are normally coordinated to iron, whilst distal ligands usually involve stabilisation of the substrate, an example being His64 in the binding site of sperm whale myoglobin (Fig. 3B).³⁹

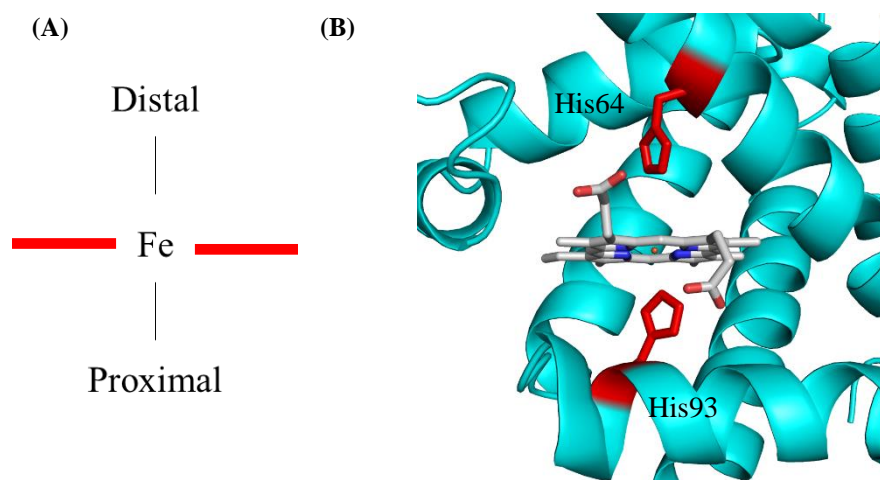


Figure 3. Conventional description of heme binding sites. **(A)** Categorisation of heme ligands, where the distal ligand is shown above and the proximal ligand is shown below the heme plane. **(B)** An example of a conventional heme binding site in sperm whale myoglobin (PDB ID: 1JP6). One of the axial ligands for heme is His93 (proximal) and the second axial ligand is either water or oxygen (H_2O/O_2). The distal histidine, His64 coordinates H_2O/O_2 .³⁹

The majority of heme binding sites in the PDB are buried inside the protein and provide heme with a hydrophobic pocket with little solvent accessibility. This is consistent with ‘permanent’ binding of heme to proteins, when acting as a cofactor.⁶ In contrast, when heme is binding transiently, it may bind on the surface of a protein. This is the case in the predicted binding site of the heme chaperone protein GAPDH, where the binding site is suspected to be in between the surfaces of the subunits of the GAPDH tetramer (PDB ID: 1ZNQ). In this case, two His53 residues, from two different protein subunits, coordinate the heme iron.³² Another example in which heme binding sites may differ from those associated with permanent binding is the presence of cysteine-proline (CP) motifs, which are found in numerous heme proteins as part of heme regulatory motifs (HRMs).⁴⁰ Although not always directly involved in heme binding, CP motifs are important for the functionality of many heme proteins, for example heme-oxygenase 2 and Rev-erb β .⁴⁰

Transient heme binding requires the heme-protein binding affinity to be lower than when heme is acting as a cofactor. Factors which will affect this include the number of non-bonding interactions (hydrophobic interactions, π - π stacking via the porphyrin ring, electrostatic interactions and hydrogen bonds)¹⁵ and the shape, size and polarity of the heme binding pocket.^{6, 41}

As novel binding modes are being uncovered, the importance of identifying transient heme binding sites is becoming increasingly important. Given that transient heme binding sites may be unlike those for traditional, high affinity heme binding proteins, efforts to predict and understand such binding modes will require different approaches than those which have been used previously. The existing tools developed specifically to predict heme binding are discussed in Section 1.2.3.^{34, 42-44}

1.2.3. Using bioinformatics to predict heme binding

Traditional methods of identifying heme proteins includes mutagenesis studies, the use of affinity columns and the identification of highly conserved histidine residues.^{32, 35} These experimental methods have been pivotal in identifying proteins that bind with a high affinity to heme. However, when heme is acting in its regulatory role, it tends to bind to proteins with lower affinities. As a result, it is necessary to employ new methods for this purpose. Methods used recently to predict heme binding include proteomic analyses and machine learning techniques.^{18, 35} These techniques are best used in combination with experimental methods yet can potentially speed up the identification by minimising time spent in the laboratory. These use various techniques such as assessing 3D structures and sequences, scanning of ligand binding templates, machine learning and more recently, deep learning.⁴⁵ Many of the techniques used are general for any case of protein-ligand binding, however there are tools which have been specifically developed to identify heme binding (Table 1). These all arise from the desire to understand the dynamic role of heme in the cell and more recently are attempting to focus on transient binding (for example, the tool named HeMoQuest).³⁴

Tool name	Examples of tools available to predict heme binding
HemeBind ⁴³ and HemeNet ⁴⁶	Structures of heme proteins used to train machine learning algorithms. These tools focus on permanent binding and do not distinguish between heme types.
TargetS ⁴⁷	Not focussed on heme binding; uses primary sequence; considers ligand properties, secondary structure, and evolutionary information; trained with a dataset of 233 structures of heme binding proteins from BioLIP database. ⁴⁸
SCMHBP ⁴²	Evaluates heme binding of 400 dipeptides and 20 amino acids which is transferred onto protein sequences.
HEMEsPred ⁴⁴	Considers permanent binding and can distinguish between heme types sequence- and structure-based predictions; gives probability of heme binding to each residue.
HeMoQuest ³⁴	Large novel dataset; predicts transient binding and the associated binding affinities; can consider solvent accessibility (WESA).

Table 1. Summary of the heme binding prediction tools available in the literature, adapted from Paul George *et al.* (2020).³⁴

When using these tools, sequence information (*e.g.*, a FASTA file) or structural data (*e.g.*, a PDB file) must be submitted as an input, depending on the tool being used. The output given to the user is a list of protein residues and a prediction of which residues may be binding to heme, sometimes with an associated probability. However, this information is not sufficient on its own in order to understand the heme binding site. For example, these tools do not give a structure associated with the heme binding predictions. As a result, we have decided to use a well-established bioinformatics tool, namely *ProFunc*, in collaboration with its creators.⁴⁹ ProFunc provides an output which can be visualised using PyMol, showing the 3D orientation of heme within its binding site.

ProFunc was developed in the Thornton group at the European Bioinformatics Institute in 2005 and is easily accessible via its webserver (<http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/>). Its multi-method approach combines sequence and structural information from various databases (Protein Data Bank, UniProt and InterPro)⁵⁰⁻⁵² to predict protein function. Upon submitting a PDB file, ProFunc performs tasks that include sequence searches, fold matching, surface cleft analysis, residue conservation and 3D template searches in order to identify functional motifs or structural homologues.⁴⁹

For the purposes of this thesis, we have used ProFunc to perform ligand searches within protein structures. In this case, the web server uses templates for heme binding sites generated by looking at structures where heme is present as a cofactor.⁵⁰ To do this, ProFunc uses the heterogeneity, abbreviated to HET (non-identifiable residues in a structure, such as a heme ligand), section of the PDB. Templates consist of two to five residues in a specific 3D arrangement.⁵³ ProFunc scans templates against the submitted PDB file of a given suspected heme binding protein and identifies matches. The server outputs a summary of possible ligands, categorised by the probability of binding (based on a heme score produced by ProFunc, from ‘certain’ to ‘long shot’) and the predicted binding sites, which can be viewed in PyMol.⁵⁴

ProFunc has been used in combination with AlphaFold. AlphaFold is a machine learning tool which can predict 3D protein structure from an amino acid sequence.⁵⁵⁻⁵⁸ For the purposes of this thesis, AlphaFold models have been submitted to ProFunc to generate the predicted heme binding sites. The full details of the methodology for this work is given in Section 2.1 and the corresponding results are given in Section 3. Along with the structure, AlphaFold gives levels of confidence from the prediction, namely ‘very confident’, ‘confident’, ‘low’, and ‘very low’. These levels of confidence have been an important consideration for the results obtained in this

thesis and have been given in the results where relevant.

1.2.4. Combining bioinformatics and experimental approaches

The use of bioinformatics is essential to facilitate studies of heme binding. Tools used to predict ligand binding to proteins, such as those described in Section 1.2.3, give a relatively quick output and require only a structure and or function; no previous knowledge of heme binding to the protein is required. These tools can search large databases to provide information on suspected binding residues, the location of the heme binding site and predictions of binding affinities for proteins of interest. However, this work must be paired with experimental studies (*e.g.*, mutagenesis studies and measurements of the affinity of protein-ligand interactions) as tools such as those described in the thesis are not yet reliable alone. Moreover, it is necessary to consider not only heme binding to proteins, but within the context of the cell. For example, heme binding will be highly dependent on the availability of heme in different cellular compartments. To this aim, heme sensors are being developed to understand how the molecule is distributed and regulated.

1.3. Detecting heme in the cell

The first quantifications of heme in the cell involved lysis of the cell and subsequent quantification, giving an estimate of total heme content.^{1, 59} However, it has since become important to understand not just the total heme content in the cell but also the portion of it which is bioavailable and not permanently sequestered by heme proteins in order to meet dynamic changes of cellular heme requirements. Cellular heme availability can be studied using optical heme sensors. The basis for heme sensors arises from the ability of heme to quench the emission of light from fluorescent reporters via resonance energy transfer (RET). RET is a photophysical interaction between chromophores wherein energy from the electronic excited states of a donor is transferred to the ground state of an acceptor.⁶⁰⁻⁶¹

1.3.1. Resonance energy transfer

For RET to occur, there must be spectral overlap between the emission of the donor molecule and the absorption of the acceptor. Importantly, the donor molecule emits at a shorter wavelength than the wavelength at which the acceptor absorbs.⁶¹ RET is also highly dependent on the distance, r , between the donor and acceptor molecules. When the distance, r , is small enough, the fluorescence of the donor is quenched via RET via the appropriate decay pathways. The Förster distance, R_0 , can be used to relate the efficiency of resonance energy transfer (RET

efficiency, E) with the distance between chromophores (Eq. 1). When $r = R_0$, the RET efficiency is at 50%.⁶⁰⁻⁶³ RET efficiency is dependent on the distance, r , (Eq. 2).

$$R_0^6 = 0.2109 \times \frac{\kappa^2 \Phi_D^0}{n^4} \times J \quad (1)$$

Equation 1. The expression for Förster distance, R_0 . In this expression, κ is the orientation factor, Φ_D^0 is the fluorescence quantum yield of the donor (without transfer), n is the refractive index and J is an integral representing the overlap between the donor emission spectrum and acceptor absorption spectrum.⁶⁰ An equation for J is given by Equation 8 (Appendix II).

$$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6} \quad (2)$$

Equation 2. RET efficiency in terms of Förster distance and distance between chromophores, r .

This phenomenon can be applied to heme sensing by creating recombinant fusion proteins which can carry out RET as a means to detect heme in the cell.

1.3.2. RET-based heme sensors

At its most simplistic, RET-based sensors work as follows. Heme binding to the sensor induces a conformational change in the protein bringing the fluorescent protein, FP, (acting as the donor) closer to heme (acceptor) such that RET can take place, and fluorescence is quenched.

This phenomenon can be applied to heme sensing by creating recombinant fusion proteins which can carry out RET. This involves the conjugation of one or more fluorescent proteins to a heme binding protein. One category of RET-based sensors involves the conjugation of a heme protein to one fluorescent protein (Fig. 4A). Heme binding (to the heme protein) induces a conformational change such that heme is brought closer to the fluorescent protein and quenches the fluorescence. This occurs because heme provides an additional relaxation pathway for the excited state of the fluorophore (Fig 4B). Examples of this include the first heme sensor in the literature, a recombinant protein consisting of EGFP and cytochrome b_{562} .⁶⁴⁻⁶⁶ Another example is the genetically encoded mAPXmEGFP sensor developed in the Raven group, which has been used as the basis for the redox-sensitive sensor described in this thesis (Sections 1.3.4 and 4).⁶³

A second category of RET-based sensors works as above but includes an additional fluorophore which acts as an ‘internal standard’ (Fig 5A). The internal standard is not sensitive to the binding of heme, such that its fluorescence is not quenched upon binding (Fig 5B). This is used to quantify the reduction in fluorescence intensity, for ratiometric measurements of heme

concentrations. An example of this is the fusion protein consisting of cyt *b*₅₆₂ and two fluorescent proteins, EGFP and mKate2. The mKate2 protein acts as the internal standard as its fluorescence is not quenched by heme binding.⁶⁷⁻⁶⁸

The final category is the use of fluorescence resonance energy transfer (FRET), which is a type of RET. Sensors of this type involve a fusion protein containing two fluorescent proteins (Fig. 6A). FRET is the resonance energy transfer that occurs between these two fluorophores; it is the quenching as a result of FRET which is measured in this case (Fig. 6B). An example is the CISDY sensor, consisting of two bacterial chaperone proteins, each conjugated to a fluorescent protein. Heme binding induces heterodimerisation between the chaperone proteins, and thus increases FRET between the fluorophores.⁶⁹

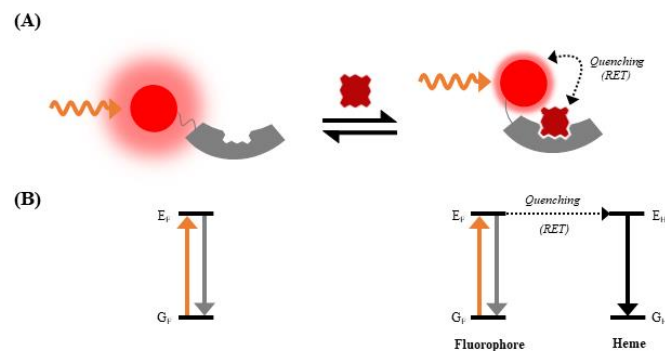


Figure 4. The first example of RET-based sensors. (A) The sensor is a recombinant protein comprising a heme protein (grey) fused to a fluorescent protein (bright red). The curly arrow (orange) shows excitation of the fluorescent protein and heme is shown in dark red. (B) Energy decay pathway for fluorophore when using this type of sensor where the excited, E, and ground, G, states of the acceptor, A, and donor, D, are shown as subscripts. Adapted from Gallio *et al.* (2021).

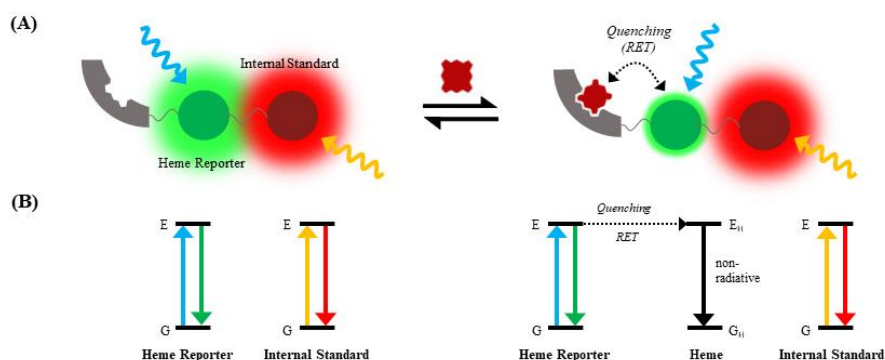


Figure 5. Modified RET-based sensors to include an internal standard for ratiometric measurements. (A) Fusion protein contains a heme protein (grey), fluorescent protein (green) and internal standard (red). The curly arrow (blue) shows excitation of the fluorescent protein. (B) Energy decay pathway unchanged from that in Fig. 4 when using this sensor. Adapted from Gallio *et al.* (2021).

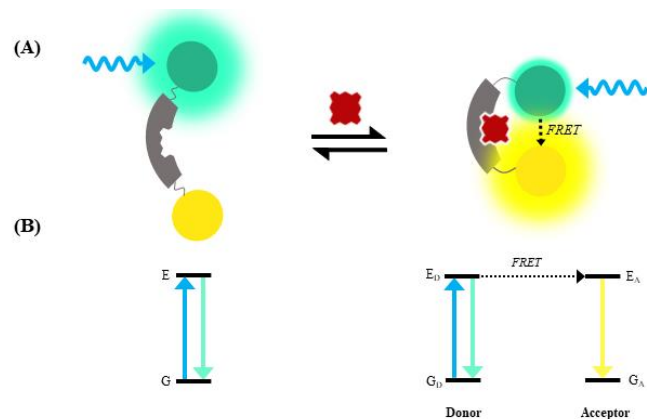


Figure 6. FRET-based sensor where RET occurs between two fluorophores. **(A)** Sensor construct comprises a heme binding domain (grey) and two fluorophores (yellow and cyan). The curly arrow (blue) shows excitation of one of the fluorophores and heme is shown in red. **(B)** Energy decay pathway for excited fluorophore, where the excited (E) and ground (G) states of the acceptor (A) and donor (D) are shown as subscripts. Adapted from Gallio *et al* (2021).

1.3.3. Lifetime-based measurements of fluorescence

An important consideration is how the fluorescence, and the quenching of fluorescence, is measured when using heme sensors. Intensity-based measurements and fluorescence lifetime measurements have both been used.¹ Fluorescent lifetime measurements are not affected by inner filtering and photobleaching, unlike intensity-based measurements.⁶³ Images showing fluorescent lifetimes can be created, where the image contrast is based on the lifetime in each region of the sample. This is known as fluorescent lifetime imaging microscopy (FLIM). FLIM gives measurements independent of probe concentration and is more reliable than intensity measurements.^{61, 63} Equation 2 can be rewritten to summarise the efficiency of RET in terms of the fluorescent lifetimes of the sensor with and without heme bound (Eq. 3).

$$E = 1 - \frac{\tau_{\text{holo}}}{\tau_{\text{apo}}} \quad (3)$$

Equation 3: RET efficiency in terms of fluorescent lifetimes for the *holo*- and *apo*-proteins, where *holo*-protein is the heme-bound protein and *apo*-protein is not bound to heme.

1.3.4. The need for a redox-sensitive sensor

We can expect the redox state of heme to be changing in the cell over time, however this is not yet fully understood. The ability to precisely measure the changing oxidation state of heme in live cells will be a huge step forward in understanding the role of heme in cellular processes and will improve the accuracy of measurements of heme concentrations.^{16, 67} It is then desirable to design a redox-sensitive heme sensor to deploy for *in cellulo* measurements. Following on

from the sensor design illustrated in Figure 4, the fluorescent and heme proteins can be chosen such that RET will be different upon ferric and ferrous heme binding. In addition to the fundamental requirement for a RET-based sensor – that there is spectral overlap between the emission spectrum of the fluorescent protein and the absorption spectrum of the heme binding domain – a sensor capable of differentiating between ferric and ferrous heme, will display differential overlap between ferric and ferrous heme with the emission spectrum of the fluorescent tag. Preliminary calculations show that a fusion protein consisting of mKate2 and a myoglobin variant, namely green myoglobin, will satisfy this condition (Section 4.1). Our attempt to develop gMbmKate2 as a redox-sensitive sensor is described in this thesis.

1.4. Thesis objectives

1.4.1. Exploration of using bioinformatics methods to predict heme binding in cells

The list of roles for the heme molecule, and thus the list of known heme proteins, are continuously growing. As such, many methods for predicting heme proteins are being explored. These vary from the traditional methods such as homology modelling, mutational analysis, and measurements of heme binding affinity, to novel approaches, including the use of machine learning and proteomic analyses. It is likely that a combination of several of these methods will be required to unravel the problem, including the use of template-based bioinformatics tools, which has been examined in this thesis.

In collaboration with template-based function predicting tool, this work demonstrates the use of a structure predicting tool, AlphaFold. The predicted binding sites for heme in various heme proteins within the AlphaFold models have been examined and have been compared to known binding residues and crystal structures, where possible. This work demonstrates that whilst an older, template-based prediction tool such as ProFunc has varying accuracy, especially for transiently binding heme proteins, it can be useful to identify important heme binding motifs in a structure to guide experimental research.

1.4.2. Development of a novel, redox-sensitive sensor for detecting heme in cells

The second objective for this thesis is the development of a novel heme sensor which builds on existing genetically encoded RET-based sensors and can be used in tandem with bioinformatics work in order to understand heme dynamics. The use of heme sensors, deployed in live cells, is required to understand how heme is distributed in different cellular compartments, how the concentration of heme changes over time and under different conditions, and with the redox-sensitive sensor, will be able to show the varying oxidation state

of heme. The sensor will thus be able to shed light on the redox dependence of cellular regulation which will be pivotal to the understanding of heme in the cell. This will provide important context to results obtained using bioinformatics tools.

2. Materials and Methods

2.1. Bioinformatics

2.1.1. Prediction of heme binding sites

Heme binding sites were predicted using the bioinformatics tool ProFunc.⁴⁹ ProFunc can be accessed via its webserver (<http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/>) and requires submission of a protein structure. The server can perform 3D template searches against protein structures to identify features such as ligand-binding sites. Heme binding templates were scanned against AlphaFold models (as PDB files) to identify potential heme binding sites.^{53, 56} The AlphaFold models were obtained using the corresponding UniProt accession number, given in Table 3.⁵¹ The results were obtained by Roman Laskowski at EMBL-EBI (European Molecular Biology Laboratory – European Bioinformatics Institute). ProFunc results included a heme binding score and a predicted binding site for each submitted protein.

2.1.2. Visualisation and analysis of binding sites

The predicted binding sites were visualised in the molecular visualisation software PyMOL.⁵⁴ The heme binding sites in the AlphaFold model were viewed along with the matched template residues. The AlphaFold models were manually scanned for histidine and cysteine residues near the heme ligand, and if present, were highlighted as binding residues. Existing structures in the PDB were overlaid with the ProFunc file using the alignment tool in PyMOL, where appropriate. When comparing the amino acid sequences of proteins between different species, a pairwise sequence alignment was used (EMBOSS Needle).⁷⁰

2.2. Preparation of competent DH5 α *Escherichia coli* (*E. coli*) cells

E. coli cells were plated on LB-agar and incubated overnight at 37 °C (Thermo Fisher, Heratherm Compact Microbiological Incubator). A single colony was inoculated in 2x YT medium and grown overnight in a shaking incubator at 37 °C, 150 rpm (Stuart SI600C Shaking Incubator). Then 2x YT medium (50 mL) was inoculated with the culture (1 mL) and grown at 37 °C, 150 rpm until OD₆₀₀ = 0.6. The cells were incubated on ice for 20 minutes and then harvested by centrifugation at 4,000 rpm for 10 minutes at 4 °C (Beckman Coulter Allegra X-30R, Benchtop Centrifuge). After discarding the supernatant, the pellet was resuspended in 0.1M CaCl₂ (10 mL) and incubated on ice for 30 minutes. The cells were harvested at 4,000 rpm for 10 minutes at 4 °C and resuspended in 0.1M CaCl₂ containing 15% glycerol (2.5 mL). The cell suspension was divided into 100 μ L aliquots, which were frozen in liquid nitrogen and then stored at -80°C.

2.3. Transformation into competent *E. coli* cells

Competent *E. coli* cells were inoculated with 50 – 100 ng of DNA and incubated on ice for 30 minutes. The cells were incubated at 42°C in a heat block (Stuart Scientific, Block Heater) for 45 seconds, then placed on ice for 2 minutes. Cells were inoculated in LB media (1 mL) and then placed in a shaking incubator at 150 rpm for a minimum of 1 hour at 37 °C. Reconstitution culture (100 µL) was plated on LB-agar (1% w/v agar) containing the appropriate antibiotics (ampicillin or kanamycin, depending on the resistance gene present in the plasmid). Plates were incubated overnight at 37 °C and then stored at 4 °C. The *E. coli* cells transformed in this way were DH5α and BL21 (DE3). The former was used to extract plasmid DNA and the latter was used to express proteins.

2.4. Preparation of sperm whale myoglobin variant (H64Y/V68F)

The plasmid containing the insert for the sperm whale myoglobin variant, pEMBL19-Mb(H64Y/V68F), was a kind gift from John Olson at Rice University.⁷¹⁻⁷² The plasmids were obtained from DH5α cells using the QIAprep Spin Miniprep Kit according to the manufacturer's instructions and the plasmid concentration was checked using UV-Visible (PerkinElmer Lambda 40 UV/Vis Spectrophotometer) using the Equation 4, given in Section 2.6. The primers used for sequencing and the results are given in Appendix III. There are published protocols for the expression and purification of this protein, which were followed as described.⁷³

2.4.1. Protein expression

The extracted DNA was used to transform *E. coli* BL21 (DE3) cells (Section 2.3). A single colony was selected from an LB_{amp} agar plate and inoculated in LB broth containing 100 µg/mL ampicillin. A starter culture was incubated overnight in a shaking incubator at 150 rpm at 37 °C.

LB medium (6x 500 mL, 20 g/L) was prepared and autoclaved in 2L baffled flasks. Ampicillin (200 mg/L) was added to the medium, which was then inoculated with starter culture (5 mL) and incubated for 24 hours at 150 rpm at 37 °C.

2.4.2. Protein purification

The cells were harvested by centrifugation at 4000 rpm for 15 minutes at 4 °C (Sorvall RC5C Plus Refrigerated Centrifuge) The cell pellets were resuspended with lysis buffer (50 mM Tris, 1 mM EDTA, 0.5 mM DTT, pH 8). A tablet of EDTA-free protease inhibitor, a micro-spatula of lysozyme and a micro-spatula of DNase were added to the suspension. The suspension was

sonicated with a 30 second cycle at 40% amplitude until lysed (Branson Digital Sonifier SFX 550). The cellular debris was separated from the homogenate by centrifugation at 19,000 rpm for 30 minutes. The pellet was discarded, and the supernatant, containing the soluble protein fraction, was filtered through a syringe (0.2 μ M).

Ammonium sulphate ((NH₄)₂SO₄) was added slowly to the cell lysate, to a saturation of 60%, over 30 minutes with stirring at 4 °C. The solution was stirred at 4 °C for 2 hours and then centrifuged at 4000 rpm for 10 minutes. The supernatant was then brought to a saturation of 95% with the slow addition of (NH₄)₂SO₄ with stirring at 4 °C and incubated for a further 1-2 hours. The solution was dialysed twice overnight using 14 kDa cut-off tubes against 20 mM Tris pH 8/1 mM EDTA. The solution was centrifuged at 4000 rpm for 10 minutes, and the pellet was resuspended in a minimum volume of lysis buffer. An excess of hemin solution was added to the suspension.

The suspension was concentrated to 3 mL using 10 kDa centrifugal concentrators and spun at 13,000 rpm. Size-exclusion chromatography used to obtain a homogenous solution. The supernatant was run through a HiLoad 16/600 Superdex 75 pg column equilibrated in 20 mM Tris pH 9.0/1 mM. The purity of the protein was assessed using SDS-PAGE.

The appropriate fractions, where the protein was at the correct molecular weight (17 kDa)⁷⁴ were pooled, concentrated to 3 mL, and loaded on a DEAE-Sepharose ion exchange column equilibrated and eluted with 20 mM Tris pH 8.4/1 mM EDTA. The UV-Visible spectrum was measured in the range 200-700 nm and the protein concentration was calculated as described in Section 2.6.

2.5. Extraction of heme from *holo*-myoglobin

Heme can be extracted from heme-bound myoglobin (referred to as the *holo*-protein) to form *apo*-myoglobin. The protein sample was run through a PD-10 column with Sephadex 0.25 resin equilibrated with deionised water. The extraction of heme from *holo*-myoglobin followed a previously published method.⁷⁵

The salt-free protein sample was acidified with dilute hydrochloric acid. The colour of the protein changed from green to orange/red. An equal amount of methyl ethyl ketone (MEK) was added to the protein and the mixture was shaken gently. The top organic layer turned pink/red. The mixture was incubated on ice for 5 minutes, and the organic layer was removed with a Pasteur pipette. The protein was dialysed four times, each overnight at 4 °C (1L of 6mM NaHCO₃/1 mM EDTA, 2X 1L of 6 mM NaHCO₃ and 2L of 50 mM KPi pH 7).

In contrast to the *holo*-protein, the *apo*-protein has no characteristic Soret peak or Q-bands in its UV-spectrum. The UV-Visible spectrum was measured to confirm successful heme extraction and to calculate the protein concentration.

2.6. Ultraviolet-Visible absorption spectroscopy

Ultraviolet-visible (UV-Visible) absorption spectroscopy was used to confirm the presence of proteins and DNA in biological samples. Solutions of known volumes are made up in a cuvette and the absorption is measured over wavelengths of 200 – 700 nm. The concentration, c , (M) of a sample can be calculated by relating absorbance, A , with the molar extinction coefficient, ϵ , ($M^{-1} \text{ cm}^{-1}$) and path length, l (cm). This relationship is demonstrated by the Beer-Lambert law:

$$A = \epsilon cl \quad (4)$$

Equation 4. The relationship between absorbance, A , with the molar extinction coefficient, ϵ , and path length, l .

The path length was kept consistently at 1 cm in all experiments. To calculate protein concentrations, the absorbance at 280 nm was used, because this is the wavelength at which aromatic residues in proteins absorb.

The extinction coefficients used for the protein Mb(H64Y/V68F) were given as $\epsilon_{280 \text{ nm}} = 15,200 \text{ M}^{-1} \text{ cm}^{-1}$ and $\epsilon_{412 \text{ nm}} = 90 \text{ mM}^{-1} \text{ cm}^{-1}$.^{71, 76} The absorbance at 280 nm was used to calculate the concentration of the *apo*-protein and the absorbance at 412 nm was used to calculate the concentration of the *holo*-protein. The extinction coefficient used to calculate the concentration of hemin was $\epsilon_{385} = 58.4 \text{ mM}^{-1} \text{ cm}^{-1}$.⁷⁷

The concentration of DNA (ng/ μL) was calculated using the absorbance at 260 nm, which is the wavelength at which DNA helix absorption occurs. This is related to l (cm), the *conversion factor* for DNA (50 $\mu\text{g}/\text{mL}$) and the *dilution factor* used in the cuvette, as follows:

$$c = \frac{A_{260 \text{ nm}}}{l} = \text{conversion factor} \times \text{dilution factor} \quad (5)$$

Equation 5. The equation used to determine the concentration of DNA using the absorbance at 260 nm.

2.7. Spectrophotometric titration of hemin into *apo*-myoglobin

The regeneration of holo-myoglobin from apo-myoglobin requires the addition of hemin (Fe(III)-PPIX chloride) (Fluka, BioChemika). Stock solution was prepared by dissolving a tiny micro spatula of hemin in 0.1M NaOH (50 μL) and diluted with water (1 mL). The solution

was spun down on a table-top centrifuge at 13,000 rpm for 1 minute and the pellet discarded. The supernatant (500 μ L) was diluted with water (500 μ L) and centrifuged again at 13,000 rpm. Hemin stock was made fresh at the beginning of each day and kept on ice. Titrations were recorded using UV-Visible spectroscopy, as described in Section 2.7.

Hemin stock (100 – 300 μ M) was added to 900 μ L of the *apo*-protein (1 – 10 μ M). Each addition of hemin corresponded to 1/10th of a molar equivalent to the *apo*-protein, until two equivalents were reached. The titrations were performed in a 1 cm quartz cuvette against a reference cuvette containing 900 μ L of water. Measurements were obtained in the range 200 – 700 nm.

2.8. Estimations of the affinity of heme binding

The binding affinity of ferric heme to the mutant myoglobin, Mb(H64Y/V68F), was estimated. Hemin was titrated into the *apo*-protein (Section 2.7). MCR-ALS software was used to deconvolute absorption spectra into the three components, namely *apo*-protein, *holo*-protein and free hemin.⁷⁸ A fitting program was then used to estimate the binding affinity. The fitting program was designed by Andrew Hudson (University of Leicester) and has been described by Leung *et al.* (2019).⁷⁹

2.9. Agarose gel electrophoresis

Agarose gel electrophoresis was used to assess the sizes and purities of protein samples. Agarose was dissolved in 1x TAE buffer (40 mM Tris/0.4 mM EDTA pH 8/20 mM acetic acid) to a concentration of 0.7%, or 1% when using low melting point agarose. The solution was microwaved (950W) at 30 second intervals until a clear solution was obtained. The solution was left to cool slightly before the addition of a dye, peqGREEN, and then left to set. Digested DNA was prepared as described in Section 2.10 and 10 ng was mixed with 5x dye (2 μ L). The samples were loaded alongside a DNA ladder into the wells of the gel and run at 80 V for 30 minutes in TAE buffer. The gel was then visualised under UV-light. The DNA ladders used were 1 kb Plus DNA Ladder and Quick-Load Purple 1 kb DNA Ladder (New England BioLabs, NEB).

2.10. SDS-PAGE gel electrophoresis

SDS-PAGE analysis was carried out to assess the size and purity of protein samples. The acrylamide gel consisted of a 12% stacking gel and a 4% resolving gel. Solutions were prepared by the addition of 2x sample buffer (diluted from Thermo Fisher 10x Invitrogen BlueJuice) in a 1:1 ratio, spun down on a table-top centrifuge for 10 seconds and boiled at 100°C for five

minutes. Samples were run in SDS running buffer (25 mM Tris/190 mM glycine/3.5 mM SDS) at 160V. Gels were stained using Instant Blue (Expedeon) for 1 hour and then washed with water. The protein standards used were SeeBlue Plus2 (Invitrogen) and Color Prestained Protein Standard, Broad Range (NEB).

2.11. Cloning

2.11.1. PCR amplification

Tube Number	Temperature (°C)	Time (s)
1	98	120
2	98	10
3	60, 61.3, 63.3, 65.6, 67.8, 69.5	15
4	72	60
5	72	120
6	4	



X 25

Table 2. PCR protocol for the amplification of the gMb gene from pEMBL19-Mb(H64Y/V68F).

For the amplification of the mutant myoglobin gene from pEMBL19-Mb(H64Y/V68F), the SensoQuest Lab Cycler was used in the BrisSynBio laboratory (University of Bristol, School of Chemistry). The amplification primers were designed to include restriction sites AclI and NotI, contained in the TOPO mKate2 vector (Addgene plasmid #68441; <http://n2t.net/addgene:68441>; RRID:Addgene_68441), for subsequent insertion.⁸⁰ The primers are given in full detail in Appendix III. Six reaction mixtures were prepared on ice with PrimerSTAR DNA Polymerase, forward and reverse primers (10 μ M, 1 μ L each), pEMBL19-Mb(H64Y/V68F) (100 ng) and deionised water (9.5 μ L). Each mixture was subjected to a different annealing temperature (T_a) in the range 60 – 69.5°C (Table 2).

The PCR products were run on an agarose gel and the bands were viewed under UV light. The PCR products were excised from the gel and purified using QIAquick Gel Cleanup Kit (QIAGEN) following manufacturer's instructions.

2.11.2. Restriction Digestion Cloning

The PCR product with the highest T_a was selected for cloning into TOPO mKate2. The gene was digested with the restriction enzymes NotI-HF and AclI (NEB). The reaction mixtures

were prepared with the PCR product (20 μ L), 10 x rCutSmart buffer (NEB) (5 μ L), NotI-HF (1 μ L), AclI (1 μ L) and deionised water (23 μ L) and incubated at 37 °C for 1 hour.

The TOPO mKate2 vector (1 μ g) was digested with the same restriction enzymes and buffer. Both digestion products were purified using QIAquick PCR & Gel Cleanup Kit (QIAGEN) following manufacturer's instructions. The products were ligated in a 1:3 ratio of plasmid to insert using Quick Ligase (NEB). The reaction mixtures were prepared on ice with digested TOPO digested mKate2 (1 μ L), gMb insert (3 μ L), Quick Ligase Reaction Buffer (2X) (10 μ L), water (6 μ L) and Quick Ligase (1 μ L) (NEB) and incubated at room temperature for 15 minutes. The ligation product was transformed directly into competent DH5 α as described in Section 2.3. Sanger sequencing was used to confirm the identity of the cloning product and the results are given in Appendix VIII.

3. Identification and Analysis of Heme Binding Sites

This Section describes the use of ProFunc, a protein function predicting tool, in combination with AlphaFold. AlphaFold is a machine-learning based tool which predicts protein structure using sequence information (Section 1.2.3). Heme binding sites have been predicted for 14 human proteins, each with an associated heme score. The binding sites have been analysed, and the uses and limitations of the methodology have been assessed.

3.1. Summary of results

ProFunc was described in Section 1.2.3 and has been employed in this thesis to predict heme binding in a test set of human proteins, following the workflow given in Figure 7.⁴⁹ The proteins were selected to contain known and suspected heme binding proteins. Structural information was submitted to ProFunc in the form of AlphaFold models (as PDB files),^{56, 58} as described in Section 2.1.1. To test the capability of this method, we first examined the predicted binding sites of four proteins with structurally determined heme binding sites (in the PDB), namely STEAP1, Rev-erb β , PGRMC1, and HO-2. The predicted binding sites for these proteins were used as a justification for the method and are detailed in Section 3.2.

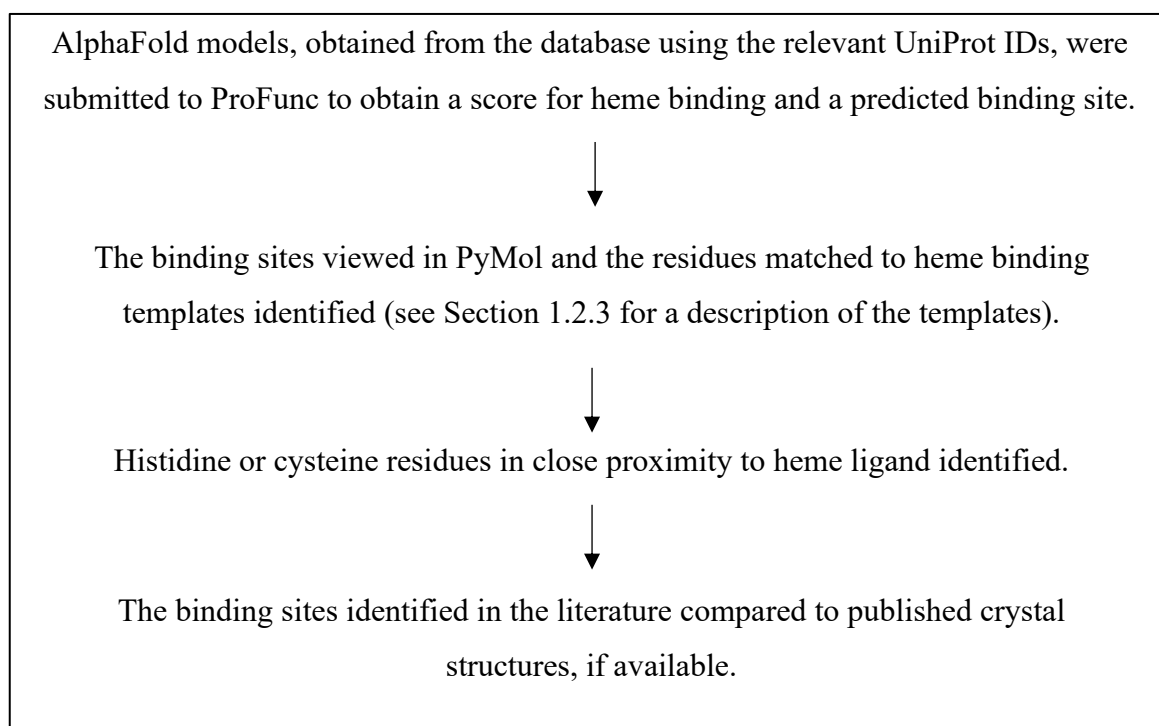


Figure 7. The workflow used to obtain the results shown in Table 3.

The results obtained for the 14 proteins in the test set have been summarised in Table 3. The Table shows the name of the protein and its UniProt ID,⁵¹ alongside its associated ‘heme score’. These scores are an assessment of heme-binding, generated by ProFunc, and range from 106 (CLOCK protein) to 740 (HO-2). All proteins except HO-2 have low scores (≤ 460) for heme binding, meaning ProFunc has assigned heme as an unlikely ligand for the protein. HO-2 is the only protein ranked as a ‘certain’ ligand for heme binding, with a score of 740. However, there is evidence in the literature that all these proteins bind heme. In fact, there are published crystal structures, with heme bound, for four of those proteins (STEAP1, Rev-erb β , PGRMC1 and HO-2) (Fig. 8). Whilst the scores for heme binding clearly do not reflect what is known about heme proteins, a predicted binding site for each protein was generated and assessed.

3.2. Predictions for proteins for which there are heme-bound crystal structures available: STEAP1, Rev-erb β , PGRMC1 and HO-2

This Section analyses the predicted binding sites for four heme proteins with known heme binding sites (Rev-erb β ,⁸¹ PGRMC1,⁸² HO-2,⁸³ and STEAP1⁸⁴). In this case, there are published heme-bound crystal structures for these proteins in the PDB (PDB IDs: 6WMQ, 4X8Y, 2QPP, and 6Y9B, respectively). As a result, there is a consensus on the heme binding site including the location of heme and the binding residues. The predicted site from ProFunc and the crystal structure from the PDB were compared to assess the use of ProFunc as a prediction tool to provide an understanding of the reliability of predictions of binding sites in proteins for which there is no known structure. This was done by aligning the predicted sites from ProFunc with the corresponding crystal structures, such that direct comparisons can be made (Fig. 8).

In all four cases, ProFunc has identified the most likely heme binding site consistent with the heme-bound crystal structures, as shown by the alignment in the right panels of Figure 8. While this is reassuring, it is not surprising given that these heme-bound structures are in the PDB and that ProFunc uses heme binding templates generated from heme-bound structures in the PDB. It is important to note, also, that the accuracy of AlphaFold and ProFunc increase as entries in the PDB database increase. These four proteins thus act as a good control to confirm that ProFunc works well to predict the binding site, when provided with accurate structural data.

Category	Protein	Protein Function	UniProt Accession Number	ProFunc Score for Heme Binding
Heme-bound crystal structure available	STEAP1	Metalloreductase (reduces Fe ³⁺ and Cu ²⁺); NAD ⁺ acceptor	Q9UHE8	217.5
	Reverb-β	Transcriptional repressor	Q14995	460.0
	PGRMC1	Part of progesterone-binding protein complex	O00264	129.8
	HO-2	Catalyses heme degradation	P30519	740.0
No heme-bound crystal structure	GAPDH	Catalyses the first step of the glycolytic pathway	P04406	160.1
	BACH1	Transcriptional regulator, regulated by heme	O14867	115.6
	BACH2	As BACH1, and plays a role in immunity	Q9BYV9	130.2
	p53	Tumour suppressor; heme binds P53-DNA complex and triggers nuclear export and degradation of P53	P04637	145.6
	Rev-erb α	Transcriptional repressor; plays a role in regulation and metabolism	P20393	295.6
	PER2	Transcriptional activator, important in regulation of circadian clock	O15055	108.6
	CLOCK	Transcriptional activator, important in regulation of circadian clock	O15516	106.2
No published crystal structure	IRP2	Posttranscriptional regulator of iron metabolism via RNA-binding	P48200	135.6
	NPAS2	Transcriptional activator, important in regulation of circadian clock	Q99743	135.1
	ALAS1*	Catalyses the formation of aminolevulinic acid (ALA); heme binding induces ALAS1 degradation	P13196	157.0

* There is a crystal structure available in the PDB, but not for the human protein.

Table 3. Summary of the AlphaFold models, obtained using the relevant UniProt IDs, submitted to ProFunc and an estimated score for heme binding (as determined by ProFunc) for a number of heme proteins, along with a brief description of protein function. All scores are colour-coded with red for ‘certain’ binding and blue when the ligand is considered a ‘long shot’ (*i.e.*, high unlikely) ligand for the protein. Other categories exist in addition to long shot and certain, *e.g.* ‘possible’, however none of the proteins in the test set fell into these categories. All proteins in the test set are human and their full names are given in Abbreviations.

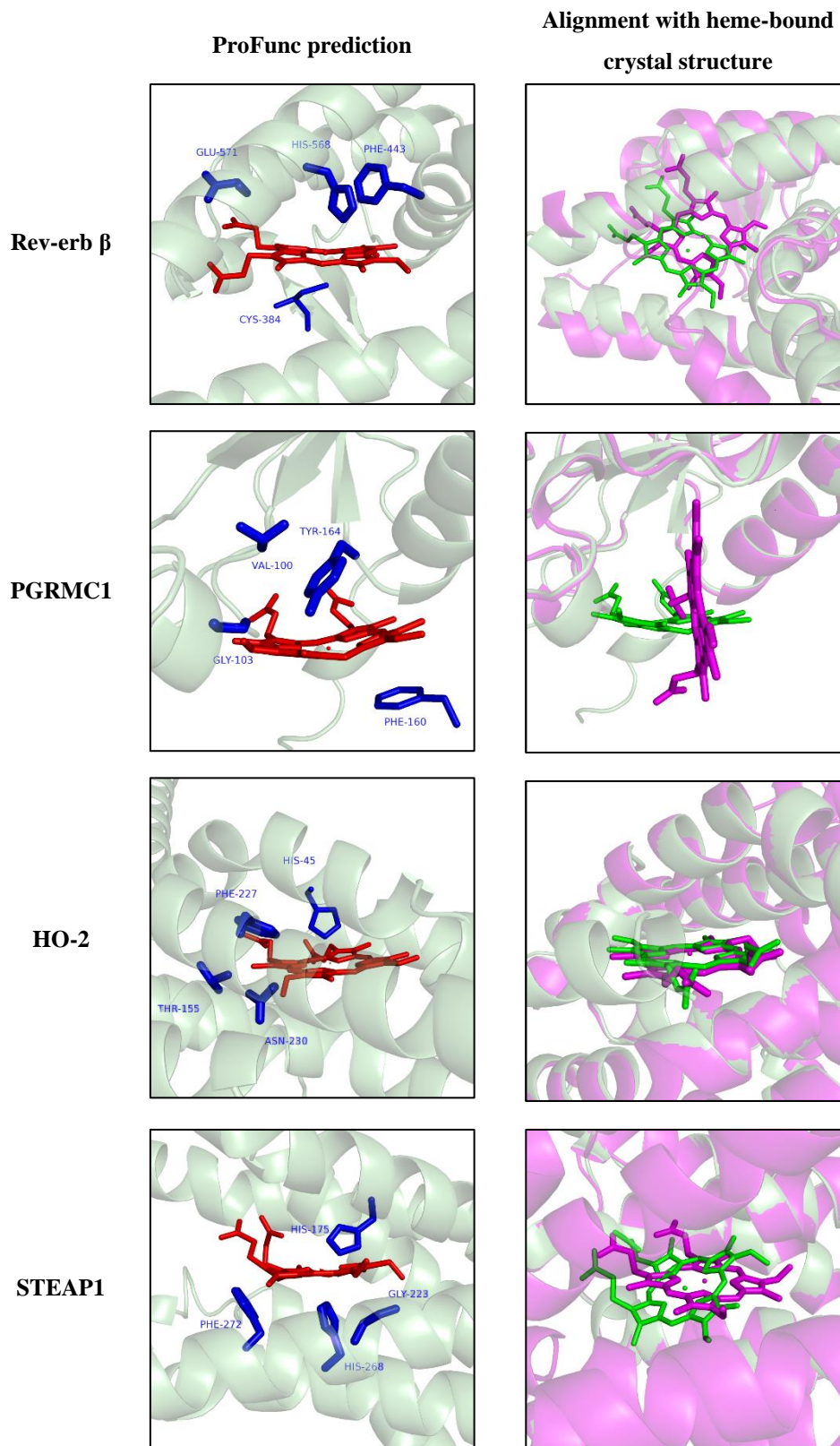


Figure 8. A comparison of the ProFunc model and crystal structure for four proteins for which the heme-bound structure is available in the PDB. Each left panel of the Figure shows the predicted heme binding sites obtained from ProFunc using AlphaFold models from the PDB. Each right panel shows a comparison of the model on the left (green) with the crystal structure from the PDB of each of the human proteins (magenta). The proteins examined and the corresponding heme-bound crystal structures used for the alignments are as follows: Rev-erb β (6WMQ), PGRMC1 (4X8Y), HO-2 (2QPP), and STEAP1 (6Y9B).

3.3. Predictions for proteins with available crystal structures, without heme bound: GAPDH, BACH1, BACH2, p53, Rev-erb α , PER2, CLOCK, and IRP2.

The heme binding sites predicted by ProFunc in the AlphaFold models of proteins which have a published structure in the PDB, for the *apo*-protein *i.e.*, without heme bound, are shown in Figure 9. The ProFunc models can be compared to the published structures to assess the practical use of AlphaFold for this purpose. A summary of the predicted binding sites for each of the proteins is given in Table 4. ProFunc was also used to assess three other proteins, non-human, however these have not been examined here. These names of those proteins (with the corresponding UniProt ID in brackets) were L-serine kinase SbnI (Q2G1M5), Ecdysone-induced protein 75B (P13055), Hormone receptor 51 (A1ZA01), and Tryptophanyl-tRNA synthetase (B7Z6G7).

3.3.1. GAPDH

GAPDH is a heme protein which has gained significant interest for its recently uncovered role as a heme chaperone and has been described in Section 1.1.3.^{12, 14, 31-33} The ProFunc model shows Thr153, Thr154 and Leu157 interacting with the α -hydrophobic edge of the heme, as shown in Figure 9 (see Section 1.1.1 for a description of the different hydrophobic edges of heme). The residue His179 is the closest histidine to heme implicating it as an axial ligand. The heme is buried in the protein, and in a region of high secondary structure. However, His53 has been identified as the most likely binding residue by Stuehr *et al.*³² Their work involved identification of the conserved histidine residues across several species, mutating them to alanine and assessing the reduction in heme binding in the mutants compared to wild-type GAPDH. They suggest that the heme ligand sits in between subunits of the GAPDH tetramer, binding two His53 residues as axial ligands, each from different subunits. However, His53 was not identified in the ProFunc model and very far from the predicting binding site. The residue His179 was investigated as for His53, but was not determined to be the most likely heme binding residue in mutagenesis studies.³²

3.3.2. BACH1

In the ProFunc model for the transcription factor BACH1 (Fig. 9), the identified heme binding residues are Gln634, Gln636, Ile652 and Cys625. The glutamine residues, Gln634 and Gln636, are located on an alpha helix, and Ile652 is in a region of low secondary structure. The cysteine residue, Cys625, may be an axial ligand given its proximity to heme.

Protein	Predicted heme binding residues in AlphaFold model	Nature of predicted site in AlphaFold model	Binding residues identified from literature	PDB ID for human <i>apo</i> -protein	Ref.
GAPDH	His179, Thr153, Thr154, Leu157 (interacting with hydrophobic edge of heme)	Region of high secondary structure; residues are part of β -sheet.	His53 – coordination of two H53 residues from two different protein subunits	1ZNQ	32
BACH1	Gln 363, Cys625, Gln634, Ile652	Binding residues are part of α -helical structure.	Six possible CP motifs involved in heme binding	2IHC	85-86
BACH2	Thr83, Gly88, Leu90, Cys112	Heme ligand sandwiched between two α -helices.	Five possible CP motifs involved in heme binding, four of which are present in intrinsically disordered regions (IDR).	3OHU	85-86
p53	Pro98, Ser99, His168, Ser269	Some secondary structure (β -sheet) but identified residues are all part of region with low or no secondary structure (except for Ser269).	Cys275-Ala-Cys277	3D09	87-88
Rev-erb α	Phe477, His602, Glu604	Highly structured region (α -helical).	Cys418-Pro419/His602	3N00	89
PER2	Ala350, Thr335, Thr334, Cys1130	Low or no secondary structure in heme binding site.	Cys841/Pro842	6OF7	90
CLOCK	His144, Asn156, Phe157, Tyr210	Low or no secondary structure in heme binding site.	His/His144	6QPJ	23
IRP2	His128, Leu130 and Ser579	Low or no secondary structure in heme binding site.	Cys201-Pro-Phe-His204	6VCD	91

Table 4. Summary of predicted binding sites in AlphaFold models of human proteins for which there is a crystal structure for the *apo*-protein available only, and relevant comparisons to binding residues in the literature. Residues identified by ProFunc as matching a heme binding template given in bold; other residues manually identified qualitatively by analysing binding site in PyMol. A relevant PDB ID for each of the the *apo*-proteins are given. Part of the Table has been adapted from Shimizu *et al.* (2019).

In fact, the BACH1 protein is known to bind heme via CP motifs, although exactly which of six possible CP motifs (Cys224, Cys301, Cys438, Cys464, Cys495, and Cys649) in the protein remains unknown.⁸⁵⁻⁸⁶ Unfortunately, no possible CP motifs have been identified in the heme binding site in the ProFunc model (Cys625 is not part of a CP motif).

3.3.3. BACH2

As shown in the ProFunc model for BACH2 (Fig. 9), another transcription factor, the predicted binding residues are Thr83, Gly88, Leu90, and Cys112. As shown in the Figure, the residues Gly88 and Leu90 interact with the δ -hydrophobic edge of heme and are located on an alpha helix. The residue Thr83 interacts with one of the propanoate groups of heme. Additionally, Cys112 is a possible axial ligand for heme in this binding site. As for BACH1 (Section 3.3.2) there are several possible CP motifs involved in binding however exactly which of these are involved in binding is not known.⁸⁶ In this case, there are five possible CP motifs (Cys369, Cys499, Cys506, Cys603, and Cys729) in the protein and four of these are present in intrinsically disordered regions (IDR).⁸⁶ Similarly, no possible CP motifs have been identified in the heme binding site in the ProFunc model.

3.3.4. p53

The ProFunc model of p53, an enzyme known for its role as a tumour suppressor, shows heme binding via Pro98, Ser99, His168 and Ser269 (Fig. 9). The residues Pro98 and Ser99 are located within unstructured regions of the protein and Ser269 is located on a β -sheet. His168 was identified in the ProFunc model for its close proximity to the heme ligand, and is located within a disordered region of the model. In contrast, mutagenesis studies have shown that heme binding occurs using the residues Cys275-Ala-Cys277.⁸⁷⁻⁸⁸ These residues were not identified in the ProFunc model.

3.3.5. Rev-erb α

The binding residues identified by ProFunc for Rev-erb α are Phe477, His602 and Glu604, all of which are located on alpha helices (Fig. 9). In a study by Yin *et al.* (2007), His602 was also identified as a heme binding residue using mutagenesis studies.⁸⁹ Interestingly, there is a published crystal structure of the related protein Rev-erb β , which is shown in Section 3.2. In the case of Rev-erb β , ProFunc was able to successfully predict the heme binding site consistent with the crystal structure, and in this case, the ProFunc model is also consistent with mutagenesis studies described in the literature. The consistency between both proteins may be a result of similarity between Rev-erb proteins, and the fact that the AlphaFold model is deemed

to be of high confidence in this region. An explanation of confidence levels in AlphaFold models has been given in Section 1.2.3.

3.3.6. PER2

In the case of PER2, Thr334, Thr335, Ala350, and Cys1130 were identified by ProFunc as binding residues (Fig. 9). However, as shown by the low levels of secondary structure in Figure 9, heme is located in a region of very low confidence in the AlphaFold model, making this an unlikely binding site. The CP motif, Cys841-Pro842 has been identified by mutagenesis studies as important for heme binding,⁹⁰ however this is a region of very low confidence in the AlphaFold model, making it unlikely that ProFunc would successfully predict heme binding in this region.

3.3.7. CLOCK

The residues Asn156, Phe157 and Tyr210 are shown to be interacting with the α -hydrophobic edge of the heme, and His144 was identified as a possible axial ligand by ProFunc (Fig. 9). Whilst the predicted heme binding site is in a region of very low secondary structure, this binding site is consistent with the literature and His144 has been identified previously identified by Freeman *et al.* (2019).²³

3.3.8. IRP2

In the ProFunc model, heme is in a disordered region of the protein and the suspected binding residues are His128, Leu130 and Ser579 (Fig. 9). These residues are located in a 'confident' region in the AlphaFold model. The residues Cys201-Pro-Phe-His204 have been identified, but as in the case of other proteins, this is in a region of low confidence in the AlphaFold model so it is unlikely that ProFunc could make a reliable prediction consistent with the literature.⁹¹

3.3.9. Discussion of the results

There are common features of heme binding sites, which have been used to identify potential binding sites. This includes the binding of histidine and cysteine to heme as axial ligands and the presence of structural motifs such as PAS domains, or regulatory, *e.g.* CP motifs (see Introduction Section 1.2.2). Overall, histidine residues were identified by ProFunc in the predicted binding sites of five of the eight proteins that have been examined (GAPDH, p53, Rev-erb α , PER2 and CLOCK). Of these, at least two are consistent with results from mutagenesis studies (CLOCK²³ and Rev-erb α ⁸⁹). As shown in Table 4, BACH1, BACH2, Rev-erb α , PER2 and IRP2 are thought from experimental work to bind heme in part using CP motifs. However, CP motifs have not been identified in the ProFunc models. It is possible that regulatory motifs such as these are harder to identify using template-based methods such as

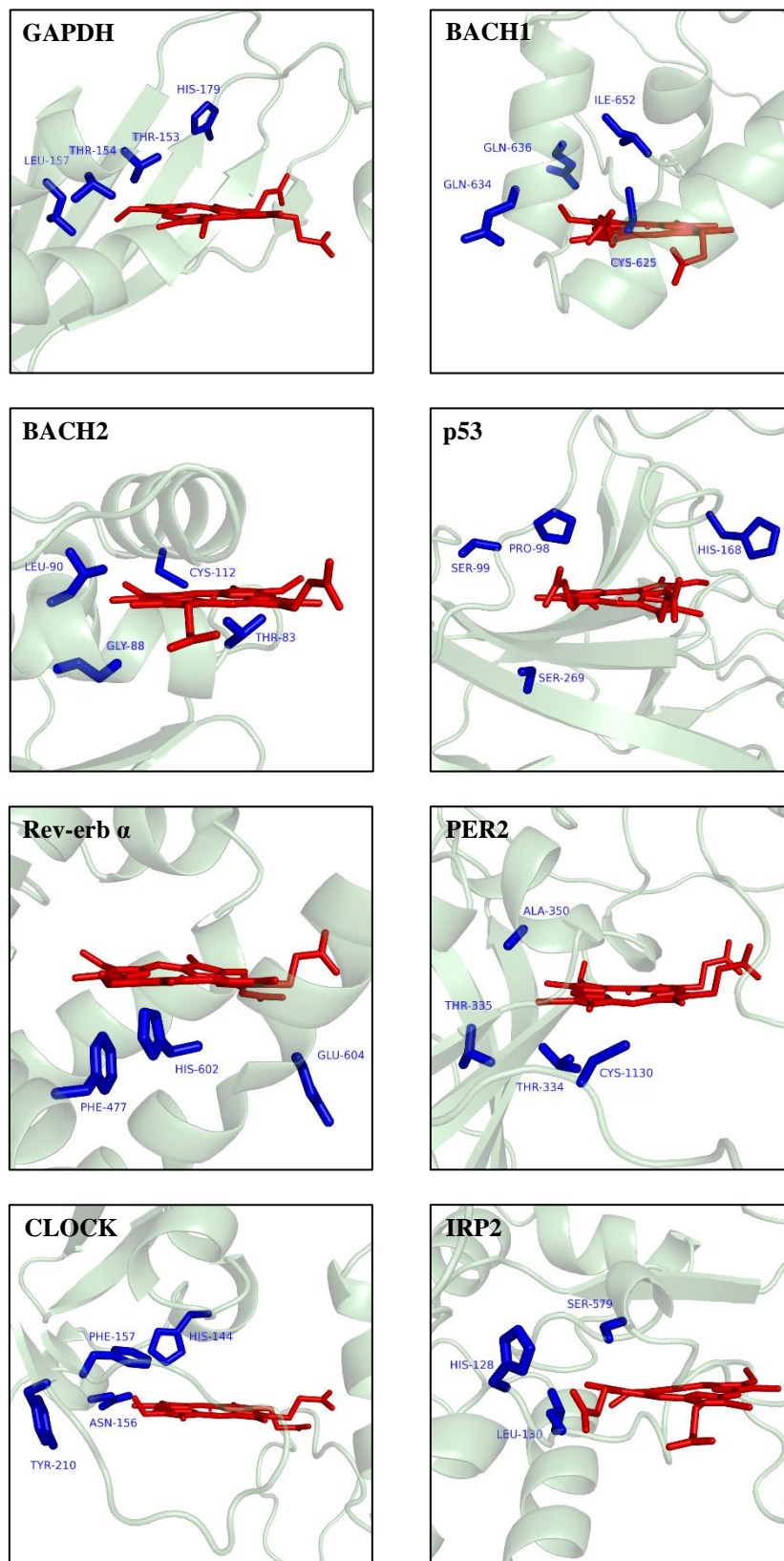


Figure 9. Predicted heme binding sites in the AlphaFold models for proteins without a heme-bound structure in the PDB (GAPDH, BACH1, BACH2, p53, Rev-erb α , PER2, CLOCK and IRP2). There are published crystal structures for the *apo*-proteins, and the relevant PDB IDs are given in Table 4.

ProFunc, given that they are less well-characterised by crystallographic studies in the context of heme binding.

Other factors to consider include the propanoate groups of the heme, which are polar and often provide interactions required for heme binding. An example of this can be found in the predicted binding sites of BACH2 and CLOCK, where the polar residues threonine and asparagine may interact with the propanoate groups of heme. Also, heme is largely hydrophobic and thus is expected to form interactions with non-polar residues in the binding site. Examples of possible hydrophobic interactions in the predicted binding sites involve those with leucine (GAPDH, BACH2), isoleucine (BACH1), glycine (BACH2), proline (p53), phenylalanine (Rev-erb α , CLOCK) and alanine (PER2).

Overall, Section 3.3 demonstrates the importance of the quality of the structural information being submitted to ProFunc. In the cases for which the AlphaFold model is of higher confidence, shown by higher levels of secondary structure, in the predicted heme binding site (*e.g.*, GAPDH, BACH1, BACH2, and Rev-erb α) or there are similar proteins with heme-bound structures in the PDB (*e.g.*, Rev-erb α), the predicted heme binding sites tend to be consistent with literature studies. In cases such as PER2 and IRP2, the binding residues identified from mutagenesis studies are in residues of low or very low confidence in the AlphaFold models, such that it is unlikely that ProFunc could make a reliable prediction of heme binding consistent with experimental studies.

3.4. Predictions for proteins without available crystal structures: ALAS1 and NPAS2

The binding sites predicted by ProFunc in the AlphaFold model of NPAS2 and ALAS1 are shown in Figure 10. Currently, there are no published crystal structures for these proteins at all. In cases such as these, the use of structure prediction methods such as AlphaFold are clearly required.

3.4.1. ALAS1

The binding site identified in the ALAS1 model consists of His259, Arg261 and Ala265. The histidine residue is likely to interact with the heme ion as an axial ligand. Arg261 interacts with one of the propanoate groups and Ala259 interacts with the δ -hydrophobic edge of the heme molecule. The heme ligand is located between three alpha helices in a well-structured region of the model. Mutagenesis studies implicate a CP motif positively involved in heme binding, namely Cys108-Pro109 in the human protein.⁹² However, this motif is in a region of low or

very low confidence in the AlphaFold model, so was unlikely to be identified by ProFunc (UniProt ID: P13196). This further demonstrates that the quality of the predictions is highly dependent on the quality of the protein structure being used.

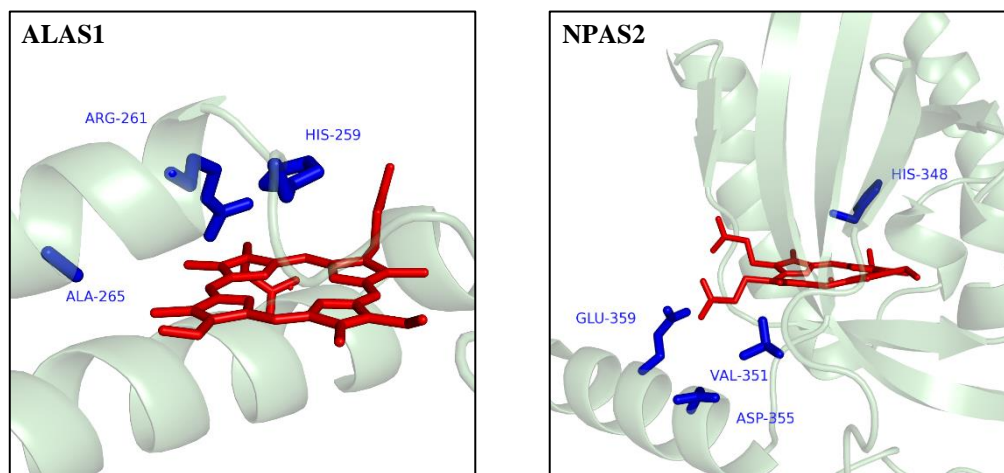


Figure 10. Predicted binding sites from ProFunc in the AlphaFold models of ALAS1 and NPAS2. There are no crystal structures of the human protein in the PDB.

3.4.2. NPAS2

NPAS2 is an important protein involved in the regulation of the circadian clock via its binding to DNA and is known to bind heme in a 1:1 stoichiometry via His/Cys or His/His at its axial sites.^{2, 40, 93} In the AlphaFold model submitted to ProFunc, the residues Val351, Asp355, Glu359 and His348 were identified as heme binding residues. These residues are shown to interact with the propanoate groups of heme. Asp355 and Glu359 are located on an alpha helix, and Val351 is on a linker between this helix and the rest of the protein.

Although ProFunc has identified a region typical of a PAS domain within the AlphaFold model,²³ the β -sheet is in fact bisecting the heme molecule (Fig. 10). Whilst this is unlikely, this does not discount the prediction altogether, given that proteins are mobile structures. It is possible that heme uses the predicted binding residues (Val351, Asp355, Glu359, and His348) to bind in the pocket, albeit in a different orientation or slightly different position. However, NPAS2 is known to have biaxial coordination with histidine and/or cysteine residues. Mutational and spectroscopic analyses have identified His119 and Cys170/His171 as the likely axial ligands for heme binding in the mouse NPAS2 protein.^{2, 93-94} A pairwise sequence alignment was used to compare the NPAS2 protein between human and mouse species. This showed 90.6% sequence similarity and confirms that residue numbering for the protein is the

same in both species (Fig. 11A). Subsequently, the binding residues from the literature (His119, Cys170, and His171) were searched for and highlighted in the same file as the predicted binding site from ProFunc (Fig. 11B). The binding residues from both possible binding sites (from the literature and the prediction) have thus been shown within the one AlphaFold model for NPAS2 and can be directly compared.

The AlphaFold model shows two PAS-like domains in the NPAS2 protein. The binding residues implicated by the literature are found in one of these domains, and the binding residues from the ProFunc prediction are found in the other. So, whilst the binding site predicted by ProFunc is not consistent with the literature,⁹³ the fact that it is found within a known binding domain is reassuring and suggests that this methodology could be used to identify heme binding motifs.

In the cases of both NPAS2 and ALAS1, the use of ProFunc in combination with AlphaFold has predicted binding sites inconsistent with experimental methods, as shown by the discrepancies with the binding sites given in the literature (Fig. 11B). It is important, however, to consider the reliability of the AlphaFold model, given that ProFunc is searching for binding sites within the AlphaFold model. In these two cases, there is no crystal structure in the PDB to make judgements of the accuracy of the AlphaFold model, however, further studies should investigate other structural prediction tools to assess their use for this purpose.

3.5. Conclusions

The results of this section demonstrate a new methodology for predicting heme binding in proteins, even in cases where a protein does not have a crystal structure available in the PDB. We have shown that AlphaFold and ProFunc can be used in combination, to identify potential heme binding sites. To understand the usefulness of this methodology, the test set of 14 proteins was split into three categories, depending on the characterisation of the heme binding site. The results show that the use of ProFunc and AlphaFold in tandem work well to predict binding residues, consistent with experimental research obtained from the literature, when there is extensive structural information (*i.e.*, a crystal structure of the heme-bound protein or an AlphaFold model with high levels of confidence) for the protein being investigated. This is expected, given that many bioinformatics tools which rely on databases such as the PDB become more accurate as more structures are deposited. In cases where there is no crystal structure for the protein or the AlphaFold model is of a lower confidence, the predicting binding sites in the ProFunc models are less consistent with the literature. Overall, this work is

promising and as it develops, may contribute significantly to our understanding of heme proteins and their binding sites.

(A)

NPAS2_HUMAN	101	TTDGSIIYVSDSITPLLGHLP	SDVMDQNLNLFPEQEHSEVYKILSSHML	150
NPAS2_MOUSE	101	TTDGSIIYVSDSITPLLGHLP	ADVMDQNLNLFPEQEHSEVYKILSSHML	150
NPAS2_HUMAN	151	VTDSPSPEYLKSDSDLEFYCH	LLRGS LNPKEFPTYEYIKFVGNFRSYNNV	200
NPAS2_MOUSE	151	VTDSPSPEFLKSDNDLEFYCH	LLRGS LNPKEFPTYEYIKFVGNFRSYNNV	200
NPAS2_HUMAN	201	PSPSCNGFDNTLSRPCRVP	LGKEVCFIATVRLATPQFLKEMCIVDEPLEE	250
NPAS2_MOUSE	201	PSPSCNGFDNTLSRPFCHV	PLGKDVCFIATVRLATPQFLKEMCVADEPLEE	250
NPAS2_HUMAN	251	FTSRHSLEWKFLFLDHRAP	PIIGYLPFEVLGTSGYDYHIDDELLARCH	300
NPAS2_MOUSE	251	FTSRHSLEWKFLFLDHRAP	PIIGYLPFEVLGTSGYNYHIDDELLARCH	300
NPAS2_HUMAN	301	QHLMQFGKGS CCYRFLTKG	QQWIWLQTHYYITYHQWNSKPEFIVCTHSV	350
NPAS2_MOUSE	301	QHLMQFGKGS CCYRFLTKG	QQWIWLQTHYYITYHQWNSKPEFIVCTHSV	350
NPAS2_HUMAN	351	VSYADVRRERRQELALEDP	PPSEALHSSALKDKGSSLEPRQHFNTLDVGAS	400
NPAS2_MOUSE	351	VSYADVRRERRQELALEDP	PTEAMHPSAVKEKDSLEPPQPFNALDMGAS	400

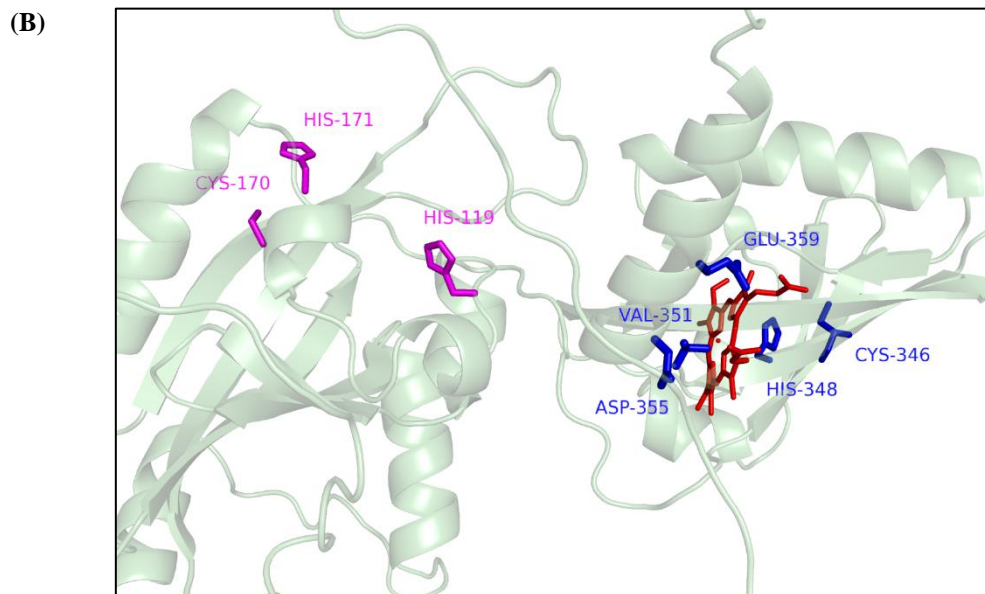


Figure 11. (A) A section (residues 150 - 400) of the pairwise sequence alignment for NPAS2 human and mouse proteins. The residues shown in blue are those identified by using ProFunc and those in pink have been identified from the literature as involved in heme binding. The full alignment is given in Appendix I. (B) The ProFunc prediction for the heme binding site in NPAS2, showing the heme binding residues (blue) within the structure from AlphaFold, which is shown in green (UniProt ID: Q99743). The suspected heme binding residues from the literature are shown (pink) within the same ProFunc model for comparison.⁹³

4. Redox-Sensitive Sensor – Design and Development of the Construct

4.1. Theoretical basis for the redox sensor

The development of a redox-sensitive sensor builds on the mAPXmEGFP sensor developed by Leung *et al.* which was described in Section 1.3.2.⁶³ Instead of the heme binding protein, APX, a sperm whale myoglobin mutant (H64Y/V68F) has been used, abbreviated to gMb. The gMb protein is a double mutant of wild-type sperm whale myoglobin and binds heme with a very high affinity.⁷¹⁻⁷² The mutant has a green colour and a characteristic UV-Visible spectrum.⁷¹⁻⁷² The expression, purification and characterisation of this protein is detailed in Section 4.2. In place of mEGFP, a far-red fluorescent protein, mKate2 has been selected (Section 4.3), such that the sensor can be denoted as gMbmKate2.⁹⁵ The first steps of development of the sensor are given in Section 4.4.

The basis for choosing this recombinant fusion protein results from the different spectral overlap between ferric and ferrous heme (when bound to the sensor) and mKate2 (Fig. 12). The Förster distances for the theoretical resonance energy transfer between gMb, bound to ferric or ferrous heme, and mKate2 were determined to be 3.16 and 2.61 nm respectively. These values were calculated from the absorption spectra of ferric and ferrous gMb, and the relative fluorescence emission spectrum of mKate2 obtained from FPbase (FPbase ID: DBBO8). The absorption spectra were converted to molar absorptivity using Equation 4 (Section 2.6) and the integrands for the spectral overlaps shown in Figure 12 were calculated using Equation 8 as detailed in Appendix II. Subsequently the calculated spectral overlaps were input into Equation 1 (Section 1.3.1) to obtain the Förster distances for mKate2 with ferric and ferrous gMb. The orientation factor was assumed to be 0.7 (simulating freely rotating chromophores),⁹⁶ the quantum yield was taken as 0.4 (as given on FPbase) and the refractive index as 1.33.⁹⁷

Using the calculated Förster distances, the efficiency of resonance energy transfer can be calculated (Eq. 6). Subsequently, the lifetimes of the ferric and ferrous *holo*-proteins at each value for r – where r is the distance between heme (bound to gMb) and the fluorescent protein – in the range 2.1 – 3.4 nm, were calculated (Table 5, Appendix II). The difference in the lifetimes upon binding of ferric and ferrous heme to gMbmKate2 can be used to differentiate between ferric and ferrous heme binding in cells. The calculations in Table 5, Appendix II, show that the difference in lifetimes will be maximised for $r = 2.8$ - 2.9 nm. To accurately differentiate between ferric and ferrous heme, the difference in fluorescence lifetime must be

maximised. Lifetime measurements will be made using FLIM in live cells and the relative concentrations of ferric and ferrous heme can then be calculated using a multiexponential fitting model for the measured mean fluorescence lifetime of the sensor.⁶³

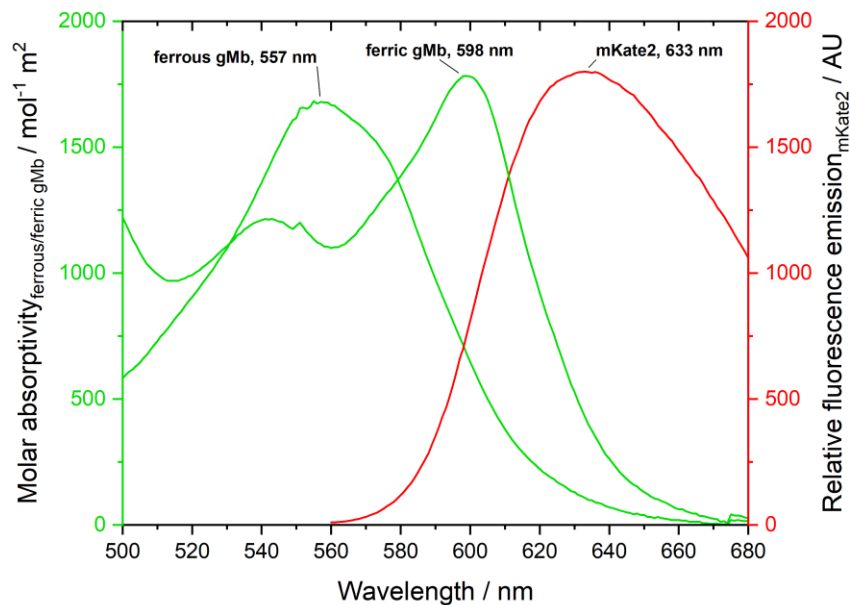
$$E_{ferrous/ferric} = \frac{1}{1 + \left(\frac{r}{R_{0, Fe^{II/III}}} \right)^6} \quad (6)$$

Equation 6. RET efficiency, E , in terms of the distance between fluorescent protein and heme, r , and the Forster distances, R_0 , for ferric (Fe^{III}) or ferrous (Fe^{II}) heme.

$$E_{ferrous/ferric} = 1 - \frac{\tau_{holo Fe^{II/III}}}{\tau_{apo}} \quad (7)$$

Equation 7. RET efficiency, E , in terms of the fluorescent lifetimes, τ , of the *holo*- and *apo*- proteins.

(A)



(B)

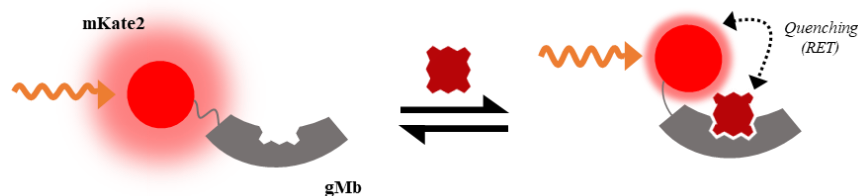


Figure 12. (A) Spectra illustrating the different overlap between the fluorescence emission spectrum of mKate2 (red) and the molar absorptivity of gMb when bound to ferric and ferrous heme (green), which will allow for ferric and ferrous heme to be distinguished in cells. The ferric and ferrous spectra are labelled on the plot and the relevant wavelengths have been identified. (B) The proposed gMbmKate2 sensor. The curly arrow (orange) indicates excitation of the fluorescent protein, mKate2 (bright red circle). Upon heme (dark red) binding to gMb (grey), the fluorescence of the FP is quenched via RET, as shown.

4.2. Sperm whale myoglobin mutant (H64Y/V68F)

4.2.1. Protein expression and purification

The sperm whale myoglobin variant, gMb, has been described in Section 4.1. Before proceeding with construction of the recombinant sensor, the gMb was expressed, purified, and characterised. This protein has been well-characterised and protocols for its expression and purification are available in the literature, which were closely followed.⁷³

The pEMBL19-Mb(H64Y/V68F) plasmid (Appendix III) encoding the gene for the protein was purified from cells which were previously gifted to the Raven laboratory (John Olson, Rice University). The UV-Visible spectrum is given in Figure 13. This was repeated for two separate cell colonies. The concentrations were calculated to be 125 and 155 ng/ μ L from the absorbance at 260 nm. Both samples were at a suitable concentration and purity (no other absorptions in range 200 – 700 nm) so either could have been chosen to proceed with. Sample 1 ($A_{260 \text{ nm}} = 0.063$) was transformed into competent *E. coli* cells (BL21 (DE3)) and the protein was expressed and purified.

Several steps were required to purify the protein, and these have been detailed in the Materials and Methods (Section 2.4.2). The protein does not have a purification tag, however there are detailed protocols in the literature which were used. Size exclusion chromatography was the penultimate step and was followed by ion exchange. Two rounds were required to obtain satisfactory purity and the elution profiles for each of the purifications were obtained (Fig. 14).

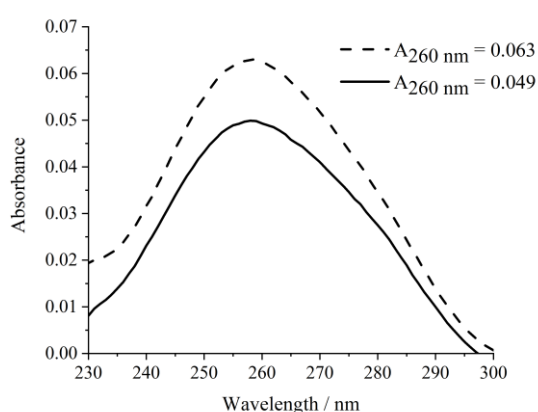


Figure 13. UV-Visible spectrum of pEMBL19-Mb(H64Y/V68F). The absorbance at 260 nm, A_{260} , for the samples were 0.063 and 0.049, and corresponded to concentrations of 155 and 125 ng/ μ L respectively. The sample given by the dotted line was used to transform BL21 (DE3) *E. coli* cells for subsequent expression of the protein.

Fractions containing the target protein were identified by absorption at 412 nm which is attributed to the Soret peak of the protein. In the first elution profile, from the first round of gel filtration, the relative purity of the sample was low. The absorbance due to protein was several times that of the absorbance due to gMb (elution volume 80 mL). There was also an additional absorbance at 412 nm at an elution volume of 40 mL, possibly due to aggregate formation. Protein fractions eluted in the range 70 – 90 mL were identified as containing gMb. These fractions were collected and pooled.

A second round of size exclusion chromatography was applied to the pooled fractions. The subsequent elution profile shows a significantly higher concentration of the desired protein with respect to all other proteins in the sample. A range of samples from the appropriate elution volumes were obtained and run on an SDS-PAGE gel (Figure 15A). The fractions containing gMb as the dominant or only protein were identified by size of the band (17 kDa) as well as the distinct green colour.⁷³ The band due to the presence of gMb has been identified. The fractions were pooled and collected before being purified by ion exchange to obtain the green protein. Finally, SDS-PAGE was used to confirm the purity of the protein (Figure 15B).

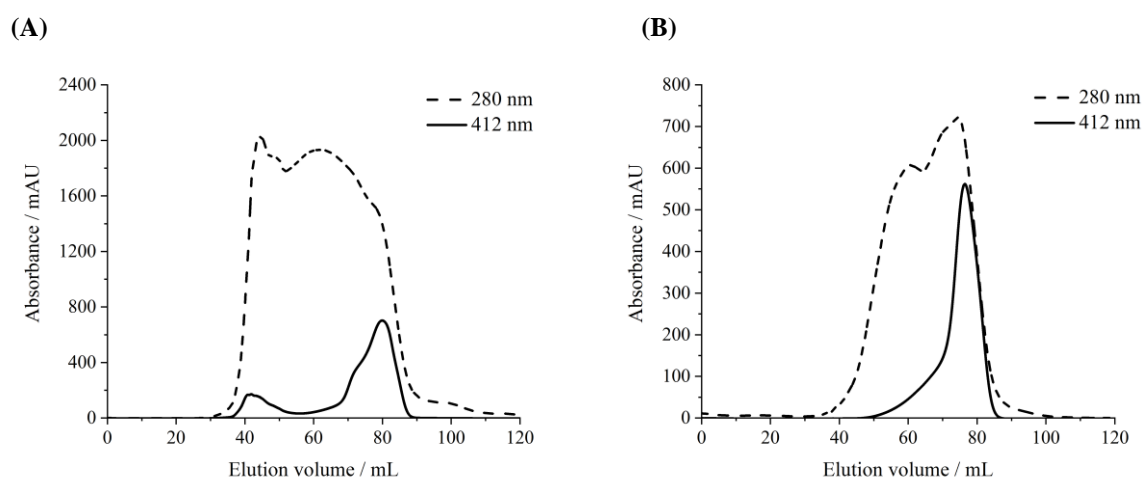


Figure 14. Elution profiles from the size exclusion chromatography during the purification of gMb. The volumes at which the protein fractions were eluted is given against absorptions at 280 and 412 nm. The absorbance at 280 nm gives an indication of the total protein content and the absorbance at 412 nm gives an indication of the gMb content. Overall, this can be used to assess the relative purity of the sample. (A) The first size exclusion chromatography and (B) the second.

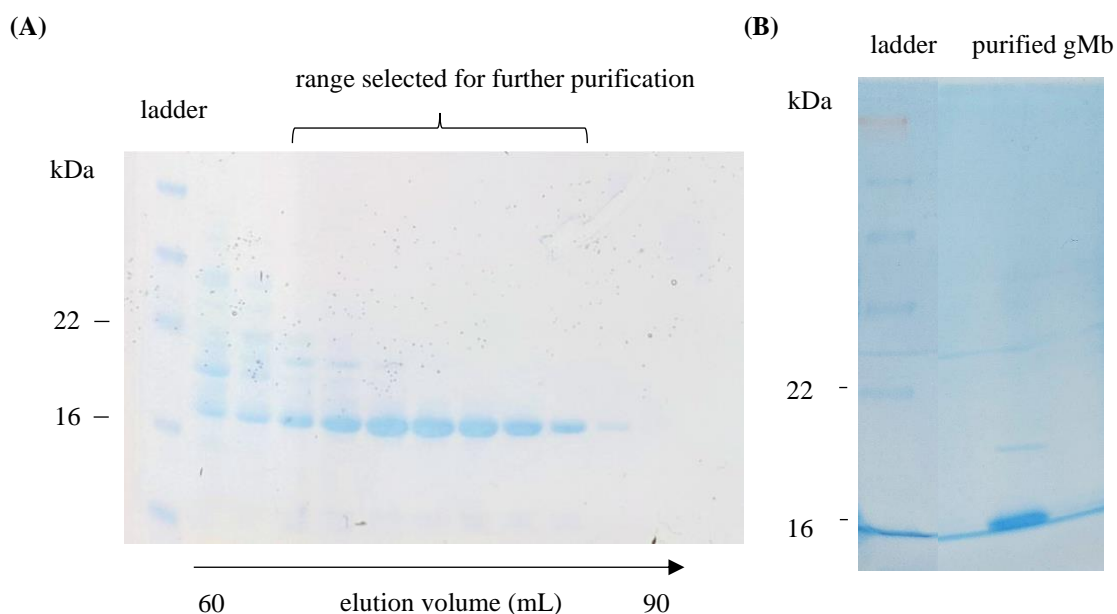


Figure 15. (A) SDS-PAGE gel after the second gel filtration. The range of elution volumes (60 – 90 mL) is indicated on the Figure. Fractions were selected at regular intervals in this range. The range showing the wells selected for further purification is shown. All fractions in this range were collected, pooled, and then purified using ion exchange. (B) SDS-PAGE following ion exchange, the final step of purification, confirming the purity of gMb. The ladder used in both cases was SeeBlue Plus2 (Invitrogen).

4.2.2. Protein characterisation

The purified protein was further characterised using UV-Visible spectroscopy. When the myoglobin mutant (H64Y/V68F) is bound to heme (*i.e.*, the *holo*-protein), it can be characterised by its spectral properties (Fig. 16) in addition to its green colour. It shows absorption at 280 nm due to the presence of aromatic residues in the protein, a Soret peak at 412 nm (with an extinction coefficient, $\epsilon_{412} = 90 \text{ mM}^{-1} \text{ cm}^{-1}$) and a series of Q-bands at 486 nm, 541 nm and 600 nm.⁷¹ The concentration of the protein was calculated using $A_{412 \text{ nm}} = 0.156$ to be 24.8 μM .

Heme can be extracted from myoglobin proteins using methyl ethyl ketone (MEK)⁷⁵ and the protocol followed has been described in Section 2.5 of Materials and Methods. Following extraction of heme from the *holo*-protein, the protein turns from green to colourless. In addition, the characteristic absorptions in the spectrum disappeared, and only a signal at 280 nm ($\epsilon = 15,200 \text{ M}^{-1} \text{ cm}^{-1}$) remained (Fig. 17A). This was used to confirm successful extraction of heme from the protein and to calculate the concentration of the apo protein (36.8 μM). A small Soret peak at 412 nm can still be observed however the relative absorbance, compared to the signal at 280 nm, was low enough to proceed (Fig. 17B). This shows that the method did

not extract heme from 100% of the *holo*-protein.

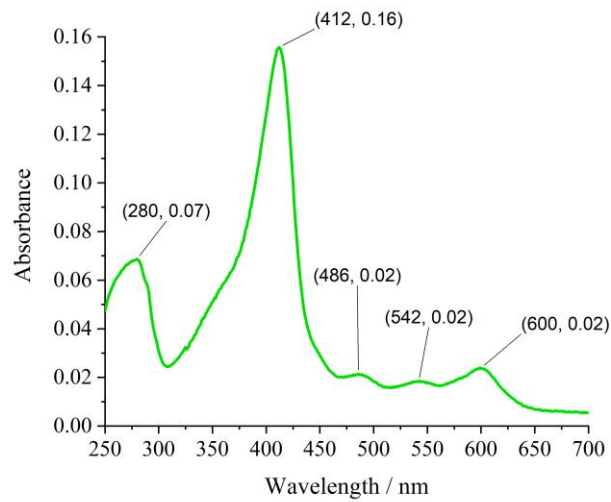


Figure 16. UV-Visible spectrum of gMb following purification. The spectrum shows the characteristic features of the protein, matching previous studies by Olson *et al.*⁷¹⁻⁷²

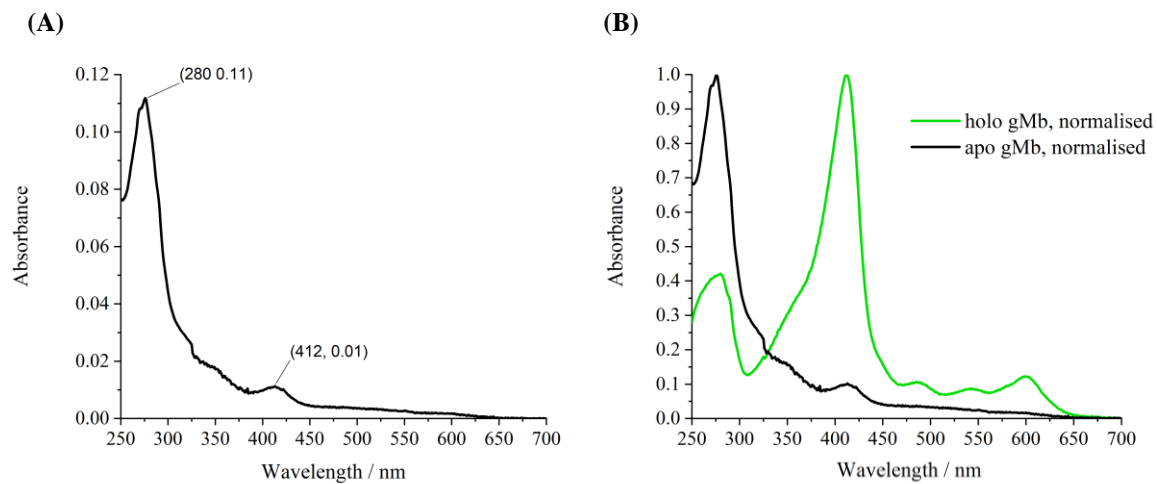


Figure 17. Extraction of heme from gMb to form the *apo*-protein. **(A)** UV-Visible spectrum of *apo*-protein. **(B)** Normalised spectra before and after heme extraction, showing the change in relative intensity of signals at 280 and 412 nm.

4.2.3. Determination of binding affinity using heme titrations

The extraction of heme is very easily reversed by the simple addition of hemin and was used to re-form the *holo*-protein. Hemin was added as $1/10^{\text{th}}$ of an equivalent molar concentration to the protein. There is one binding site for heme in the protein. Following the addition of each aliquot of hemin, the spectrum was recorded (Figure 18A). The Soret peak and Q-bands gradually appeared with each addition. The absorption at 412 nm by gMb is shown to increase with the concentration of hemin and does not plateau. Unlike in other titration experiments, such as difference titrations, the absorbance of hemin is not accounted for in the recorded spectra. Given that hemin shows absorption at 412 nm (Appendix V), each addition will increase the apparent intensity of the Soret peak, even when the *holo*-protein is fully formed.

By the end of the titrations, hemin was in excess, and the spectrum consisted of a mixture of *apo*-protein, *holo*-protein, and free hemin. The titration spectra were deconvoluted into three pure spectra using a software called MCR-ALS (Fig. 18C) and subsequently applied to a fitting program to obtain an estimate of binding affinity of heme for the protein.⁷⁹ Full details have been given in Sections 2.7 and 2.8 and the volumes and concentrations used for the titrations are given in Appendix IV.

The hemin trace in the deconvoluted spectra (Fig. 18C) is characteristic of the UV-Visible spectrum of free hemin (Appendix V) with a broad absorption, peaking at 385 nm. The blue *holo*-protein spectrum corroborates that of purified gMb (Fig. 16) with absorptions at 280, 385, 486, 541 and 600 nm. The relative intensity of absorbance at 280 nm compared to 412 nm is higher in the theoretical spectrum of pure *holo*-protein. The *apo*-protein trace has a dominant absorption at 280 nm, as expected (Fig. 17A), however the Soret peak and Q-bands of the *holo*-protein are present, although at a much lower intensity. Also, the grating observed in Figure 18A has been exaggerated in the theoretical pure *apo*-spectrum. This is likely to result from the data input into the programme. There was a very small absorption at 412 nm in the *apo*-protein (0.011 A.U.) and this has been identified by the software as a significant absorption signal. The deconvoluted spectra were subsequently input into a fitting programme designed in the Hudson group (University of Leicester)⁷⁹, which predicted the binding affinity of heme to gMb as 40 pM. This can be compared to affinity measurements in the literature (10 fM).⁷² Whilst these measurements are not consistent, they are both very high affinities associated with permanent binding.

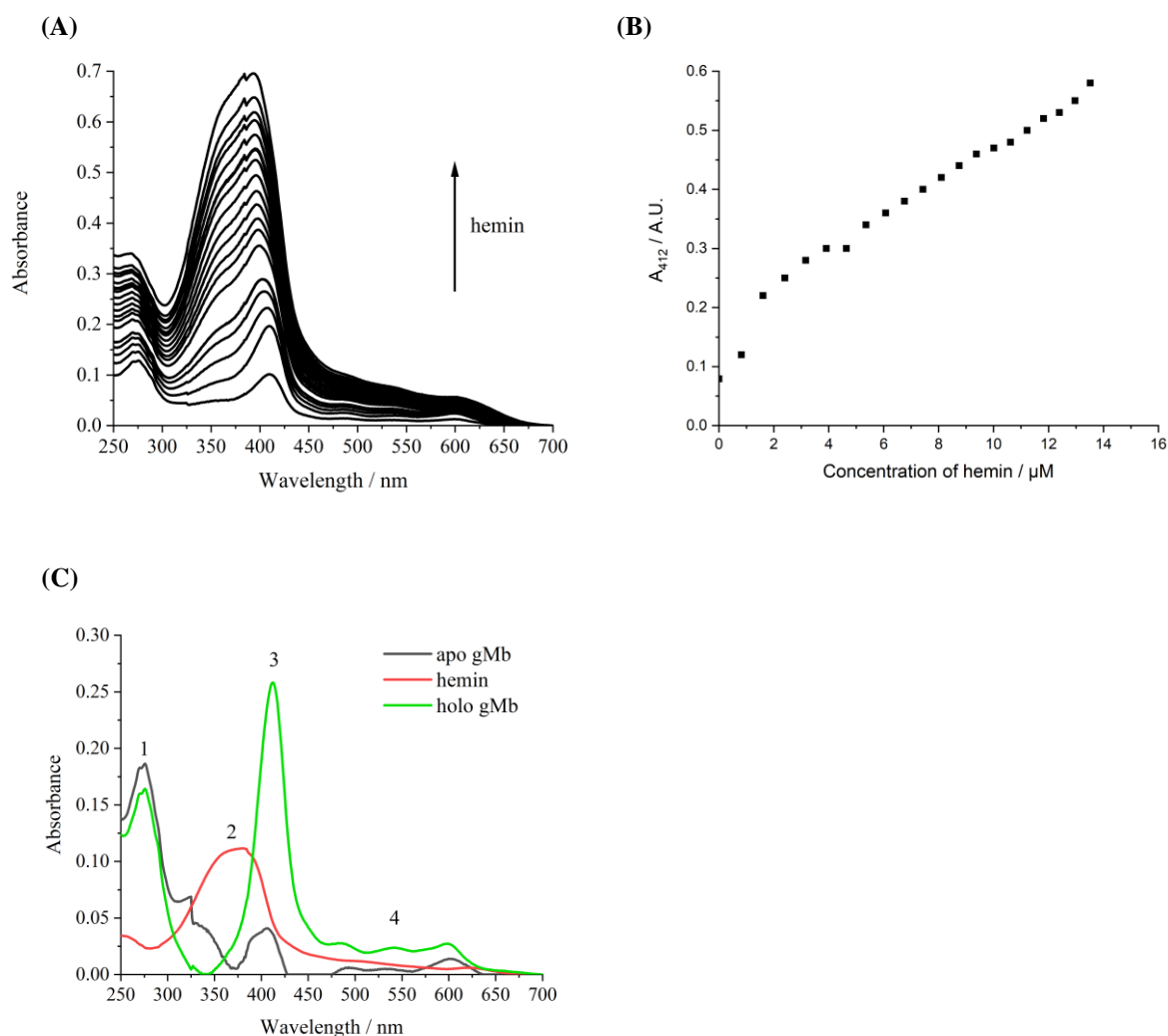


Figure 18. UV-Visible spectra showing hemin titration into gMb and deconvolution of the spectra. **(A)** Titration of hemin into the apo-protein. The stock concentration of hemin used in this experiment was $75 \mu\text{M}$ and the concentration of hemin in the cuvette was $0.82 - 13.52 \mu\text{M}$ (Appendix IV). The initial *apo*-protein concentration was $8.2 \mu\text{M}$. With each addition, the absorption of the Soret peak (due to formation of the holo-protein) increased, as did the absorption at 385 nm due to free heme concentration increasing. A small increase in absorption at 280 nm was also observed. The data has been adjusted such that the absorbances at 700 nm are zero; this modification is required for deconvolution of the spectra. Also, note there is a slight grating in the spectra because of the instrument used. **(B)** Plot showing the increase in $A_{412 \text{ nm}}$ with increasing hemin concentration. Each square point on the plot is the absorbance recorded at 412 nm following a hemin addition. **(C)** Deconvolution of spectrum using MCR-ALS programme into *apo*-protein (black line), *holo*-protein (green line) and hemin, or free heme (red line). Important regions of the spectra have been labelled as follows: the general protein absorption at 280 nm (1), free heme absorption at 385 nm (2), Soret peak of gMb at 412 nm (3), and Q-bands from gMb (4).

4.3. Red fluorescent protein - mKate2

The plasmid TOPO mKate2 (Addgene plasmid #68441) (Appendix VI) containing the gene for the far-red fluorescent protein mKate2⁹⁵ was purchased as an agar stab and extracted from cells. Five samples of plasmids were extracted from five individual minipreps and the corresponding concentrations calculated for each sample was calculated as given in the Figure. The plasmid was checked using restriction digestions (Fig. 19A). The expected size of the plasmid is 4.3 kb and digested bands appear between 4 and 5 kb.

For each plasmid, the undigested sample and singly digested sample are shown alongside. The plasmids were digested with NotI restriction enzyme. Upon single digestion, the bands appear sharper, and the size of the DNA can be more easily identified. The size was confirmed in the expected range 4 – 5 kb for all the bands. There are no other identifiable bands from the digested samples, confirming the purity. Any of the four samples could be chosen, and Sample 4 was selected to proceed with.

Sample 4 was digested again to further confirm the identity of the DNA (Fig. 19B). This was important to confirm the identity of the plasmid before proceeding with using it to obtain the sensor construct. As before it was digested at the unique site, NotI. It was also digested at another unique site, HindIII and doubly digested with both restriction enzymes. The single digestions confirm the presence of those unique sites, as the bands appear sharper compared to the undigested well. The double digestion confirms that the sites are at the correct distances from one other and the single digestions confirm the overall size of the plasmid at 4.3 kb. A slight shift from 4.3 kb to a lighter mass was apparent following double digestion, confirming that the plasmid was cut at two positions. The two restriction sites are 843 bp apart (Appendix VI), and this is consistent with the agarose gel (Fig. 19B). This was sufficient to confirm the identity of the plasmid to proceed with cloning.

Expression of the red fluorescent protein, mKate2 was attempted directly from the TOPO mKate2 plasmid. The expression was induced using IPTG (*lac* operon controlled) under various conditions but there were very low expression levels. The mKate2 gene will be cloned into a different plasmid (used previously by members of the Raven group) and expression will be reattempted.

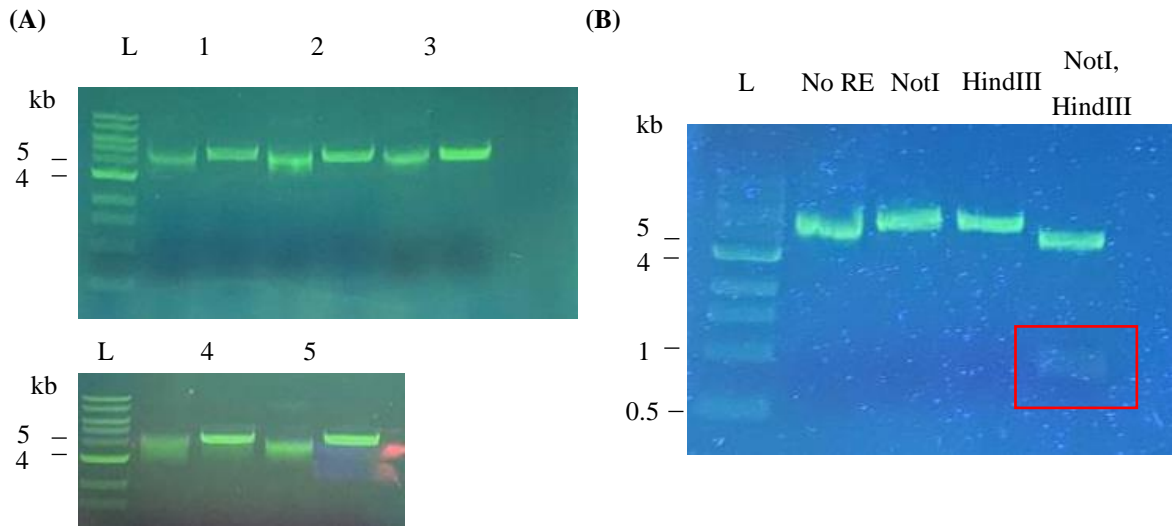


Figure 19. (A) DNA agarose gel of five samples (1 – 5) of TOPO mKate2 extracted from cells with concentrations 300, 500, 450, 350 and 350 ng/ μ L respectively. The undigested DNA was run alongside the singly digested samples (restriction enzyme NotI) to check the size of the plasmid. (B) Agarose gel showing single digestions (NotI and HindIII) and double digestion (NotI/HindIII) of the TOPO mKate2 plasmid, alongside undigested (No RE) (Sample 4). The small fragment (< 1 kb) confirms the identity of the TOPO mKate2 plasmid and is indicated by the red box. The ladder, L, used for both gels was Quick-Load Purple 1 kb DNA Ladder (N0552).

4.4. Constructing the redox sensor

For construction of the sensor, the gene for gMb was cloned into the TOPO mKate2 vector using restriction digestion and ligation cloning. A summary of the cloning protocol is shown in Figure 20.

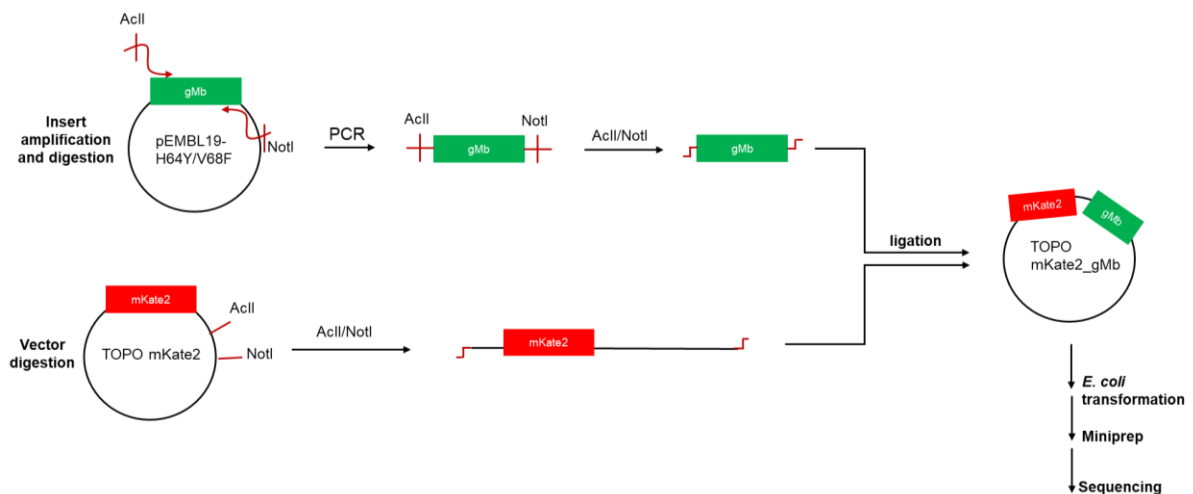


Figure 20. Cloning protocol for the insertion of the gMb gene into TOPO mKate2 for the formation of the gMbmKate2 sensor.

First, the gMb gene was amplified using PCR. The gene was amplified with primers containing restriction sites for AclI and NotI. Details of the primers is given in Appendix VI. The primers were designed such that gMb has a stop codon but no start codon. Six tubes were prepared identically, and different annealing temperatures were used. The full detail of the protocol is given in Section 2.11.1. The PCR products were run on an agarose gel for analysis. The PCR reaction was successful for five of the six samples (Fig. 21A). The sample with the highest annealing temperature (69.5 °C) was used to proceed with cloning. This is because a higher annealing temperature corresponds to highest specificity of primer binding and thus is associated with the highest purity.

The DNA was digested with the appropriate restriction enzymes to produce sticky ends. The TOPO mKate2 vector was digested with the same restriction enzymes to produce complementary sticky ends. The digested products were then ligated and directly transformed into *E. coli* cells. Following a successful transformation, a single colony was selected. The colony was used for DNA preparation and the plasmid was purified and run on an agarose gel (Fig. 21B). The agarose gel shows the size of the plasmid undigested, with two separate single digestions (5 kb in both cases) and a double digestion (4.5 kb and 0.5 kb). The release of the 0.5 kb insert, consistent with the size of gMb (Fig. 22), confirms successful cloning. Further verification was obtained via Sanger sequencing of the cloned region. This was aligned with known sequence of the vector backbone to obtain the full sequence (Fig. 22).

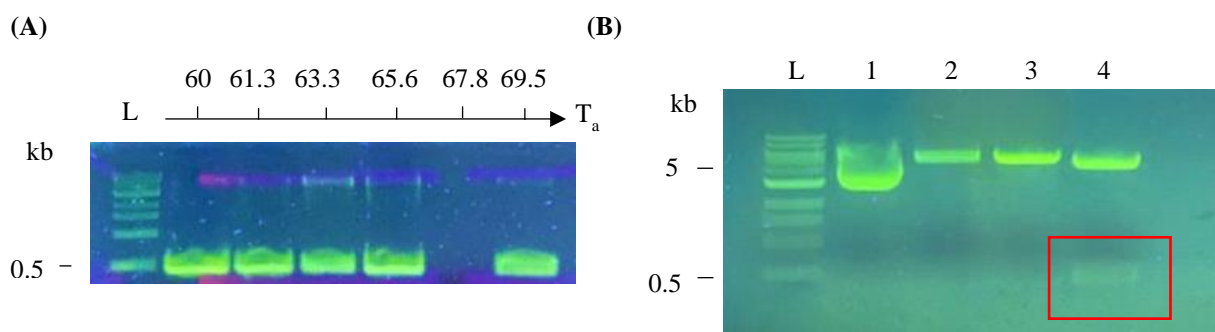


Figure 21. (A) Agarose gel of PCR amplification of gMb. Six reaction tubes were run at different annealing temperatures. The PCR protocol is shown on the left. Size of gMb gene is 468 bp. The PCR product with the highest annealing temperature was excised from the gel and purified, digested, and ligated with TOPO mKate2. (B) Agarose gel confirming successful ligation. (1) undigested product (2) single digestion with NotI (3) single digestion with AclI (4) double digestion NotI/AclI to release gMb insert (shown in red box). The ladder, L, used for both gels was Quick-Load Purple 1 kb DNA Ladder (N0552).

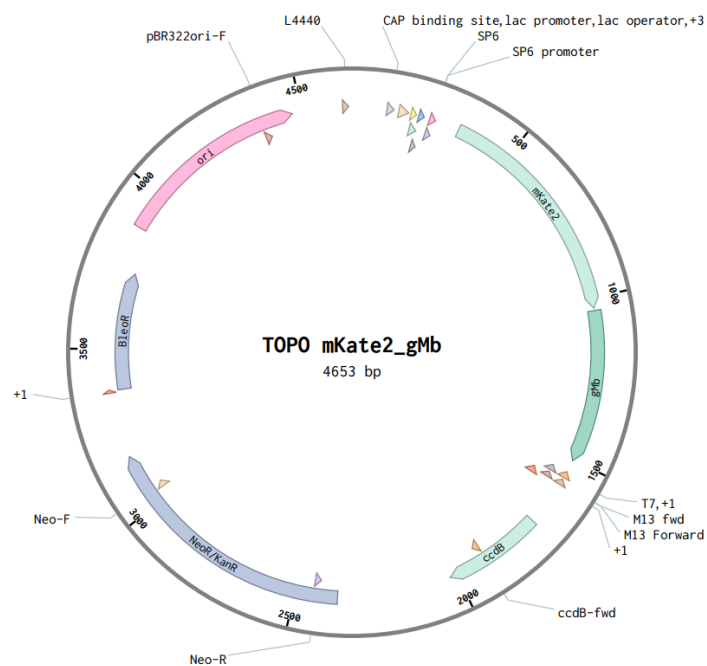


Figure 22. Plasmid map of gMbmKate2. Sanger sequencing was used to confirm successful cloning (Appendix VIII).

The sequencing results show that the genes for mKate2 and gMb are not in frame, *i.e.* there are four bases between the stop codon of mKate2 and the beginning of the gMb sequence. This means that the gMb gene will not be translated with the correct amino acid sequence. The number of bases between the two genes must be a multiple of three in accordance with the triplet code of DNA. This can be achieved by the deletion of one base, for example. Additionally, a stop codon (TGA) is still present in the mKate2 gene; this must be removed such that translation does not end after mKate2 without translating the gMb gene. These problems may be solved simultaneously, by creating primers for amplification of the entire plasmid, excluding the mKate2 stop codon and ensuring the genes are in frame.

4.5. Conclusions

The myoglobin mutant (H64Y/V68F), gMb, was successfully purified and expressed; however further studies are required to confirm the binding affinity of heme to the protein. The heme titration measurements were recorded under aerobic conditions, such that ferric heme (oxidised, Fe^{3+}) binding to gMb was being measured. The next step would be to quantify ferrous heme (reduced, Fe^{2+}) binding to gMb to quantify the difference in binding affinity between the two oxidation states. These measurements are important in the characterisation of the difference in heme binding to myoglobin and will be required in the characterisation of fluorescent lifetime measurements when using the redox sensor. Additionally, given the

reducing cellular environment, measurement of binding of ferrous heme may reflect the physiological environment of the cell more accurately.

The theoretical basis for using gMbmKate2 as a heme sensor has been established (Section 4.1). The next steps will be to reattempt expression of mKate2, possibly by cloning the gene into another plasmid, for which mKate2 expression is better characterised *i.e.* there are working protocols are available. Once the sequence of the sensor construct is edited slightly (Section 4.4) to ensure the recombinant gMbmKate2 protein will be translated as required, it can be expression and characterised in *E. coli*. The construct will then be cloned into a plasmid for mammalian expression. This will enable the quantification of heme concentrations *in vivo*, using the work done for the mAPXmEGFP sensor as a guide, including fluorescence lifetime intensity measurements (FLIM).⁶³

5. Perspectives and Future Work

This thesis has discussed two projects, both of which will ultimately contribute to understanding of the complex dynamics of heme in the cell. The bioinformatics work presented in Section 3 is the start of an exploration into the techniques available for predicting heme binding and combines an older tool, ProFunc, with AlphaFold, which is newer and uses complex machine learning to predict the 3D structure of proteins from amino acid sequence alone. We have shown that tools such as ProFunc can facilitate experimental research into heme binding by identifying structural motifs, and that AlphaFold enables these tools to be used when there is no crystal structure of the target protein available.

The next steps for this work will be to further assess and quantify the predicted heme binding sites further. This includes considerations of steric clashes and hydrophobicity. Techniques such as protein-ligand docking, or energy minimisations can be used to optimise the predicted heme binding sites, and there are also other tools that can be used in order predict possible binding residues (Table 1). The most recent heme-specific tool is HeMoQuest, which aims to predict transient, low-affinity binding to proteins. This tool aims to predict transient binding to heme, and thus may be better suited to identifying regulatory heme proteins such as GAPDH (Section 1.2.2). Overall, a combination of these tools, along with traditional methods (Section 1.2.3) will facilitate progress into the identification of novel heme binding sites in the future.

In Section 4, the first steps of the development of a redox-sensitive sensor were described. There is currently no sensor that can differentiate between ferric and ferrous heme in live cells, and this will be an important step forward for the heme community. The myoglobin mutant (H64Y/V68F), which will be used in this sensor, has been successfully expressed, purified, and characterised following protocols in the literature and other members of the Raven group. The novel sensor construct has been successfully cloned using restriction-digestion techniques. The next step will be to perform extra cloning work to ensure the genes for both the mKate2 and gMb are in frame, and then the fusion protein can be expressed in *E. coli*. Following on from this, the sensor will be cloned into a vector for mammalian expression and will be characterised using FLIM experiments during my PhD in the Raven group. Overall, these projects present a further exploration into the complex dynamics of heme in the cell, focussing on heme binding, and provide the first steps into developing a novel method of quantifying heme in the cell.

References

1. Gallio, A. E.; Fung, S. S. P.; Cammack-Najera, A.; Hudson, A. J.; Raven, E. L., Understanding the Logistics for the Distribution of Heme in Cells. *JACS Au* **2021**, *1* (10), 1541-1555.
2. Shimizu, T.; Lengalova, A.; Martínek, V.; Martínková, M., Heme: emergent roles of heme in signal transduction, functional regulation and as catalytic centres. *Chemical Society Reviews* **2019**, *48* (24), 5624-5657.
3. Fischer, H., On haemin and the relationships between haemin and chlorophyll. 1930.
4. Wang, J.; Niemevz, F.; Lad, L.; Huang, L.; Alvarez, D. E.; Buldain, G.; Poulos, T. L.; De Montellano, P. R. O., Human Heme Oxygenase Oxidation of 5- and 15-Phenylhemes. *Journal of Biological Chemistry* **2004**, *279* (41), 42593-42604.
5. Edsall, J. T., Blood and hemoglobin: The evolution of knowledge of functional adaptation in a biochemical system. *Journal of the History of Biology* **1972**, *5* (2), 205-257.
6. Smith, L. J.; Kahraman, A.; Thornton, J. M., Heme proteins-Diversity in structural characteristics, function, and folding. *Proteins: Structure, Function, and Bioinformatics* **2010**, *78* (10), 2349-2368.
7. Roumenina, L. T.; Rayes, J.; Lacroix-Desmazes, S.; Dimitrov, J. D., Heme: Modulator of Plasma Systems in Hemolytic Diseases. *Trends in Molecular Medicine* **2016**, *22* (3), 200-213.
8. Barupala, D. P.; Dzul, S. P.; Riggs-Gelasco, P. J.; Stemmler, T. L., Synthesis, delivery and regulation of eukaryotic heme and Fe-S cluster cofactors. *Archives of Biochemistry and Biophysics* **2016**, *592*, 60-75.
9. Perutz, M. F.; Kendrew, J. C.; Watson, H. C., Structure and function of haemoglobin. *Journal of Molecular Biology* **1965**, *13* (3), 669-678.
10. Bilska-Wilkosz, A.; Iciek, M.; Górny, M.; Kowalczyk-Pachel, D., The Role of Hemoproteins: Hemoglobin, Myoglobin and Neuroglobin in Endogenous Thiosulfate Production Processes. *International Journal of Molecular Sciences* **2017**, *18* (6), 1315.
11. Lin, Y.-W.; Wang, J., Structure and function of heme proteins in non-native states: A mini-review. *Journal of Inorganic Biochemistry* **2013**, *129*, 162-171.
12. Chakravarti, R.; Aulak, K. S.; Fox, P. L.; Stuehr, D. J., GAPDH regulates cellular heme insertion into inducible nitric oxide synthase. *Proceedings of the National Academy of Sciences* **2010**, *107* (42), 18004-18009.
13. Kühn, T.; Imhof, D., Regulatory Fe^{II/III}Heme: The Reconstruction of a Molecule's Biography. *ChemBioChem* **2014**, *15* (14), 2024-2035.
14. Dai, Y.; Sweeny, E. A.; Schlanger, S.; Ghosh, A.; Stuehr, D. J., GAPDH delivers heme to soluble guanylyl cyclase. *Journal of Biological Chemistry* **2020**, *295* (24), 8145-8154.
15. Wißbrock, A.; Paul George, A. A.; Hans; Kühn, T.; Imhof, D., The molecular basis of transient heme-protein interactions: analysis, concept and implementation. *Bioscience Reports* **2019**, *39* (1).

16. Hanna, D. A.; Martinez-Guzman, O.; Reddi, A. R., Heme Gazing: Illuminating Eukaryotic Heme Trafficking, Dynamics, and Signaling with Fluorescent Heme Sensors. *Biochemistry* **2017**, *56* (13), 1815-1823.
17. Reddi, A. R.; Hamza, I., Heme Mobilization in Animals: A Metallolipid's Journey. *Acc Chem Res* **2016**, *49* (6), 1104-10.
18. Burton, M. J.; Cresser-Brown, J.; Thomas, M.; Portolano, N.; Basran, J.; Freeman, S. L.; Kwon, H.; Bottrill, A. R.; Llansola-Portoles, M. J.; Pascal, A. A.; Jukes-Jones, R.; Chernova, T.; Schmid, R.; Davies, N. W.; Storey, N. M.; Dorlet, P.; Moody, P. C. E.; Mitcheson, J. S.; Raven, E. L., Discovery of a heme-binding domain in a neuronal voltage-gated potassium channel. *Journal of Biological Chemistry* **2020**, *295* (38), 13277-13286.
19. Shimizu, T.; Huang, D.; Yan, F.; Stranova, M.; Bartosova, M.; Fojtíková, V.; Martínková, M., Gaseous O₂, NO, and CO in Signal Transduction: Structure and Function Relationships of Heme-Based Gas Sensors and Heme-Redox Sensors. *Chemical Reviews* **2015**, *115* (13), 6491-6533.
20. Burton, M. J.; Kapetanaki, S. M.; Chernova, T.; Jamieson, A. G.; Dorlet, P.; Santolini, J.; Moody, P. C. E.; Mitcheson, J. S.; Davies, N. W.; Schmid, R.; Raven, E. L.; Storey, N. M., A heme-binding domain controls regulation of ATP-dependent potassium channels. *Proceedings of the National Academy of Sciences* **2016**, *113* (14), 3785-3790.
21. Kapetanaki, S. M.; Burton, M. J.; Basran, J.; Uragami, C.; Moody, P. C. E.; Mitcheson, J. S.; Schmid, R.; Davies, N. W.; Dorlet, P.; Vos, M. H.; Storey, N. M.; Raven, E., A mechanism for CO regulation of ion channels. *Nature Communications* **2018**, *9* (1).
22. Minegishi, S.; Sagami, I.; Negi, S.; Kano, K.; Kitagishi, H., Circadian clock disruption by selective removal of endogenous carbon monoxide. *Scientific Reports* **2018**, *8* (1).
23. Freeman, S. L.; Kwon, H.; Portolano, N.; Parkin, G.; Venkatraman Girija, U.; Basran, J.; Fielding, A. J.; Fairall, L.; Svistunenko, D. A.; Moody, P. C. E.; Schwabe, J. W. R.; Kyriacou, C. P.; Raven, E. L., Heme binding to human CLOCK affects interactions with the E-box. *Proceedings of the National Academy of Sciences* **2019**, *116* (40), 19911-19916.
24. Airola, M. V.; Du, J.; Dawson, J. H.; Crane, B. R., Heme Binding to the Mammalian Circadian Clock Protein Period 2 Is Nonspecific. *Biochemistry* **2010**, *49* (20), 4327-4338.
25. Raghuram, S.; Stayrook, K. R.; Huang, P.; Rogers, P. M.; Nosie, A. K.; McClure, D. B.; Burris, L. L.; Khorasanizadeh, S.; Burris, T. P.; Rastinejad, F., Identification of heme as the ligand for the orphan nuclear receptors REV-ERB α and REV-ERB β . *Nature Structural & Molecular Biology* **2007**, *14* (12), 1207-1213.
26. Small, S. K.; Puri, S.; O'Brian, M. R., Heme-dependent metalloregulation by the iron response regulator (Irr) protein in *Rhizobium* and other Alpha-proteobacteria. *BioMetals* **2009**, *22* (1), 89-97.
27. Girvan, H. M.; Munro, A. W., Heme sensor proteins. *J Biol Chem* **2013**, *288* (19), 13194-203.

28. Yuan, X.; Rietzschel, N.; Kwon, H.; Nuno, A. B. W.; Hanna, D. A.; Phillips, J. D.; Raven, E. L.; Reddi, A. R.; Hamza, I., Regulation of intracellular heme trafficking revealed by subcellular reporters. *Proceedings of the National Academy of Sciences* **2016**, *113* (35), E5144-E5152.
29. Abshire, J. R.; Rowlands, C. J.; Ganesan, S. M.; So, P. T. C.; Niles, J. C., Quantification of labile heme in live malaria parasites using a genetically encoded biosensor. *Proceedings of the National Academy of Sciences* **2017**, *114* (11), E2068-E2076.
30. Tristan, C.; Shahani, N.; Sedlak, T. W.; Sawa, A., The diverse functions of GAPDH: Views from different subcellular compartments. *Cellular Signalling* **2011**, *23* (2), 317-323.
31. Hannibal, L.; Collins, D.; Brassard, J.; Chakravarti, R.; Vempati, R.; Dorlet, P.; Santolini, J.; Dawson, J. H.; Stuehr, D. J., Heme Binding Properties of Glyceraldehyde-3-phosphate Dehydrogenase. *Biochemistry* **2012**, *51* (43), 8514-8529.
32. Sweeny, E. A.; Singh, A. B.; Chakravarti, R.; Martinez-Guzman, O.; Saini, A.; Haque, M. M.; Garee, G.; Dans, P. D.; Hannibal, L.; Reddi, A. R.; Stuehr, D. J., Glyceraldehyde-3-phosphate dehydrogenase is a chaperone that allocates labile heme in cells. *Journal of Biological Chemistry* **2018**, *293* (37), 14557-14568.
33. Ghosh, A.; Chawla-Sarkar, M.; Stuehr, D. J., Hsp90 interacts with inducible NO synthase client protein in its heme-free state and then drives heme insertion by an ATP-dependent process. *The FASEB Journal* **2011**, *25* (6), 2049-2060.
34. Paul George, A. A.; Lacerda, M.; Syllwasschy, B. F.; Hopp, M.-T.; Wißbrock, A.; Imhof, D., HeMoQuest: a webserver for qualitative prediction of transient heme binding to protein motifs. *BMC Bioinformatics* **2020**, *21* (1), 124.
35. Vasiliki-Dimitra C Tsolaki, S. K. G.-S., Athina I Tsamadou, Stefanos A Tsiftoglou, Martina Samiotaki, George Panayotou, Asterios S Tsiftoglou, Hemin accumulation and identification of a heme-binding protein clan in K562 cells by proteomic and computational analysis. *J Cell Physiol.* **2021**.
36. Lukacik, P.; Owen, C. D.; Harris, G.; Bolla, J. R.; Picaud, S.; Alibay, I.; Nettleship, J. E.; Bird, L. E.; Owens, R. J.; Biggin, P. C.; Filippakopoulos, P.; Robinson, C. V.; Walsh, M. A., The structure of nontypeable *Haemophilus influenzae* SapA in a closed conformation reveals a constricted ligand-binding cavity and a novel RNA binding motif. *PLOS ONE* **2021**, *16* (10), e0256070.
37. Che, S.; Liang, Y.; Chen, Y.; Wu, W.; Liu, R.; Zhang, Q.; Bartlam, M., Structure of *Pseudomonas aeruginosa* spermidine dehydrogenase: a polyamine oxidase with a novel heme-binding fold. *The FEBS Journal* **2021**.
38. Dawson, J. H., Probing structure-function relations in heme-containing oxygenases and peroxidases. *Science* **1988**, *240* (4851), 433-9.
39. Culbertson, D. S.; Olson, J. S., Role of heme in the unfolding and assembly of myoglobin. *Biochemistry* **2010**, *49* (29), 6052-63.

40. Fleischhacker, A. S.; Sarkar, A.; Liu, L.; Ragsdale, S. W., Regulation of protein function and degradation by heme, heme responsive motifs, and CO. *Critical Reviews in Biochemistry and Molecular Biology* **2022**, *57* (1), 16-47.
41. Mauk, A. G.; Moore, G. R., Control of metalloprotein redox potentials: what does site-directed mutagenesis of hemoproteins tell us? *JBIC Journal of Biological Inorganic Chemistry* **1997**, *2* (1), 119-125.
42. Liou, Y.-F.; Charoenkwan, P.; Srinivasulu, Y.; Vasylenko, T.; Lai, S.-C.; Lee, H.-C.; Chen, Y.-H.; Huang, H.-L.; Ho, S.-Y., SCMHBP: prediction and analysis of heme binding proteins using propensity scores of dipeptides. *BMC Bioinformatics* **2014**, *15* (Suppl 16), S4.
43. Liu, R.; Hu, J., HemeBIND: a novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinformatics* **2011**, *12* (1), 207.
44. Zhang, J.; Chai, H.; Gao, B.; Yang, G.; Ma, Z., HEMEsPred: Structure-Based Ligand-Specific Heme Binding Residues Prediction by Using Fast-Adaptive Ensemble Learning Scheme. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2018**, *15* (1), 147-156.
45. Zhao, J.; Cao, Y.; Zhang, L., Exploring the computational methods for protein-ligand binding site prediction. *Computational and Structural Biotechnology Journal* **2020**, *18*, 417-426.
46. Liu, R.; Hu, J., Computational Prediction of Heme-Binding Residues by Exploiting Residue Interaction Network. *PLoS ONE* **2011**, *6* (10), e25560.
47. Yu, D.-J.; Hu, J.; Yang, J.; Shen, H.-B.; Tang, J.; Yang, J.-Y., Designing Template-Free Predictor for Targeting Protein-Ligand Binding Sites with Classifier Ensemble and Spatial Clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2013**, *10* (4), 994-1008.
48. Yang, J.; Roy, A.; Zhang, Y., BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Research* **2012**, *41* (D1), D1096-D1103.
49. Laskowski, R. A.; Watson, J. D.; Thornton, J. M., ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Research* **2005**, *33* (Web Server), W89-W93.
50. Berman, H. M., The Protein Data Bank. *Nucleic Acids Research* **2000**, *28* (1), 235-242.
51. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **2019**, *47* (D1), D506-D515.
52. Blum, M.; Chang, H.-Y.; Chuguransky, S.; Grego, T.; Kandasamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; Richardson, L.; Salazar, G. A.; Williams, L.; Bork, P.; Bridge, A.; Gough, J.; Haft, D. H.; Letunic, I.; Marchler-Bauer, A.; Mi, H.; Natale, D. A.; Necci, M.; Orengo, C. A.; Pandurangan, A. P.; Rivoire, C.; Sigrist, C. J. A.; Sillitoe, I.; Thanki, N.; Thomas, P. D.; Tosatto, S. C. E.; Wu, C. H.; Bateman, A.; Finn, R. D., The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* **2020**, *49* (D1), D344-D354.
53. Wallace, A. C.; Laskowski, R. A.; Thornton, J. M., Derivation of 3D coordinate templates for searching structural databases: Application to ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Science* **1996**, *5* (6), 1001-1013.

54. Schrodinger, LLC, The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8. 2015.
55. Jones, D. T.; Thornton, J. M., The impact of AlphaFold2 one year on. *Nature Methods* **2022**, *19* (1), 15-20.
56. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583-589.
57. Thornton, J. M.; Laskowski, R. A.; Borkakoti, N., AlphaFold heralds a data-driven revolution in biology and medicine. *Nature Medicine* **2021**, *27* (10), 1666-1669.
58. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Zidek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S., AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **2022**, *50* (D1), D439-D444.
59. Morrison, G. R., Fluorometric Microdetermination of Heme Protein. *Analytical Chemistry* **1965**, *37* (9), 1124-1126.
60. Andrews, D. L.; Bradshaw, D. S., Resonance Energy Transfer. *digital Encyclopedia of Applied Physics* **2009**, 533-554.
61. Lakowicz, J. R., Principles of Fluorescence Spectroscopy. 2006; Vol. 1.
62. Jones, G. A.; Bradshaw, D. S., Resonance Energy Transfer: From Fundamental Theory to Recent Applications. *Frontiers in Physics* **2019**, *7* (100).
63. Leung, G. C. H.; Fung, S. S. P.; Gallio, A. E.; Blore, R.; Alibhai, D.; Raven, E. L.; Hudson, A. J., Unravelling the mechanisms controlling heme supply and demand. *Proceedings of the National Academy of Sciences* **2021**, *118* (22), e2104008118.
64. Takeda, S.; Kamiya, N.; Arai, R.; Nagamune, T., Design of an Artificial Light-Harvesting Unit by Protein Engineering: Cytochrome *b*₅₆₂-Green Fluorescent Protein Chimera. *Biochemical and Biophysical Research Communications* **2001**, *289* (1), 299-304.
65. Takeda, S.; Kamiya, N.; Nagamune, T., A novel protein-based heme sensor consisting of green fluorescent protein and apocytochrome *b*₅₆₂. *Analytical Biochemistry* **2003**, *317* (1), 116-119.
66. Arpino, J. A. J.; Czapinska, H.; Piasecka, A.; Edwards, W. R.; Barker, P.; Gajda, M. J.; Bochtler, M.; Jones, D. D., Structural Basis for Efficient Chromophore Communication and Energy Transfer in a Constructed Didomain Protein Scaffold. *Journal of the American Chemical Society* **2012**, *134* (33), 13632-13640.

67. Hanna, D. A.; Harvey, R. M.; Martinez-Guzman, O.; Yuan, X.; Chandrasekharan, B.; Raju, G.; Outten, F. W.; Hamza, I.; Reddi, A. R., Heme dynamics and trafficking factors revealed by genetically encoded fluorescent heme sensors. *Proceedings of the National Academy of Sciences* **2016**, *113* (27), 7539-7544.
68. Hanna, D. A.; Hu, R.; Kim, H.; Martinez-Guzman, O.; Torres, M. P.; Reddi, A. R., Heme bioavailability and signaling in response to stress in yeast cells. *J Biol Chem* **2018**, *293* (32), 12378-12393.
69. Song, Y.; Yang, M.; Wegner, S. V.; Zhao, J.; Zhu, R.; Wu, Y.; He, C.; Chen, P. R., A genetically encoded FRET sensor for intracellular heme. *ACS chemical biology* **2015**, *10* (7), 1610-1615.
70. Madeira, F.; Pearce, M.; Tivey, A. R. N.; Basutkar, P.; Lee, J.; Edbali, O.; Madhusoodanan, N.; Kolesnikov, A.; Lopez, R., Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research* **2022**, *50* (W1), W276-W279.
71. Hargrove, M. S.; Singleton, E. W.; Quillin, M. L.; Ortiz, L. A.; Phillips, G. N.; Olson, J. S.; Mathews, A. J., His64(E7)→Tyr apomyoglobin as a reagent for measuring rates of heme dissociation. *Journal of Biological Chemistry* **1994**, *269* (6), 4207-4214.
72. Hargrove, M. S.; Barrick, D.; Olson, J. S., The Association Rate Constant for Heme Binding to Globin Is Independent of Protein Structure. *Biochemistry* **1996**, *35* (35), 11293-11299.
73. Springer, B. A.; Sligar, S. G., High-level expression of sperm whale myoglobin in *Escherichia coli*. *Proceedings of the National Academy of Sciences* **1987**, *84* (24), 8961-8965.
74. Chung, Y.-B.; Yang, H.-J.; Hong, S.-J.; Kang, S.-Y.; Lee, M.; Kim, T. Y.; Choi, M.-H.; Chai, J.-Y.; Hong, S.-T., Molecular cloning and immunolocalization of the 17 kDa myoglobin of *Clonorchis sinensis*. *Parasitology Research* **2003**, *90* (5), 365-368.
75. Teale, F. W., Cleavage of the haem-protein link by acid methylethylketone. *Biochim biophys acta* **1959**, *35*, 543.
76. Light, W. R.; Rohlfs, R. J.; Palmer, G.; Olson, J. S., Functional effects of heme orientational disorder in sperm whale myoglobin. *J Biol Chem* **1987**, *262* (1), 46-52.
77. Seki, T.; Kunichika, T.; Watanabe, K.; Orino, K., Apolipoprotein B binds ferritin by heme-mediated binding: evidence of direct binding of apolipoprotein B and ferritin to heme. *BioMetals* **2008**, *21* (1), 61-69.
78. De Juan, A.; Jaumot, J.; Tauler, R., Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Anal. Methods* **2014**, *6* (14), 4964-4976.
79. Leung, G. C. H.; Fung, S. S. P.; Dovey, N. R. B.; Raven, E. L.; Hudson, A. J., Precise determination of heme binding affinity in proteins. *Analytical Biochemistry* **2019**, *572*, 45-51.
80. Akama-Garren, E. H.; Joshi, N. S.; Tammela, T.; Chang, G. P.; Wagner, B. L.; Lee, D.-Y.; Rideout Iii, W. M.; Papagiannakopoulos, T.; Xue, W.; Jacks, T., A Modular Assembly Platform for Rapid Generation of DNA Constructs. *Scientific Reports* **2016**, *6* (1), 16836.

81. Mosure, S. A.; Strutzenberg, T. S.; Shang, J.; Munoz-Tello, P.; Solt, L. A.; Griffin, P. R.; Kojetin, D. J., Structural basis for heme-dependent NCoR binding to the transcriptional repressor REV-ERB β . *Science Advances* **2021**, 7 (5).
82. Kabe, Y.; Nakane, T.; Koike, I.; Yamamoto, T.; Sugiura, Y.; Harada, E.; Sugase, K.; Shimamura, T.; Ohmura, M.; Muraoka, K.; Yamamoto, A.; Uchida, T.; Iwata, S.; Yamaguchi, Y.; Krayukhina, E.; Noda, M.; Handa, H.; Ishimori, K.; Uchiyama, S.; Kobayashi, T.; Suematsu, M., Haem-dependent dimerization of PGRMC1/Sigma-2 receptor facilitates cancer proliferation and chemoresistance. *Nat Commun* **2016**, 7, 11030.
83. Bianchetti, C. M.; Yi, L.; Ragsdale, S. W.; Phillips, G. N., Comparison of Apo- and Heme-bound Crystal Structures of a Truncated Human Heme Oxygenase-2. *Journal of Biological Chemistry* **2007**, 282 (52), 37624-37631.
84. Oosterheert, W.; Gros, P., Cryo-electron microscopy structure and potential enzymatic function of human six-transmembrane epithelial antigen of the prostate 1 (STEAP1). *Journal of Biological Chemistry* **2020**, 295 (28), 9502-9512.
85. Hira, S.; Tomita, T.; Matsui, T.; Igarashi, K.; Ikeda-Saito, M., Bach1, a heme-dependent transcription factor, reveals presence of multiple heme binding sites with distinct coordination structure. *IUBMB Life* **2007**, 59 (8), 542-551.
86. Igarashi, K.; Kurosaki, T.; Roychoudhuri, R., BACH transcription factors in innate and adaptive immunity. *Nature Reviews Immunology* **2017**, 17 (7), 437-450.
87. Shen, J.; Sheng, X.; Chang, Z.; Wu, Q.; Wang, S.; Xuan, Z.; Li, D.; Wu, Y.; Shang, Y.; Kong, X.; Yu, L.; Li, L.; Ruan, K.; Hu, H.; Huang, Y.; Hui, L.; Xie, D.; Wang, F.; Hu, R., Iron metabolism regulates p53 signaling through direct heme-p53 interaction and modulation of p53 localization, stability, and function. *Cell Rep* **2014**, 7 (1), 180-93.
88. Shen, J.; Sheng, X.; Chang, Z.; Wu, Q.; Xie, D.; Wang, F.; Hu, R., The heme-p53 interaction: Linking iron metabolism to p53 signaling and tumorigenesis. *Mol Cell Oncol* **2016**, 3 (1), e965642.
89. Yin, L.; Wu, N.; Curtin, J. C.; Qatanani, M.; Szwergold, N. R.; Reid, R. A.; Waitt, G. M.; Parks, D. J.; Pearce, K. H.; Wisely, G. B.; Lazar, M. A., Rev-erb alpha, a heme sensor that coordinates metabolic and circadian pathways. *Science* **2007**, 318 (5857), 1786-9.
90. Yang, J.; Kim, K. D.; Lucas, A.; Drahos, K. E.; Santos, C. S.; Mury, S. P.; Capelluto, D. G.; Finkielstein, C. V., A novel heme-regulatory motif mediates heme-dependent degradation of the circadian factor period 2. *Mol Cell Biol* **2008**, 28 (15), 4697-711.
91. Nishitani, Y.; Okutani, H.; Takeda, Y.; Uchida, T.; Iwai, K.; Ishimori, K., Specific heme binding to heme regulatory motifs in iron regulatory proteins and its functional significance. *J Inorg Biochem* **2019**, 198, 110726.
92. Lathrop, J. T.; Timko, M. P., Regulation by heme of mitochondrial protein transport through a conserved amino acid motif. *Science* **1993**, 259 (5094), 522-5.

93. Uchida, T.; Sagami, I.; Shimizu, T.; Ishimori, K.; Kitagawa, T., Effects of the bHLH domain on axial coordination of heme in the PAS-A domain of neuronal PAS domain protein 2 (NPAS2): conversion from His119/Cys170 coordination to His119/His171 coordination. *J Inorg Biochem* **2012**, *108*, 188-95.
94. Smith, A. T.; Pazicni, S.; Marvin, K. A.; Stevens, D. J.; Paulsen, K. M.; Burstyn, J. N., Functional Divergence of Heme-Thiolate Proteins: A Classification Based on Spectroscopic Attributes. *Chemical Reviews* **2015**, *115* (7), 2532-2558.
95. Shcherbo, D.; Christopher; Galina; Elena; Tatiana; Aleksandr; Vladislav; Vladimir; Kristin; Patrick; Lukyanov, S.; Andrey; Michael; Dmitriy, Far-red fluorescent tags for protein imaging in living tissues. *Biochemical Journal* **2009**, *418* (3), 567-574.
96. Khrenova, M.; Topol, I.; Collins, J.; Nemukhin, A., Estimating Orientation Factors in the FRET Theory of Fluorescent Proteins: The TagRFP-KFP Pair and Beyond. *Biophysical Journal* **2015**, *108* (1), 126-132.
97. Hosoda, H.; Mori, H.; Sogoshi, N.; Nagasawa, A.; Nakabayashi, S., Refractive Indices of Water and Aqueous Electrolyte Solutions under High Magnetic Fields. *The Journal of Physical Chemistry A* **2004**, *108* (9), 1461-1464.

Appendix II: Additional information on the preliminary calculations for redox sensor.

$$J = \frac{\int F_D(\lambda) \times \epsilon_A(\lambda) \times \lambda^4 d\lambda}{\int F_D(\lambda) d\lambda} \quad (8)$$

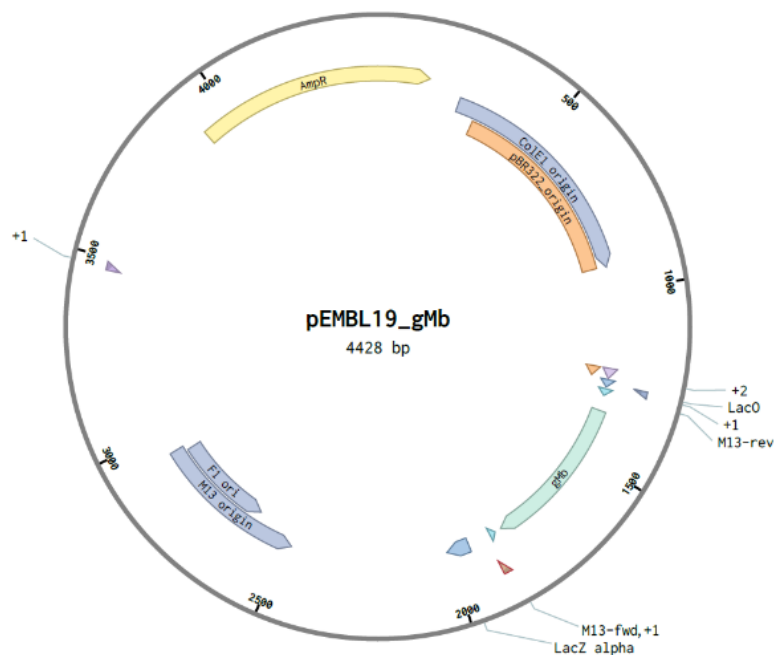
Equation 8. The spectral overlap (J) was calculated from the molar absorptivity spectrum of the acceptor, gMb, and the relative emission spectrum of the donor, mKate2. The fluorescence intensity of the donor at wavelength λ is denoted by $F_D(\lambda)$, the molar absorptivity of the acceptor at wavelength λ is denoted by $\epsilon_A(\lambda)$. The denominator accounts for the fact that a relative emission spectrum was used for mKate2.

r / nm	(A)		(B)			
	E_{ferric}	$E_{ferrous}$	τ_{apo} / ns	$\tau_{holo ferric} / \text{ns}$	$\tau_{holo ferrous} / \text{ns}$	$\Delta\tau_{holo} / \text{ns}$
2.1	0.92	0.79	2.50	0.20	0.53	0.34
2.2	0.90	0.74	2.50	0.25	0.66	0.41
2.3	0.87	0.68	2.50	0.32	0.80	0.48
2.4	0.84	0.62	2.50	0.40	0.94	0.54
2.5	0.80	0.56	2.50	0.49	1.09	0.60
2.6	0.76	0.51	2.50	0.59	1.23	0.65
2.7	0.72	0.45	2.50	0.70	1.38	0.68
2.8	0.68	0.40	2.50	0.81	1.51	0.70
2.9	0.63	0.35	2.50	0.93	1.63	0.70
3	0.58	0.30	2.50	1.05	1.74	0.69
3.1	0.53	0.26	2.50	1.17	1.84	0.67
3.2	0.48	0.23	2.50	1.29	1.93	0.64
3.3	0.44	0.20	2.50	1.41	2.01	0.60
3.4	0.39	0.17	2.50	1.52	2.07	0.56

Table 5. (A) The efficiency of energy transfer upon ferric and ferrous heme binding to the sensor have been given as E_{ferric} and $E_{ferrous}$, and were calculated using Equations 1 (Section 1.3.1) and 6 (Section 4.1). (B) The fluorescent lifetimes for the heme-bound sensor, τ_{holo} , were calculated for both ferric and ferrous heme using Equation 7 (Section 4.1) and τ_{apo} (given as 2.5 ns, which is the lifetime for mKate2, obtained from FPbase ID: DBBO8). The relative concentrations of ferric and ferrous heme in live cells can then be calculated using a multiexponential model, following methods used by Leung *et al.* for characterisation of the mAPXmEGFP sensor. The difference in fluorescence lifetimes for the sensor with ferric and ferrous heme bound, $\Delta\tau_{holo}$, demonstrates that it should be possible to differentiate between both ferric and ferrous heme when r is in the range 2.1-3.4 nm. The value for $\Delta\tau_{holo}$ is maximised when $r = 2.8 - 2.9$ nm (shown in bold). At these values, the difference in fluorescence lifetimes upon binding to ferric or ferrous heme is 0.7 ns.

Appendix III: Plasmid and sequence information for sperm whale Mb(H64Y/V68F).

(A)



(B)

	ggt	ctg	tct	gaa	ggt	gaa	tgg	cag	ctg	ggt	ctg	cat	ggt	tgg	gct	aaa	ggt	gaa	gct
1	V	L	S	E	G	E	W	Q	L	V	L	H	V	W	A	K	V	E	A
	gac	gtc	gct	ggt	cat	ggt	cag	gac	atc	ttg	att	cga	ctg	ttc	aaa	tct	cat	ccg	gaa
20	D	V	A	G	H	G	Q	D	I	L	I	R	L	F	K	S	H	P	E
	ctg	gaa	aaa	ttc	gat	cgt	ttc	aaa	cat	ctg	aaa	act	gaa	gct	gaa	atg	aaa	gct	tct
40	L	E	K	F	D	R	F	K	H	L	K	T	E	A	E	M	K	A	S
	gat	ctg	aaa	aaa	tac	ggt	ggt	acc	ttc	cta	act	gcc	cta	ggt	gct	atc	ctt	aag	aaa
60	D	L	K	K	Y	G	V	T	F	L	T	A	L	G	A	I	L	K	K
	ggg	cat	cat	gaa	gct	gag	ctc	aaa	ccg	ctt	gcg	caa	tcg	cat	gct	act	aaa	cat	aag
80	G	H	H	E	A	E	L	K	P	L	A	Q	S	H	A	T	K	H	K
	ccg	atc	aaa	tac	ctg	gaa	ttc	atc	tct	gaa	gcg	atc	atc	cat	ggt	ctg	cat	tct	aga
100	P	I	K	Y	L	E	F	I	S	E	A	I	I	H	V	L	H	S	R
	cca	ggt	aac	ttc	ggt	gct	gac	gct	cag	ggt	gct	atg	aac	aaa	gct	ctc	gag	ctg	ttc
120	P	G	N	F	G	A	D	A	Q	G	A	M	N	K	A	L	E	L	F
	aaa	gat	atc	gct	gct	aag	tac	aaa	gaa	ctg	ggt	tac	cag	ggt	taa	tga			
140	K	D	I	A	A	K	Y	K	E	L	G	Y	Q	G	-	-			

(C)

Forward primer (M13-rev): AGCGGATAACAATTTTCAC; Reverse primer (M13-fwd):
CCCAGTCACGACGTTGTAAAACG; Internal gMb primer: TAACTTCGGTGCTGACGC

Figure 24. (A) Plasmid map for pEMBL19-Mb(H64Y/V68F), which encodes for sperm whale gMb. (B) Sequence information for gMb showing codons and corresponding amino acids. The mutated residues (H64Y and V68F) have been shown in red. (C) Primers used for sequencing.

Appendix IV: The full set of data collected from heme titration into gMb.

# addition	V _{cuvette} /uL	V _{hemin added} /uL	Cumulative V _{hemin added} /uL	Abs ₄₁₂	<i>In situ</i> [hemin]/uM	<i>In situ</i> [P]/uM	<i>In situ</i> [hemin]/starting [P]	<i>In situ</i> [hemin]/ <i>in situ</i> [P]	<i>In situ</i> -starting [P] deviation	% <i>in situ</i> [P] over starting [P]
1	909.90	9.90	9.90	0.12	0.82	7.66	0.10	0.11	0.54	0.93
2	919.80	9.90	19.80	0.22	1.61	7.58	0.20	0.21	0.62	0.92
3	929.70	9.90	29.70	0.25	2.40	7.50	0.29	0.32	0.70	0.91
4	939.60	9.90	39.60	0.28	3.16	7.42	0.39	0.43	0.78	0.90
5	949.50	9.90	49.50	0.30	3.91	7.34	0.48	0.53	0.86	0.90
6	959.40	9.90	59.40	0.30	4.64	7.26	0.57	0.64	0.94	0.89
7	969.30	9.90	69.30	0.34	5.36	7.19	0.65	0.75	1.01	0.88
8	979.20	9.90	79.20	0.36	6.07	7.12	0.74	0.85	1.08	0.87
9	989.10	9.90	89.10	0.38	6.76	7.05	0.82	0.96	1.15	0.86
10	999.00	9.90	99.00	0.40	7.43	6.98	0.91	1.07	1.22	0.85
11	1008.90	9.90	108.90	0.42	8.10	6.91	0.99	1.17	1.29	0.84
12	1018.80	9.90	118.80	0.44	8.75	6.84	1.07	1.28	1.36	0.83
13	1028.70	9.90	128.70	0.46	9.38	6.78	1.14	1.38	1.42	0.83
14	1038.60	9.90	138.60	0.47	10.01	6.71	1.22	1.49	1.49	0.82
15	1048.50	9.90	148.50	0.48	10.62	6.65	1.30	1.60	1.55	0.81
16	1058.40	9.90	158.40	0.50	11.22	6.59	1.37	1.70	1.61	0.80
17	1068.30	9.90	168.30	0.52	11.82	6.52	1.44	1.81	1.68	0.80
18	1078.20	9.90	178.20	0.53	12.40	6.46	1.51	1.92	1.74	0.79
19	1088.10	9.90	188.10	0.55	12.97	6.41	1.58	2.02	1.79	0.78
20	1098.00	9.90	198.00	0.58	13.52	6.35	1.65	2.13	1.85	0.77

Table 6. The full set of titration data collected to produce the results in Section 4.2.3. The concentration of hemin stock was 75 uM and the initial concentration of protein (*i.e.* apo-protein before addition #1) was 8.2 uM.

Appendix V: UV-Visible spectrum of hemin stock used for heme titrations.

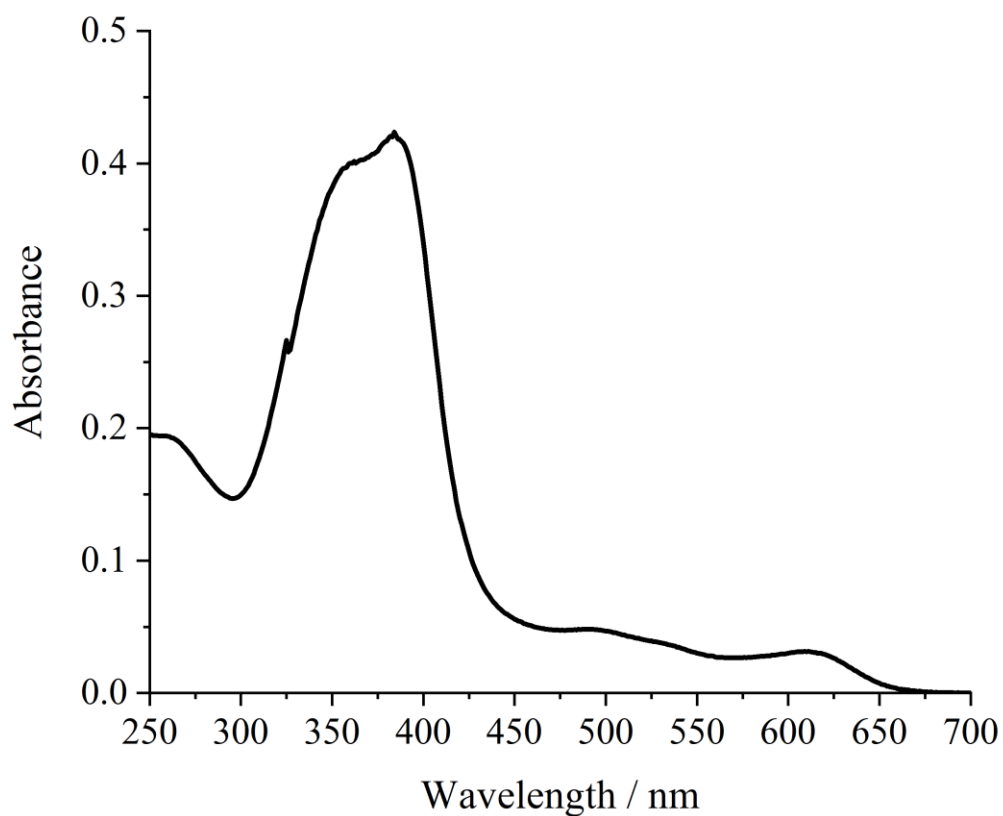


Figure 25. UV-Visible spectrum of hemin stock, with a concentration of 75 μM . The stock was produced as described in Section 2.7 and the concentration was calculated using Equation 4 (Section 2.6) using the following values: $A_{385} = 0.44$; $\epsilon_{385} = 58.4 \text{ mM}^{-1} \text{ cm}^{-1}$; $l = 1 \text{ cm}$.

Appendix VI: Plasmid map for TOPO mKate2 plasmid, used for cloning of sensor.

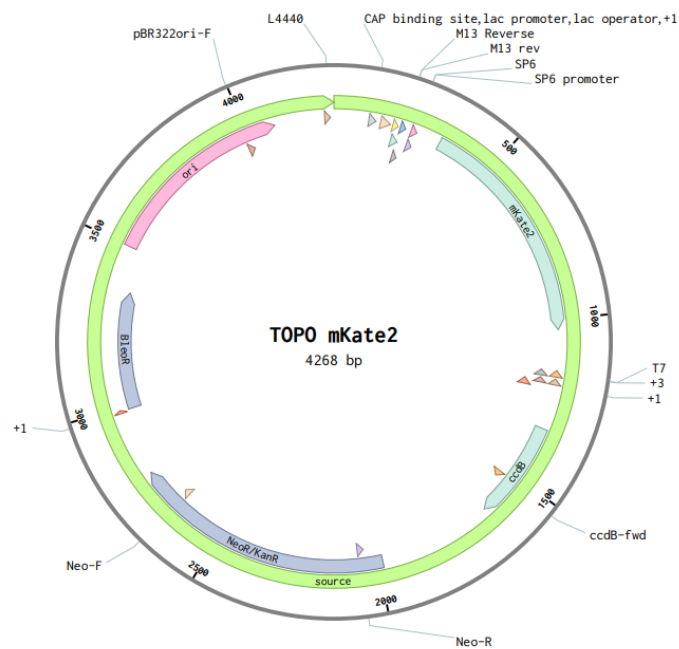


Figure 26. Plasmid map for TOPO mKate2 plasmid. Unique restriction sites used to confirm the identity of plasmid were NotI (position 1120 bp) and HindIII (position 277 bp). The distance between these sites is 843 bp (< 1 kb). The unique restriction sites used for cloning gMb gene into TOPO mKate2 with restriction digestion/ligation cloning techniques were NotI and AclI (position 1034 bp). The cut sites for the relevant restriction enzymes are NotI (GC/GGCCGC); AclI (AA/CGTT); HindIII (A/AGCTT).

Appendix VII: Primers used during PCR amplification of gMb gene.

FORWARD 5' to 3'

AGTGC **AACGTT** **GTTCTGTCTGAAGGTGAATG**

GC content: 45%

$T_m = 61.7$ °C

REVERSE 5' to 3'

ATATC **GCGGCCGC** **TCATTAACCCTGGTAACC**

GC content: 55%

$T_m = 65.7$ °C

gMb sequence – start codon excluded, stop codons included. Reverse primer is complementary to sense strand

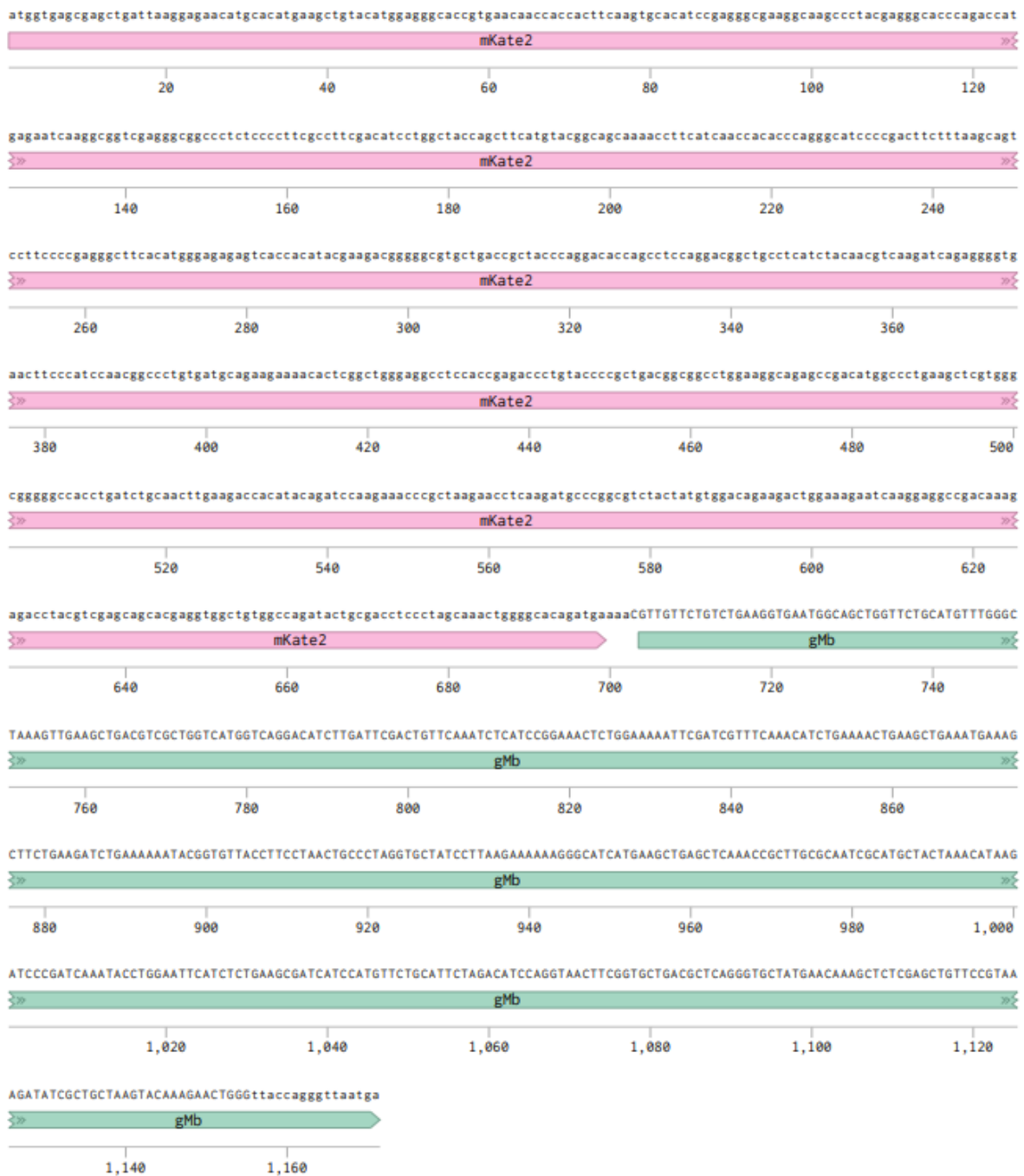
AclI recognition site sense

NotI recognition site antisense, written 5' to 3'

Additional bases included at 5' end of primers to ensure successful digestions. GC content and T_m calculations performed with Oligo Calc (<http://biotools.nubic.northwestern.edu/OligoCalc.html>), as well as checks for self-complementarity within the primers.

Appendix VIII: Sequence of gMbmKate2 construct.

(A)



(B)

Forward primer 5' to 3' – AACGTTGTTCTGTCTGAAGGTG

GC content: 45%; $T_m = 64^\circ\text{C}$

Reverse primer 5' to 3' – TCTGTGCCCCAGTTTGCT

GC content: 56%; $T_m = 67^\circ\text{C}$

