Characterizing Heme Pockets

Patrick Finnerty

TO BE COMPLETED UPON AGREEMENT OF DISCUSSION/CONCLUSION

Metalloproteins compose approximately 40 percent (look up how to do percents in latex) of all known proteins, and use some metallic group to accomplish their chemistry. One such metallic group is heme. Heme is a member of the porphyrin family, which are able to catalyze a broad range of reactions. Heme in particular catalyzes many different reactions and is present in many proteins. However, the underlying structural requirements to host heme in a protein are not well studied.

In this study, all heme or heme-c containing proteins as of xx were downloaded and processed in order to determine underlying structural characteristics these proteins may have in common. Parameters that were examined include: xx. Overall, we found: xx. These results may have implications for protein engineering; or if I fucked up this illustrates the difficulty of the field and demonstrate the wide range of acceptable environments of heme; it may therefore be more appropriate to take a more hands-on approach until perhaps other computational methods evolve to better examine structure-function relationships.

Table of Contents

# Lay Summary

I investigated how heme, a molecule involved in many biological processes, binds to proteins. I did this by…

# Introduction

Proteins that contain metal ion cofactor(s) are known as metalloproteins, and they compose approximately 40% of all proteins. The metal ion can be free, or bound within a ligand. One such ligand is heme, a complex of iron and a porphyrin ring. Heme is employed by many metalloproteins to catalyze a broad range of reactions.

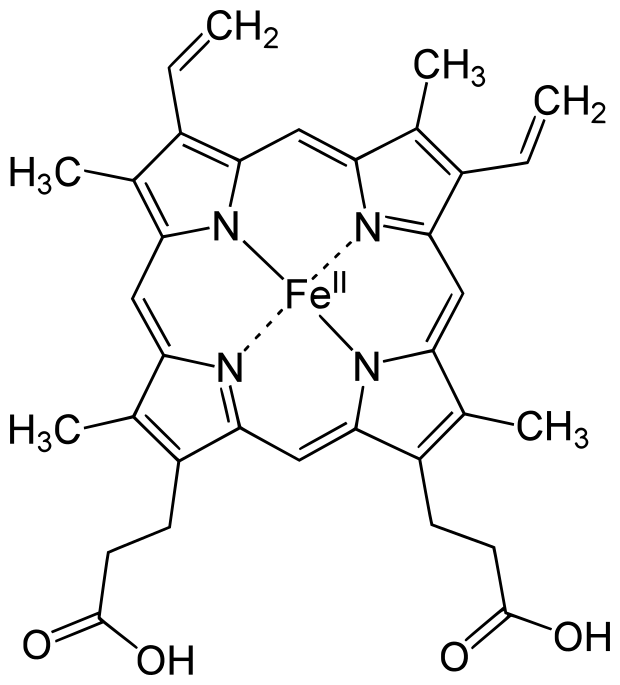
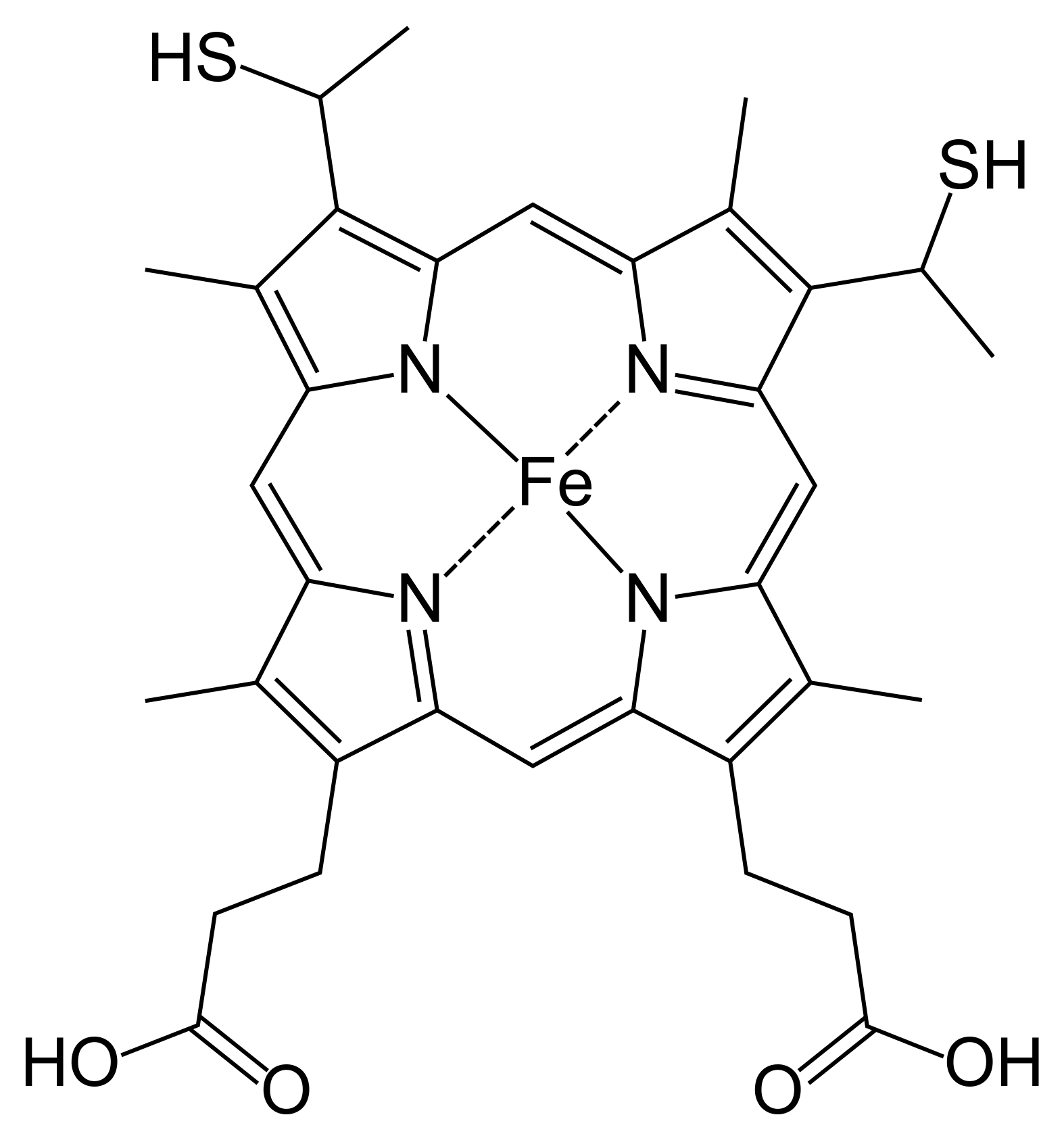
Proteins containing heme are known as hemoproteins - hemoglobin and myoglboin are well-known examples, using heme to store and transport oxygen. Other examples hemoproteins are peroxidases, catalases (a type of peroxidase), nitric oxide synthases, heme oxygenases, and cytochrome p450s. Peroxidases and catalases catalyze oxidation-reduction reactions using a histidine-bound heme, with catalases in particular specializing in the decomposition of hydrogen peroxide. Nitric oxide synthases catalyze the reaction of L-arginine into nitric oxide, which is vitally important to cellular signaling ([Poulos](#ref-Poulos2014) ([2014](#ref-Poulos2014))). Free heme molecules may be released upon degradation of hemoproteins (especially hemoglobin), however, heme is prooxidative and therefore toxic to cells and must be cleared. Heme oxygenases assist in the degradation of heme, and are regarded as potential therapeutics, due to anti-inflammatory effects([Araujo, Zhang, and Yin 2012](#ref-Araujo2012)).

The enzymes with arguably the most potential applications, cytochrome P450s function as powerful monooxygenases. They participate in many reactions: capable oxidizing a wide range of substrates, serving to oxidize carbohydrates, steroids, fatty acids; catalyzing hormone degradation and synthesis; and degrading the majority of drugs([Poulos](#ref-Poulos2014) ([2014](#ref-Poulos2014))). Due to their extraordinary utility and range of reactions, cytochrome p450s are of great interest in the protein engineering field. Cytochrome P450s have the potential to be used in industrial biocatalysis, e.g. in pharmaceutical production, bioremediation of environmental pollutants([Du et al. 2017](#ref-Du2017); [Lalonde 2016](#ref-Lalonde2016)). The limiting factor preventing its deployment has been the struggle to increase enzymatic efficiency and therefore yield of processes employing the enzyme([Girvan and Munro 2016](#ref-Girvan2016); [Z. Li et al. 2020](#ref-Li2020)).

Echoing the many applications of its host enzymes, the heme molecule itself has its own variety. Thus far, only heme-b has been discussed (although it is the most abundant and most employed). There are several types of heme, structurally and chemically different, that are used to achieve different chemical reactions: in this study, we examined heme-b, heme-c, siroheme, and verdoheme. Their structures are shown in Figures 1-5.

All types of heme are a coordination complex composed of an iron atom coordinated and bound to a porphyrin ring. Porphyins are composed of four pyrrole subunits (pentagonal structures of four carbons and a nitrogen) that are bound together via methine (i.e. carbon) bridges (FIXME get a citation for this stuff). Porphyrins are considered macrocycles, molecules with large ring structures. The ring within porphyrins is heterocyclic, and considered aromatic, and therefore acts as a large resonant structure capable of transferring electrons. Coupled with an iron atom to enable reduction-oxidation reactions and the macrocycle of the porphyrin ring, the overall heme complex is therefore highly suited for reactions involving electron transfer.

**Aware I need to scale some of the figures and make this nice and uniform**

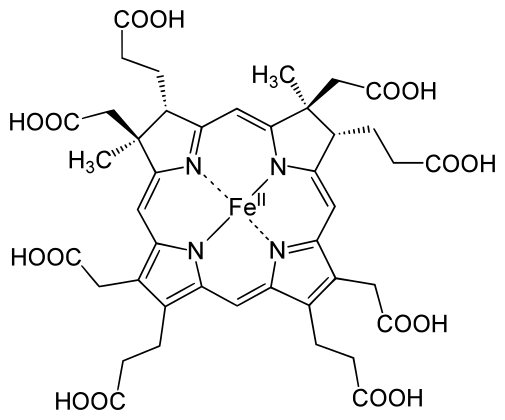
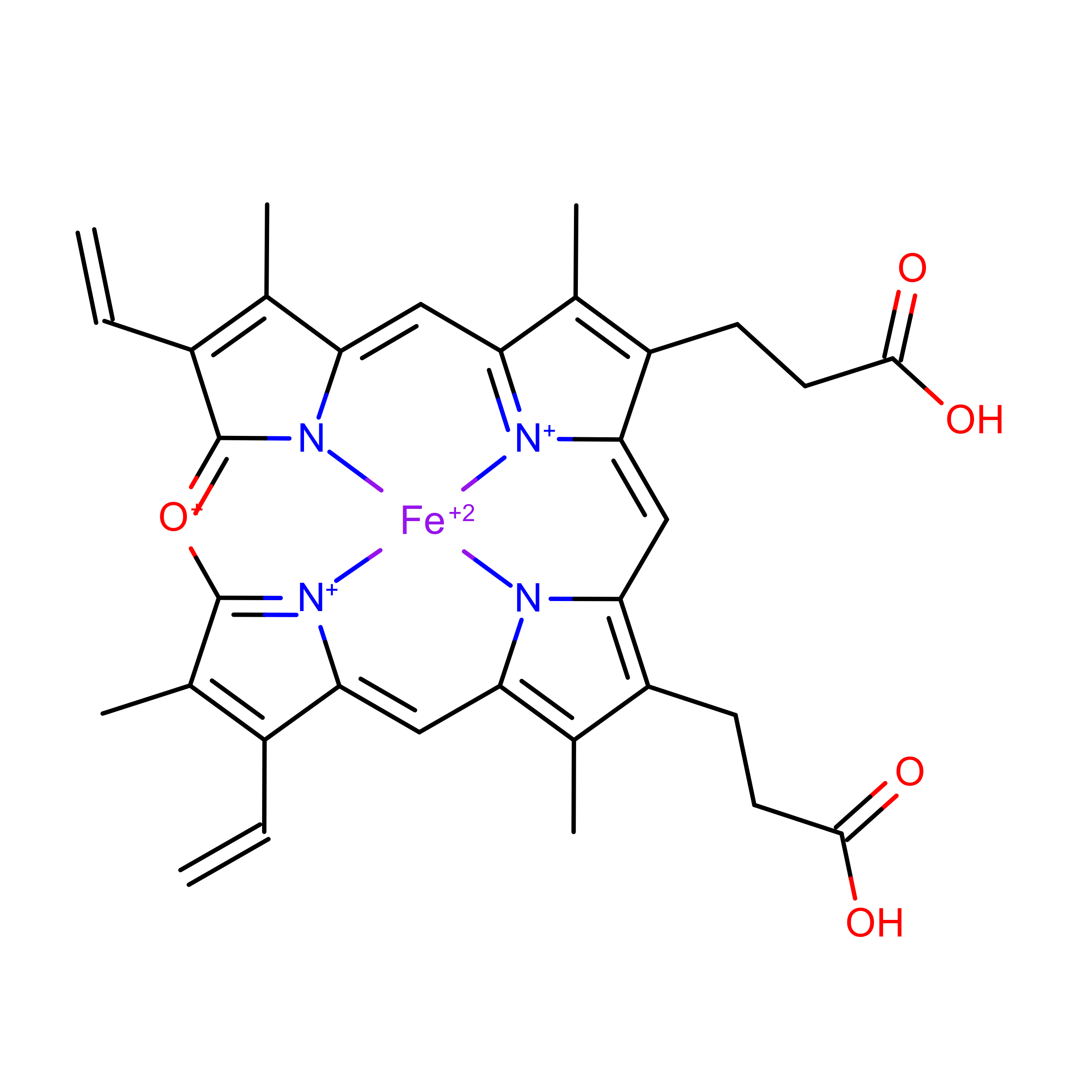
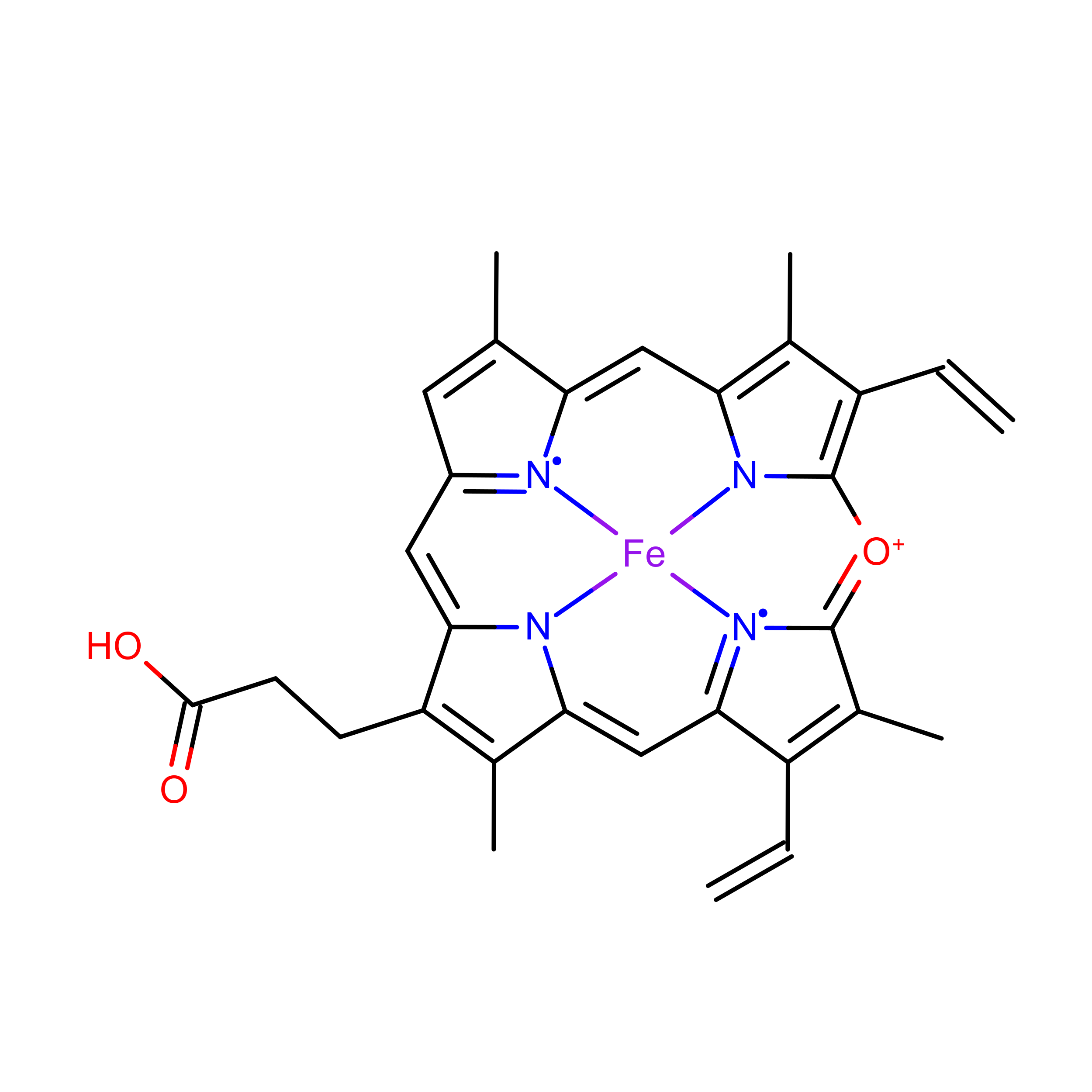


Figure 0.3: Siroheme (SRM)

##Types of Heme

The most common heme is heme-b. It is employed by the vast majority of hemoproteins. It is composed of an iron and porphyrin ring complex, with the addition of two propionate groups. These two propionate groups interact with polar amino acids in the binding pocket. The iron atom is usually coordinated to a histidine or cysteine, depending on the enzyme.

Heme-c is derived from heme-b. It is fairly similar to heme, with two notable excetions: the iron atom binds, with few exceptions, covalently to cysteine residues in the binding pocket; and its two vinyl groups form thioester bonds with amino acids in the protein binding pocket. Its function is much more specific than heme-b, mostly serving as an electron carrier rather than catalyzing a plethora of reactions. The reason for this is not abundantly clear, but several studies suggest that because of its covalent bonding, heme-c has an electronic potential that is can be far lower and in general more broad, and more specifiable, than heme-b. ([Bowman and Bren 2008](#ref-Bowman2008); [Kleingardner and Bren 2015](#ref-Kleingardner2015))

Siroheme is even more limted in its applications, but highly specialized for its role. It is still an iron atom-porphyrin coordination complex, but it contains far more carboxyl and propionate groups than the other types of heme, making it highly polar. It is used exclusively in sulfite and nitrite reductases, which catalyze the reduction of the sulfates and nitrates plants uptake from the ground, providing the sources of nitrogen and sulfur used to produce nitrogen and sulfur-containing amino acids([Tripathy, Sherameti, and Oelmüller](#ref-Tripathy2010) ([2010](#ref-Tripathy2010))). The reason for the use of siroheme in this reaction rather than heme-b is not completely understood. But one study suggests that the bridge that siroheme forms between its catalyic iron atom, and the protein matrix envrionment (which also necessarily involves another cofactor, a cluster of cubane for electron transfer and provision) is more efficient at channeling electrons than the bridge that could be formed by heme.([Brânzanic, Ryde, and Silaghi-Dumitrescu 2019](#ref-Branzanic2019))

Lastly, verdoheme is an intermediate product in the degradation of heme-b by heme oxygenase. When heme oxygenase degrades heme-b, biliverdin, carbon monoxide, and iron are produced; verdoheme is the precusor to bilverdin([Lai et al. 2010](#ref-Lai2010); [Sato et al. 2007](#ref-Sato2007)). While a product of prior reactions wthin heme oxygenase, verdoheme appears to be oriented and bound differently ([Lad, Ortiz De Montellano, and Poulos 2004](#ref-Lad2004)). The two structures used in the study, VEA and VER, are either partially oxidized or partially oxidized and contain one less propionate group.

In summary, heme molecules are varied, and used in a diverse set of reactions. Several enzymes have the potential to be of great value either in biocatalysis or pharmaceutical applications. There are multiple methods employed to design molecules, but rational design in particular (basically, the mutation of certain residues based on an understanding of the structure-function relationships) is at least partially hampered by an incomplete understanding of the binding environment for heme. The importance of the binding environment was noted in a study seeking to design *de novo* heme-c based enzymes, and found the binding environment likely to be of importance in modulating redox potential ([Ishida et al. 2004](#ref-Ishida2004)).

A fairly recent study conducted a structural analysis of 125 hemoprotein chains([T. Li, Bonkovsky, and Guo](#ref-Li2011) ([2011](#ref-Li2011))). The study suggested hemoproteins undergo small conformational changes during binding; and that apo-form (ligand-containing) proteins may therefore be suitable for bioinformatics-based prediction and protein design. Additionally, the heme binding environment of the protein dataset used in the study was found to be rich in aromatic and nonpolar residues.

The aforementioned study was published in 2011 - since then the PDB has been populated with far more hemoproteins. The 125 protein chains used in the study were sourced from a small dataset of proteins (FIXME: Double check but their Table 2 is like 20 proteins), compared to the amount of hemoproteins now available in the PDB. The focus of the study was on conformational differences induced by heme-binding. And as for the binding environment, the focus was largely on interactions with the coordinating iron atom rather than interactions that occur with the porphyrin ring. **could add the distances arbitrary definition thing but I hesitate to write anything that could be construed as uh, negative**

The interactions heme forms with its binding environment have been well documented. And many studies have demonstrated the importance of these interactions and have either demonstrated or at least furnished theoretical mechanisms by which the heme molecule functions in reactions involving hemoproteins. However, there has not been significant work dedicated to examining the heme-binding environment itself.

In this study, we present research focused on elucidating the binding environment of multiple heme molecules: heme-b (HEM), heme-c (HEC), siroheme (SER), and verdoheme (VEA/VER). UCSF Chimera was used to both extract and predict properties of a diverse set of hemoproteins. R was used to analyze the results. A robust and high-throughput framework was constructed to process the datasets for each heme molecule, requiring only inputs of which ligand is to be examined per dataset.

The properties extracted and predicted of the heme molecules were (I need to figure out how to do a bullet list in this format): +Amino acid frequencies within binding pockets +Volumes +Solvent accessible and excluded surface areas of heme and binding pockets +Distances of amino acids from the Fe atom of heme +Planar angles of amino acids in binding pocket v. heme +CA-CB-Fe angles of amino acids in binding pocket v. heme

These results may be of use in rational design of hemoproteins in future studies, or at least improve the understanding of the heme binding environment.

**WILL INSERT PARAGRAPH ABOUT THE MEANING OF THESE RESULTS IN THE DESIGN ARTIFICAL METALLOPROTEINS, AND BRIEFLY INTRODUCE THEM EARLIER. ‘AS MENTIONED PREVIOUSLY, THERE ARE MANY OPPORTUNITIES TO EMPLOY IMPROVED HEMOPROTEINS, AND A MORE THOROUGH UNDERSTANDING OF THE HEME BINDING ENVIRONMENT WOULD AID IN THEIR DESIGN…’**

# 1 Methods

## 1.1 Datasets

A list of PDBs was assembled that represented either a representative sample of a variety of proteins, with a resolution better than 3A, (HEM and HEC) or, all proteins containing these ligands were downloaded from the PDB (in the case of SRM, VER, VEA). Not all downloaded PDBs were appropriate for this study (e.g. contained “wobble” structures) and therefore the amount of PDBs was culled. The datasets are current as of 16 August 2021.

The size of the datasets actually used in the study were as follows: HEM (n=58), HEC(n=14), SRM (n=9), VER (n=2) and VEA (n=2) for a combined n=4 for VERDOHEME.

The name of all proteins used in the study and their source organism are provided tables within Appendix 5.1.

## 1.2 Preprocessing

Many of the PDBs downloaded were multimeric structures. While many of the scripts employed in the study may function with multimeric structures, the number of subunits per protein (FIXME! better way to say this?) would skew results in favor of multimeric proteins with more subunits. The information gleaned from similar subunits would also not be of utility in this study.

Therefore all downloaded PDBs were converted to monomeric structures. This was achieved by saving a single chain (chain A) of each PDB and eliminating all other chains. The single chain was then saved as a PDB and used in all subsequent scripts.

## 1.3 Processing Monomers

UCSF-Chimera was used to generate all data in this study. Multiple scripts were employed to achieve a high-throughput process where all monomeric PDBs could be processed in the same session.

Chimera was used to predict the following qualities: Volume of the ligand binding pocket, accessible and excluded surface area of the ligand, and accessible and excluded surface area of the binding pocket. These calculations require a population of atoms to be selected for the calculation.

Atoms were selected within a distance cutoff, to be considered as “interacting” with the ligand or forming the binding pocket. Distance cutoffs from the ligand of 5A and 7A were chosen; for the predicted qualities, the algorithms were run twice to get values at 5A and 7A. For the distance and angle calculations, only the 7A distance cutoff was used, as the cutoff does not factor into any calculations and may be set during analysis.

as these are selected arbitrarily, data from the 5A and 7A runs are overlaid in the figures reported in Appendix ??. Data tables are also provided in Appendix ??.

### 1.3.1 Amino Acid Frequency

Amino acids within the bounds of the lower and upper distance cutoff were selected and recorded. These were then counted for frequency per residue.

### 1.3.2 Volume Calculations

Volume of the binding pocket was predicted via Surfnet, run with default parameters of Grid Interval = 1.0 and Distance Cutoff = 10.0 (the latter option does not relate to the distance cutoff from the ligand(FIXME! find source for this, appears 100% to be true?)).

### 1.3.3 Surface Area Calculations

Excluded and accessible surface areas of both the ligand and the binding pocket were calcualted using Chimera’s “surf” algorithm, available as “Measure Volume and Area” via the GUI.

### 1.3.4 Distance Calculations

Distances of amino acids from the ligand could not be calculated accurately nor precisely in a direct way. Instead, distances for each atom composing a residue were calculated. The distances of all atoms within a residue were averaged, and this value was taken as the mean distance of the entire residue and used in subsequent steps.

The data produced in this step therefore include the mean distance of each amino acid. This is traceable, and the angular data below are cross-referenced with this list of distances. All data shown in figures (FIME! Also for tables?) are multidimensional and may be filtered for distance.

### 1.3.5 Planar Angle Calculations

Individual residues and the ligand were defined as axes The angle between each residue’s axis and the axis of the ligand were calculated. Each axis functions essentially as a separate plane. (FIXME! Include a picture of what this looks like?) This employed the “define axis,” and “angle” functions of Chimera; the Axes/Planes/Centroids Structural Analysis function of Chimera via GUI.

### 1.3.6 CA-CB-Fe Calculations

Residues within the distance cutoff were examined one by one. The angle of between each residue’s carbon alpha (CA) and carbon beta (CB) and the Fe of the ligand was calculated, using the “angle” function of Chimera. The ligand nor the Fe atom were compared with themselves.

## 1.4 Import to R

All data were imported to R and processed from text files into organized data formats. R was used to cross-reference angle and distance data. All plots and tables were constructed using R and imported directly to this document using Rmarkdown.

# 2 Results and Discussion

–>

## 2.1 Analysis of Nearby Residues of Natural Porphyrins

The first part of the study aimed at providing statistics on the amino acid propensity to interact with hemes in natural proteins. We studied heme-b, heme-c, siroheme and verdoheme. Because we are not looking only at the iron environment, but instead at the environment of the entire microcycle, we did the analysis for any amino acid with potential contact with the heme. This was defined as any AA having at least one atom within the cutoff distances of 5 and 7 Angtroms (A).

Amino acid frequencies were obtained for distance cutoffs of 5A and 7A - these figures and data are shown in **FIXME ADD APPENDICES LATER** The trends in these data are very similar and therefore only the data pertaining to the 7A distance cutoff are discussed below.

## 2.2 AA Frequency

### 2.2.1 Heme-b

#### 2.2.1.1 Amino Acid Frequencies in Binding Pocket

Figure 2.1 plots the frequency of each residue within 7A of heme-b. Immediately below is Figure 2.2, which plots the frequency of each residue within the entire monomer. **I’m thinking put tables in back they dont add much value here** Data tables are available in Appendix…

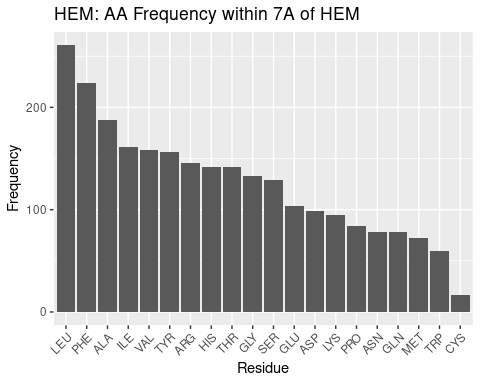


Figure 2.1: HEM: AA Frequency within 7A

**I use ‘surprising, striking’ a lot in this discussion. I’ll reword this to have some variety later. The results are at least interesting!**

Beginning at the left of the figure and moving right, large, nonpolar amino acids appear most frequently within 7A: LEU and PHE; ILE appears less frequently than these two amino acids but nonetheless is in high frequency. Small, nonpolar amino acids ALA and VAL also appear very frequently. As the majority of the heme-b molecule is made up of the nonpolar porphyrin ring, these amino acids are therefore likely in such high frequency to provide the nonpolar interactions/environment with the pyrole groups and methyl and vinyl groups.

Tyrosine, arginine, histidine appear next most frequently. The two propionate groups on heme are used to coordinate the heme in the binding pocket. These polar residues are therefore likely interact with the propionate groups, providing the polar interactions necessary, in addition to the nonpolar interactions above, to provide as hospitable of a binding environment as possible to coordinate the heme. In additon, the arginine and histidine groups are positively charged (at pH 7),further facilitating interaction with the electronegative proprionate groups of heme-b. It should be noted histidine is one of the residues that coordinates the iron atom, and this may therefore inflate its frequency in the binding pocket.

Glycine is a small residue and cannot form significant interactions within its environment; however, its frequency, or lack thereof (compared to background frequency, discused later), suggests the binding pocket may not require as much flexibility or… spatial considerations as in the rest of the protein. This would logically follow from the need for conserved binding sites.

Next appear serine, glutamate (glutamic acid) and aspartate (aspartic acid) and lysine. These are polar residues, and glutamate and aspartate are negatively charged; lysine is polarpositively charged. The negative charge is unlikely of importance in interaction with heme-b, however these polar amino acids likely again interact with the propionate groups on heme; only, infrequently. What is most interesting is why lysine is in such low abundance relative to the other polar, positively charged residues, arginine and histidine. Perhaps lysine’s fairly linear structure prevents it from fitting into the binding pocket; however, arginine is also somewhat linear and features prominently. The exact reason for why this could be is beyond the scope of this study.

Proline is a small nonpolar amino acid in low frequency; the trend for heme-b, at least, appears to be to favor large nonpolar amino acids in the binding pocket. This may suggest that a large amount of nonpolar interactions, per residue, is favored in the binding pocket, perhaps because of the limited space available to position residues to interact with heme.

Asparagine and glutamine are both medium-sized polar amino acids; given the trends already discussed it is surprising these are not in greater abundance. But as with proline, it may simply be a matter of maximizing the benefit of the interactions that may be formed with the heme; while asparagine and glutamine are polar, amino acids like arginine and histidine are both polar and positively charged, capable of stronger interactions with the electronegative propionate groups.

Methionine and tryptophan appear very infrequently in the binding pocket. All nonpolar amino acids already mentioned do not possess a sulfur (thio something?) bond in their structure; perhaps it is less favorable for the sulfur atom to interact with the porphyrin ring than another carbon **HELP I’M AT A LOSS ON EXPLAINING THIS ONE**. Tryptophan is very surprising to find as second-to-least frequent. It is a large nonpolar amino acid - but perhaps its single, potential hydrogen bond, although weak, is enough to prefer completely nonpolar residues. Or, with its size, it is preferable to have more numerous, smaller nonpolar residues that can favorably interact with the porphyrin while reducing steric hindrance of other residues in the environment (taking up less space).

Cystine appears most infrequently of all the amino acids in the binding pocket. This is quite surprising - cystine is highly evolutionarily conserved to coordinate the iron in the binding pocket. Perhaps the sample of PDBs used in this study mostly use histidine to coordinate the iron - but this would only account for one residue in the binding pocket per pdb. Therefore these results suggest that while cystidine may be well suited to coordinate the iron in heme, it is poorly suited to form any nonpolar interactions with the porphyrin ring, leaving the task up to other, more suitably/intensely nonpolar amino acids.

Moving away from discussing individual amino acid populations, what is especially notable of the data for heme-b is that nonpolar residues appear in much greater frequency than polar residues. Nonpolar interactions with heme are therefore more numerous than polar interactions; quite logical, given there are only two polar propionate groups on a large porphyrin ring that is otherwise nonpolar. Their multiplicity may also suggest that they are potentially of greater importance than previously thought. At the very least, these results suggest that polar interactions and coordination of the iron atom, while necessary for heme binding, are insufficient, and that nonpolar intercations and the population of nopolar residues in the binding pocket should be considered when examining the binding environment of heme.

#### 2.2.1.2 Comparison with Background Amino Acid Frequencies

While the frequencies of amino acids in the binding pocket have been discussed, it may also be of interest to compare against the background amino acid frequency, the general frequency of amino acids within the entire monomer. The degree to which this may affect the significance of the frequencies of the amino acids in the binding pocket is unclear - those amino acids are still employed and placed such as to bind the heme, rather than being a random assortment of residues. However, a in depth examination of simlarities and differences may reveal that some amino acids may simply be extremely highly conserved by chance and by virtue of their numerous population, rather than some chemical benefit.

Figure ?? documents the frequencies of amino acids overall within the monomer.

Leucine and alanine as in the binding pocket frequencies are highly frequent in the overall monomer. This may suggest their prevalence in the binding pocket may simply be due to a high pulation of leucine and alanine in hemoproteins.

However, after these two amino acids the tendencies in frequency for the binding pocket and the monomer at large diverge.

Glycine is in high frequency - likely due to more complex geometry e.g. helices outside the binding pocket. In interest of brevity, the remaining frequencies are summed up thus: the same trends that appear to exist in the binding pocket do not appear to exist in the monomer at large. While the order of frequencies in conserved binding pockets can be rationalized, justifying the overall frequencies in monomers invites significant speculation.

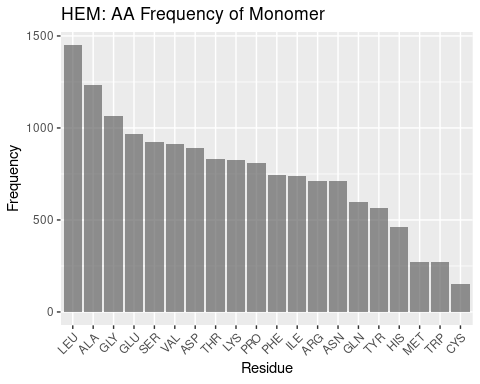


Figure 2.2: HEM: AA Frequency of Monomer

#### 2.2.1.3 Distributon of Amino Acids over Distance

After an exhaustive exploration of the relative frequencies of amino acids in the binding pocket, the figure below is fairly straightfoward. Figure 2.3 plots the distribution of amino acids in the binding pocket against their distance from the iron of the heme.

We find that only a few residues come in close contact (<4A) of the heme: Cys, His, Tyr. Most residues center their distribution at around 6A, although Lys seems more biased than the remaining residues to be a bit closer. Cystidine and histidine may be at least in part explained to be close due to their use as coordnating residues; histidine, being in greater frequency, may also be this close due to favorable interactions with the porphyrin ring.

The proximity of tyrosine however, is lkely more notable. It cannot form coordination bonds with the heme iron, but tyrosine residues do interact with the propionate groups; and these results suggest that of all potentially interacting polar/positively charged residues, tyrosine is the most likely at least to be in close proximity to the heme molecule. Whether this illustrates an extreme imortance of tyrosine to interact with propionate groups, or instead the need for tyrosine to be in close proximity in order to form such interactions, is beyond the scope of this study.

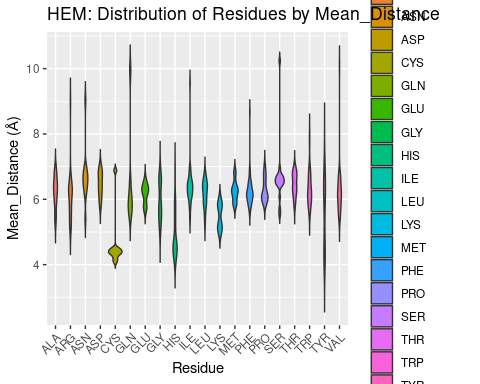


Figure 2.3: HEM: AA Distances

### 2.2.2 Heme-c

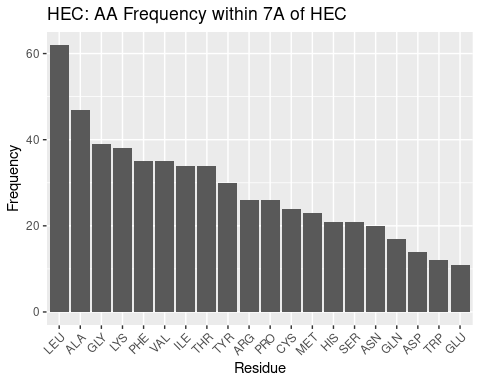


Figure 2.4: HEC: AA Frequency

Leucine and alanine again are highly frequent for HEC, followed by quite similar trends and therfore HEC will not be as thoroughly discussed as HEM. The most notable differences may be that GLY and CYS are in far higher frequency than in heme. Heme-c almost always covalently binds to CYS, and this may explain that frequency: but as for the high frequency of glycine, perhaps the covalent binding of CYS is sufficient for other interactions to be of ‘lower priority,’ and the flexibility and shape of the binding pocket to be prioritized, therefore leading to more glycines being included in order to shape the pocket favorably. A note on this last part: **I feel like this is grasping at straws without much support, add more qualifiers or remove**

#### 2.2.2.1 Comparison to background frequency

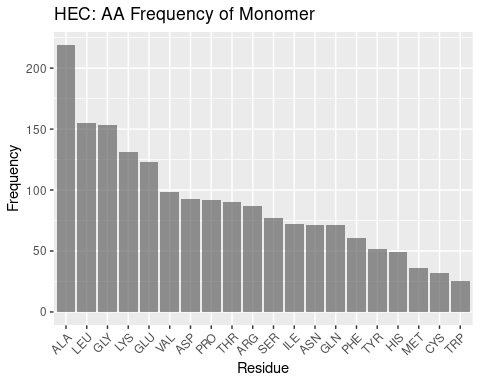


Figure 2.5: HEC: AA Frequency of Monomer

Generally, the heme-c monomer is similar to the heme-b monomer, with a high frequency of alanine and leucine, followed by a divergence in the frequency of amino acids and therefore a struggle to form any meaningful discussion when it comes to comparing the binding pocket frequencies against background frequencies.

#### 2.2.2.2 AA Distribution v. Distance

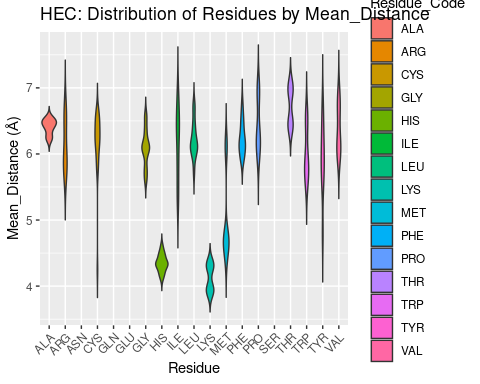


Figure 2.6: HEC: AA Distances

The distribution of amino acids over distance from the heme iron for HEC is similar to HEM, with some exceptions. Cys, His, Tyr again are amongst the closest residues to HEC, likely for the same reasons of very strong polar interactions or coordination. Additionally, cysteine forms covalent, thioester bonds with heme-c, providing further justification for its proximity. However, for heme-c, lysine and methionine also are very proximal. The methionine residues are nonpolar and **HELP, NOT SO SURE WHY THEY’D BE SO CLOSE** Lysine is polar and positively charged, and therefore in the case of HEC being covalently bonded, and therefore reliably, specifically, positioned, the environment may be such that a lengthy, polar, charged residue may also be positioned well enough to be consistently nearby and forming interactions with the propionate groups. **GRASPING AT STRAWS AGAIN** …. Maybe just leave it that for heme-c in particular, lysine and methionine appear favored. For good reason, they’re nonpolar. But it’s unclear why for heme-c these residues are employed but others aren’t. I’ve left other sections like that.

### 2.2.3 Verdoheme

**Got some conflicting stuff here**

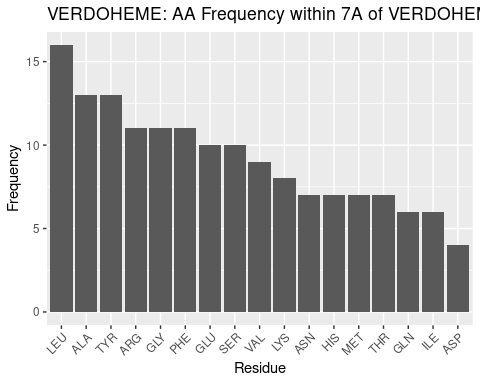


Figure 2.7: VERDOHEME: AA Frequency

Verdoheme is dissimilar from HEM and HEC above. This is fairly surprising, given that verdoheme is an intermediate in the binding pocket for heme within heme oxygenases. The results discussed below may be attributable to the small sample size of verdoheme PDBs (n=4, combining VEA and VER), and should be appreciated with some skepticism. Nonetheless, the results will be discussed.

Leucine and alanine are again most frequent, but after this results diverge. Tyrosine and arginine are next most frequent - surprising, given that this is still the same pocket that bound heme. The data for heme-b indicate more frequent nonpolar residues before tyrosine. It is possible that heme’s reorientation during its degradation moves it closer to another region of the binding pocket. Chemically, it may be that as heme is oxidized, there is greater need for polar interactions; this would help to explain the high frequency of polar residues.

Glycine is the next most frequent - it is in lower frequency, relatively, for heme-b…. **I’m gonna hold off on discussing this further until I double check the paper that described heme’s reorientation.**

#### 2.2.3.1 Comparison to background freq

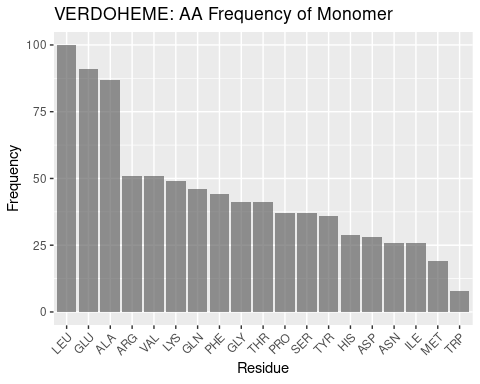


Figure 2.8: VERDOHEME: AA Frequency of Monomer

#### 2.2.3.2 AA Distribution over distance

**I’ll double check why so few data appear here. I suspect it may be simply that there is not enough data/atoms to pull and form a decent enough graph.** The highly conserved histidine for hemoproteins is exclusively within 5A for verdoheme. This result again suggests that at least some of the data for verdoheme may be highly biased because of the small sample size - heme-b data included a greater range for histidine. Nonetheless, this data suggests that verdoheme may be of different orientation during heme-b degradation, but is not spatially displaced far from the conserved binding site.

Glycine’s proximity to verdoheme is just weird WTF.

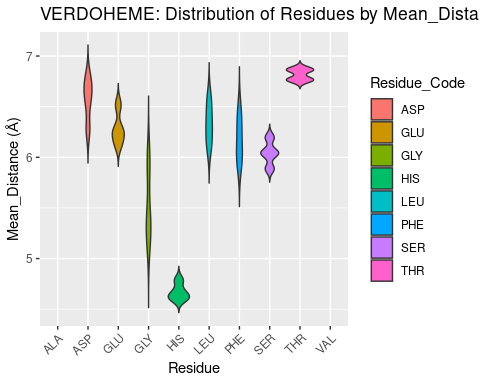


Figure 2.9: VERDOHEME: AA Distances

### 2.2.4 Siroheme

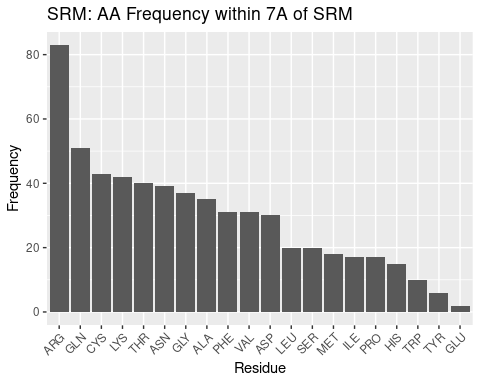


Figure 2.10: SRM: AA Frequency

Siroheme, with a structure highly dissimilar to the other heme molecules examined, should be expected to have a different amino acid frequency profile - and indeed we confirm this in our results.

Nonpolar residues are not the most abundant in the siroheme binding pocket. In fact, disproportionately frequent to the rest of the residues in the binding pocket is arginine. Siroheme is saturated with carboxyl and propionate groups; the entire porphyrin ring surrounded by polar, electronegative groups. And therefore a polar, positively charged amino acid such as arginine is reasonable to expect in the binding pocket - what is striking, however is the extreme preference for arginine; such a profile does not exist for the other hemes.

Arginine is followed by other polar amino acids: glutamine, cystine, lysine (positively charged), threonine, and asparagine; a more homogenous trend than seen for the other heme molecules, in that there are no nonpolar residues at all in this first… group of frequencies. **reword**. Though these results could be expected, they demonstrate the extent to which siroheme’s binding pocket is dominated by polar residues. The preference for arginine out of all polar amino acids may be attributed to its positive charge, and very low pKa; lysine also has a positive charge and a low pKa (~10.5), but arginine’s is much, much lower (~13.8) **fixme add citation** and it is able to form an additional hydrogen bond, and therefore reasonably dominates amongst the polar amino acids. Cysteine is used to coordinate the iron of siroheme, and while this did not significantly affect the frequency for other heme molecules, it is still possible this inflates the value for cysteine for siroheme.

After this group of polar amino acids, glycine is the next most frequent. Glycine has been situated at about a median frequency for other heme molecules, so perhaps its frequency here, slightly above the median, is of note. Only speculation is possible; perhaps ensuring the dominance of polar amino acids in the binding pocket requires extenive folding in the protein, therefore favoring glycine residues.

Finally we come to several nonpolar amino acids: alanine, phenylalanine, and valine. These amino acids define roughly the median of the frequency data. With all the polar groups on siroheme, it might be expected that only polar interactions would be desirable. However, the not miniscule frequency of these residues suggests nonpolar interactions still occur in the binding pocket; the porphyrin ring remains, as well as methyl groups and the small nonpolar portion of the carboxyl and propionate groups. It is perhaps in these areas that the nonpolar residues interact.

After these nonpolar residues the remaining frequencies do not follow a clear trend but regardless are discussed. After aspartate the remaining frequencies are considerably lower. This may be an artefact of a small sample size, or may suggest the remaining residues form, if any, less favorable interactions with the heme.

Aspartate appears next most frequently; it is a polar, negatively charged amino acid (at pH 7). Siroheme is saturated with other electronegative groups, and the repulsion of htese charges perhaps explains while aspartate, despite being a polar residue, appears not very frequently in the binding pocket.

Leucine is the first of the residues of diminished frequency. It is nonpolar. It, and, skipping a frequency, methionine, isoleucine, and proline, appear less frequently, and therefore are likely disfavored from forming the relatively few nonpolar interactions that do occur. Why is not clear - other small, nonpolar residues, and other lengthy nonpolar residues appear in the pocket in greater frequency. **double check pKa?**

Serine appears just less frequently than leucine, and in this context may likely be considered a polar residue that is not as strongly polar or positively charged and therefore less preferred to include in the binding pocket to form polar interactions with siroheme as other residues.

Histidine appears quite infrequently. As with siroheme, other, more strongly polar and perhaps less bulky residues are likely preferred.

Tryptophan is the least frequent nonpolar residue. The presence of a weak hydrogen bond and its size may preclude its inclusion in the binding pocket in lieu of more uniformly nonpolar residues that take up less space.

Tyrosine and glutamate are the least frequent polar residues. This is in stark opposition to the other heme molecules - tyrosine seemed to be favored for other heme molecules to form interactions with the propionate groups. Glutamate is also extremely infrequent, even in spite of its similarity to aspartate. Both are electronegative at pH 7 - glutamate’s extra carbon may provide sufficient steric hindrance to render it less favored. In either case, the infrequency of these residues and the tendencies of other, more intensely polar or nonpolar amino acids to be more populous, suggests tyrosine and glutamate, in the siroheme binding environment, do not interact strongly enough to be favored over other polar residues. **I’m gonna need to get the pka table maybe**

#### 2.2.4.1 Comparing to SRM AA Background Freq

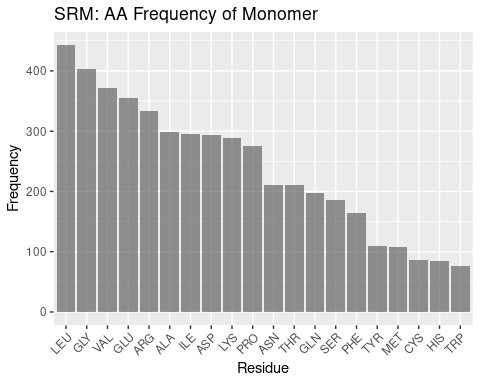


Figure 2.11: SRM: AA Frequency of Monomer

Compared to the other heme molecules, siroheme’s binding pocket amino acid frequencies are even more different than the background frequencies. Arginine is far and away the most frequent amino acid in the binding pocket - leucine is the most populous amino acid in the monomer overall, seeming to follow a trend amongst the hemoproteins examined so far. Again discussing the remainder of the frequencies of the monomer would be pure conjecture, but it is worthwhile to note that the pocket frequencies certainly appear unique against the background.

#### 2.2.4.2 Distance stuff

Residues appear less uniformly distributed over distance for siroheme binding pockets when compared against the distribution for other heme molecules. Cysteine is the only residue that comes within 5A of siroheme; it is used to coordinate the iron in siroheme, so this result is expected. The lack of other residues being within 5A, differing from other heme molecules, suggests the many carboxyl and propionate groups on siroheme prevent, or preclude the need for closer interaction except for coordinating residues.

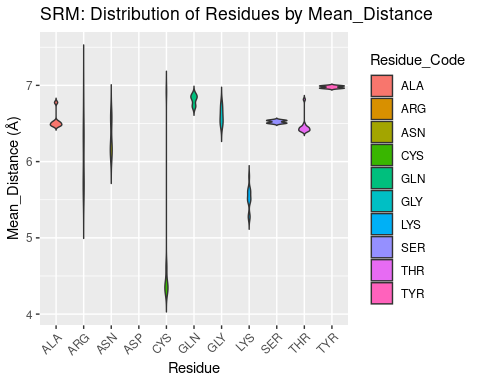


Figure 2.12: SRM: AA Distances

## 2.3 Volume Discussion

Figures can be found in Appendix 4.2.

**worthwhile to add SD measures? ’x% values fall within…** The utility of this result is somewhat dubious, at least within the context of this study. Volume results were rather spread out, with close agreement only found for heme-b. In general, volume for all heme molecules regardless of distance cutoff averaged about 1200 A³. This is somewhat contrived, and the result is not useful for elucidating the binding environment further; perhaps for other studies this result, or its lack of precision, may be informative.

## 2.4 Surface Areas

### 2.4.1 Surface Area of Heme Molecules

**Just going off solvent accessible, since that’s really all that’s of chemical importance. Can mention the excluded data is in the appendix**

Both solvent accessible and solvent excluded surface areas were calculated for heme molecules and binding pockets. The results are extremely similar and only solvent accesible surface area, a measure more practically interpreted into chemical phenonema, is discussed; figures and data for solvent excluded surface areas are available in Appendix (FIXME: insert reference)

**worth it to add SD etc? I think maybe, if time. Would leave raw data in appendix and add summary statistics in this section**

The solvent accessible surface area for all heme molecules themselves centers around values of 1000 A². This result is reasonable, given the similarity in size and structure of all heme molecules, in spite of the attached groups. Figures are shown below; extreme outliers have been removed from these figures but full data tables are available in (FIXME add appendix number). The outliers are likely artefacts of the method used to calculate surface area and potential conflicts with the method used to convert multimeric proteins to monomers. **worth to include this last statement?**

## 2.5 Surface Area of Binding Pockets

The surface area of binding pockets is more varied than the heme surface areas.

Heme-b and verdoheme, being highly similar molecules, with the same propionate groups, and one the derivative of the other, have quite similar surface areas, centering around 10,000-11,000 A². This is useful as a baseline to discuss the surface area of the binding pockets of the other two heme molecules below.

The surface area of the binding pocket of heme-c is considerably lower than that of heme-b and verdoheme. Its values center around 7500 A². Heme-c is bound covalently to the hemoprotein, forming thioether bonds with cysteine residues at two sites, excluding these sites from interacting with water molecules. **is this all that explains the reduction in SA? forgive my ignorance!**

The surface area of siroheme’s binding pocket is far greater than that for other heme molecules - values center around 21000 A². Siroheme’s extra groups do not appear to affect its own surface area, per above. However, it is effectively a very polar molecule and appropriately the binding pocket is highly saturated with very polar amino acids, as seen in the amino acid frequency analysis. The binding pocket is therefore completely different from the other heme molecules, and these populous, polar amino acids favorably interact with aqueous solvent, negating the need to bury any hydrophobic residues and reduce surface area.

## 2.6 Angular Data

\*\*I think I’ll just stick all of this in the appendix, including highlighting the clusters of data. Nothing can be discussed from it - it is interesting to note and perhaps I’ll include some examples below very briefly, but otherwise, I don’t know what value would be added to the report of “verdoheme always has his interact with a planar angle of 116 degrees”

Figures can be found in 4.7

These data, for all ligands, except potentially for heme-c, largely serve to compare as noise for the next section. The planar angles of all residues, falling within the upper distance cutoff of 7A, are plotted.

In the notable exception of heme-c, Figure **FIXME: insert figure name** seems to suggest that GLU, MET and LYS have fairly specific planar angles with the ligand. Lysine is effectively the median of amino acid frequency for heme-c, methionine is even less frequent and glutamine is the least frequent amino acid. For the latter two amino acids their tight range of planar angles is therefore likely an artefact of a small sample size of amino acids. However, for lysine the tight range of angles may be significant; this is dicussed further below.

## 2.7 Planar Angles of Closest Residues

Figures can be found in 4.8

Here, the three closest residues to the ligand in each PDDB and their planar angle to the ligand are plotted. Data are summarized below, and discussed below.

HEM has a fairly inconclusive set of data for this measure. GLU and GLN nearby HEM do appear to fall within a tight range, though, of approximately 75 degrees and 80 degrees respectively.

The data for residues nearby HEC diverge from what is found for all residues around HEC. The most agreement is found for ILE and LYS, with angles concentrated about 50 degrees and 75 degrees, respectively.

## 2.8 All CA-CB-Fe Angles

Figures can be found in 4.9

## 2.9 CA-CB-Fe Angles of Closest Residues

Figures can be found in 4.10

## 2.10 Limitations of the Study

Limited sample size

Limited experimental data to reference to verify

NO experimental data in this study to verify, all theoretical

Only one software package/few algorithms used to calculate all these properties. Others were evaluated but none are compared w.

Algorithms may introduce bias based on how they work e.g. all the bubbles

Arbitrary selection of parameters; some based on rule of thumb or visual evaluation but all or almost all arbitrary

Unknown if the qualities measured are truly the most critical for the heme binding. Some papers suggest other properties may also be important but cannot be calculated, at least right now, e.g. ionic bonding strength etc.

Visual examination itself to OK the parameters/algorithms can introduce bias

# 3 Conclusion

A knowledge gap in the binding environment for heme exists in the present literature. A high-throughput framework employing UCSF Chimera was constructed to process diverse sets of hemoproteins and output information about their binding pockets: amino acid frequencies and distances from heme, volume, surface area, angles. Data was gathered and predicted from representative and varied datasets for heme-b, heme-c, verdoheme, and siroheme, and their respective hemoproteins. R was used to analyze all data.

The results of this study are limited by the small sample size, but suggest that binding pockets for hemoproteins have some requirements for binding that may be overlooked, e.g. a high population of nonpolar amino acids for heme-b, heme-c, and verdoheme which may provide much of the interactions necessary to bind heme; in the case of siroheme there is a preponderance of polar amino acids. In either case both polar and nonpolar amino acids appear in force, suggesting both types of interactions are important to bind heme. Volumes are within reason. Surface area results reinforce, demonstrate the significance of the amino acid frequency results and show pockets have lesser or greater surface area depending on the involvemnt of covalent bonds or preferences for polar amino acids. Angular data, while not able to be interpreted into useful results, is produced and a framework provided for future study.

These results may be useful in the rational design of hemoproteins in the future, with the importance of nonpolar interactions in particular likely of great interest. The framework constructed can be applied to any list of PDBs and their respective ligands, thereby facilitating similar research for other proteins.

# (APPENDIX) Appendix

# 4 Figures

## 4.1 AA Frequency

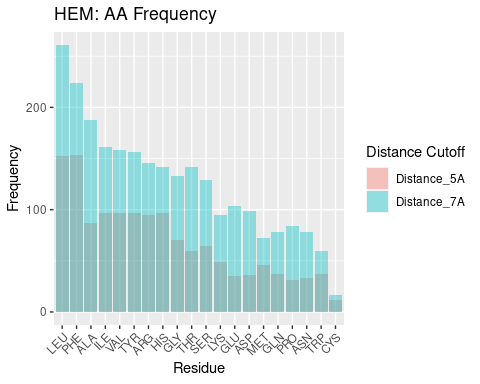


Figure 4.1: HEM: AA Frequency

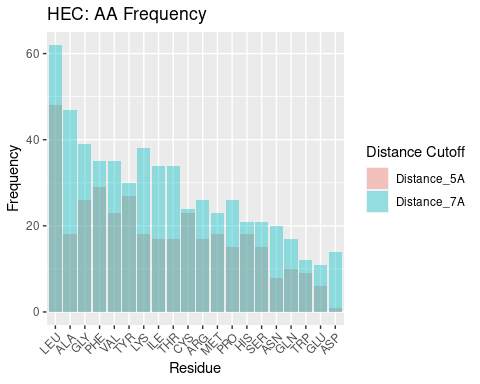


Figure 4.2: HEC: AA Frequency

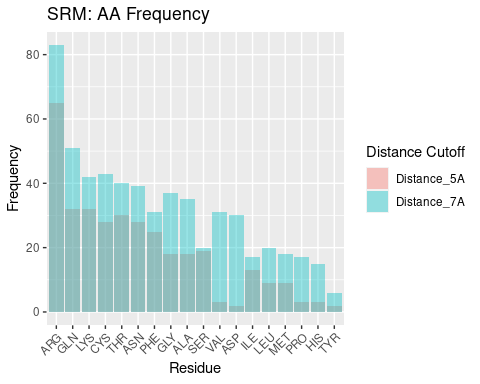


Figure 4.3: SRM: AA Frequency

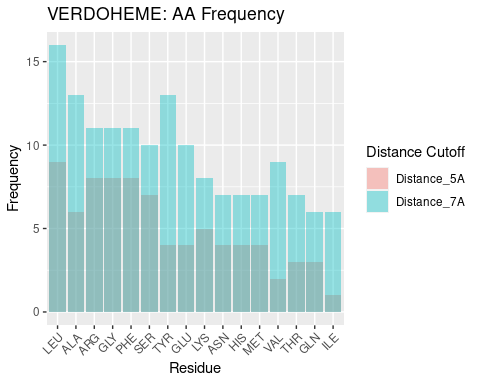


Figure 4.4: VERDOHEME: AA Frequency

## 4.2 Volume

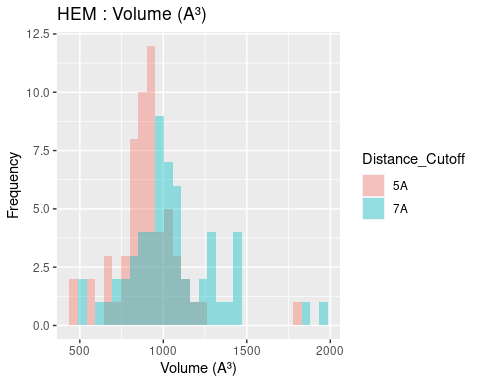


Figure 4.5: HEM: Volume

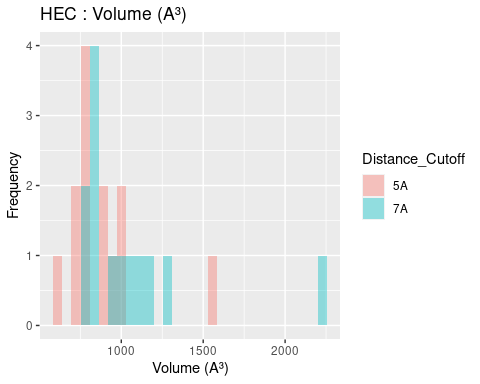


Figure 4.6: HEC: Volume

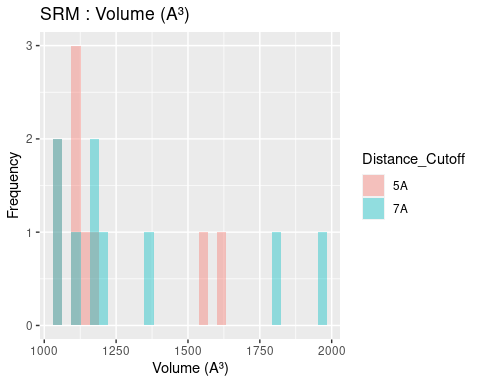


Figure 4.7: SRM: Volume

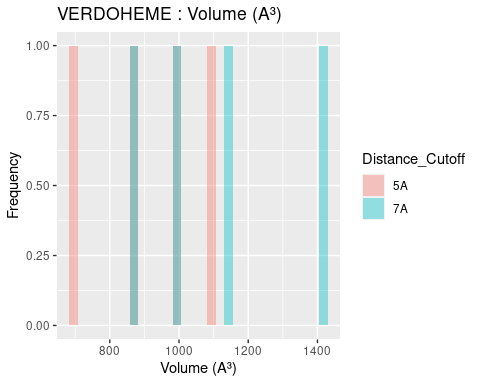


Figure 4.8: VERDOHEME: Volume

## 4.3 Ligand Excluded Surface Area

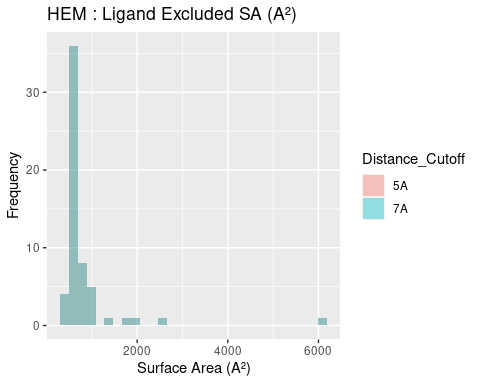


Figure 4.9: HEM: Ligand Excluded Suface Area

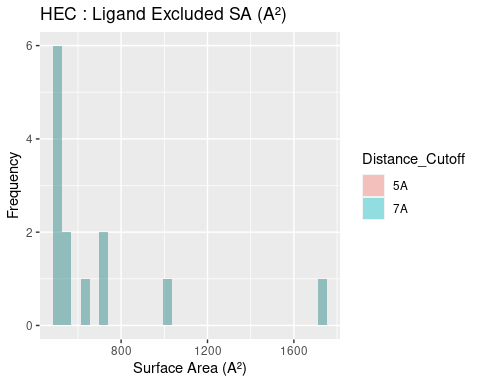


Figure 4.10: HEC: Ligand Excluded Suface Area

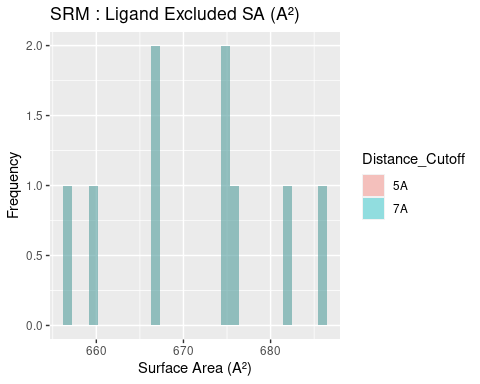


Figure 4.11: SRM: Ligand Excluded Suface Area

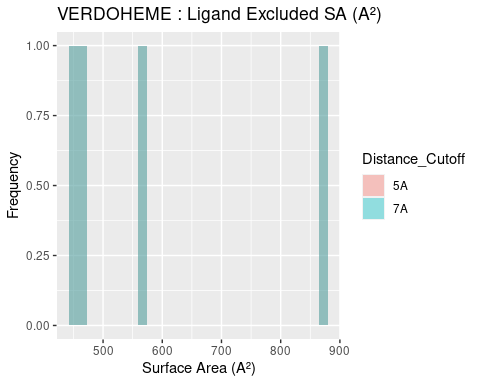


Figure 4.12: VERDOHEME: Ligand Excluded Suface Area

## 4.4 Ligand Accessible Surface Area

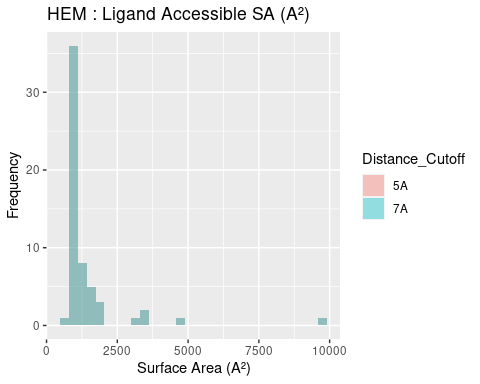


Figure 4.13: HEM: Ligand Accessible Surface Area

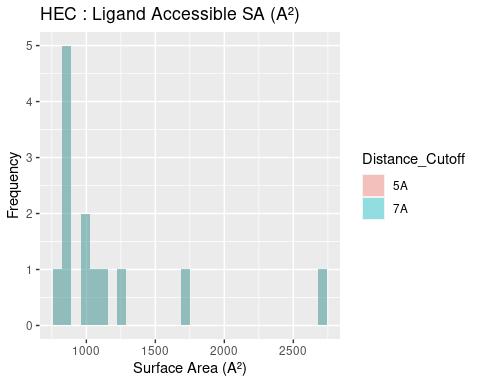


Figure 4.14: HEC: Ligand Accessible Surface Area

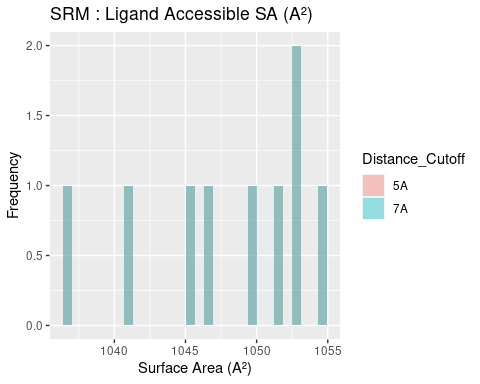


Figure 4.15: SRM: Ligand Accessible Surface Area

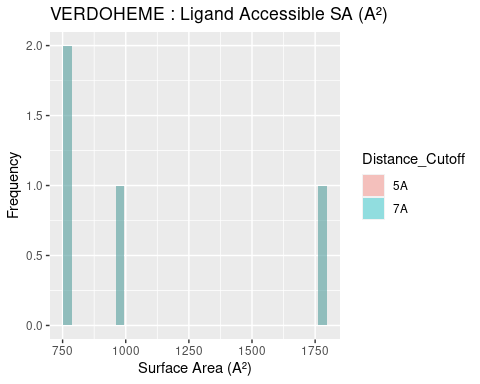


Figure 4.16: VERDOHEME: Ligand Accessible Surface Area

## 4.5 Pocket Excluded Surface Area

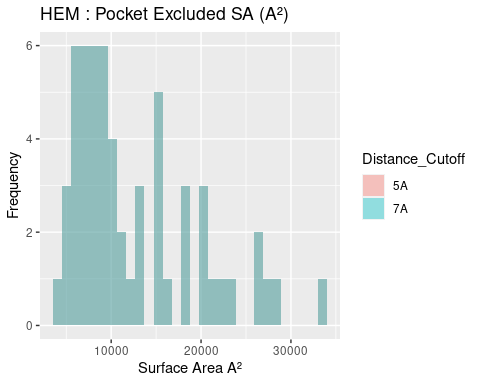


Figure 4.17: HEM: Pocket Excluded Surface Area

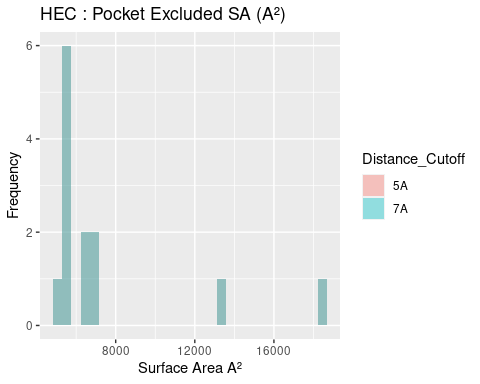


Figure 4.18: HEC: Pocket Excluded Surface Area

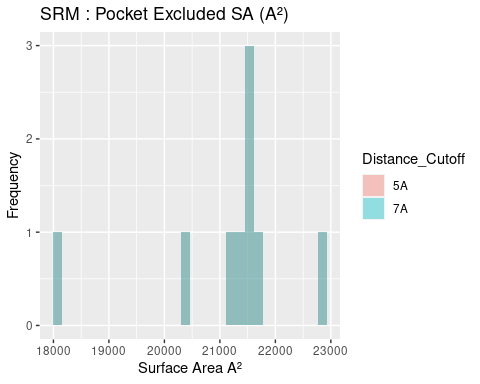


Figure 4.19: SRM: Pocket Excluded Surface Area

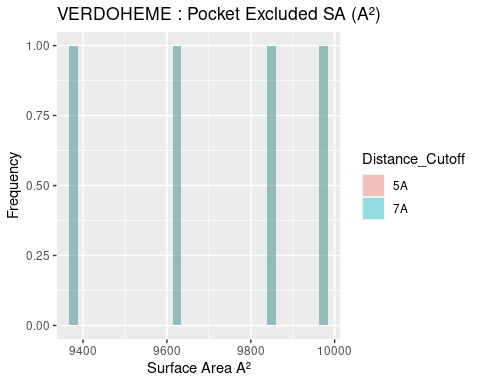


Figure 4.20: VERDOHEME: Pocket Excluded Surface Area

## 4.6 Pocket Accessible Surface Area

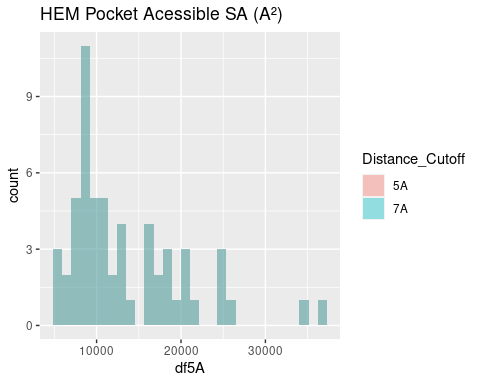


Figure 4.21: HEM: Pocket Accessible Surface Area

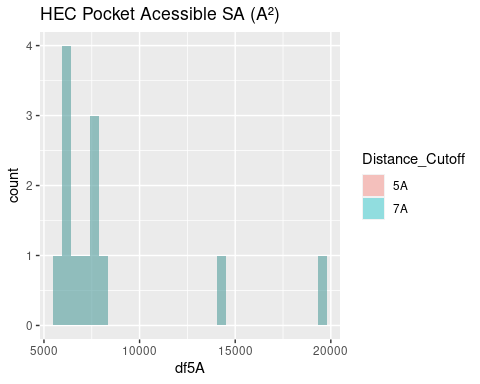


Figure 4.22: HEC: Pocket Accessible Surface Area

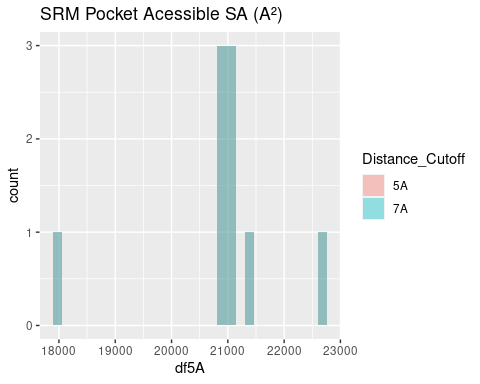


Figure 4.23: SRM: Pocket Accessible Surface Area

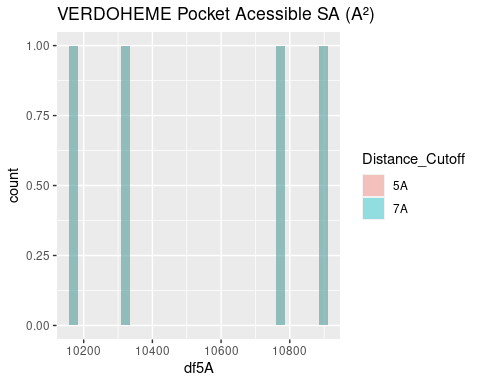


Figure 4.24: VERDOHEME: Pocket Accessible Surface Area

## 4.7 All Planar Angles

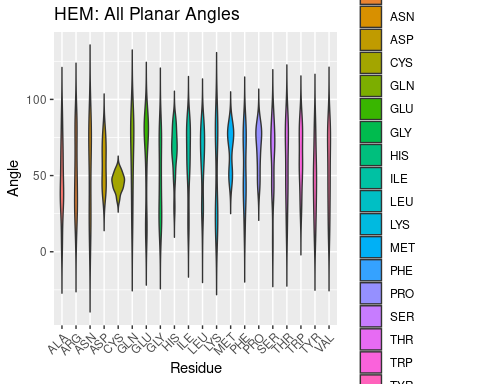


Figure 4.25: HEM: All Planar Angles

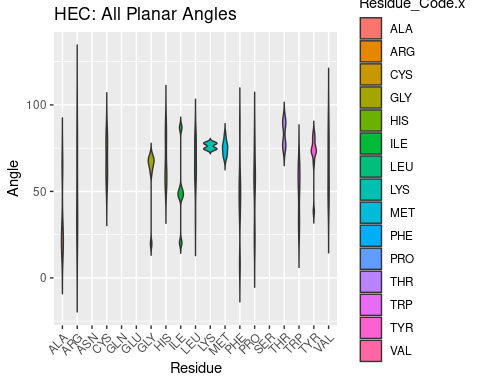


Figure 4.26: HEC: All Planar Angles

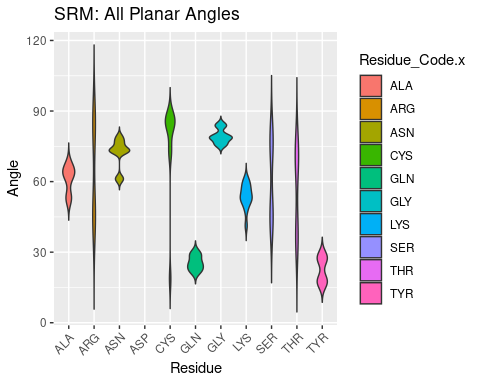


Figure 4.27: SRM: All Planar Angles

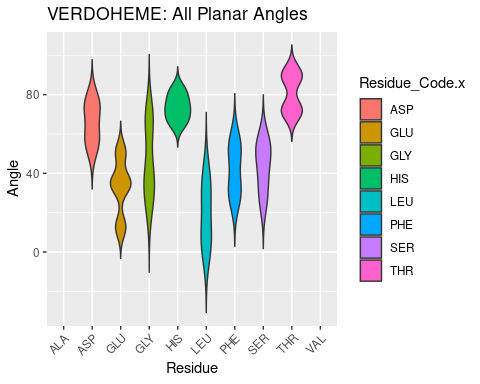


Figure 4.28: VERDOHEME: All Planar Angles

## 4.8 Planar Angles of Closest Residues

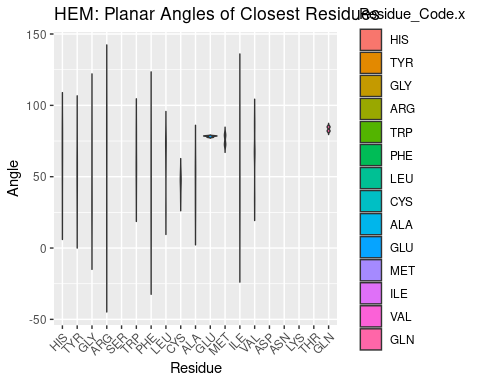


Figure 4.29: HEM: Planar Angles of Closest Residues

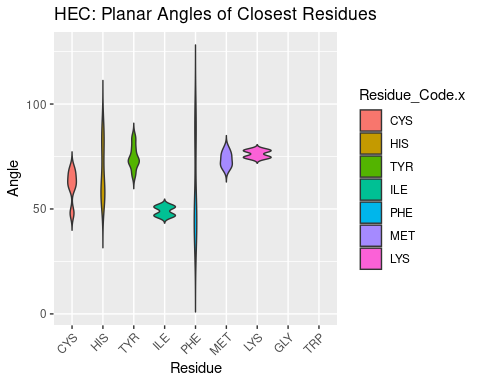


Figure 4.30: HEC: Planar Angles of Closest Residues

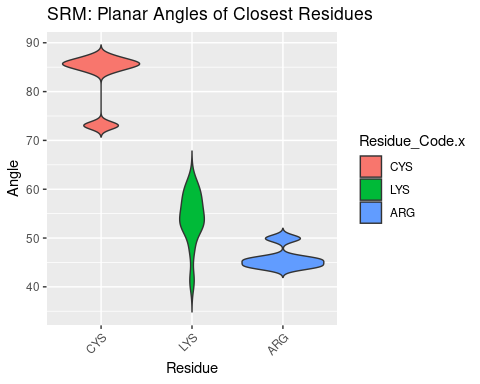


Figure 4.31: SRM: Planar Angles of Closest Residues

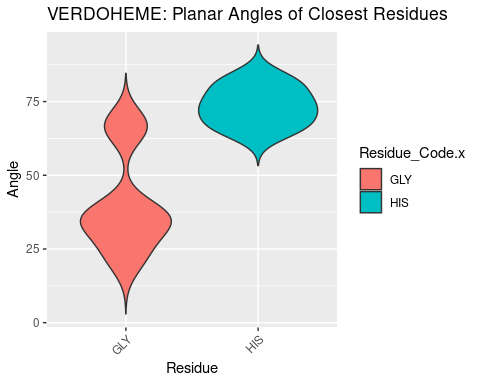


Figure 4.32: VERDOHEME: Planar Angles of Closest Residues

## 4.9 All CA-CB-Fe Angles

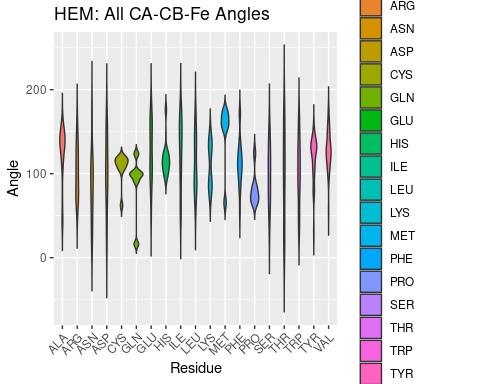


Figure 4.33: HEM: All CA-CB-Fe Angles

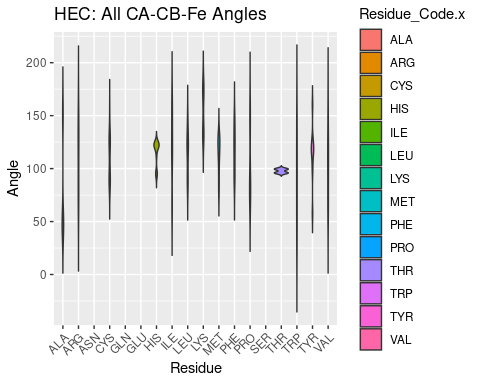


Figure 4.34: HEC: All CA-CB-Fe Angles

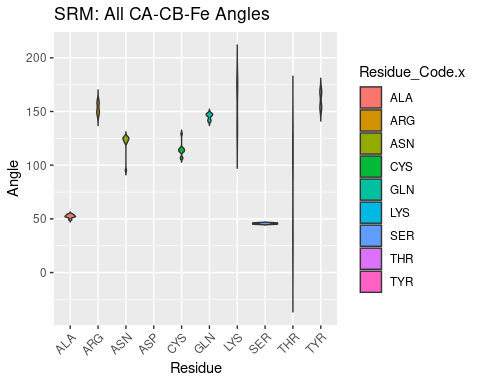


Figure 4.35: SRM: All CA-CB-Fe Angles

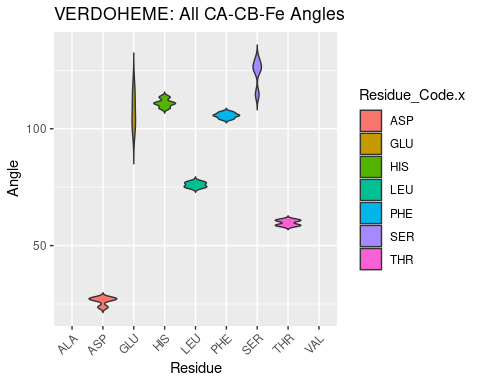


Figure 4.36: VERDOHEME: All CA-CB-Fe Angles

## 4.10 CA-CB-Fe Angles of Closest Residues

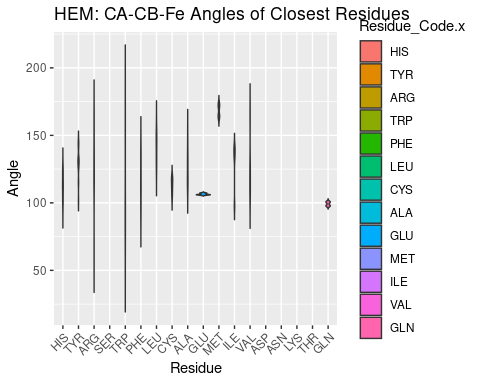


Figure 4.37: HEM: CACBFe Angles of Closest Residues

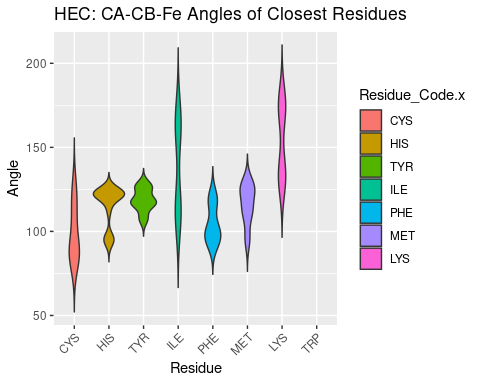


Figure 4.38: HEC: CACBFe Angles of Closest Residues

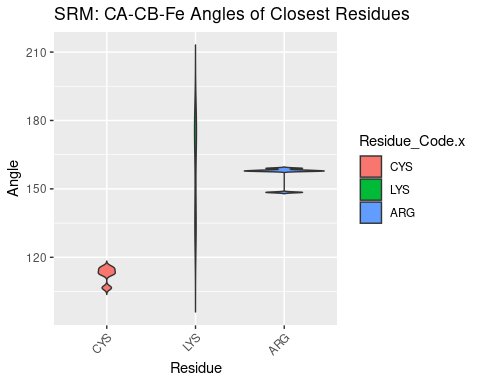


Figure 4.39: SRM: CACBFe Angles of Closest Residues

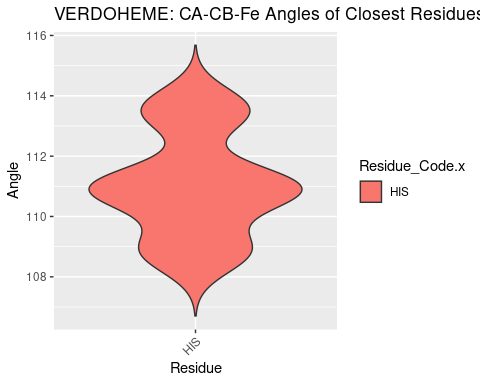


Figure 4.40: VERDOHEME: CACBFe Angles of Closest Residues

# 5 Tables

## 5.1 Molecule Names and Source Organisms

## 5.2 Volume and Surface Areas

## 5.3 All Planar Angles

## 5.4 Planar Angles of Closest Residues

## 5.5 All CA-CB-Fe Angles

## 5.6 CA-CB-Fe Angles of Closest Residues

# References

Araujo, Jesus A., Min Zhang, and Fen Yin. 2012. “Heme oxygenase-1, oxidation, inflammation, and atherosclerosis.” *Frontiers in Pharmacology* 3 JUL (July): 1–17. <https://doi.org/10.3389/fphar.2012.00119>.

Bowman, Sarah E. J., and Kara L. Bren. 2008. “The chemistry and biochemistry of heme c: Functional bases for covalent attachment.” *Natural Product Reports* 25 (6): 1118–30. <https://doi.org/10.1039/b717196j>.

Brânzanic, Adrian M. V., Ulf Ryde, and Radu Silaghi-Dumitrescu. 2019. “Why does sulfite reductase employ siroheme?” *Chemical Communications* 55 (93): 14047–49. <https://doi.org/10.1039/c9cc05271b>.

Du, Lei, Sheng Dong, Xingwang Zhang, Chengying Jiang, Jingfei Chen, Lishan Yao, Xiao Wang, et al. 2017. “Selective oxidation of aliphatic C-H bonds in alkylphenols by a chemomimetic biocatalytic system.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (26): E5129–37. <https://doi.org/10.1073/pnas.1702317114>.

Girvan, Hazel M., and Andrew W. Munro. 2016. “Applications of microbial cytochrome P450 enzymes in biotechnology and synthetic biology.” *Current Opinion in Chemical Biology* 31: 136–45. <https://doi.org/10.1016/j.cbpa.2016.02.018>.

Ishida, Manabu, Naoshi Dohmae, Yoshitsugu Shiro, Tadatake Oku, Tetsutaro Iizuka, and Yasuhiro Isogai. 2004. “Design and synthesis of de Novo cytochromes c.” *Biochemistry* 43 (30): 9823–33. <https://doi.org/10.1021/bi049546e>.

Kleingardner, Jesse G., and Kara L. Bren. 2015. “Biological Significance and Applications of Heme c Proteins and Peptides.” *Accounts of Chemical Research* 48 (7): 1845–52. <https://doi.org/10.1021/acs.accounts.5b00106>.

Lad, Latesh, Paul R. Ortiz De Montellano, and Thomas L. Poulos. 2004. “Crystal structures of ferrous and ferrous-NO forms of verdoheme in a complex with human heme oxygenase-1: Catalytic implications for heme cleavage.” *Journal of Inorganic Biochemistry* 98 (11): 1686–95. <https://doi.org/10.1016/j.jinorgbio.2004.07.004>.

Lai, Wenzhen, Hui Chen, Toshitaka Matsui, Kohei Omori, Masaki Unno, Masao Ikeda-Saito, and Sason Shaik. 2010. “Enzymatic ring-opening mechanism of verdoheme by the heme oxygenase: A combined X-ray crystallography and QM/MM study.” *Journal of the American Chemical Society* 132 (37): 12960–70. <https://doi.org/10.1021/ja104674q>.

Lalonde, Jim. 2016. “Highly engineered biocatalysts for efficient small molecule pharmaceutical synthesis.” *Current Opinion in Biotechnology* 42: 152–58. <https://doi.org/10.1016/j.copbio.2016.04.023>.

Li, Ting, Herbert L Bonkovsky, and Jun Tao Guo. 2011. “Structural analysis of heme proteins: Implications for design and prediction.” *BMC Structural Biology* 11. <https://doi.org/10.1186/1472-6807-11-13>.

Li, Zhong, Yuanyuan Jiang, F. Peter Guengerich, Li Ma, Shengying Li, and Wei Zhang. 2020. “Engineering cytochrome P450 enzyme systems for biomedical and biotechnological applications.” *Journal of Biological Chemistry* 295 (3): 833–49. <https://doi.org/10.1016/s0021-9258(17)49939-x>.

Poulos, Thomas L. 2014. “Heme enzyme structure and function.” *Chemical Reviews* 114 (7): 3919–62. <https://doi.org/10.1021/cr400415k>.

Sato, Hideaki, Yuichiro Higashimoto, Hiroshi Sakamoto, Masakazu Sugishima, Kenichi Takahashi, Graham Palmer, and Masato Noguchi. 2007. “Electrochemical reduction of ferrous -verdoheme in complex with heme oxygenase-1.” *Journal of Inorganic Biochemistry* 101 (10): 1394–99. <https://doi.org/10.1016/j.jinorgbio.2007.05.016>.

Tripathy, Baishnab C., Irena Sherameti, and Ralf Oelmüller. 2010. “Siroheme: An essential component for life on earth.” *Plant Signaling and Behavior* 5 (1): 14–20. <https://doi.org/10.4161/psb.5.1.10173>.