1
Results and Discussion

–>

1.1 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.

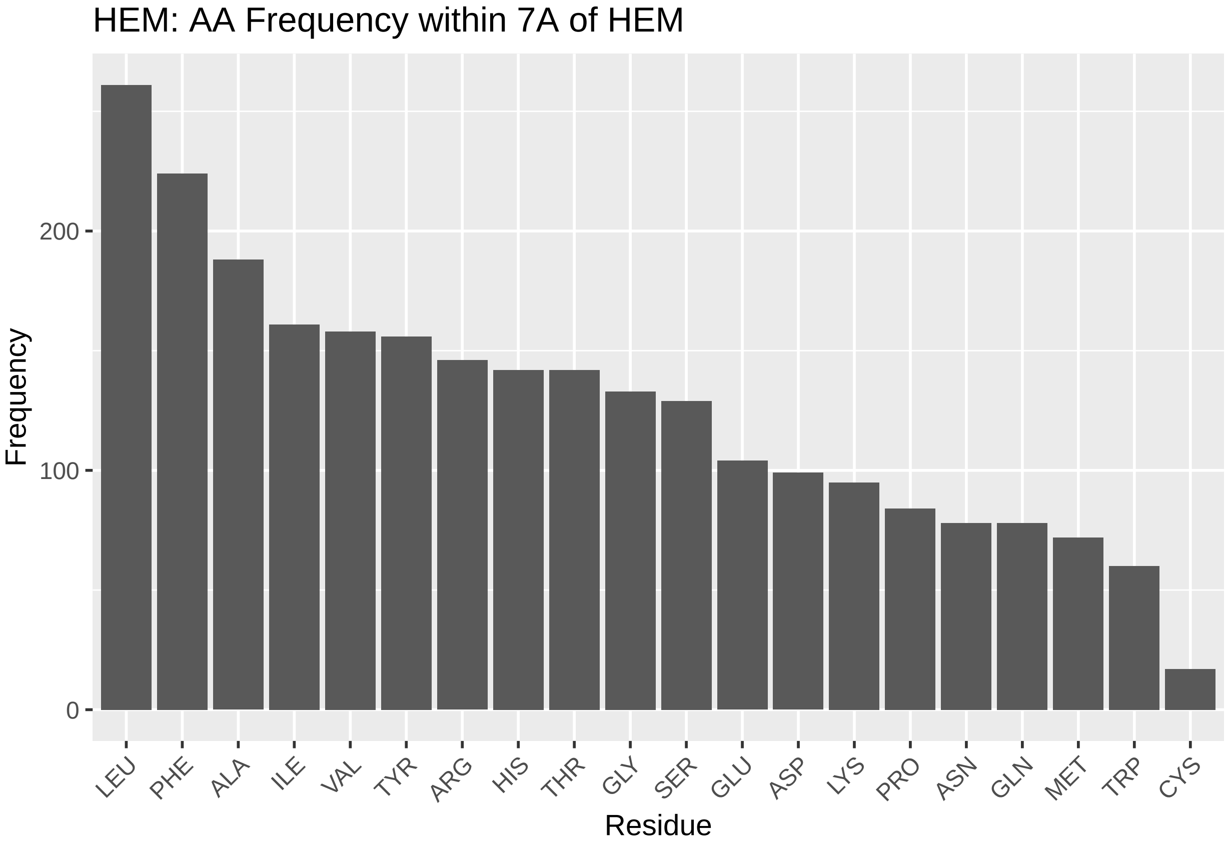

Figure 1.1: HEM: AA Frequency within 7A

1.2 AA Frequency

1.2.1 Heme-b

Amino Acid Frequencies in Binding Pocket

Figure 1.1 plots the frequency of each residue within 7A of heme-b. Immediately below is Figure 1.2, 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…

Table 1.1: HEM AA Freq

Residue	Freq
LEU	261
PHE	224
ALA	188
ILE	161
VAL	158

Table 1.1: HEM AA Freq *(continued)*

Residue	Freq
TYR	156
ARG	146
HIS	142
THR	142
GLY	133
SER	129
GLU	104
ASP	99
LYS	95
PRO	84
ASN	78
GLN	78
MET	72
TRP	60
CYS	17

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.

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.

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 1.3 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

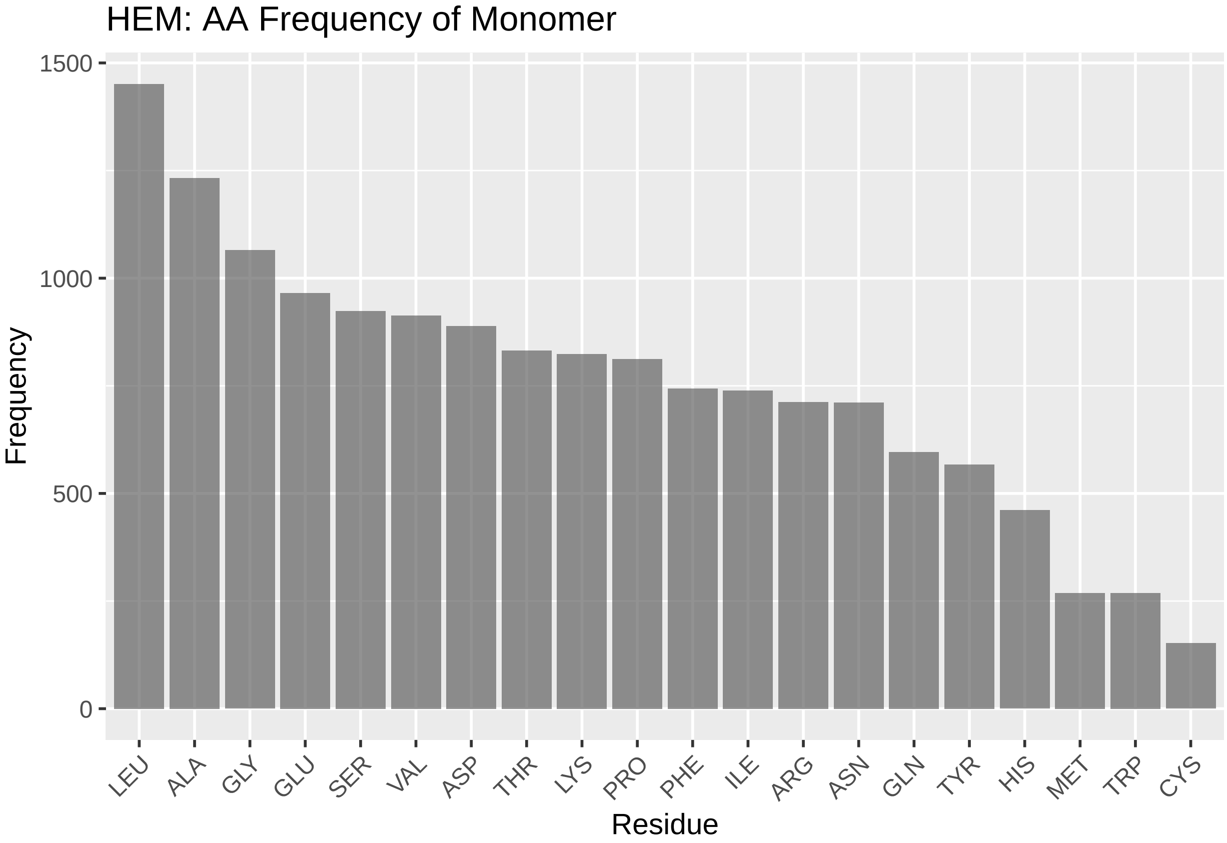

Figure 1.2: HEM: AA Frequency of Monomer

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.

1.2.2 Heme-c

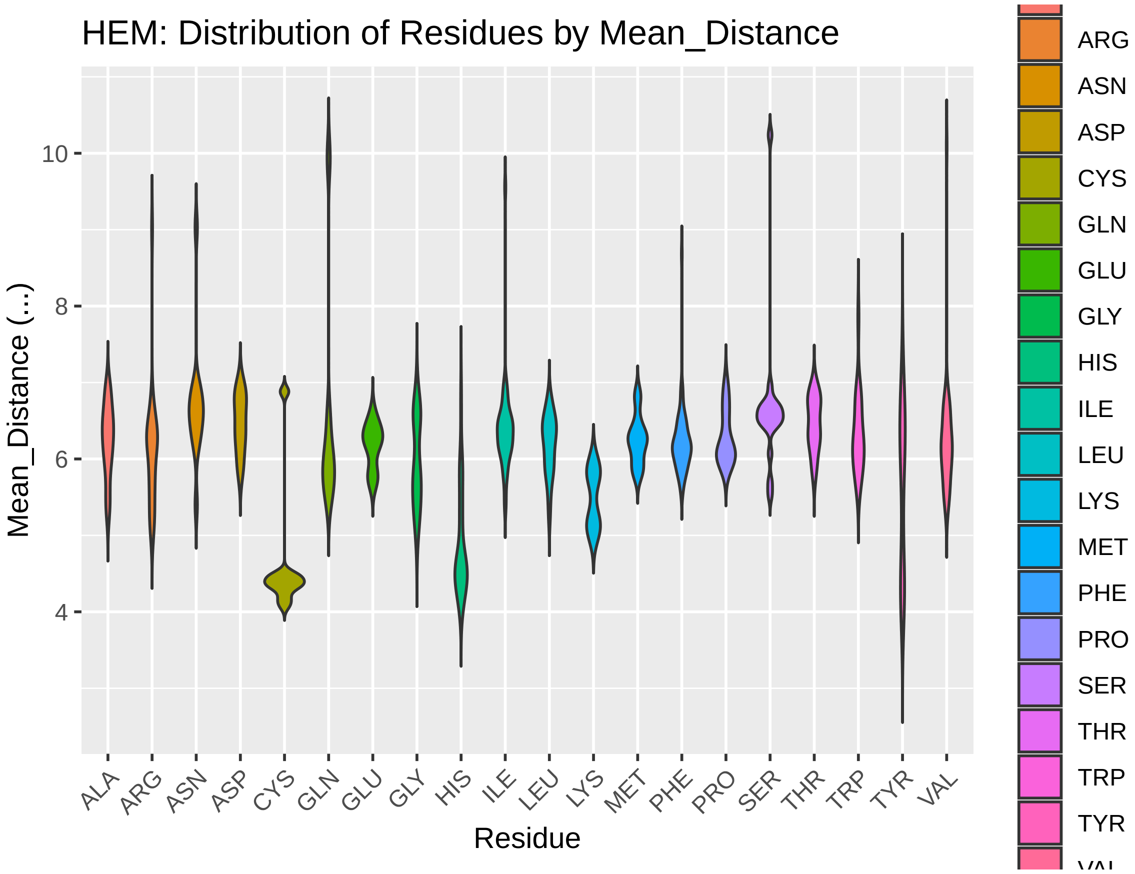

Figure 1.3: HEM: AA Distances

Table 1.2: HEC AA Freq

Residue	Freq
LEU	62
ALA	47
GLY	39
LYS	38
PHE	35
VAL	35
ILE	34
THR	34
TYR	30
ARG	26
PRO	26
CYS	24
MET	23
HIS	21
SER	21

Table 1.2: HEC AA Freq *(continued)*

Residue	Freq
ASN	20
GLN	17
ASP	14
TRP	12
GLU	11

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**

Comparison to background frequency

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.

AA Distribution v. Distance

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

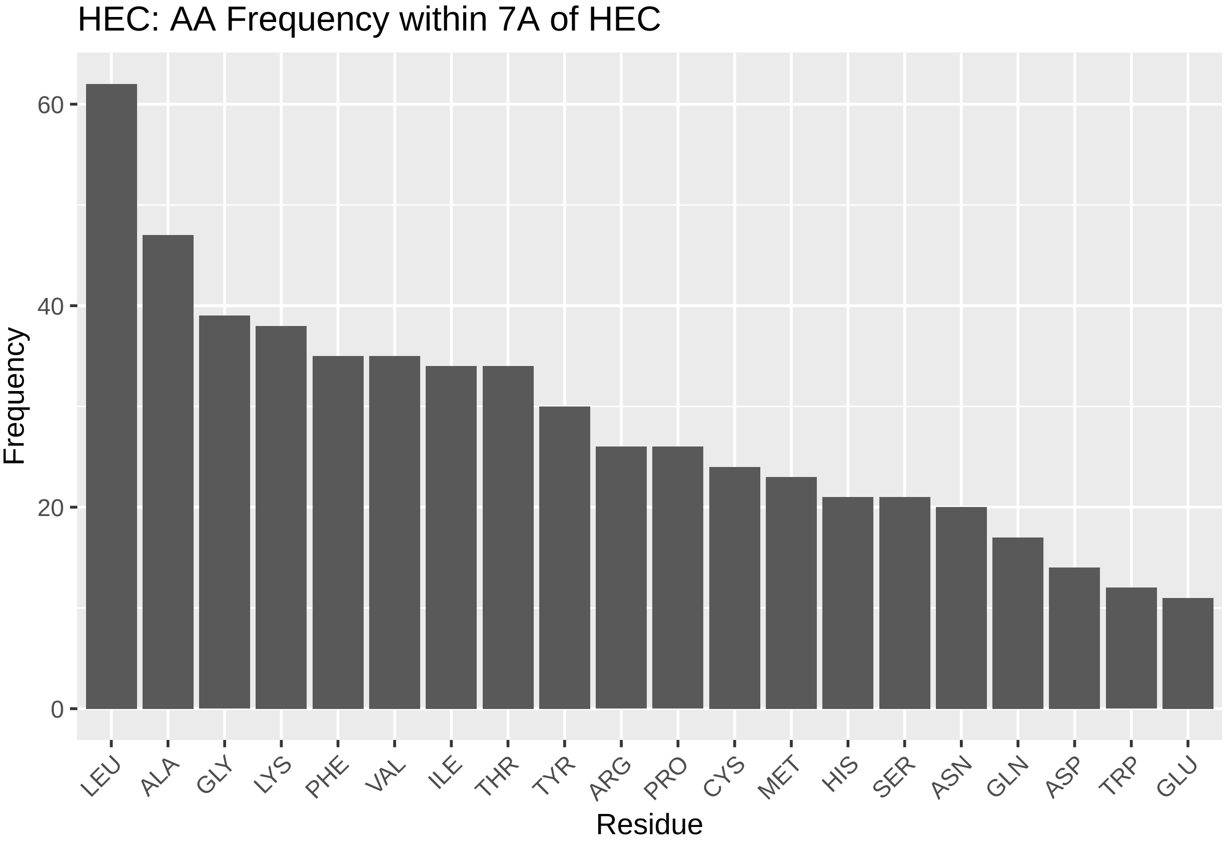

Figure 1.4: HEC: AA Frequency

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.

1.2.3 Verdoheme

Got some conflicting stuff here

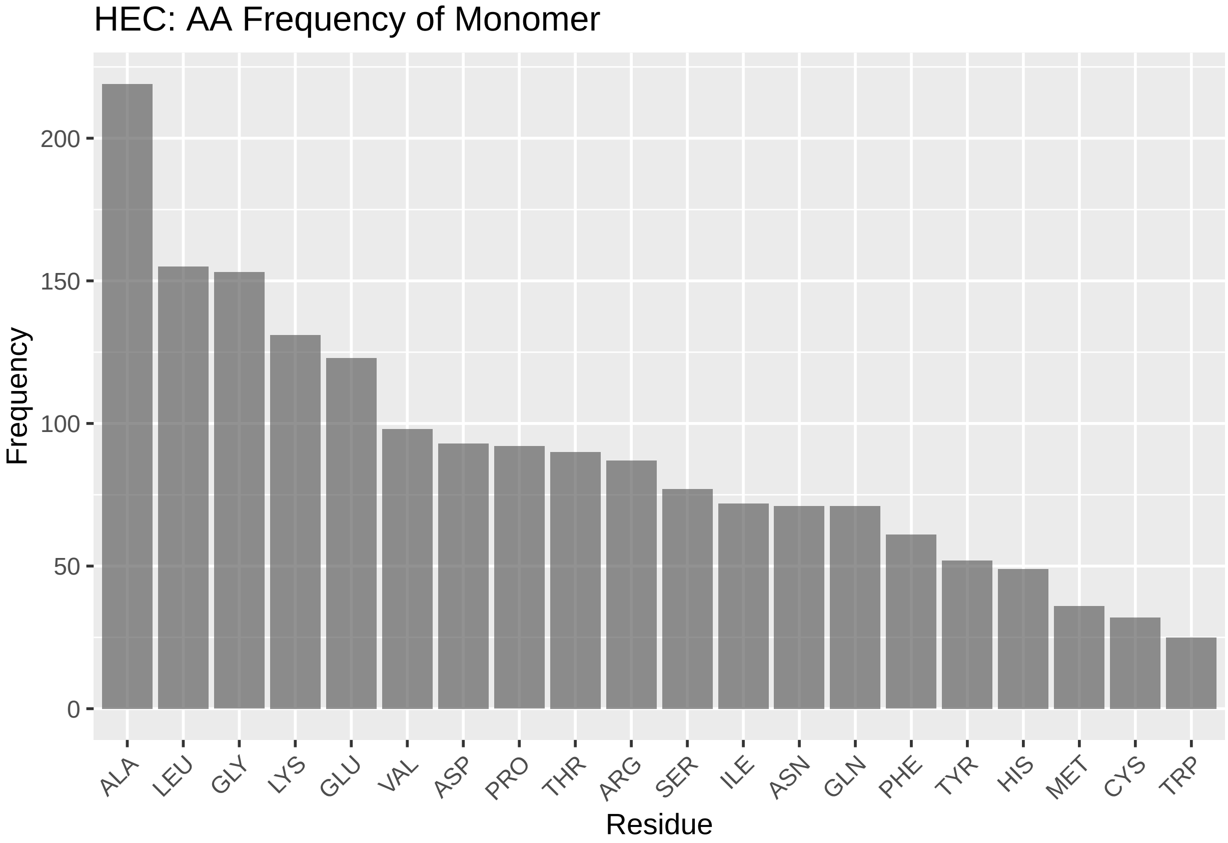

Figure 1.5: HEC: AA Frequency of Monomer

Table 1.3: VERDOHEME AA Freq

Residue	Freq
LEU	16
ALA	13
TYR	13
ARG	11
GLY	11
PHE	11
GLU	10
SER	10
VAL	9
LYS	8
ASN	7
HIS	7
MET	7
THR	7
GLN	6
ILE	6

Table 1.3: VERDOHEME AA Freq *(continued)*

Residue	Freq
ASP	4

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.**

Comparison to background freq

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.

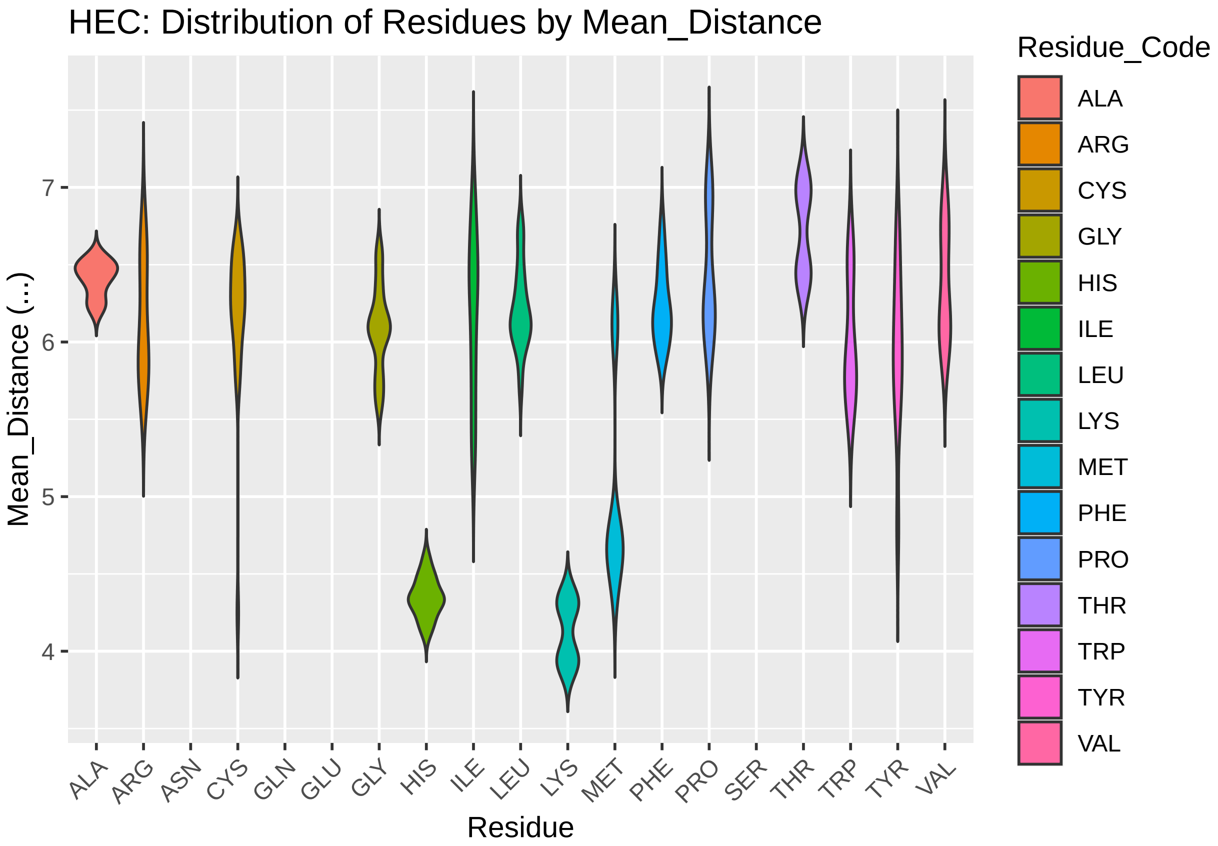

Figure 1.6: HEC: AA Distances

Glycine's proximity to verdoheme is just weird WTF.

1.2.4 Siroheme

Table 1.4: SRM AA Freq

Residue	Freq
ARG	83
GLN	51
CYS	43
LYS	42
THR	40
ASN	39
GLY	37
ALA	35
PHE	31
VAL	31
ASP	30
LEU	20

Table 1.4: SRM AA Freq *(continued)*

Residue	Freq
SER	20
MET	18
ILE	17
PRO	17
HIS	15
TRP	10
TYR	6
GLU	2

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,

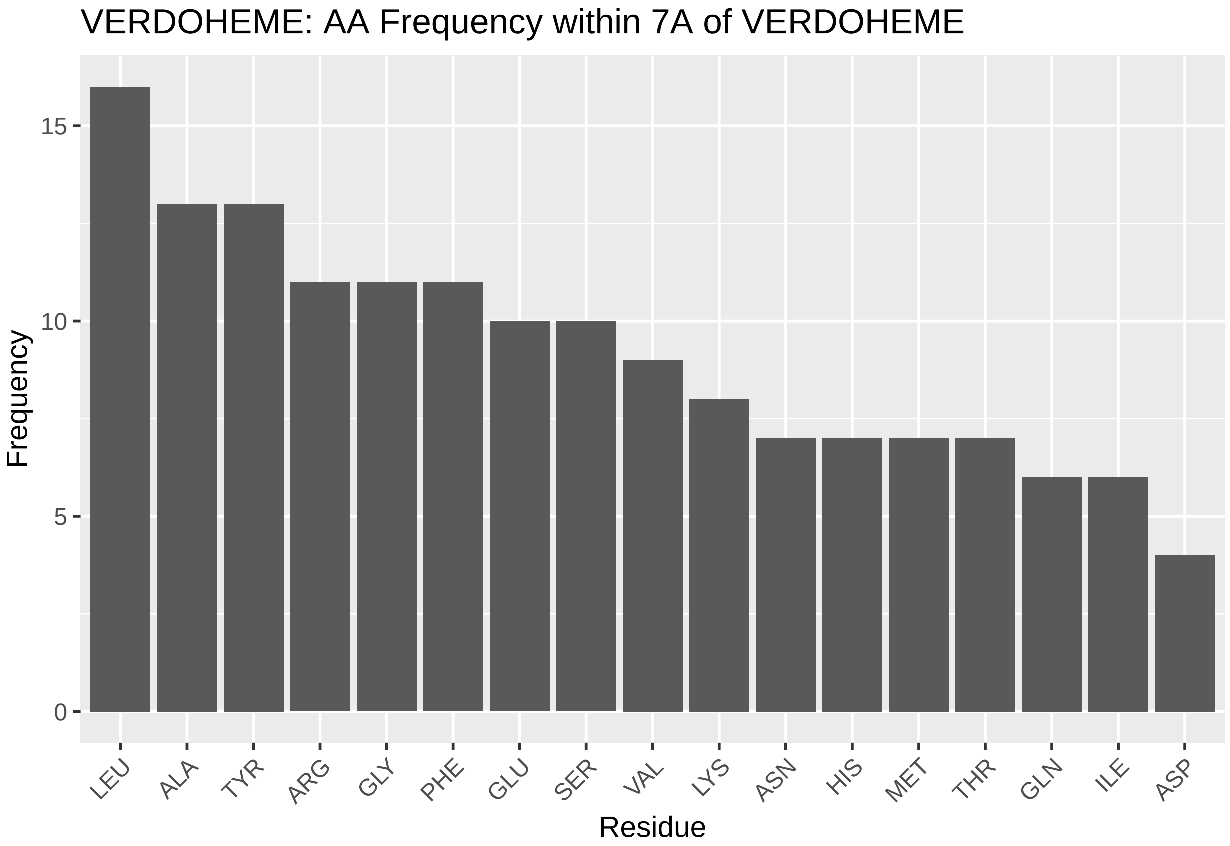

Figure 1.7: VERDOHEME: AA Frequency

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.

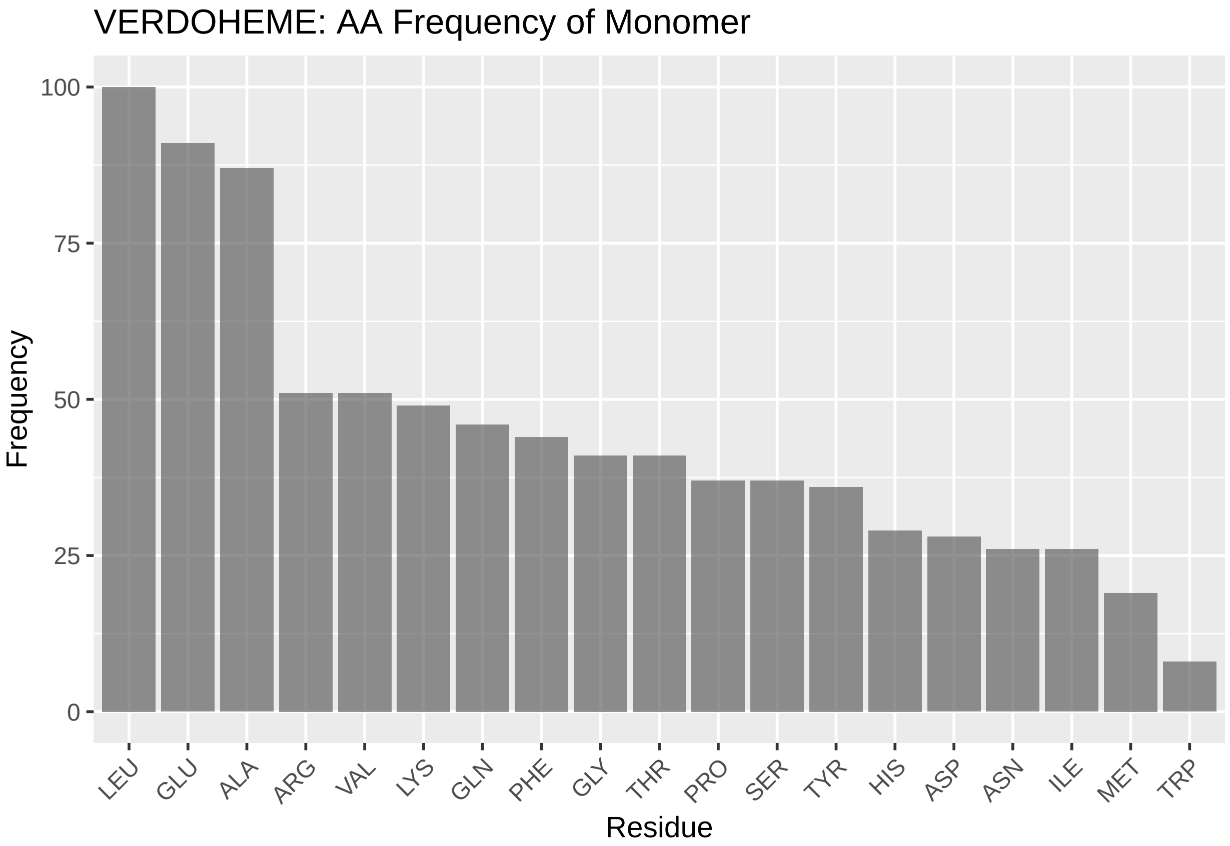

Figure 1.8: VERDOHEME: AA Frequency of Monomer

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?**

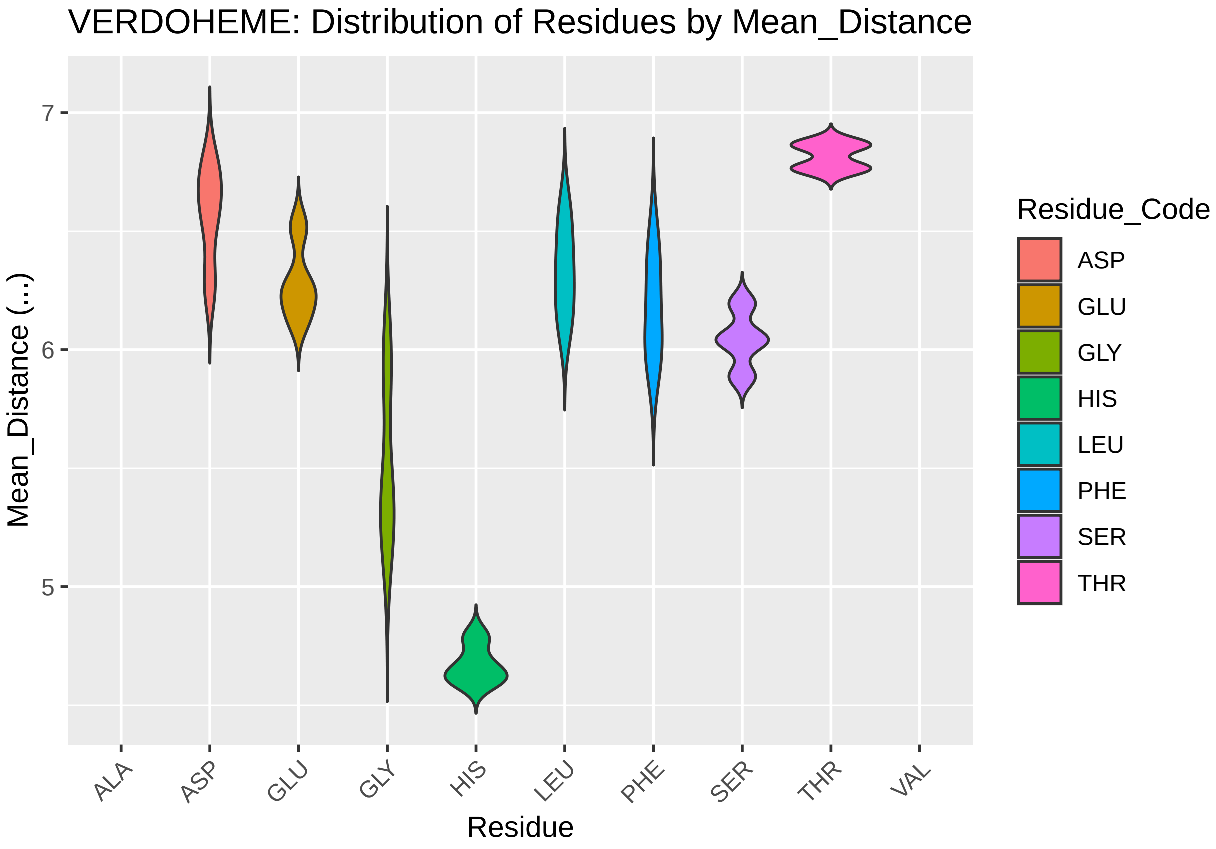

Figure 1.9: VERDOHEME: AA Distances

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

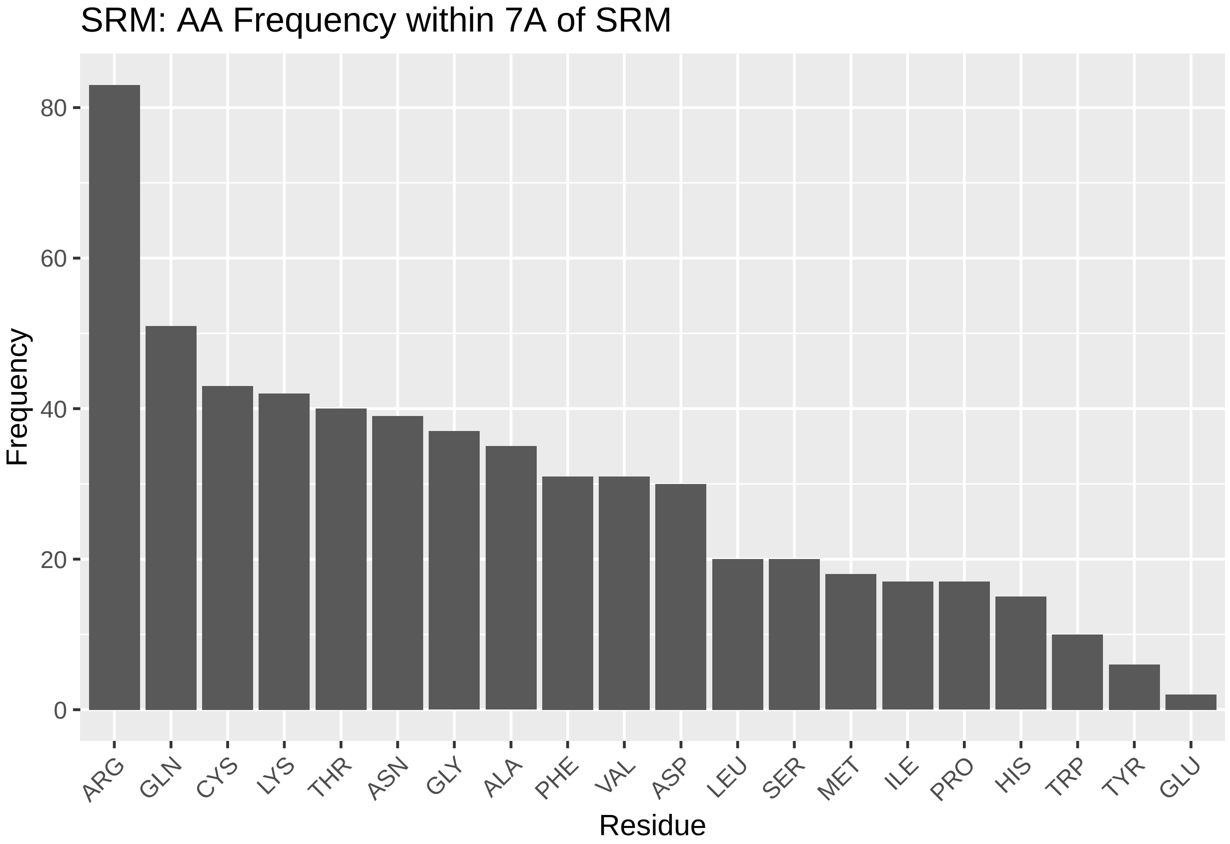

Figure 1.10: SRM: AA Frequency

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**

Comparing to SRM AA Background Freq

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.

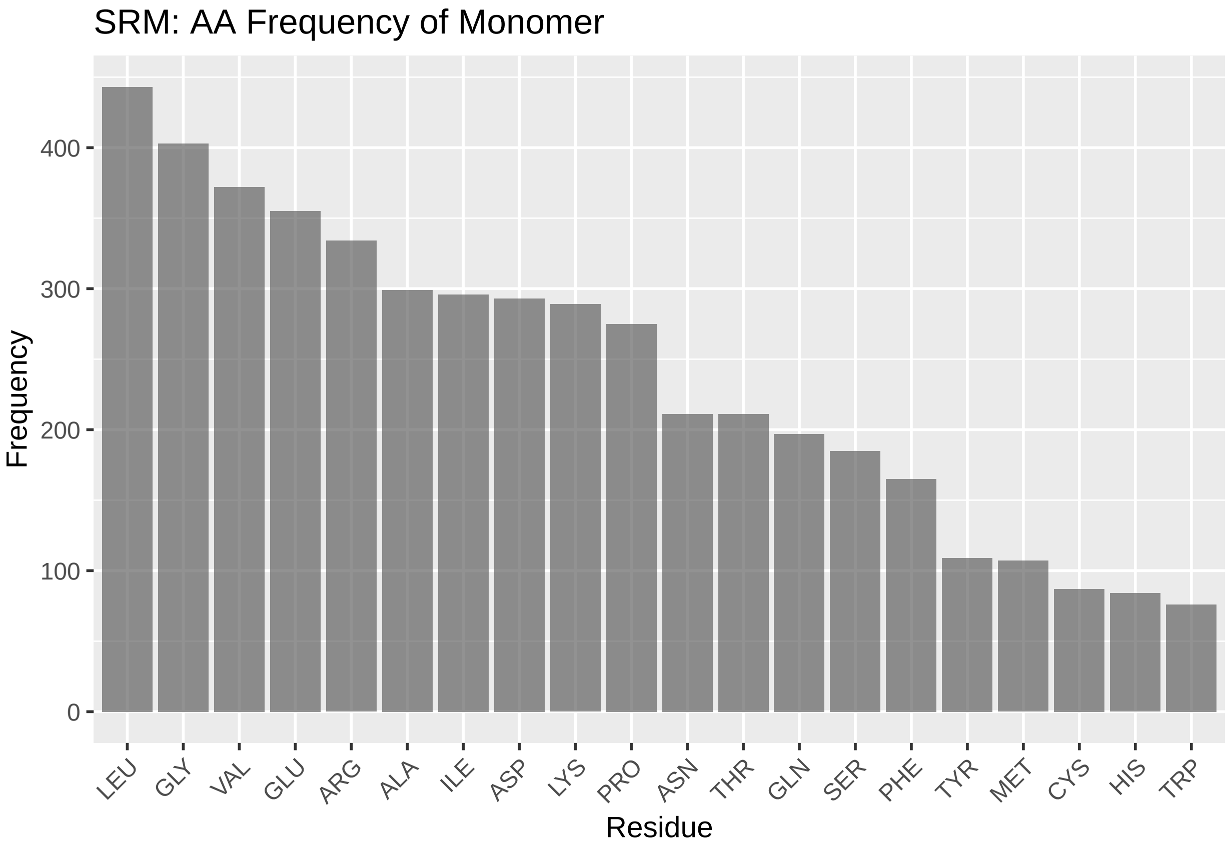

Figure 1.11: SRM: AA Frequency of Monomer

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.

1.3 Volume Discussion

Figures can be found in Appendix **??**.

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

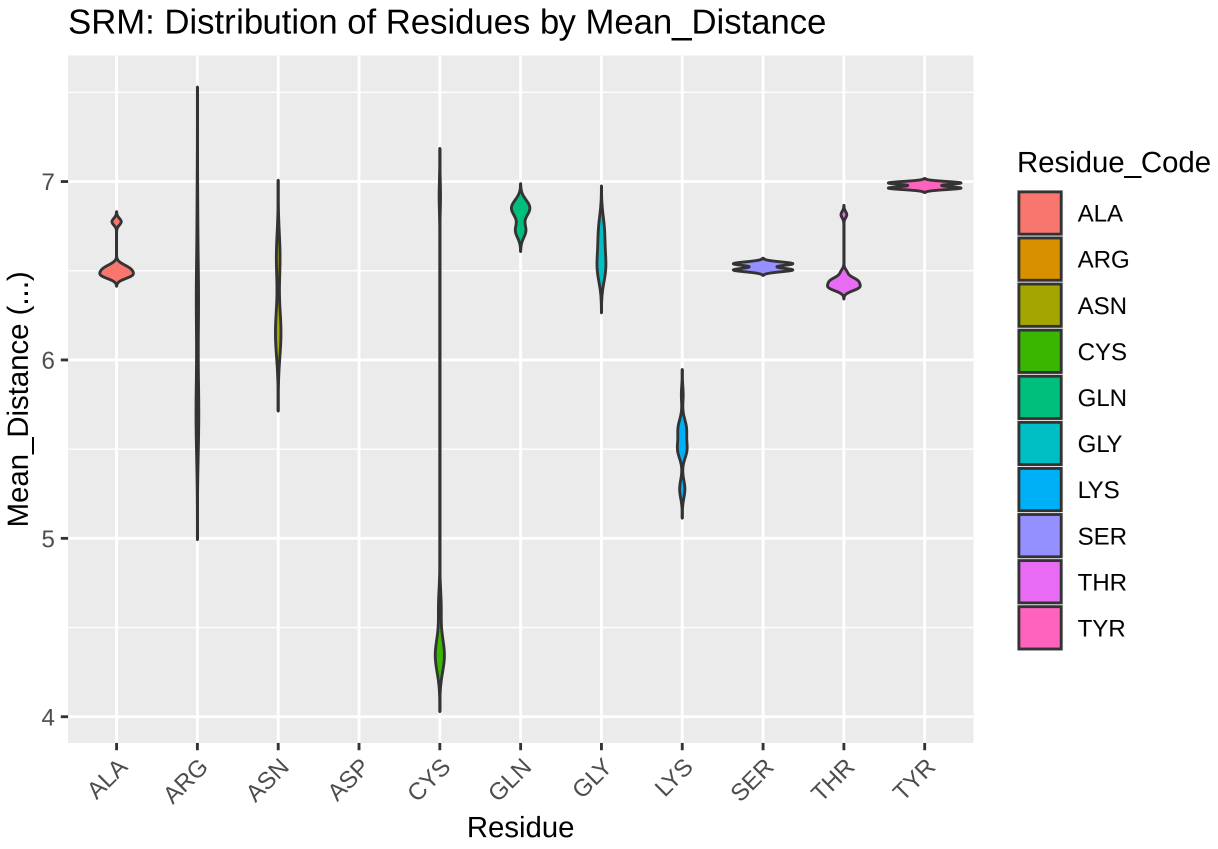

Figure 1.12: SRM: AA Distances

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 A^3. 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.

1.4 Surface Areas

1.4.1 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 A^2. 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?**

1.5 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 A^2. 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 A^2. 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 A^2. 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.

1.6 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 **??**

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.

1.7 Planar Angles of Closest Residues

Figures can be found in **??**

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.

1.8 All CA-CB-Fe Angles

Figures can be found in ??

1.9 CA-CB-Fe Angles of Closest Residues

Figures can be found in ??

1.10 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