I'm gonna need like 10 Irish Car Bombs, or margaritas when this is all over
UAB

Pat the Great and Powerful

29 June 2021

Abstract

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.

See? Not so bad of a worst-case scenario. Just, an unusual sentiment to see in modern science.

Lay Summary

Proteins are biological molecules that perform essential functions in our bodies and in all living things. They are responsible for everything from transporting oxygen in our blood to photosynthesis in plants. They can be extracted from living things and cultured in laboratories and factories. They can then be used as drugs, or be used to perform functions in industrial processes.

Many proteins require additional molecules to perform their function, and these molecules must be bound, docked to the molecule like a boat in a harbor. The molecules can be bound to the protein inside a specialized space, or pocket, within or on the surface of the protein.

One class of proteins is hemoproteins - proteins that use the molecule heme to perform their function. The heme is critical to the proper function of these proteins. Heme enables many specific chemical reactions to be performed, or assisted by the protein. In hemoglobin, in blood, the heme molecule allows the overall protein of hemoglobin to carry and transport oxygen around the body.

But the specifics of the pocket that binds heme in these hemoproteins is not well understood. What are the conditions inside the pocket? Is it a very specific size or can it vary? Is there anything in common among the pockets of many different hemoproteins? These are the questions this research hoped to answer; or at the very least, provide some data and lay the groundwork for others to continue the research later.

This investigation was not conducted in a lab. Rather, using various software packages and the 3D structures of hemoproteins published in a database, various properties of these hemoproteins were calculated. The data produced were analyzed with various statistical methods to extract potentially useful information.

Overall, the following was found: Grad school sucks but bravas are delicious.

Acknowledgments

In case anyone reads this in the future, some context may be appreciated: I attended and completed this Master's during the COVID-19 global pandemic from September 2020 to September 2021.

Thanks professors

Thanks lab

Thanks UAB

Thanks Spain, and Catalonia, allowing me in and then also having public health measures unlike Donny's America

Thanks classmates

Thanks fam, friends

Thanks to the media and the creators of media that facilitated the survival of my sanity through the pandemic.

Finally, I'd like to quote a well-known artist from California. He was referencing his own work, but I wholly identify with his appreciation for the subject of his esteem:

"Last but not least, I wanna thank me. I wanna thank me for believing in me. I wanna thank me for doing all this hard work. I wanna thank me for having no days off. I wanna thank me for, for never quitting. I wanna thank me for always being a giver, and trying to give more than I receive. I wanna thank me for trying to do more right than wrong. I wanna thank me for just being me at all times."

– Calvin Cordozar Broadus Jr.

Contents

List of Figures

List of Tables

Introduction

Proteins may catalyze reactions, and many require ligands to enable their chemistry. A significant portion of all proteins, approximately 40%, require a metallic group as a ligand in order to function correctly - these proteins are known as metalloproteins.

One of these metallic groups is heme. Heme is a member of the porphyrin family, a group of molecules capable of catalyzing a broad range of reactions. Heme can catalyze many different reactions and is present in many proteins. However, the underlying structural requirements to host heme in a protein are not well understood. [MAY ADD CITATIONS]

There have only been a handful of studies dedicated to understanding the structure-chemical relationship between heme and the proteins that use heme for their chemistry (these proteins are known as hemoproteins).

In the most significant previous work, approximately 125 hemoproteins were studied[2]. Although pdbs were thoroughly examined and the datasets were culled, the sample size of this study is very small compared to the amount of hemoproteins available in the pdb a decade later (10,000 HEM-containing proteins and xx). The dataset is also limited in that there is a somewhat homogenous group of proteins examined (?). The characteristics examined were limited to: xx.

It is hypothesized that the following characteristics all have an impact on the binding of heme and function of the hemoprotein: XXXXXXXXX.

In this study, some of these characteristics were examined. They include: XX. The remainder are thus far not feasible to calculate.

All of these characteristics have implications in the field of protein engineering or basic research into hemoproteins. Examples of the uses of these results include [SUPER BLOOD STUDY] and [OTHER PROTEIN ENGINEERING STUFF]. Not sure how much we can reference those other papers besides doing that besides in the conclusion.

Notable results from some of the prior studies include: xx and xx. These characteristics are also examined in this dataset, while some are not due to different study approaches.

Methods

Datasets

**Remember to alter how this header's size appears... later. Several datasets were constructed to examine the full ... (this sentence belongs in the intro, I think) The primary dataset of heme-containing proteins (HEM) was composed by finding approximately 30 (specify exact number) of types of proteins that are present in the PDB. This included heme oxygenase, myoglobin, cytochrome P450, among many others.

Datasets: HEM, HEC(?), SRM, VER/VEA.

Four datasets were constructed for this study.

The primary dataset of heme-containing proteins (HEM) was constructed by searching for 30 different classes of proteins; this enabled a dataset of diverse proteins to account for any structural deltas to achieve different chemistry. Additional samples of each class of protein were then added to the dataset, bringing the total to XX. The PDBs were restricted to the following criteria to ensure quality: XX. A full list of proteins and their source organism used in the study is available in table XX.

The heme-c dataset followed the same criteria (XX). Similar proteins were searched for as HEM, depending on availability in the PDB/possibility with chemistry. This dataset was anticipated to be fairly similar to HEM, and so only contains XX samples. The full table is available in table XX.

The siroheme dataset (SRM) contains fewer samples than HEM or HEC due to the limited structures available. A search for 'SRM' as of 26 July 2021 produced 52 structures. Not all of these structures contain siroheme. A full-text search for "siroheme" produces many more results, but very few are complex with siroheme. SAH appears commonly used to complex the relevant siroheme proteins, however examination of whether this guarantees acceptable results/estimates of SA/V etc. is outside the scope of this study. No quality criteria were employed, but all proteins are within: XX.

The verdoheme dataset (VER/VEA) is very limited. A search for "ver-

doheme" as of 26 July 2021 produces 12 results. From these results only 4 usable proteins are available. All PDBs fall within XX criteria.

Some PDBs in HEM-dataset contain PDBs where there is a double-molecule representation of heme. If this has not been taken care of ** FIX ME!!!* then write something here.

The scripts used in this study were modified depending whether HEM/HEC/SRM/VER/VEA were being processed, and depending on the distance from the ligand of interest being examined (i.e. 5-7A). This is discussed in further detail below MAYBE (FIXME!).

Confirming data quality and details/PDB detail table

All PDBs used in the study were scanned/text-parsed with a python script. This script grabbed a bunch of relevant qualities, like molecule purpose, source organism, resolution (XX FIX), and PDB code to confirm. The data produced are in table X.

Preprocessing/before chimera/monomers

Many of the PDBS downloaded are multimeric structures. These were all processed into monomeric structures, by selecting a single chain (chain A) and eliminating all others chains in each PDB. This makes examination easier and more representative when data are aggregated; multimers with more pockets would otherwise skew the data and be the majority represented in the results.

All scripts below were paused while running for visual examination; in rare cases the process of conversion to monomers resulted in errors processing, especially for volume measurements. This is discussed below if this issue was not corrected (FIXME!! XX).

Examination in Chimera/acquiring results

Multiple scripts were written and applied in Chimera. These scripts are divided up based on what results are produced.

In the scripts the choice of 5A or 7A as a distance from the ligand is arbitrarily chosen. This is a distance that generally accounts for all residues

able to interact with the ligand. The results are presented in both 5A and 7A sets to account for the variability introduced by these cutoffs.

Volume

Volume of the binding pockets for each ligand are calculated using surfnet. Atoms within 5-7A of the ligand are selected and the pocket volume they form with the ligand is calculated. The surfnet algorithm works by... making triangles along the molecular surface, I guess, until the distance cutoff.

Volume of the ligands is also calculated, using a different method; the above method with surfnet is not possible to employ for single molecules outside the pocket. First the ligands were isolated from their pdb. The molecular solvent/surface area was calculated (Accesible and excluded). The surface area and the volume of the resulting... blob, is given by Chimera. This is not the same method as employed by surfnet to acquire volume and represents a limitation in the study (FIXME! IDK IF THIS SHOULD GET DISCUSSED HERE)

Images of this operation are available in XX.

Surface Area

Surface area of the pockets is calculated using a similar method as above. The atoms within 5-7A of the ligand are selected. The 'surf' operation is applied. This creates a surface... IDK how this algorithm works (FIXME!). Both accessible and excluded solvent area are outputted.

Surface area of the ligands is calculated as noted above in Volume section.

Images of this operation are available in XX.

Distances

Distances (NOT THE ANGLEDIST stuff) were calculated by selecting all atoms within 5-7A of Fe in the ligands. Each atom's distance to the Fe was calculated by using the distance operation in Chimera (confirm this is precisely what we do), which simply draws a line between the atom and the Fe atom. (FIXME! IN DISCUSSION, DISCUSS WHY METAL COORDINATION IS GARBAGE)

The residue each atom is in is reported. Therefore, each

Angles Residues - Heme Plane

Angles Fe-CA-CB

Amino Acid Frequency in Pockets

title

Importing to R and stastical analysis

FIgures wer also made in R.

- Download from PDB using the script they've provided at RCSB for many, many files
- Use UCSF Chimera to determine:
  - Volume
  - SA
  - Nearby AA
- R to process raw data and produce tables
- Whatever other software we use to achieve the other results. E.g. E, or availability to solvent etc. likely will stick w Chimera I suspect. Or somehow implement the Python script to open both chimera for the first part of what we've done or for something else later. The script we've written is a python script, not a chimera script. We're initializing it with chimera and excluding the necessary code... to initialize chimera and specify chimera to receive the commands

Equipment

Results

Attempt to reference **??** figure: **??** References

Discussion

Limitations of this brilliant work: 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 Visual examination itself to OK the parameters/algorithms can introduce bias Precision of algorithms needs to be evaluated** or at least IDK how PRECISE they are

Conclusion

this is just as master's and in basic research don't feel the need to replicate what took some god forsaken, sad, overworked, impoverished PhD students years + with help of their PIs and with generous word fluff to hide fuck ups. 0 -¿ thesis in approx. 3-4 months during global catastrophe is nifty

Bibliography

[1] Chris E Cooper et al. "Engineering tyrosine residues into hemoglobin enhances heme reduction, decreases oxidative stress and increases vascular retention of a hemoglobin based blood substitute". In: *Free Radical Biology and Medicine* 134 (June 2019), pp. 106–118. ISSN: 18734596. DOI: 10.1016/j.freeradbiomed.2018.12.030.

[2] Ting Li, Herbert L Bonkovsky, and Jun Tao Guo. "Structural analysis of heme proteins: Implications for design and prediction". In: *BMC Structural Biology* 11 (2011). ISSN: 14726807. DOI: 10.1186/1472-6807-11-13.

Appendices

Figures

Figure A.1: HEM AA Frequency 7A

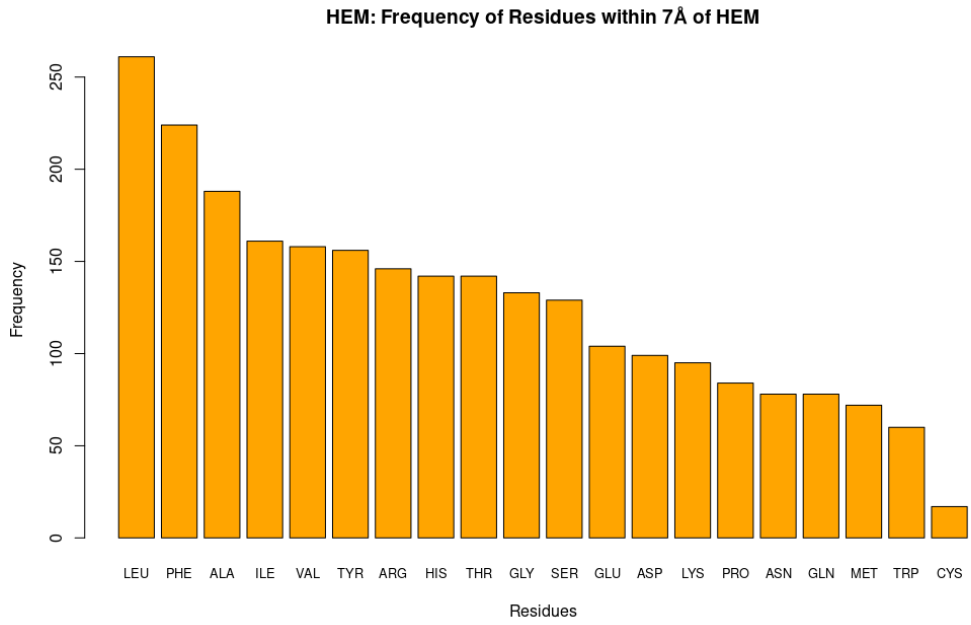


Figure A.2: HEC AA Frequency 7A

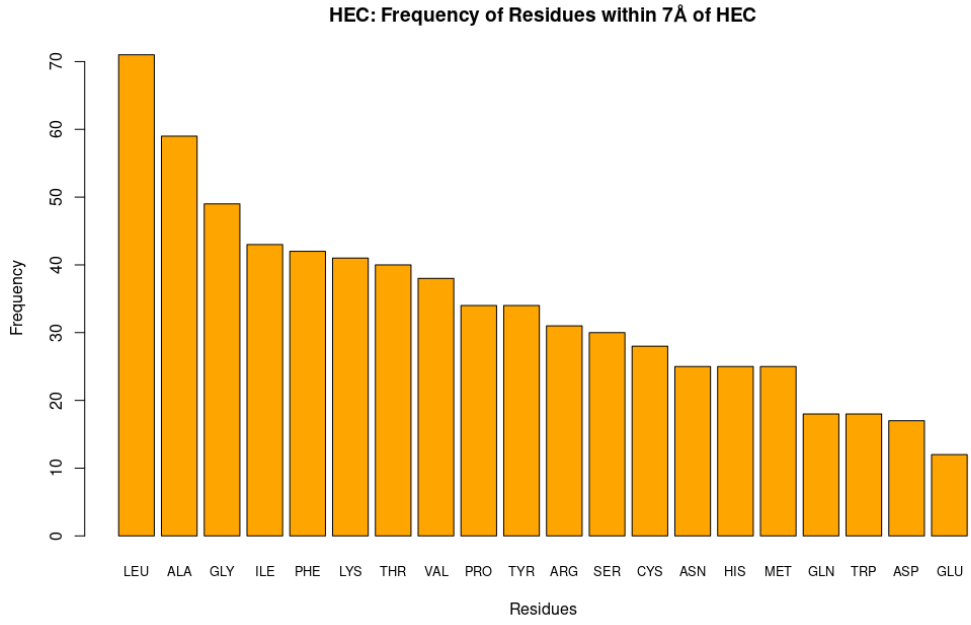


Figure A.3: SRM AA Frequency 7A

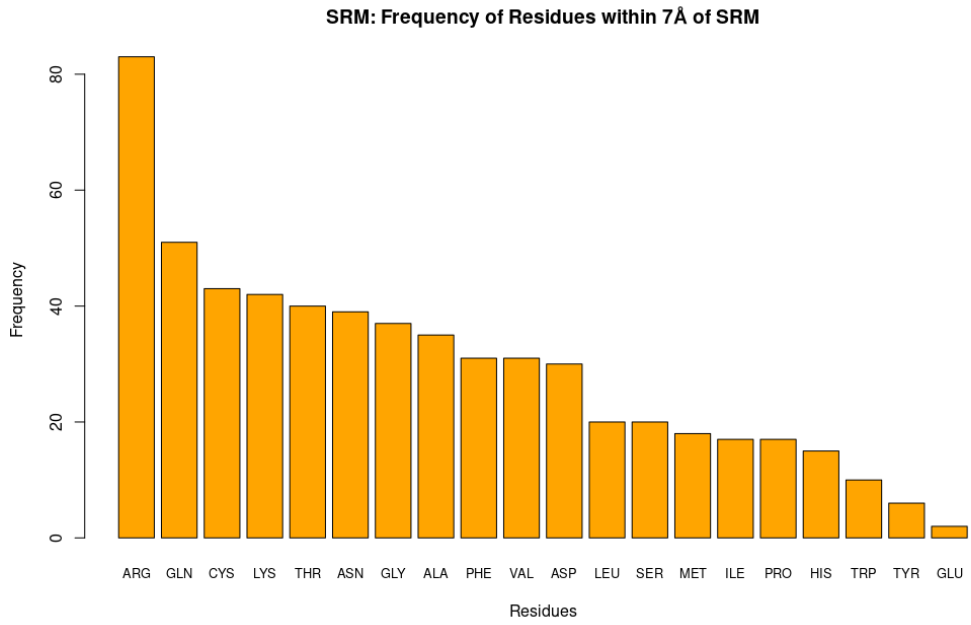


Figure A.4: VERDOHEME AA Frequency 7A

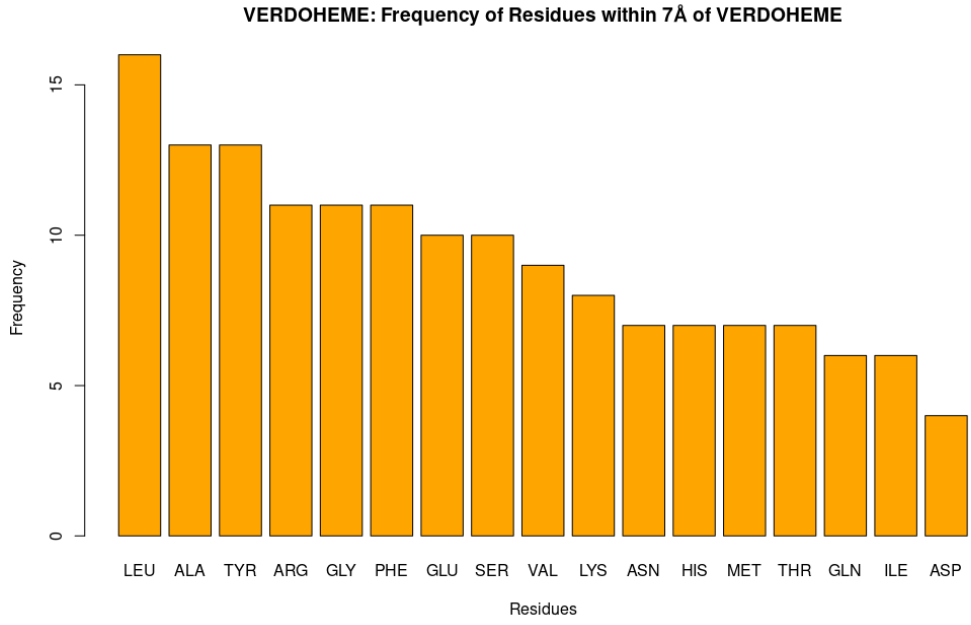


Figure A.5: HEM CACBFe Data

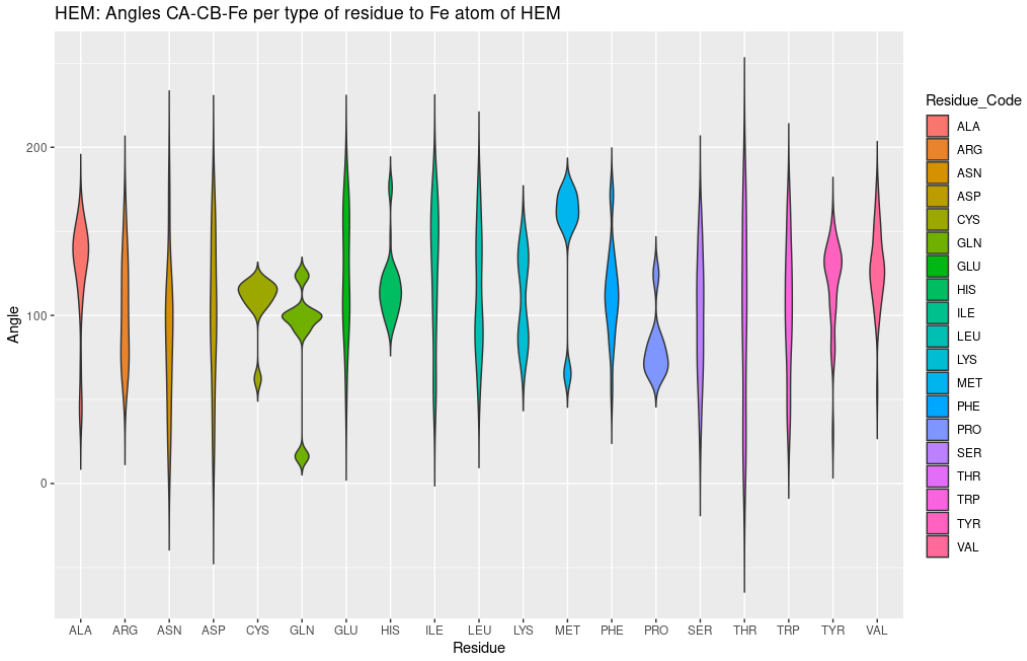


Figure A.6: HEC CACBFe Data

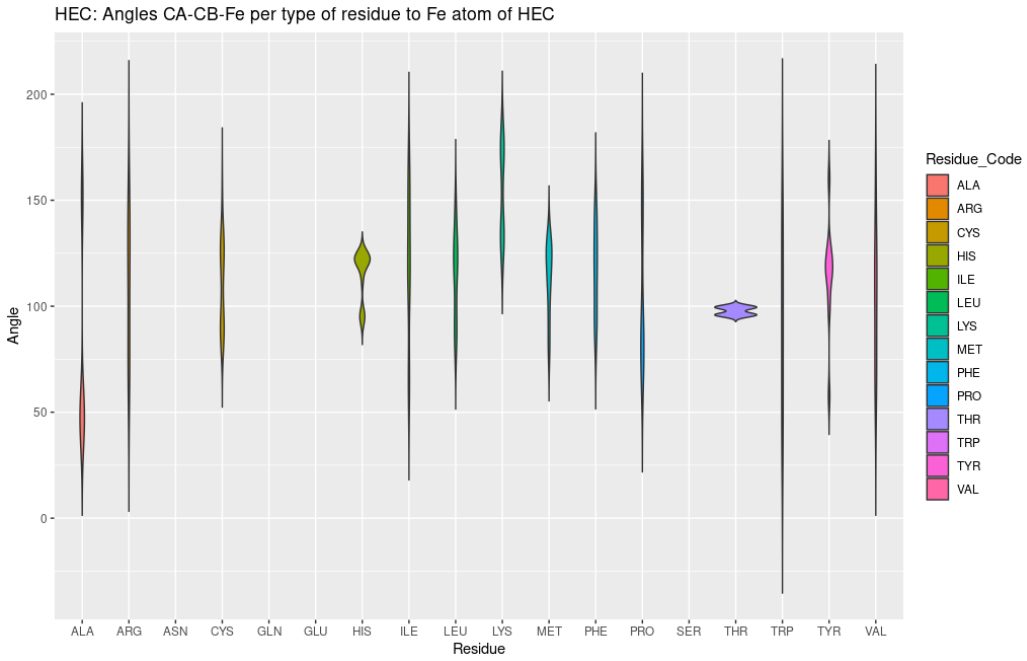


Figure A.7: SRM CACBFe Data

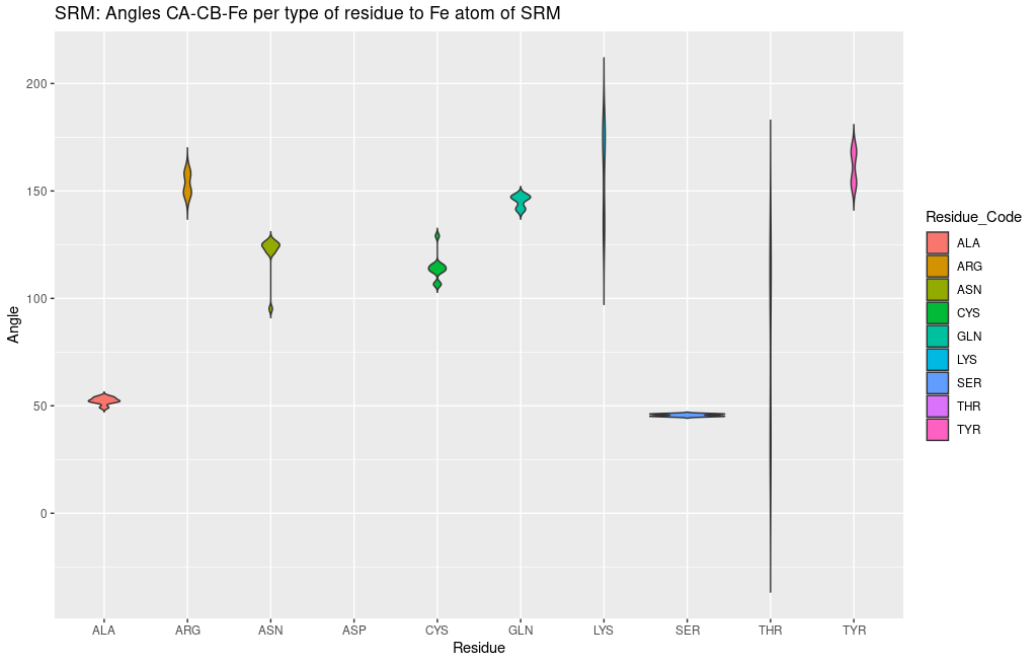


Figure A.8: VERDOHEME CACBFe Data

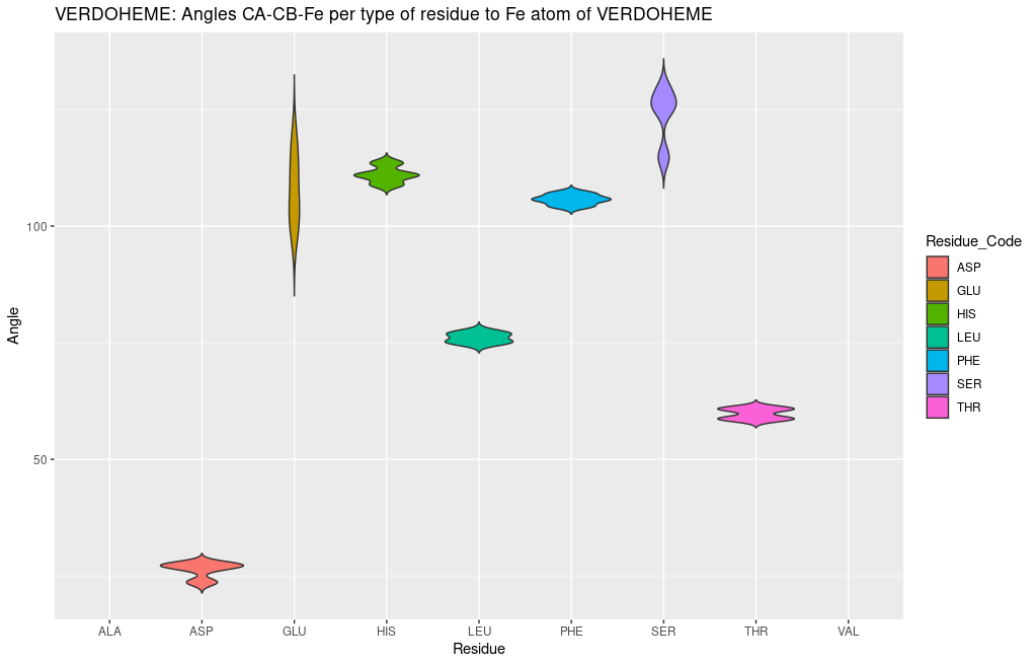


Figure A.9: HEM Closest Residue Data

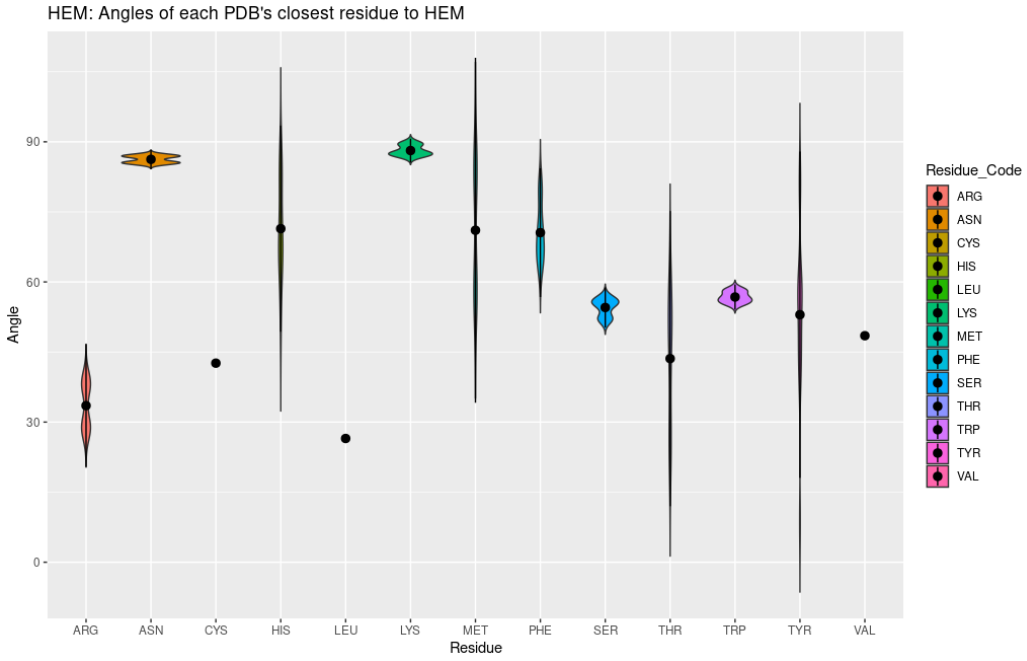


Figure A.10: HEC Closest Residue Data

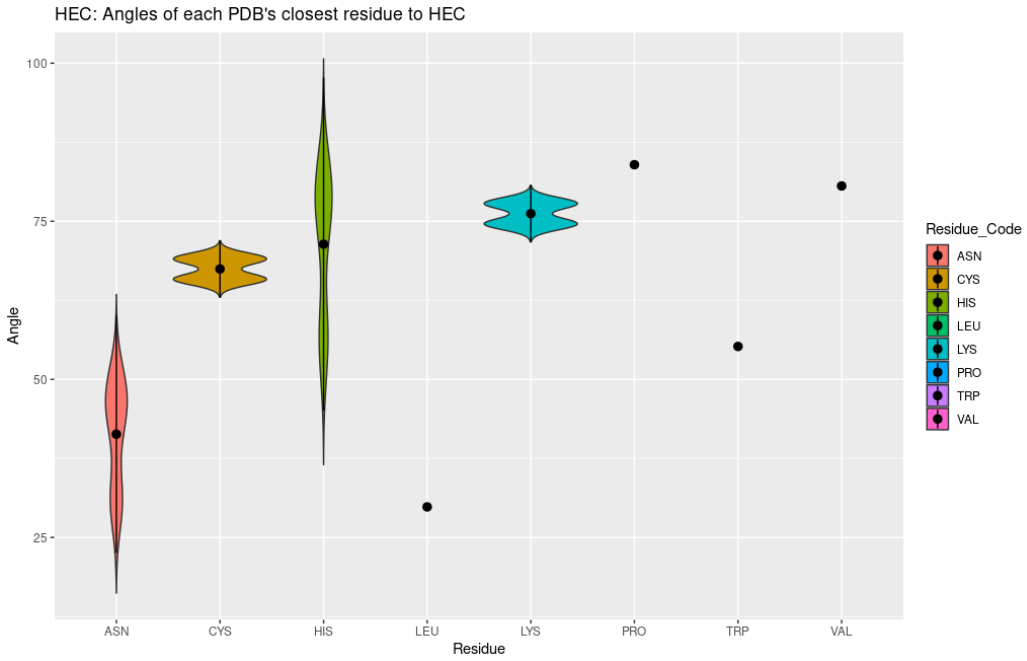


Figure A.11: SRM Closest Residue Data

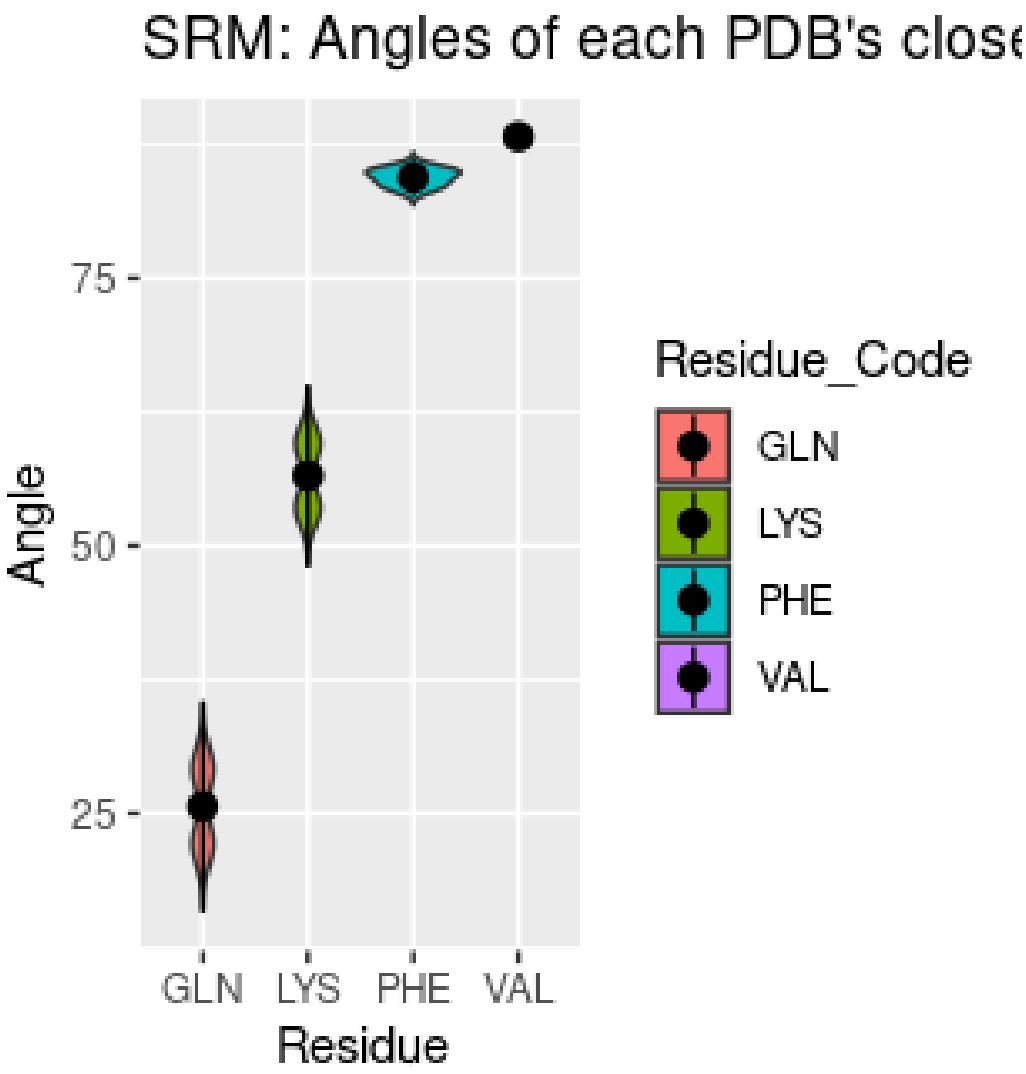


Figure A.12: VERDOHEME Closest Residue Data

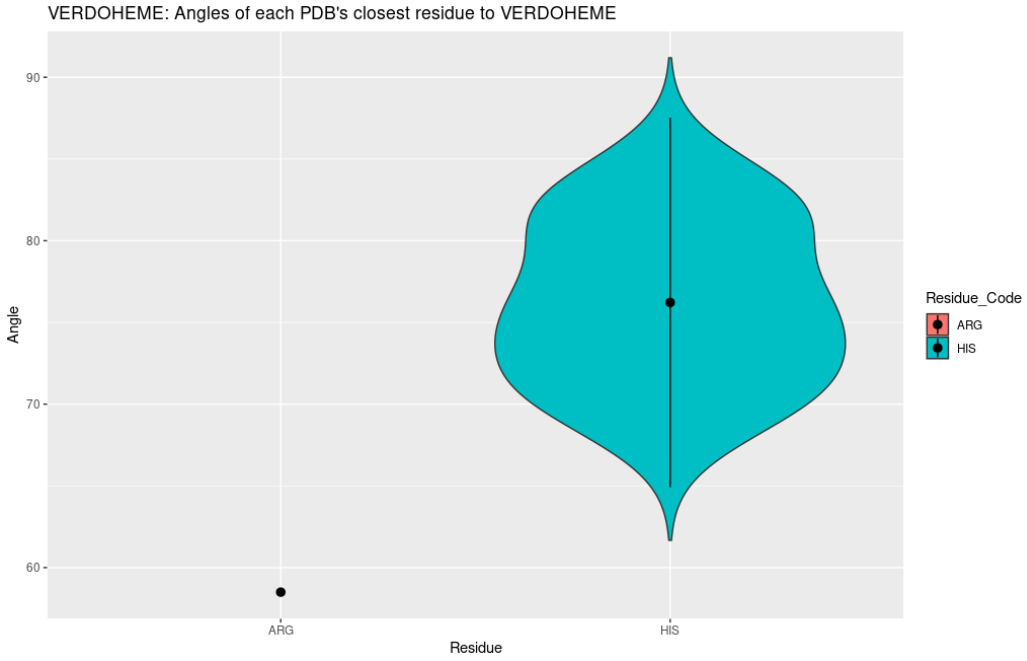


Figure A.13: HEM Coordinating Residue Data

Figure A.14: HEC Coordinating Residue Data

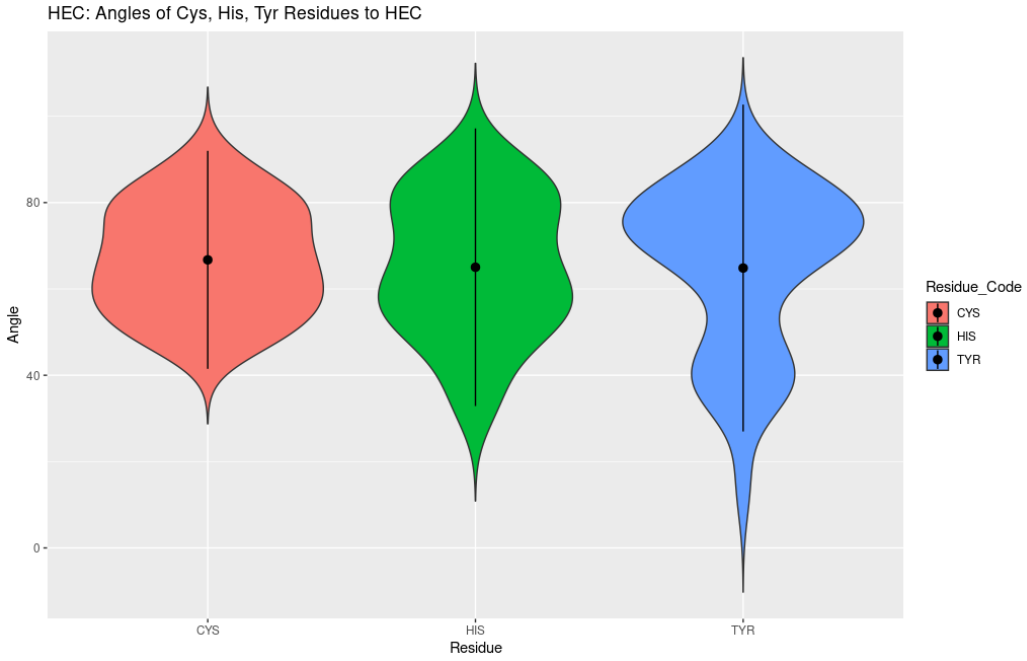


Figure A.15: SRM Coordinating Residue Data

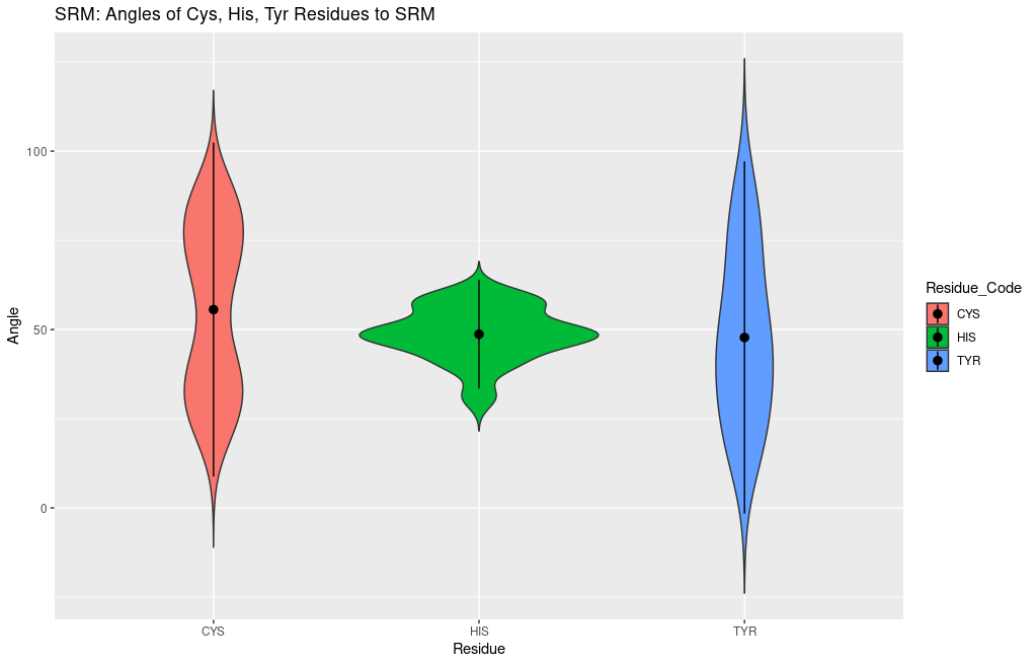


Figure A.16: VERDOHEME Coordinating Residue Data

Figure A.17: HEM Ligand Accessible Surface Area

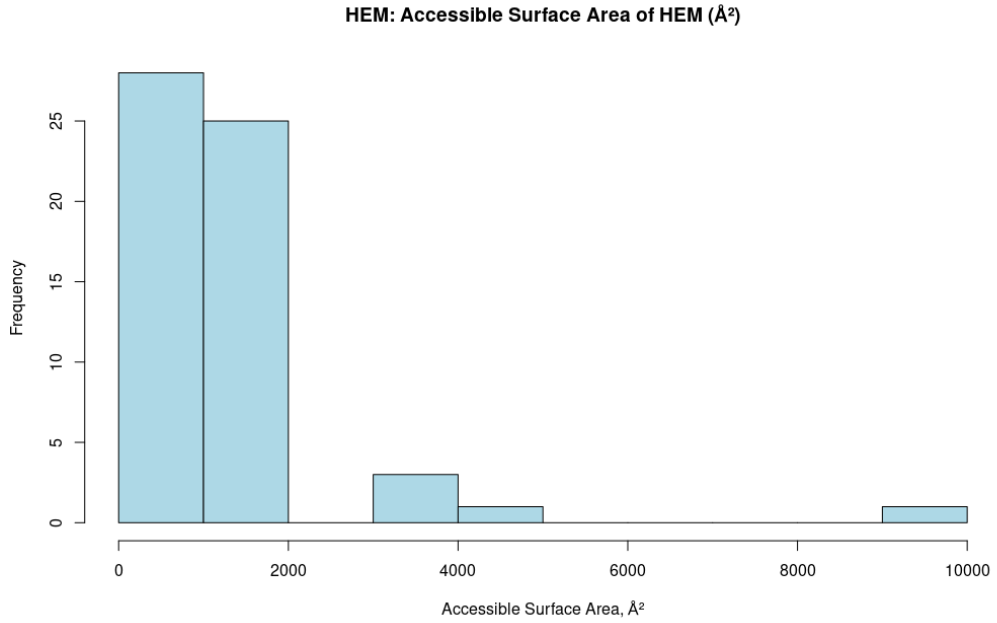


Figure A.18: HEC Ligand Accessible Surface Area

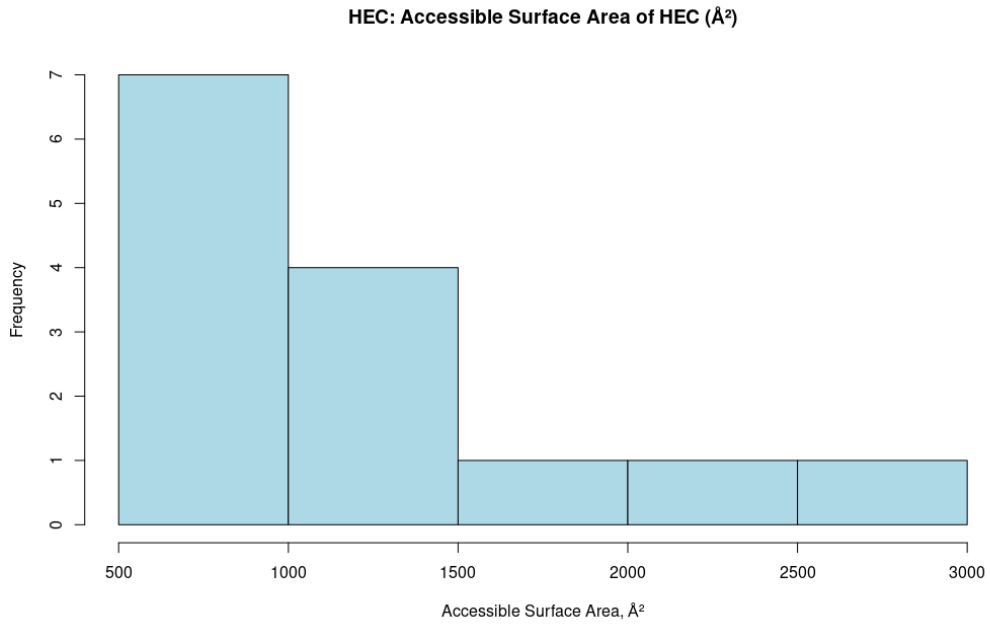


Figure A.19: SRM Ligand Accessible Surface Area

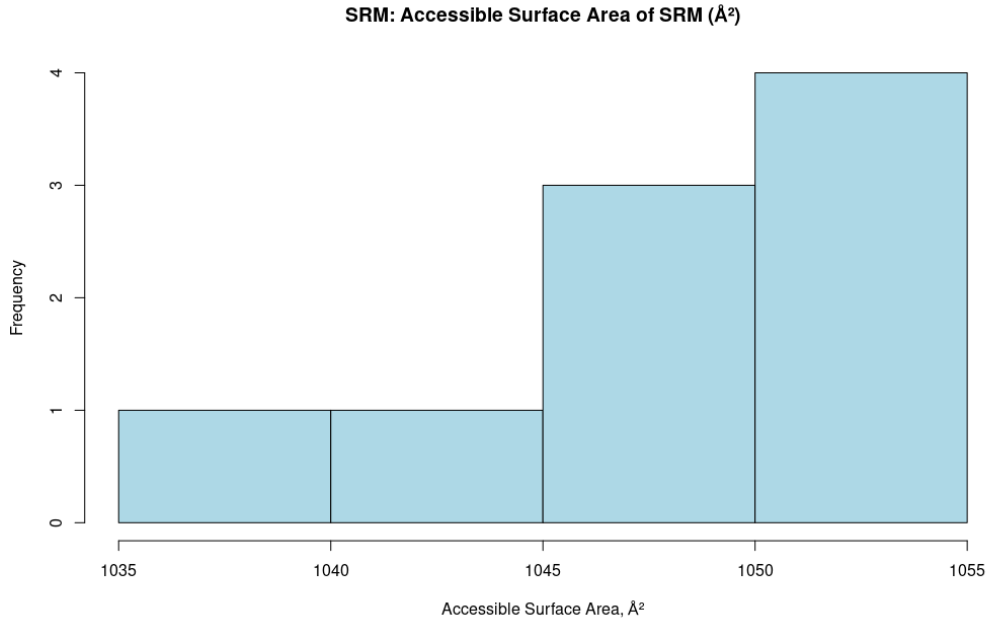


Figure A.20: VERDOHEME Ligand Accessible Surface Area

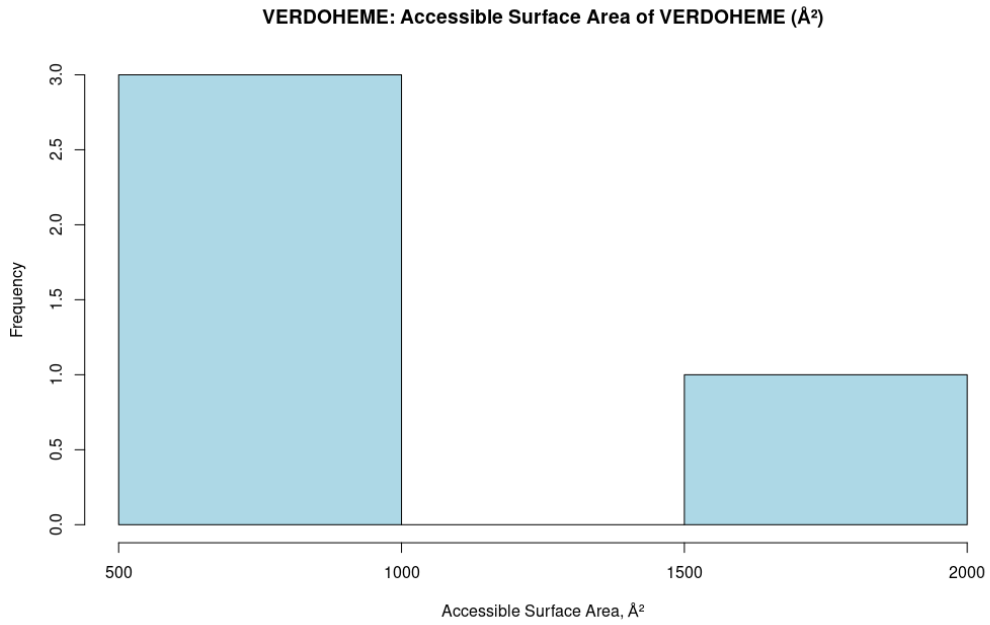


Figure A.21: HEM Ligand Excluded Surface Area

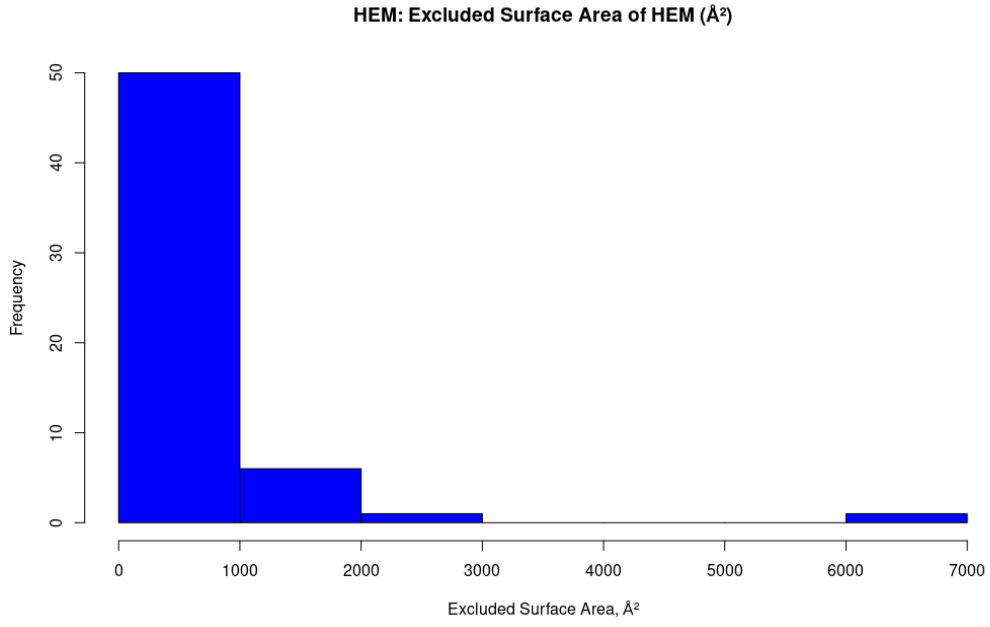


Figure A.22: HEC Ligand Excluded Surface Area

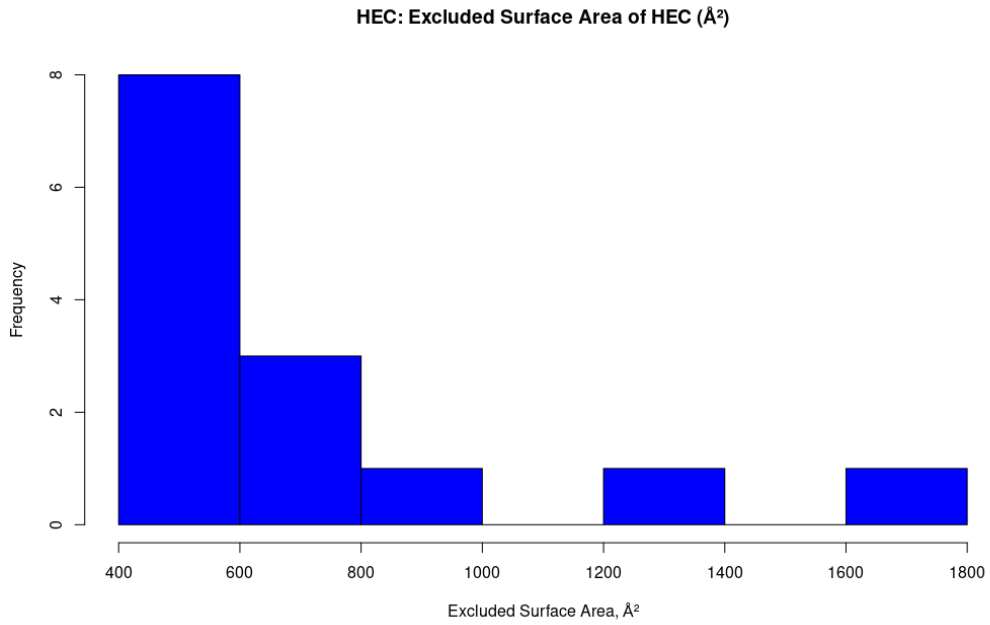


Figure A.23: SRM Ligand Excluded Surface Area

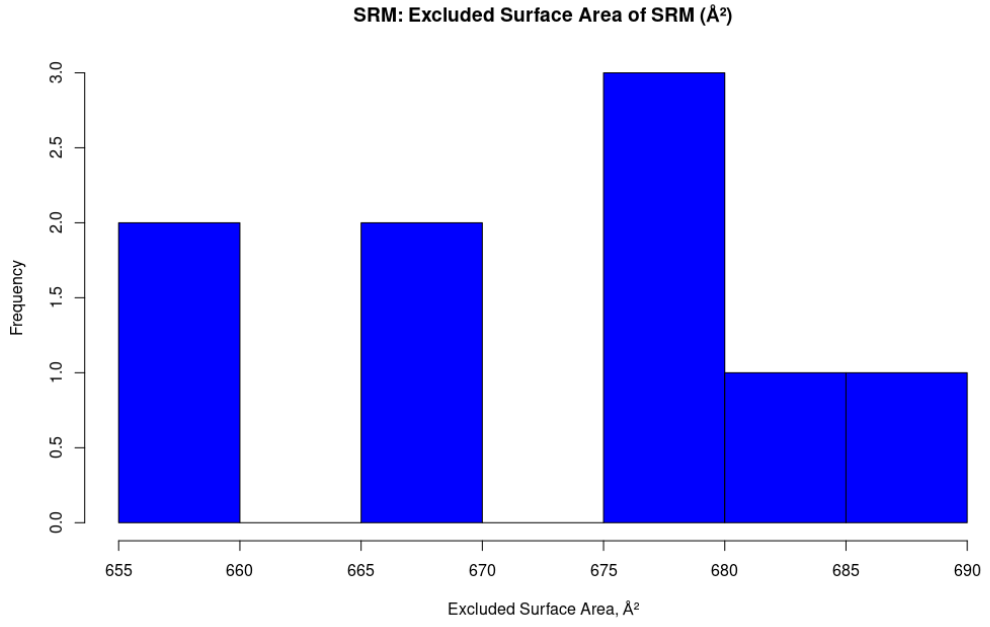


Figure A.24: VERDOHEME Ligand Excluded Surface Area

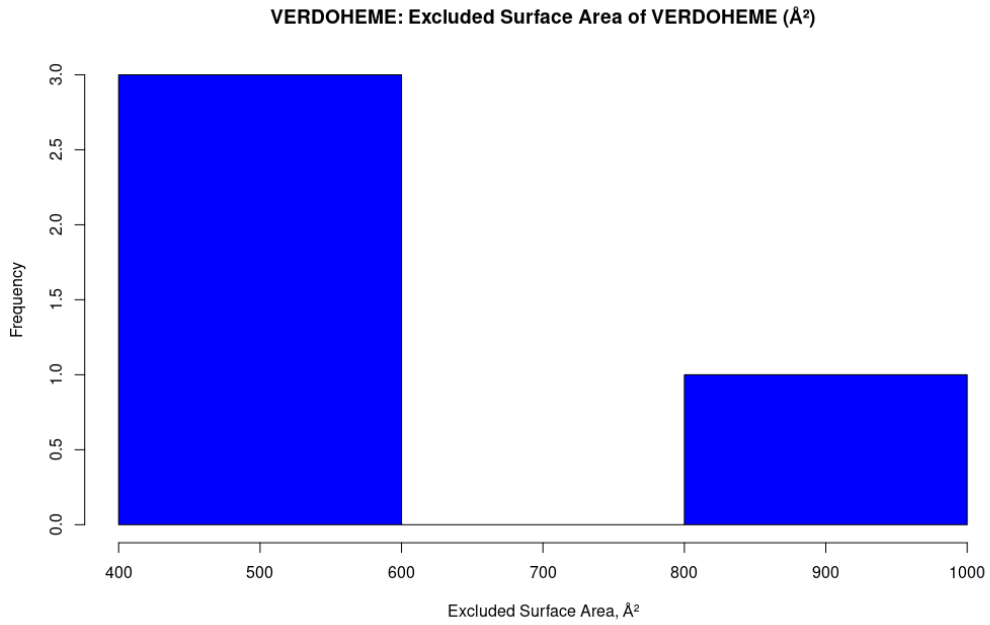


Figure A.25: HEM Planar Angles

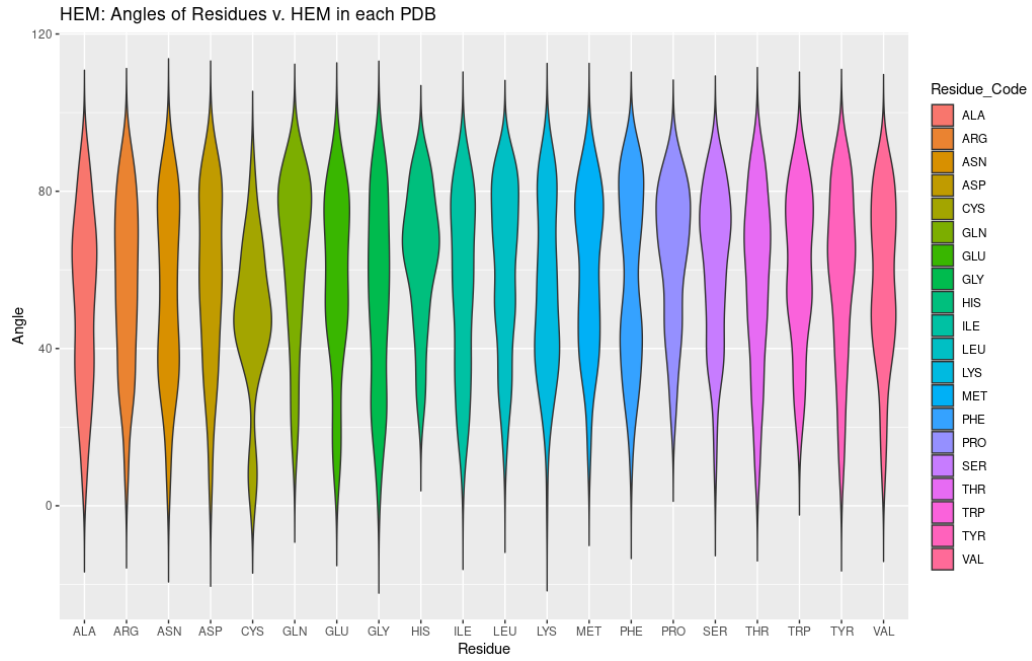


Figure A.26: HEC Planar Angles

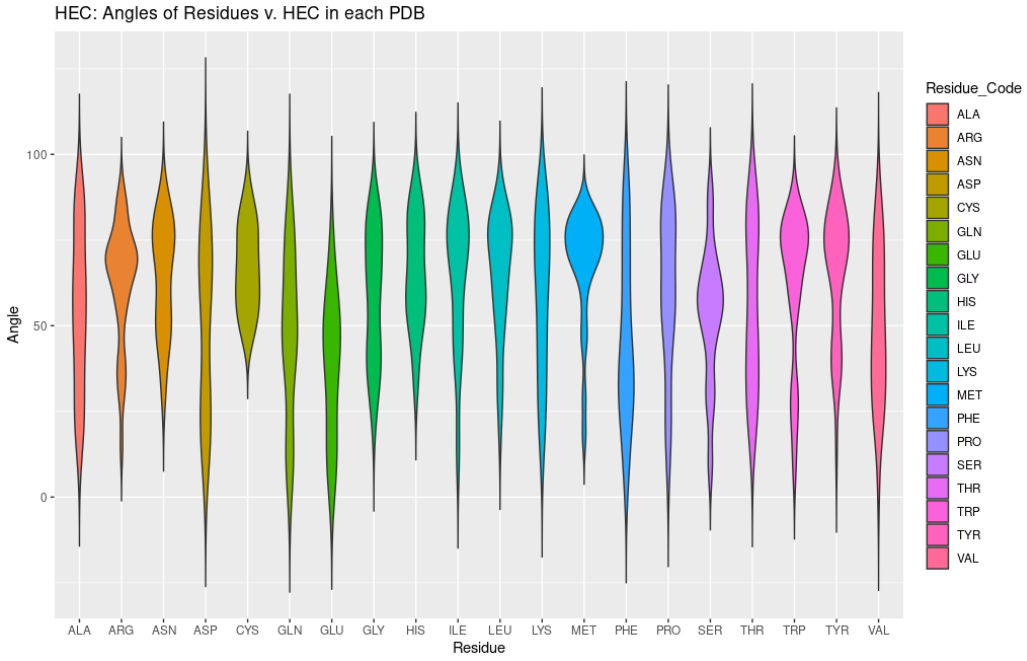


Figure A.27: SRM Planar Angles

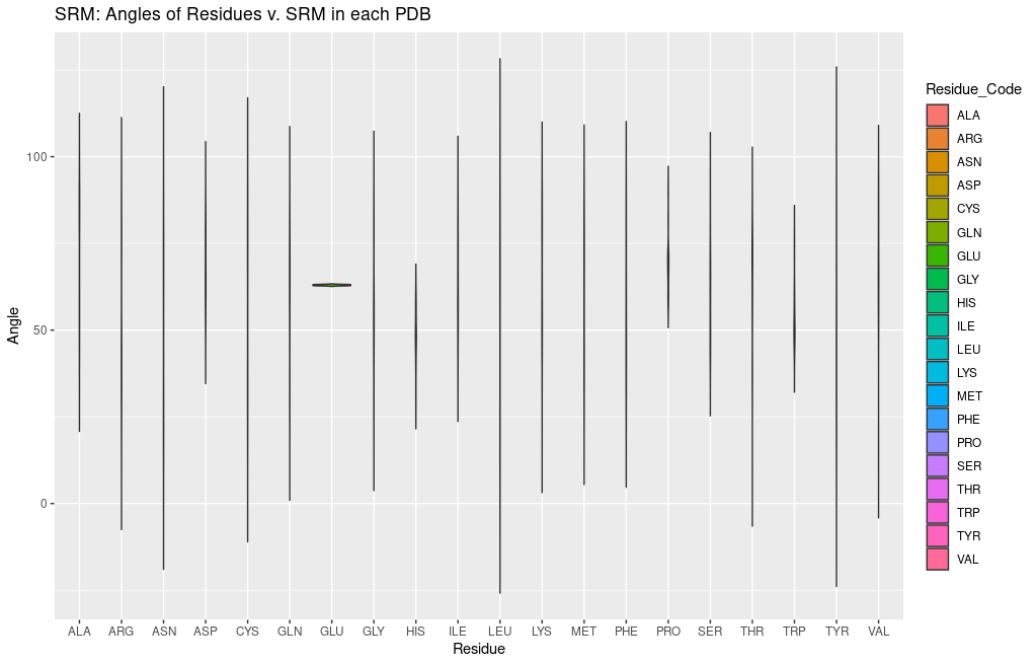


Figure A.28: VERDOHEME Planar Angles

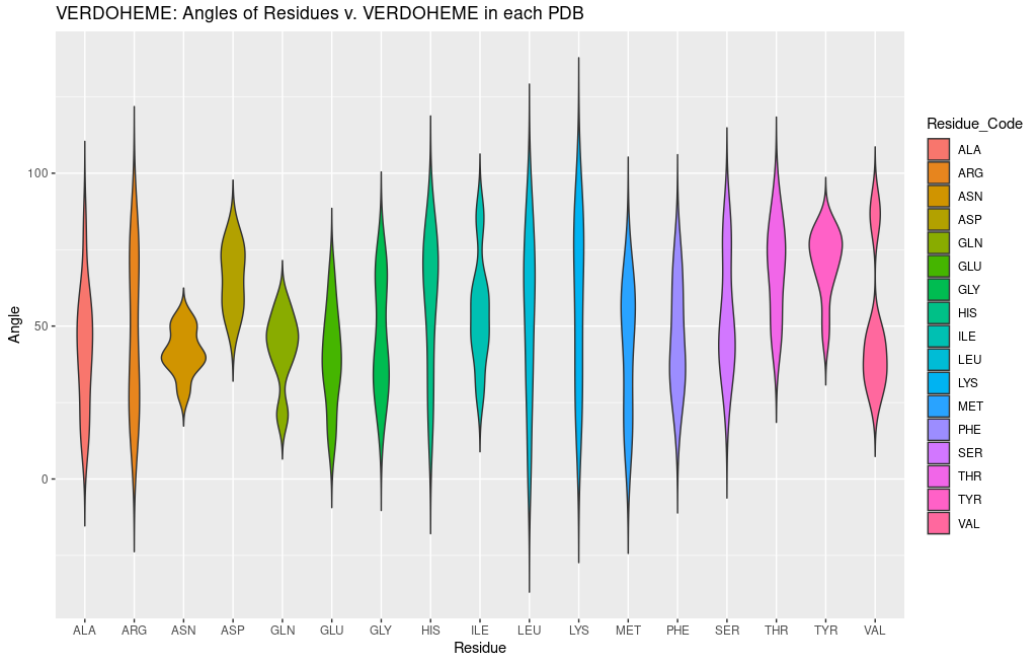
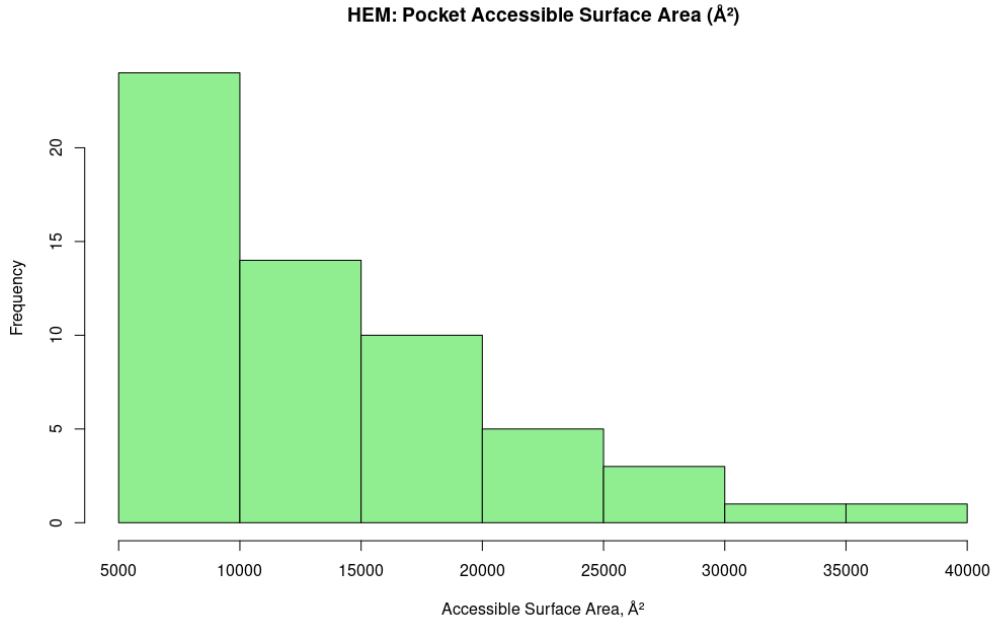


Figure A.29: HEM Pocket Accessible Surface Area



Tables

B.1 AA Frequency

B.2 CACBFe Data

B.3 Likely coordinating residues data

B.4 Minimum Distance Residues Data

B.5 Planar Angles Data

B.6 Other Data