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 **??**.

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.