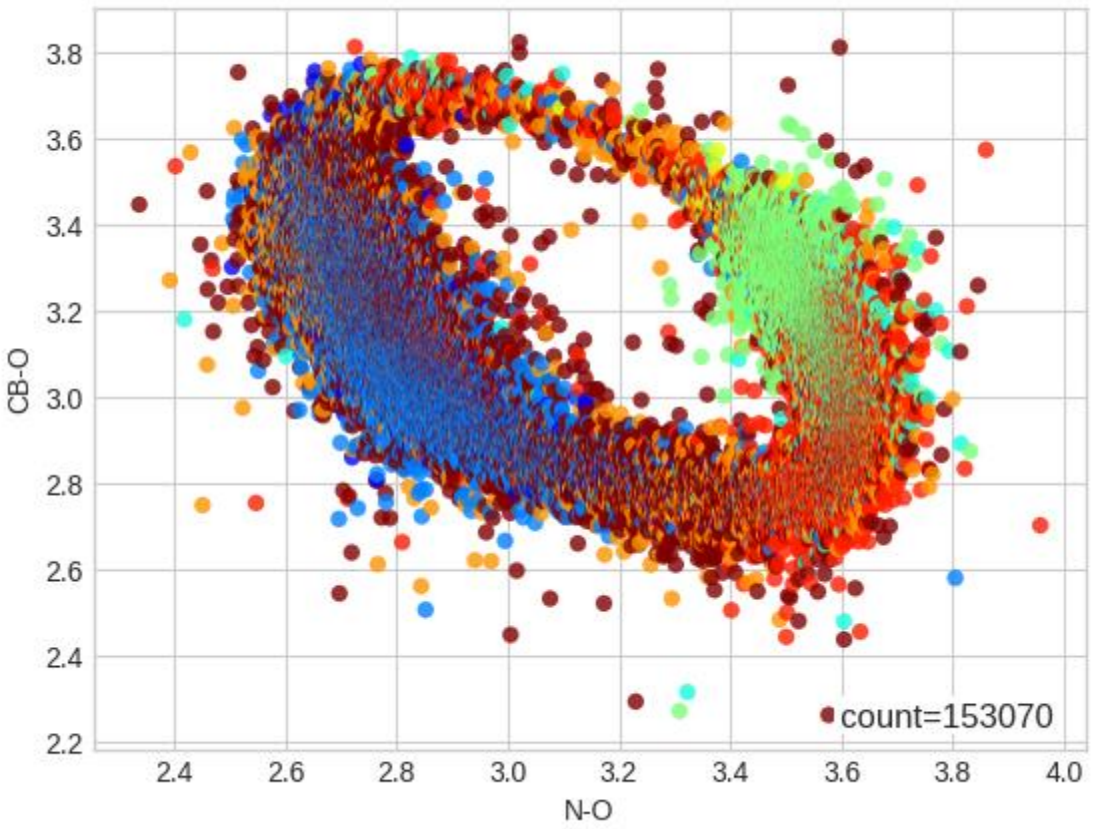
MSc Bioinformatics and Systems Biology

What new things can we learn from ultrahigh-resolution protein crystal structures?

by

Rachel Alcraft



Correlation of atom distances N-O and CB-O, <= 1.1 Å, coloured on secondary structure (R.Alcraft 2020)

Supervisor: Mark Williams

Date: 24th August 2020

Birkbeck College

Department of Biological Sciences

Birkbeck, University of London

Malet Street, London WC1E 7HX

United Kingdom

Abstract

Ultrahigh resolution structures from the Protein Data Bank (PDB, Berman et al, 2002) have been analysed using novel software to calculate 170 different geometric measures, including many not previously examined in large scale analysis. These results have been made accessible through a website on the Birkbeck College servers, with geometric analyses and correlations available on demand using python CGI scripts, NumPy and matplotlib.

The protein atomic positions have been analysed for recommended updates to stereochemical geometry that could be warranted on resolution, and any further insights that can be gained into protein structure based on the increased confidence in atom placement using the ultrahigh-resolution data set.

Interesting correlations have been found that define geometrically necessary regions, for example the intra residue one-four distances between nitrogen-oxygen and carbonβ-oxygen form a parametric relationship with the dihedral angle PSI underlying (see title picture). A correlation page of a selection of these geometric relationships can be found on the website (Correlations Page).

Exploration of the multimodal aspect of the distributions has yielded areas of correlation that suggest secondary structure features still to be explored, with the changing perspective from correlations viewed on secondary structure further hinting at categories for unidentified secondary structures.

Where geometry is irregular, there is the possibility of sites of genuine functional interest. The suggested novel correlations can pick out geometrically unusual sites which could enable those sites to be further analysed for validity and structural and functional interest.

A true confidence in the atom placement of a structure for analysis of functional sites requires knowledge of experimental evidence: exploration of electron density has yielded a novel method for comparison of normalised density matrices that could lead to an evaluation of structures against their electron density.

The combination of geometric feature analysis and electron density analysis has the future potential to enable fast detection of functional features of proteins; to use electron density directly to examine geometry; and to provide insight into the nature of atomic bonds.

Table of Contents

Abstract ii
Table of Contents iii
Figures and Tables iv
Glossary and Abbreviations vi
1. Introduction 8
1.1 Background 8
1.2 Aim and objectives 11
2. Materials and Methods 12
2.1 Geometric Data Generation 12
2.2 Geometric Data Reports 15
2.3 Electron Density 20
3. Results 23
3.1 Summary of structures and residues 23
3.2 Validation 24
3.3 Results for bond lengths and angles 27
3.4 Results for new insights 32
3.5 Results for electron density 60
4. Discussion and Conclusions 63
4.1 Bond lengths and angles 63
4.2 New insights into geometrical features 64
4.3 Electron density 68
4.4 Resolution 69
4.5 Overall research aim 70
4.6 Further work 71
4.7 Implications of the research 72
References 73
Appendices 77

Figures and Tables

Table 2.1.2.1 Selection of geometric measures calculated 13

Figure 2.2.1 Website header with menu options 15

Figure 2.2.1.1 Choices for the distributions page 15

Figure 2.2.2.1 Choices for the resolutions page 17

Figure 2.2.3.1 Choices for the Correlation page, generates sixteen correlations considered interesting 17

Figure 2.2.4.1 Results for the close contacts for histograms on Distributions page 18

Figure 2.2.4.2 Contact Maps, 4 contact views for 6mry 19

Figure 2.3.1.1 Trilinear interpolation, overall method, depicts the coordinates in the cube around the point to interpolate 20

Figure 2.3.1.2 Trilinear interpolation, the linear fraction calculated from the cosine rule 20

Table 3.1.1 The included structures broken down by resolution and dataset 23

Figure 3.2.1.2 Correlation plots for 1i1w showing the suspicious residue 180 (circled) 25

Figure 3.2.1.3 The geometric outlier Ser 180 in 1I1W, planar carbonyl not as expected, with electron density. 25

Table 3.2.1.4 Calculations to investigate possible labelling error of occupants A and B, 1I1W residues 179-181. 25

Figure 3.2.1.5 Correlations plotted for occupant combinations for 1i1w residues 180 and 181 26

Figure 3.2.1.6 Chimera image shows geometrically unlikely region in 1i1w and the active site 26

Table 3.3.1.1 Bond-length comparison for the highest resolution structures, compared with E&H and Jaskolski. 27

Table 3.3.1.2 Bond lengths on resolution, Jaskolski vs PSU-Beta HQ set and HIGH/2019 data 28

Table 3.3.1.3 Tau value comparisons using Jaskolski Table 3 (Jaskolski et al, 2007). 29

Figure 3.3.1.4 Violin plots show tau distributions for all but pro and gly against individual amino acids in the HQ set. 29

Figure 3.3.1.5 Bimodality in tau correlated with psi and different favoured tau regions for ser, leu, ile and thr. 30

Figure 3.3.2.1 Violin plots for PHI, PSI and OMEGA – PRO, CYS and GLY 31

Figure 3.3.2.2 Violin plots for N-O and CB-O for PRO, ILE and ASN. 31

Figure 3.3.2.3 Violin plots for CB-N and O1N-CB for PRO, ILE and ASN. 31

Figure 3.4.1.1 The Ramachandran plot from PSU-ALPHA, 3 views, with the secondary structure key. 32

Figure 3.4.1.2 "Correlations" page in PSU-View for <=1.2 Å, HQ set 33

Figure 3.4.1.3 Geometric correlations, CHI1 vs CHI2, graduated on resolution, HQ set, resolution <=1.2Å 33

Figure 3.4.1.4 Comparing "the square plot", the ellipse, PHI/C1n-CB and Ramachandran on secondary structures. 34

Figure 3.4.2.1 Distributions for PRO, HIS and MET graduated on resolution, rvalue, rfree and bfactor 35

Figure 3.4.2.2 Resolution shows contours of probability density in scatter plots 36

Figure 3.4.2.3 The rarity effect: compares distributions for resolution; bfactor; second letter of the name. 37

Figure 3.4.3.1 Geometry apparently influenced by refinement software 39

Figure 3.4.3.2 Refinement software and the influence on geometry 40

Figure 3.4.4.1 Scatter, density trace and probability density for PSI/N-O, resolution <= 1.25Å 41

Figure 3.4.5.1 Cis and trans peptide bonds, with OMEGA shown in cis formation 42

Figure 3.4.5.2 Cis and trans peptide bonds directly correlate with C-alpha distance. 43

Figure 3.4.5.3 Comparing residues identified as omega-cis and Cα-cis 43

Figure 3.4.6.1 Unique conformations of proline and glycine, hydrogens not shown 44

Figure 3.4.6.2 Cis/trans regions as CA-CA1C/CA1N-CA, orange is proline 44

Figure 3.4.6.3 Cis/trans correlations comparisons for proline 45

Figure 3.4.6.4 Shows proline cis/trans effect on other geometry, coloured on resolution. 45

In each plot, the cis region is on the left and the trans region on the right. 45

Figure 3.4.6.5 Five CHI dihedrals of the proline ring 46

Figure 3.4.6.6 Proline CHI correlations in all combinations shown as probability density 47

Figure 3.4.6.7: Proline ring conformations' PCA analysis on CHI1-5 48

Table 3.4.6.8 Four examples each of the two proline ring conformations, green is down- blue is up-pucker 48

Figure 3.4.6.9 Two proline ring conformations illustrated from 1dy5 (Chimera) 48

Figure 3.4.6.10 Some correlations showing the up/down pucker and cis/trans peptide bond of proline 49

Table 3.4.6.11 The ratios of proline's cis/trans peptide bond to up/down pucker states and conditional probabilities 50

Figure 3.4.7.1 Cα measures calculated by PSU-Beta and a simple Cα skeleton 51

Figure 3.4.7.2 Probability overlays scatter plot for c-alpha pseudo-ramachandran, for ALA, GLY and PRO 52

Figure 3.4.7.3 Violin plots for Cα angles along the chain 53

Figure 3.4.7.4 Angles along Cα and backbone for PRO graduated on CAP-CA as a proxy cis/trans. 53

Figure 3.4.7.5 shows the relationship between O1N-CA distance and peptide bond, with possible models 53

Figure 3.4.7.6 Angles along Cα and correlated against PSI for GLY graduated on secondary structure. 54

Figure 3.4.8.1 Chi1 compared for 3 amino acids at different resolutions and with 2 different kde settings 55

Figure 3.4.8.2 Artificial CHI1 for alanine uses atom HB1 56

Figure 3.4.8.3 Alanine's HB1 naming in structure 3X2M 56

Figure 3.4.9.1 Close contacts between N and O of atom pairs up to 6.1Å apart, counts given for each resolution 58

Table 3.4.9.2 Donors and Acceptors for amino acids at resolution <= 1.1Å, rvalue <= 0.16, rfree <= 0.3 59

Figure 3.4.9.3 Proline – 2 examples when it is in close contact with other residues 59

Figure 3.5.1 The electron density of 11 tyrosine rings from 1us0 superposed 60

Figure 3.5.2 Peptide bond in structure 1ejg. 61

Figure 3.5.3 Tyrosine difference density and difference superposition for 1us0 62

Figure 4.2.1.1 The secondary structure regions shown in different correlation plots 64

Figure 4.2.3.3 Histidine - trimodal CHI2 66

Figure 4.2.4.1 Comparing close contact and non-close contact residues in the Ramachandran plot at <=0.9Å 67

This shows residues in close contact have a distinct region in the Ramachandran plot. 67

Glossary and Abbreviations

Term	Definition
Angstrom (Å)	Metric unit equivalent to 10^{-10} m, atomic scale measurement
Bfactor	Or Debye-Waller factor/temperature factor/atomic displacement parameter – the degree to which electrons are spread out, measured in $Å^2$ it gives a measure of local atom placement uncertainty.
Chi angles	Measurement of angle of rotation over sidechain bonds through the planar dihedral angles, successive atoms chosen along the chain for CHI1-5.
Cis-peptide	The Cαs are on the same side of the peptide bond
Dihedral angle	The angle between two planes.
DSSP	A program to calculate secondary structure, used to refer to secondary structures as calculated and classified by this program (Joosten et al, 2015; Kabsch & Sander, 1983).
E&H	Engh and Huber created the original geometric parameters for protein refinement.
PDBe	The Protein Data Bank in Europe, structure data, structure factors and electron density ccp4 files are freely available (PDBe)
Fo and Fc	The electron density measures obtained from the experiment (Fo) and implied by the model (Fc)
Occupancy	The likelihood of an alternative atom placement, multiple occupant positions will add up to 1.
Omega	Angle of right-handed rotation around C-N bond, measured by the angle between the planes CA-C-N^{+1} and C-N^{+1}-CA^{+1}
Peptide bond	Chemical bond formed between the carboxyl group of one amino acid and the amino group of another. A water molecule is released in the reaction and the linked amino acids are known as residues.
PDB	Protein Data Bank based in the US, structure data, structure factors and electron density maps (dsn6 format) are freely available (PDB)
Phi	Angle of right-handed rotation around N-CA bond, measured by the angle between the planes C^{-1}-N-CA and N-CA-C
Psi	Angle of right-handed rotation around CA-C bond, measured by the angle between the planes N-CA-C and CA-C-N^{+1}
Resolution	Smallest distance between crystal lattice planes resolved in the diffraction pattern
Rfree	Calculated like Rvalue below, but using a small subset of experimental data that has been withheld from refinement to act as an independent quality check.
Rvalue	$R = 100 \cdot \Sigma \mid \lvert Fo \rvert - \lvert Fc \rvert \mid / \Sigma \lvert Fo \rvert$ A measure of the difference between the structure factors calculated from the model and those from the experimental data.
Sp2 hybridised	Electron shells form three planar bonds at 120º
Sp3 hybridised	Electron shells form four tetrahedral bonds at 109.5º
Trans-peptide	The Cαs are on opposite sides of the peptide bond

Acknowledgements

I am grateful for: the inspiring teaching on the course from Adrian Shepherd, Maya Topf, Irilenia Nobeli and Andrew Martin; co-students Laura Phillips and Justin Barton's motivation and support; Dave Houldershaw's help with database problems even on bank holidays; but mostly to my supervisor Mark Williams for his time and for giving me confidence to think and explore ideas.

1. Introduction
1.1 Background

This project aims to analyse the extent to which ultrahigh-resolution structures can provide us with updated and new information on the geometry, structure and function of proteins.

In 1915, William Henry Bragg and his son William Lawrence Bragg became the first pair to be awarded a Nobel prize for the analysis of crystal structures through X-rays, theorising diffraction through atomic planes from the spherical/elliptical shape of Von Laue spots (Perutz, 1990). Many discoveries and Nobel prizes have followed, including: Pauling (in 1954 for the nature of the chemical bond and the alpha-helix); Kendrew and Perutz (in 1962 for globular proteins); Crick, Watson and Wilkins (in 1962 for the structure of DNA, with insight based on Rosalind Franklin's X-ray images); Dorothy Hodgkin (in 1964 for the structure of Vitamin B12), Anfinsen (in 1972 for folding of protein chains) (International Union of Crystallography, IUCR). In a review of X-ray crystallography in 1957, Crick and Kendrew refer to protein structure as "the geometrical aspects – the arrangements of atoms in space" (Crick and Kendrew, 1957) in contrast to the common meaning of sequence and polypeptide connections. They claimed X-ray crystallography alone of the techniques at that time could elucidate this atomic geometry – still needing the sequence, about which they say "It is likely to be a very long time before X-ray analysis can obtain by itself the amino acid sequence of a protein." (Crick and Kendrew, 1957). This has come to pass, the efforts to solve an X-ray structure without sequence are increasing with ultrahigh-resolution solutions, but an estimated 80% of structures are solved with molecular replacement involving knowledge of sequence and fragment structure (McCoy et al, 2017). Since Crick's time, the explosion in sequencing technology means the need for the amino acid sequence for a structure does not add significant difficulty - a direct solution still appeals.

Solving a structure requires iteration between experimental data and proposed structures until a satisfactory agreement is achieved. The amplitude of the diffracted x-rays correspond to the darkness of the diffraction spots, but the experimental data does not include the phase of the x-ray waves, so there is no direct calculable solution - as Crick and Kendrew say (1957), it cannot be solved by a "mathematical sausage machine". To ascertain the electron density the structure factors are transformed with a selection of phases derived from an initial model of the structure, followed by efforts to imply the structure, until a good agreement between calculated and observed diffraction spots is achieved and thus a satisfactory set of phases is implied.

The stereochemical restrictions on a structure are very tight, with bonds, angles, dihedrals, hydrogen bonds and van der Waal interactions all allowed within specific ranges, the rules for which are derived from small molecules in the Cambridge Structural Database (Groom et al, 2016) and the established Engh and Huber values (Engh & Huber, 1991, 2001). These values are used to balance the experimental evidence with the energetically possible and favoured geometry to refine the structure. Where

there is missing evidence, the knowledge of the sequence can help fill it in, relying on the stereochemical parameters. In low resolution structures where there is no experimental evidence for hydrogens (as the electron density of hydrogen at 1 electron is very small), all the hydrogens will be assumed to be located at fixed positions away from the atoms to which they are bonded (if they are included at all). Thus, in such a case all the hydrogens will appear to be geometrically uniform. It can be unclear: how a structure's atoms are placed; how valid they are; and how much experimental evidence there is.

Measures for the quality of the whole structure can be given in the form of root-mean-square deviation from expected stereochemical values (Wlodawer, 2007), with local variability being given by bfactors for each atom which measure the mobility of the atom. A high bfactor means less evidence for the correct placement of the atom. This mathematical puzzle is hugely complex, with refinement software making decisions on which stereochemical restraints to apply. This has the effect that experimental evidence of bonds and atoms is superposed by beliefs we already hold.

One such belief is that atoms are spherical - that there is a uniform electron distribution around a nucleus with charge density dependent only on distance from the nucleus. This simple model is necessary at low resolution – when there is low experimental evidence to support contrary decisions. A non-spherical multipole method can be used at high resolution when the electron density can support an anisotropic model. This is essential for hydrogen - the nature of the single hydrogen electron and the strong covalent bond means that hydrogen's electron density peak is far from the nucleus and a spherical method cannot correctly place the hydrogen atom. With ultrahigh-resolution structures at sub-atomic resolution of <=1Å, hydrogen positions can be determined, adding to the understanding of protein function through elucidation of protonation states and hydrogen bonding. For structures deposited in the Protein Databank (PDB, Berman et al, 2002) at <=0.7Å the electron density can directly provide information on the bonding of catalytic sites (Blakeley, 2015).

Diisopropyl-fluorophosphatase was solved to 0.85Å with hydrogens (Elias et al, 2013, pdb code 3o4p) and without (Koepke et al, 2003, pdb code 1pjx). The structure with hydrogens has helped to determine the protonation state around the active site through the position of hydrogen atoms in water molecules in the vicinity, leading to the possible identification of a catalytic site. A cholesterol oxidase protein solved to 0.74Å (Zarychta et al, 2015, pdb code 4rek) has the structural feature of a tunnel to reach the active site, the single residue gate keeper visible only at high resolution. A human aldose reductase-inhibitor complex solved to 0.66Å (Howard et al, 2004, pdb code 1us0) shows evidence of a departure from the spherical atom model and deviation from stereochemical expectations in active sites. They suggest these geometric subtleties can only be treated with confidence at high resolution when refinement parameters are relaxed and are essential in drug design. The crambin structure, solved to 0.54Å (Jelsch et al, 2000, pdb code 1ejg), was refined with three methods: spherical; non-spherical models and charge-density refinement leading to a suggestion for the development of methods to understand redox potential of metalloproteins in combination with quantum mechanical

calculations. An iron-sulfur protein was solved to 0.48Å (Hirano et al, 2016, pdb code 5d8v) with evidence for non-planar peptide bonds around active site cysteine residues bound to iron, with non-spherical atomic density evidenced in the same region.

These examples point to the increasing identification of protein function through structure from the solving of ultrahigh-resolution structures, and the ability to analyse atomic models through the ultra-high-resolution data. Much data and many tools exist to facilitate exploration of protein structural data at all resolutions. The Protein Data Bank in Europe (PDBe, Velankar et al, 2009) contains the structure's atomic coordinate files, along with their structure factors where available, and the electron density in ccp4 format as density and density difference. Janet Thornton's group looked at stereochemical quality of protein coordinates in 1992 (Morris et al, 1992) and noted that in higher resolutions, structures have a higher incidence of cis-peptides: thought due to greater confidence from electron density quality. A geometric tool is freely available at Duke University, MolProbity (Williams et al, 2018), to examine structures on a variety of features such as Ramachandran plot, cis-peptides, and Cβ deviations, as well as the facility to change a model to remove outliers - useful as part of the refinement process. This builds on ProCheck (Laskowski et al, 1993) which includes the validation plots for CHI1/CHI2, the Ramachandran plot as well as deviations for omega and c-alpha chirality.

Many opensource libraries are available to explore this data include the python library BioPython (Cock et al, 2009; Hamelryck et al, 2003) and the database and web application, the Protein Geometry Database, created in 2009 (Berkholz et al, 2009) to evaluate protein structure on backbone geometry and conformations.

The explosion of solved x-ray crystal structures at atomic level lends itself to the possibility of some update of current tools and knowledge. With the added confidence in atom placement, can there be a revision of the stereochemical restraints? Can we use this confidence to find structural features or new refinement parameters? In 2007, Jaskolski (Jaskolski et al, 2007) reviewed the Engh and Huber restraints using the 10 ultrahigh-resolution structures deposited at the time, with some updates suggested. The increase in ultrahigh-resolution structures since then means that a further analysis of these recommendations is warranted.

Ultimately, we may need to go back to the electron density for any anomalies or uncertainty, or to seek improvements in understanding geometric features of proteins - the electron density is the final arbiter (Wlodawer, 2007). In that case we might ask: what we can learn directly from the comparison of electron density of ultrahigh-resolution structures? This issue is complicated by the absence of standardisation of units used in the electron density matrices - they cannot be easily compared to each other. There are numerous attempts to solve this, a recent effort converts the arbitrary electron density to numbers of electrons (Yao et al, 2019) with an accompanying python library.

With the ultrahigh-resolution confidence in atom placement; the explosion of atomic detail solved structures; the availability of electron density: what new structural, functional or geometric features can we discover?

1.2 Aim and objectives

As stated, the aim of this project is to analyse the extent to which higher resolution structures can provide us with updated and new information on the geometry, structure and function of proteins, and to provide insights into and recommendations concerning those features. This can be broken into four objectives:

Objectives:

1.2.1 Bond lengths and angles

The Engh and Huber (1991) restraints were re-analysed in 2007 (Jaskolski et al, 2007) using the 10 highest resolution structures at the time. With the explosion in high-resolution structures, can these be again reviewed, this time with hundreds of structures?

1.2.2 New insights into geometric features and correlations

Given data for hundreds of ultrahigh-resolution structures, can any new insights be found? Geometric measures will be analysed statistically: investigating correlation; linear regression; PCA analysis; normality and modality. Do these insights provide any new recommendations and insights for structure validation or analysis?

1.2.3 Electron density analysis

Many structures have electron density and structure factors deposited, which provides an opportunity for direct analysis of the experimental evidence for structural features and provide further insight. In particular, the superposition of electron density of structural features over multiple observations will be attempted to see if atomic detail is enhanced by this method (Jelsch et al, 2000).

1.2.4 Resolution and geometry

Reviewing the geometry, new insights and electron density above: what can we learn about the importance of resolution for protein structural analysis?

2. Materials and Methods

A web-viewer was developed (PSU-Beta WebViewer) to explore geometric data based on the protocol developed in the MSc Biocomputing 2 module. The web-viewer allows browsing of geometric parameters including histograms, scatter plots and probability density plots, using a database developed to store all the geometric data. For the database, a non-homologous data set of ultrahigh-resolution structures was obtained, along with a comparator set of lower resolution structures taken from 2019. 170 geometric features were defined, including bonded and non-bonded lengths, angles and dihedrals.

2.1 Geometric Data Generation

2.1.1 Generation of geometric measures

A C++ program named PSU-Beta (Protein Structure Utility, version b) was created, consisting of shared libraries and executables. The executables perform 4 steps of the process, after a list of structures was generated from the Protein Data Bank (PDB, Berman et al, 2002) using the advanced search facility to search for: structures <=1.3Å (high set); structures deposited in 2019 (2019 set).

- Remove similarity - removes structures above 90% homologous, keeps highest resolution.
- Annotate structure - annotates the structure with e.g. resolution, rvalue, number of residues. Some structures are rejected at this stage: 30 or fewer residues; any structure with nucleotides.
- Create geometry - calculates the geometry given the measures specified in 2.1.2. Chosen data is extracted from the pdb files during this process pertaining to the atoms and residues, i.e. coordinates, bfactors, occupancy, chain name etc, the data extracted can be seen in the database tables, see 2.1.3. At this stage decisions are made about structural features that impact the geometry: negative amino acids are not included; where there identical chains but no NCS model only one chain is kept; where there are multiple occupants multiple models are built with the occupancy recorded.
- Contact map – for each structure, all n-residues are cross references against each other in the structure (n^2 calculations). Adjacent residues are excluded from the considered residue pairs. The distance is calculated between the specified atom pairs of interest: SG-SG, CB-CB, CA-CA and N-O. If the distance is < 6.1Å it is saved to the database.

The library and executables can be found on the project GitHub (PSU-Beta C++).

2.1.2 Geometric Measures calculated

The following geometric measures have been calculated. Note the following conventions:

- Distance e.g. CA-C, there are 2 atoms, it may not be a physical bond.
- Angle e.g. CA-C-O, or an alias e.g. TAU - there are 3 atoms, it may not be between bonded atoms.
- Dihedral angle e.g. N-CA-C-O or an alias e.g. PSI - 4 atoms, dihedral, improper or non-bonded.

The naming conventions from the pdb structures have been used, but note also, the convention of 1N for 1 backwards in the N-prime direction, and 2C meaning 2 forwards in the c-prime direction. So, CA1N is the previous residue's Cα and CA1C is the next residue's Cα.

Below in Table 2.1.2.1 is a selection of geometric measures calculated, (total list can be found in Appendix 4). Those with the type 1-4 are atoms with 3 bonds between them (they measure the rotation around the middle bond).

Alias	Type	Description
C-O, CA-C, C-N1C, C1N-N, N-CA	Bond length	Main chain distance
OMEGA: CA-C-N1C-CA1C PHI: C1N-N-CA-C PSI: N-CA-C-N1C	Dihedral	Main chain dihedral
TAU: N-CA-C TAU1N: C1N-N-CA TAU1C: CA-C-N1C	Angle	Main chain angle
CA-CB-CG CB-CA-C N-CA-CB	Angle	Side chain and main chain angle
C-C1C, CA-CA1C, CA1N-CA, C1N-C, N-N1C, N1N-N	Distance	Inter residue distance
CA-CA1C-CA2C CA1N-CA-CA1C CA2N-CA1N-CA	Angle	Inter residue angle
CA2N-CA1N-CA-CA1C	Dihedral	Inter residue dihedral
CB-O, N-O	One-Four	Intra residue 1-4
C-CB1C, CB1N-N, C1N-CB, O1N-CA, CB-N1C	One-Four	Inter residue 1-4

Table 2.1.2.1 Selection of geometric measures calculated

2.1.3 Database

A MySQL Server version: 5.5.65-MariaDB database was implemented, using SQL and the python pandas library for creation and population. Scripts and table specifications can be found on this GitHub link: PSU-Beta Database. The design was considered carefully, changing from a long table, using the Entity-Attribute-Value anti-pattern, to a wide approach which is easier and faster to query on. Tables are: (GitHub link to definitions)

- protein_structure_v1 - each structure including resolution, number of residues, author, refinement software.
- protein_set_v1 - the validity of each structure and the assigned set.
- protein_atom_v1 - atom coordinates, occupancy and B-factor for every atom.
- geo_contact_v1 – atom pair contact distances calculated at < 6.1Å.
- geo_high_v1 - geometric values for the high-resolution data asset.
- geo_2019_v1 - geometric values for the 2019 dataset.
- geo_calcs_v1 – the geometric measures used in the system.

2.1.4 Secondary Structure

The Linux version of DSSP was sourced via “sudo apt-get install dssp”, installing mkdssp 3.0.0 (https://swift.cmbi.umcn.nl/gv/dssp, Joosten et al, 2015; Kabsch & Sander, 1983). The library was accessed via BioPython (Cock et al, 2009; Hamelryck et al, 2003) to calculate the dssp secondary structure for each residue and update the database.

An alternative secondary structure implementation is in column `ss_psu` which contains a very rough estimation of secondary structure based on Ramachandran region from the MSc Structural Bioinformatics course notes.

Both are accessible from the “Correlations” page of the website (PSU-View Correlations) as Hue Choice “SS DSSP” and “Ramachandran Area” respectively.

2.2 Geometric Data Reports

The data is viewed via a web browser, using python, CGI scripts, pandas and matplotlib, called PSU-View (Protein Structure Utility - View).

Figure 2.2.1 Website header with menu options

The website can be found here http://student.cryst.bbk.ac.uk/~ab002/thesis.html and has three main pages to explore: Distributions, Resolution and Correlations, with also Contact Maps, a database explorer and a geometry calculator.

2.2.1 PSU-View Distributions

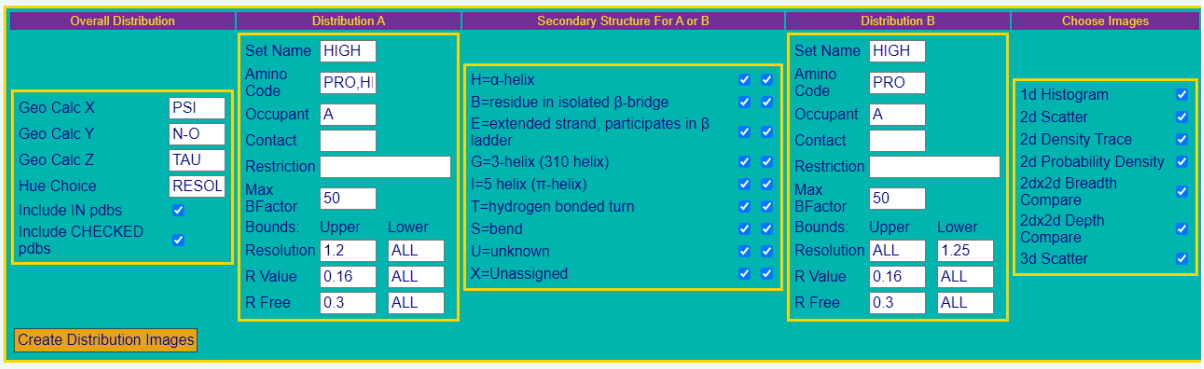


Figure 2.2.1.1 Choices for the distributions page
This shows the selection of options available when viewing geometric data, the results of this selection can be found in Appendix 12

All geometric parameters, or pairs and triples of geometric parameters, can be viewed in several ways, as chosen from the far-right column "Choose Images".

1d Histogram – matplotlib hist with 50 bins. Outliers are shown and statistics are given for the distribution using scipy.stats shapiro, skew and kurtosis.

2d Scatter – matplotlib scatter with the measures chosen in Geo Calc X and Geo Calc Y. The scatter points are graduated on the value in Hue Choice, which defaults to RESOLUTION. If a non-numeric hue is chosen, it is encoded as numeric values with the key given, this leads to automatic assignment of the colours - except for the choice "dssp" which has a fixed assignment of colours.

2d Density Trace – As above for scatter, but with single hued points with transparency of 0.05. This gives an indication of the frequency of the points as well as the location.

2d Probability Density – The gaussian_kde function is used from scipy.stats to form a smooth normalised surface over the scatter point data using Gaussian kernels. The bandwidth is chosen to be 0.10 (see Appendix 19 for some results to demonstrate this choice), with 12 contours. The probability density can be used with the other plots to show the most probable areas: information will be absent for less probable, but still possible, multi-modal distributions.

2d x 2d Breadth Compare – A second distribution can be defined in the column Distribution B. The 2 distributions are represented as a 2d histogram with 2 colour shades: highly populated and slightly populated. The 2 distributions are mathematically compared via numpy arrays of the images to produce a difference image: where they are both highly or slightly populated the image is empty, where they are both populated to a different extent the image is pale grey, where 1 is populated but not the other, the difference image retains that shade. The difference image shows the differing breadths of the images, and the differing locations. The difference image header shows a "masked image metric" in the form of percentages for the left- and right-hand images, calculated by comparing the numerical colour values in the numpy arrays. For example, 90:10 would mean that 90% of the left hand image was also occupied by the right hand, but only 10% of the right hand image was also occupied by the left – the left hand image is presumably much smaller. Identical images would be 100:100, no overlap at all would be 0:0.

2d x 2d Depth Compare – As above, a difference image is created based on distributions A and B. For each, the gaussian_kde function (see 2d Probability Density above) is created, and those numpy arrays are directly subtracted to create a difference array. The hue for all distributions is kept consistent in colour and intensity, so that in the difference image the negative values reflect distribution A and the positive values reflect distribution B. Due to the normalisation of the gaussian_kde the difference image is not substantially effected by the size of the distribution and reflects where distributions have genuinely different probability – comparing the Ramachandran plots for residues in different secondary structures, for example, will show this clearly.

3d Scatter – As for 2d Scatter, but 3d and including the Geo Calc Z measure. The distribution is shown with the axes arranged in three different perspectives.

2.2.2 PSU-View Resolution

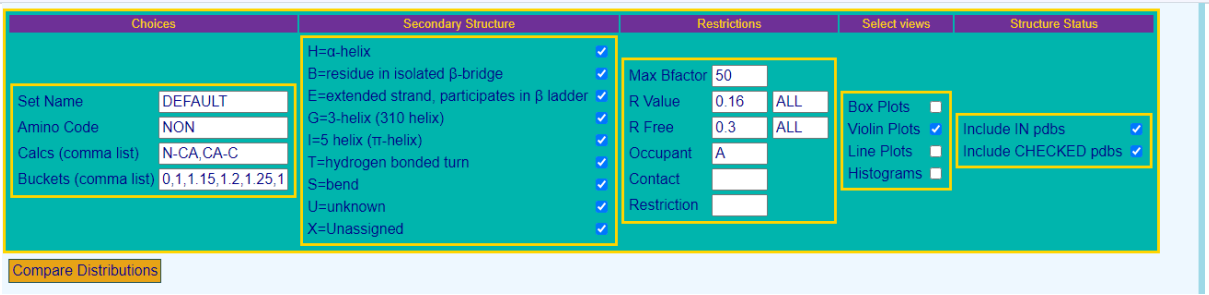


Figure 2.2.2.1 Choices for the resolutions page
The violin plots for calculations CB-O and N-O can be found in Appendix 7

This compares the distributions over different resolution buckets. Distributions can be displayed as boxplots, violin plots, line plots and histograms, using the seaborn library functions boxplot, violinplot, lineplot and distplot respectively. The violin plot uses a kernel density smoothing, the bandwidth is chosen as 0.10. The lineplot uses a kde with cos kernel and silverman rule of thumb.

Multiple resolution buckets can be entered with a comma delimited list, the bucket being between each entered resolution (upper value inclusive).

Calcs is a comma delimited list of all the measures required - they should be chosen to have the same unit and approximate range (Å and º do not compare well on the same axis).

2.2.3 PSU-View Correlation

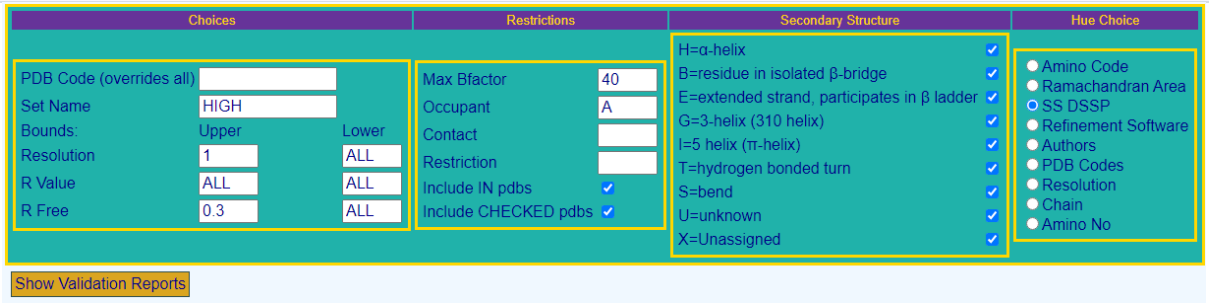


Figure 2.2.3.1 Choices for the Correlation page, generates sixteen correlations considered interesting

Over the course of the project, some correlations have been found to be interesting either geometrically or for validation purposes. This page showcases a selection of 16 of these correlations. It includes standard correlations like the Ramachandran plot and CHI1/CHI2, and some novel such as PSI/N-O and CHI1/CA-CB-CG for proline. The hue can be chosen from the Hue Choice column on the far left, with other choices consistent with the previous pages. The page enables a rapid validation of a single pdb. See Appendix 14 for an example of the selected correlations.

2.2.4 Atom close contacts

The database for close contacts is utilised in 3 ways that facilitate exploration of distributions of inter residue close contacts or restrict on inter-residue close contact.

Use 1: As a restriction on residues in all reports

In the Distributions, Resolutions and Correlations pages, there is a box Contact. In this box you can enter the possible contacts for a restriction on those contacts. This restricts the residues in the data set to those with close contacts on <3.6Å. The options are:

- N-O for any residues where the N is in close contact with another residue's O (> 2 apart)
- O-N for any residues where the O is in close contact with another residue's N (> 2 apart)
- S-S for any cysteine residues where the S is in close contact with another S
- CB-CB for any residue in close contact with another Cβ
- CA-CA close contact between the Cα and another Cα
- XN-O means NOT N-O, the complement set to N-O
    - The X can be used on all the above

Use 2: As a geo measure in histograms

When looking at 1 dimensional data either on the Resolutions page, or for histograms on the Distributions page, the "Geo Calc X" or the Calcs can be entered as, for example C@N-O (or C@SG-SG etc). The C@ being a notation for contact. This will show a distribution of all close contacts directly from the contact database (which is restricted at 6.1Å not 3.6Å). See example below.

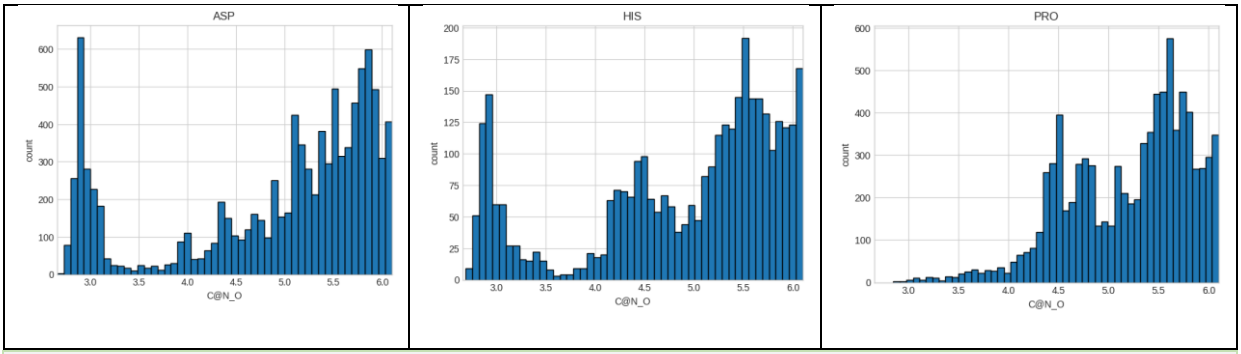


Figure 2.2.4.1 Results for the close contacts for histograms on Distributions page

Use 3: Contact Maps

Contact maps can be viewed directly for a single pdb. This feature is not yet used for further analysis in this study. The Contact Map page shows the close contact map for all the atom pairs calculated, SG-SG, CA-CA, CB-CB and N-O. The data is shown on colour and size for diminishing distance,

with the effect of 3d along the backbone, consideration of which holds future promise. An example is shown below for 6mry, chosen as having the largest number of SG-SG contacts in the database.

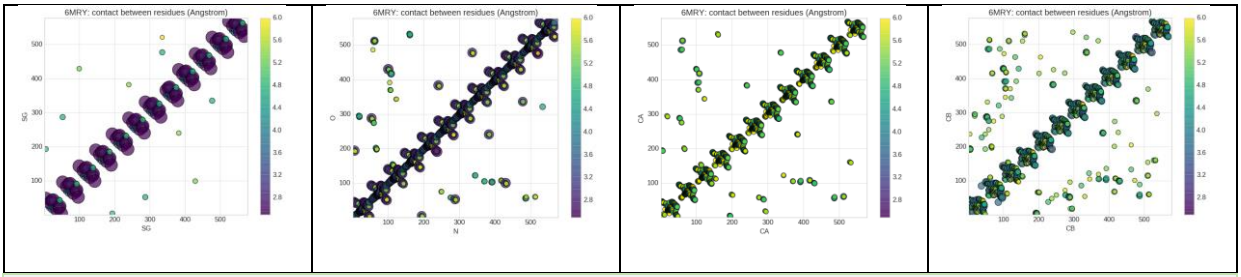


Figure 2.2.4.2 Contact Maps, 4 contact views for 6mry

2.2.5 PCA data analysis

PCA analysis of proline ring conformations was performed using R with a protocol derived from the MSC Bioinformatics statistics module practical. See link for script: R PCA Analysis

2.2.6 Validation

Validation of the data was performed using the reports produced by PSU-View and detailed above. The histogram report produced in the Distributions page shows the outliers. This was used to clean data of any evident errors in either structure or code. Two code errors were detected in this process:

- Not recognising breaks in protein chains, i.e. where residues are missing due to poor electron density. This was fixed.
- Incorrectly handling mutation insertions, column 27 of an atom row in the pdb file. Where mutations are found in the structure the different residues at that point are entered at the same residue number with the order in which they are found in column 27. The few structures not handled by this were removed from the dataset.

The Correlations page was also used to detect structures that fell into areas that would seem to be geometrically impossible. These structures were further investigated individually and either rejected due to evident mistakes in the deposited structure or annotated as “Checked”. An example is detailed in Results Section 3.1, with a full list in Appendix 3.

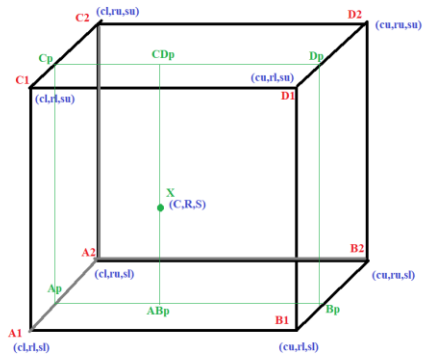
2.3 Electron Density

A publicly available python library, pdb_eda, was evaluated for electron density functionality (Yao et al, 2019). This library was used, with some changes, along with BioPython (Cock et al, 2009; Hamelryck et al, 2003) in the novel python library PSU-ED, PSU-ED on GitHub. Speed and memory considerations were foremost.

The library provides access to the density matrix 2Fo-Fc and the difference matrix Fo-Fc.

2.3.1 Interpolating density matrices

The pdb_eda library was evaluated for the density retrieval from the matrix. Their point density was based on a nearest neighbour approach and was not smooth. Their smoothed density, using a spherical average of chosen radius, was too slow. The decision was made to implement a trilinear interpolation for smoother but faster point density. See Figure 2.3.1.1 for the overall method and 2.3.1.2 for the linear interpolation step.



Given a non-integer point in the density matrix (C,R,S) converted from atom space (X,Y,Z), a cube is formed around it from the integer coordinates of the floor and ceiling of each point.

The density is found for each vertex of the cube, labelled A1-D2.

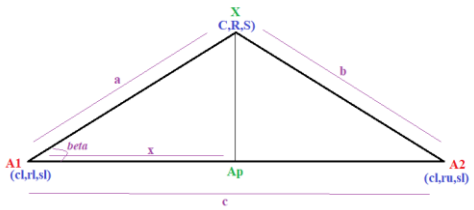
4 linear interpolations are performed along the cube sides A1-A2, B1-B2, C1-C2 and D1-D2 to find Ap, Bp, Cp and Dp.

2 linear interpolations - across cube sides Ap-Bp and Cp-Dp find central surface points ABp and CDp.

Final interpolation through the cube between ABp-CDp finds the interpolated density X at (C,R,S).

Figure 2.3.1.1 Trilinear interpolation, overall method, depicts the coordinates in the cube around the point to interpolate

Each interpolation step finds both the interpolated density and the coordinates of the new interpolated point. The interpolation is always performed with the central (C,R,S) coordinates and the 2 points being interpolated. Each interpolation step is performed as below, which is given as an example for points A1, A2 and X in the diagram above.



Interpolation Step

The ratio x/c is calculated using the cosine rule

$$\frac{x}{c} = \frac{a^2+c^2-b^2}{2c^2}$$

This ratio is applied to linearly interpolate densities:

$$\text{Ap} = A1 + \frac{x}{c} \times (A2 - A1)$$

It is also applied to the coordinates such that

$$(\text{cp,rp,sp,}) = (\text{cl} + \frac{x}{c} \times (\text{cl} - \text{cl}), rl + \frac{x}{c} \times (\text{ru} - \text{rl}), \text{sl} + \frac{x}{c} \times (\text{sl} - \text{sl}))$$

Figure 2.3.1.2 Trilinear interpolation, the linear fraction calculated from the cosine rule

2.3.2 Normalising density matrices

The library pdb_eda has a novel approach to the problem of normalisation of density matrices – they convert the arbitrary density to number of electrons per unit volume using a method that finds the density and density difference in spherical areas around each atom in the structure and compares that to the expected number of electrons for the structure, finding a conversion factor that can be applied across the whole density matrix. The method was evaluated, and some reservations are found, specifically the use of the atomic coordinates of the solved structure to make this calculation: errors in the placement of the atomic coordinates could lead to an incorrect summation for the electron density; keeping the electron density independent of the solved structure leaves available the ability to solve the structure from the electron density. However, the final decision not to use this method was based on speed – for some structures, for example 5gsm, the calculation of a single conversion factor takes 45 minutes, raising the prospect of it taking 15 days to calculate the conversion factors for 500 structures.

The pdb_eda method relies on the calculation of a single conversion factor. If this is a reasonable method, the density matrix values must all be distributed equivalently such that a single scale factor renders them similar. Thus a simpler method of normalising the density matrices was evaluated using the approach that the density distribution would be similar in all proteins, and that the median value in a specific density matrix would approximately correspond to the same thing in all density matrices. The density matrices are scaled by a linear factor such that the median is 50, an arbitrary choice that allows comparison of density between structures. This is a first attempt at normalisation and is under review: the results are promising. This normalisation method has been used in generating superpositions of electron density in this report. An analysis of a selection of density and density difference matrices at different resolutions are given in Appendix 5.

2.3.3 Superposition of density matrices

A method was developed to find similar atom configurations and superpose the electron density to draw out features of geometry, bond and shape that may not be visible or reliable on an individual basis.

A cube is created of a configurable size and gap. Three atoms are defined for each sample – a central atom, a linear atom and a planar atom. Transformations are applied to the original cube, so that, net-like, it can capture a cube of required space in the sample structure. The cube is manipulated via translations and rotations such that the central atom is at the origin, the linear atom is on the x-axis, and the planar atom lies flat against the x-y plane. Density is then retrieved for every point on the cube's grid, using the normalised interpolated methods described above.

PSU-ED results are produced in an html document, showing all x, y and z slices generated from matplotlib image libraries. Results are given for each individual sample and the superposition.

Results can also be viewed in 3d through a Mathematica notebook, adapted from code written by Mark Williams - Mathematica Notebook.

3. Results

3.1 Summary of structures and residues

There are 2 sets of data, non-homologous at 90%

- HIGH with resolutions <=1.3Å
- 2019 is all structures deposited in 2019 >1.3Å

The 2 datasets facilitate comparison of resolutions, with the non-high set chosen as being from 2019 to minimise variability in the data (choosing recent structures suggests a consistency in versions of refinement software and methods that could reduce variability).

The 2019 data set has not undergone manual validation.

Following Jaskolski (2007), a high quality (HQ) subset of the HIGH structures is defined by applying these filters:

- rfree <= 0.3
- rvalue <=0.16
- bfactor <= 50
- Checked excluded

The structure count is given in Table 3.1.1.

HIGH				2019	
Resolution	HQ set	All	"CHECKED"	Resolution	All
0.5	2	2	0	1.3	71
0.6	3	3	0	1.4	133
0.7	9	10	0	1.5	155
0.8	47	47	1	1.6	208
0.9	139	142	2	1.8	278
1.0	455	510	3	1.9	313
1.1	822	980	5	>2.0	1491
1.2	1015	1415	6		
1.3	942	1508	4		
Total	3434	4617	21		2649

Table 3.1.1 The included structures broken down by resolution and dataset
The resolutions are rounded to 1 d.p. for this analysis

3.2 Validation

A process of validation was undergone to check all structures for extreme outliers. The Ramachandran plot is the standard approach (Ramachandran et al, 1963) to identifying unusual backbone geometry. However, in developing the analysis reported here other well-defined correlations have been found between backbone geometric features that (arguably) better highlight outliers. The results of this inspection of the high-resolution data set can be found in Appendix 3 with an example described below.

3.2.1 Structure 1i1w, 0.89Å Thermostable Xylanase

Figure 3.2.1.1b below shows a clear geometric outlier that is not evident on the Ramachandran plot – see the point at (PSI,N-O) = (10,3.05). In this plot there is a relationship seen between the Ramachandran plot and the PSI/N-O plot - the secondary structure colours illuminate this, see in particular the brown areas that are undefined in the Ramachandran plot but appear ordered in PSI/N-O. This geometric order in PSI/N-O facilitates the identification of unlikely residues.

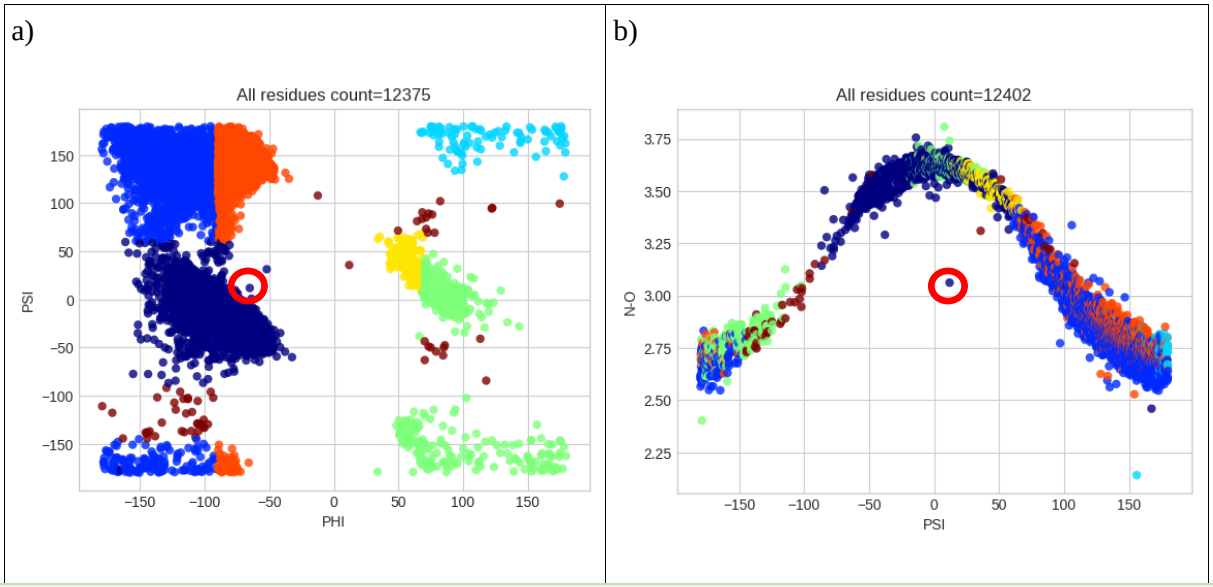


Figure 3.2.1.1 A residue in a geometrically unlikely area is picked out on both plots.
Structures between 0.8-0.9 Å.
a) *Shows the Ramachandran plot with secondary structure regions denoted by different colours*
b) Shows the geometric correlation PSI vs N-O, with the same data set and colouring as (a)

Many clear outliers have been checked by hand; this point is in structure 1I1W. The residues at 180 (SER) and 181 (TYR) both have occupants A and B, and the occupant A for SER seems to deviate visually from standard geometry – see Figure 3.2.1.3. Visually it can be seen that the PSI angle is as reported, close to planar. However, the carbonyl C is expected to be sp^2 hybridised, planar with angles of 120°. It clearly deviates from this. The same residue is an outlier in other correlation plots (Figure 3.2.1.2).

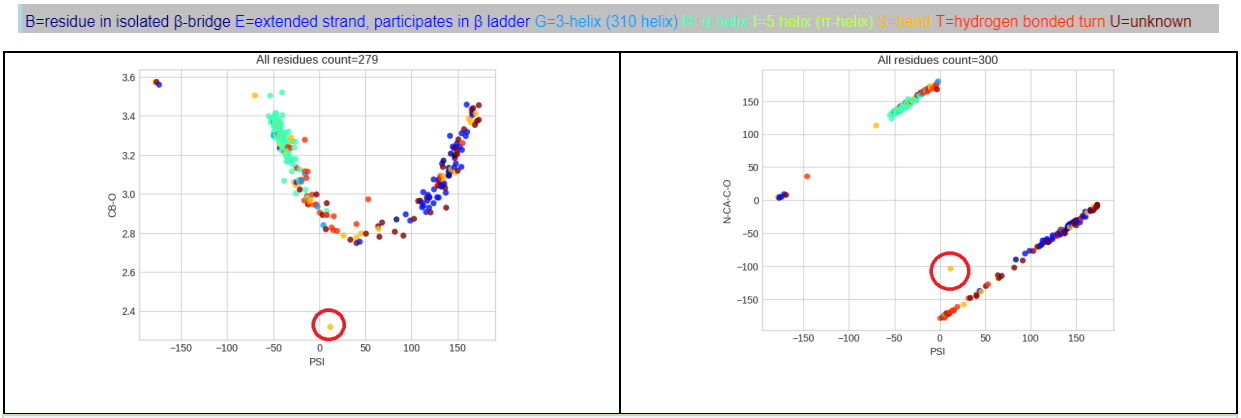


Figure 3.2.1.2 Correlation plots for 1i1w showing the suspicious residue 180 (circled)

I have examined the electron density in Chimera (Figure 3.2.1.3) (Pettersen et al, 2004), verified the N-O distance manually and cross-calculated PHI, PSI and N-O as if the A&B occupants had been mixed up, to check for simple labelling errors – see Table 3.2.1.4.

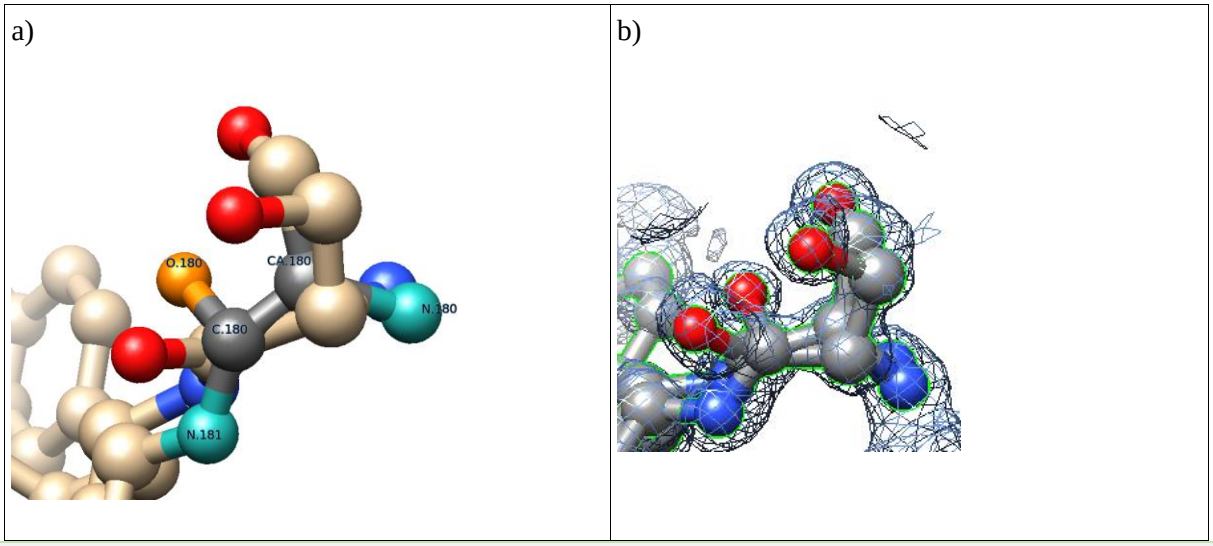


Figure 3.2.1.3 The geometric outlier Ser 180 in 1I1W, planar carbonyl not as expected, with electron density.
a) Shows atoms of interest highlighted in orange/turquoise/grey.
b) The electron density shows evidence for atom placement.
Images produce in Chimera.

Residue 0 ALA.179	Residue 1 SER.180	Residue 2 TYR.181	PHI	PSI	N-O
One occupant C (-5.452,12.011,19.188)	Occupant A N (-5.289,13.095,19.649) CA (-5.409,14.625,19.282) C (-4.663,15.095,18.398) O (-3.802,15.745,19.274)	Occupant A N (-3.804,14.715,17.422)	-65.14º	11.92º	3.062
	Occupant A	Occupant B N (-3.031,14.664,18.432)	-65.14º	-44.38º	3.062
	Occupant B N (-4.912,13.362,19.958) CA (-5.680,14.311,18.905) C (-4.097,15.066,18.404) O (-4.480,16.189,17.660)	Occupant A	-119.73º	106.70º	3.669
	Occupant B	Occupant B	-119.73º	26.06º	3.669

Table 3.2.1.4 Calculations to investigate possible labelling error of occupants A and B, 1I1W residues 179-181.

The combinations for N-O/PHI/PSI do not suggest the occupants have been mixed up as other combinations of the values do not fall on an ordered region, see Figure 3.2.1.5 below for the different combinations plotted.

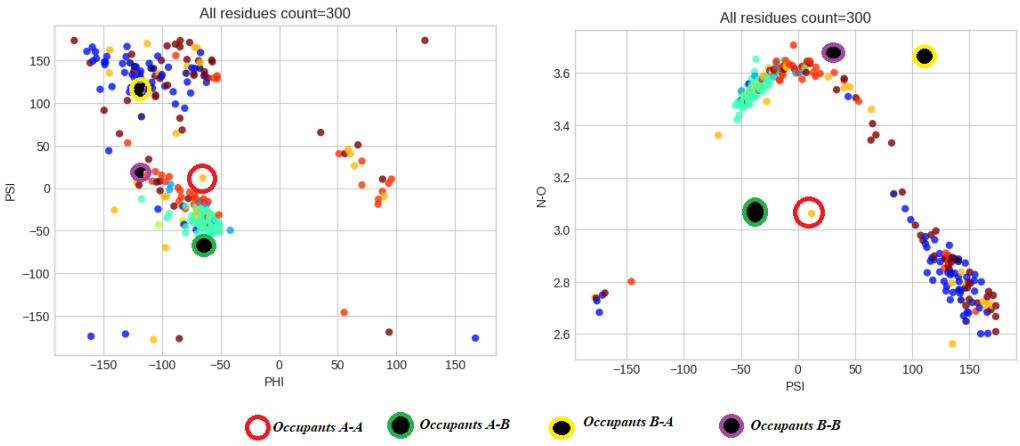


Figure 3.2.1.5 Correlations plotted for occupant combinations for 1i1w residues 180 and 181
The points are clearly in error on PSI/N-O but seem ordered on the Ramachandran plot

The deposition paper for this structure (Figure 3.2.1.6) does not make any specific mention of these residues as sites of interest (Natesh et al, 2003). It seems that there is a mistake in the solved structure despite the high resolution of 0.89Å and level of attention that has gone into it.

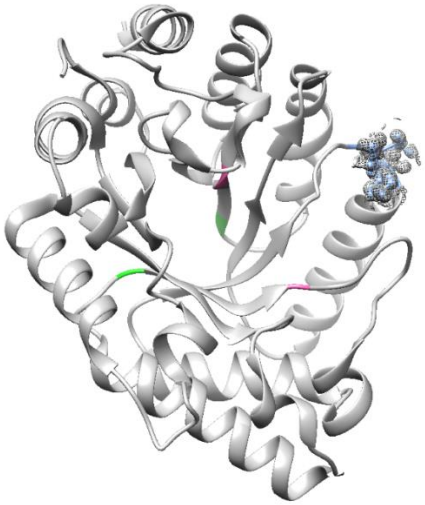


Figure 3.2.1.6 Chimera image shows geometrically unlikely region in 1i1w and the active site Residues Ala179, Ser180 and Tyr181 are shown in blue, with the active site residues Glu131 and Glu237 in pink and the salt bridge Arg124-Glu232 in green.

Th process described takes time and was performed for the high-resolution data due to that data being the primary goal of the project. It was not possible to spend that level of detail on the 2019 data set. Where structures are found to be correctly representing the deposited structure with no evidence to reject (other than geometric unlikeliness), they are marked CHECKED in the database. This enables them to be filtered out or specifically analysed when required. This process shows that even at 0.89Å mistakes are made, and the identification can be time consuming.

3.3 Results for bond lengths and angles

3.3.1 Jaskolski and E&H backbone comparison

In 1991 Engh and Huber published standards for bond length and angle restraints (Engh & Huber, 1991) used in refinement of protein structure, with further updates in 2001. This was reviewed in 2007 (Jaskolski et al, 2007) using 10 ultrahigh-resolution structures, with some recommendations for updates. Below the results from those previous publications (Jaskolski [Table 2], 2007) are compared to the results obtained from the HQ and HIGH datasets.

Table 3.3.1.1 shows a summary of the data from the HQ set (upper row) and HIGH set (lower row) against Jaskolski's screened (upper row) and all structures (lower row). Note the stability of the median and iqr even in the HIGH dataset due to the reduced influence of rare deviation from the most probable value.

		E&H	Jaskolski	This study		
Data	**Measure**	**Mean (sd)**	**Mean (sd)**	**Mean (sd)**	**Median (iqr)**	**Sample size**
Screened/HQ	N-CA	1.458 (19)	1.454 (12)	1.455 (9)	1.455 (9)	1098
All/HIGH	(exc GLY,PRO)		1.456 (15)	1.455 (12)	1.455 (11)	2324
Screened/HQ	CA-C	1.525 (21)	1.527 (13)	1.525 (10)	1.525 (11)	1162
All/HIGH	(exc GLY)		1.526 (14)	1.526 (13)	1.526 (12)	2469
Screened/HQ	C1N-N	1.329 (14)	1.334 (13)	1.331 (11)	1.331 (10)	1202
All/HIGH	(exc xxx-PRO)		1.334 (18)	1.333 (13)	1.332 (12)	2549
Screened/HQ		1.231 (20)	1.234 (12)	1.233 (10)	1.234 (11)	1285
All/HIGH	C=O		1.234 (13)	1.234 (11)	1.234 (12)	2736

Table 3.3.1.1 Bond-length comparison for the highest resolution structures, compared with E&H and Jaskolski.

Table 3.3.1.2 compares the results from Jaskolski and this study by resolution, (Jaskolski [Table 4], 2007). I have included additionally to Jaskolski the bond length N-CA as the results clearly show a reduction in bond length per with higher resolution to 1.455Å. This change agrees with Jaskolski's overall value and suggests a need for change to the E&H accepted value of 1.458Å. Note that the standard deviation and interquartile range of my data is in almost all cases lower than the EH and Jaskolski data.

Resolution	C1N-N mean (sd)			C=O mean (sd)			N-CA mean (sd)	
	Jaskolski	HQ set	HIGH Median	Jaskolski	HQ set	ALL Median	HQ set	ALL Median
<=0.8	1.334 (18)	1.332 (11) [1291]	1.333 (12) [3008]	1.234 (13)	1.233 (10) [1381]	1.234 (12) [3220]	1.455 (10) [1179]	1.455 (11) [2735]
0.8<=0.9	1.333 (16)	1.332 (11) [2710]	1.331 (13) [11849]	1.236 (13)	1.234 (11) [2882]	1.234 (14) [12572]	1.458 (11) [2488]	1.457 (14) [10859]
0.9<=1.0	1.332 (14)	1.330 (10) [16479]	1.330 (12) [42719]	1.236 (13)	1.233 (11) [17490]	1.234 (14) [50201]	1.457 (13) [15185]	1.458 (14) [43516]
1.0<=1.1	1.329 (13)	1.329 (9) [35273]	1.329 (10) [100066]	1.233 (13)	1.233 (11) [37497]	1.233 (12) [106543]	1.458 (12) [32583]	1.458 (12) [92417]
1.1<=1.2	1.330 (12)	1.329 (9) [35746]	1.329 (9) [129831]	1.236 (12)	1.233 (11) [38058]	1.233 (11) [138138]	1.459 (12) [33127]	1.459 (12) [120146]
1.2<=1.3	1.329 (10)	1.329 (8) [43004]	1.330 (8) [214592]	1.233 (11)	1.232 (12) [45837]	1.233 (11) [228271]	1.460 (11) [39854]	1.459 (11) [199177]
1.3<=1.4	1.329 (9)	1.333 (5) [271]	1.332 (8) [24572]	1.232 (11)	1.235 (8) [292]	1.236 (7) [26075]	1.459 (6) [251]	1.46 (10) [22643]
1.4<=1.5	1.329 (16)	1.325 (9) [927]	1.332 (8) [32179]	1.232 (11)	1.238 (16) [999]	1.236 (7) [34220]	1.477 (10) [857]	1.461 (16) [29875]
1.5<=1.6	1.329 (7)	1.331 (4) [151]	1.333 (8) [39660]	1.234 (11)	1.237 (6) [162]	1.236 (8) [42155]	1.463 (7) [144]	1.461 (14) [37108]
1.6<=1.7	1.329 (7)	None	1.333 (8) [34143]	1.233 (11)	None	1.235 (7) [36319]	None	1.46 (11) [31615]
1.7<=1.8	1.329 (7)	None	1.332 (7) [58713]	1.233 (11)	None	1.236 (7) [62111]	None	1.46 (11) [54693]
EH Value	1.329 (14)			1.231 (20)			1.458 (19)	

Table 3.3.1.2 Bond lengths on resolution, Jaskolski vs PSU-Beta HQ set and HIGH/2019 data
The count is given below each PSU-Beta data in square brackets. The HQ set is manually cleaned of outliers to the resolution of 1.3. The median column shows median (iqr) instead of mean(sd). The absence of data at the lower resolutions in the HQ set is due to the strict requirement of bfactors and rvalues.

Table 3.3.1.3 shows tau values (Jaskolski [Table 3], 2007), depicting these distributions as violin plots in Figure 3.3.1.4 using the HQ set at 3 different resolutions - the first aggregated for all but pro and gly: then pro; gly; his; met; and trp. It is interesting to note that there has not been a consensus on the nature of the tau distribution, with some dispute concerning bimodality (Jaskolski, 2007) where they suggest that although wide, the tau distribution is not bimodal, illustrating this with a histogram in their paper. The results in this study show evidence of bi/multi-modality and different characteristics for each amino acid (see Appendix 7 for the results for individual amino acids). The aggregation of the amino acids with different modalities distorts the view. These differences are demonstrated with violin plots in Figure 3.3.1.4 and using the depth compare in Figure 3.3.1.5.

a)

Type	GLY	PRO	Other
E&H	112.50 (2.90)	111.80 (2.50)	111.20 (2.80)
PDB+ (Jaskolski)	113.91 (2.23)	112.48 (2.19)	110.72 (2.22)
PDB- (Jaskolski)	113.80 (2.28)	112.44 (2.31)	110.61 (2.41)
HQ set	113.41 (2.34)	112.61 (2.03)	110.41 (2.21)
HQ set (median/iqr)	113.61 (3.40)	112.94 (3.20)	110.39 (3.13)
HIGH set	113.50 (2.33)	112.35 (2.27)	110.51 (2.29)
HIGH set (median/iqr)	113.77 (3.60)	112.45 (3.38)	110.56 (3.20)

b)

Amino Acid	ILE	VAL	HIS	LYS	TRP	CYS	GLN	LEU	ARG	THR	MET	TYR	GLU	SER	ALA	PHE	ASP	ASN	PRO	GLY
count	66	111	24	80	18	22	39	86	47	99	33	45	71	63	116	38	77	63	64	123
mean	108.8	109.2	109.53	110.15	110.24	110.28	110.29	110.3	110.36	110.49	110.54	110.74	110.79	110.86	110.86	111.25	111.36	111.47	112.61	113.41

Table 3.3.1.3 Tau value comparisons using Jaskolski Table 3 (Jaskolski et al, 2007)
HQ set, resolution < 0.8Å
a) *This study (in pink) agrees with the Jaskolski data, both for the HQ set and for all data. The median is seen to be a good alternative to the mean.*
b) The mean for each amino acid shows a wide spread of tau values in the HQ set <0.8Å

Table 3.3.1.3(a) shows agreement between the Jaskolski tau values and this study. However, further breakdown in (b) shows that each amino acid is quite different. The observations are few at this high resolution, so it can become difficult to glean information with certainty on an individual amino acid basis.

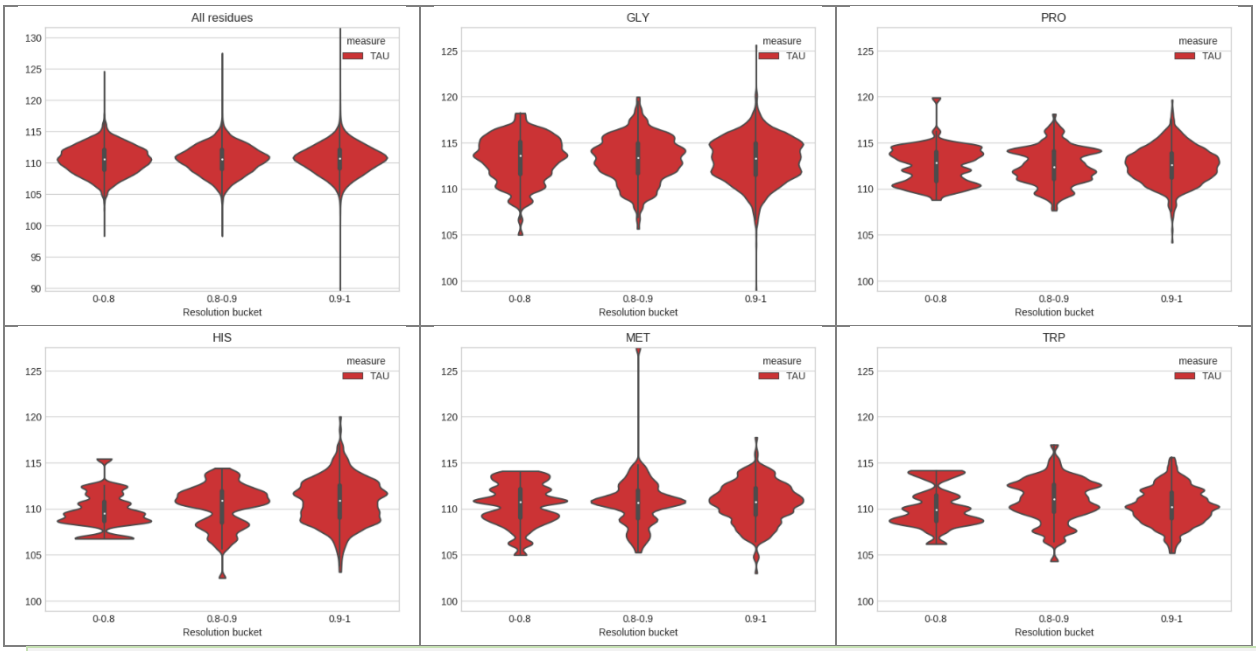


Figure 3.3.1.4 Violin plots show tau distributions for all but pro and gly against individual amino acids in the HQ set.
The median is the white dot, the interquartile range the central thick black bar, the thin black line defines outliers at quartile += 1.5 IQR, with the thickness of the plot showing the distribution
This demonstrates that putting all the amino acids together obscures the different modalities of the amino acids.
Seaborn kde smoothing is used: violinplot(kde=0.10)

The violin plots are highly dependent on the kde settings and do not tell a reliable story for these distributions. A better demonstration of tau differences between different amino acids and an apparent

bimodality is through using the depth compare image (see Methods 2.2.1). In Figure 3.3.1.5 a correlation is shown for TAU against PSI for the amino acids ILE, SER, THR and LEU. PSI leads to a clear bimodality, but additionally, the tau areas are slightly different for each PSI region. The depth compare shows a difference image between isoleucine and threonine, and serine and leucine. In both cases the different amino acids clearly favour different regions of tau. Tau seems to have a subtle bimodality, where the different PSI regions associate with a slightly different TAU , but when TAU is viewed alone in 1-dimension these modalities instead appear as a spread.

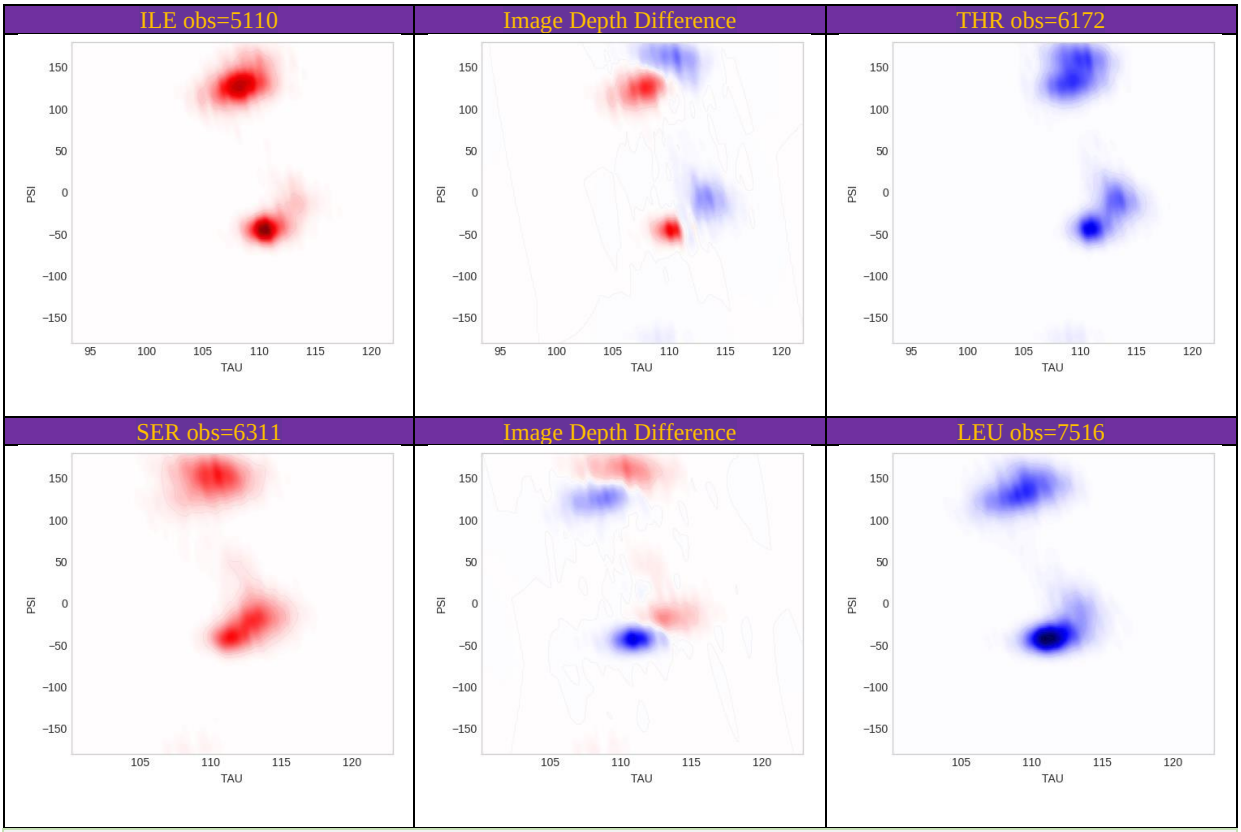


Figure 3.3.1.5 Bimodality in tau correlated with psi and different favoured tau regions for ser, leu, ile and thr. This image demonstrates both the bimodality of tau and the different regions favoured by different amino acids. The correlation with PSI makes the bimodality clearer and links it to structure.
Residues selected from HQ set at resolution <=1.2Å

3.3.2 Other geometric measures

A selection of distance, angle and dihedral distributions are below, picked out to demonstrate the multi modal nature of the distributions, as well as the differences between amino acid types.

Figure 3.3.2.1 shows the distributions for the main chain dihedral angles PHI, PSI and OMEGA. They are given for PRO, CYS and GLY as an indication of how the amino acids differ, e.g. the almost symmetry of glycine is clearly shown. The full results are given in Appendices 9 and 10 for PHI and PSI.

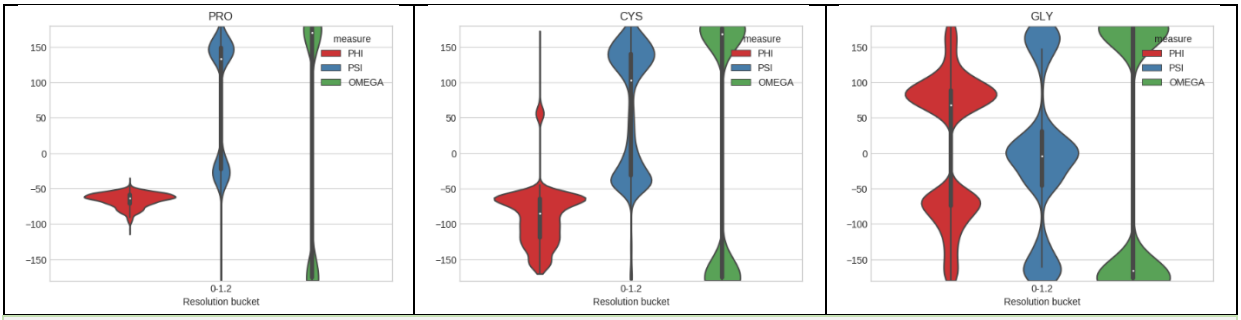


Figure 3.3.2.1 Violin plots for PHI, PSI and OMEGA – PRO, CYS and GLY
Resolution <=1.2Å for the HQ set. This demonstrates the bi/multi modal nature of the mainchain dihedral angles and the distinct character of the different amino acids.

Figure 3.3.2.2 shows the distributions of distances between N-O and CB-O for PRO, ILE and ASN, 1-4 intra residue distances. These non-bonded distances reflect features of the backbone geometry, both representing a twist around the CA-C bond. The amino acids are clearly different with N-O distinctly bimodal: the results for all amino acids are in Appendix 8.

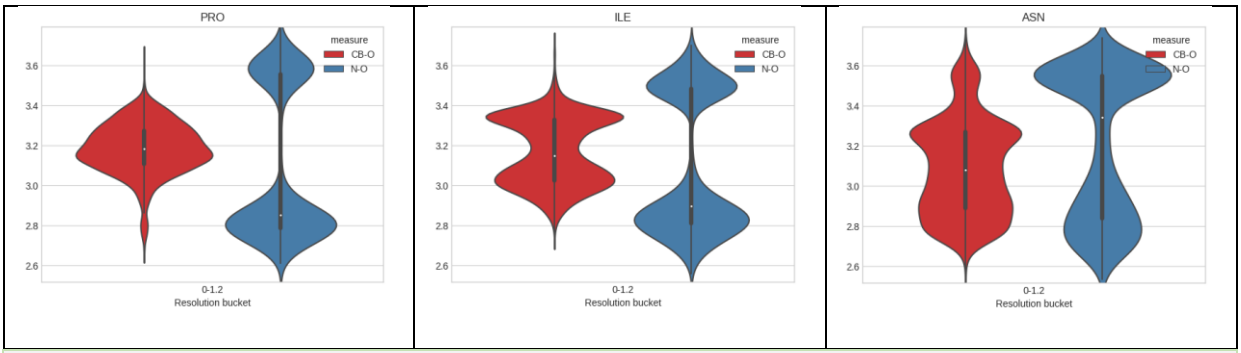


Figure 3.3.2.2 Violin plots for N-O and CB-O for PRO, ILE and ASN.
Resolution <=1.2Å for the "hq set". This demonstrates the bi/multi modal nature of the intra 1-4 measures and the distinct character of the different amino acids.

Figure 3.3.2.3 shows the distributions of distances between the previous Cβ and N (also reflecting the CA-C bond) and the previous O and Cβ for PRO, ILE and ASN, reflecting the sidechain conformation - again demonstrating the amino acid-distinct, multimodal non-normal nature of these distributions.

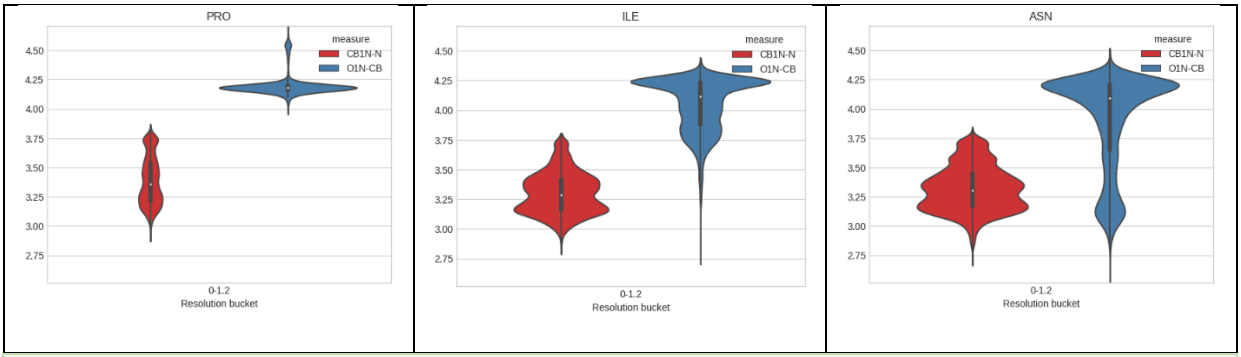


Figure 3.3.2.3 Violin plots for CB-N and O1N-CB for PRO, ILE and ASN.
Resolution <=1.2Å for the HQ set. This demonstrates the bi/multi modal nature of inter residue measures and the distinct character of the different amino acids.

3.4 Results for new insights

3.4.1 Geometric Correlations

To provide insight into the interrelatedness of the geometric measures, further analysis using scatter plots and probability density diagrams was undertaken.

The standard plot in structural bioinformatics is the Ramachandran plot (Ramachandran et al, 1963), shown below produced in PSU-View as a scatter diagram graduated on resolution, as a probability density plot, and as a scatter diagram graduated on secondary structure.

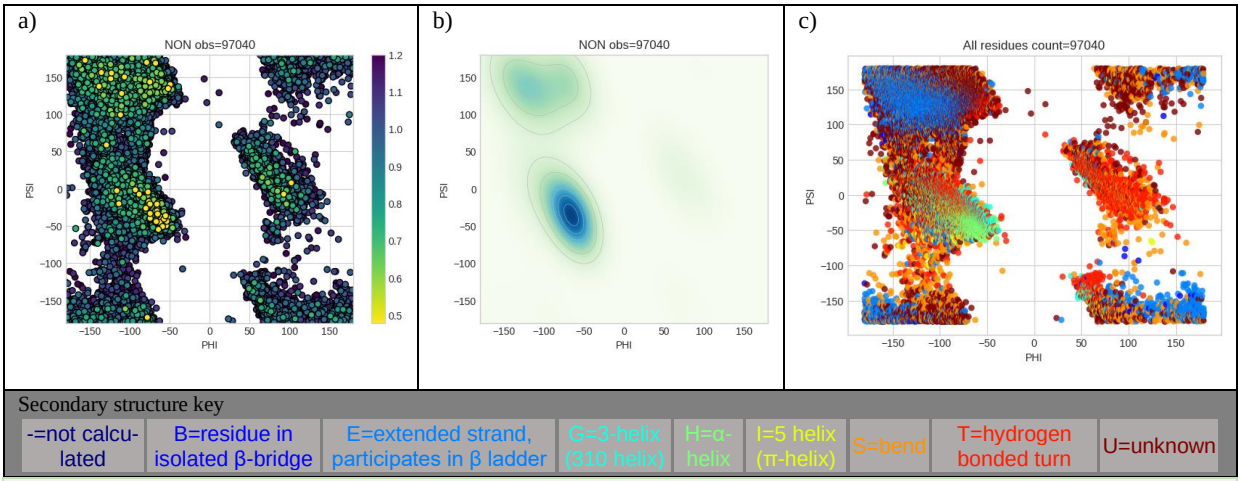


Figure 3.4.1.1 The Ramachandran plot from PSU-ALPHA, 3 views, with the secondary structure key.
All HQ set residues selected with resolution <=1.2
a) *The Ramachandran plot with resolution as the hue*
b) The probability density of the Ramachandran plot
c) The Ramachandran plot with secondary structure as the hue

The web viewer's correlation page contains several plots that are standard: PHI/PSI the Ramachandran plot; CHI1/CHI2 as suggested as another validation tool by Rose (2019); OMEGA/TAU which is suggested as correlating to secondary structure features in the 1i1w deposition paper (Natesh et al, 2003). Some are simply validation plots for extreme values, such as CA-C/N-CA to check the main chain lengths. Exploration of the data has yielded some novel plots that provide interesting correlations and suggest areas of geometric necessity, see Figure 3.4.1.2. For example PSI/N-O and PSI/CB-O which can be further viewed as a parametric sine curve for N-O/CB-O with PSI underlying; and the "square plot" of CA2N-CA1N-CA/CA-CA1C-CA2C which correlates angles made by shifting frames of Cαs.

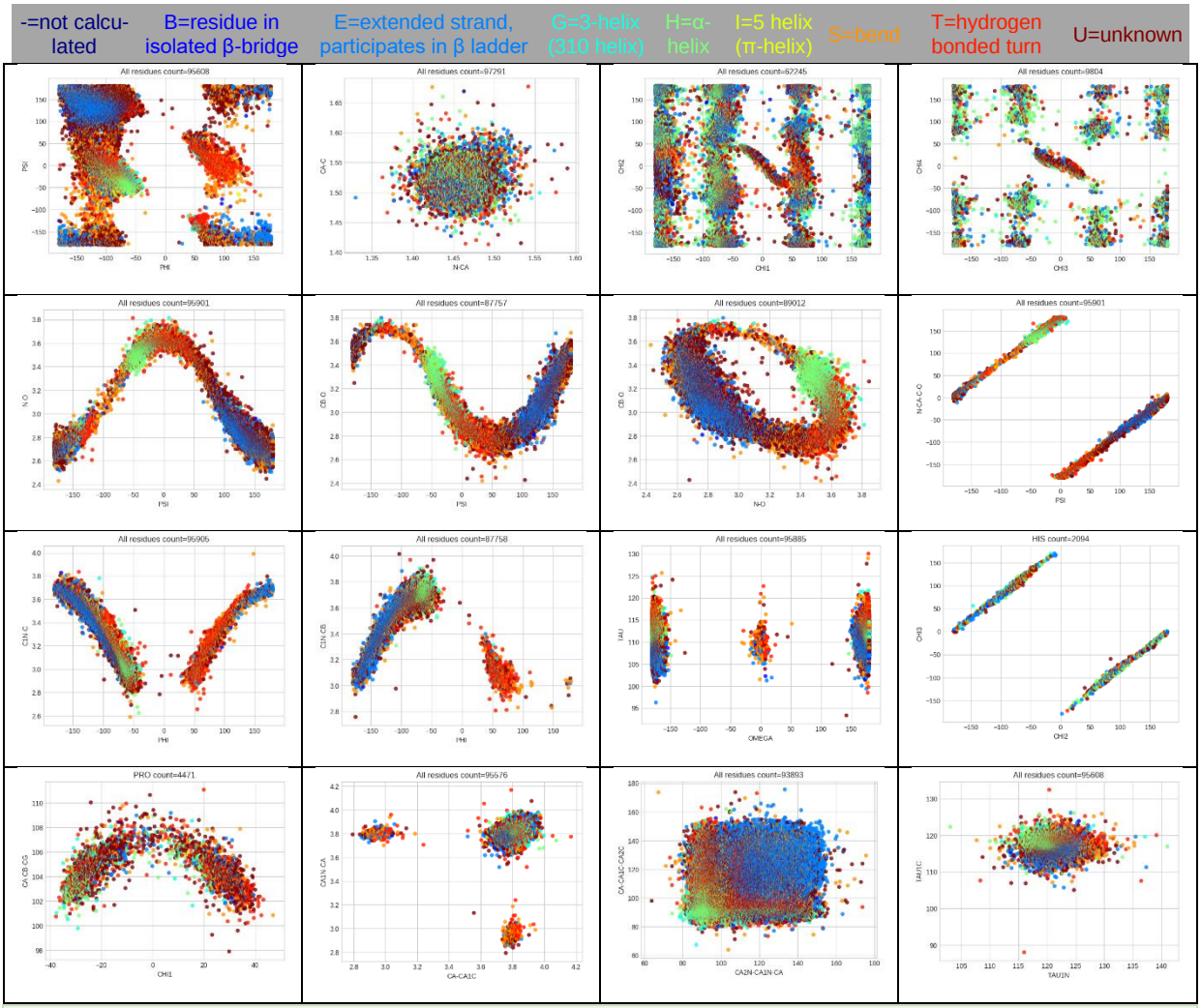


Figure 3.4.1.2 "Correlations" page in PSU-View for <=1.2 Å, HQ set
This shows the selected set of correlation plots chosen as interesting or useful for validation, and the dssp hue demonstrates the locations of secondary structure and how they relate across the correlations.

Although CHI1/CHI2 shows all amino acids together, proline forms a distinct area which can be distinguished in Figure 3.4.1.2 and is shown clearly in Figure 3.4.1.3. Additionally for proline, a novel geometric correlation is shown on the correlations page above in Figure 3.4.1.2, (bottom left) correlating CHI1 against the angle CA-CB-CG, demonstrating the twist over the CA-CB bond (CHI1) distorts the CA-CB-CG angle uniquely to proline.

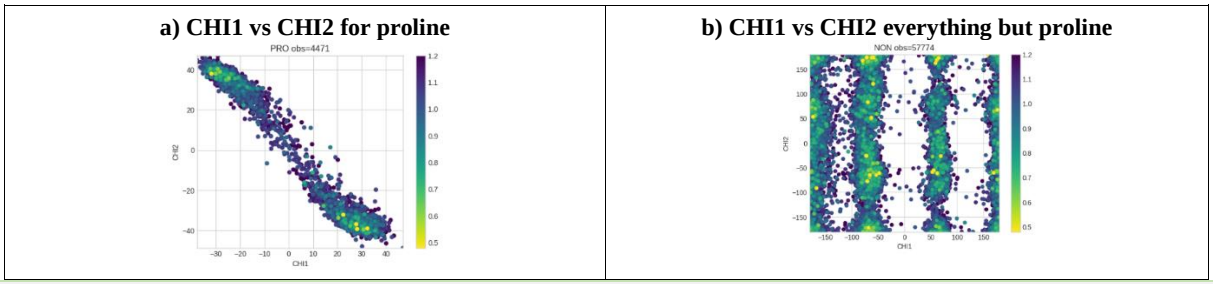


Figure 3.4.1.3 Geometric correlations, CHI1 vs CHI2, graduated on resolution, HQ set, resolution <=1.2Å

The “square plot” bounds the possible values of angles along 3 successive Cαs at between 80Å and 150º approximately, associating strongly with secondary structure. Figure 3.4.1.4 below shows four “square plots” in different secondary structure groups.

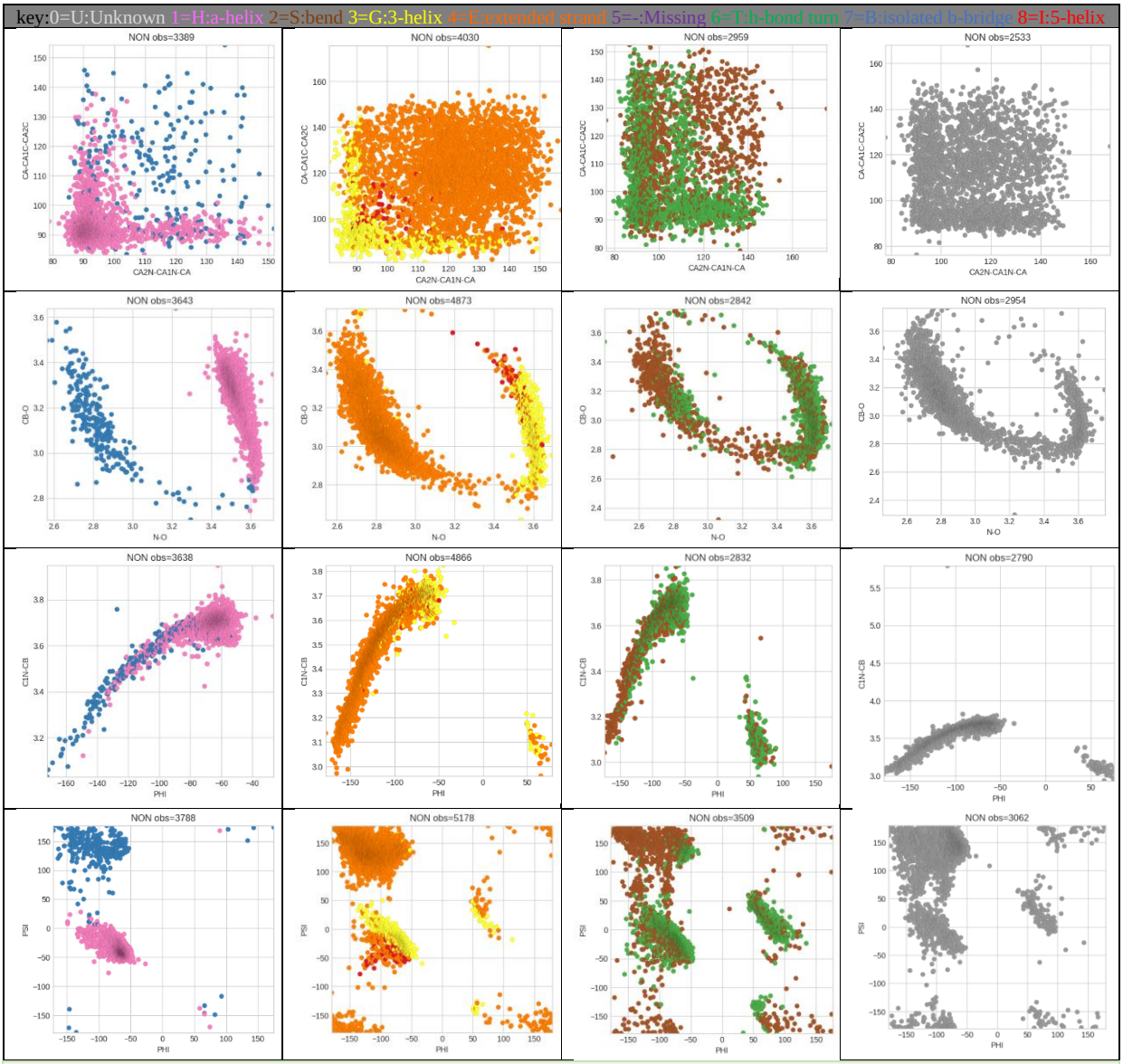


Figure 3.4.1.4 Comparing “the square plot”, the ellipse, PHI/C1n-CB and Ramachandran on secondary structures.
HIGH set <=0.9Å
This figure separates the secondary structures into 4 groups for ease of identifying the regions they occupy. The last group is unknown – dssp did not make an assignment. The changing views of the correlation plots make it seem possible these could be identified and assigned.

The different secondary structures have clear similarities in subsets, e.g. 3-helix and a-helix form a group distinct from extended strand and s-bend. The hydrogen bonded turn is distinct, and 5-helix also has distinct regions.

3.4.2 Rarity effect

There is a strong association shown between rvalue, rfree, bfactor and resolution, as can be seen below in Figure 3.4.2.1 showing the same three N-O/CB-O plots coloured according to changes in these 4 variables (bfactor is calculated as the maximum for the structure, not the atom/residue bfactor).

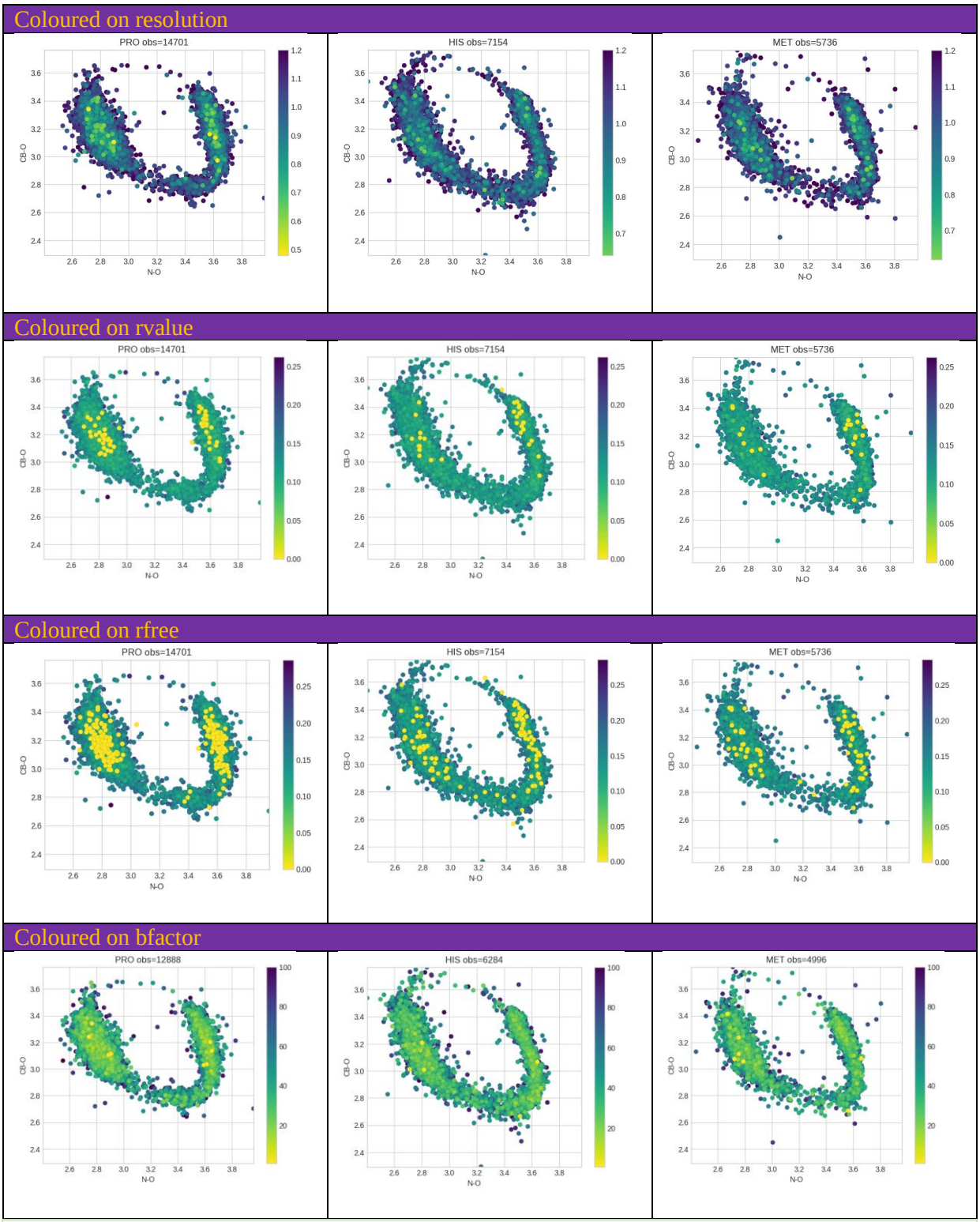
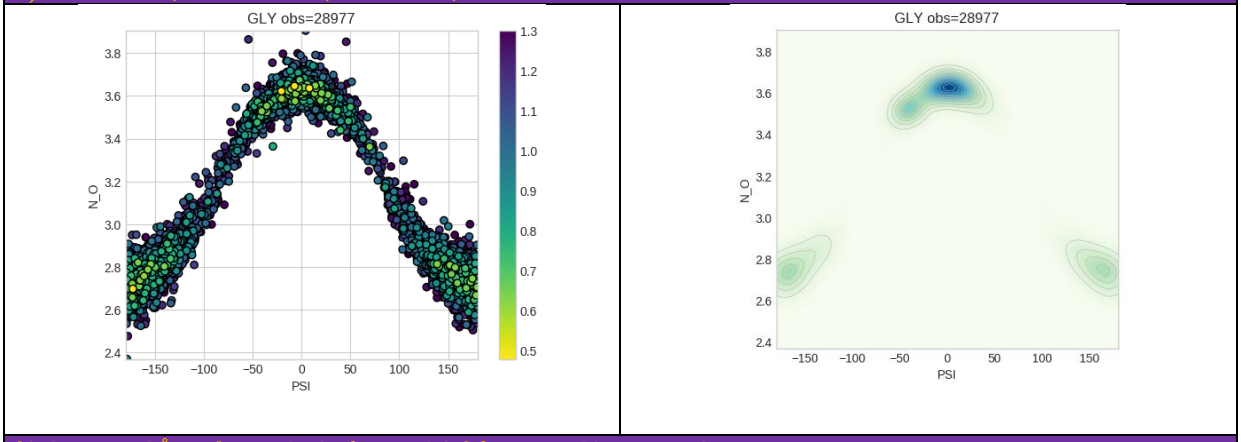


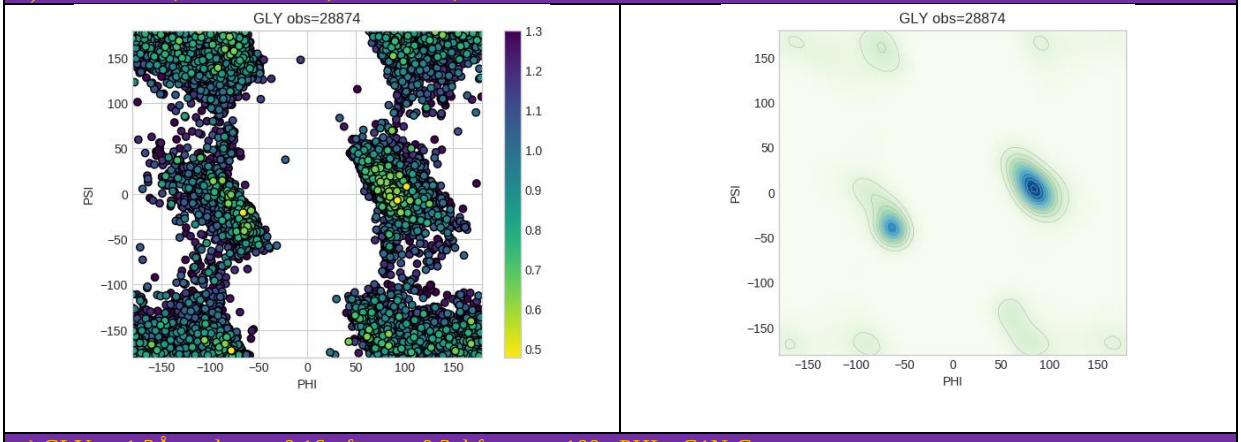
Figure 3.4.2.1 Distributions for PRO, HIS and MET graduated on resolution, rvalue, rfree and bfactor HIGH set resolution <=1.2Å, bfactor distribution with additional restriction on bfactor<=100Å^2

There is also a correspondence between probability density and resolution, to such a strong degree that the resolutions almost seem to directly map to the probability density contours. See Figure 3.4.2.2 comparing a scatter plot against a probability density where the resolution is the scatter colour gradient.

a) GLY <=1.3Å, rvalue <= 0.16, rfree <= 0.3, bfactor <=100 : PSI v N-O



b) GLY <=1.3Å, rvalue <= 0.16, rfree <= 0.3, bfactor <=100 : PHI v PSI



a) GLY <=1.3Å, rvalue <= 0.16, rfree <= 0.3, bfactor <=100 : PHI v C1N-C

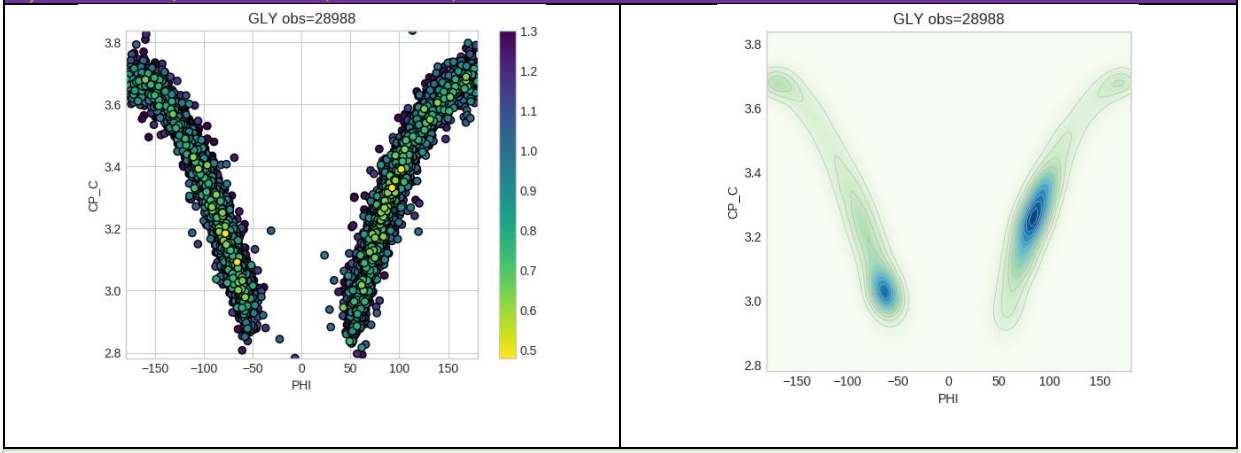


Figure 3.4.2.2 Resolution shows contours of probability density in scatter plots
Probability density uses 12 contours and a gaussian kde from scipy.stats with bandwidth of 0.10

It would be enticing to believe this means that higher resolutions show the correct geometry, but there is no evidence for that. There are fewer structures at lower resolutions, so this demonstrates only that

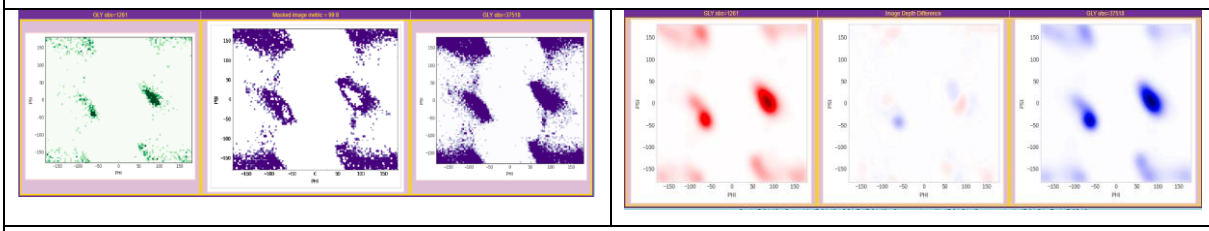
samples are more likely to be found where they are most probable, and thus with a smaller sample they will appear closer to the more probable areas – the rarity effect.

To check this, I used the depth and breadth compare facility on PSU-View to examine the Ramachandran Plot for glycine at resolution <=1.3Å in three circumstances:

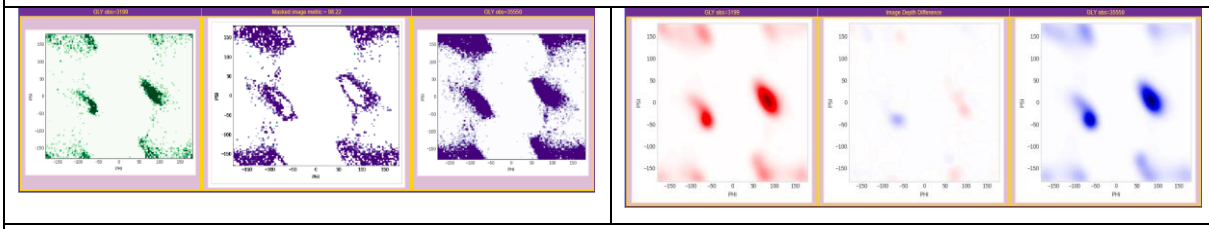
- At resolution <= 0.9Å versus >0.9Å and <= 1.3Å
- For max bfactor of 30, versus for bfactor between 30 and 100
- Any random sample, which I chose to be structures with second letter 'A' versus structures where second letter is not 'A' (recent structures are assigned in sequential order of remaining pdb codes, the first character is numeric and associated with deposition date).

The results can be seen in Figure 3.4.2.3.

a) Left hand image is structures <= 0.9Å versus between 0.9 and 1.3Å



b) Structures with a max bfactor of 30 versus those between 30 and 100



c) Left hand: structures with second pdb letter = 'A' versus those that are not 'A'

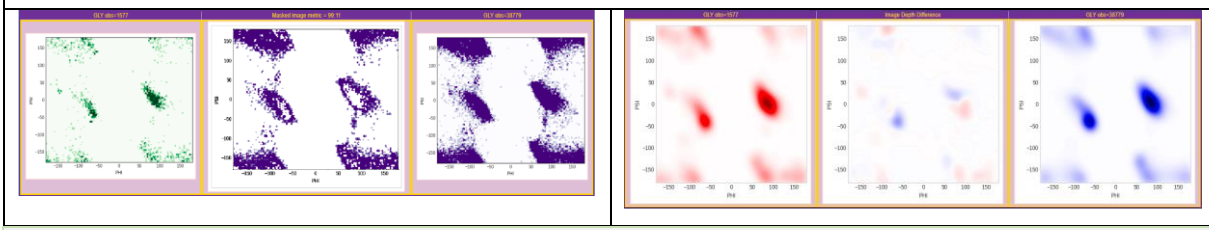


Figure 3.4.2.3 The rarity effect: compares distributions for resolution; bfactor; second letter of the name. The left hand trio show a simplified 2d histograms with the image difference in the middle where if both have significant density it is blank, pale grey for both having different amounts, and the colour of the distribution if one has significant density but not the other none.
The right hand trio contains normalised probability density plots for both with a scipy.stats gaussian implementation with bandwidth of 0.10 and 12 contours. The middle difference images are roughly the same for all 3. These images show that the smaller distribution, whether it is smaller due to being high resolution or a random selection, has a distribution closer to the most probable areas.

The first trio of images shows the smaller distribution on the left, the larger distribution on the right, and the centre images shows the overlap as white and the areas in only one distribution in that colour. The masked image metric, see method section 2.2.1, shows the proportion of each image fully covered by the other, so for the left hand image it is nearly 100% in all

three, for the right hand between 8 and 22% which seems related to the number of residues in the smaller distribution. In all cases the smaller distributions track the more probable regions, even for structures with a second letter “A”. The second trio of images shows the difference in probability density, normalised to remove the discrepancy of distribution size. There is no evident difference between the probability densities of the distributions in any of the pairs, nor in the difference images between the three.

It is important to be aware of this rarity effect when considering the effects of resolution on geometric data. It would be a mistake to draw incorrect conclusions from the data based on any aspect that reflected rarity effect rather than a true difference.

3.4.3 Refinement process

Differing refinement methods may lead to a bias towards different geometric features from parameters and method (Wilson et al, 1998). All residues in HIGH were examined on the correlations page, with refinement software encoded by the hue. For all results see Appendix 12, with Figure 3.4.3.1 below showing 3 of the plots which appear to show that some of the refinement methods have broader distributions of values than others.

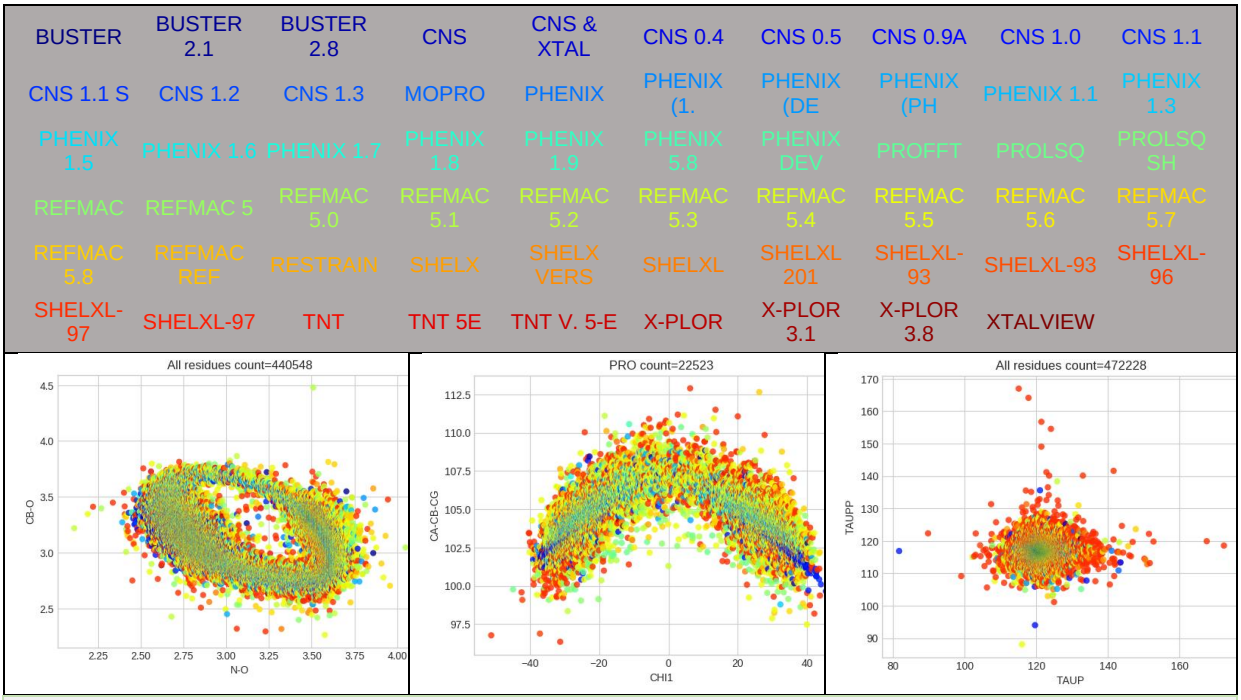
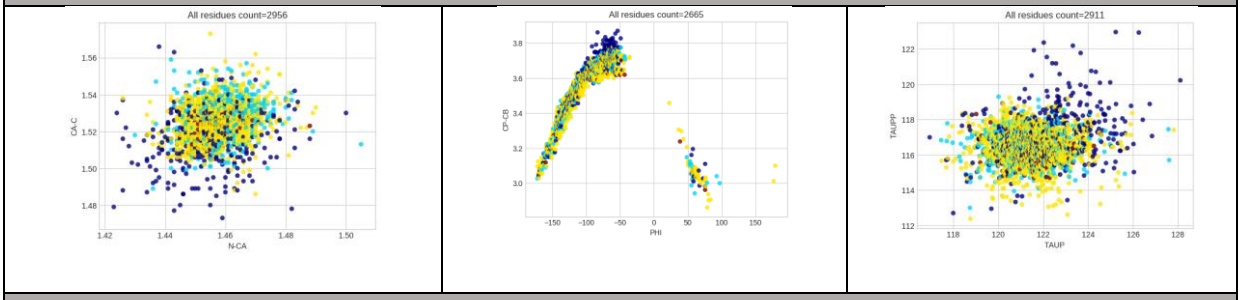


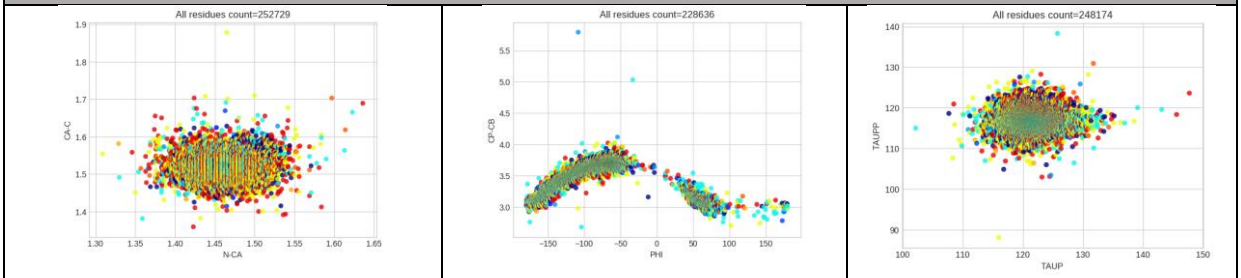
Figure 3.4.3.1 Geometry apparently influenced by refinement software

Five further examples are shown in Figure 3.4.3.2 below for a narrowed down selection of refinement software- versions of XPLOR, CNS, SHELX, PHENIX and REFMAC. There are clear differences in the spread of distributions: XPLOR (not versioned) shows clear areas in dark blue that are not shared by the other XLPOR versions; SHELX-97 has a more relaxed TAU restriction than other SHELX versions; REFMAC 5.1/5.2 has extreme CP-CB values compared to other REFMAC versions. There are substantially more observations for REFMAC than XPLOR but extending beyond 130º tau is common whereas 128º is REFMAC's limit. The extremes are also stretched by REFMAC for N-CA where 1.5Å is only just off centre but for REFMAC there are only 2 observations >1.49Å. Considering the N-CA mean value recommendation in this study of 1.455Å, only PHENIX and SHELX versions centre on this value, with the other refinement software biased to larger values. These different distributions based on refinement software will cause bias in the geometric values measured from refined structures, adding to the appeal of analysing geometric features directly from the experimental evidence.

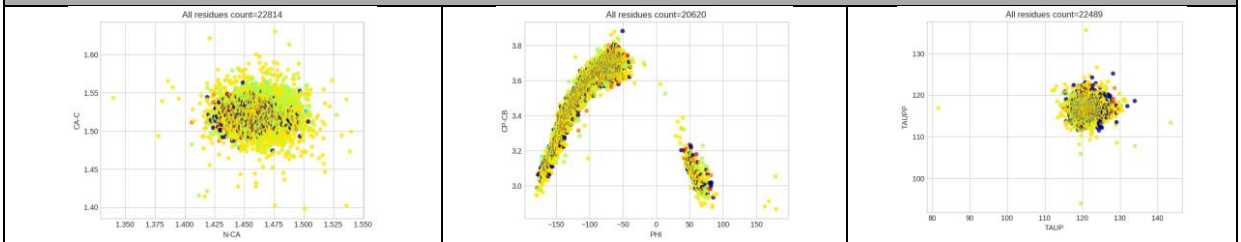
X-PLOR X-PLOR 3.1 X-PLOR 3.8 XTALVIEW



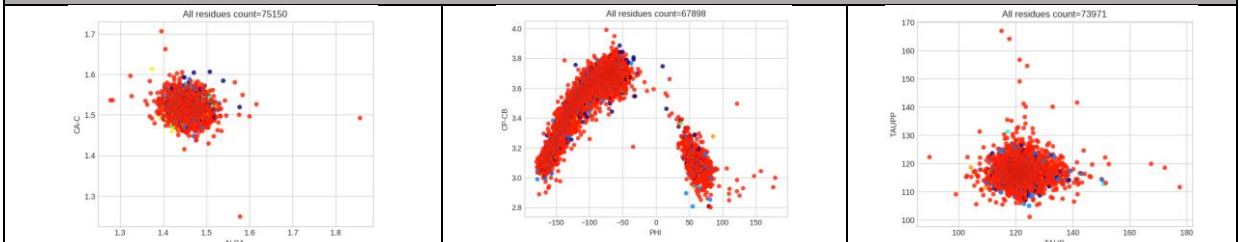
REFMAC REFMAC 5 REFMAC 5.0 REFMAC 5.1 REFMAC 5.2 REFMAC 5.3 REFMAC 5.4 REFMAC 5.5 REFMAC 5.6 REFMAC 5.7 REFMAC 5.8 REFMAC REF



CNS CNS & XTAL CNS 0.4 CNS 0.5 CNS 0.9A CNS 1.0 CNS 1.1 CNS 1.1 S CNS 1.2 CNS 1.3



SHELX SHELX VERS SHELXL SHELXL 20 SHELXL-93 SHELXL-93 SHELXL-96 SHELXL-97 SHELXL-97



PHENIX PHENIX (1. PHENIX (DE PHENIX (PH PHENIX 1.1 PHENIX 1.3 PHENIX 1.5 PHENIX 1.6 PHENIX 1.7 PHENIX 1.8 PHENIX 1.9 PHENIX 5.8 PHENIX DEV

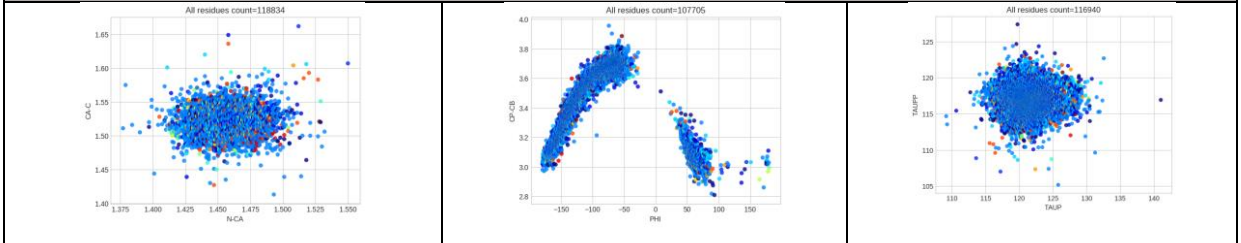


Figure 3.4.3.2 Refinement software and the influence on geometry
These plots demonstrate a difference in the geometry of structures refined with different software, notably N-CA

3.4.4 Energy

The geometric correlations we have seen could just be considered as inevitable consequences of the geometry of the structures - the movement of the atoms are constrained by forces of attraction and repulsion. But - it also demonstrates this very fact: PSU-View contains thousands of observations of protein atomic position deposited over many decades in many locations by many people using different equipment. And yet, the absolute geometry of these structures is preserved – a demonstration of the veracity of the model of interatomic forces.

The geometric plots also show two additional energetic features: energetically favoured locations and potential energy barriers in transitions - Figure 3.4.4.1.

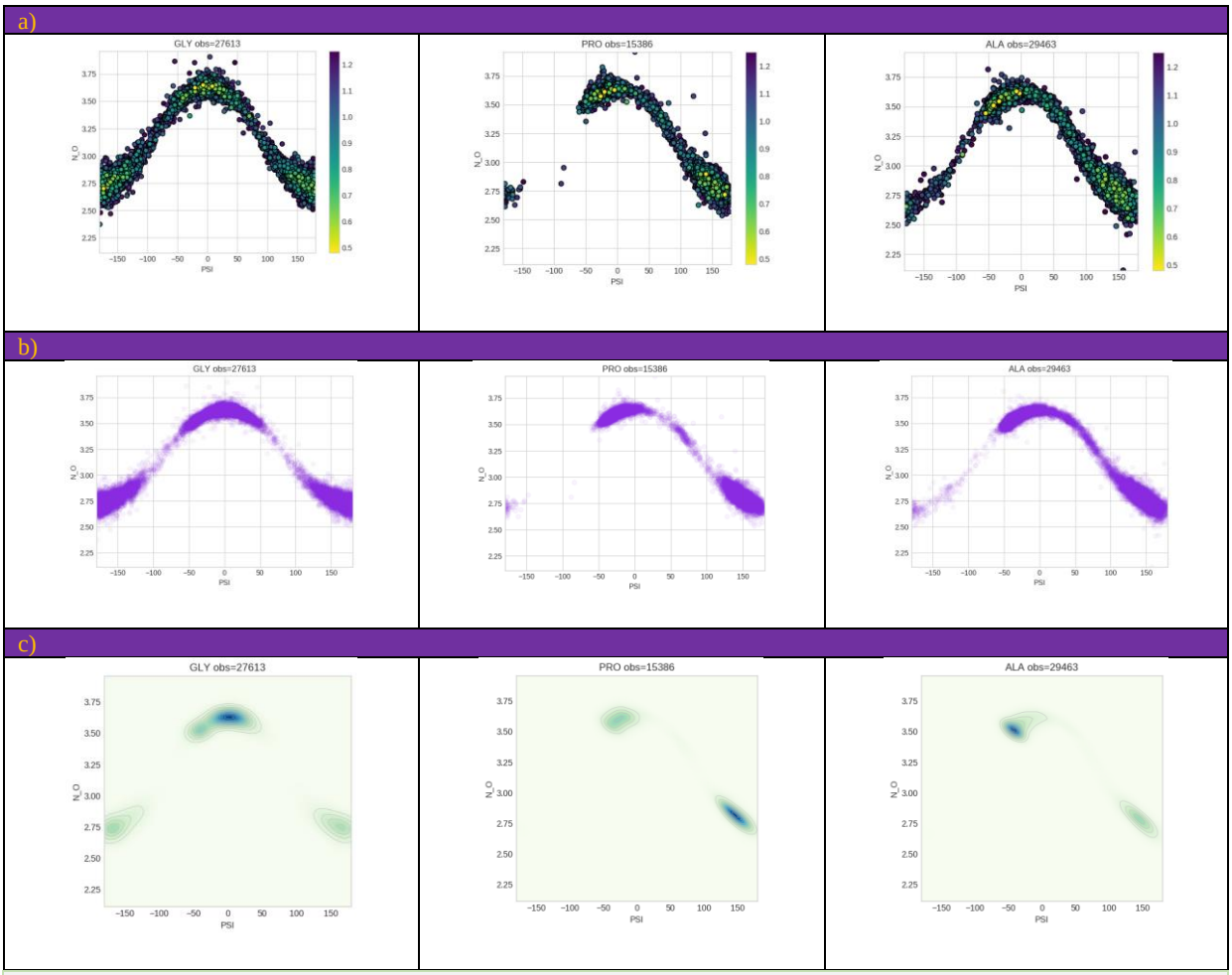


Figure 3.4.4.1 Scatter, density trace and probability density for PSI/N-O, resolution <= 1.25Å
The scatter plot is graduated on resolution; the probability density uses scipy.stats gaussian kernel with bandwidth 0.10; the density trace is a mono-colour scatter plot with an opacity of 0.05 to contrast the likely regions and the impossible regions.
a) *Scatter plot, coloured on resolution, the width of the areas indicates the energetically favoured regions, with the extend of the width giving an indication of bond strength at that location*
b) Density trace, the faintness of the lines indicates the energy barrier in transition
c) Probability density, the energetically favoured regions

These are graphical illustrations of these ideas and are not measurable or quantifiable from these correlation plots.

3.4.5 Cis and trans peptide bonds

The peptide bond is planar and can be found mostly in two conformations at 0º and 180º. The trans formation is the more common due to steric hindrance of the side chains (Birkbeck PPS Course).

There has been discussion about higher resolution structures showing more cis residues (Morris et al, 1992), and most cis residues are cis-proline: XXX-PRO with the bond cis. The identification of cis/trans at different resolutions has been undertaken.

Figure 3.4.5.1 illustrates the two conformations, showing distances that can be used to characterise the two states.

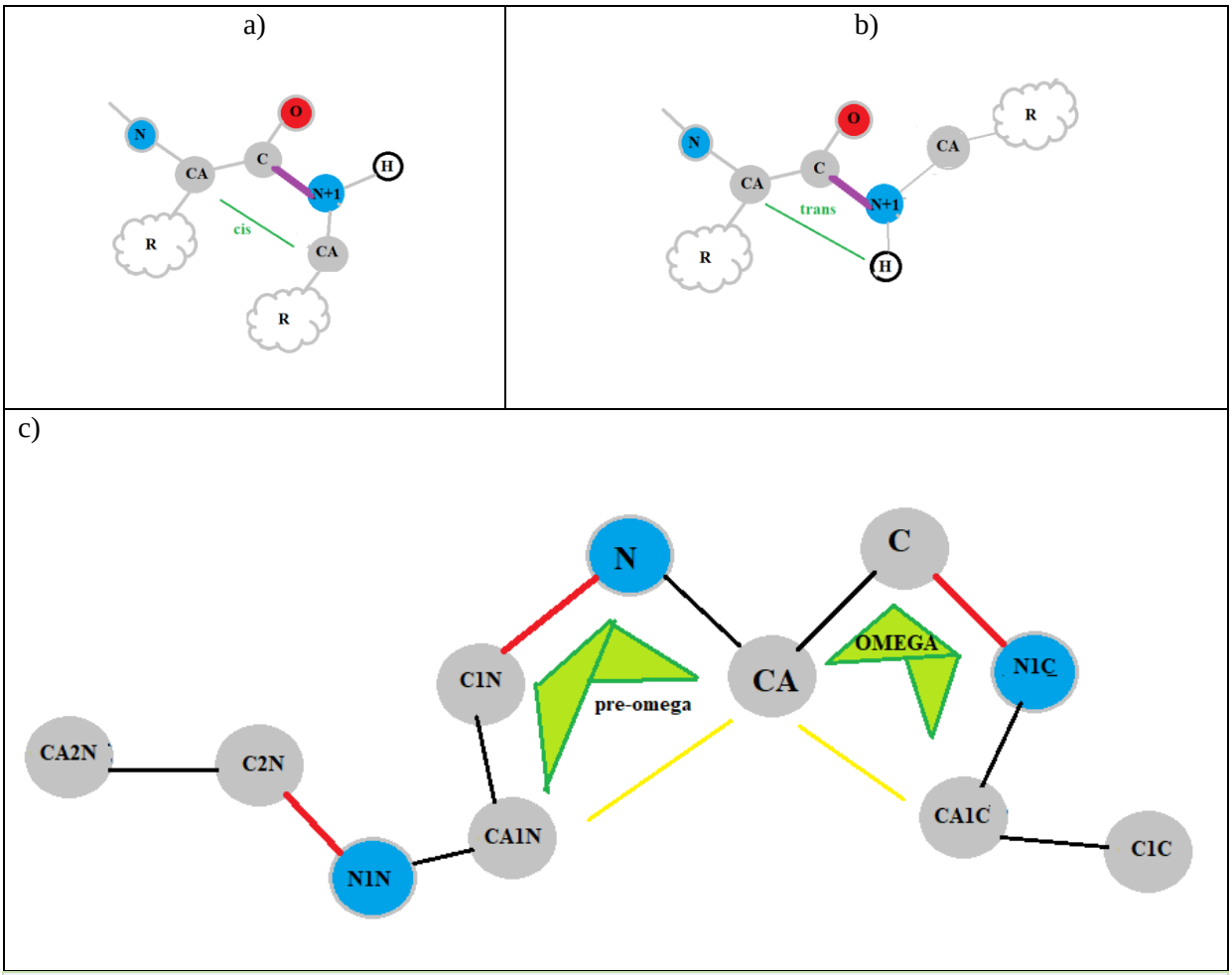


Figure 3.4.5.1 Cis and trans peptide bonds, with OMEGA shown in cis formation
a) Shows the peptide bond C-N+1 in the cis state
b) Shows the peptide bond C-N+1 in the trans state
c) Shows the peptide bonds on either side of the residue and highlights the relationship between OMEGA and the Cα distances

PSU-Beta notates omega as CA-C-N1C-CA1C, and as Figure 3.4.5.1(c) shows, omega measures the peptide bond twist for the following peptide bond of a given residue. This convention is due to the relationship with proline – an omega cis suggests the following residue is proline. The diagram above suggests a relationship with the Cα distance which could provide useful information for both an alter-

native description of the cis/trans switch (potentially useful in low resolution models) and an indication of the nature of the preceding peptide bond nature in this study – although the pre-omega is easy to calculate it has not been calculated in this study. The potential identification of pre-cis will be useful for proline correlations.

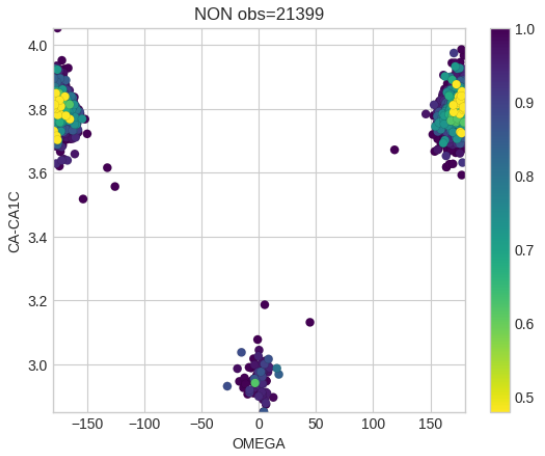


Figure 3.4.5.2 Cis and trans peptide bonds directly correlate with C-alpha distance.
HQ set, resolution <=1.0Å
CA-CA1C <= 3.2 appears to correlate with omega = 0º

There is a correlation between distance CA-CA1C and omega; if CA-CA1C<3.2Å the bond is cis.

We can extend this analysis to the previous peptide bond and say that where CA1N-CA <= 3.2 we have a preceding cis peptide bond. This assumption will be used in considering geometric features in later sections. Subsequent to this analysis, it was found in prior literature (Kleywegt, 1997).

The identification of the Cα distance as a direct identification of the cis-peptide bond does not guarantee that the bond has been identified as such, there is discussion in the literature about more cis-bonds at higher resolutions. The OMEGA/CA-CA1C plot has been analysed to 1.3Å to look for a correlation between resolution and cis bonds - Figure 3.4.5.3, coloured on resolution, rfree and bfactor.

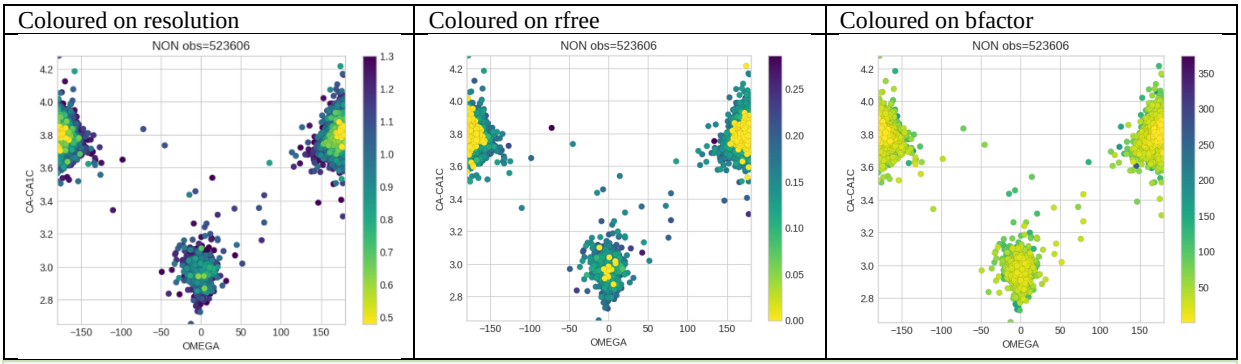


Figure 3.4.5.3 Comparing residues identified as omega-cis and Cα-cis
HIGH set <= 1.0-1.3Å, there are residues omega-cis but not Cα-cis. It is not clear that higher resolution associates with cis, but lower resolution (and rfree) does associate with less ordered regions.

These results suggest that at lower resolutions there are residues that are not cis when measured by the CA-CA1C distance but have been placed (omega=0º) in a cis formation.

3.4.6 Proline

Proline and glycine are special cases in protein structure and geometry due to their unique conformations.

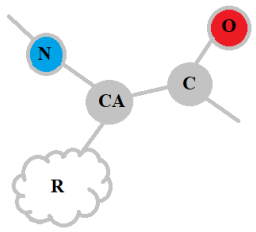
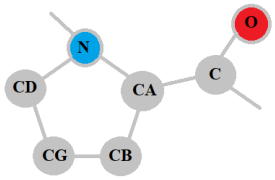
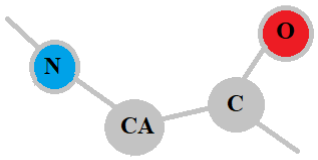
a) General structure of an amino acid with a sidechain R	b) Proline's sidechain uniquely wraps back onto the backbone	c) Glycine has no sidechain
		

Figure 3.4.6.1 Unique conformations of proline and glycine, hydrogens not shown

In Figure 3.4.6.1, the sidechain of proline is shown to wrap onto the backbone, while glycine has no sidechain, removing the hydrogen bonding potential of sidechains from these residues. These properties effect the role proline and glycine play in structural features. Proline is examined further.

In Figure 3.4.6.2 the correlation between CA-CA1C and CA1N-CA is shown as a scatter plot with the residues coloured by amino acid, where orange is proline. As noted in the previous section, the Cα distances can stand in as a proxy for omega, or the cis/trans nature of the peptide bond. In effect the figure below correlates the peptide bonds on either side of a residue for cis- and trans- formations.

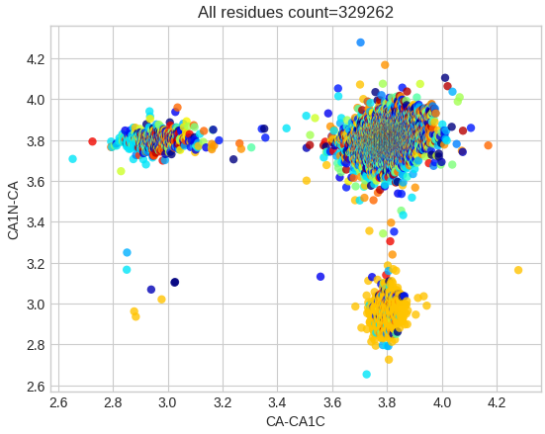


Figure 3.4.6.2 Cis/trans regions as CA-CA1C/CA1N-CA, orange is proline
Resolution <= 1.25Å, max bfactor 100, rfree <= 0.3Å

The most probable values for the CA-CA distance between residues is ≈3.8Å – trans/trans. But there is evidently another area of the scatter plot that is dominated by proline that would correlate to cis/trans on either side of the residue. Figure 3.4.6.3 narrows down the correlation plots on proline only, with regions > 3.6Å and <3.15Å analysed separately on secondary structure, showing an association of secondary structure with the cis- trans- peptide bonds.

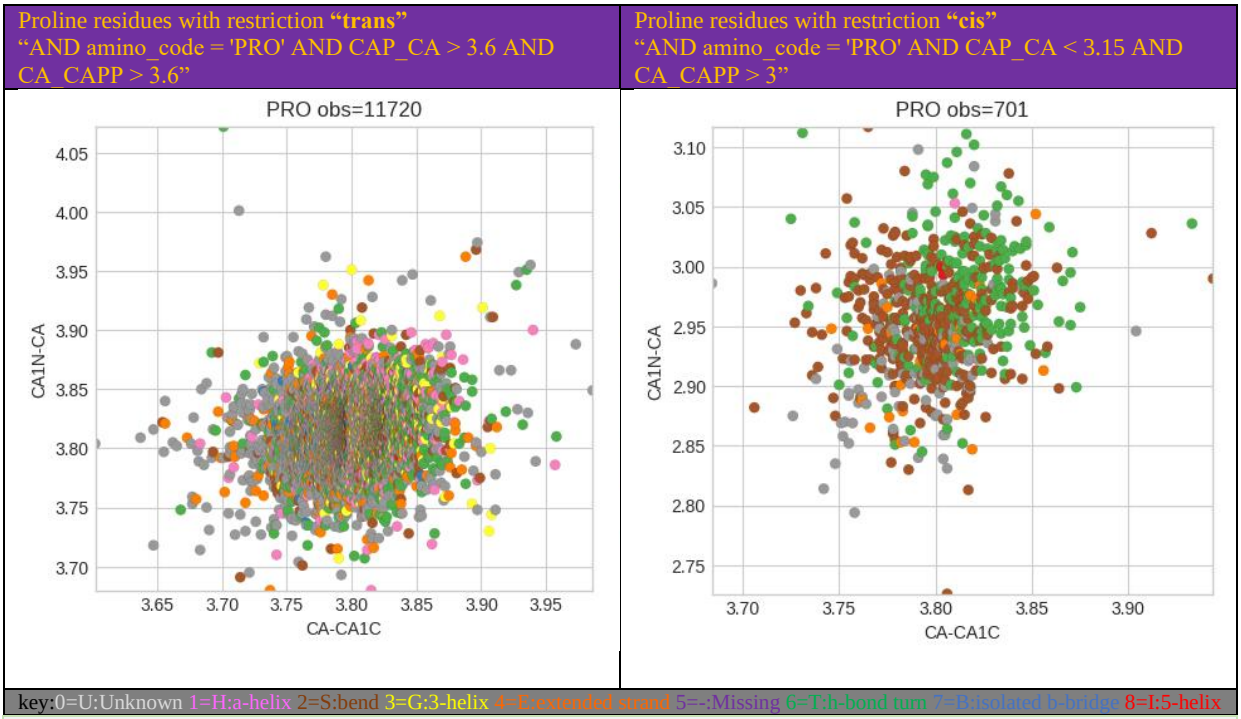


Figure 3.4.6.3 Cis/trans correlations comparisons for proline
The left-hand figure shows trans-proline, the right-hand shows cis-proline dominated by s-bend and h-bonded turn.

The cis-peptide bonds preceding proline are found largely in bends and hydrogen-bonded turns and rarely in any other type of secondary structure. The trans/trans area is dominated by unknown (dssp did not make an assignment), blurring the secondary assignment distinctions.

The cis-peptide bond associates with other geometric features. See Figure 3.4.6.4 for correlations that show effected geometry from this region.

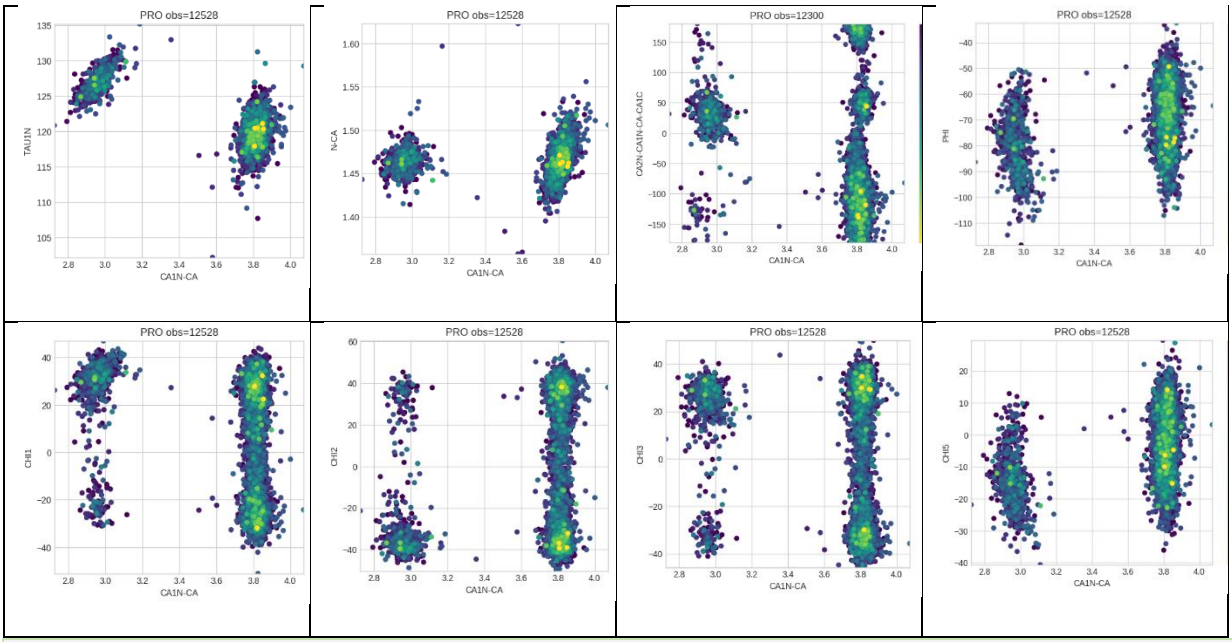


Figure 3.4.6.4 Shows proline cis/trans effect on other geometry, coloured on resolution.
HIGH set, resolution <=1.2Å, rvalue<=0.3Å
In each plot, the cis region is on the left and the trans region on the right.

The cis regions have fewer observations and seem to show some clear differences in the correlated geometric value. TAU1N, the backbone angle C1N-N-CA, shows the clearest bimodal region with the cis/trans formation. The cis- formation for CHI angles 1-3 shows a preference for CHI value: cis is CHI1 positive; CHI2 negative; CHI3 positive.

It has been suggested that the proline ring takes two puckering conformations (Wu, 2013). The five dihedral CHI angles for proline can describe the ring conformation as they successively rotate planar pairs of the ring. Figure 3.4.6.5 depicts the 5 CHI angles of proline.

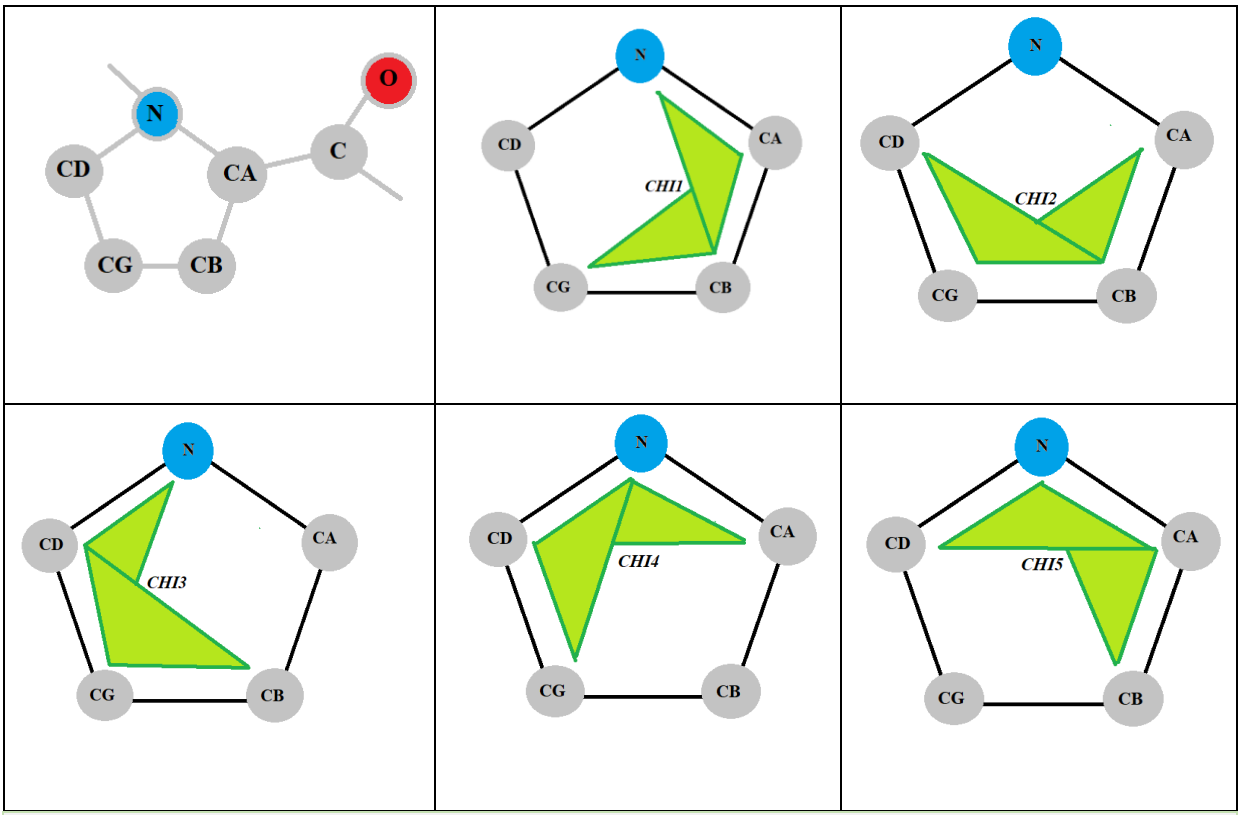


Figure 3.4.6.5 Five CHI dihedrals of the proline ring

To investigate the possible conformations the probability density plot of each of the CHI angles were correlated against each other to create Figure 3.4.6.6. The existence of two probable regions in every plot strongly suggests two conformations of the proline ring but does not guarantee it.

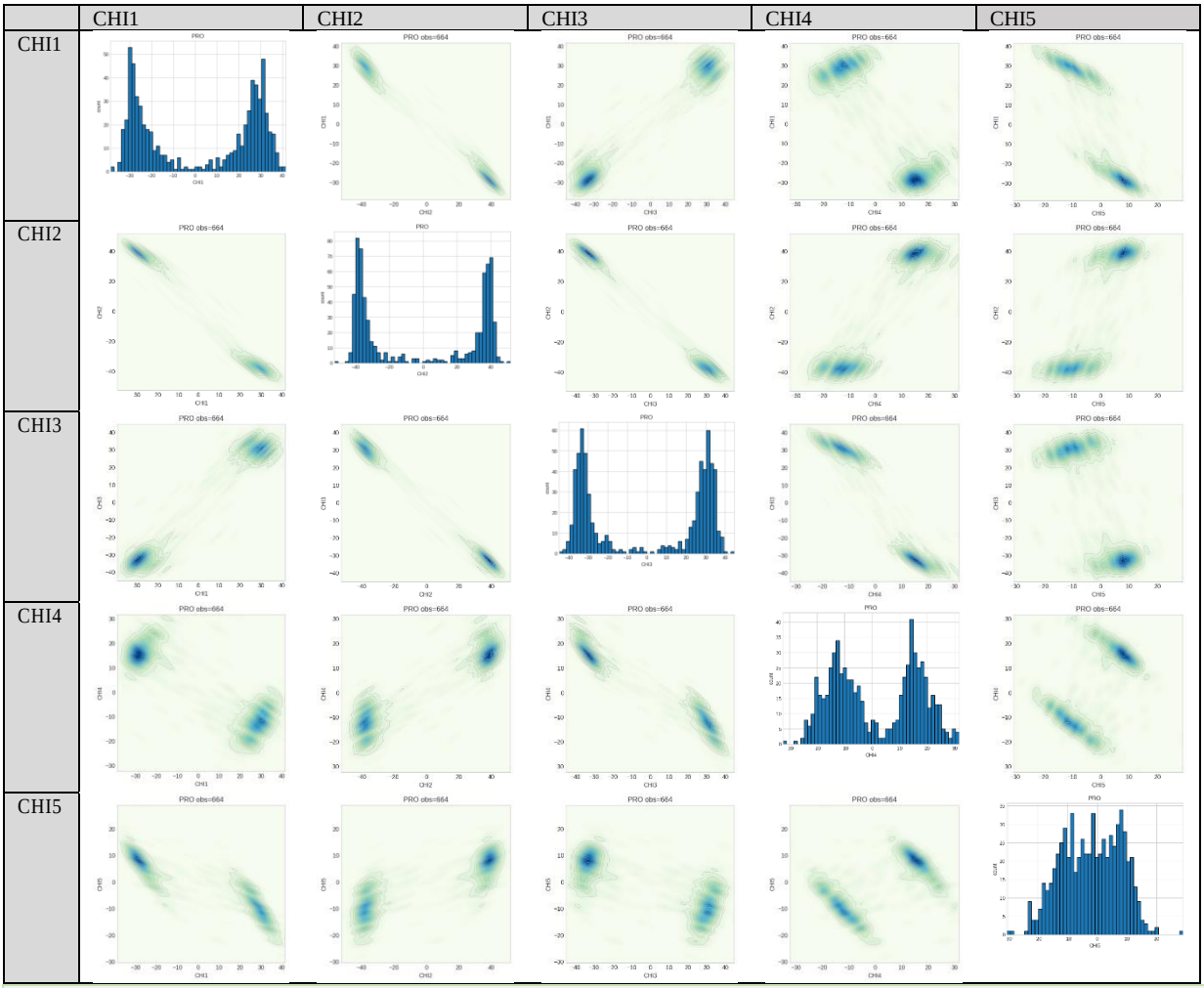


Figure 3.4.6.6 Proline CHI correlations in all combinations shown as probability density
Histogram of each CHI shown for diagonal axis
Residues taken as resolution <=0.9Å, max bfactor 100. Kde is 0.10 with 12 contours.
2 regions in each plot suggest 2 conformations, but do not guarantee it

There was further investigation using PCA analysis. This analysis was performed with three sets of data: all high-resolution proline residues; those with a resolution of <=0.9Å; those with a resolution between 0.9 and 1.2Å. See Figure 3.4.6.7 for the components and clusters for each resolution bucket. The sql queries, data and R markdown can be found on GitHub here: Proline PCA analysis.

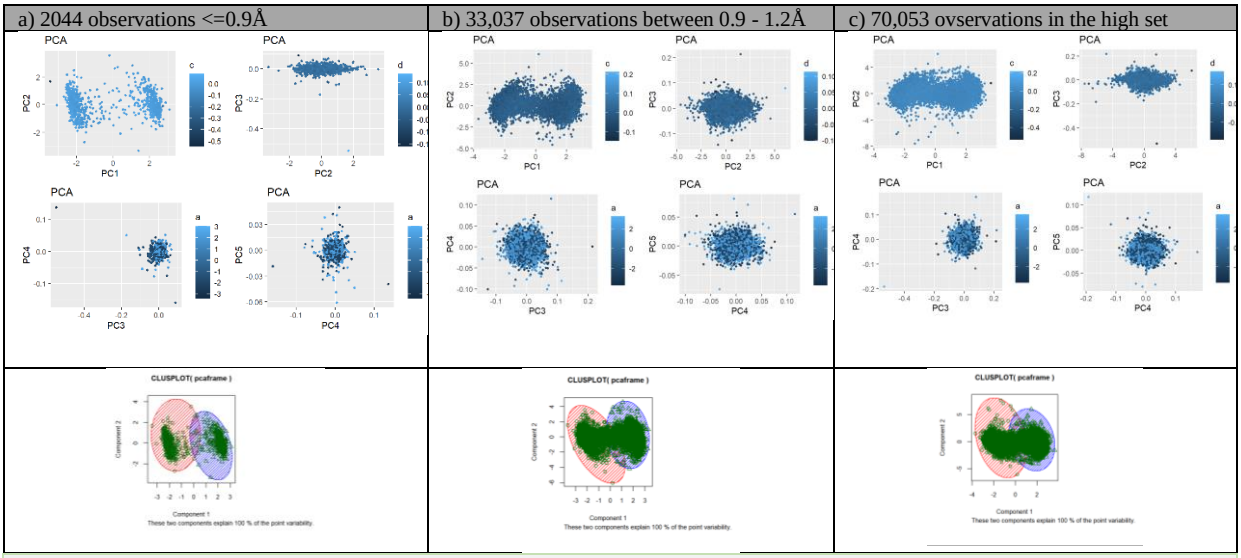


Figure 3.4.6.7: Proline ring conformations' PCA analysis on CHI1-5
The first row compares the 4 components, the last row shows that 2 clusters explain 100% of variability

In Figure 3.4.6.7 there are clearly two clusters for the higher resolution samples - as the resolution lowers the data is still explained by two clusters, but the evidence is less clear. The two clusters indicate two conformations for the proline ring.

Eight representative samples have been investigated from the database, four from the PCA 1 cluster and from 4 groups from the PCA3 cluster (-ve, 0, +ve). See Table 3.4.6.8.

Row	PDB	Amino	PC1	PC2	CHI1	CHI2	CHI3	CHI4	CHI5
1	1cbn	A5	-2.29	-0.29	29.997	-40.012	33.783	-16.415	-8.420
16	1dy5	A114	-2.36	-0.34	30.728	-41.901	34.997	-17.123	-8.243
1369	3x2m	A66	-2.16	-0.06	28.942	-37.711	31.238	-13.202	-9.617
1952	6eio	A91	-2.22	-0.00	29.878	-38.826	31.883	-13.040	-10.311
14	1dy5	A42	2.28	0.02	-27.620	37.449	-34.069	16.852	5.995
52	1gci	A5	2.26	-0.08	-27.039	37.073	-32.503	16.151	6.945
1334	3ui4	A93	2.20	0.22	-25.110	37.046	-33.799	18.300	4.318
1734	4u9h	L234	2.07	0.87	-21.020	35.254	-35.660	23.369	-1.137

Table 3.4.6.8 Four examples each of the two proline ring conformations, green is down- blue is up-pucker

The table shows that the two conformations seem to be simple inversions of each other, with two examples illustrated in Figure 3.4.6.9 from Chimera.

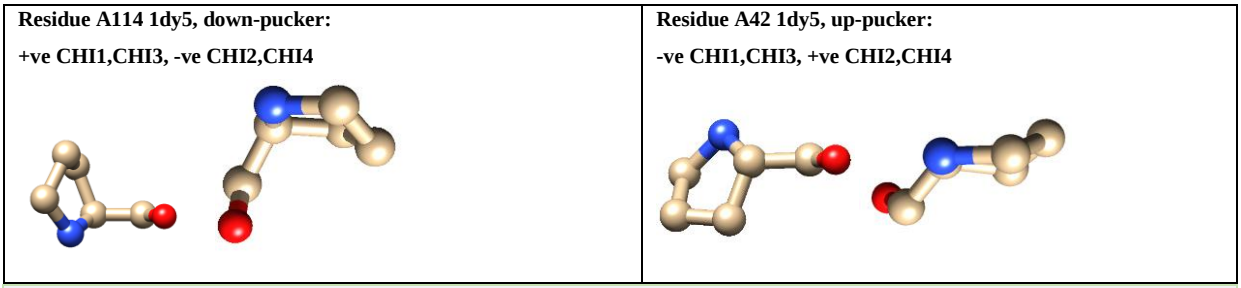


Figure 3.4.6.9 Two proline ring conformations illustrated from 1dy5 (Chimera)

There is a down-pucker and up-pucker of CG, which is achieved by all the angles inverting. There is discussion that the cis-trans- state influences the ring puckering (Vitagliano et al, 2001). Any of the CHI values can stand in for the puckering state, and we have seen already in Figure 3.4.6.4 that the

CHI1, CHI2 and CHI3 angles associate with the cis- trans- formation. Figure 3.4.6.4 suggests that the cis-formation associates with the CHI1 positive values, or down-pucker.

Using CHI1 to stand in for the down-up-puckering state of the proline ring, and the c-alpha distance to stand in for the cis-trans-peptide bond, Figure 3.4.6.10 shows a selection of correlation plots graduated on the left by CHI1 and on the right by CA1N-CA. Ideas for plots omega/phi and C1N-C/phi taken from the literature (Vitagliano et al, 2001).

These correlations show that the up-pucker has a higher PHI, shorter C1N-C bond and lower TAU1N. The cis-peptide has a distinctly higher TAU1N angle and a greater C1N-C bond.

Up-down-pucker state by proxy of CHI1, yellow=up	Cis/trans state by proxy of CAP-CA, yellow = cis

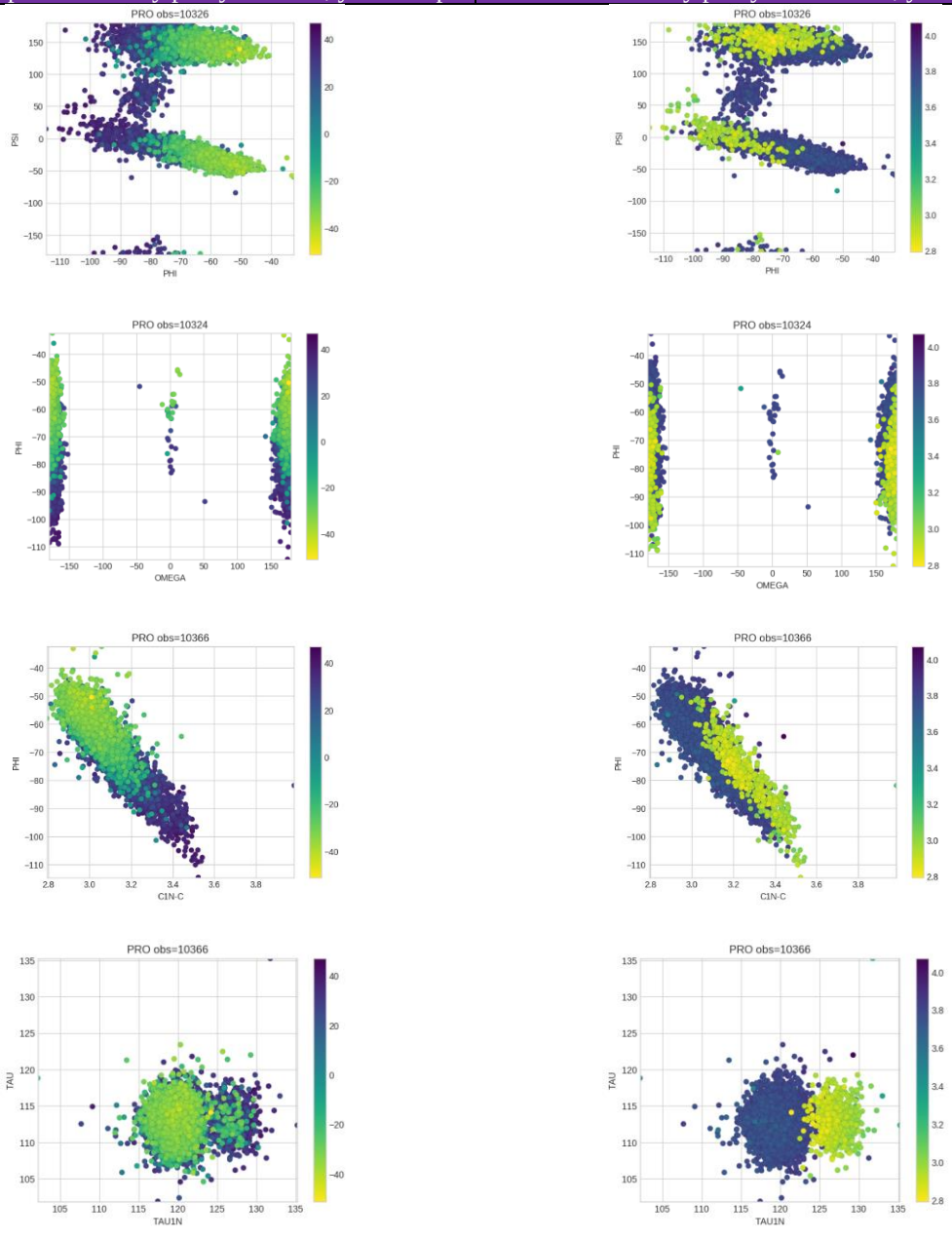


Figure 3.4.6.10 Some correlations showing the up/down pucker and cis/trans peptide bond of proline Resolution <= 1.2Å, rvalue <= 0.16Å, rfree <=0.3Å, bfactor <=100Å^2, checked pdbs excluded

This time using the stand-ins of CHI2 positive as an up-pucker, and low CA1N-CA as a cis-peptide bond, proportions can be analysed at different resolutions, Table 3.4.6.11 shows the observation counts for these states at different resolutions and analyses the conditional probabilities.

a)	Total Count	Cis/Down AND CAP_CA < 3.2 AND CHI2 <0 0	Cis/Up AND CAP_CA < 3.2 AND CHI2 > 0	Trans/Down AND CAP_CA >= 3.2 AND CHI2 < 0	Trans/Up AND CAP_CA >= 3.2 AND CHI2 > 0
1.2<=1.3	10556	521	75	4695	5265
1.1<=1.2	6301	292	56	2731	3222
1.0<=1.1	4878	233	33	2140	2472
0.9<=1.0	2309	132	16	1045	1116
<=0.9	680	41	8	314	317
b)	1.2<=1.3	1.1<=1.2	1.0<=1.1	0.9<=1.0	<=0.9
P(Down)	0.4941	0.4798	0.4865	0.5097	0.5265
P(Up)	0.5059	0.5202	0.5135	0.4903	0.4735
P(Cis)	0.0565	0.0552	0.0545	0.0641	0.0721
P(Trans)	0.9435	0.9448	0.9455	0.9359	0.9279
P(Down|Cis)	0.8742	0.8391	0.8759	0.8919	0.8367
P(Down|Trans)	0.4714	0.4588	0.4640	0.4836	0.5024
P(Up|Cis)	0.1258	0.1609	0.1241	0.1081	0.1633
P(Up|Trans)	0.5286	0.5412	0.5360	0.5164	0.4976
P(Cis|Down)	0.0999	0.0966	0.0982	0.1121	0.1293
P(Cis|Up)	0.0140	0.0171	0.0132	0.0141	0.0255
P(Trans|Down)	0.9001	0.9034	0.9018	0.8879	0.8855
P(Trans|Up)	0.9860	0.9829	0.9868	0.9859	0.9752

Table 3.4.6.11 The ratios of proline's cis/trans peptide bond to up/down pucker states and conditional probabilities
a) *Shows the observations in each state at different resolutions for the HIGH set*
b) Shows the conditional probabilities coloured on blue = high, yellow=medium and green=low

Table 3.4.6.11 shows that if proline is cis it is much more likely to be down: P(Down|Cis) = 0.8742. However, if proline is down there is only around a 0.1 chance it will be cis. If it is up, there is a 0.01 chance it will be cis. The resolution comparisons show an apparent increase in cis as the resolution increases, though there are too few observations at <=0.9Å for confidence in the data when looking at cis proportions.

3.4.7 Calpha-Calpha Modelling

We have seen already in the cis-trans section 3.4.5 that there is a correlation between Cα distances and the cis or trans nature of the peptide bond. Additionally, this study's results include Cα angles and a pseudo-dihedral, detailed below in Figure 3.4.6.1.

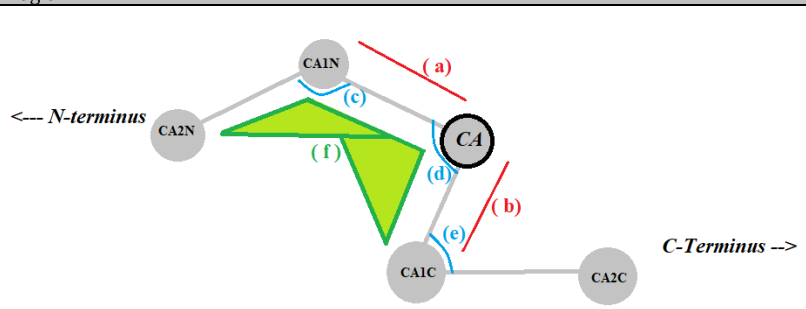
Type	Measure	Diagram
Distance	a) CA1N-CA	
Distance	b) CA-CA1C	
Angle	c) CA2N-CA1N-CA	
Angle	d) CA1N-CA-CA1C	
Angle	e) CA-CA1C-CA2C	
Pseudo-dihedral	f) CA2N-CA1N-CA-CA1C	

Figure 3.4.7.1 Cα measures calculated by PSU-Beta and a simple Cα skeleton

C-Alpha values are useful in validation and c-alpha modelling (Asachi et al 2020; Kleywegt, 1997). The pseudo-dihedral leads to the angle being measured between the planes CA2N-CA1N-CA and CA1N-CA-CA1C. A pseudo c-alpha Ramachandran plot can be created with this dihedral and the angle CA2N-CA1N-CA (Asachi et al, 2020) or CA1N-CA-CA1C and CA-CA1C-CA2C (novel).

These plots are shown in Figure 3.4.7.2 for alanine (to stand in as representative of all amino), glycine and proline. In this plot, the probability density overlays the scatter plot (coloured on secondary structure). The probable areas of glycine and proline clearly differ. The most probable area for alpha helices is at around (angle,dihedral)=(90º,50º) and b-sheets around (angle,dihedral)=(120º,-150º) the a-helix corresponds to the most probable regions in the density plot for alanine.

The preferred regions of glycine and proline are different to the amino acids in general, specifically they show areas of high probability density that are neither the a-helix nor b-sheet regions. For glycine there is a region around (angle,dihedral)=(110º,25º) in the CA2N-CA1N-CA plot, an area around (angle,dihedral)=(90º,10º) in the CA1N-CA-CA1C plot, and (angle,dihedral)=(90º,-110º) in both. Proline favours an area around (angle,dihedral)=(115º,-100º) in the CA2N-CA1N-CA plot, both this area and the same areas as glycine at around (angle,dihedral)=(90º,-110º) in CA1N-CA-CA1C and in the final plot CA-CA1C-CA2C proline favours primarily the area (angle, dihedral) =(90º,-110º).

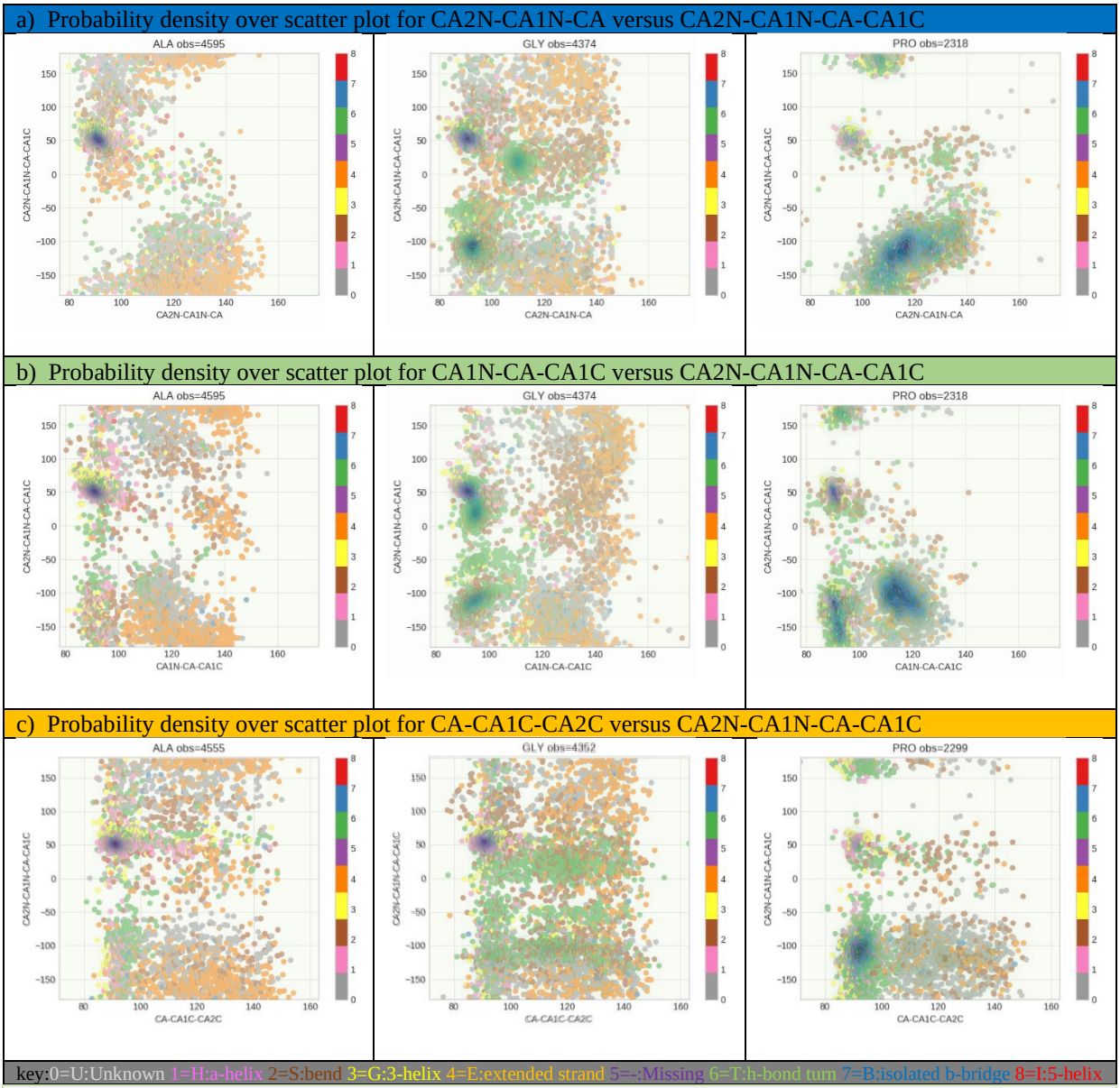


Figure 3.4.7.2 Probability overlays scatter plot for c-alpha pseudo-ramachandran, for ALA, GLY and PRO Taken for Ala, Gly and Pro residues <=1.0Å, rvalue <= 0.16Å, rfree <= 0.3Å, bfactor <=100Å^2
This shows the different areas of probability for proline, glycine and alanine for the pseudo-Ramachandran plot used in Cα modelling, and the secondary structures associated with these areas.

Further analysis of the trio of c-alpha angles yields differences for the relative distributions of glycine and proline. Unlike alanine, angles preceding and following glycine and proline have distinct distributions from the angle centred at the residue itself. See Figure 3.4.7.3 for the results for PRO, GLY and ALA, with Appendix 11 contain the full set of results for each residue type.

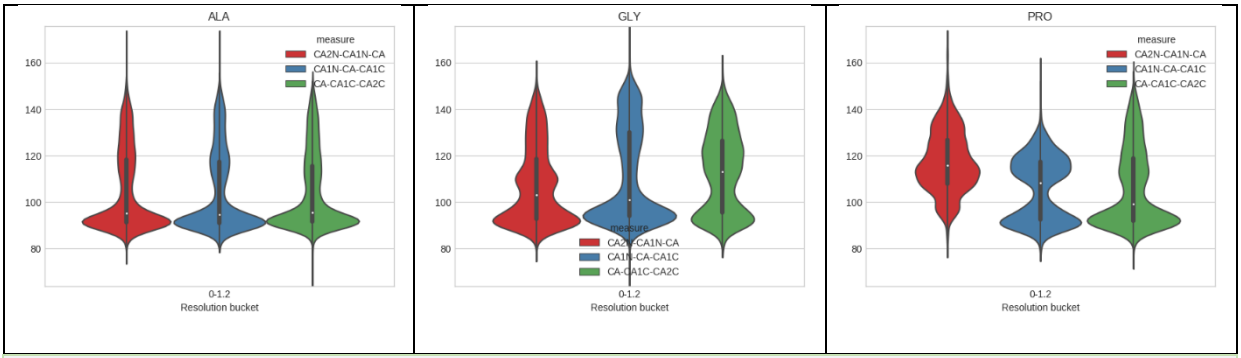


Figure 3.4.7.3 Violin plots for Cα angles along the chain
Taken for Ala, Gly and Pro residues <=1.2Å, rvalue <= 0.16Å, rfree <= 0.3Å, bfactor <=50Å²
The angle distributions are shown to clearly differ for glycine and proline depending on the position the residue in the 3 residue motif. Thus the N- or C- terminus direction of the chain impacts the angles for glycine and proline strongly.

In the case of proline, we know already it favours a cis pre-peptide bond over other amino acids, which could account for this. Looking at these angles for proline, graduated on the Cα distance which stands in as a proxy for whether it is pre-cis, we see the results Figure 3.4.7.4.

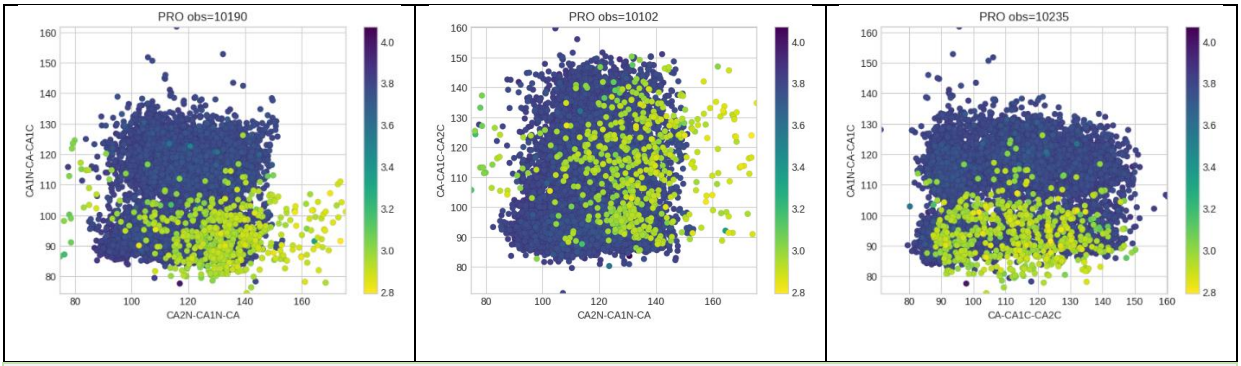


Figure 3.4.7.4 Angles along Cα and backbone for PRO graduated on CAP-CA as a proxy cis/trans.
Yellow = cis, taken for Pro residues <=1.2Å, rvalue <= 0.16Å, rfree <= 0.3Å, bfactor <=100Å²
This shows cis-pro can be found in certain regions of these correlation plots.

Notably the CA2N-CA1N-CA angle has cis values >150º and <80Å that are not seen for trans. The angle is affected by twists on the main chain bond in the 3 residues. The cis-bond also directly associates with a long O1N-CA, an inevitable feature of the cis-peptide bond. See Figure 3.4.7.5 for the relationship between the O1N-CA distance, the peptide bond and the Cα angle, and for 2 possible models that show the angle extremes of < 90º and >160º.

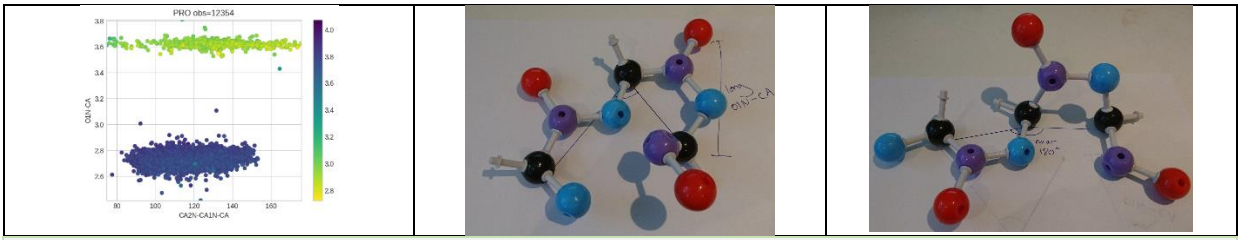


Figure 3.4.7.5 shows the relationship between O1N-CA distance and peptide bond, with possible models
The correlation shows the relationship between O1N-CA, the Cα angle, and cis (yellow)/trans (purple)
The extremes of the Cα angle are demonstrated: <90º and near linear

Glycine does not share the cis/trans feature with proline; glycine has the unique feature among the amino acids that it can rotate 360º around the N-CA bond without steric hindrance. The negative PSI values correspond to this unique feature and the correlation of PSI against each of the three Cα angles elucidates it, see Figure 3.4.7.6. There is a clear geometric correlation between PSI and CA1N-CA-CA1C along with clear secondary structure regions: the results of this correlation for each amino acid individually can be found in Appendix 21.

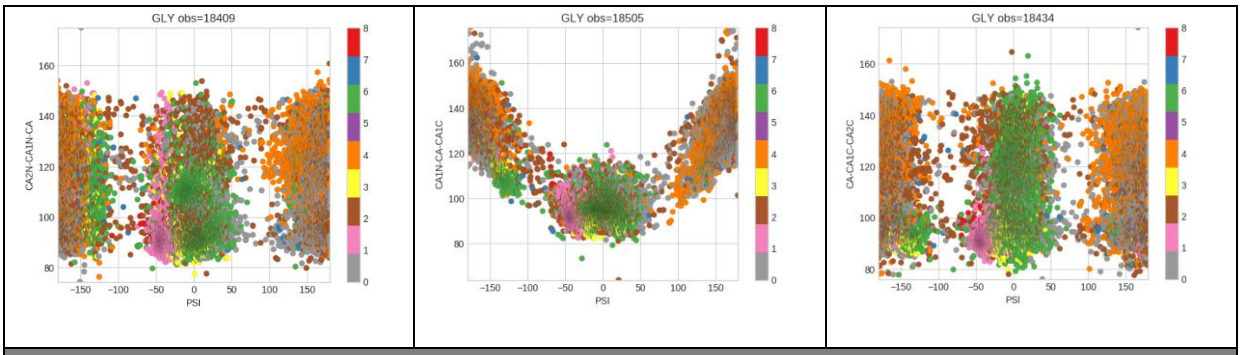


Figure 3.4.7.6 Angles along Cα and correlated against PSI for GLY graduated on secondary structure.
Taken for Gly residues <=1.2Å, rvalue <= 0.16Å, rfree <= 0.3Å, bfactor <=100Å^2
The novel plot PSI/CA1N-CA-CA1C in the centre is clearly geometric with strong secondary structure regions.

3.4.8 CHI1 and resolution

CHI1 distributions were analysed at different resolution buckets.

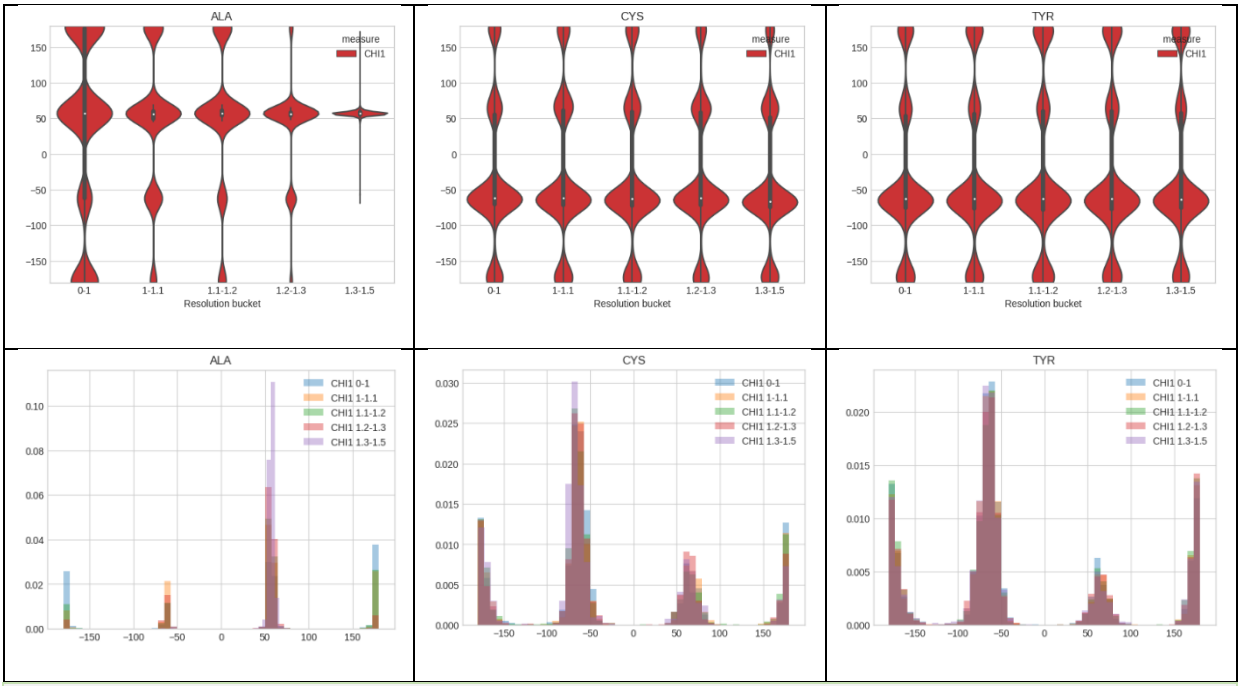


Figure 3.4.8.1 Chi1 compared for 3 amino acids at different resolutions and with 2 different kde settings Top row, kde gaussian, bandwidth 0.1, Bottom row, histogram may have some evidence of wider values at higher resolution (blue bars). There is an unexpected result for alanine due to the definition of CH1 including HB1. The lack of any hydrogen experimental evidence at the lower resolutions means this plot gives a pure view of the transition from forcefield to experimental evidence.

In general, at the higher resolutions the distribution seems less precise, with some residues not changing much. Appendix 15 contains the results for all amino acids, with three chosen in Figure 3.4.8.1. Appendix 16 contains the summary statistics including observation count. As resolution increases there is an anticipated effect of the interplay between the forcefield in refinement used to establish the most energetically stable atom positions and the experimental evidence. The results here are subtle or, depending significantly on experimentally changing kde settings – but what has happened to alanine?

The alanine CHI1 definition is not standard, it is the only CHI1 definition to include hydrogen. As the only hydrogen in this data, what we see for alanine is a single effect. At the lower resolution there is no experimental evidence at all for hydrogen, so the forcefield is used entirely to position HB1, always in the same place.

As resolution improves and there is experimental evidence something occurs to the naming of the hydrogens, since there is no way to choose which hydrogen is HB1, HB2 or HB3, see Figure 3.4.8.2, as all three hydrogens are sterically identical. Note, the definition seems to be incorrect, defined as C-CA-CB-HB1 (it should be N-CA-CB-HB1).

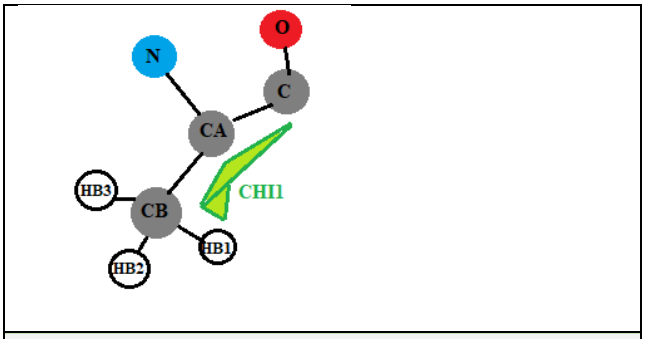
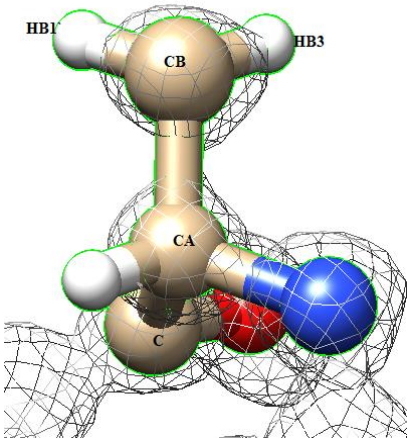


Figure 3.4.8.2 Artificial CHI1 for alanine uses atom HB1

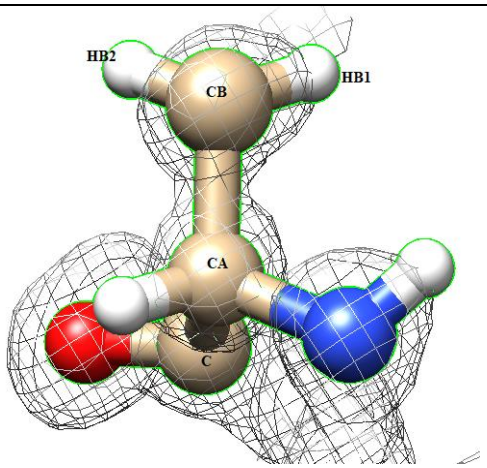
The CHI1 resolution violin plot for alanine (Figure 3.4.8.1) shows that at the highest resolution there are positions chosen for HB1 in all 3 of the hydrogen locations, which leads to the tri-modal violin plot with dihedral angles around 0/180º, 60º and −60°.

A possible explanation for this is that HB1 is chosen to be the atom with the greatest experimental evidence - the hydrogen with the greatest electron density. This would represent a human factor, manual or programmed, impacting the atom positions.

Looking at structure 3X2M, see Figure 3.4.8.3, this does not seem to be the case. For residue 59, with a CHI1 of 57.1º, the HB3 seems to have just slightly more electron density – there is no HB2. In residue 9, with a CHI1 of -64.8º there does seem to be slightly more electron density on the HB1 – this time there is no HB3.



Atom bfactors:
HB1=4.39
HB3=4.25



Atom bfactors:
HB1=4.87
HB2=5.59

Figure 3.4.8.3 Alanine's HB1 naming in structure 3X2M
The empty electron density is the peptide bond – the atoms are hidden. Neither electron density nor bfactors indicates how the hydrogens are named. This shows the difficulty and inconsistency in naming HB1 even within the same structure.

A manual calculation of the lowest 10 resolution and top 10 resolution results from the alanine CHI1 data was performed, results are recorded in Appendix 22. The spreadsheet with the calculations can be found on GitHub: CHI1 calculation.

Due to the low number of hydrogens recorded generally in the structures, hydrogen has not been a feature of investigation in this project. Inadvertently, this CHI1 result provides an insight into hydrogen placement at high resolution that suggests an interesting opportunity for further investigation.

3.4.9 Hydrogen bonds

PSU-BETA does not identify hydrogen bonds but has the facility to analyse geometric features based on close contacts between CB-CB, CA-CA, S-S (for cysteine) and N-O as a (presumed) donor or acceptor. The close contact is taken from a database table containing all contacts <6.1Å. The analysis in 3.4.9.1b below is further restrained to atom pair contacts <3.6Å.

Close contact distributions for N-O were analysed at five resolution buckets, with the results in Figure 3.4.9.1. There is an apparent increase in the relative number found at close contact 2.8Å as the resolution omproves.

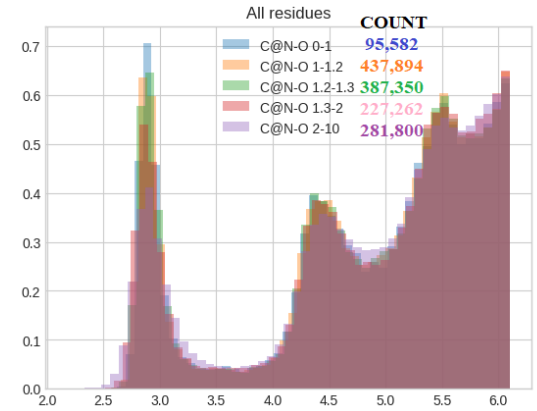
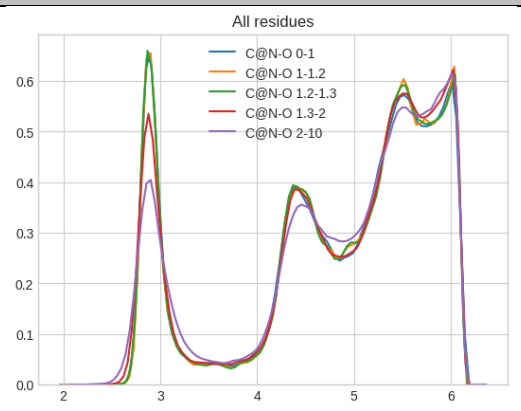
a) Overlayed bar chart of the distribution proportion	b) Line plot of the distributions
	

Figure 3.4.9.1 Close contacts between N and O of atom pairs up to 6.1Å apart, counts given for each resolution
a) As a histogram the high resolution clearly shows at around 2.8Å
b) As a line plot with kde smoothing, using cos kernel and silverman rule of thumb.
Using the HIGH set for <= 1.3Å and the 2019 set at lower resolutions, with rfree at 0.3Å with other values unrestrained.

At a higher resolution, the structures are solved with a greater accuracy to atom placement and seem to have a greater proportion of close contacts at around 2.8Å, suggesting that at higher resolutions there may be a clearer view of hydrogen bonding, and that it may be greater than suggested at lower resolutions.

Further analysis was performed of each amino acid and the fulfilment of hydrogen bonding potential as a donor, or acceptor, or not, see Table 3.5.2 below. The analysis was performed on a data set for resolution <=1.1Å looking at close contact with another atom at <3.6Å.

Amino acid	Total	Donor	Not donor	%donor	Acceptor	Not acceptor	%acceptor
ALA	11,975	9,645	2,330	81	8,867	3,108	74
CYS	1,995	1,569	426	79	1,457	538	73
ASP	8,121	5,901	2,220	73	5,722	2,399	71
GLU	7,210	5,531	1,679	77	5,069	2,141	70
PHE	4,990	4,069	921	82	3,695	1,295	74
GLY	11,177	8,685	2,492	78	6,644	4,533	59
HIS	3,052	2,322	730	76	2,115	937	69
ILE	6,986	5,856	1,130	84	5,428	1,558	78
LYS	7,291	5,634	1,657	77	5,041	2,250	69
LEU	10,568	9,017	1,551	85	7,857	2,711	74
MET	2,293	1,980	313	86	1,735	558	76
ASN	6,325	4,708	1,617	74	4,238	2,087	67
PRO	6,270	2,088	4,182	33	3,698	2,572	59
GLN	4,804	3,865	939	81	3,449	1,355	72
ARG	5,681	4,483	1,198	79	3,985	1,696	70
SER	8,088	5,689	2,399	70	5,462	2,626	68
THR	8,399	5,955	2,444	71	5,811	2,588	69
VAL	9,455	7,701	1,754	81	6,997	2,458	74
TRP	2,095	1,684	411	80	1,534	561	73
TYR	4,768	3,821	947	80	3,543	1,225	74

Table 3.4.9.2 Donors and Acceptors for amino acids at resolution <= 1.1Å, rvalue <= 0.16, rfree <= 0.3
The total is all residues – all the candidates for hydrogen bonding
Donor - all residues' N <=3.6Å toanother O, Acceptor – a;; residues O found <=3.6Å to another N

The simplified nature of this analysis of close contacts between atom pairs <3.6Å with more than one residue between does not give a real view of hydrogen bonding, nor does it show any indication of hydrogen bonding fulfilled in solution or complex. It does give an indication of the propensity for close contact that could lead to hydrogen bonding, with the expected result that proline does not (cannot) hydrogen bond with nitrogen as a donor. Two scenarios of proline being in close contact are given in Figure 3.4.9.3.

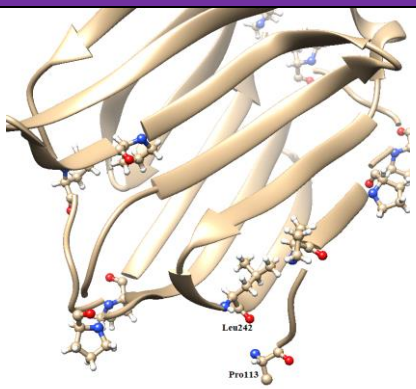
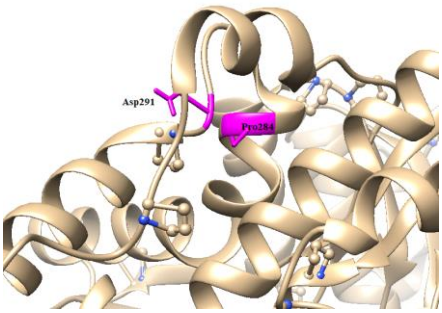
6qln pro113 is the start of the chain, in close contact with leu242 in a beta sheet, distance=2.897Å	3vif pro284 in a helix is in close contact with asp291 in a loop, distance=3.034Å
	

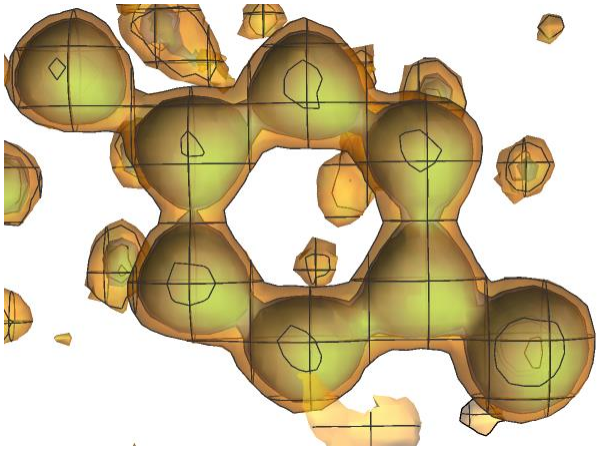
Figure 3.4.9.3 Proline – 2 examples when it is in close contact with other residues

3.5 Results for electron density

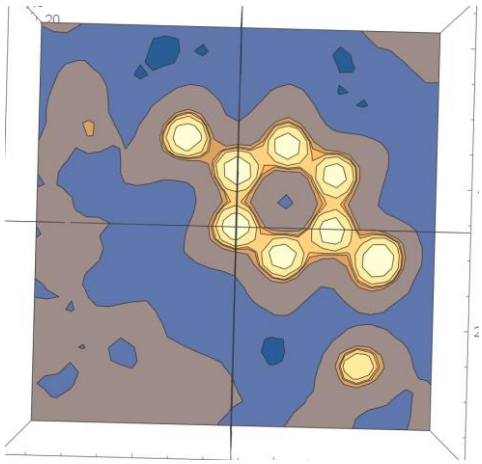
To examine the concept of electron density superposition, investigations were designed that might successfully show interesting information.: a planar set of atoms, thus the tyrosine ring was chosen; an exploration of the peptide bond, the idea taken from Jelsch (2000); a large number of samples across multiple structures under specific geometric constraints – GLN-GLN hydrogen bonds were chosen (Escobedo et al, 2019). Superposition residues were chosen with specified criteria using the PSU-Beta database.

The results for the tyr ring are in Figure 3.5.1, the structure 1us0 was chosen (Howard et al, 2004, an ultrahigh resolution structure of 0.66Å), the results consist of all 11 tyr tings superposed

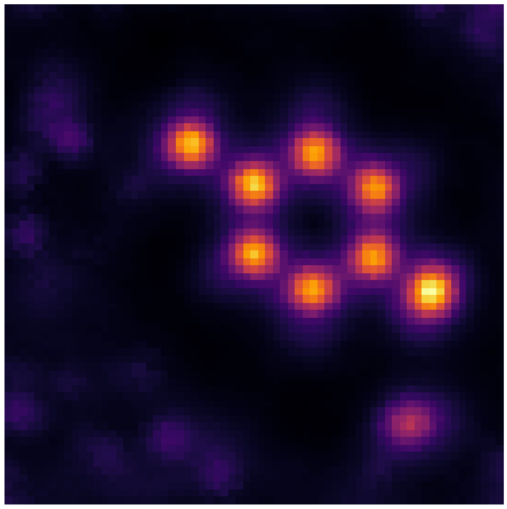
a) A 3d view of the superposition structure with insight into bonds. Image produced in Mathematica.



b) A central cross section showing the 3 atoms used in the superposition method. Image produced in Mathematica.



c) Electron density superposition visualised with matplotlib. Image produced from PSU-ED using matplotlib.



d) Simplified depiction of the tyrosine ring in the same orientation.

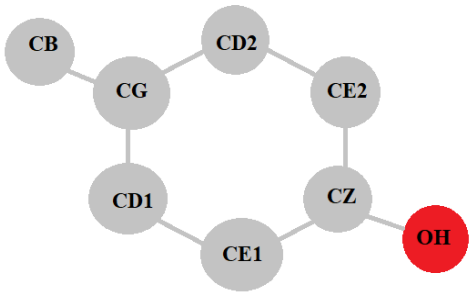


Figure 3.5.1 The electron density of 11 tyrosine rings from 1us0 superposed

The atoms used for this superposition were (central, linear, planar)=(CD1, CG, CD2). The success of the method is clear from the sliced image produced in Mathematica (Figure 3.8.1 b), with the central CD1 atom clearly central, the linear CG atom clearly on the x-axis (which is displayed vertically), and the planar atom CD2 bringing the structure flat to the xy plane so the cross section of the planar ring forms the xy plane.

The 3d superposition images (a, b) show a differing bond between the OH-CZ and the CG-CB, the oxygen bond being much thicker, the bonds between CD1-CG and CZ-CE2 appear to pull the electron density into a teardrop shape.

The results for glutamine were based on 145 glutamine residues across the PSU-Beta database, all the residues that fulfilled the criteria of having GLN in close contact with another GLN with 2 residues between. The results are given in Appendix 13 and are complicated, requiring further analysis.

The peptide bond result consists of all residues in 1ejg superposed, see Figure 3.5.2. The atoms for superposition were chosen as the planar atoms over the bond C-N+1-O. The peptide bond has clear information visible, including the bump of the protonated nitrogen.

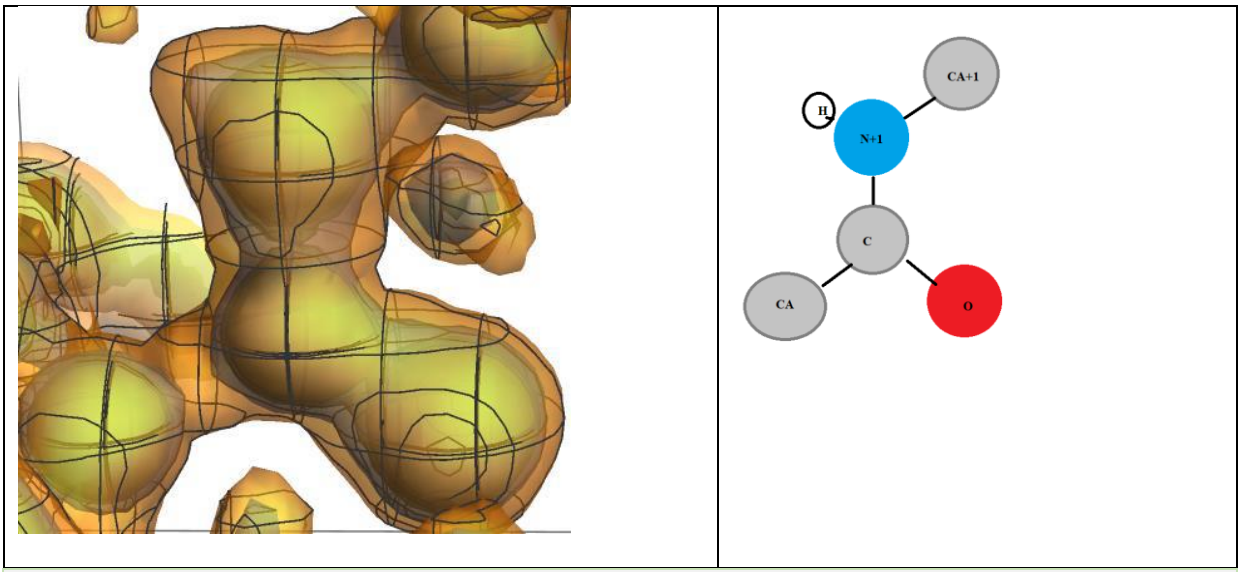
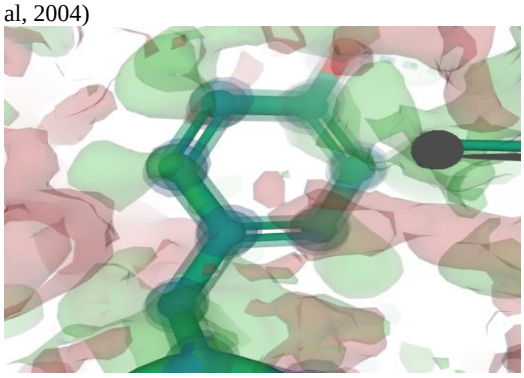


Figure 3.5.2 Peptide bond in structure 1ejg.
Shows the electron density from the peptide bond for all residues in 1ejg superposed

The superposition method was also applied to the difference matrices. In Figure 3.5.3 there is an example from the electron density application of the PDB showing a single TYR ring of the structure 1us0. Next to it is this study's superposition of the electron density for the 11 rings.

a) Protein Data Bank electron density for 1us0 (Howard et al, 2004)



b) PSU-ED image of difference superposition for 11 TYR tings in 1us0

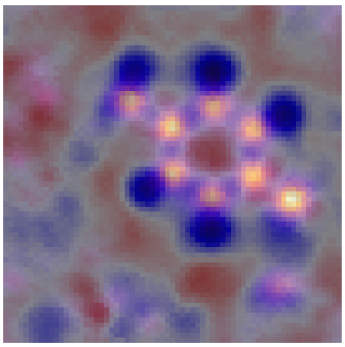


Figure 3.5.3 Tyrosine difference density and difference superposition for 1us0
a) *shows an image from the pdb website if a single residue.*
b) shows the superposed difference density for the 11 TYR rings, with the atoms at 40% transparency over the top. The protons in the difference image can be seen around the outside of the tyrosine ring.

The images are equivalent red-red and green-blue and can be seen to have a similar area around the outside of the ring - something missing from the model that exists in the electron density. This is indicative of the protonation state of the atoms in the tyrosine ring.

Additional analysis of the difference matrices shows a distribution of values at different resolutions, see Appendix 5. This shows that the differences are always similarly distributed no matter the resolution - that is the final model is always roughly equally different to the electron density. A comparison of the way the superposition of the density and differences change on resolution can be found in Appendix 6. The results files can be browsed in GitHub: PSU-ED Results

4. Discussion and Conclusions

4.1 Bond lengths and angles

In 2007, Jaskolski (Jaskolski et al, 2007) reviewed the most common restraints used in protein structure refinement for evidence of needed updates. They used the deposited 10 highest-resolution well-ordered structures and concluded that there was evidence for some change, notably C-N and N-Cα-C (tau). These findings have been reviewed using PSU-Beta's HQ set containing 3,434 structures <=1.3Å resolution. The results here are broadly in agreement with the Jaskolski (2007) values with refinements (see Table 3.3.1.1).

- As the resolution improves the bond lengths decrease, for N-CA this seems to still be shortening at the highest resolution bucket. The other C1N-N and C=O median lengths have settled, although the distributions at the highest resolution have perhaps widened, perhaps as a result of competition between forcefield and experimental evidence.
- For C=O the results agree with Jaskolski, suggesting a C=O bond length of 1.234Å over the E&H value of 1.231Å
- The results also suggest the N-CA bond length could be 1.455Å rather than 1.458Å
- The results do not agree with the Jaskolski value of 1.334Å for C1N-N, finding 1.332Å which is closer to the original E&H value of 1.329Å
- The average results from this study's dataset have lower variation (sd and iqr) than previous estimates of Jaskolski and E&H.
- The median is a good alternative to the mean, showing consistent values over the resolutions without outlier bias (which can come from refinement error even in the highest structures).

Jaskolski (2007) suggest that there is discussion on bimodality of the tau angle but no evidence of it. Here, the results clearly showing bimodality in the tau angle in relation to the PSI dihedral causing an overlap in modalities that appears as a spread in 1 dimension. There is also a clear difference when broken down on individual amino acids (Figure 3.3.1.5).

Another backbone pre-tau angle TAU1N has a distinct bimodality for proline in relation to the cis-trans-peptide bond – see Figure 3.4.6.10, which reinforces a prior idea in the literature (Kleywegt, 1997).

4.2 New insights into geometrical features

4.2.1 Correlations

The outliers found in new correlation plots, outlined in results section 3.1, demonstrate that even the highest resolution structures are susceptible to errors in atomic placement, even when there is good electron density in that area.

There is evidence that deviations from ideal atom positions can be indications of interesting structural features. For example, while the planarity of omega is known to deviate 20º (Jaskolski, 2007), but it has been suggested that an omega >30º could be an active site (Berkholz, 2009), or a conserved site (Berkholz, 2012). The additional use of correlation plots shows that omega values' deviation from planar can depend on the tau value. A tau of 110º appears to just be in the geometrically common region for an omega of 150º, but is doubtful with a tau of 100º.

The geometric correlation plots make the unusual geometries easier to identify. Currently, these values are manually inspected for evidence in the electron density (Berkholz, 2012) which makes the process of elucidating interesting or invalid geometric features time consuming and labour intense. The addition of a numerical indication of experimental evidence in these correlation plots on a per residue or calculation basis would be a valuable addition to these analyses.

An interesting feature of the correlation plots is the changing view of the secondary structures such as in the Ramachandran plot versus the "square" and "elliptical" plots, see Figure 4.2.1.1.

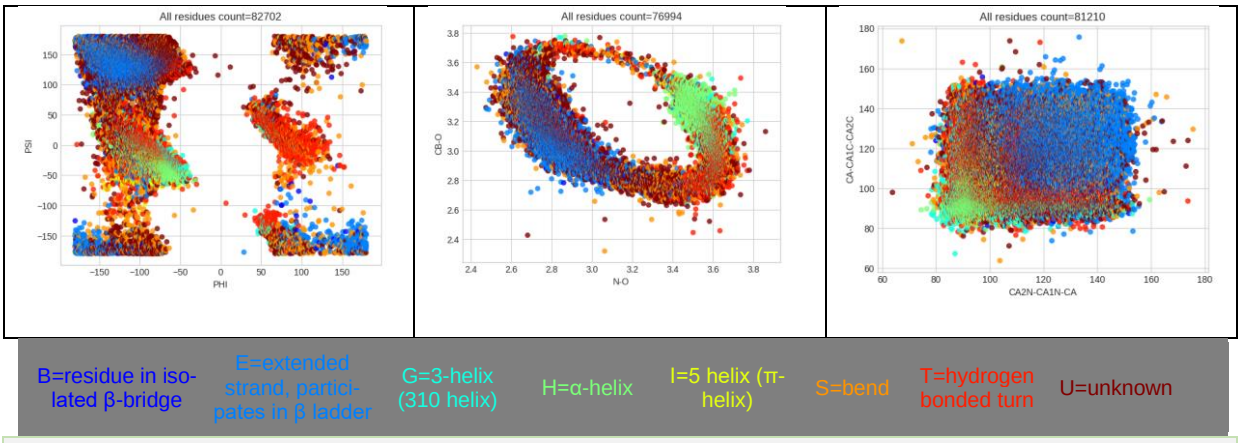


Figure 4.2.1.1 The secondary structure regions shown in different correlation plots

In Figure 4.2.1.1 there is the enticing feeling that the 'unknown' secondary structure, as defined by dssp, must be identifiable from the multiple plots. The brown unknown regions can be seen to inhabit specific regions in the other plots and further analysis must surely elucidate ways to identify existing secondary structures (or sub-groups or new ones).

The correlations show more than geometric accuracy. Where the plots are geometric in character, such as PSI/N-O (Figure 3.4.4.1) faint trace lines demonstrate the existence of forces between atoms

with an indication (not quantifiable) of the energy barrier; with no trace in a large dataset the geometric location would seem impossible due to the crystallographic state or steric hindrance; in the more probable regions the larger the spread of points the weaker the bond (not quantifiable).

4.2.3 Structural features

The correlation between Cα distance and omega established a direct relationship between them, showing that Cα-Cα <3.2Å means a cis-peptide bond, as suggested by Kleywegt (1997). This is useful in identifying pre-cis bonds, where the pre-omega has not been calculated - omega is traditionally calculated as the post-peptide bond because the cis identification means almost certainly that a proline will follow

Analysis for cis and proline suggests (Williams, 2015) that 5% of all prolines follow a cis-peptide bond. These numbers were calculated in Table 3.4.6.11 and approximately agree with Williams (2015), with the resolutions having some effect on the result, from 5.65% at 1.2-1.3Å to 6.41% at 0.9-1.0Å.

The identification of the Cα distance as direct identification of the cis-peptide bond does not guarantee that the bond has been identified as such. A comparison of OMEGA against CA-CA1C (Figure 3.4.5.3) shows that a very few peptide bonds are identified at omega-cis that are not distance-cis. These certainly warrant investigation; my expectation is that these are errors. The ability to colour the scatter points on experimental evidence would help identify this.

Analysis for proline shows the relationship between some multimodal distributions and the cis-trans peptide bond, notably the TAU1N angle. The PCA analysis on proline distinctly shows 2 clusters, correlating to 2 conformations of the proline ring, an up- and down-pucker state with all the CHI angles inverting, allowing the CHI angles to stand in for the pucker state of the ring. This agrees with a 2013 study on proline (Wu, 2013) that shows the same change in CHI angle between the two states: they found that CHI2 is linearly correlated with the puckering amplitude. Vitagliano et al (2001) describe the states with the formula:

- Up = CHI1 + CHI3 – CHI2 – CHI4 > 40º
- Down = CHI1 + CHI3 – CHI2 – CHI4 < 40º

The use of CHI1/CHI2 to stand in for the pucker state has led to some analysis of the multimodal proline distributions in the correlation plots, with interesting results in the traditional Ramachandran plot and the PHI/C1N-C showing a relationship with PHI and the pucker state (Vitagliano et al 2001).

The correlation plots have been discussed as indicating energy barriers between states. The CHI plots for proline, Figure 3.4.6.6, suggest that in the trans state proline can move quite freely between up and down pro, with a slight preference for down; in the cis state the energy barrier is high, but not impossible, for a transition to the up state. There is no evident movement between cis and trans states.

The analysis of proline suggests that PSU-Beta can usefully perform further ring analysis. For example, histidine shows tri-modality of its CHI2 angle (Figure 4.2.3.3) and has three protonation states that can be analysed on the ring geometry, looking at bond lengths and angles (Malinska et al, 2015). The states ND1 protonated, NE2 protonated and ND1&NE2 protonated are already known to have bond lengths and angles for different protonation states. High resolution structures may be able to examine this relationship.

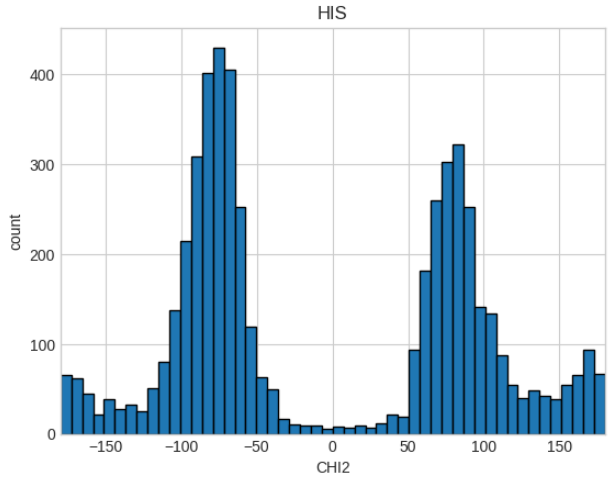


Figure 4.2.3.3 Histidine - trimodal CHI2

The successful analysis of proline suggests that a complete specification of the angles, dihedrals and bond lengths around histidine could provide useful in a full geometric specification of the protonation states.

4.2.4 Hydrogen bonds

Hydrogen bonds were largely omitted from this study. Geometric analysis holds promise for finding different types of hydrogen bonds on some geometric features, e.g. linear, 3-centred and bifurcated hydrogen bonds (Kuster et al, 2015), but the absence of hydrogen placement from most structures meant that this study limited analysis to close contacts. However, the alanine CHI1 result (Figure 3.4.8.1) shows the experimental evidence of improved confidence in placement of HB1 as resolution improves, suggesting that an analysis of the hydrogens in the structures would be an interesting study. My experience of uncertain atom placements and manual exploration of the electron density suggests that electron density features will need to be first added to the database.

The close contact analysis is limited here to those within the proteins themselves, this study has not looked at proteins in complex or bonding with water. In a 1994 study, McDonald and Thornton (1994) examined fulfilment of hydrogen bonding potential in proteins, with the finding that as the resolution improves, the percentage of nitrogen and oxygen atoms that fail to hydrogen bond falls. That is surely a feature of refinement rather than reality: this study finds most effects of resolution are rarity effect or reduction in refinement constraints given experimental evidence. McDonald and Thornton (1994) suggest this is evidence of better-quality structures, correlating also to better areas of the Ramachandran plot. This study finds that close contact atoms correlate to different areas of the Ramachandran plot. See Figure 4.2.4.1 for a comparison of the Ramachandran plot on some close contact high resolution structures against non-close contact high resolution structures. Note, difference image has a clear area of the Ramachandran plot more populated by close contact atoms and an area not populated by close contacts.

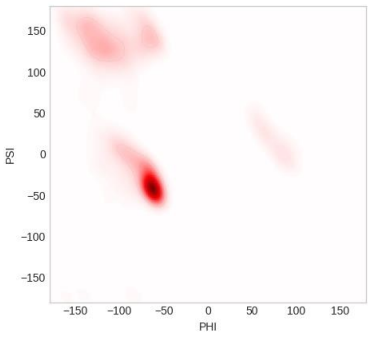
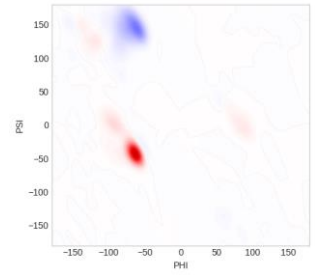
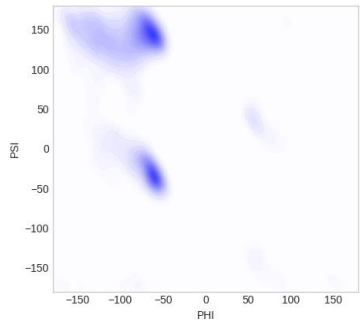
Observations = 10493 Resolution <=0.9 Å Nitrogen within 3.6 Å of an oxygen	Difference image for close contacts	Observations = 3392 Resolution <=0.9 Å Nitrogen NOT within 3.6 Å of oxygen
		

Figure 4.2.4.1 Comparing close contact and non-close contact residues in the Ramachandran plot at <=0.9Å This shows residues in close contact have a distinct region in the Ramachandran plot.

4.2.5 C-alpha modelling

Some c-alpha analysis results are reported in section 3.4.6, with the result from 3.4.5 reporting that the Cα distance directly correlates with the cis- or trans- nature of the peptide bond, which agrees with the results from Kleywegt (1997) in which he classifies the Cα distances into five classes: short, cis (2.8-3.0Å), poor, trans (3.7-3.9Å) and long. He notes that the small percentages of cis-peptide bonds (and the other non-trans categories) make them impossible to use as validation criteria as the tiniest deviation appears as an unacceptable outlier. Reflecting this point, Table 3.4.6.11 has so few observations at <=0.9Å it is hard to draw statistically inference from the values. Though the results may not work to validate the numbers statistically, a cis-distance should guarantee a cis-omega in the structure. As already discussed, this is found not to be the case in the lower resolutions, see Figure 3.4.5.3 with a tiny number of non cis-Cα lengths described as cis-peptide bonds.

The calculated Cα geometric measures have included angles and a pseudo-dihedral, enabling the production of the pseudo-Ramachandran plots (Asachi et al, 2020; Kleywegt, 1997) based on CA2N-CA1N-CAC and CA2N-CA1N-CA-CA1C, see Figure 3.4.7.2. The probability density agrees with the results from Kleywegt (1997) when all residues are taken together, but here the results have been further broken down per amino acid and show a strong difference for glycine and proline. This is not surprising, but important for the Cα model. Additionally, in Figure 3.4.6.3 the dihedral was analysed over 3 successive Cα angles along the backbone, with a distinct change in the character of the plots and probabilities. These Cα angles were analysed for each amino acid as violin plots (see Appendix 11). Distinct characteristics are seen for glycine and proline that could be important in the early stages of model building in crystallography (Kleywegt, 1997).

4.3 Electron density

Any question we have on a deposited structure comes back to the electron density. If there is experimental evidence, we have surety on atom placement. If not, then forcefields in refinement are used. The better the evidence, the less the reliance on forcefields, and the truer our understanding.

The electron density element of this project represents a proof of concept for future work. The challenges were mathematical and conceptual: can density from different structures in different configurations be captured in the same orientation and overlayed? Would that mean anything? Can density matrices be compared when they are based on different units with no known conversion?

The problem of comparing density matrices has a solution under review that is simple and statistical – assuming that the median density would always be the same and thus using a linear scale factor. Early analysis shows this is promising, see Appendix 5 for a comparison of this normalisation approach over 3 different resolution buckets.

The results have been surprising in their elegance at an early stage. The overlaying of all 11 tyrosine rings from the ultrahigh-resolution structure 1us0 (Howard et al, 2004) when viewed as 3d contours in Mathematica results in a view of the bonds so clear that the difference between the character of the bond between CG-CD1 and CG-CB can be seen, as can the size of the oxygen atom, Figure 3.5.1.

Jelsch, structure 1ejg, compared different refinement methods: spherical, multipolar, and a varying of average electron density parameters for the polypeptide main chain (Jelsch et al, 2000). In this paper the average difference density for polypeptide bonds was calculated. I repeated this difference analysis using the difference density from the PDBe database, and additionally performed the analysis with the density matrix, overlaying the density of 45 residues. This yields a distinctive view of the peptide bond, see Figure 3.5.2 in which the character of the peptide bond appears different to the N-C bond. The protonation of the nitrogen is apparent, with the distinct bump of electron density that is visible at

this high level of resolution. This technique has promise in elucidation of interesting information on the nature of the electrons in bonds and orbitals from ultrahigh-resolution structures. When used in conjunction with the database to extract atoms with similar geometric features the overlay could yield interesting information, particularly on planar parts of structures. The accuracy that this may give could also add further evidence to bond lengths and angles.

The information is not always easy to interpret. The effort to examine the nature of hydrogen bonding in GLN-GLN contacts in α-helices (Appendix 13) yielded more information than I can understand. I need to review what information will be interpretable and how to suitably define it.

On an individual structure and residue basis, the possibility that the density matrices can be compared also gives promise to an automated procedure for checking anomalies against experimental evidence, i.e. with the addition to the database of a normalised electron density for each atom in the dataset.

4.4 Resolution

An important figure in this study is Figure 3.4.2.2 – the rarity effect. This figure shows the enticing correlation between resolution and probability density, such that the resolutions almost map onto the probability density contours. Figure 3.4.2.1 shows the same effect for bfactor and rfree, but then look at Figure 3.4.2.3 and see the same effect for all structures whose second letter is ‘A’. The rarity effect is a simple statement of the obvious: when there are few observations, they are more likely to be found where they are more likely to be found. The resolutions happen to track the gradients because at each successively higher resolution there are successively fewer observations; rvalue, rfree and bfactor also follow this effect.

It is important when analysing any high-resolution data to keep in mind this rarity effect, it has its uses and its dangers. When looking for a mean value, the higher resolution structures will track the likely areas with their accurate atom placement, so for the bond lengths and angles they can give increasingly accurate values that may be used in refinement. However, these distributions can be multimodal. For the geometric trace correlation plots such as PSI/N-O, all possible areas provide essential information on the energetics of the structures. There are suggestions that resolution effects hydrogen bonding (McDonald and Thornton, 1994); that restraints should be changed depending on bfactor (Jaskolski, 2007) and that as resolution improves the CHI1 distribution “becomes more tightly clustered into these three idealized energy wells.” (Morris et al, 1992). The more probable areas are not better: they do not represent a better structure or a better refinement; the forcefield needs such parameters, but the best geometric value is the one that most accurately reflects the experimental evidence and the secondary structure.

The relaxing of the forcefield with better experimental evidence, as found in the CHI1 alanine results for HB1, is of interest for the elucidation of truer geometry. There is not clear evidence that the resolution limit has been reached – that we have all the experimental evidence needed at say, 0.85Å and there is no need to go further. The bond length analysis seems to show a continued improvement in results to 0.8Å for N-CA (Table 3.3.1.2), the close contact analysis shows distinct improvement at 0-1Å over 1.0-1.2Å, and mistakes are found in the ultrahigh-resolution structure 1i1w at 0.89Å. It may be that there is a cost benefit at a certain resolution, but the relatively low number of structures found at the very high resolutions (14 in total at a resolution higher than 0.8Å) means there is no certainty that such a point has been reached.

The impact of refinement software on the final structures was shown in section 3.4.3, with the interesting result in Figure 3.4.3.2 of successive releases of X-PLOR software versions showing tighter geometric correlations. When looking at geometry from the solved structures, these effects have inevitable influence: the improvement in experimental evidence and the relaxation of restraints can only improve this situation.

4.5 Overall research aim

One of the original purposes of this study was to review some of the refinement parameters: there is evidence here (Table 3.3.1.2) to agree with the Jaskolski (2007) recommendations for change to the C=O bond length from 1.231Å to 1.234Å, and additional evidence that the N-CA bond length is too wide at 1.458Å when 1.455Å seems to be a stable value at high resolutions. Generally, the amino acids form such characteristically different distributions that considering them all together, or considering them without glycine and proline, obscures important information.

The large-scale analysis of data at high resolution has enabled statistical analysis of geometric measures to give a detailed view of some of the distributions. The multimodal nature of the geometry of protein structures is clear: the bi-modality of tau is evident; the parametric relationship between N-O and CB-O with PSI underlying is elegant; the “square plot” is almost amusing.

We have seen that mistakes are made in even high resolution structures, and the correlation plots that have been identified in this study are a recommended tool for analysing the sanity of a structure over and above the refinement parameters and Ramachandran plot that are traditional. The identification of the geometric plots that cannot be deviated from adds some remarkably simple checks on structural integrity. They also aid an understanding of the energetics of positions and transitions that provides further insight into the protein structures.

None of this can really matter without experimental evidence. The method developed to compare density matrices provides promise for automation of analysis of outliers, and the overlay method suggests promise for a tool for further analysis of geometry directly on electron density sidestepping the uncer-

tainty of the refinement process and directly using the experimental evidence for the geometry of ultrahigh-resolution structures.

4.6 Further work

There are technical aspects to the project that could be improved: a review of the database's wide versus entity-value-attribute model of the database; the possibility of a database in the cloud; moving all my visual studio C++ code to a Linux CMake environment.

There are features that would be useful to add: the addition of hydrogen bonds by a calculated method; a specification of geometric measures for some specific interesting features like the histidine ring; or tyrosine ring; or looking at the sp2- or sp3- hybridised nature of certain carbon atoms.

Features could be added to make the user interface more friendly: an easier way to drill down to residues from a plot; an easy way to find the coordinates of any atoms in a residue on a plot; a link to the pdb information. The pre-calculation of the geometric measures is fast but given all atom coordinates are stored in the database the facility to request geometric measures not pre-calculated would be highly flexible and interesting. It would be interesting to be able to upload a pdb text file to the website and request correlation calculations.

The literature contains many different definitions of bond types and atom types: a mapping of a section of these from the literature to the database, with an analysis of geometry based on these definitions would be a useful addition.

There are some features that only now are possible or apparent: the ability to add a normalised electron density to every atom in the database could mean the ability to look at correlations on experimental evidence. This feature would help distinguish outliers on the geometric plot that are experimentally valid, potentially facilitating the discovery of structural features. The bfactor could also be added (currently available only as a hue on a per structure basis), but the electron density is the final arbiter (Wlodawer, 2007).

There are investigations from this data that suggest themselves: the identification of the secondary structures of some features in the 'unknown' secondary structure from dssp that appear in well-ordered areas of the correlation plots; the establishment of what makes a non-geometric point truly wrong or potentially interesting.

The correlations so far have been picked out by eye: with 170 calculations, correlating all of them is 28,730 to check manually. The non-linear nature of the correlations means that linear regression and PCA analysis are rarely appropriate. A method could be developed and automated, with an algorithm picking out correlations either mathematically or visually (e.g. machine learning).

The electron density portion of the work holds promise: for the possibility of analysing geometric features of proteins without refinement which requires work on the analysis and manipulation of the density space directly; for the investigation of the impact of atomic models in refinement which requires the rebuilding of density from structure factors with different models; for the elucidation of the true nature of the atomic bonds through the method of density overlay; for the possibility of linking existing structures to their experimental evidence directly for better analysis on the importance of geometric features; for the enticing challenge of a direct solution to the solving of a crystallographic structure – the invention of a mathematical sausage machine (Crick & Kendrew, 1957).

4.7 Implications of the research

The work in this study has implications for the validation of structures in terms of the correlation plots described in Appendix 14 and available on the website - Correlations Page .

The analysis of Cα features for angles has potential application to Cα modelling and the early model stage of crystallographic refinement.

There is potential as a teaching tool: the clarity of the geometric plots and the conceptual understanding of what they must mean provides evidential insight into the nature of atoms and bonds; the correlations provide insight into protein structure.

The identification of geometrically unusual features linked to the possibility of active sites is an idea that could be explored by the linking of the correlation plots to electron density for the integrated analysis of experimental evidence for geometrically suspicious residues.

The electron density work has the promise to elucidate more information on bonding, hydrogen bonding and atomic geometry: either through direct analysis of the electron density or when combined with the large geometric database to pick out similar features for overlay. This has the potential to add insights into bonds and the functional sites of proteins and to elucidate the nature of atoms and bonds.

References

Antonyuk, S. V., Strange, R. W., Sawers, G., Eady, R. R., & Hasnain, S. S. (2005). Atomic resolution structures of resting-state, substrate- and product-complexed Cu-nitrite reductase provide insight into catalytic mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(34), 12041–12046. https://doi.org/10.1073/pnas.0504207102

Asachi, P., Nguyen, U. N. T., Dang, J., Spencer, R. K., Martin, H. S., & Zuckermann, R. N. (2020). Skeletides: A modular, simplified physical model of protein secondary structure. *3D Printing and Additive Manufacturing*, *7*(2), 61–69. https://doi.org/10.1089/3dp.2019.0121

Crick, F. H. C., & Kendrew, J. C. (1957). X-Ray Analysis and Protein Structure. *Advances in Protein Chemistry*, *12*, 133–214. https://doi.org/https://doi.org/10.1016/S0065-3233(08)60116-3

Berkholz, D. S., Driggers, C. M., Shapovalov, M. V., Dunbrack, R. L., & Karplus, P. A. (2012). Non-planar peptide bonds in proteins are common and conserved but not biased toward active sites. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(2), 449–453. https://doi.org/10.1073/pnas.1107115108

Berkholz, D. S., Krenesky, P. B., Davidson, J. R., & Karplus, P. A. (2009). Protein Geometry Database: A flexible engine to explore backbone conformations and their relationships to covalent geometry. *Nucleic Acids Research*, *38*(SUPPL.1), 320–325. https://doi.org/10.1093/nar/gkp1013

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., … Zardecki, C. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, *58*(6 I), 899–907. https://doi.org/10.1107/S0907444902003451

Blakeley, M. P., Hasnain, S. S., & Antonyuk, S. V. (2015). Sub-atomic resolution X-ray crystallography and neutron crystallography: Promise, challenges and potential. *IUCrJ*, *2*, 464–474. https://doi.org/10.1107/S2052252515011239

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Elias, M., Liebschner, D., Koepke, J., Lecomte, C., Guillot, B., Jelsch, C., & Chabriere, E. (2013). Hydrogen atoms in protein structures: High-resolution X-ray diffraction structure of the DFPase. *BMC Research Notes*, *6*(1). https://doi.org/10.1186/1756-0500-6-308

Engh, R.A. & Huber, R. (1991). Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement. *Epidemiologia e Prevenzione*, *34*(5-6 Suppl 3), 27–31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21220832

Engh, R. A. & Huber, R. (2001). International Tables for Crystallography,Vol. F, edited by M. G. Rossman & E. Arnold, pp. 382– 392. Kluwer Academic Publishers, Dordrecht.

Escobedo, A., Topal, B., Kunze, M. B. A., Aranda, J., Chiesa, G., Mungianu, D., … Salvatella, X. (2019). Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09923-2

Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C. (2016). The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, *72*(2), 171–179. https://doi.org/10.1107/S2052520616003954

Hamelryck, T., & Manderick, B. (2003). PDB file parser and structure class implemented in Python. *Bioinformatics*, *19*(17), 2308–2310. https://doi.org/10.1093/bioinformatics/btg299

Hirano, Y., Takeda, K., & Miki, K. (2016). Charge-density analysis of an iron-sulfur protein at an ultra-high resolution of 0.48 Å. *Nature*, *534*(7606), 281–284. https://doi.org/10.1038/nature18001

Howard, E. I., Sanishvili, R., Cachau, R. E., Mitschler, A., Chevrier, B., Barth, P., … Podjarny, A. (2004). Ultrahigh resolution drug design I: Details of interactions in human aldose reductase-inhibitor complex at 0.66 Å. *Proteins: Structure, Function and Genetics*, *55*(4), 792–804. https://doi.org/10.1002/prot.20015

Jaskolski, M., Gilski, M., Dauter, Z., & Wlodawer, A. (2007). Stereochemical restraints revisited: How accurate are refinement targets and how much should protein structures be allowed to deviate from them? *Acta Crystallographica Section D: Biological Crystallography*, *63*(5), 611–620. https://doi.org/10.1107/S090744490700978X

Jelsch, C., Teeter, M. M., Lamzin, V., Pichon-Pesme, V., Blessing, R. H., & Lecomte, C. (2000). Accurate protein crystallography at ultra-high resolution: Valence electron distribution in crambin. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(7), 3171–3176. https://doi.org/10.1073/pnas.97.7.3171

Joosten, R. P., Te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, *39*(SUPPL. 1), 411–419. https://doi.org/10.1093/nar/gkq1105

Kabsch, W., & Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, *22*, 2577–2637.

Kleywegt, G. J. (1997). Validation of protein models from C(α) coordinates alone. *Journal of Molecular Biology*, *273*(2), 371–376. https://doi.org/10.1006/jmbi.1997.1309

Koepke, J., Scharff, E. I., Lücke, C., Rüterjans, H., & Fritzsch, G. (2003). Statistical analysis of crystallographic data obtained from squid ganglion DFPase at 0.85 Å resolution. *Acta Crystallographica - Section D Biological Crystallography*, *59*(10), 1744–1754. https://doi.org/10.1107/S0907444903016135

Kuster, D. J., Liu, C., Fang, Z., Ponder, J. W., & Marshall, G. R. (2015). High-resolution crystal structures of protein helices reconciled with three-centered hydrogen bonds and multipole electrostatics. *PLoS ONE*, *10*(4), 1–37. https://doi.org/10.1371/journal.pone.0123146

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, *26*(2), 283–291. https://doi.org/10.1107/s0021889892009944

Malinska, M., Dauter, M., Kowiel, M., Jaskolski, M., & Dauter, Z. (2015). Protonation and geometry of histidine rings. *Acta Crystallographica Section D: Biological Crystallography*, *71*, 1444–1454. https://doi.org/10.1107/S1399004715007816

McCoy, A. J., Oeffner, R. D., Wrobel, A. G., Ojala, J. R. M., Tryggvason, K., Lohkamp, B., & Read, R. J. (2017). Ab initio solution of macromolecular crystal structures without direct methods. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(14), 3637–3641. https://doi.org/10.1073/pnas.1701640114

McDonald, I. K., & Thornton, J. M. (1994). Satisfying Hydrogen Bonding Potential in Proteins. Journal of Molecular Biology, 238(5), 777–793. doi:10.1006/jmbi.1994.1334

Morris, A. L., MacArthur, M. W., Hutchinson, E. G., & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*, *12*(4), 345–364. https://doi.org/10.1002/prot.340120407

Natesh, R., Manikandan, K., Bhanumoorthy, P., Viswamitra, M. A., & Ramakumar, S. (2003). Thermostable xylanase from Thermoascus aurantiacus at ultrahigh resolution (0.89 Å) at 100 K and atomic resolution (1.11 Å) at 293 K refined anisotropically to small-molecule accuracy. *Acta Crystallographica - Section D Biological Crystallography*, *59*(1), 105–117. https://doi.org/10.1107/S0907444902020164

Perutz, M. F. (1990). How W. L. Bragg invented X-ray analysis. *Acta Crystallographica Section A*, *46*(8), 633–643. https://doi.org/10.1107/S010876739000410X

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, *7*(1), 95–99. https://doi.org/10.1016/S0022-2836(63)80023-6

Rose, G. D. (2019). Ramachandran maps for side chains in globular proteins. *Proteins: Structure, Function and Bioinformatics*, *87*(5), 357–364. https://doi.org/10.1002/prot.25656

Velankar, S., Best, C., Beuth, B., Boutselakis, C. H., Cobley, N., Sousa da Silva, A. W., … Kleywegt, G. J. (2009). PDBe: Protein Data Bank in Europe. *Nucleic Acids Research*, *38*(SUPPL.1), 308–317. https://doi.org/10.1093/nar/gkp916

Vitagliano, L., Berisio, R., Mastrangelo, A., Mazzarella, L., & Zagari, A. (2001). Preferred proline puckerings in cis and trans peptide groups: Implications for collagen stability. *Protein Science*, *10*(12), 2627–2632. https://doi.org/10.1110/ps.ps.26601a

Williams, C. J., Using C-Alpha Geometry to Describe Protein Secondary Structure and Motifs (PhD thesis) (2015)

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., … Richardson, D. C. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*, *27*(1), 293–315. https://doi.org/10.1002/pro.3330

Wilson, K. S., Butterworth, S., Dauter, Z., Lamzin, V. S., Walsh, M., Wodak, S., … Rullmann, J. A. C. (1998). Who checks the checkers? Four validation tools applied to eight atomic resolution structures. *Journal of Molecular Biology*, *276*(2), 417–436. https://doi.org/10.1006/jmbi.1997.1526

Wlodawer, A. (2007). Stereochemistry and Validation of Macromolecular Structures. *Methods Molecular Biology, 1607*, 68–75. https://doi.org/10.1007/978-1-4939-7000-1

Wu, D. (2013). The puckering free-energy surface of proline. *AIP Advances*, *3*(3). https://doi.org/10.1063/1.4799082

Yao, S., & Moseley, H. N. B. (2019). *A chemical interpretation of protein electron density maps in the worldwide protein data bank Software and full results available at :* https://www.biorxiv.org/content/10.1101/613109v1

Zarychta, B., Lyubimov, A., Ahmed, M., Munshi, P., Guillot, B., Vrielink, A., & Jelsch, C. (2015). Cholesterol oxidase: Ultrahigh-resolution crystal structure and multipolar atom model-based analysis. *Acta Crystallographica Section D: Biological Crystallography*, *71*, 954–968. https://doi.org/10.1107/S1399004715002382

Appendices

Appendix 1: Omega at four resolution buckets 78

Appendix 2: Validation plots of the high-resolution set before and after 79

Appendix 3: Structures that have been individually manually checked 83

Appendix 4: All calculated geometric measures 85

Appendix 5: Density and difference matrix comparison over resolutions 92

Appendix 6: Density and difference images at different resolutions 102

Appendix 7: Violin Plots for Tau 103

Appendix 8: The bimodal nature of N-O, and CB-O 104

Appendix 9: PHI distributions per amino acid 105

Appendix 10: PSI distributions per amino acid 106

Appendix 11: Comparing Cα angles along the chain 107

Appendix 12: Distribution reports from website 108

Appendix 13: 146 GLN Superposition 112

Appendix 14: Correlation page for all HIGH residues, on refinement method 113

Appendix 15: CHI1 for different resolutions buckets 114

Appendix 16: Summary statistics for CHI1 distributions 115

Appendix 17: Distribution close contact differences for hydrogen bond donors 117

Appendix 18: Distribution close contact differences for hydrogen bond acceptors 118

Appendix 19: KDE Bandwidth settings comparison for probability density 119

Appendix 20: Proline dominated region in cis/trans correlation plot 120

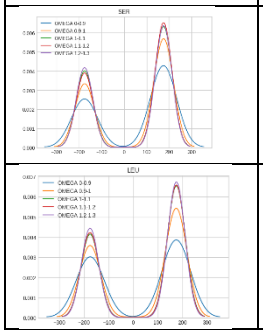
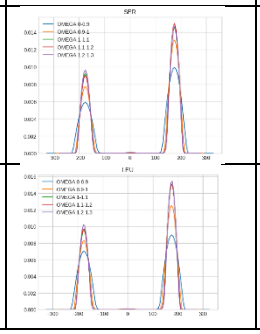
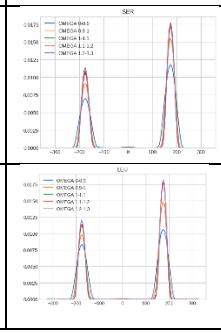
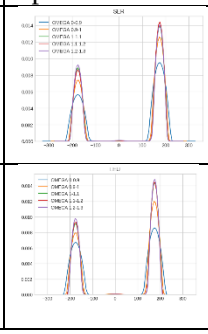
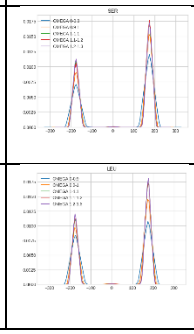
Appendix 21: Correlation of PSI versus CA1N-CA-CA1C for all amino acids 121

Appendix 22: Alanine CHI1 and hydrogen placement 122

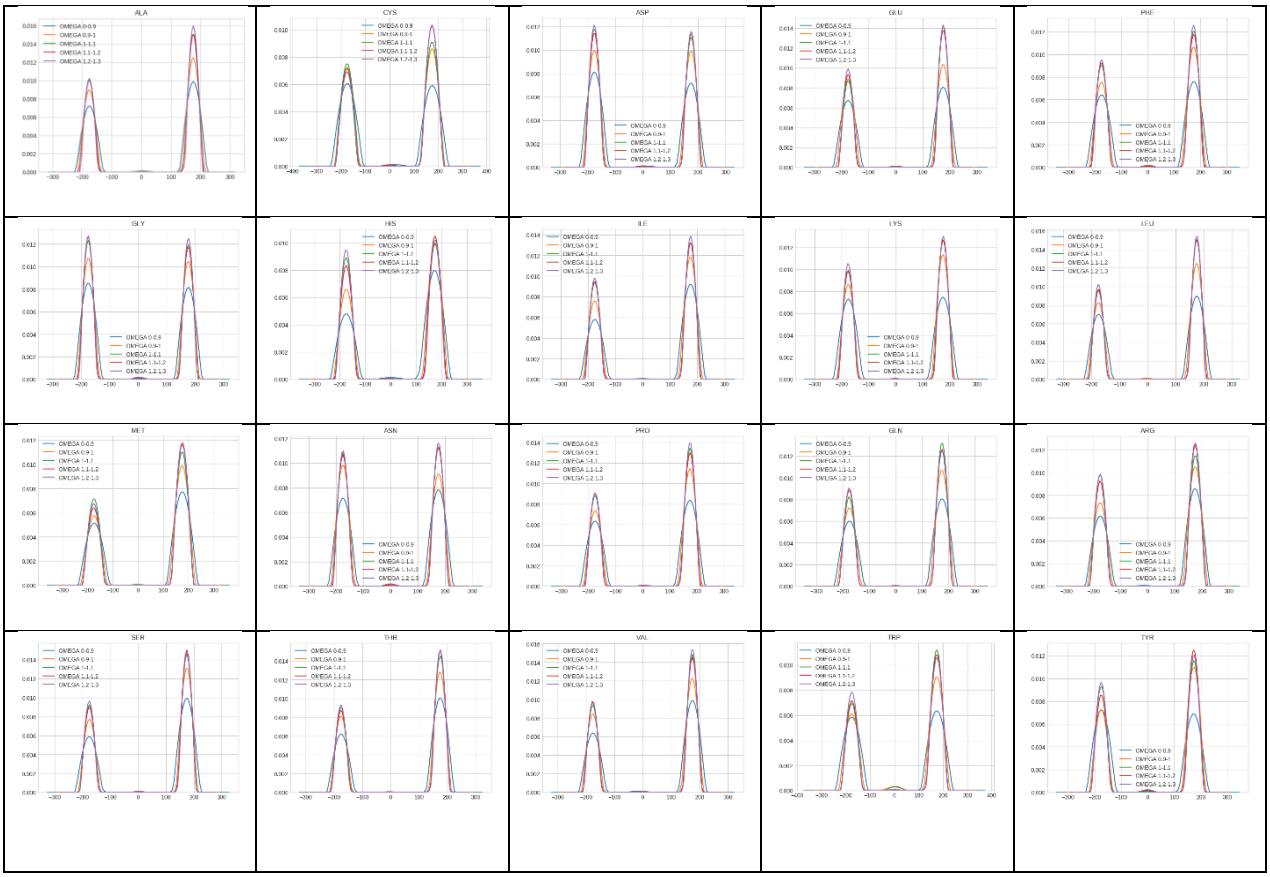
Appendix 1: Omega at four resolution buckets

Omega compared at four resolution buckets for bfactor<=100Å^2, rvalue <=0.16Å, rfree <=0.3Å.

A comparison is made between 5 kernel smoothing methods, using the silverman rule of thumb, with cos chosen for the method used.

Gaussian	Cos	biw	epa	tri
				

Omega per amino acid

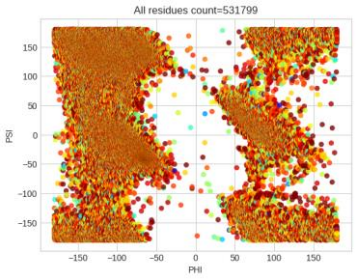


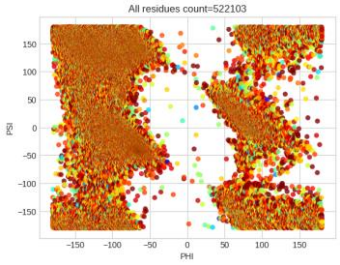
Appendix 2: Validation plots of the high-resolution set before and after

The validation report for the high-resolution structures clearly shows geometrically impossible features. These have some correspondence to the Ramachandran plot but are clearer. The plots on the right is after some structure validation. Although some structures are excluded from the right-hand, they cannot be removed as there is no evidence to do this as the structures are interpreted correctly as given. They are suspicious structures, information in Appendix 3. All plots are coloured on resolution.

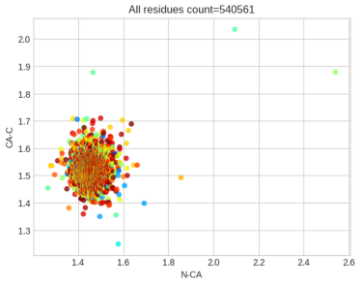
Key = Resolution (Å) 0.48 - 0.75 – 0.95 – 1.05 – 1.15 – 1.25 – 1.3

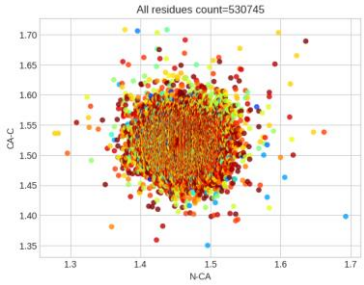
a) Ramachandran plot



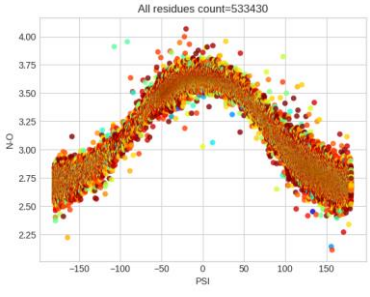


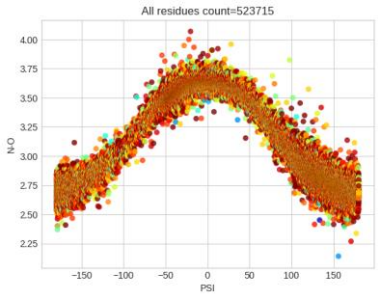
b) N-CA vs CA-C



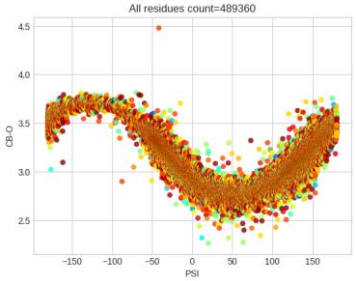


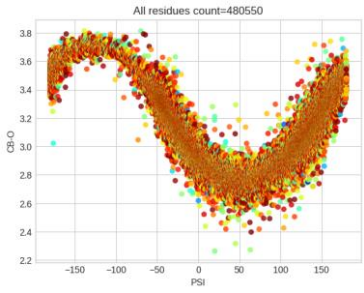
c) PSI vs N-O



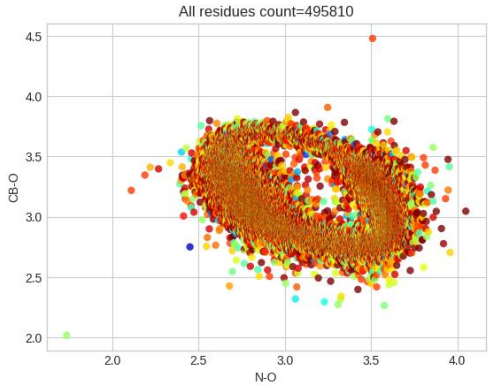


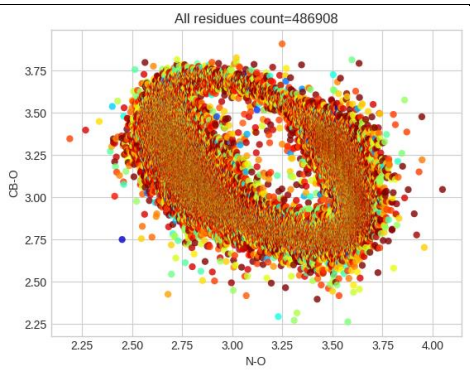
d) PSI vs CB-O



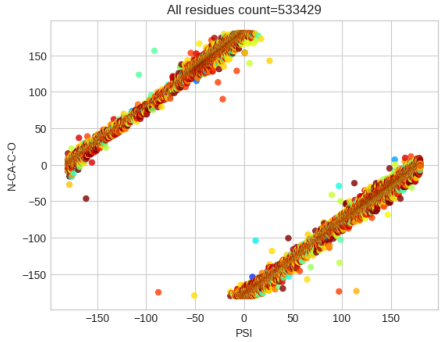


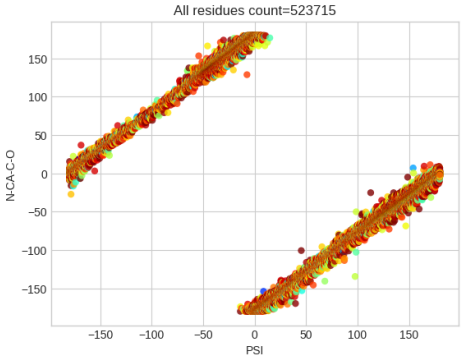
e) N-O vs CB-O



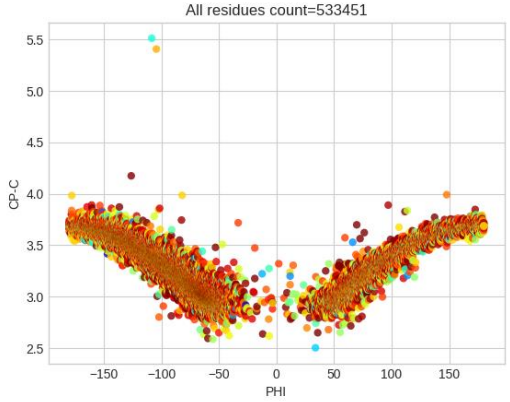


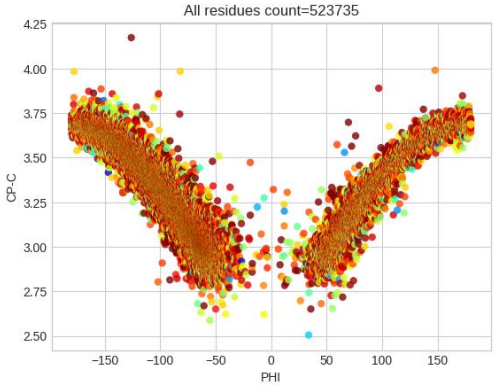
f) PSI vs NA-CA-C-O



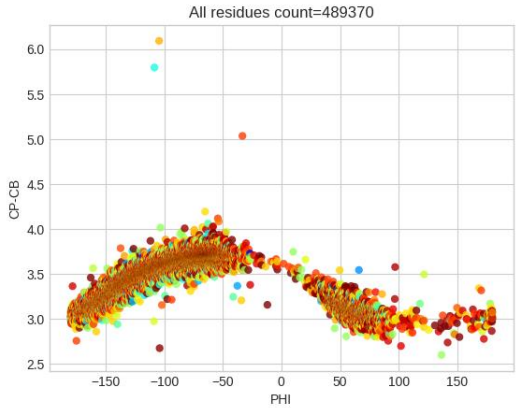


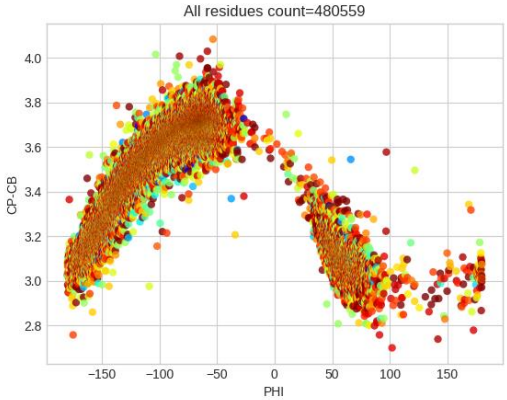
g) PHI vs C1N-C



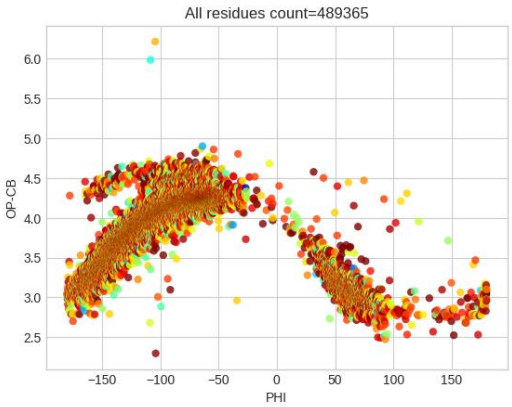


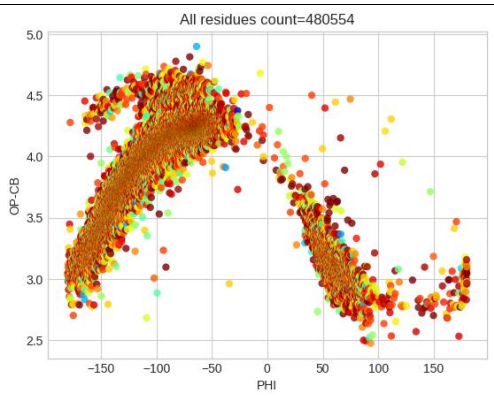
h) PHI vs C1N-CB



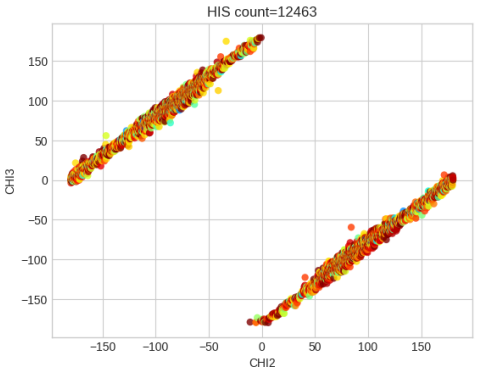


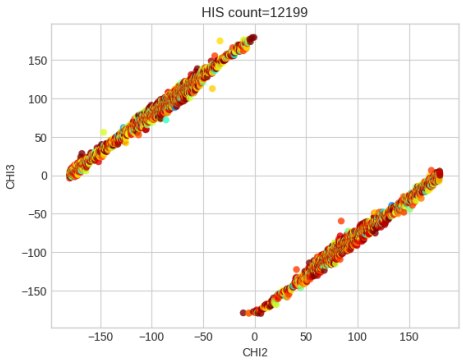
i) PHI vs O1N-CB



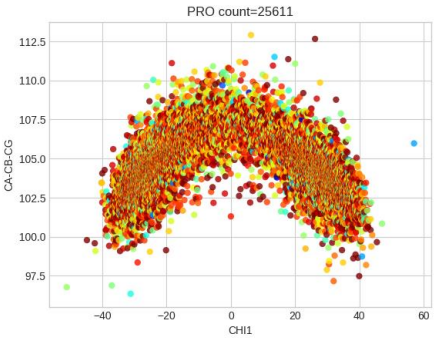


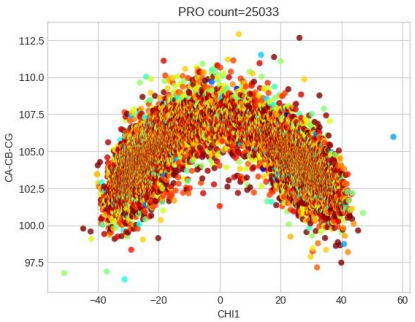
j) CHI2 vs CHI3 for HIS only



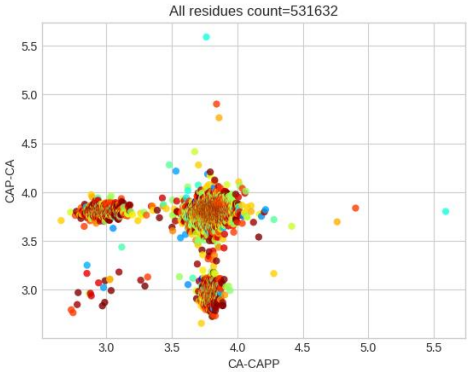


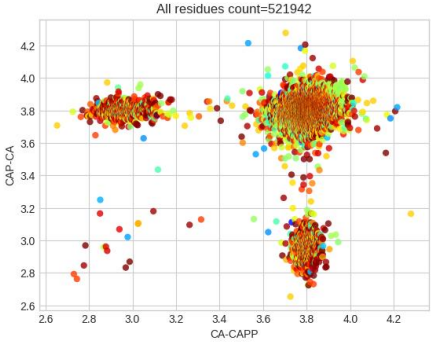
k) CHI1 vs CA-CB-CG PRO only



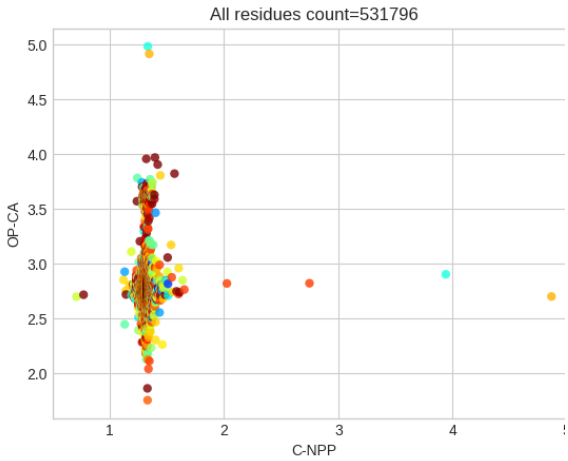


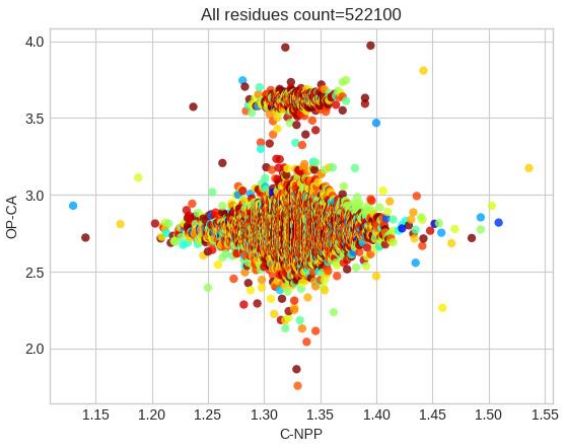
l) CA-CA1C vs CA1N-CA



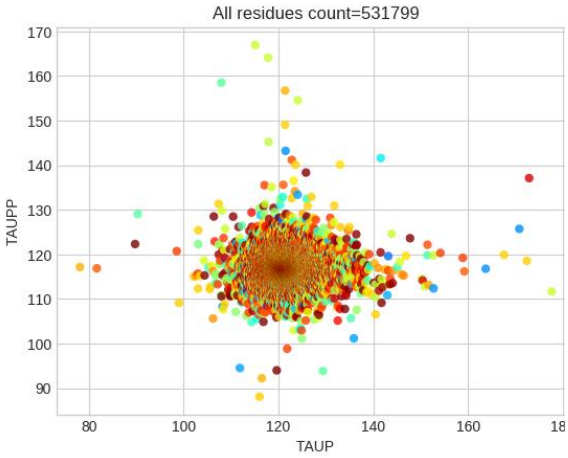


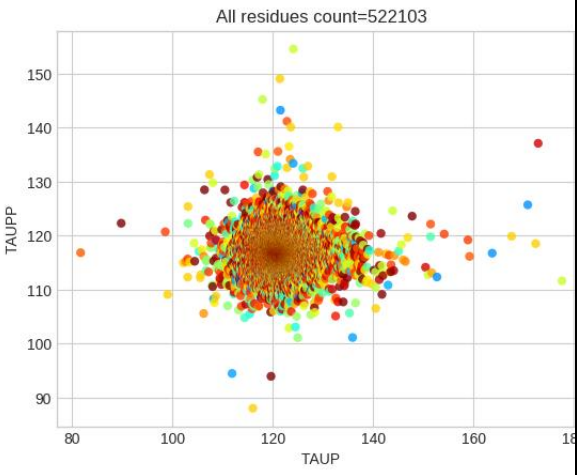
m) C-N1C vs O1N-CA





n) TAU1N vs TAU1C





Note a name change to the geometric measures, where P used to refer to the previous and PP used to refer to the next, so TAUP is now TAU1N and TAUPP as now TAI1C etc.

Appendix 3: Structures that have been individually manually checked

The following structures were found to contain geometrically unusual features in the initial validation results.

Table A.3.1 Structures with invalid geometric features

Structure	Geometric Concern	Investigation	Decision
2BW4	CP-C distance is 5.5Å Errors seen in most validation plots.	Looked at the structure in Chimera. Verified my calculations manually. It seems that the occupants A and B may have been mixed up around residue 195.	I have no evidence to remove, this is the structure as given and I interpret correctly. Marked as 'CHECKED'. It is surprising that this passed basic refinement checks.
1W0N	There is a TAUP > 140° and a TAUPP around 127° which are extreme outliers.	The residue 42, ASN, has an A and a B occupant. Manually verified - the reported TAUP is 143.1° and TAUPP is 127.7°	I have no evidence to remove. I interpret correctly. Marked as 'CHECKED'.
1I1W	A spot in the middle of PSI N-O, clearly geometrically unusual. It does not appear invalid on the Ramachandran plot.	SP^2 hybridized carbonyl on A:180 is non planar and irregular. It is supported by the electron density.	No evidence to remove, mark as 'CHECKED'.
5GJI	CA-CA distance between residues 394 (ASP) and 395 (PRO) < 3Å	Only 1 occupant, bfactor < 10 Manually verified distance as 2.962Å	Nothing evidently wrong. No change to status on 'IN', consider this to be an accepted area to be investigated. Xxx-PRO is accepted in standard bond lengths as an excluded case for C-N. (ref Jaskolski Table 2).
2vk2	CP-CB distance > 5 Å between residues 292 to 293	Only 1 occupant, bfactor of 292.LYS 70 and of 293.LYS 20. Manually verified as 5.034Å	It is the CP-N bond that is extreme at 2.748 Å. I have no evidence to reject so marked 'CHECKED'
1G2B	47-48 CP-N extreme at 4.876 Å	The chain begins at 48 so code incorrectly assumes a continuation from 47 to 48.	Marked as 'OUT' as not handled, need to add rules to code to identify this cases (when rules are understood)
2j9j	50-51 for both chains A and B are extreme outliers. A high at 1.636 and B low at 0.708	The bfactor is < 20, there are A and B occupants.	Visually strange, no evidence to reject, marked as 'CHECKED', possible mix-up of occupants.
1es9	57.ILE to 58.TRP extreme CP-N of 0.77	No occupant, bfactor about 50, visually odd.	No evidence to reject, marked as 'CHECKED'
1w32	A.86.SER to 87.SER 3 extreme CP-N of 1.652	3/2 occupants, low bfactor, big bundle of atoms at the bond.	No evidence to reject, marked as 'CHECKED'.
4p40	A.324.LYS to 325.VAL extreme CP-N of 1.602	Verified value, looks ok.	No evidence to reject, marked as 'CHECKED'.
2gec	B.86.PRO to 87.VAL extreme CP-N of 1.609	Looks ok apart from long bond.	No evidence to reject, marked as 'CHECKED'
1n62	E.617.GLY to 618.LEU extreme CP-N of 1.603	Enormous structure over 19,000 atoms. Looks ok apart from long bond.	No evidence to reject, marked as 'CHECKED'
6k05	A.118.GLY to 119.ARG high CP-N of 1.55	No occupants, reasonable bfactors.	No evidence to reject, marked as 'CHECKED'
5k26	B.78.SER to 79.GLY CP-N is 1.542	No occupants, reasonable bfactors.	No evidence to reject, marked as 'CHECKED'
1zl0	B.88.GLY to 89.TYR	A and B GLY occupants. Huge structure. Low bfactors.	No evidence to reject, marked as 'CHECKED'

1mn8	A.3.GLN, N-O extreme low of 1.737	This is the chain beginning at it is clearly disordered.	Marking as ‘OUT’ as atoms are in the same space. In future my system could make a decision to start at a different residue.
1mj4	A.37.VAL CB-O extreme value of 4.479	Occupants A and B look like the CB and CG2 have been mixed up.	I am confident this is wrong, but will mark ‘CHECKED’ as I need evidence for an error.
3jvl	A.346.ALA low N-O of 2.111	Chain beginning but not obvious what is wrong	No evidence to reject, marked as ‘CHECKED’
1bxo	A.279.ASP high N-CA of 2.095	The C-O bond is also extreme. Odd looking residue with hydrogens, and bfactor around 80.	No evidence to reject, marked as ‘CHECKED’
1lu4	A.1093.ALA long N0CA of 1.856	Nothing evidently wrong apart from the long bond.	No evidence to reject, marked as ‘CHECKED’
3ned	A.65.PHE high CA-C of 1.878	CA has only 1 occupant bu C has A.B,C occupants, chain break after.	No evidence to reject, marked as ‘CHECKED’
6rqq	A.145.GLY, 146.ASP, 147.PRO CA-CA between both pairs in the low area < 3.4	Near chain beginning, GLY-ASP-PRO-PRO, feels like nothing is wrong.	Leaving as ‘IN’, out of area but looks like it is fine. GLY and PRO are excluded in standard checks for N-CA and CA-C (Jaskolski, E&H ref).
1zq5	A.134.ASN, 135.GLY to 136.LYS to 137.VAL CA-CA in the chains both in <3.4 area	Near a break, but atoms look ordered, bfactor < 50 no occupant.	Leaving as IN, currently consider this rare but accepted area to investigate. GLY special case,
2wlv	A.144.TYR CA-C =1.71	Chain end, nothing obvious.	No evidence to reject, marked as ‘CHECKED’
4e9s	N-O vs PSI non-geometric region. 2 residues, A.44.ALA and A.330.LEU	Nothing evident, no occupants, bfactor < 20	No evidence to reject, marked as ‘CHECKED’
1w6s	D.3072.SER to 3073.ALA is extreme C-N at 2.025	This is a huge structure with incomplete atoms. The 3023.ALA has only N	Marking as ‘OUT’ as the structure is incomplete at this area and it could be I am not handling it correctly. Further code investigation needed.
2z26	B.318.GLU to 319.GLN Out of geometric region in PHI sv CP-CB, CP-CB < 2.8 PHI about -100.	No occupants, bfactor < 50, nothing obvious.	No evidence to reject, marked as ‘CHECKED’

Appendix 4: All calculated geometric measures

The complete list of geometric measures that are calculated. The design ensures it is easy to think of a new measure of interest and add it easily.

Alias	Amino Acid	Code * = ALL	Description
C-CB1C	C-CB1C	*	Distance between C and the next CB
C-C1C	C-C1C	*	Distance between C and the next C
C-N1C	C-N1C	*	Distance between C and the next N
C-O	C-O	*	Distance between C and O
CA-C	CA-C	*	Distance between CA and C
CA-C-O	CA-C-O	*	Angle between CA-C-O
CA-CA1C	CA-CA1C	*	Distance between Ca and the next CA
CA-CA1C -CA2C	CA-CA1C -CA2C	*	Angle between this and the next 2 CAs
CA-CB-CG	CA-CB-CG	*	Angle between CA, CB and CG
CA1N-CA	CA1N-CA	*	Distance between previous CA and CA
CA1N -CA-CA1C	CA1N -CA-CA1C	*	Angle between previous, this and next CA
CA2N -CA1N -CA	CA2N -CA1N -CA	*	Angle between previous 2 CAs and this
CA2N -CA1N -CA-CA1C	CA2N -CA1N -CA-CA1C	*	Dihedral angle of CAs
CB-CA-C	CB-CA-C	*	Angle between CB, CA and C
CB-N1C	CB-N1C	*	Distance between CB and the next N
CB-O	CB-O	*	Distance between CB and O
CB1N -N	CB1N -N	*	Distance between previous CB and N
C1N -C	C1N -C	*	Distance between previous C and C
C1N -CB	C1N -CB	*	Distance between previous C and CB
C1N -N	C1N -N	*	Distance between previous C and N
HG-N1N	HG-N1N	*	Distance between HG and next N – there are not many hydrogens.
HG-O	HG-O	*	Distance HG-O
N-CA	N-CA	*	Main chain distance N-CA
N-CA-C-O	N-CA-C-O	*	Dihedral
N-CA-CB	N-CA-CB	*	Angle
N-N1C	N-N1C	*	N and next N distance
N-O	N-O	*	Distance between N and O
N1N -N	N1N -N	*	Distance between previous N and N

O-C-N1C	O-C-N1C	*	Angle spanning to next N
O-CA1C	O-CA1C	*	Distance from O to next CA
OMEGA	CA-C-N1C-CA1C	*	Classic dihedral main chain angle
O1N -CA	O1N -CA	*	Distance from previous O to CA
O1N -CB	O1N -CB	*	Distance from previous O to CB
O1N -CP-N	O1N -CP-N	*	Angle from previous O and C to N
PHI	C1N -N-CA-C	*	Classic dihedral main chain angle
PSI	N-CA-C-N1C	*	Classic dihedral main chain angle
TAU	N-CA-C	*	Main chain angle
TAU1N	C1N-N-CA	*	Main chain angle starting with previous c
TAU1C	CA-C-N1C	*	Main chain angle going to next N
The following are residue specific definitions for the CHI and improper angles. CHI definitions follow standards, the IMP1-IMP5 angles are defined by me for the purpose of being able to do comparisons. They are a selection of improper angles for each residue that seem useful (where hydrogens are generally not available).			
CHI1	C-CA-CB-HB1	ALA	
IMP1	CB-C-N-HA	ALA	
IMP2	HB3-CA-CB-HB1	ALA	
CHI1	N-CA-CB-CG	ARG	
CHI2	CA-CB-CG-CD	ARG	
CHI3	CB-CG-CD-NE	ARG	
CHI4	CG-CD-NE-CZ	ARG	
CHI5	CD-NE-CZ-NH1	ARG	
IMP1	CB-C-N-HA	ARG	
IMP2	HE-CZ-CD-NE	ARG	
IMP3	NH2-NH1-NE-CZ	ARG	
IMP4	NE-HE-NH2-CZ	ARG	
IMP5	CG-CA-HB2-HB1	ARG	
CHI1	N-CA-CB-CG	ASN	
CHI2	CA-CB-CG-OD1	ASN	
IMP1	CB-C-N-HA	ASN	
IMP2	ND2-OD1-CB-CG	ASN	

IMP3	CG-CA-CH2-HB1	ASN	
CHI1	N-CA-CB-CG	ASP	
CHI2	CA-CB-CG-OD1	ASP	
IMP1	CB-C-N-HA	ASP	
IMP2	OD2-OD1-CB-CG	ASP	
IMP3	CG-CA-HB2-HB1	ASP	
CHI1	N-CA-CB-SG	CYS	
IMP1	CB-C-N-HA	CYS	
IMP2	SG-CA-HB2-HB1	CYS	
IMP3	SG-CB-CA-C	CYS	
CHI1	N-CA-CB-CG	GLN	
CHI2	CA-CB-CG-CD	GLN	
CHI3	CB-CG-CD-OE1	GLN	
IMP1	CB-C-N-HA	GLN	
IMP2	NE2-OE1-CG-CD	GLN	
IMP3	CG-CA-HB2-HB1	GLN	
IMP4	CD-CB-HG2-HG1	GLN	
CHI1	N-CA-CB-CG	GLU	
CHI2	CA-CB-CG-CD	GLU	
CHI3	CB-CG-CD-OE1	GLU	
IMP1	CB-C-N-HA	GLU	
IMP2	OE2-OE1-CG-CD	GLU	
IMP3	CG-CA-HB1-HB2	GLU	
IMP4	CD-CG-HG2-HG1	GLU	
IMP1	C-N-HA2-HA1	GLY	
CHI1	N-CA-CB-CG	HIS	
CHI2	CA-CB-CG-ND1	HIS	
CHI3	CA-CB-CG-CD2	HIS	

IMP1	CB-C-N-HA	HIS	
IMP2	CD2-ND1-CB-CG	HIS	
IMP3	CD2-NE2-CE1-ND1	HIS	
IMP4	NE2-CE1-ND1-CG	HIS	
IMP5	CG-CA-HB2-HB1	HIS	
CHI1	N-CA-CB-CG1	ILE	
CHI2	CA-CB-CG1-CD1	ILE	
CHI3	CB-CG1-CD1-CD11	ILE	
CHI4	CA-CB-CG2-HG21	ILE	
IMP1	CB-C-N-HA	ILE	
IMP2	CG1-CG2-CA-HB	ILE	
IMP3	CD1-CB-HG12-HG11	ILE	
IMP4	HG23-CB-HG22-HG21	ILE	
IMP5	HD13-CG1-HD12-HD11	ILE	
CHI1	N-CA-CB-CG	LEU	
CHI2	CA-CB-CG-CD1	LEU	
CHI3	CB-CG-CD1-HD11	LEU	
CHI4	CB-CG-CD2-HD21	LEU	
IMP1	CB-C-N-HA	LEU	
IMP2	CD2-CD1-CB-HG	LEU	
IMP3	CG-CA-HB2-HB1	LEU	
IMP4	HD13-CG-HD12-HD11	LEU	
IMP5	HD23-CG-HD22-HD21	LEU	
CHI1	N-CA-CB-CG	LYS	
CHI2	CA-CB-CG-CD	LYS	
CHI3	CD-CE-NZ-HZ1	LYS	
IMP1	CB-C-N-HA	LYS	

IMP2	CG-CA-HB2-HB1	LYS	
IMP3	CD-CB-HG2-HG1	LYS	
IMP4	NZ-CD-HE2-HE1	LYS	
IMP5	HZ3-CE-HZ2-HZ1	LYS	
CHI1	N-CA-CB-CG	MET	
CHI2	CA-CB-CG-SD	MET	
CHI3	CG-SD-CE-HE1	MET	
IMP1	CB-C-N-HA	MET	
IMP2	CG-CA-HB2-HB1	MET	
IMP3	SD-CB-HG2-HG1	MET	
IMP4	HE3-SD-HE2-HE1	MET	
CHI1	N-CA-CB-CG	PHE	
CHI2	CA-CB-CG-CD1	PHE	
IMP1	CB-C-N-HA	PHE	
IMP2	CG-CA-HB2-HB1	PHE	
IMP3	CE2-CD2-CG-CB	PHE	
IMP4	CZ-CE1-CD1-CG	PHE	
IMP5	CG-CD2-CE2-CZ	PHE	
CHI1	N-CA-CB-CG	PRO	
CHI2	CA-CB-CG-CD	PRO	
CHI3	CB-CG-CD-N	PRO	
CHI4	CG-CD-N-CA	PRO	
CHI5	CD-N-CA-CB	PRO	
IMP1	CB-C-N-HA	PRO	
IMP2	CG-CA-HB2-HB1	PRO	
IMP3	CD-CB-HG2-HG1	PRO	
IMP4	N-CG-HD2-HD1	PRO	
CHI1	N-CA-CB-OG	SER	

CHI2	CA-CB-OG-H	SER	
IMP1	CB-C-N-HA	SER	
IMP2	OG-CA-HB2-HB1	SER	
CHI1	N-CA-CB-OG1	THR	
CHI2	CA-CB-CG2-HG21	THR	
CHI3	HG23-CB-OG1-H	THR	
IMP1	CB-C-N-HA	THR	
IMP2	CG2-OG1-CA-HB	THR	
IMP3	HG23-CB-HG22-HG21	THR	
CHI1	N-CA-CB-CG	TRP	
CHI2	CA-CB-CG-CD1	TRP	
IMP1	CB-C-N-HA	TRP	
IMP2	CG-CA-HB2-HB1	TRP	
IMP3	CZ2-CE2-CD2-CE3	TRP	
IMP4	CH2-CZ2-CE2-NE1	TRP	
IMP5	CE2-CD2-CG-CB	TRP	
CHI1	N-CA-CB-CG	TYR	
CHI2	CA-CB-CG-CD1	TYR	
CHI3	CE2-CZ-OH-HH	TYR	
IMP1	CD1-CE1-CZ-OH	TYR	
IMP2	CE2-CD2-CG-CB	TYR	
IMP3	CZ-CE1-CD1-CG	TYR	
IMP4	CE2-CZ-CE1-CD1	TYR	
IMP5	CE1-CD1-CG-CD2	TYR	
CHI1	N-CA-CB-CG1	VAL	
CHI2	CA-CB-CG1-HG11	VAL	
CHI3	CA-CB-CG2-HG21	VAL	

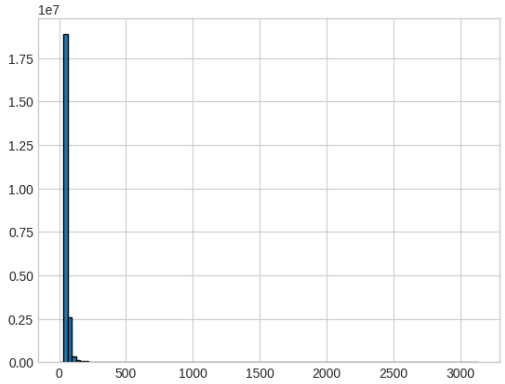
IMP1	CB-C-N-HA	VAL	
IMP2	CG2-CG1-CA-HB	VAL	
IMP3	HG13-CB-HG12-HG11	VAL	
IMP4	HG23-CB-HG22-HG21	VAL	

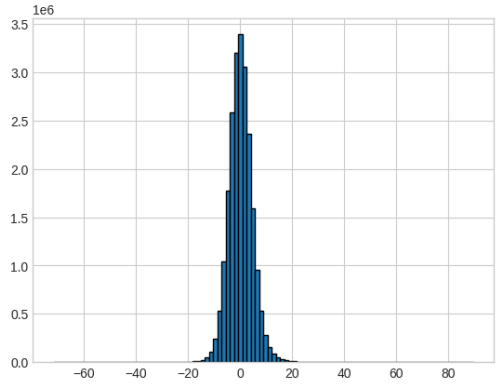
Appendix 5: Density and difference matrix comparison over resolutions

Below the difference matrices are shown as histograms to show the distributions of their densities and the density differences, where the density matrices have been standardised by comparison by my own method of median adjustment. The matrices are organised by resolution, and the highest atom is noted by each matrix for comparison of the maximum density.

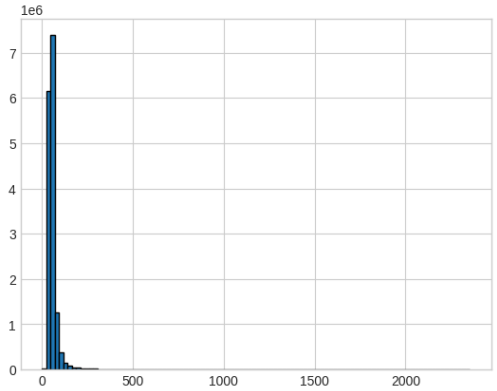
High resolution structures

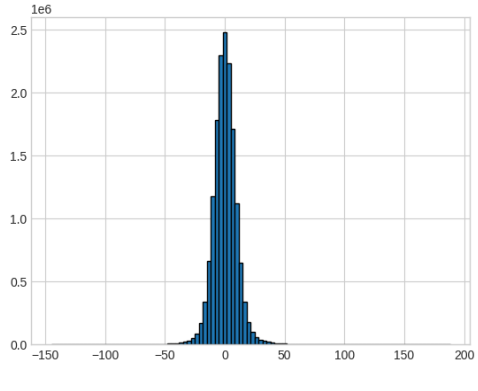
PDB=5D8V, Resolution =0.48, Max Atom = S



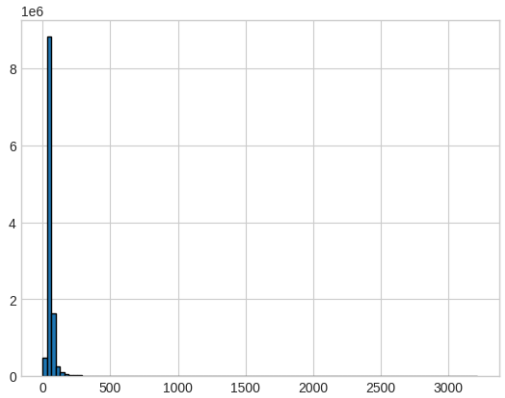


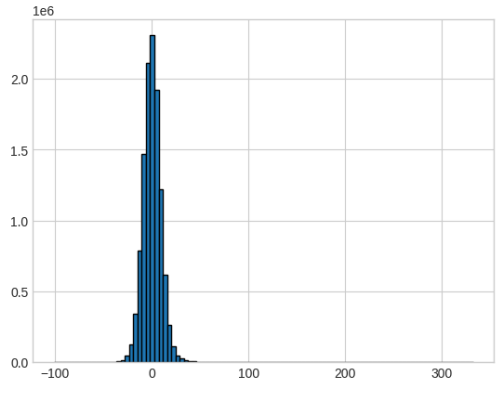
PDB=3NIR, Resolution=0.48, Max Atom=S



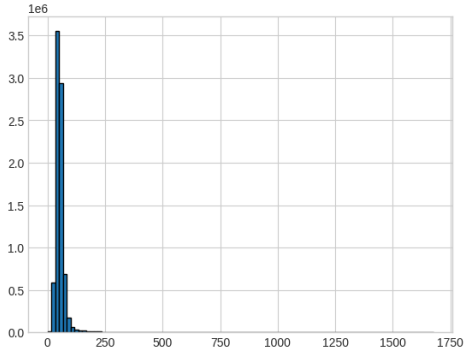


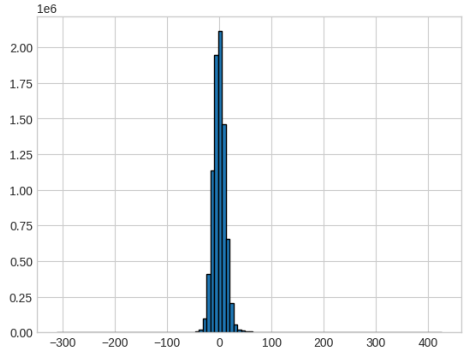
PDB=5NW3, Resolution =0.59, Max Atom =



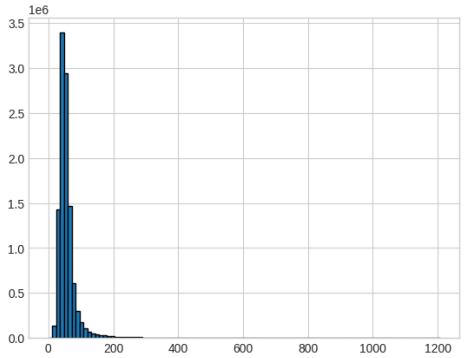


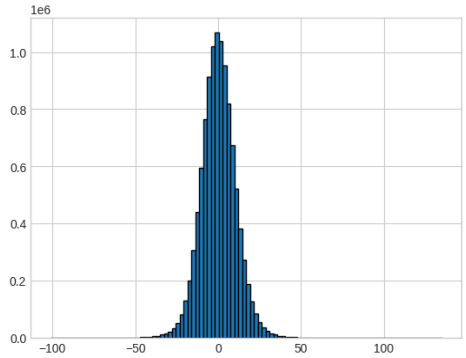
PDB=1EJG, Resolution =0.54, Max Atom =



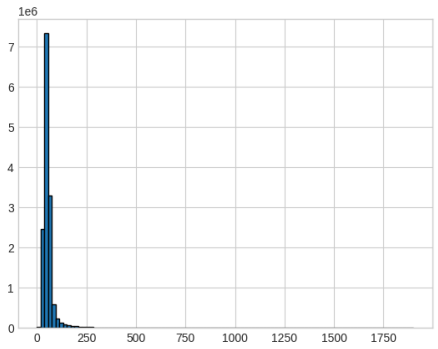


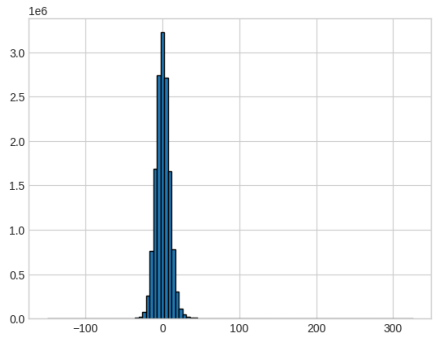
PDB=1ucs, Resolution =0.62, Max Atom =



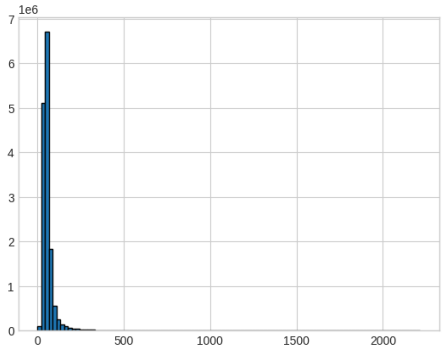


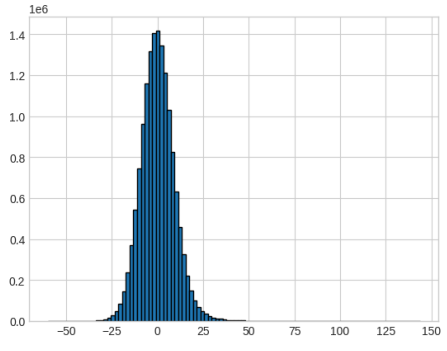
PDB=3X2M, Resolution =0.64, Max Atom =



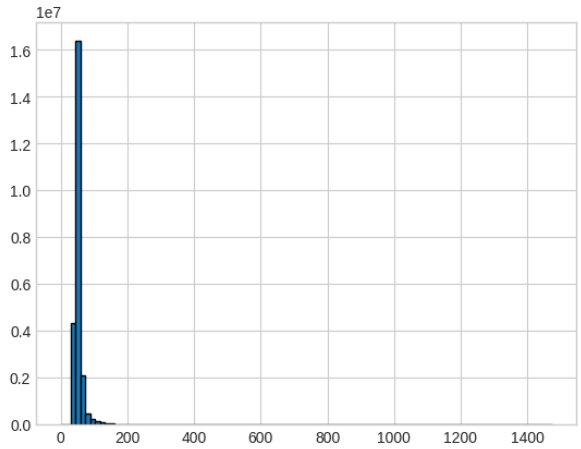


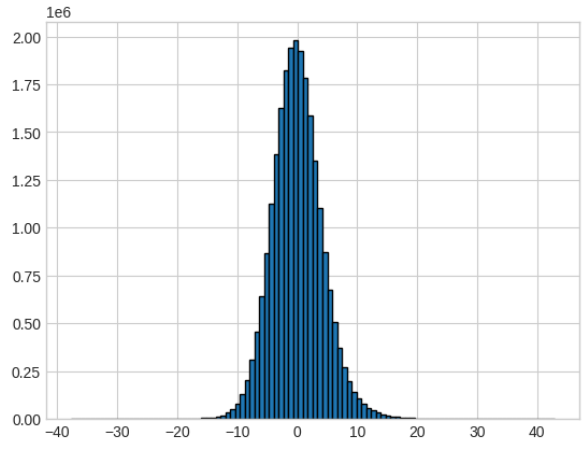
PDB=2VB1, Resolution =0.65, Max Atom =



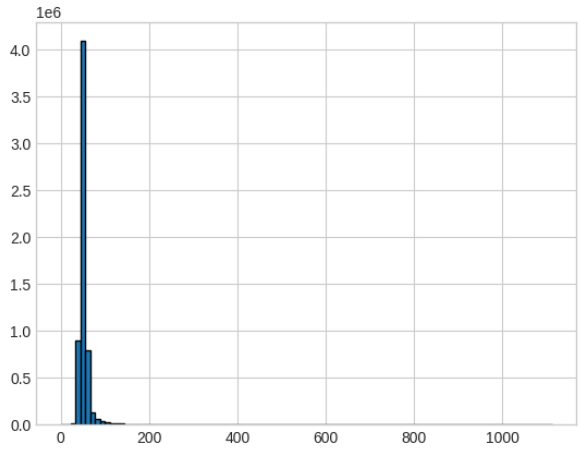


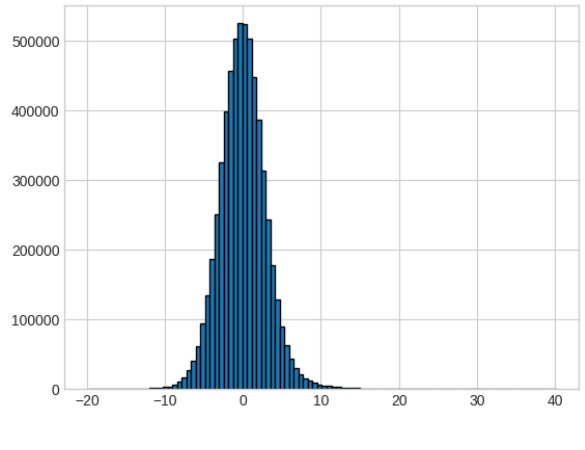
PDB=1us0, Resolution =0.66, Max Atom =



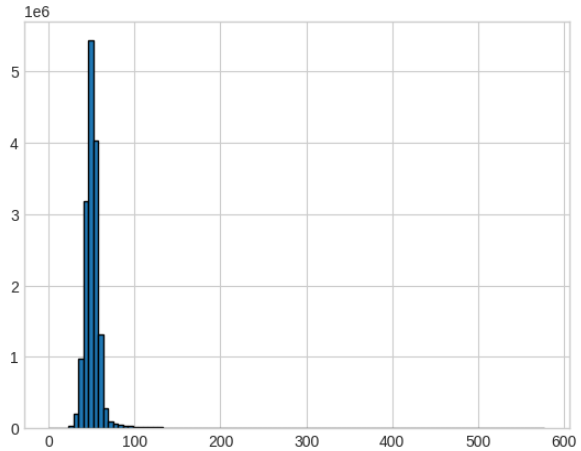


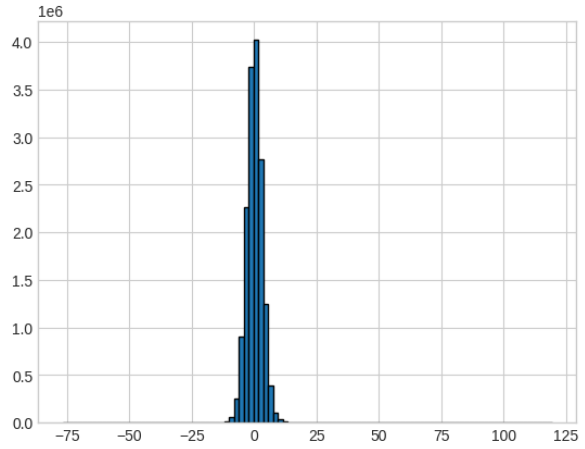
PDB=1yk4, Resolution =0.69, Max Atom =





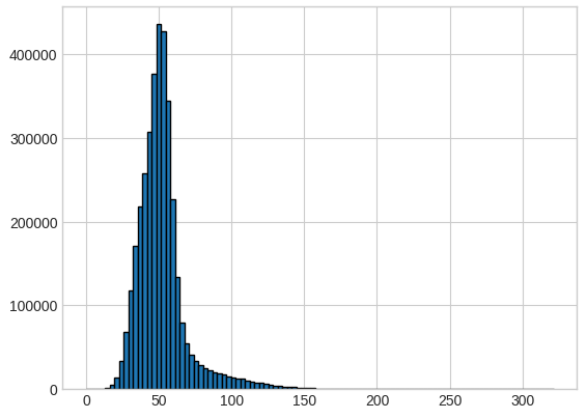
PDB=4ZM7, Resolution =0.70, Max Atom =

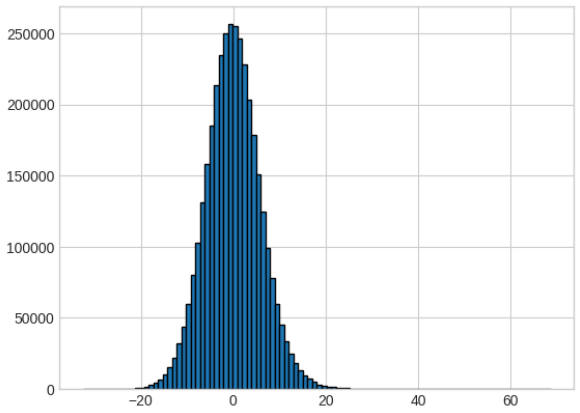




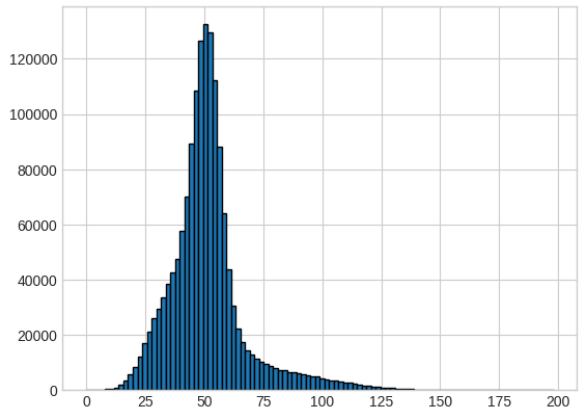
Middle resolution structures

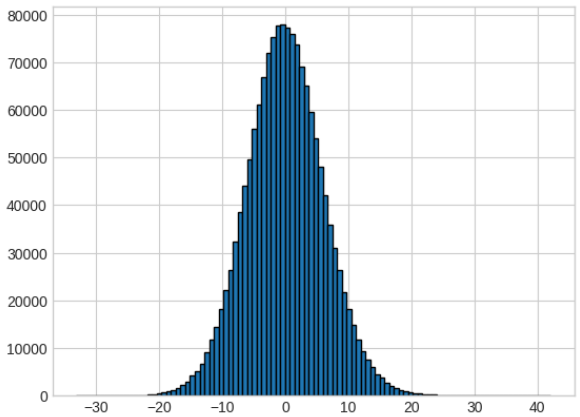
PDB=6jvv, Resolution =1.51



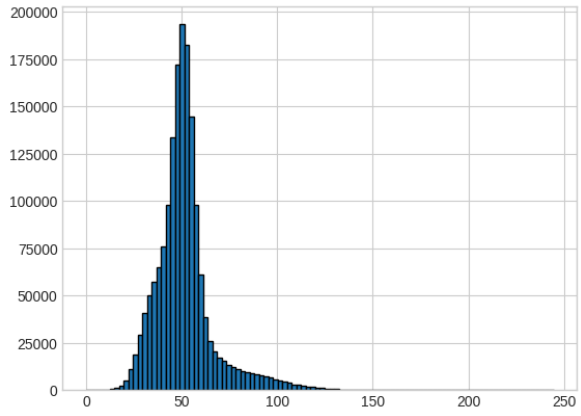


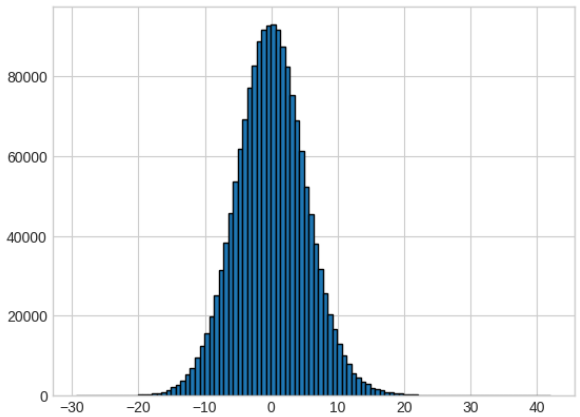
PDB=6jd0, Resolution=1.81



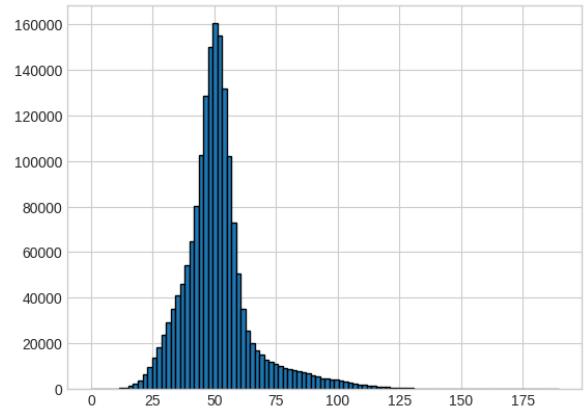


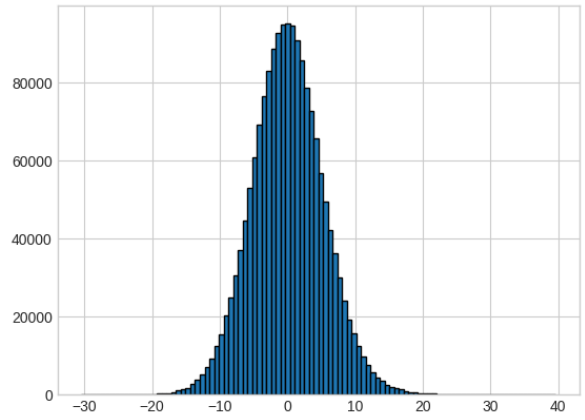
PDB=6pvz, Resolution =1.99



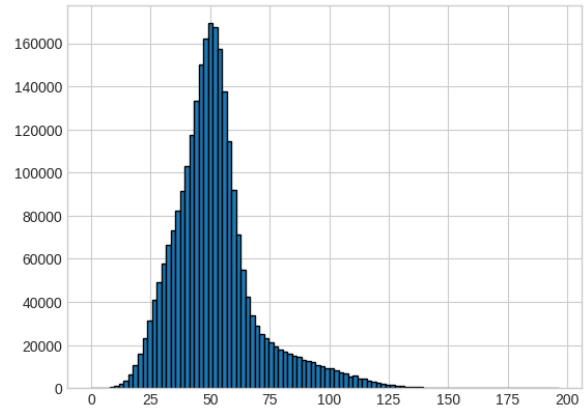


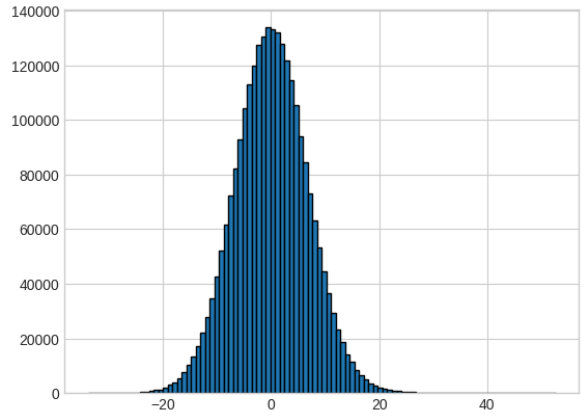
PDB=6o3x, Resolution =1.99



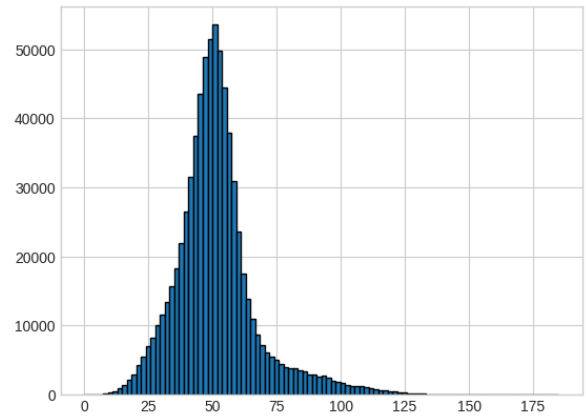


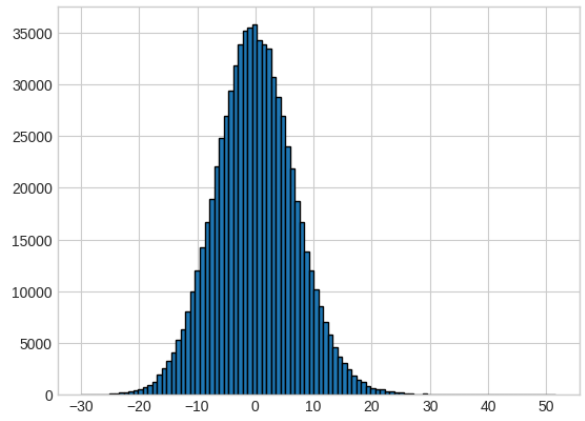
PDB=6nl4, Resolution =1.99



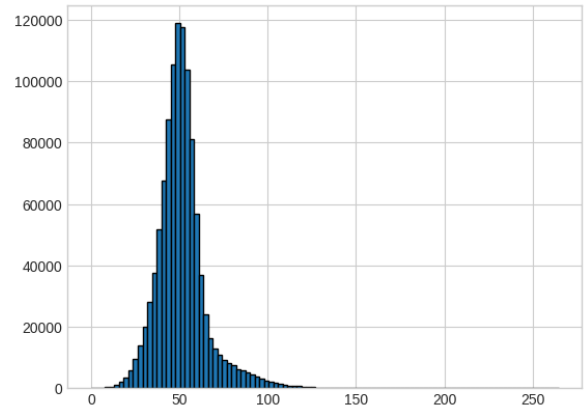


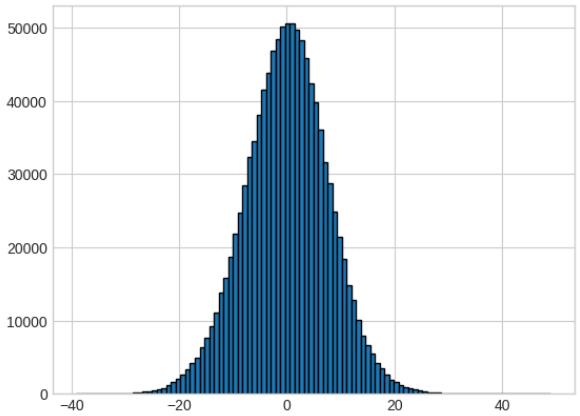
PDB=6rr2, Resolution =1.99



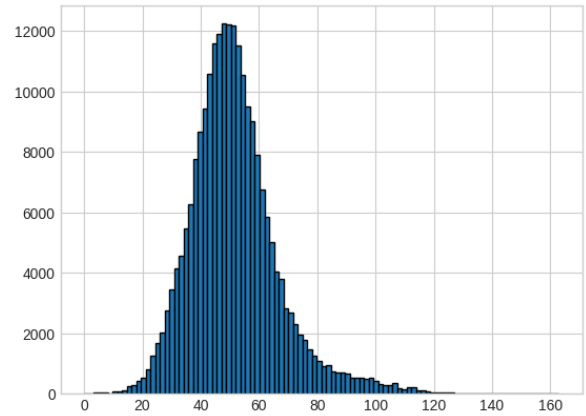


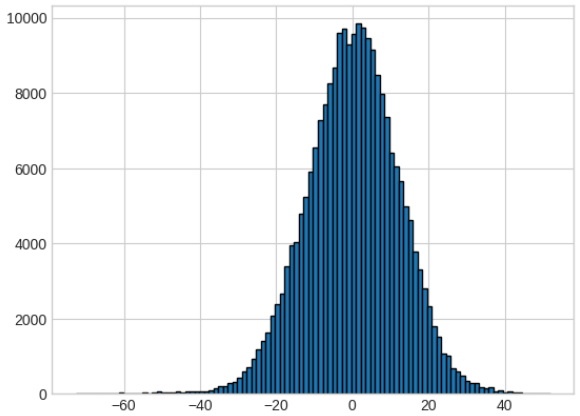
PDB=6rsl, Resolution =1.99



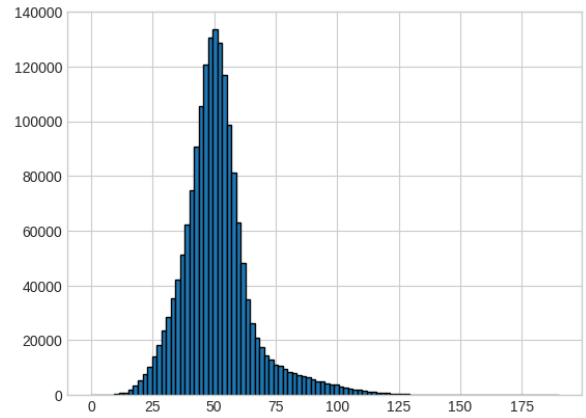


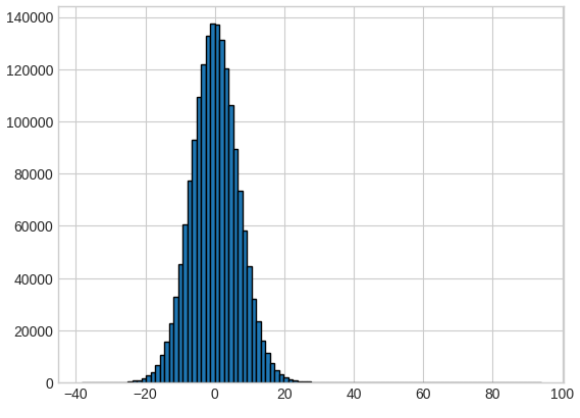
PDB=6shk, Resolution =1.99



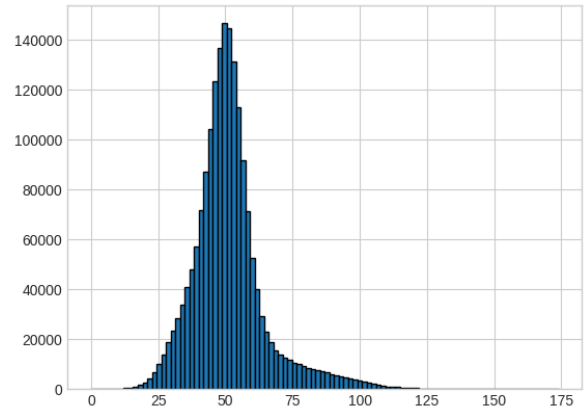


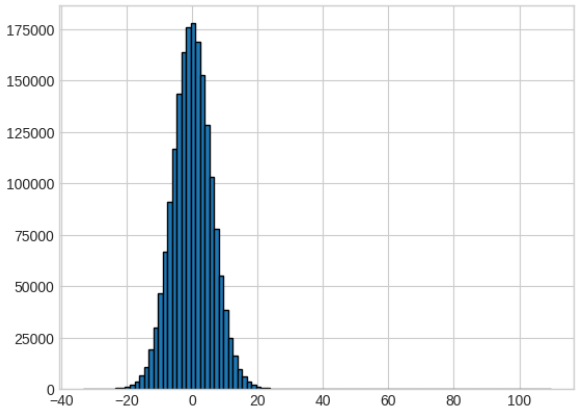
PDB=5qkw, Resolution =1.97





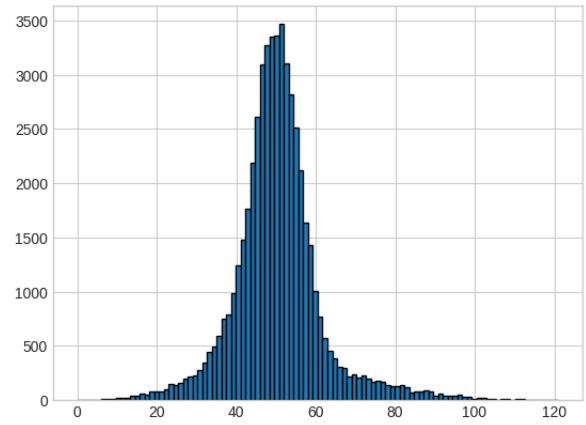
PDB=5ql3, Resolution =1.96

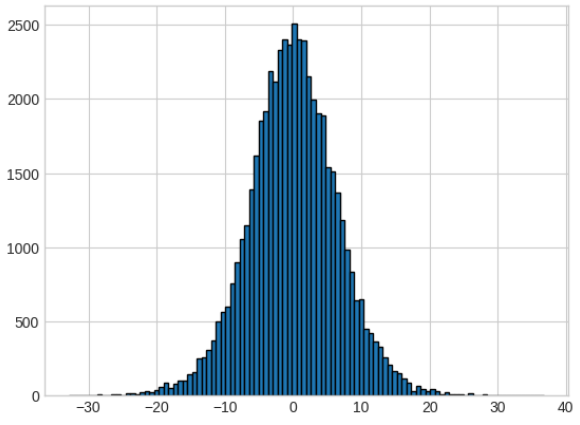




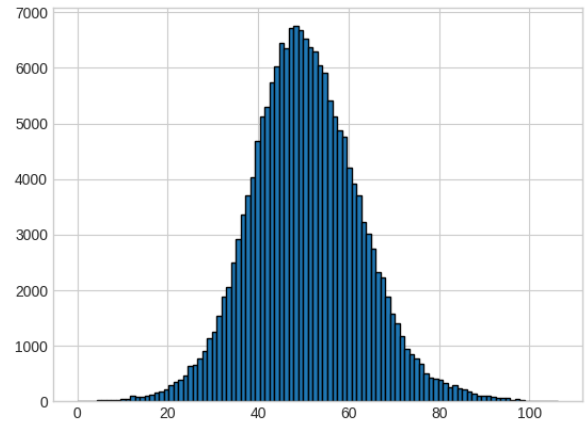
Low resolution structures

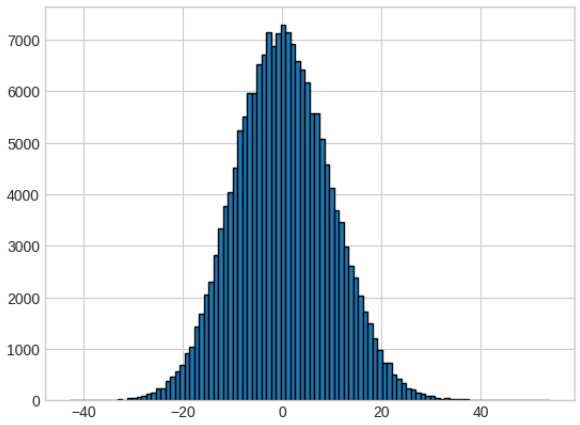
PDB=6fgz, Resolution =7.00



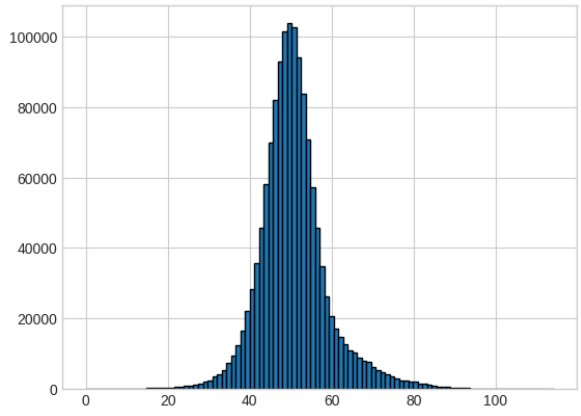


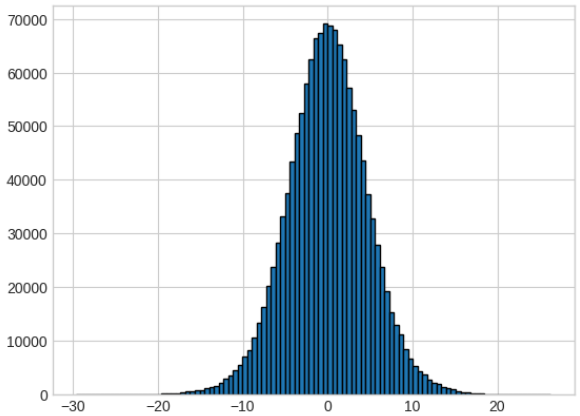
PDB=6ctd, Resolution=5.80



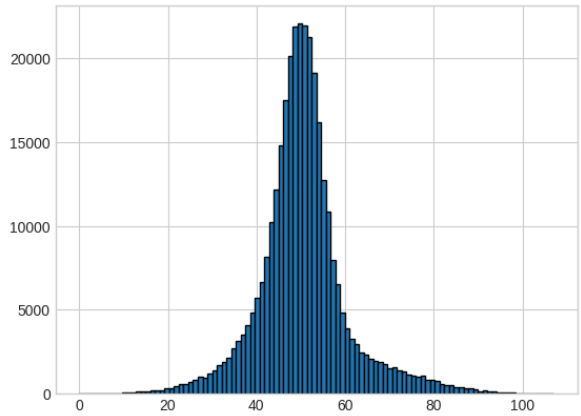


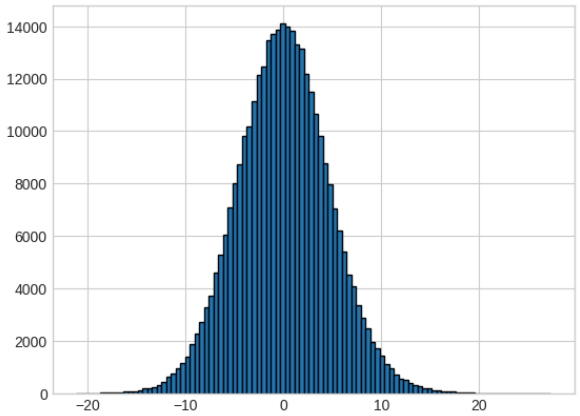
PDB=6e6b, Resolution =4.50



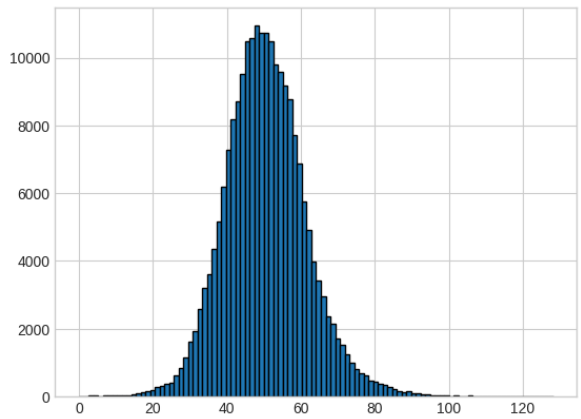


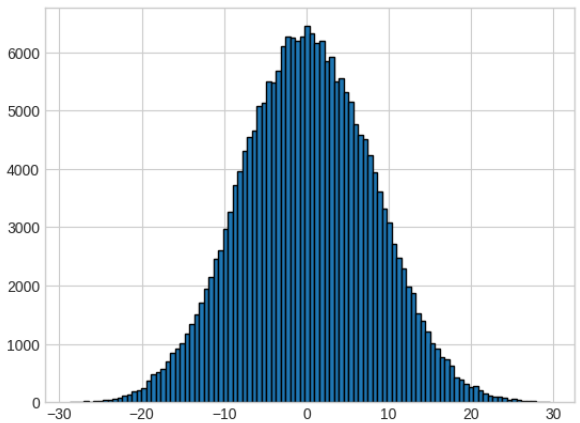
PDB=6nzi, Resolution =4.44



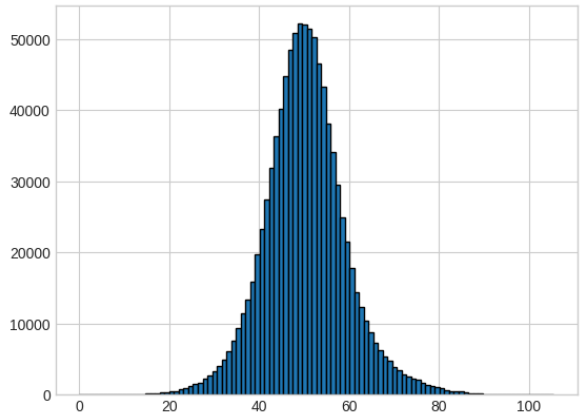


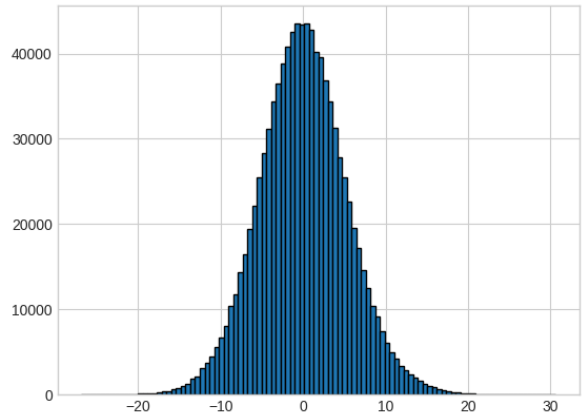
PDB=6fwf, Resolution =4.20



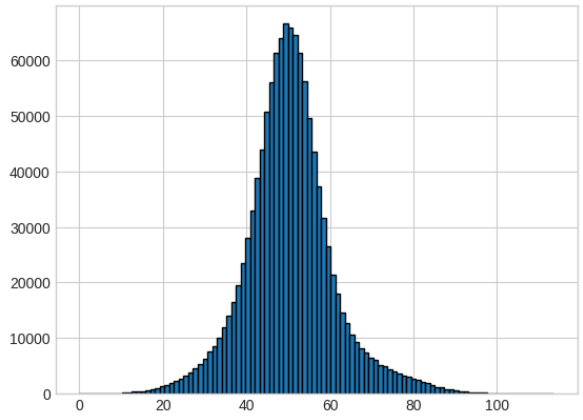


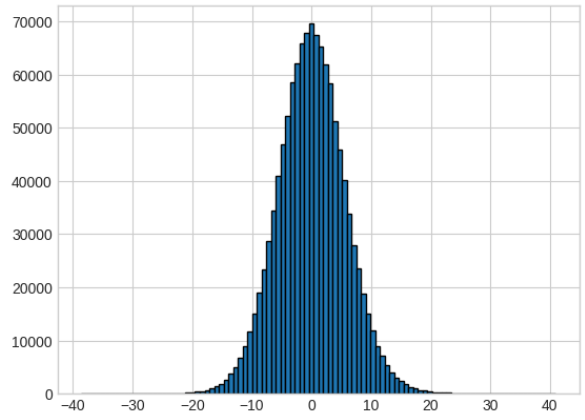
PDB=6j4a, Resolution =3.99



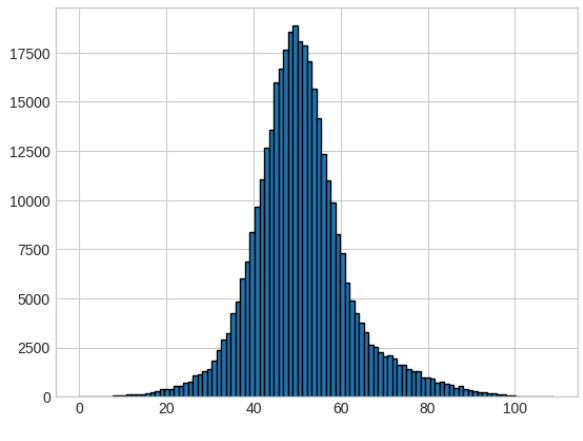


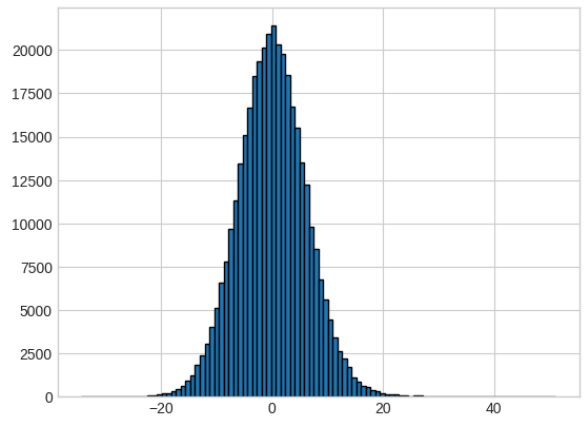
PDB=6j7g, Resolution =3.87



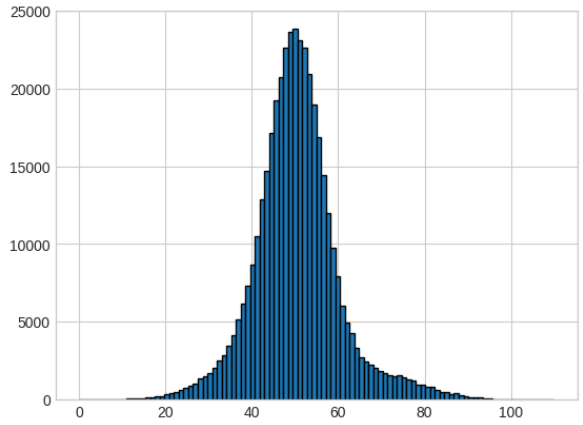


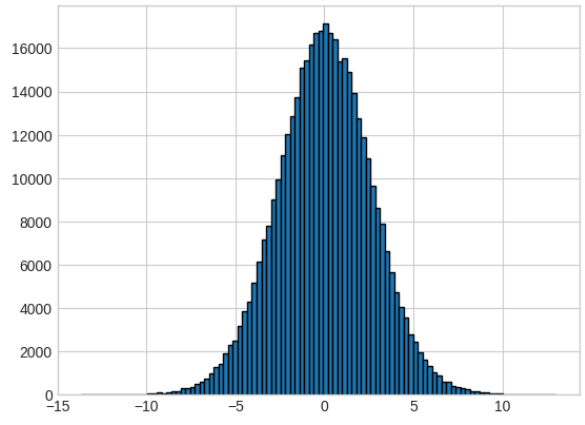
PDB=6je7, Resolution =3.90



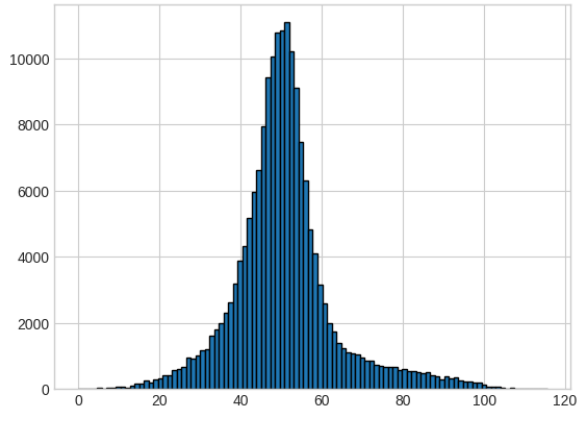


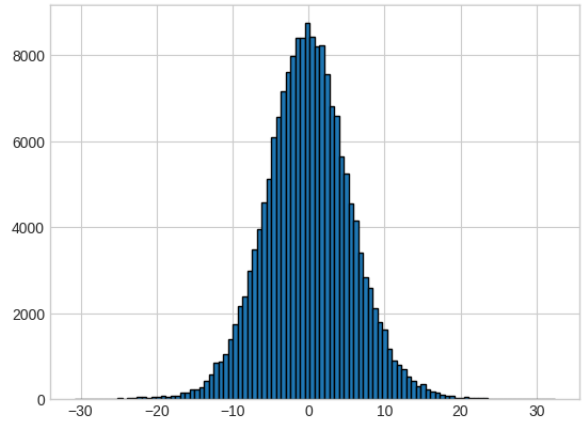
PDB=6l58, Resolution =3.90





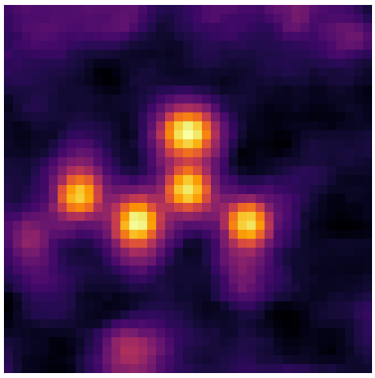
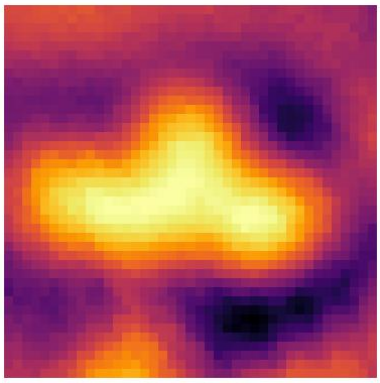
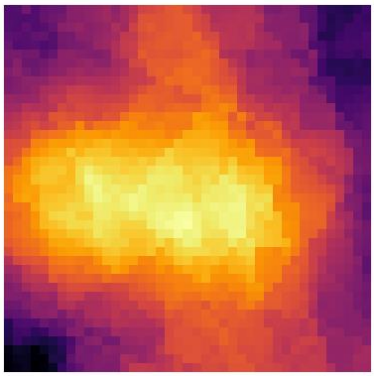
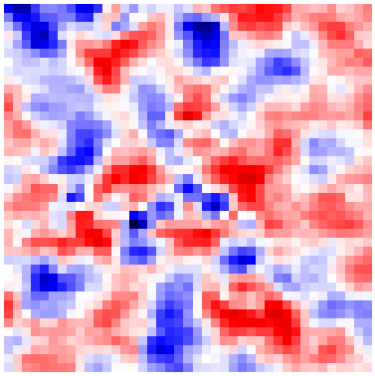
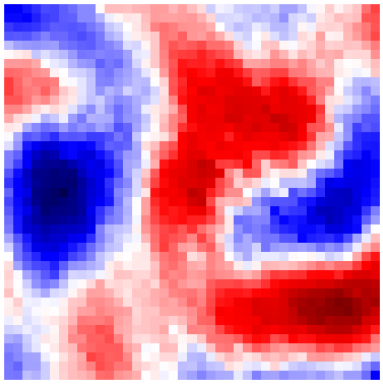
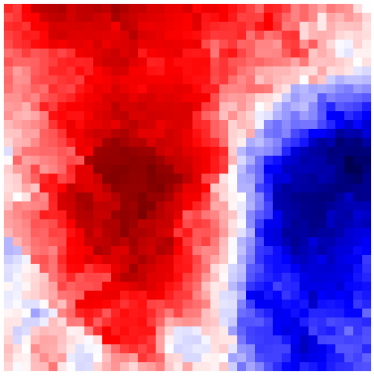
PDB=6q53, Resolution =3.70





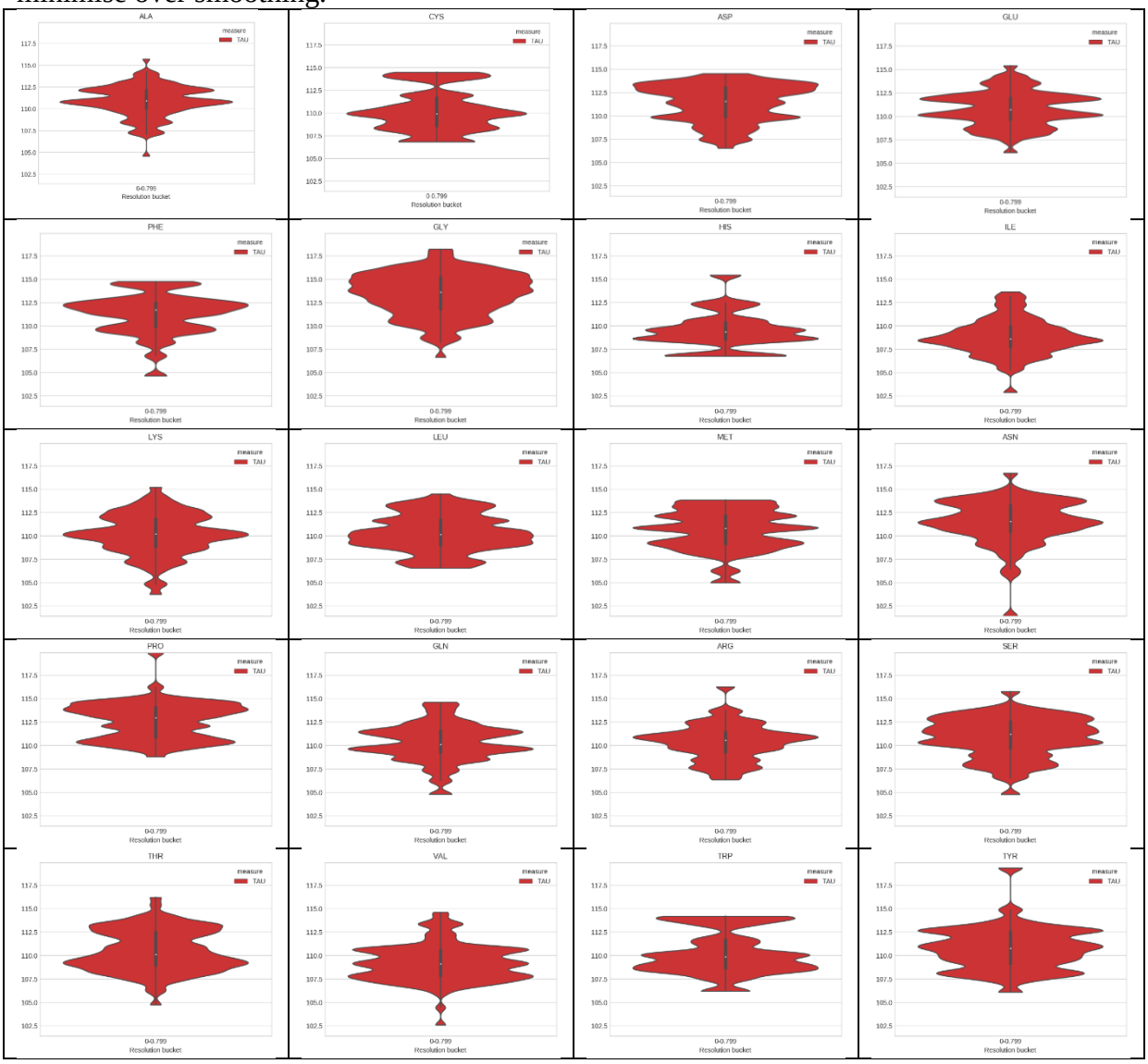
Appendix 6: Density and difference images at different resolutions

The images below are taken from electron density images overlayed. There is a clear difference in clarity at higher resolution.

High	Middle	Low
		
		

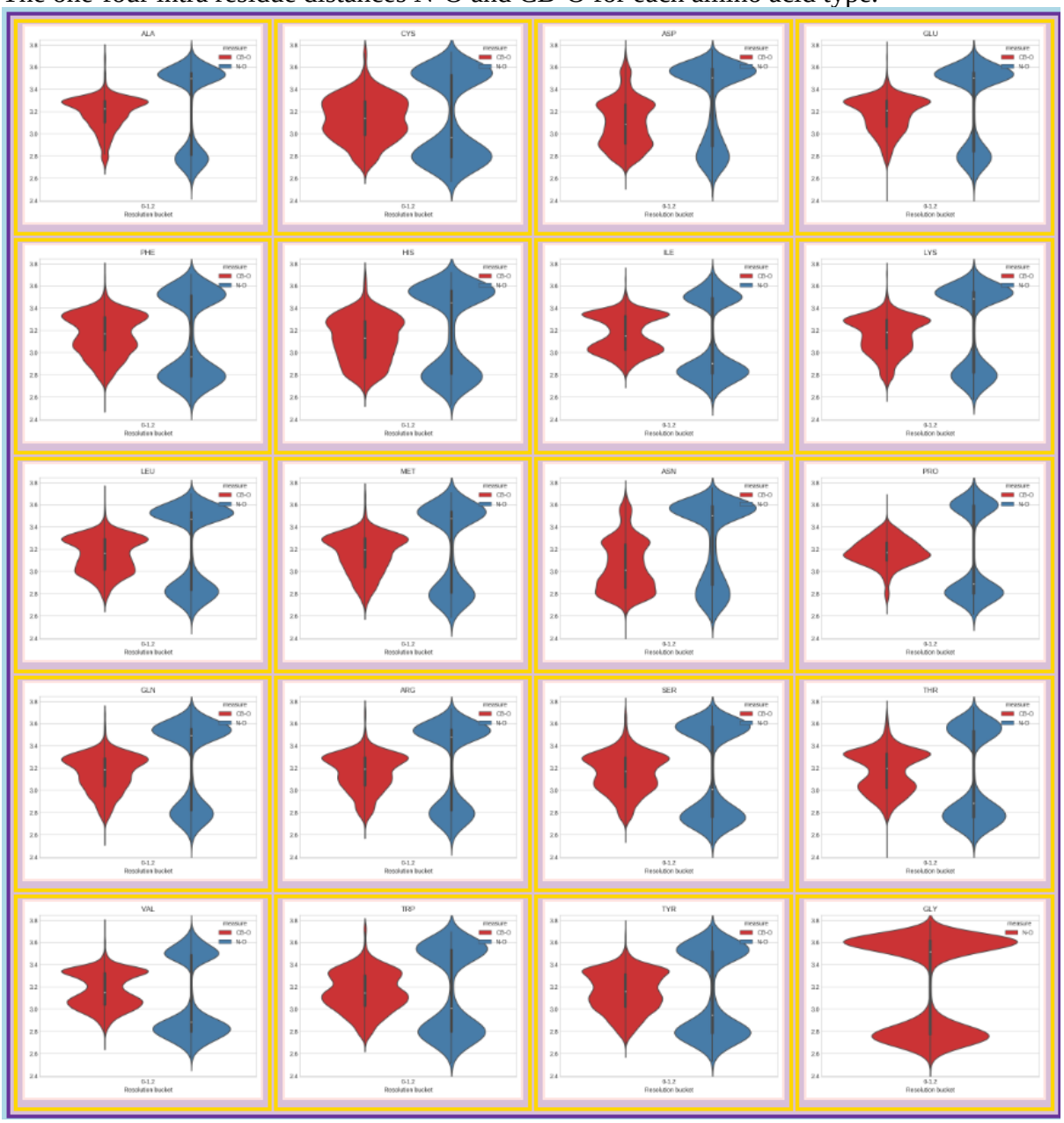
Appendix 7: Violin Plots for Tau

The violin plots below are for each amino acid, on the HQ set at < 0.8Å resolution. The violin plots provide a visualisation for the different distributions for each amino acid. KDE smoothing in seaborn violinplot(kde=0.15). They are evidently different, the kde smoothing is low to minimise over smoothing.



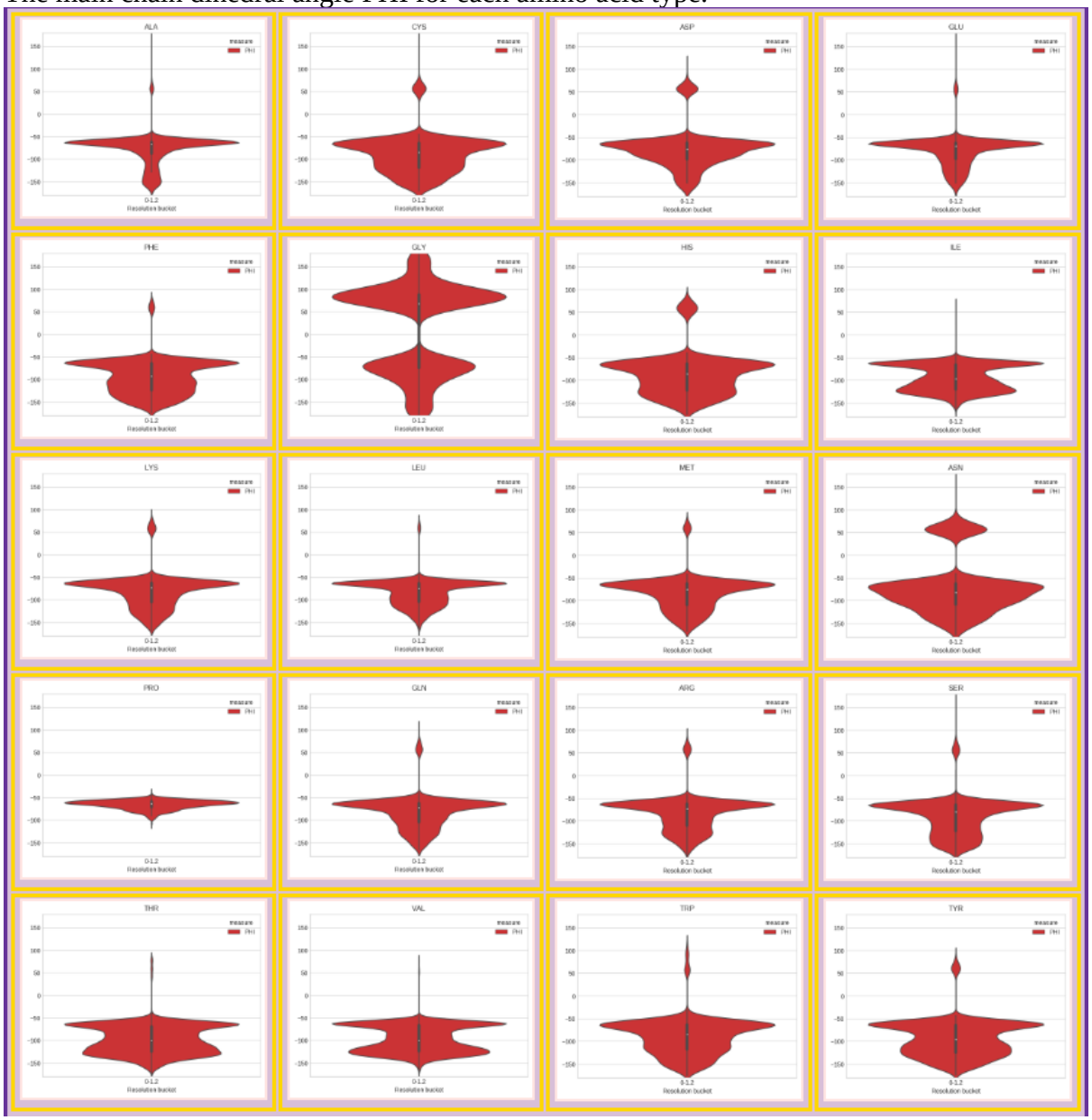
Appendix 8: The bimodal nature of N-O, and CB-O

The one-four intra residue distances N-O and CB-O for each amino acid type.



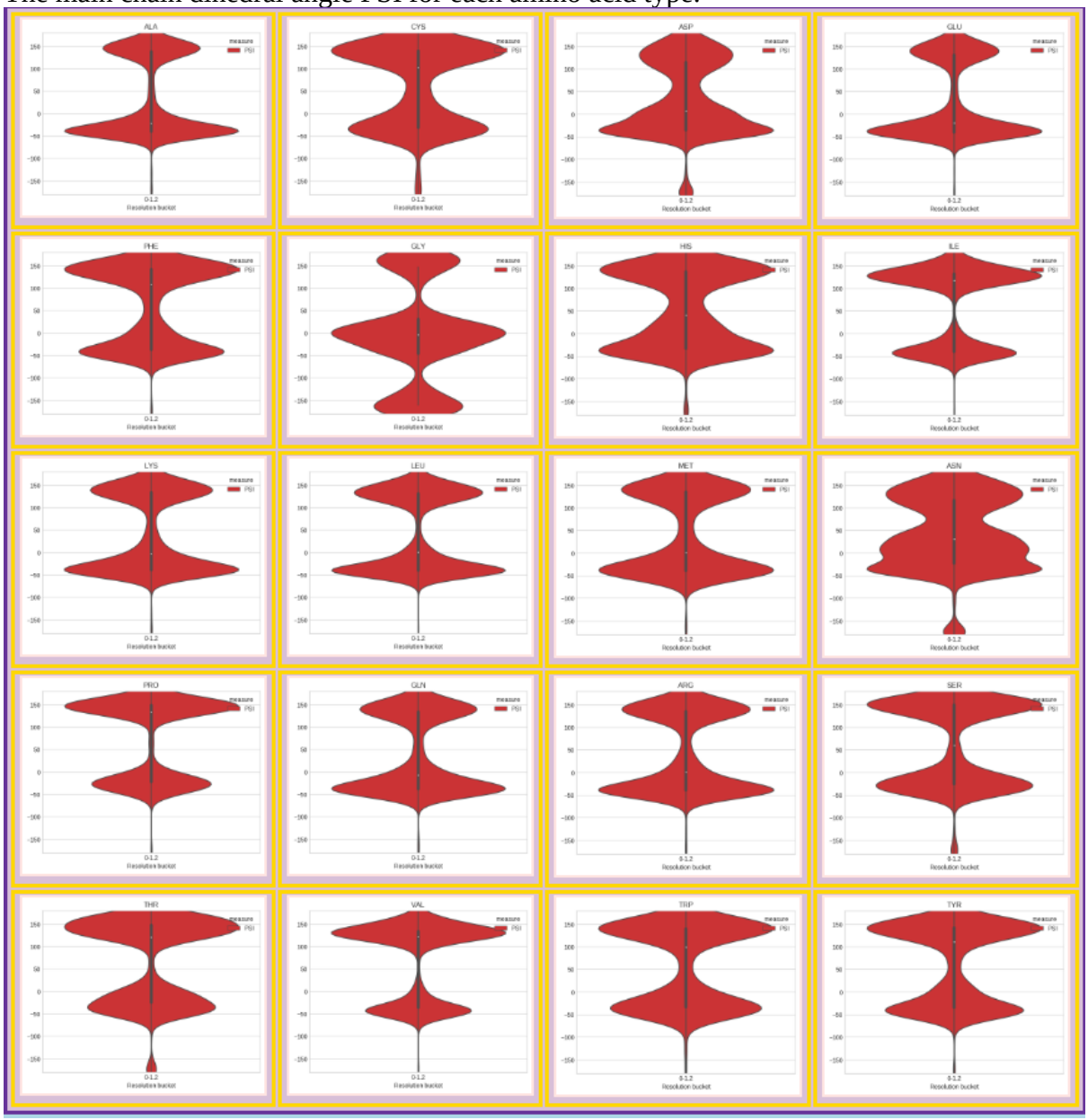
Appendix 9: PHI distributions per amino acid

The main chain dihedral angle PHI for each amino acid type.



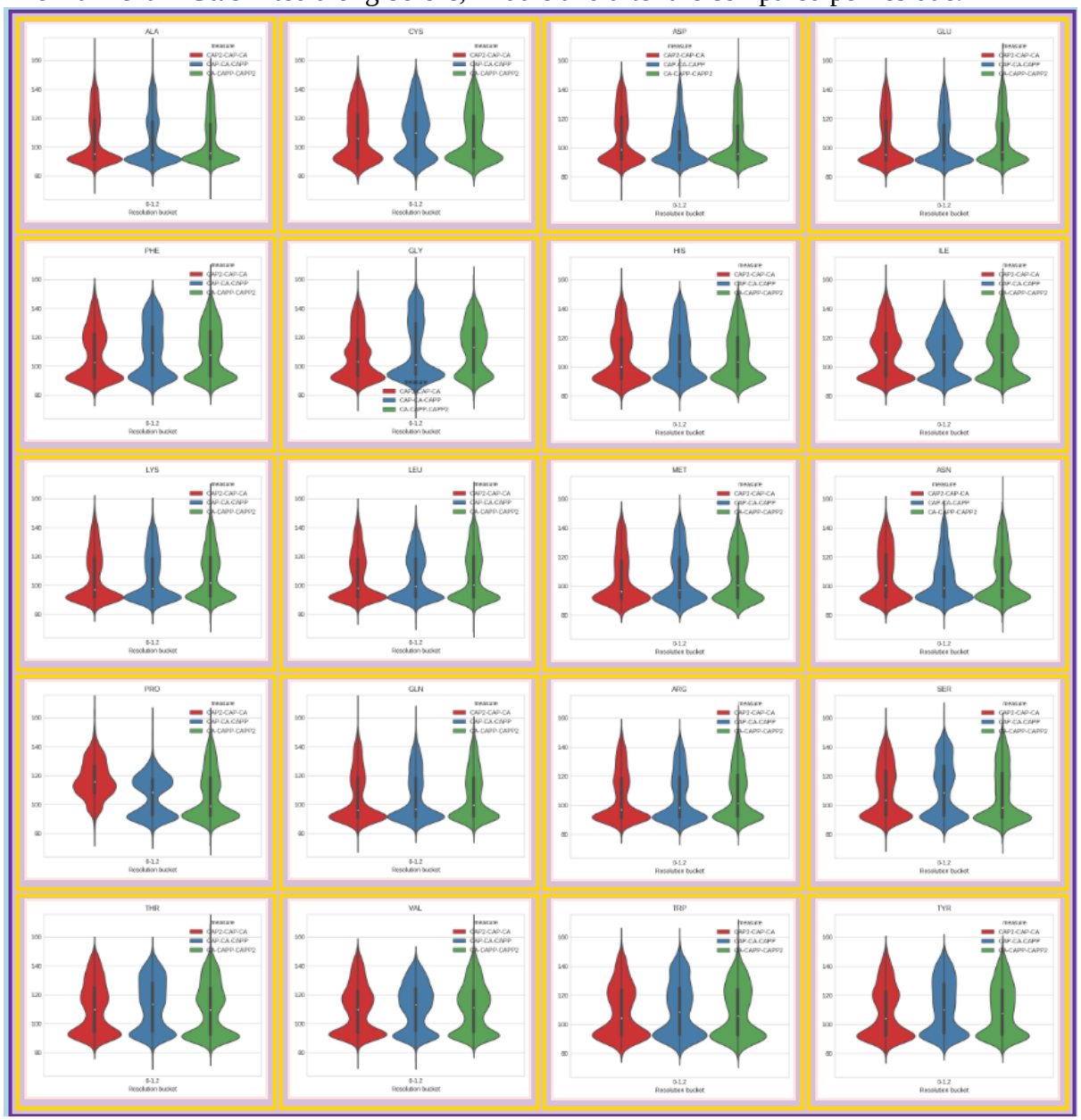
Appendix 10: PSI distributions per amino acid

The main chain dihedral angle PSI for each amino acid type.



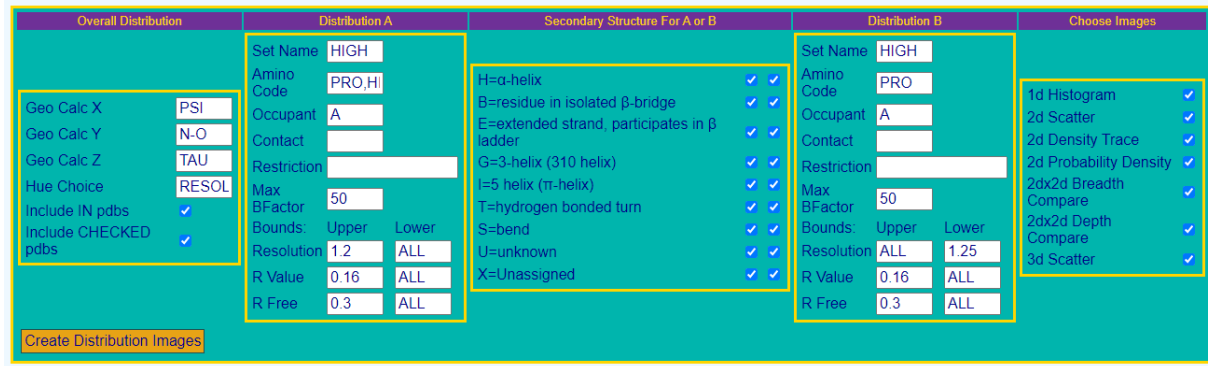
Appendix 11: Comparing Cα angles along the chain

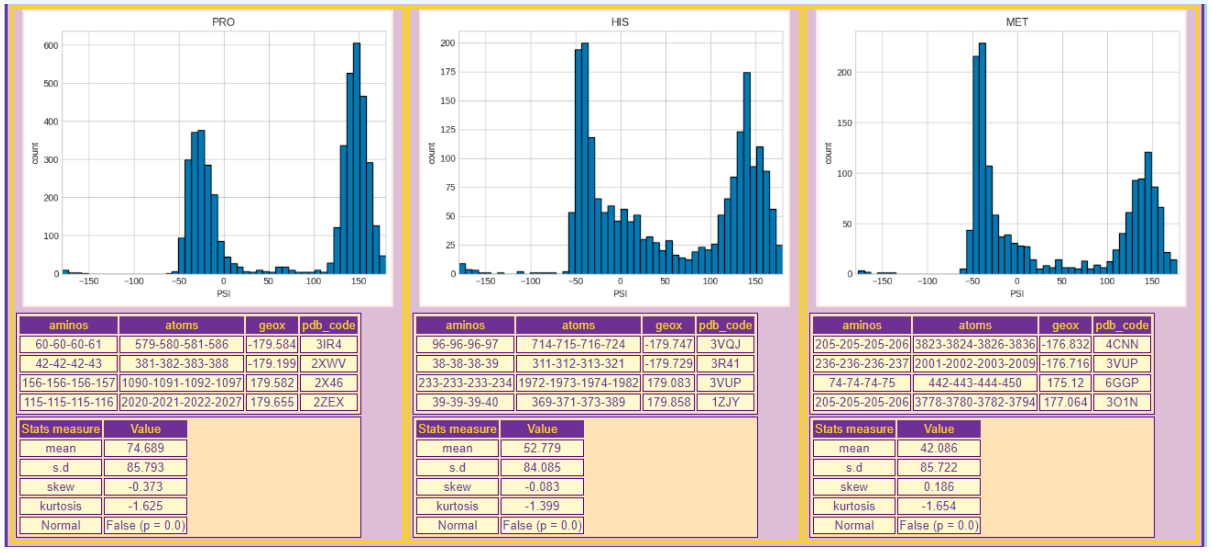
The main chain Cα shifted along before, middle and after are compared per residue.



Appendix 12: Distribution reports from website

The images below are a single page from the Distributions page given all views were checked.





aminos	atoms	geox	pdb_code
60-60-60-61	579-580-581-586	-179.584	3IR4
42-42-42-43	381-382-383-388	-179.199	2XWV
156-156-156-157	1090-1091-1092-1097	179.582	2X46
115-115-115-116	2020-2021-2022-2027	179.655	2ZEX

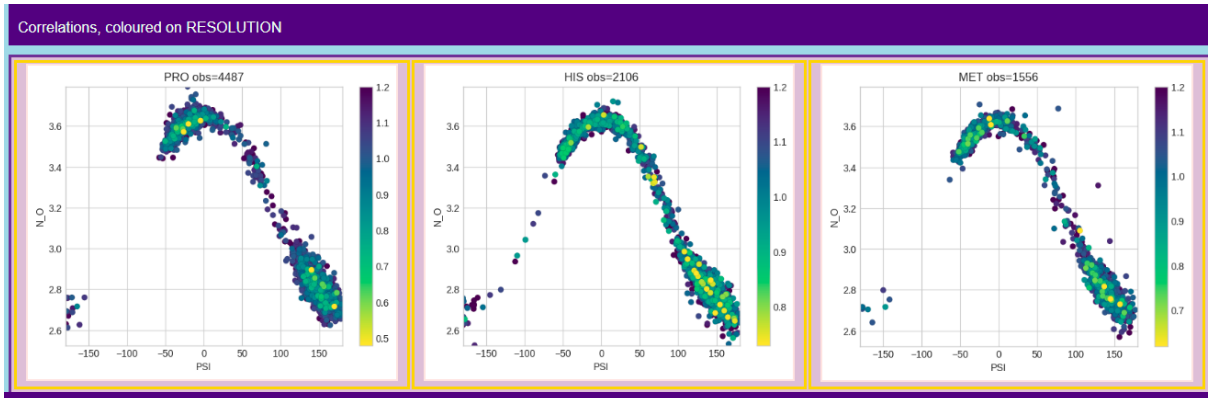
Stats measure	Value
mean	74.689
s.d	85.793
skew	-0.373
kurtosis	-1.625
Normal	False (p = 0.0)

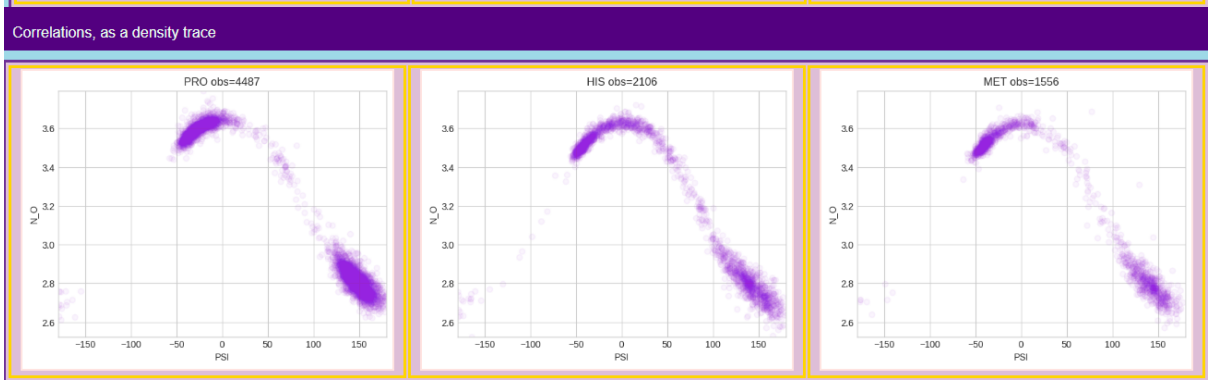
aminos	atoms	geox	pdb_code
96-96-96-97	714-715-716-724	-179.747	3VQJ
38-38-38-39	311-312-313-321	-179.729	3R41
233-233-233-234	1972-1973-1974-1982	179.083	3VUP
39-39-39-40	369-371-373-389	179.858	1ZJY

Stats measure	Value
mean	52.779
s.d	84.085
skew	-0.083
kurtosis	-1.399
Normal	False (p = 0.0)

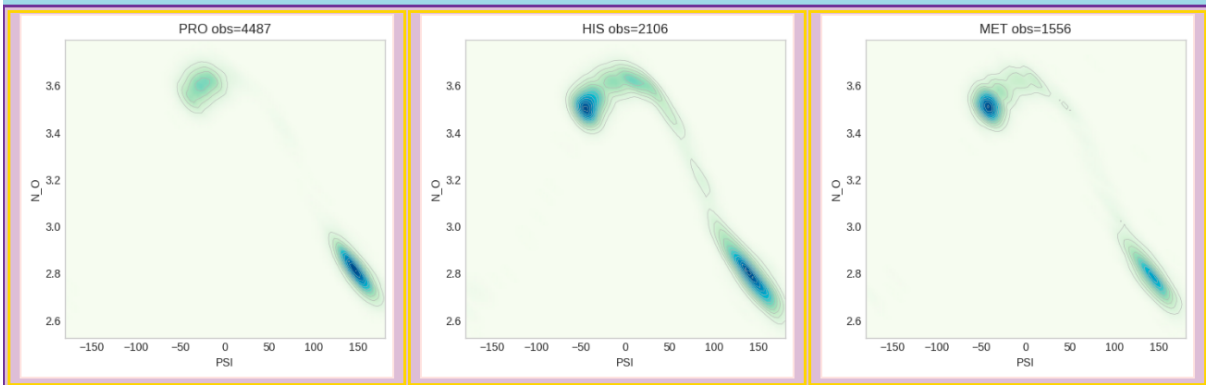
aminos	atoms	geox	pdb_code
205-205-205-206	3823-3824-3826-3836	-176.832	4CNN
236-236-236-237	2001-2002-2003-2009	-176.716	3VUP
74-74-74-75	442-443-444-450	175.12	6GGP
205-205-205-206	3778-3780-3782-3794	177.064	3O1N

Stats measure	Value
mean	42.086
s.d	85.722
skew	0.186
kurtosis	-1.654
Normal	False (p = 0.0)

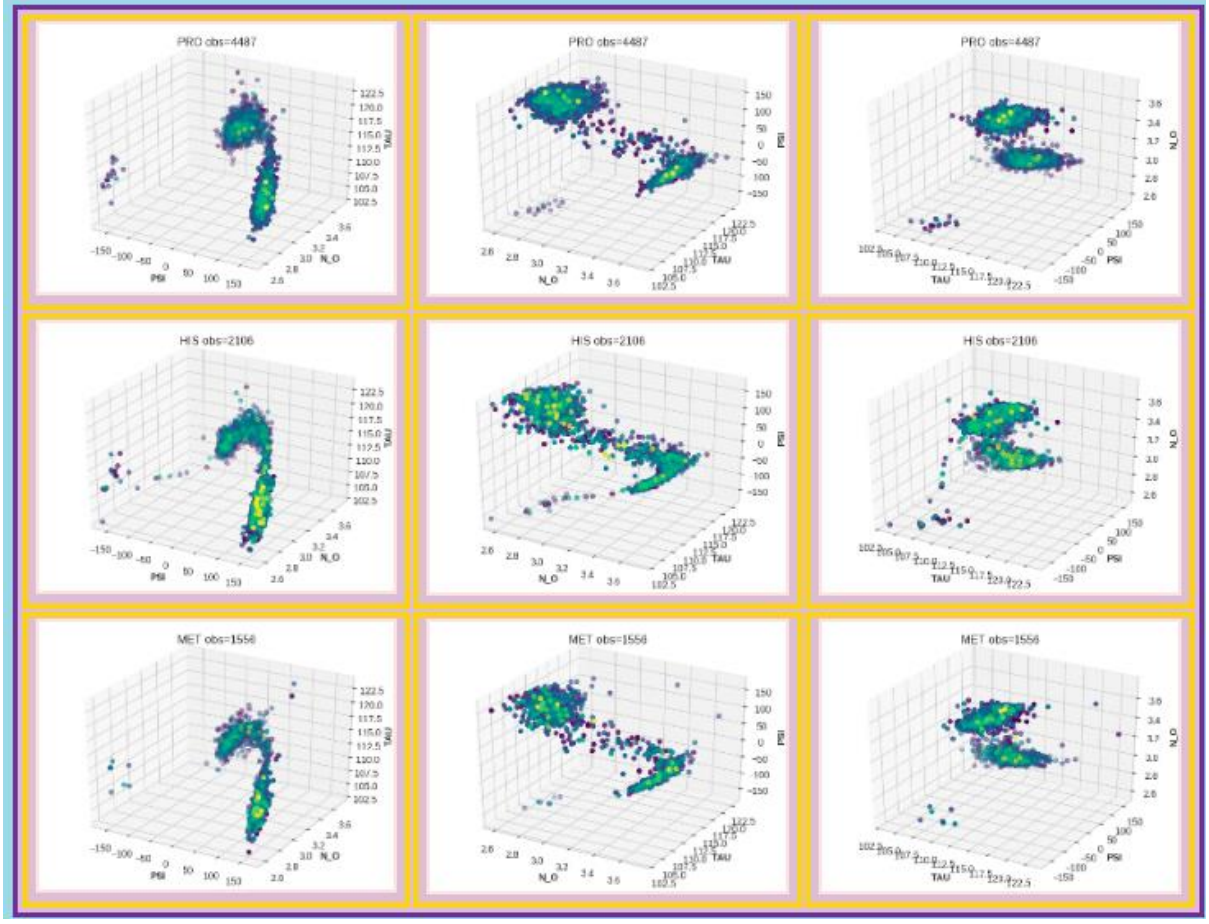


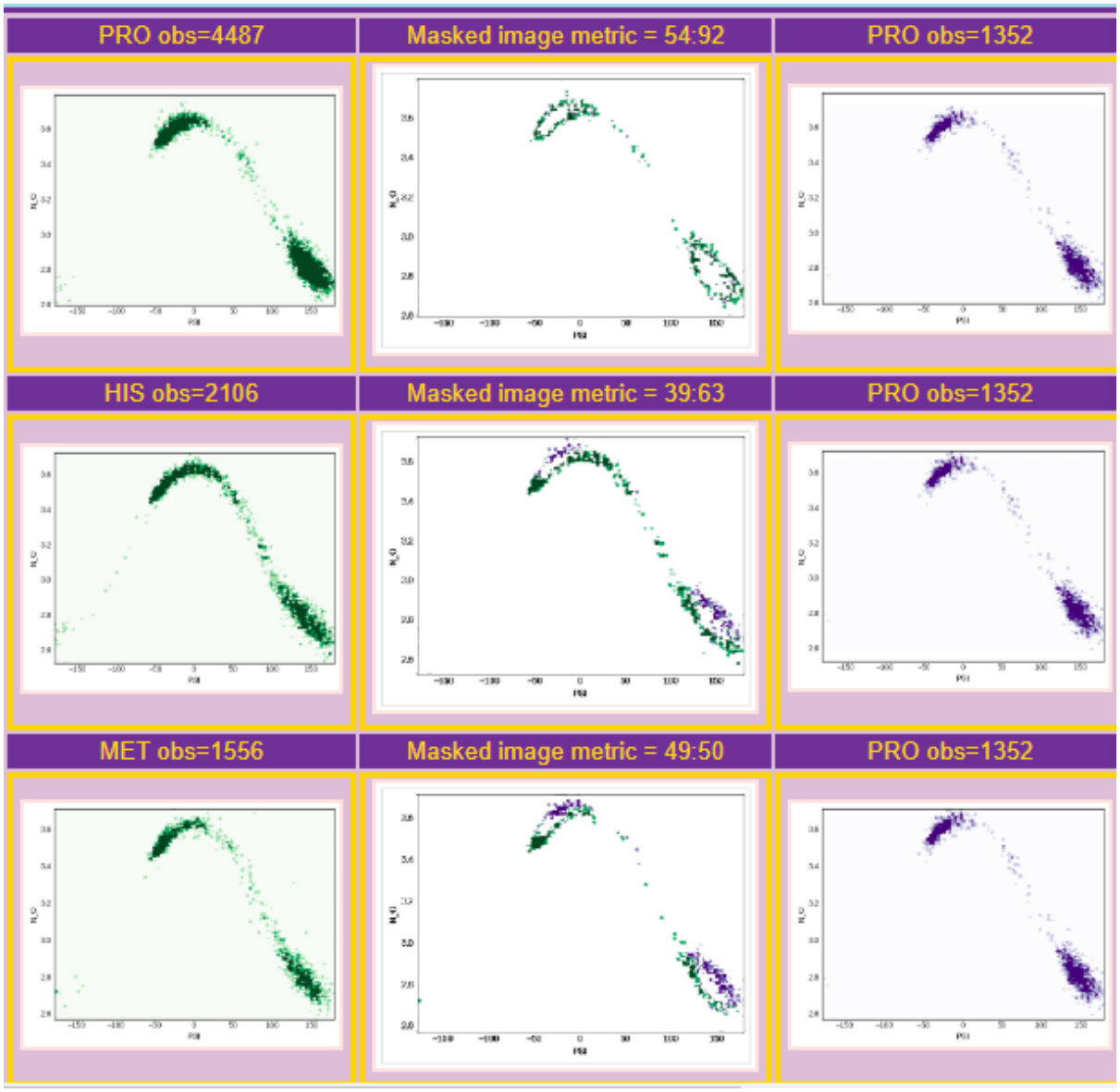


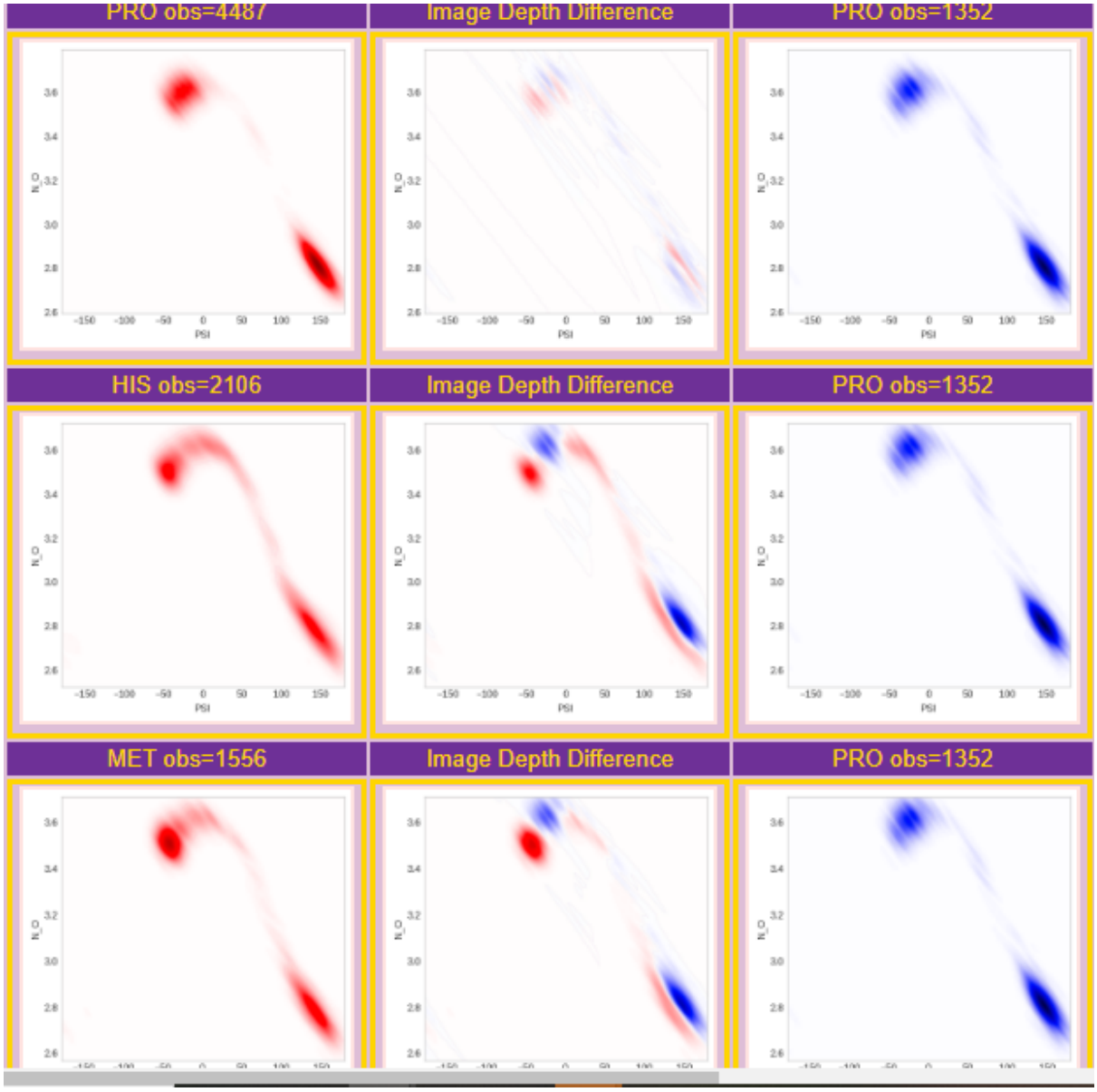
Correlations, as a probability density plot



Graduated on RESOLUTION



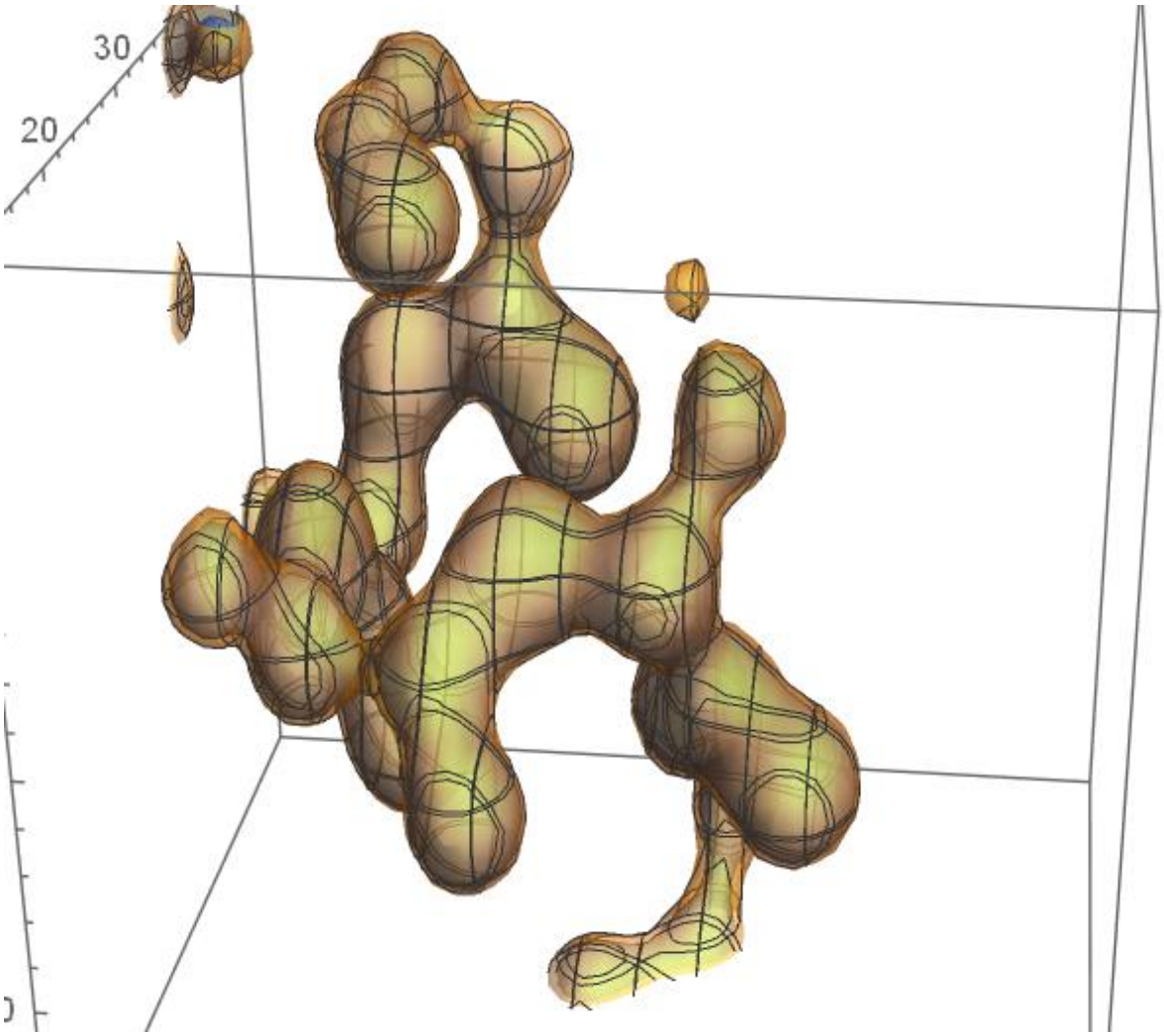

PRO obs=4487
Masked image metric = 54:92
PRO obs=1352
HIS obs=2106
Masked image metric = 39:63
PRO obs=1352
MET obs=1556
Masked image metric = 49:50
PRO obs=1352


PRO obs=4487
Image Depth Difference
PRO obs=1352
HIS obs=2106
Image Depth Difference
PRO obs=1352
MET obs=1556
Image Depth Difference
PRO obs=1352
N_O
PSI

Appendix 13: 146 GLN Superposition

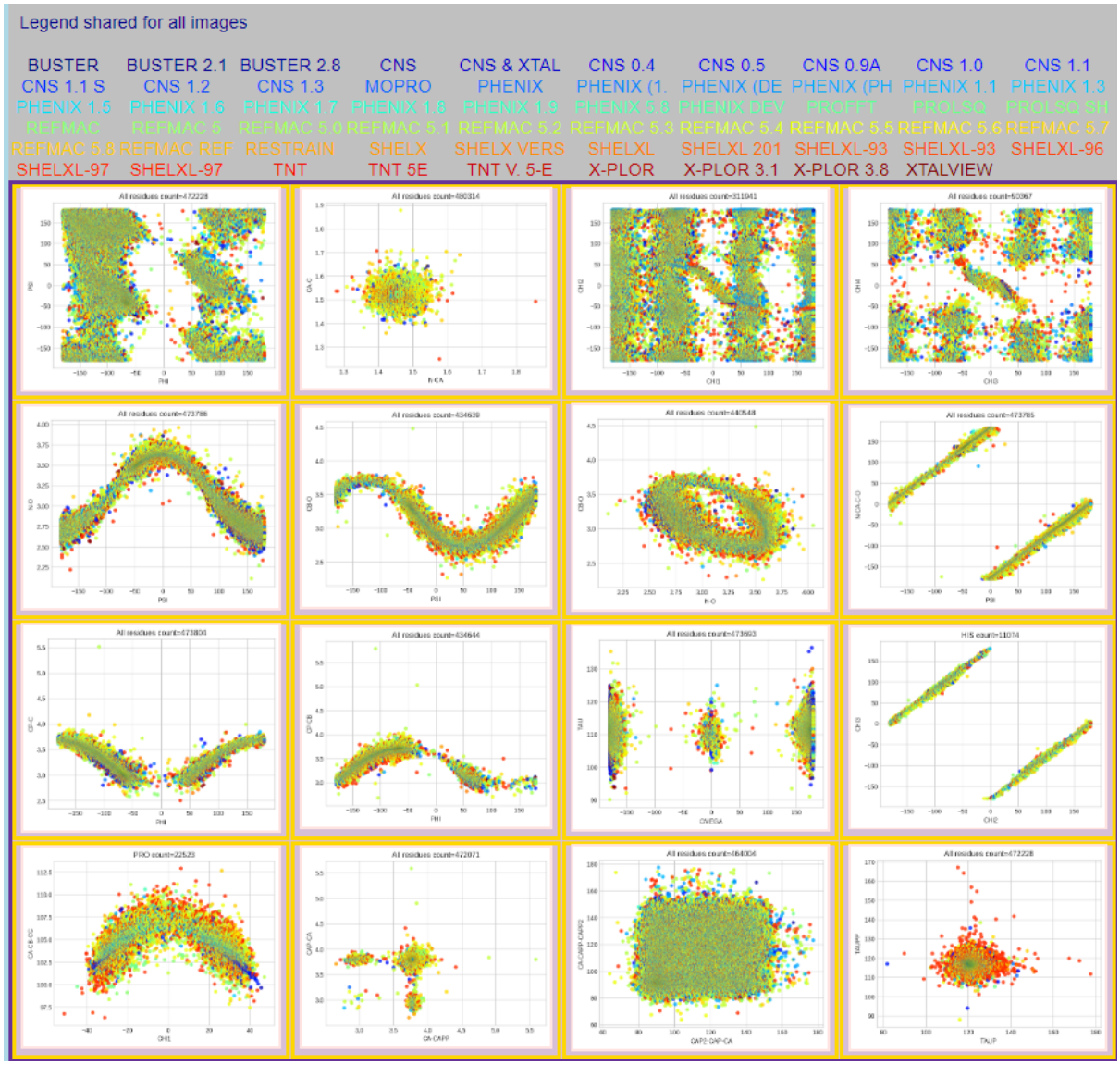
146 glutamine residues are overlayed, chosen on the close contact between N and O of residue i+-3.

Only 3 residues are i+3, the remaining are at i-3. Those at plus have also close contact at minus.



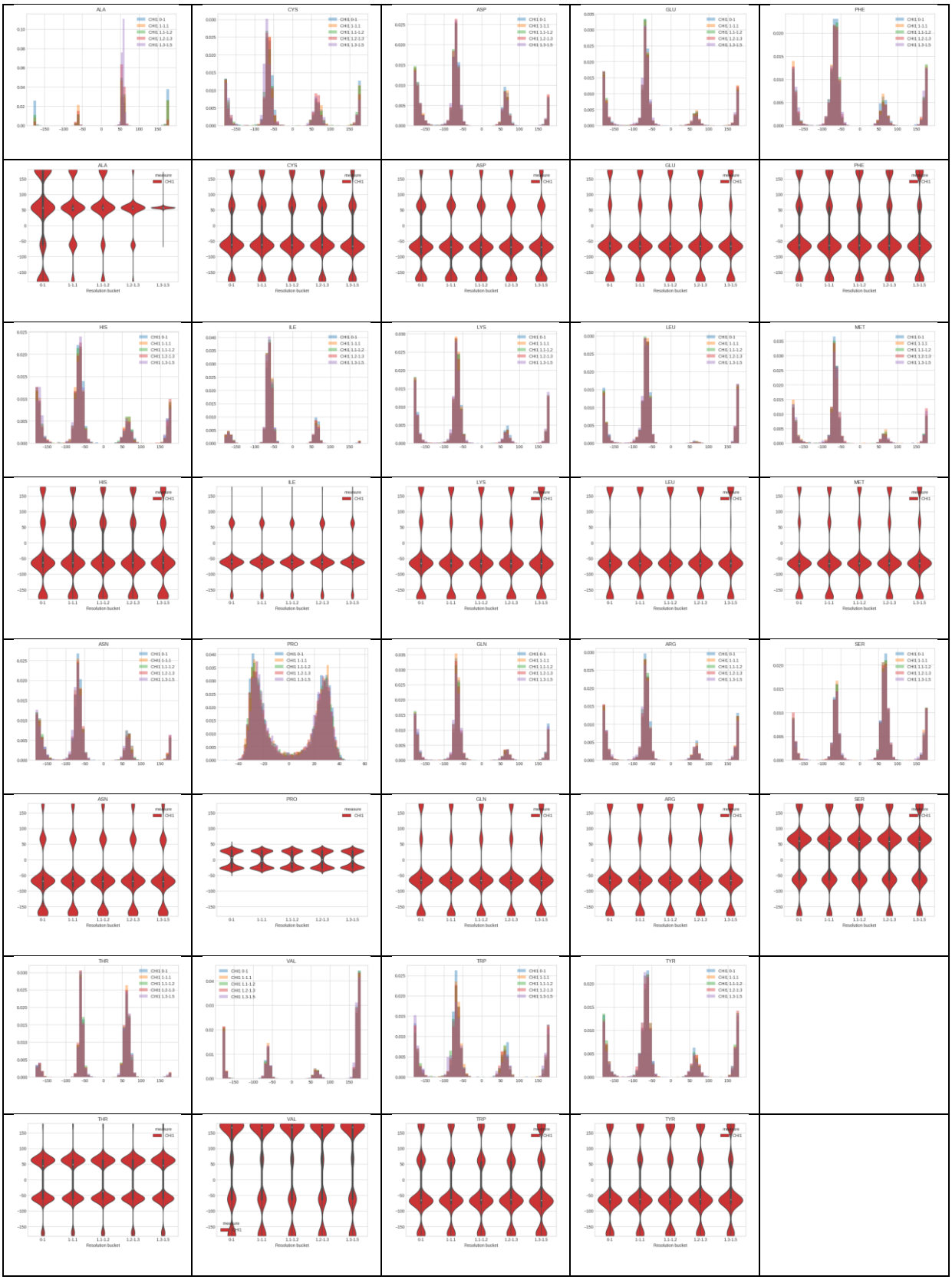
Appendix 14: Correlation page for all HIGH residues, on refinement method

The image below gives some indication of the influence of refinement software on final structure geometry.



Appendix 15: CHI1 for different resolutions buckets

The residues were chosen with all values unrestrained within each resolution band.



Appendix 16: Summary statistics for CHI1 distributions

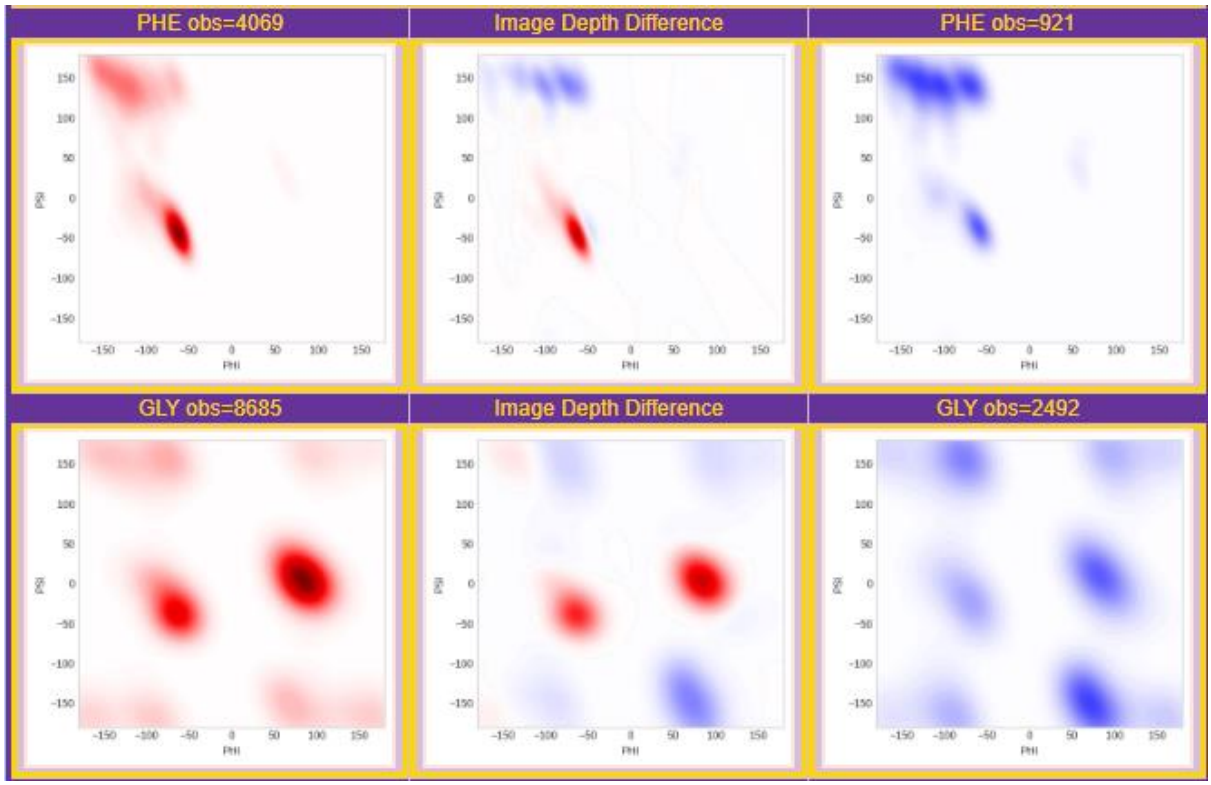
CHI1 summary statistics at different resolution buckets, including observation count, all values unrestrained.

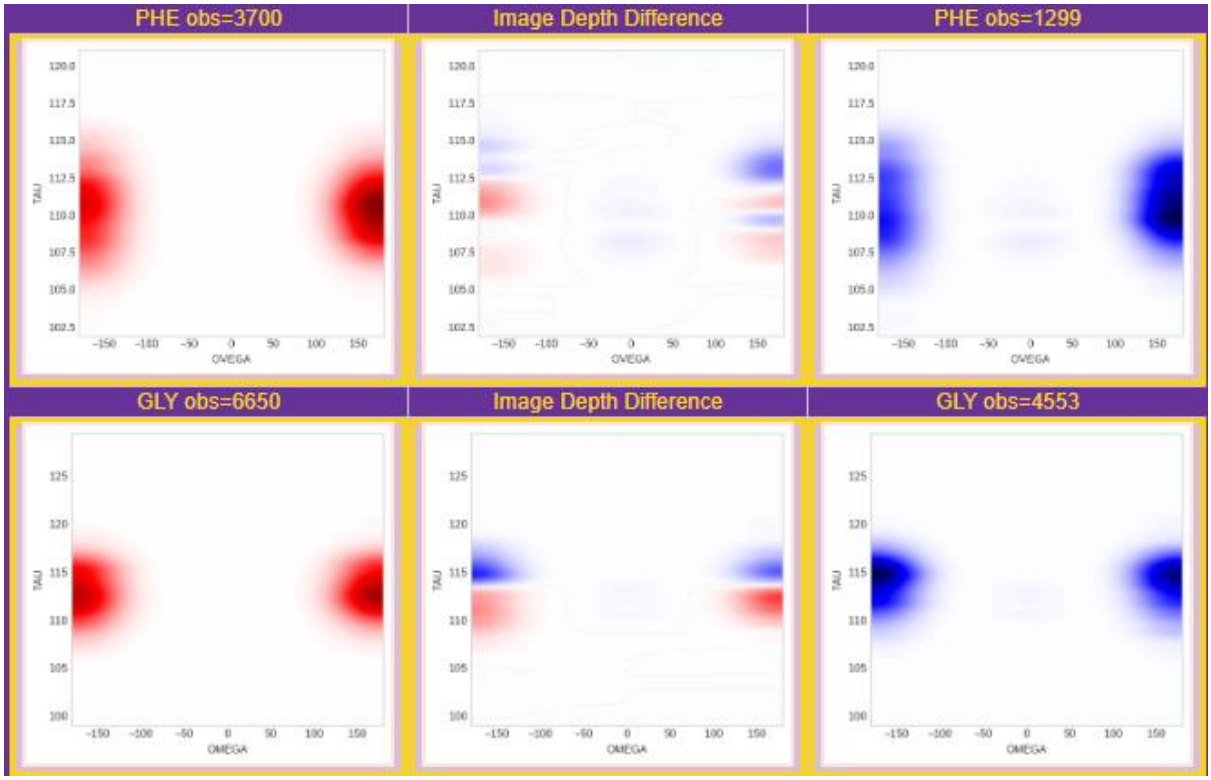
data	count	mean	std	min	25%	50%	75%	max	skew	kurtosis	normality
CHI1 ALA 0-1	1905.0	32.419	127.1978	-179.999	-61.101	57.097	176.158	179.999	-0.514	-0.963	N
CHI1 CYS 0-1	1096.0	-33.4026	100.6235	-179.881	-70.719	-61.14	55.1418	179.963	0.721	-0.149	N
CHI1 ASP 0-1	3988.0	-52.4658	95.9403	-179.999	-91.115	-68.472	48.9912	179.97	0.704	-0.134	N
CHI1 GLU 0-1	3420.0	-47.7886	101.5827	-179.951	-75.6545	-65.923	-54.9932	179.952	0.956	0.27	N
CHI1 PHE 0-1	2505.0	-28.6367	105.2743	-179.993	-74.606	-62.876	57.806	179.98	0.687	-0.437	N
CHI1 HIS 0-1	1466.0	-42.4166	102.082	-179.909	-76.9292	-63.3855	-45.4515	179.886	0.821	-0.041	N
CHI1 ILE 0-1	3358.0	-48.353	61.2114	-179.821	-66.9528	-60.278	-52.402	179.814	0.659	1.227	N
CHI1 LYS 0-1	3710.0	-47.1289	103.0462	-179.987	-76.5412	-65.2	-52.571	179.991	0.928	0.184	N
CHI1 LEU 0-1	5026.0	-45.5898	104.1677	-179.981	-76.1958	-64.9145	-55.5978	179.998	1.145	0.52	N
CHI1 MET 0-1	1171.0	-54.4526	92.587	-179.914	-74.932	-66.494	-57.911	179.991	1.177	1.174	N
CHI1 ASN 0-1	3264.0	-60.7756	89.4433	-179.999	-94.6922	-68.604	-56.5285	179.985	0.917	0.581	N
CHI1 PRO 0-1	3076.0	1.7093	26.3312	-51.094	-25.5625	6.632	27.5497	56.981	-0.032	-1.782	N
CHI1 GLN 0-1	2453.0	-50.4068	99.3198	-179.936	-74.805	-65.738	-55.499	179.972	1.035	0.562	N
CHI1 ARG 0-1	2673.0	-45.3694	102.6945	-179.946	-76.046	-65.375	-53.675	179.939	0.954	0.212	N
CHI1 SER 0-1	4272.0	17.1085	98.7951	-179.972	-63.4735	59.7555	70.5003	179.987	-0.338	-0.662	N
CHI1 THR 0-1	4445.0	-3.526	75.4911	-179.995	-60.606	44.874	62.789	179.95	-0.377	-0.668	N
CHI1 VAL 0-1	4772.0	55.7858	140.6505	-179.998	-63.9962	166.5635	174.2382	179.985	-0.595	-1.323	N
CHI1 TRP 0-1	1001.0	-29.2233	105.5755	-179.985	-75.479	-65.001	59.886	179.983	0.605	-0.527	N
CHI1 TYR 0-1	2419.0	-33.7364	105.9757	-179.982	-75.0435	-63.025	53.6915	179.983	0.75	-0.317	N
CHI1 ALA 1-1.1	2416.0	43.8719	94.0149	-180.0	49.796	56.7515	60.504	180.0	-0.558	0.118	N
CHI1 CYS 1-1.1	1494.0	-30.1012	102.7927	-179.982	-71.2345	-61.9635	61.004	179.953	0.585	-0.444	N
CHI1 ASP 1-1.1	6495.0	-54.3049	97.2969	-179.984	-137.8865	-69.134	-49.5395	179.983	0.776	-0.035	N
CHI1 GLU 1-1.1	6048.0	-44.6237	103.0096	-179.998	-75.591	-66.1455	-52.9415	179.996	0.917	0.119	N
CHI1 PHE 1-1.1	4159.0	-31.9524	108.6898	-179.999	-77.239	-63.945	57.8765	179.995	0.692	-0.494	N
CHI1 HIS 1-1.1	2553.0	-41.5679	102.2276	-179.99	-77.583	-63.514	45.477	179.989	0.77	-0.137	N
CHI1 ILE 1-1.1	5820.0	-51.2138	60.2562	-179.966	-67.307	-60.9695	-53.3325	179.964	0.76	1.809	N
CHI1 LYS 1-1.1	5716.0	-48.2375	102.6591	-180.0	-76.6462	-65.802	-54.6758	179.984	1.005	0.341	N
CHI1 LEU 1-1.1	8770.0	-45.9544	102.3056	-179.998	-74.974	-65.029	-56.2888	179.997	1.188	0.669	N
CHI1 MET 1-1.1	2077.0	-54.1756	94.743	-179.996	-76.906	-66.532	-56.336	179.989	1.086	0.877	N
CHI1 ASN 1-1.1	4866.0	-57.5559	91.6178	-179.968	-85.889	-68.583	-55.049	179.977	0.894	0.443	N
CHI1 PRO 1-1.1	5164.0	0.3098	25.8216	-42.124	-25.5663	-3.8275	26.7895	43.485	0.052	-1.771	N
CHI1 GLN 1-1.1	3942.0	-53.5592	95.4412	-179.992	-74.606	-66.1865	-57.1788	179.996	1.092	0.867	N
CHI1 ARG 1-1.1	4848.0	-45.3657	103.7176	-179.994	-77.711	-65.766	-52.968	179.993	0.938	0.13	N
CHI1 SER 1-1.1	6355.0	19.3901	99.2459	-179.99	-63.377	59.903	71.493	179.985	-0.33	-0.672	N
CHI1 THR 1-1.1	6398.0	-1.7119	75.6522	-179.77	-60.5918	48.226	62.8762	179.936	-0.37	-0.632	N
CHI1 VAL 1-1.1	7637.0	54.6134	141.0675	-179.999	-63.791	166.009	174.123	179.993	-0.578	-1.348	N
CHI1 TRP 1-1.1	1667.0	-32.6818	104.1292	-179.933	-75.8005	-64.212	54.8925	179.974	0.69	-0.359	N
CHI1 TYR 1-1.1	3730.0	-31.0476	107.6488	-179.983	-75.8035	-62.877	56.6542	179.98	0.737	-0.419	N
CHI1 ALA 1.1-1.2	1916.0	49.0782	95.5898	-179.997	54.7658	57.227	60.7405	179.999	-0.834	0.596	N
CHI1 CYS 1.1-1.2	1899.0	-32.0105	103.1261	-179.915	-72.25	-62.428	59.032	179.997	0.647	-0.373	N
CHI1 ASP 1.1-1.2	8266.0	-56.4871	96.4792	-179.995	-146.0302	-69.729	-53.4892	179.998	0.814	0.067	N
CHI1 GLU 1.1-1.2	8367.0	-50.0136	101.3109	-179.997	-77.598	-66.843	-56.156	179.998	0.98	0.333	N
CHI1 PHE 1.1-1.2	5473.0	-32.5434	107.8976	-179.993	-77.167	-63.801	55.776	179.997	0.755	-0.397	N
CHI1 HIS 1.1-1.2	3115.0	-41.3044	100.812	-179.977	-77.127	-63.68	44.763	179.958	0.763	-0.115	N
CHI1 ILE 1.1-1.2	7536.0	-53.1501	58.3971	-179.938	-67.81	-61.759	-54.3835	179.909	0.811	2.168	N
CHI1 LYS 1.1-1.2	7810.0	-49.3468	102.9235	-179.995	-77.935	-66.0165	-54.6502	179.994	1.001	0.337	N
CHI1 LEU 1.1-1.2	11584.0	-46.5778	103.0829	-179.999	-75.4868	-65.402	-56.2238	180.0	1.187	0.642	N
CHI1 MET 1.1-1.2	2461.0	-51.0949	92.9715	-179.986	-74.259	-65.951	-56.996	179.968	1.12	0.982	N
CHI1 ASN 1.1-1.2	6201.0	-58.0051	91.2802	-179.927	-89.739	-68.511	-54.337	179.995	0.863	0.37	N

CHI1 PRO 1.1-1.2	6460.0	0.2178	25.8361	-40.236	-25.5	-6.5045	26.71	43.975	0.073	-1.758	N
CHI1 GLN 1.1-1.2	5156.0	-52.2702	97.497	-179.999	-76.0592	-66.015	-56.1295	179.978	1.031	0.63	N
CHI1 ARG 1.1-1.2	6250.0	-46.4722	101.4435	-179.984	-76.899	-65.9765	-54.3075	179.998	0.99	0.314	N
CHI1 SER 1.1-1.2	8386.0	17.134	100.162	-179.999	-63.8692	59.2825	71.002	179.98	-0.324	-0.697	N
CHI1 THR 1.1-1.2	8220.0	-4.4711	75.1241	-179.994	-60.9155	-36.991	62.1618	179.947	-0.327	-0.651	N
CHI1 VAL 1.1-1.2	9912.0	59.8257	140.0804	-179.996	-63.107	167.3675	174.3103	179.991	-0.65	-1.264	N
CHI1 TRP 1.1-1.2	2085.0	-31.7823	104.2354	-179.971	-77.166	-64.521	56.371	179.996	0.673	-0.423	N
CHI1 TYR 1.1-1.2	4905.0	-29.5665	112.481	-179.996	-78.462	-62.514	60.171	180.0	0.659	-0.649	N
CHI1 ALA 1.2-1.3	1994.0	38.1184	64.4607	-179.993	54.4837	56.765	58.6252	179.987	-1.234	2.831	N
CHI1 CYS 1.2-1.3	2694.0	-32.3946	97.3674	-179.957	-71.0522	-61.724	57.9982	179.976	0.618	-0.218	N
CHI1 ASP 1.2-1.3	13879.0	-53.0832	97.3349	-179.998	-92.419	-69.219	-46.3265	179.997	0.788	-0.01	N
CHI1 GLU 1.2-1.3	14398.0	-46.7346	104.4178	-179.996	-77.7165	-66.704	-53.5692	179.999	0.927	0.094	N
CHI1 PHE 1.2-1.3	9128.0	-34.4333	106.8694	-179.999	-77.562	-64.041	53.655	179.999	0.762	-0.345	N
CHI1 HIS 1.2-1.3	5294.0	-40.321	103.8598	-179.981	-78.2415	-63.5425	47.771	179.994	0.774	-0.2	N
CHI1 ILE 1.2-1.3	13115.0	-51.1558	60.9075	-179.924	-68.0655	-61.752	-53.959	179.978	0.761	1.678	N
CHI1 LYS 1.2-1.3	12748.0	-49.4733	103.0262	-179.998	-78.7692	-66.231	-54.7358	179.996	0.998	0.321	N
CHI1 LEU 1.2-1.3	19715.0	-44.1368	103.9361	-179.996	-75.1895	-65.183	-55.97	179.995	1.16	0.514	N
CHI1 MET 1.2-1.3	4124.0	-46.7246	96.9652	-179.981	-73.6988	-65.7805	-55.446	179.994	1.091	0.685	N
CHI1 ASN 1.2-1.3	10090.0	-56.9497	91.7212	-179.965	-85.897	-68.4105	-55.039	179.981	0.924	0.504	N
CHI1 PRO 1.2-1.3	10795.0	0.7641	25.8927	-44.793	-25.3845	-0.317	26.783	45.423	0.024	-1.756	N
CHI1 GLN 1.2-1.3	8192.0	-52.6299	95.4518	-179.995	-75.7375	-66.2325	-55.7025	179.987	1.062	0.785	N
CHI1 ARG 1.2-1.3	10821.0	-46.2993	103.1037	-179.99	-77.905	-66.009	-53.007	179.983	0.952	0.184	N
CHI1 SER 1.2-1.3	13644.0	15.2049	101.5831	-179.993	-64.3282	59.383	71.167	179.998	-0.333	-0.735	N
CHI1 THR 1.2-1.3	12906.0	-4.8328	76.0717	-179.954	-61.1998	-29.05	61.938	179.98	-0.327	-0.65	N
CHI1 VAL 1.2-1.3	16545.0	57.9908	141.193	-179.998	-63.24	166.944	174.185	179.998	-0.631	-1.296	N
CHI1 TRP 1.2-1.3	3453.0	-28.1541	108.4562	-179.994	-77.02	-63.87	57.185	179.989	0.602	-0.611	N
CHI1 TYR 1.2-1.3	8189.0	-27.7505	110.3388	-179.99	-76.819	-62.527	60.585	179.987	0.669	-0.604	N
CHI1 ALA 1.3-1.5	1024.0	57.1519	7.8955	-68.456	55.7885	57.133	58.83	172.17	-6.646	186.172	N
CHI1 CYS 1.3-1.5	747.0	-41.2766	95.413	-179.801	-74.1915	-66.32	51.0895	179.984	0.719	-0.029	N
CHI1 ASP 1.3-1.5	3590.0	-55.0809	96.4131	-179.995	-97.1825	-69.4125	-53.9482	179.982	0.841	0.129	N
CHI1 GLU 1.3-1.5	3464.0	-49.0669	99.6363	-179.983	-77.365	-66.9045	-55.1665	179.97	1.044	0.501	N
CHI1 PHE 1.3-1.5	2231.0	-35.6894	108.8956	-179.983	-79.1875	-64.628	49.9455	179.981	0.799	-0.347	N
CHI1 HIS 1.3-1.5	1467.0	-45.3825	102.0759	-179.941	-82.2435	-63.545	-44.459	179.972	0.825	-0.043	N
CHI1 ILE 1.3-1.5	3419.0	-53.3145	55.224	-179.996	-67.648	-61.617	-54.5465	179.665	0.852	2.448	N
CHI1 LYS 1.3-1.5	3033.0	-45.9788	106.4564	-179.952	-80.937	-66.611	-52.707	179.991	0.954	0.067	N
CHI1 LEU 1.3-1.5	5283.0	-46.0016	102.2526	-179.987	-76.438	-65.695	-56.51	179.986	1.213	0.687	N
CHI1 MET 1.3-1.5	1285.0	-51.2512	97.1297	-179.992	-77.185	-65.839	-55.488	179.953	1.11	0.783	N
CHI1 ASN 1.3-1.5	2732.0	-59.2811	89.382	-179.99	-87.028	-69.3735	-55.5162	179.998	0.939	0.642	N
CHI1 PRO 1.3-1.5	2919.0	1.4953	25.45	-41.842	-24.1235	1.029	26.9185	45.259	0.005	-1.768	N
CHI1 GLN 1.3-1.5	2288.0	-53.156	100.4589	-179.933	-80.8205	-66.5095	-55.805	179.973	1.049	0.522	N
CHI1 ARG 1.3-1.5	3048.0	-47.3454	101.185	-179.992	-78.933	-66.296	-53.268	179.982	0.983	0.293	N
CHI1 SER 1.3-1.5	3655.0	17.775	101.0572	-179.99	-63.8485	59.155	71.985	179.996	-0.319	-0.733	N
CHI1 THR 1.3-1.5	3360.0	-2.6335	77.2086	-179.944	-60.7402	45.8	62.638	179.723	-0.383	-0.599	N
CHI1 VAL 1.3-1.5	4130.0	63.3369	138.5869	-179.997	-61.9355	166.968	173.955	179.988	-0.698	-1.2	N
CHI1 TRP 1.3-1.5	931.0	-33.9753	111.2305	-179.908	-85.555	-66.114	57.9435	179.954	0.666	-0.605	N
CHI1 TYR 1.3-1.5	2156.0	-29.6433	108.2525	-179.995	-75.612	-63.5875	57.1332	179.907	0.75	-0.45	N

Appendix 17: Distribution close contact differences for hydrogen bond donors

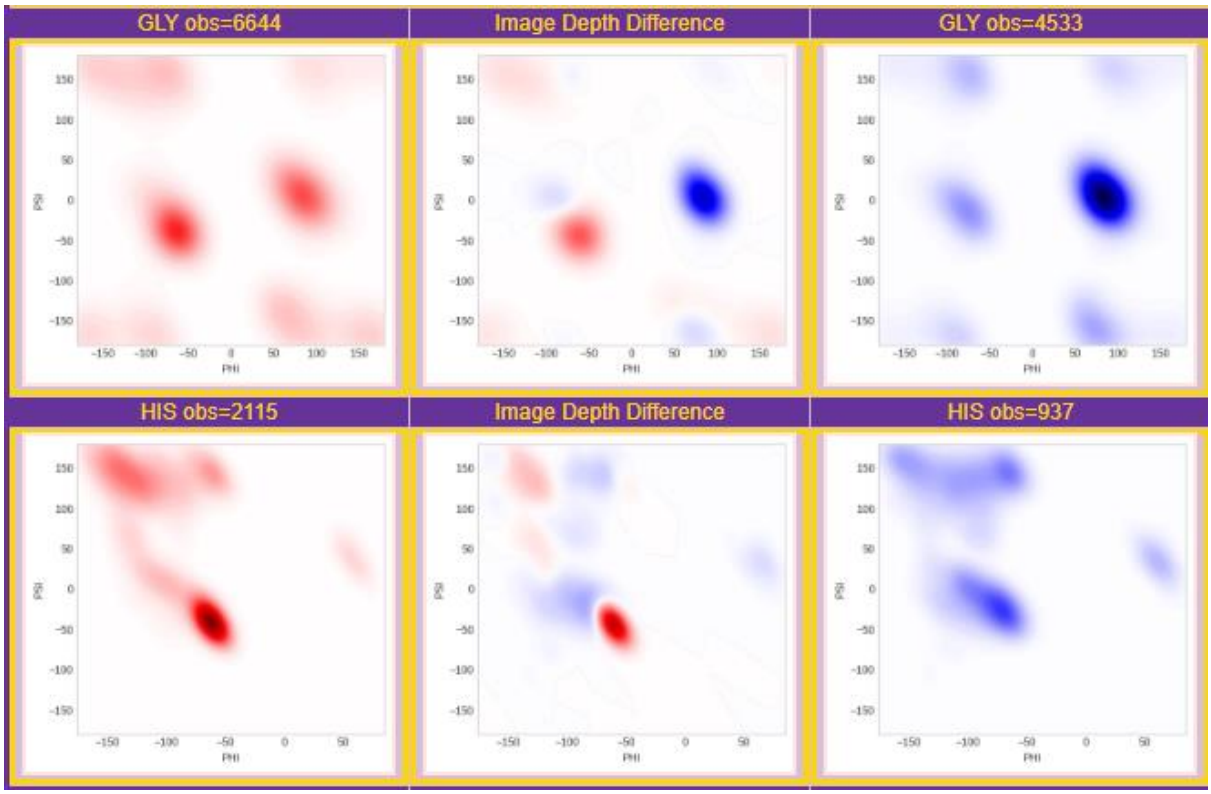
The difference images below show the distributions for N-O donors in red, and not N-O donors in blue. Where the difference image is red the distribution is skewed towards the donor, and where blue the non-donor. 2 examples, PHE and GLY, for the Ramachandran plot and Omega/Tau.

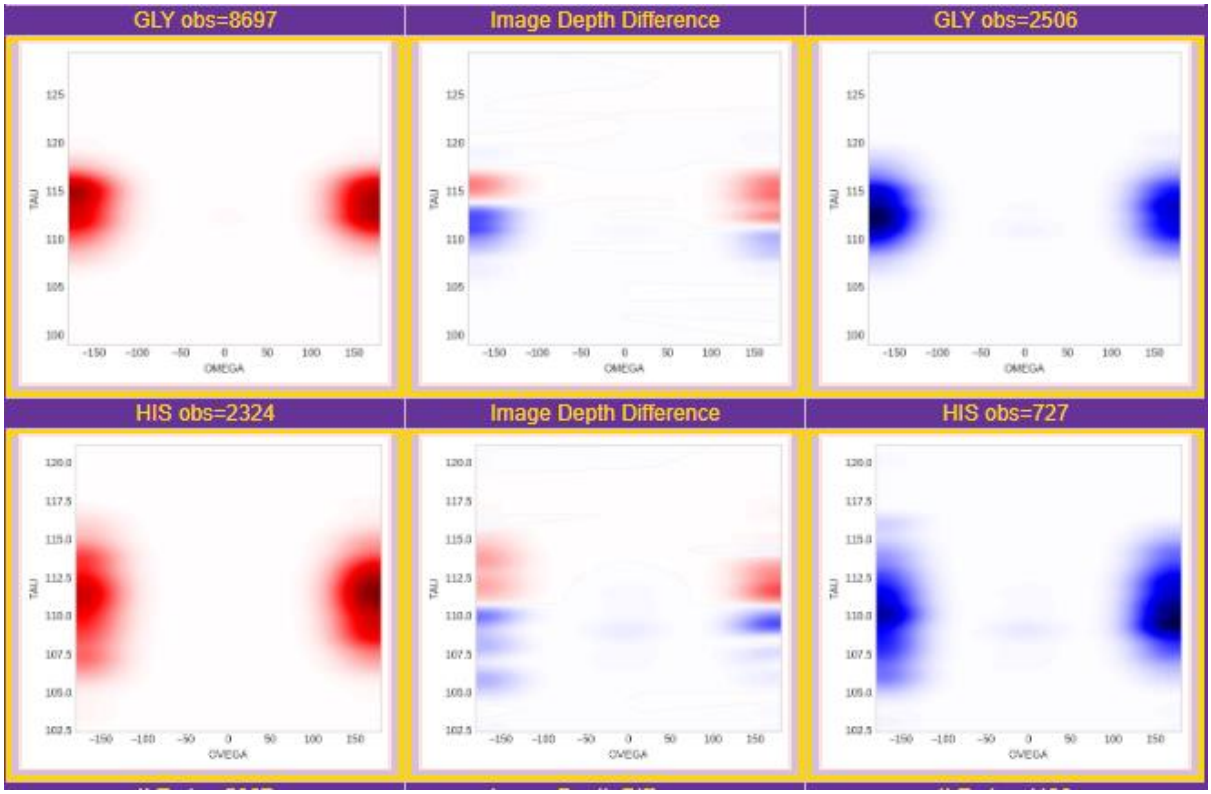




Appendix 18: Distribution close contact differences for hydrogen bond acceptors

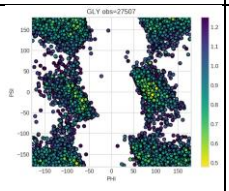
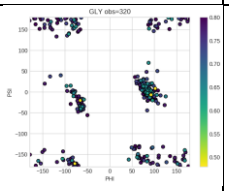
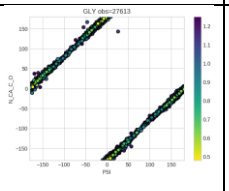
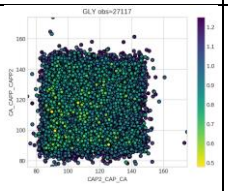
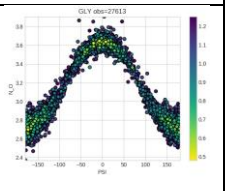
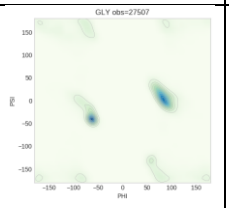
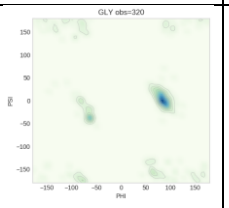
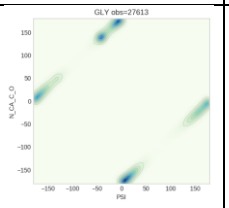
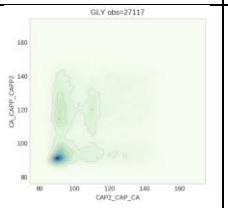
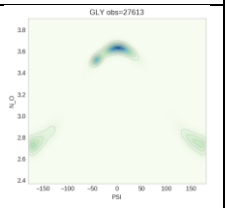
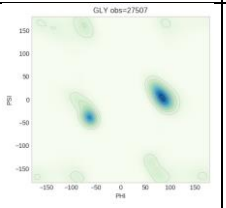
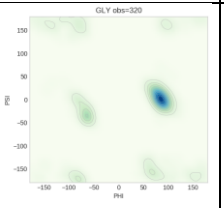
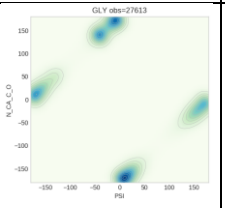
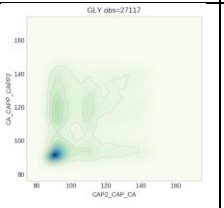
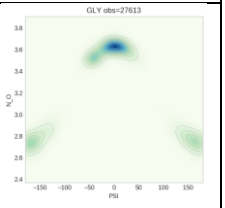
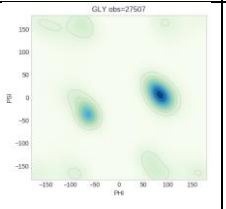
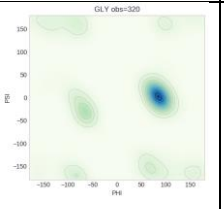
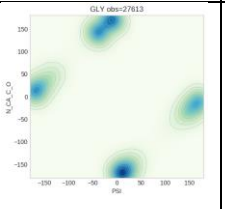
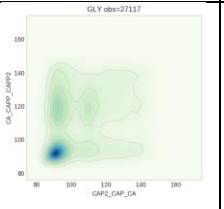
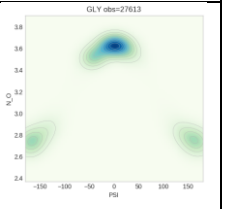
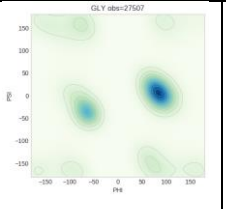
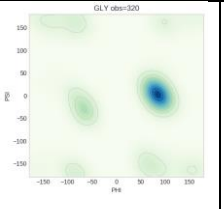
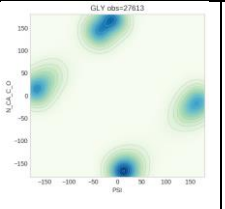
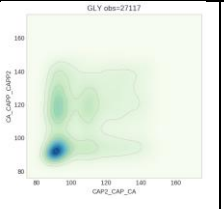
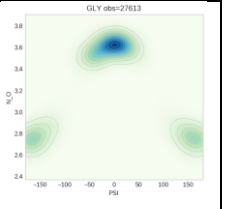
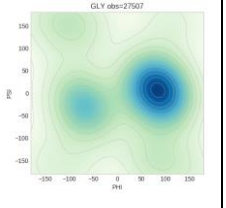
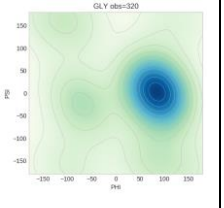
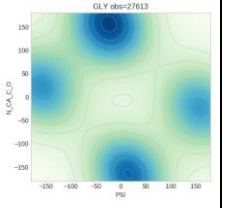
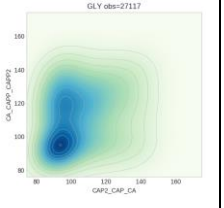
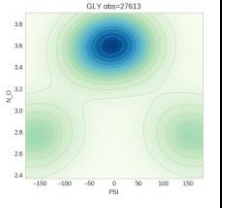
The difference images below show the distributions for N-O acceptors in red, and not N-O acceptors in blue. Where the difference image is red the distribution is skewed towards the acceptor, and where blue the non-acceptor. 2 examples, GLY and HIS, for the Ramachandran plot and Omega/Tau.





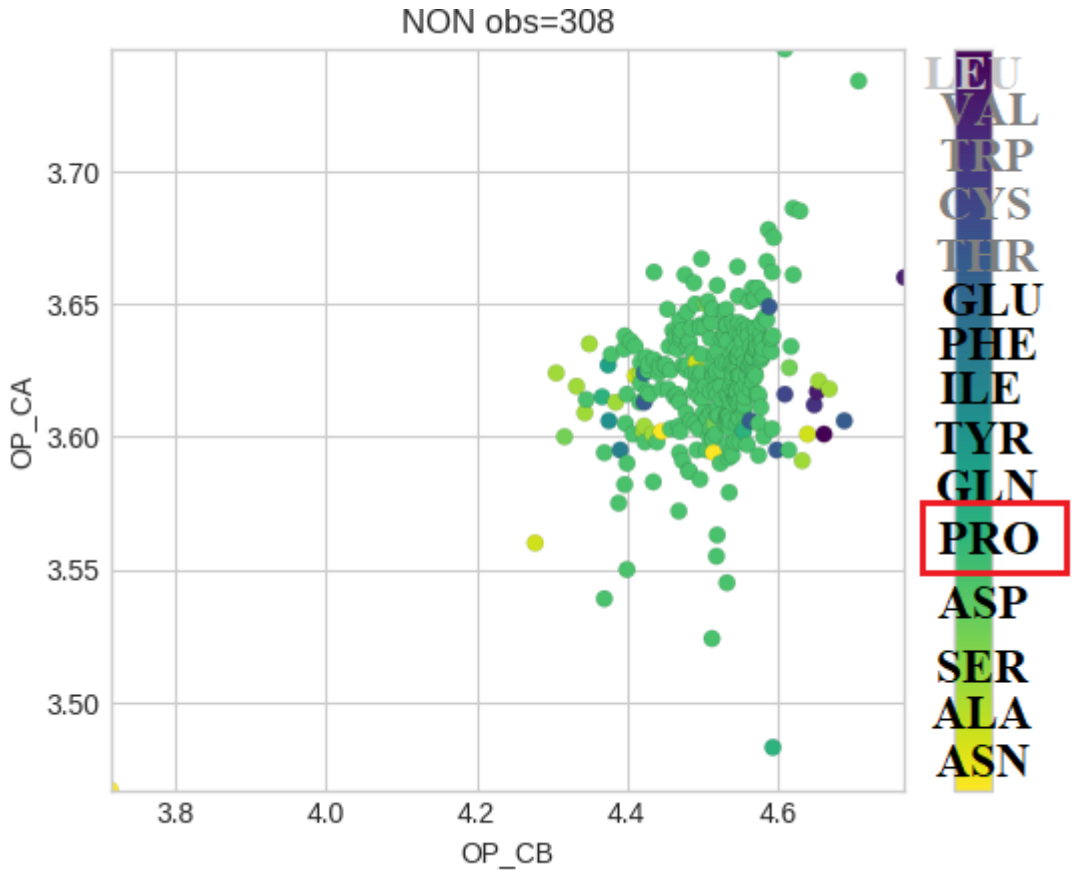
Appendix 19: KDE Bandwidth settings comparison for probability density

The kde bandwidth setting was selected as 0.10 to balance over and underfitting for the spread of distributions. The effort was made to cover areas of probability when the distribution is sparse but avoid improbable areas.

Kde=	a) GLY <=1.25	b) GLY <= 0.8	c) GLY <= 1.25	d) GLY <=1.25	e) GLY<=1.25
					
0.05					
0.10					
0.15					
0.18					
0.50					

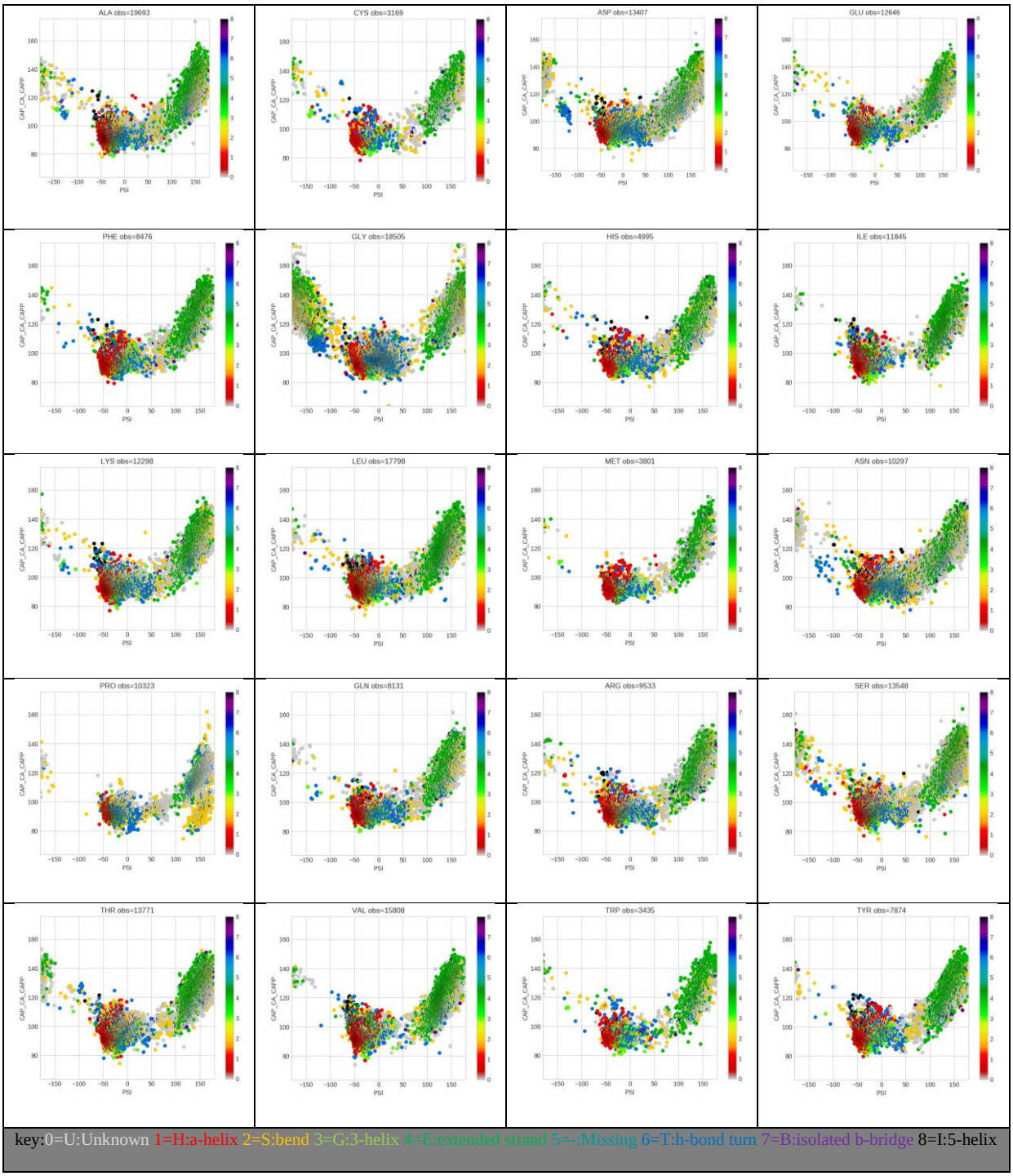
Appendix 20: Proline dominated region in cis/trans correlation plot

Residues at max bfactor 50, resolution <=1.2Å, rvalue < 0.16Å, rfree <= 0.3Å, for all residues, OP-CA > 3.4Å graduated on amino acid, shows this area almost entirely dominated by proline.



Appendix 21: Correlation of PSI versus CA1N-CA-CA1C for all amino acids

The geometric plot PSI/CA1N-CA-CA1C is shown below for each amino acid type, graduated on dssp secondary structure. The set was taken for resolution <=1.2Å, rvalue <=0.16Å, rfree <= 0.3Å and max bfactor 100.



Appendix 22: Alanine CHI1 and hydrogen placement

The definition for CHI1 is non-standard, defined as C-CA-CB-HB1 - the only hydrogen in a CHI1 angle in the system. Alanine has noticeable differences in the CHI1 distribution at high resolutions, to chich I attribute the increased accuracy of hydrogen determination at high resolution. The table below takes a selection of 10 from the highest resolution CHI1 alanine values and from the lowest 10, checking their atom coordinates and manually verifying the calculations.

The spreadsheet with the full calculations can be found on GitHub: CHI1 calculation

CHI1	atoms_CHI1	pdb_code	resolution	C-x	C-y	C-z	CA-x	CA-y	CA-z	CB-x	CB-y	CB-z	HB1-x	HB1-y	HB1-z
175.49	464-463-466-469	3NIR	0.48	2.337	-15.81	5.195	1.914	-16.19	6.614	1.293	-17.59	6.653	1.085	-17.85	7.658
-175.10	746-744-750-756	3NIR	0.48	8.087	-12.69	21.849	6.712	-13.34	21.992	6.011	-12.82	23.236	5.095	-13.33	23.378
179.07	45-44-47-50	1UCS	0.62	12.848	-0.892	21.586	13.84	-1.936	22.126	14.63	-2.507	20.966	15.263	-3.15	21.295
-179.64	753-752-755-758	1UCS	0.62	14.506	16.957	16.206	14.282	18.051	15.157	15.421	19.05	15.223	15.28	19.74	14.57
- 64.84	108-107-110-113	3X2M	0.64	-1.022	-11.75	-6.973	-1.256	-10.26	-6.778	-2.078	-10	-5.528	-2.959	-10.38	-5.654
57.10	682-681-684-686	3X2M	0.64	5.801	-17.33	-4.674	5.847	-16.82	-3.237	4.863	-17.59	-2.372	5.069	-18.53	-2.421
-179.95	2150-2149-2152-2154	3X2M	0.64	3.715	0.961	-20.79	3.377	1.031	-19.31	2.245	0.077	-19.01	2.031	0.124	-18.07
179.98	2198-2197-2200-2203	2VB1	0.65	5.514	21.258	26.135	6.363	20.077	26.584	6.418	18.995	25.512	6.959	18.265	25.822
- 60.00	2490-2489-2492-2495	2VB1	0.65	1.499	26.936	12.936	1.622	28.454	13.159	3.051	28.815	13.505	3.312	28.357	14.307
-168.75	727-726-729-732	1YK4	0.69	8.323	11.413	4.937	8.714	11.231	3.449	8.174	12.394	2.634	8.263	12.192	1.679
55.53	786-785-788-791	3WWX	1.5	12.234	2.325	21.413	12.21	2.629	19.925	12.631	1.412	19.136	12.06	0.666	19.379
59.06	5031-5030-5033-5036	3WWX	1.5	39.371	-12.37	7.304	39.988	-13.1	8.489	41.121	-13.99	8.017	41.789	-13.44	7.58
57.08	3-2-5-7	4HQS	1.5	6.827	12.248	16.535	6.442	10.804	16.24	5.071	10.461	16.858	4.413	11.073	16.52
- 60.87	476-475-478-481	4HQS	1.5	5.804	21.545	37.901	5.401	20.55	36.824	4.212	19.733	37.331	4.475	19.257	38.122
176.05	590-589-592-595	4HQS	1.5	16.317	21.137	42.127	15.26	21.591	43.124	15.212	20.626	44.281	14.588	20.952	44.934
57.22	716-715-718-721	4HQS	1.5	22.896	23.335	28.287	21.766	22.392	28.691	20.474	23.18	28.856	20.281	23.638	28.035
54.89	1984-1983-1986-1989	4HQS	1.5	23.32	10.237	34.122	24.106	9.898	35.392	23.187	9.297	36.436	22.734	8.541	36.055
58.12	879-878-881-884	4L58	1.5	-5.643	-7.901	4.336	-6.861	-7.072	4.736	-7.088	-7.132	6.234	-6.301	-6.812	6.682
53.99	481-480-483-486	4N5U	1.5	-11.68	12.677	-18.06	-11.95	11.169	-18.13	-12.09	10.717	-19.6	-11.31	11	-20.1
60.93	671-670-673-676	4N5U	1.5	-5.739	2.3	2.223	-4.737	3.055	1.384	-4.763	4.538	1.721	-4.528	4.653	2.655