Ultra-high-resolution crystallographic evidence for protein geometry variation

Introduction

Geometric parameters are essential for the refinement of proteins and the prediction of protein structure. The original parameters (E&H 1991, 2001) were derived from the CSD. Since then, there have been some reviews of these parameters (Jaskolski, 2007) based on the increasing resolution of x-ray refinement to sub-atomic resolution. As the resolution increases, there is the possibility of relaxation of refinement parameters to give a greater weight to experimental evidence. However, as E&H say "protein structures are generally solved not to build a statistically optimised protein database, but to discover biophysical functional mechanisms" (E&H 2001). We find ourselves then in an infinite spiral of looking for parameters from structures solved with the parameters.

In this study, we seek to sever this spiral by building a statistically optimised database of geometric evidence based on pure experimental evidence from the electron density of ultra-high resolution x-ray crystallographic structures solved at <=1Å. We decouple ourselves from the solved structures and go straight to the electron density for all atomic positions. In doing this we are freed from the problem of outliers: no cut off is required, all outliers are evidential in the electron density and thus all outliers provide us with geometric insights.

Method

Electron Density Topology

A method of finding: values, first and second derivatives from the electron density has been derived to enable both mathematical and visual analysis. The electron density grid points are interpolated using a multivariate interpolation method with Vandermonde matrices.
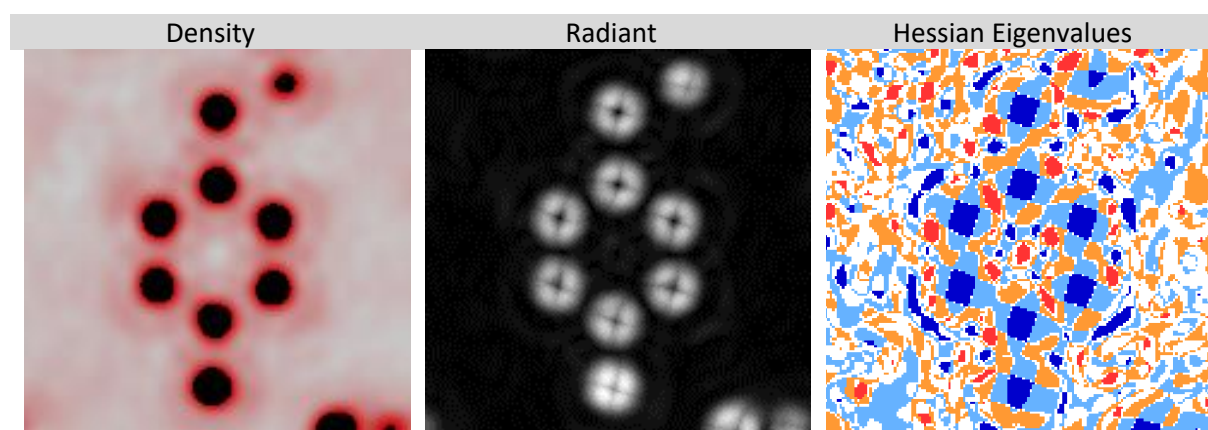

| Density | Radiant | Hessian Eigenvalues |

Figure 1 - 1us0 TYR A48, calculated with a 5th degree multivariate interpolating polynomial at 0.05Å samples. In the radiant image the atoms on the plane can be seen clearly by the black crosses – OH is not quite planar. The Hessian is calculated by the number of -ve or +ve eigenvalues of the Hessian. 3 negative values are a maximum and dark blue. 3 positive values, red, is a minimum. These are planar slices of 3d calculations.

The accuracy of the values can be determined by increasing the degree of the interpolated multivariate polynomial (stable by this method to a 7-degree polynomial) and the sample frequency of the interpolated points. In general, we find that a 3rd degree polynomial is sufficient mathematically, but a 5th degree polynomial is better visually. For visual images we sample at 0.05Å.

The visual images can be used to verify and determine the features we find mathematically – the values show us the electron density itself, but the most striking visual images are the images of the 1st derivative.

Electron density is difficult to compare across structures and within structures as the electron density depends on the stability of atoms' position and is therefore variant on secondary structure, as can be seen by these plots that shows electron density per atom type for 3 structures.

1ejg 0.54Å	5nqo 1.15Å	6q53 3.7Å
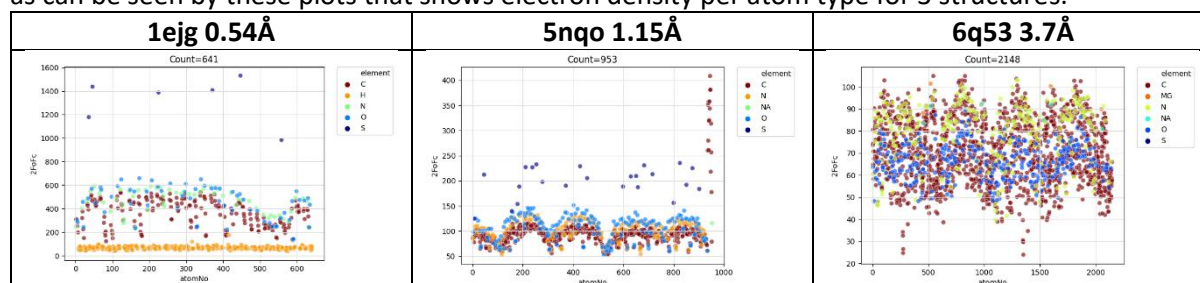
The density has a slider so that any given area of the electron density can be viewed at the appropriate intensity. The 1st derivative does not suffer the problem of value intensity and is more comparable across regions and structures. Mathematically, the maxima and minima are indistinguishable - visually they are clear. We use the L1 norm of Del of the 1st partial derivatives to produce images which we have coined the radiant of the electron density – a combination of the 1st derivative being a gradient and the way the coordinate axes lines radiate out from the maxima (see Figure 1). The solution to maxima distinction is solved mathematically with the Hessian eigenvalues but the radiant remains the most visually explicable solution.

The generation of a defensible dataset

A set of pdb structures is chosen as being <=1Å non-homologous at 90%, no nucleotides and >40 residues. We look at the original pdb data in these sets in 2 ways – UNRESTRICTED is all atoms, RESTRICTED is only those atoms that have single occupancy and whose bfactor is less than 1.3 x the average bfactor for the structure (CA atom of the residue).

Then, for each pdb structure, we choose a set of atoms which we wish to be evidential for geometric purposes, for the backbone geometric analysis we have chosen atoms of the types N, CA, C and O. For every atom of this type in each pdb, we look at the coordinates in the electron density and find at that position the value and the second derivative which is calculated via the 2nd partial derivatives and the eigen values of the hessian, where we choose whether the eigen values are 3 negatives to mark a maximum (Bader). If the atom is not at a maximum, we delete it from the pdb file. If the atom is a maximum, we build a cube around the point and look at all the values of the vertices – if any are bigger, we reject the point as not being a maximum, again deleted. The wider the cube the more permissive, several sets of data have been analysed at different widths.

We now have a reduced pdb file and a list of "bad" atoms for every structure. To examine whether these have given us a genuine separation of good and bad experimental evidence we need a more automated check of the electron density (manually loading every pdb in chimera is too time consuming). For this we have visualised the electron density's first derivative via the radiant images. Using this technique with an automated tool we can generate images of all the "bad" atoms to verify they are not defensible. Additionally, once the geometry has been calculated we can look at all the outliers to verify that despite the unusual geometry the atoms are indeed where they are purported to be.
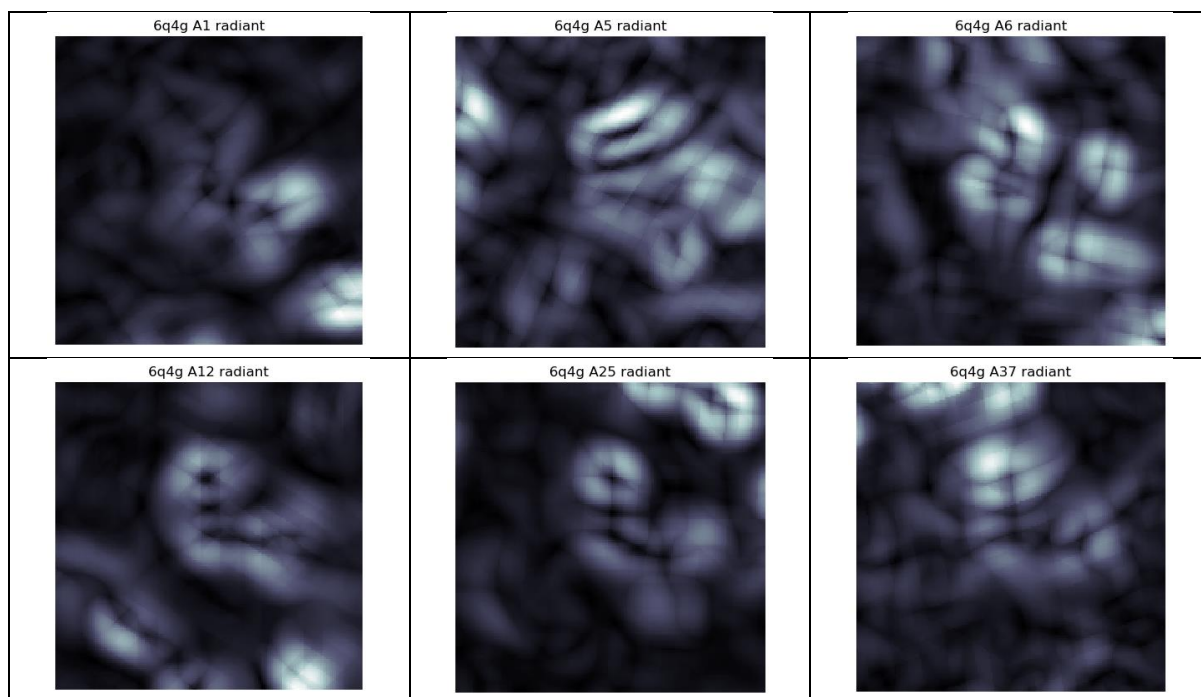
Results

Datasets

5 sets of pdb files were generated, based on electron density grid point fractions of 0.1, 0.2, 0.25, 0.3, 0.4 and 0.5. No data was calculated at 0.1 – too few atoms were found to be maxima and that tight interval to generate any geometry. At 0.5
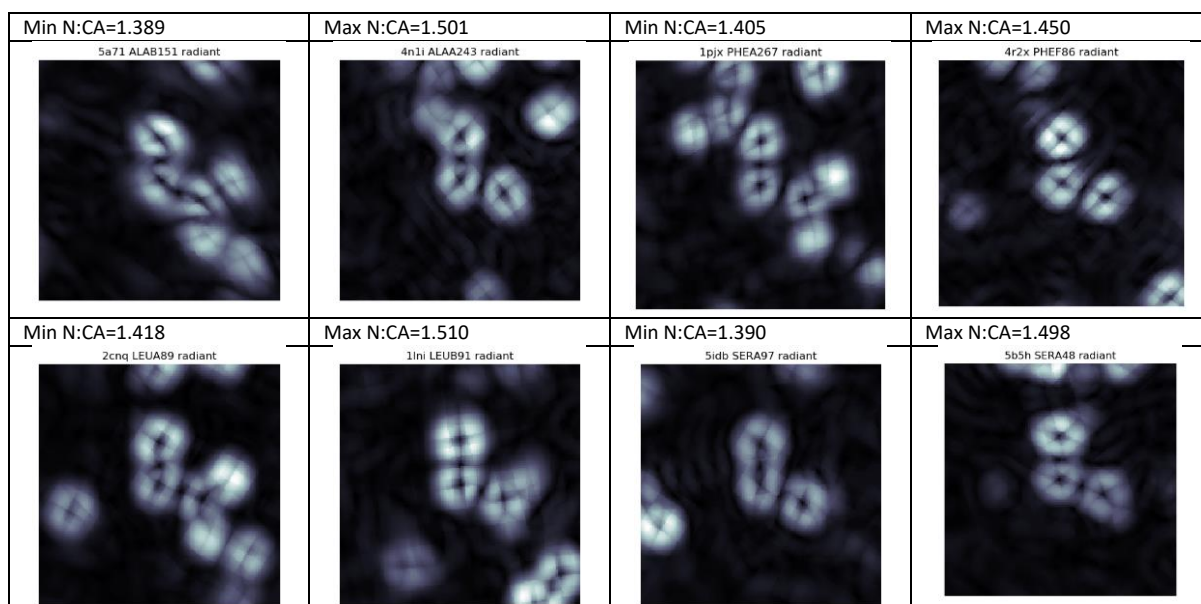
Radiant Images of Rejected Atoms
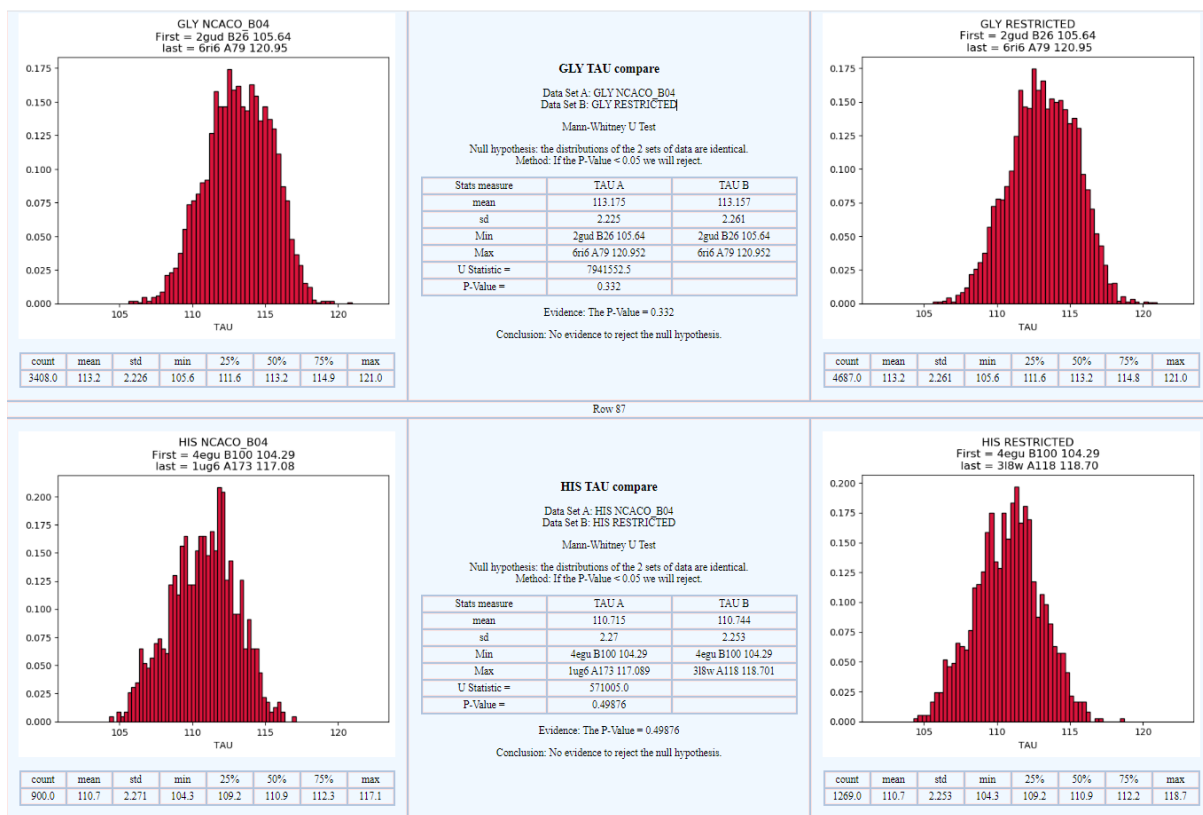
Some examples from 6q4g:

Radiant Images of Accepted Outliers

Some examples from set B03 for N:CA outliers
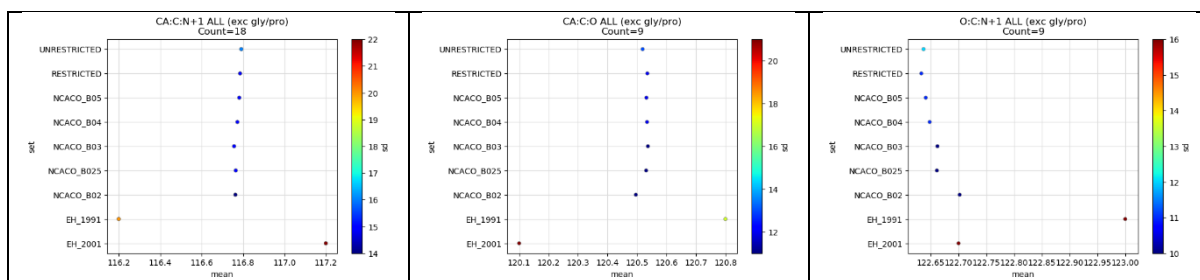
Statistical Analysis

We compare the datasets we get to the original data to check that it is representative as a distribution. There are loads of results, what is relevant to show, what is interesting?

A comparison of sets

For each of the sets we compare the means and sd against the E&H values. We are confident that the sets give similar values.
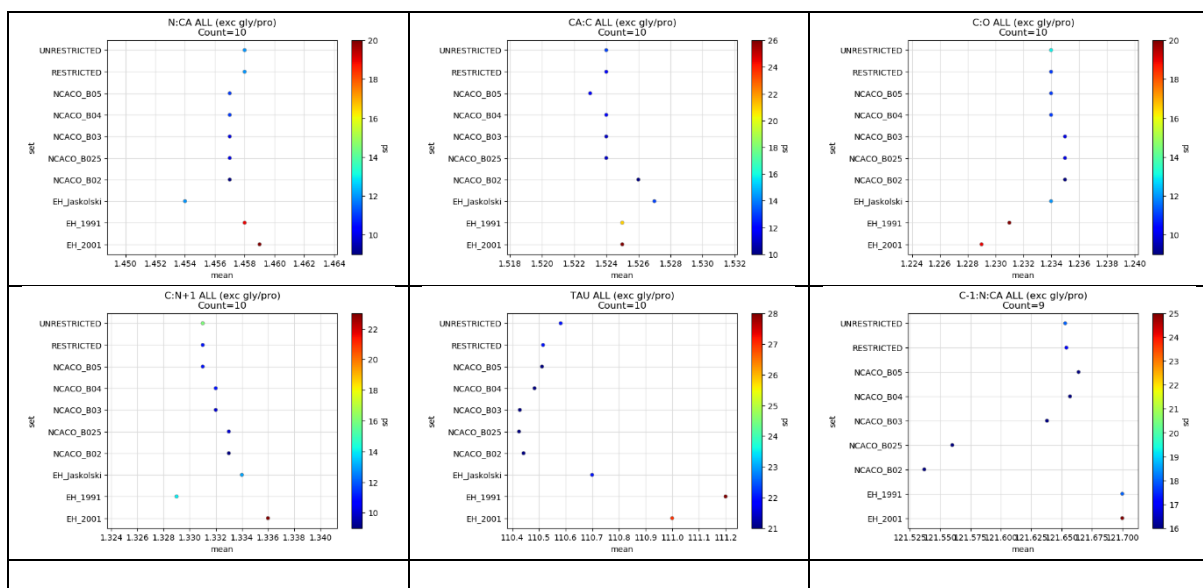
These are the observation counts we obtain for each of the sets.

AA	NCACO_B02	NCACO_B025	NCACO_B03	NCACO_B04	NCACO_B05	RESTRICTED	UNRESTRICTED
ALL	1062	8601	19579	33968	39687	47726	69595
GLY	105	882	1976	3408	4052	4687	7125
PRO	60	404	915	1684	2012	2441	3729

Some histograms

For each of the E&H bond lengths and angles, a histogram is generated with the outliers specified. These are compared to the E&H recommendations and the later Jaskolski recommendations. We have good values across the datasets so we will concentrate on the B03 set as a middle ground. All the data can be found here: Summary B03 E&H and Per Amino Acid B03 E&H

Some examples here:
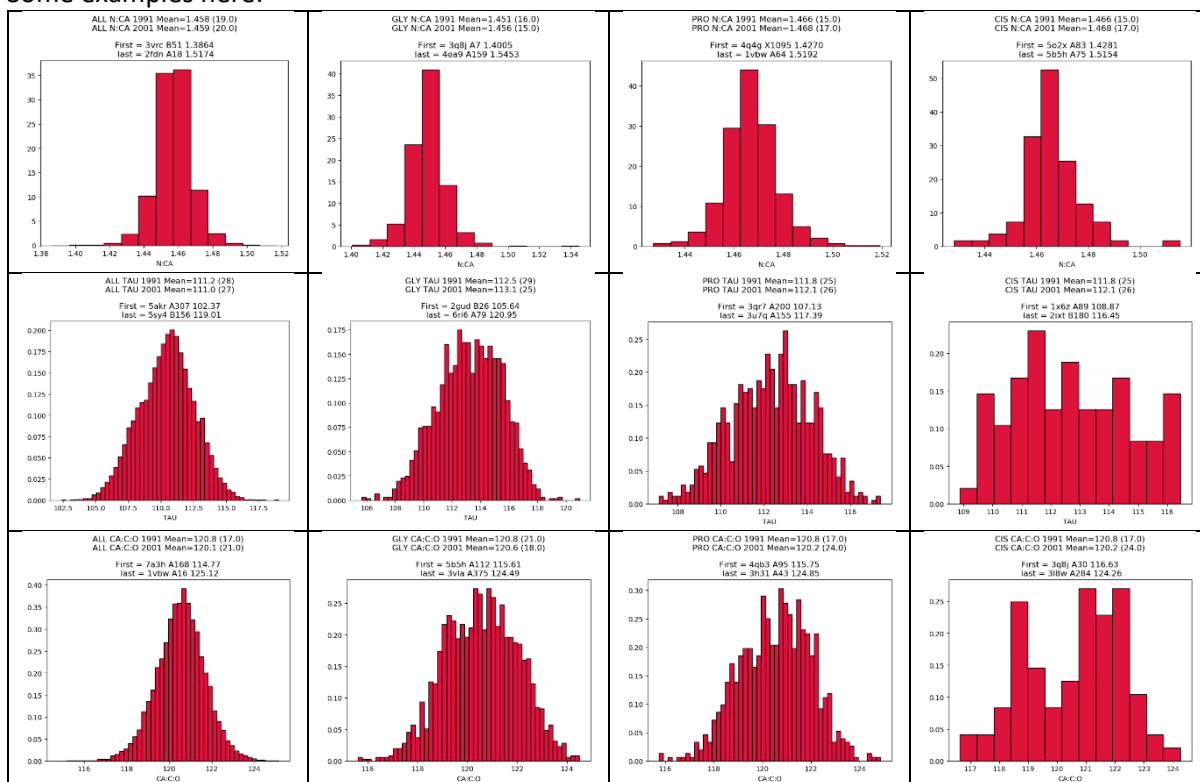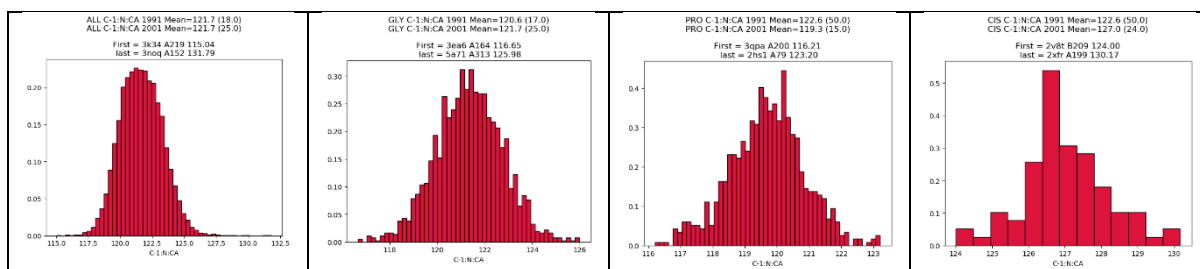
A table of results

|	geo	EH_1991	EH_2001	EH_Jaskolski	NCACO_B02	NCACO_B025	NCACO_B03	NCACO_B04	NCACO_B05	RESTRICTED	UNRESTRICTED
ALL	N:CA	1.458 (19)	1.459 (20)	1.454 (12)	1.457 (9)	1.457 (10)	1.457 (10)	1.457 (11)	1.457 (11)	1.458 (12)	1.458 (12)
GLY	N:CA	1.451 (16)	1.456 (15)		1.448 (8)	1.448 (11)	1.449 (11)	1.449 (12)	1.449 (12)	1.449 (12)	1.45 (14)
PRO	N:CA	1.466 (15)	1.468 (17)		1.468 (8)	1.468 (10)	1.467 (11)	1.467 (11)	1.467 (12)	1.467 (12)	1.467 (13)
ALL	CA:C	1.525 (21)	1.525 (26)	1.527 (13)	1.526 (10)	1.524 (11)	1.524 (11)	1.524 (12)	1.523 (12)	1.524 (12)	1.524 (13)
GLY	CA:C	1.516 (18)	1.514 (16)		1.515 (9)	1.515 (10)	1.514 (11)	1.514 (12)	1.514 (12)	1.514 (12)	1.514 (13)
PRO	CA:C	1.525 (21)	1.524 (20)		1.522 (10)	1.522 (11)	1.521 (12)	1.521 (12)	1.521 (13)	1.521 (13)	1.521 (14)
ALL	C:O	1.231 (20)	1.229 (19)	1.234 (12)	1.235 (9)	1.235 (10)	1.235 (10)	1.234 (11)	1.234 (11)	1.234 (11)	1.234 (13)
GLY	C:O	1.231 (20)	1.232 (16)		1.235 (8)	1.234 (9)	1.234 (10)	1.234 (11)	1.234 (11)	1.234 (11)	1.234 (16)
PRO	C:O	1.231 (20)	1.228 (20)		1.235 (9)	1.234 (10)	1.234 (11)	1.234 (11)	1.234 (11)	1.234 (12)	1.234 (13)
ALL	C:N+1	1.329 (14)	1.336 (23)	1.334 (13)	1.333 (9)	1.333 (10)	1.332 (10)	1.332 (11)	1.331 (11)	1.331 (11)	1.331 (16)
GLY	C:N+1	1.329 (14)	1.326 (18)		1.332 (9)	1.332 (10)	1.331 (10)	1.331 (11)	1.331 (11)	1.331 (11)	1.33 (13)
PRO	C:N+1	1.341 (16)	1.338 (19)		1.333 (11)	1.332 (10)	1.331 (11)	1.331 (11)	1.331 (12)	1.331 (12)	1.33 (13)
ALL	TAU	111.2 (28)	111 (27)	110.7 (22)	110.4 (21)	110.4 (21)	110.4 (21)	110.5 (21)	110.5 (21)	110.5 (22)	110.6 (22)
GLY	TAU	112.5 (29)	113.1 (25)		113.2 (19)	113.2 (22)	113.2 (22)	113.2 (22)	113.2 (22)	113.2 (23)	113.3 (24)
PRO	TAU	111.8 (25)	112.1 (26)		112.4 (18)	112.2 (19)	112.3 (19)	112.4 (19)	112.4 (19)	112.4 (20)	112.5 (20)
ALL	C-1:N:CA	121.7 (18)	121.7 (25)		121.5 (16)	121.6 (16)	121.6 (16)	121.7 (16)	121.7 (16)	121.7 (17)	121.7 (18)
GLY	C-1:N:CA	120.6 (17)	121.7 (25)		121.1 (14)	121.3 (14)	121.3 (14)	121.3 (14)	121.4 (15)	121.4 (15)	121.5 (20)
PRO	C-1:N:CA	122.6 (50)	119.3 (15)		120.7 (27)	120.3 (25)	120.3 (24)	120.3 (23)	120.3 (23)	120.2 (22)	120.2 (24)
ALL	CA:C:N+1	116.2 (20)	117.2 (22)		116.8 (14)	116.8 (15)	116.8 (15)	116.8 (15)	116.8 (15)	116.8 (15)	116.8 (16)
GLY	CA:C:N+1	116.4 (21)	116.2 (20)		116.5 (17)	116.6 (18)	116.6 (18)	116.6 (18)	116.6 (18)	116.6 (18)	116.7 (19)
PRO	CA:C:N+1	116.9 (15)	117.1 (28)		116.6 (18)	116.6 (17)	116.6 (17)	116.7 (17)	116.6 (17)	116.6 (18)	116.6 (19)
ALL	CA:C:O	120.8 (17)	120.1 (21)		120.5 (11)	120.5 (11)	120.5 (11)	120.5 (12)	120.5 (12)	120.5 (12)	120.5 (13)
GLY	CA:C:O	120.8 (21)	120.6 (18)		120.7 (14)	120.6 (15)	120.6 (15)	120.6 (15)	120.6 (15)	120.6 (15)	120.6 (17)
PRO	CA:C:O	120.8 (17)	120.2 (24)		120.6 (16)	120.6 (15)	120.5 (15)	120.5 (15)	120.5 (15)	120.6 (15)	120.5 (16)
ALL	O:C:N+1	123 (16)	122.7 (16)		122.7 (10)	122.7 (10)	122.7 (10)	122.6 (11)	122.6 (11)	122.6 (11)	122.6 (12)
GLY	O:C:N+1	123 (16)	123.2 (17)		122.7 (9)	122.8 (10)	122.8 (10)	122.8 (10)	122.8 (11)	122.7 (11)	122.7 (13)
PRO	O:C:N+1	122 (14)	121.1 (19)		122.7 (7)	122.8 (9)	122.8 (10)	122.8 (11)	122.8 (11)	122.8 (11)	122.8 (13)

Tau Variation

[An exploration of tau variation]
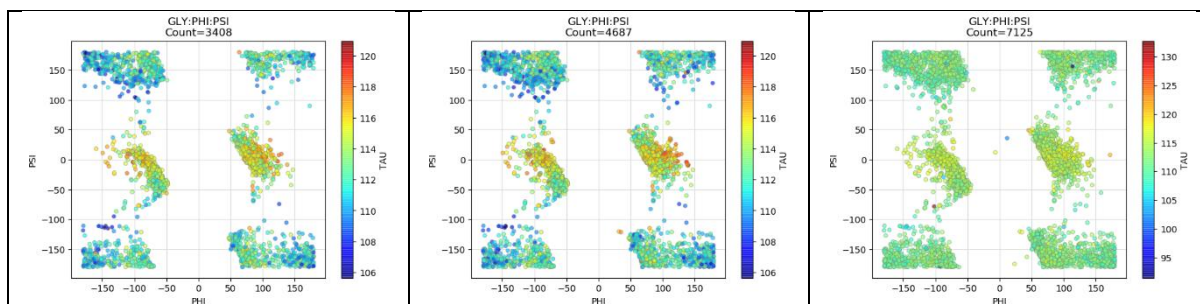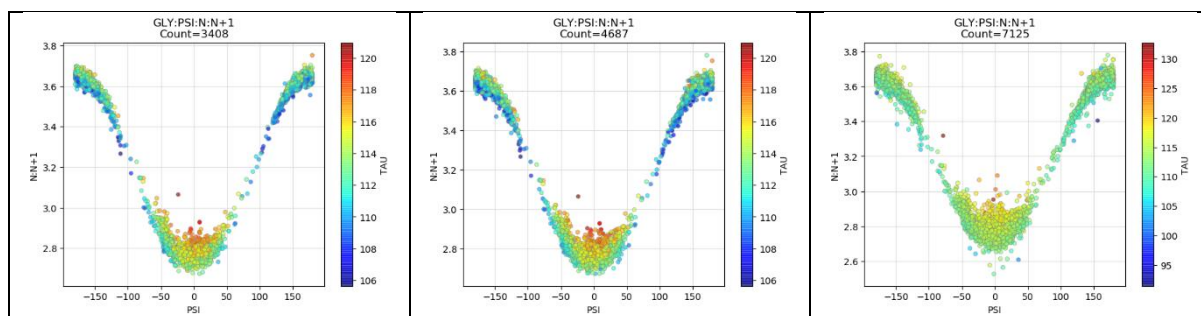
Discussion

a. So that's interesting - these are the new geometric recommendations….

For example, I have produced reports for every, or any geo you might want in this format:

GeoX	GeoY	MinHue	Min	MaxHue	Max	Mean	Sd
N:CA	ALA	5a71B151ALA	1.389	5x9lA140ALA	1.527	1.458	0.011
N:CA	ARG	2cnqA38ARG	1.414	2rbkA197ARG	1.493	1.458	0.011
N:CA	ASN	3vrcB51ASN	1.386	4uyrA74ASN	1.515	1.457	0.011
N:CA	ASP	6cdxB9ASP	1.4	3wcqA68ASP	1.515	1.458	0.012
N:CA	CYS	3k34A206CYS	1.407	2fdnA18CYS	1.517	1.455	0.012
N:CA	GLN	5a71B112GLN	1.409	5b5hA165GLN	1.506	1.458	0.011
N:CA	GLU	4r2xA124GLU	1.398	4r2xD183GLU	1.499	1.458	0.012
N:CA	GLY	3q8jA7GLY	1.401	4ea9A159GLY	1.545	1.449	0.012
N:CA	HIS	1m1qA52HIS	1.397	3nedA25HIS	1.502	1.457	0.011
N:CA	ILE	4r2xF20ILE	1.406	1vyrA107ILE	1.5	1.459	0.011
N:CA	LEU	3ea6A170LEU	1.413	5aq0B422LEU	1.521	1.458	0.011
N:CA	LYS	4q4gX1108LYS	1.397	4x6hA77LYS	1.498	1.457	0.011
N:CA	MET	1mj5A131MET	1.413	4x6hA196MET	1.499	1.459	0.012
N:CA	PHE	1pjxA267PHE	1.405	3k34A66PHE	1.509	1.458	0.01
N:CA	PRO	4q4gX1095PRO	1.427	1vbwA64PRO	1.519	1.467	0.011
N:CA	SER	5idbA97SER	1.39	5idbB107SER	1.512	1.455	0.012
N:CA	THR	5b5hA4THR	1.403	1byiA9THR	1.5	1.455	0.011
N:CA	TRP	6g1iA334TRP	1.419	1c75A87TRP	1.493	1.457	0.01
N:CA	TYR	3dhaA132TYR	1.414	1vyrA351TYR	1.499	1.457	0.011
N:CA	VAL	1m1qA57VAL	1.406	4bj0A32VAL	1.524	1.458	0.01

b. And, the outliers are interesting too. Most of them look fine….why…

c. In the future we could:

    i. look at the spreads of various parameters and use the defensible data to explore what causes the spreads: we might start with tau, based on this N:N+1 rainbow pictures and the knowledge of the spreads

    ii. We might also look at outliers and see whether unusual geometry is defensible or dodgy. Are there any fabled undiscovered active sites? Are there experimental problems?

    iii. Multiple conformations is a whole topic

    iv. We could create parameters on an ad-hoc basis for anything anyone might be interested in.

    v. Re-refine the structures and look at them again. *E&H very end of paper, last point).

d. In conclusion, different geometric parameters are recommended here which are defensible with low sd.

References

Engh, R A, and R Huber. '18.3. Structure Quality and Target Parameters', n.d., 12.

Jaskolski, Mariusz, Miroslaw Gilski, Zbigniew Dauter, and Alexander Wlodawer. '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, no. 5 (1 May 2007): 611–20. https://doi.org/10.1107/S090744490700978X.