VOLUME TWELVE

COMPUTATIONAL CRYSTALLOGRAPHY NEWSLETTER

JULY MMXXI

PDB@50, cis-PRO, top2018, chiral validation

Table of Contents

- Editor's Note 33
- Phenix News 33
- Expert Advice
  - Fitting tips #21 – What are chiral outliers and what can I do about them? 34
- Short Communications
  - Updates from the Worldwide PDB: Celebrating PDB50 and PDBx/mmCIF news 41
  - Lessons from using the Cambridge Structure Database: III – Outlier Rejection 44
- Articles
  - The effect of adding a single peptide bond class 47
  - The top2018 pre-filtered dataset of high-quality protein residues 53

Editor

Nigel W. Moriarty, NWMoriarty@LBL.Gov

Editor's Note

Some of you may be aware of the recent announcement that the Protein Data Bank (PDB) is extending the length of the codes for the PDB entries from four to eight, and for Chemical Component Dictionary (CCD) entries from three to four. Read the news release here. Some may also be aware of an Editor's Note from July 2015 promoting the use of "human readable" formatting for codes. In short, it suggested using only the appropriate case for the letter o, i and L (see table A for examples).

Table A: Examples of human readable PDB codes compared with standard representations.

Standard		Human readable	
Uppercase	Lowercase	Uppercase	Lowercase
1OI0	1oi0	1oi0	1oi0
1IJJ	1ijj	1iJJ	1ijj
4OCL	4ocl	4oCL	4ocL
5SS2	5ss2	5ss2	5ss2

A small addendum to the original specification appears in the last line. The uppercase letter S (nineteenth letter of the alphabet) can be confused with the numeral 5 (five).

So, the appropriate case for the four letter is o, i, L and s. Interestingly, this provides a mnemonic – Lois – that is easy to remember and provides the correct case for each letter.

Alternatively, one could always use lowercase for the codes with the one exception for "L", which should always be uppercase. This approach has the added advantage of

The Computational Crystallography Newsletter (CCN) is a regularly distributed electronically via email and the Phenix website, www.phenix-online.org/newsletter. Feature articles, meeting announcements and reports, information on research or other items of interest to computational crystallographers or crystallographic software users can be submitted to the editor at any time for consideration. Submission of text by email or word-processing files using the CCN templates is requested. The CCN is not a formal publication and the authors retain full copyright on their contributions. The articles reproduced here may be freely downloaded for personal use, but to reference, copy or quote from it, such permission must be sought directly from the authors and agreed with them personally.

lowering the ambiguity of seeing a code and guessing whether it is uppercase or lowercase.

Phenix News

Announcements

New Phenix Release Imminent

Developers are working on a Python3.7 version of Phenix. This version will contain many new features. In the meantime, nightly builds are available by contacting the download email.

Please note that the latest publication should be used to cite the use of Phenix: Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, Hintze B, Hung LW, Jain S, McCoy AJ, Moriarty NW, Oeffner RD, Poon BK, Prisant MG, Read RJ, Richardson JS, Richardson DC, Sammito MD, Sobolev OV, Stockwell DH, Terwilliger TC, Urzhumtsev AG, Videau LL, Williams CJ, Adams PD: Acta Cryst. (2019). D75, 861-877.

Downloads, documentation and changes are available at phenix-online.org

Expert advice

Fitting Tip #21 – What are chiral outliers and what can I do about them?

Jane Richardson and Christopher Williams, Duke University

Chirality (or handedness) is an important and pervasive feature of biology. Your hands, of course, are handed, and so are macromolecules. Proteins are made of chiral L-amino acids – a property that is manifested at larger scale in righthanded α-helices and twisted β-sheets. Nucleic acids are made of handed nucleotide components with each form of DNA or RNA double helix having a specific handedness.

Handedness reversals are very rarely seen in macromolecular structures because they are disallowed by the geometry libraries used in model-building software and are very difficult for refinement to change. Backward chirality at the Cα is already detected by extremely large Cβ deviations (although that measure does not test other chiral centers). However, we have recently encountered a few chirality outliers in deposited or in-process models, and have implemented chirality validation in Phenix and MolProbity (Prisant 2020). If any chiral, pseudo-chiral or tetrahedral-geometry outliers occur in a model, they are now noted in the summary report and are listed individually in a Phenix validation GUI table or in a MolProbity text report. They are flagged in yellow on the 3D structure in the MolProbity "multi-kin" kinemage graphics, as seen in the icon above and in most of the figures below.

In pure geometry, the choice of chirality is a binary, plus-or-minus property, but to allow for molecular flexibility and for convenient programming it is usually measured by "chiral volume": volume of the tetrahedron enclosed by the central atom and the three attached atoms of highest chemical priority (for biological macromolecules, the lowest priority atom is almost always a hydrogen). Therefore, intermediate changes in the chiral volume measure can usefully detect serious

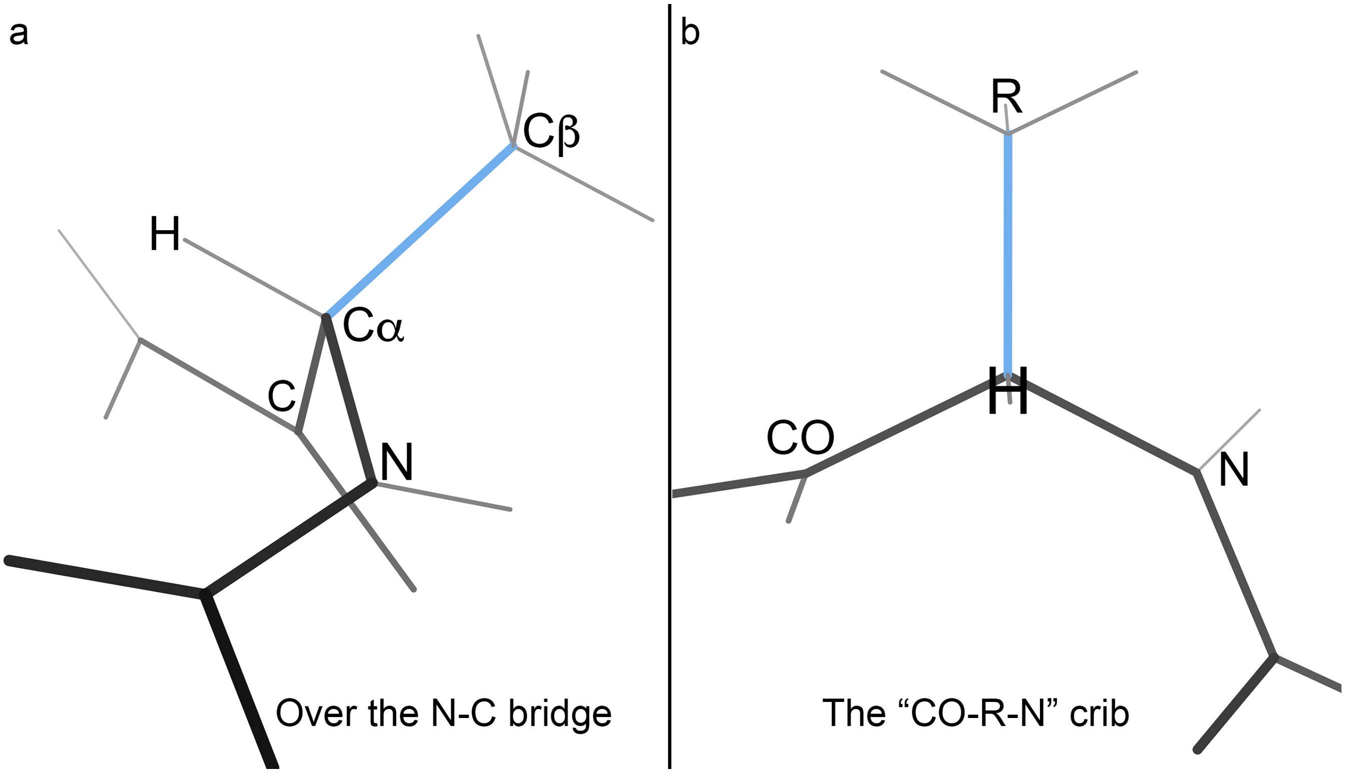


Figure 1: Mnemonics for identifying the normal L-amino acid handedness at a Cα. a) Turn the model or graphics to look from the N-terminal direction across the backbone "bridge" from N to C; the sidechain should be on your right b) Turn to look down on the Cα from the Hα; the substituents should read CO, R (the sidechain), N in the clockwise direction.

distortions of tetrahedral geometry, which are more common than chirality errors.

Definitions

A true chiral atom makes covalent bonds to four distinct atom types or branches. Common cases are:

- The protein Cα atom, bonded to the backbone N, backbone carbonyl C, sidechain Cβ, and H (Hα). Figure 1 illustrates two different mnemonics to help you distinguish normal L-amino acids from the chiral opposite D-amino acids.

- The Cβ atom of Ile or Thr, bonded to the Cα, the Hβ, the Cγ1 or Oγ1 of the long or heavy sidechain branch, and the Cγ2 methyl of the shorter branch.

- The nucleic acid C1' atom, bonded to C2' and O4'of the sugar ring, the C1' H, and the N1/N9 of the base (the other positions of substituents on the puckered sugar ring are also chiral - the C3', C5' and for RNA the C2').

- Carbohydrates are even richer in chiral centers, as are many enzyme substrates and other ligands.

A pseudo-chiral atom makes tetrahedral bonds to two distinct and two identical atoms or branches. The two identical ones are distinguished in name only, by an arbitrary consensus label (usually a number, such as Hb2 vs Hb3). Examples are:

- The Cβ of Val, with bonds to the Cα, the Hβ and the two identical Cγ methyls, which by pre-established chemical convention are labeled as Cg1 for the right-arm position and Cg2 for the left arm. Confusingly, that convention makes Val χ1 values differ in

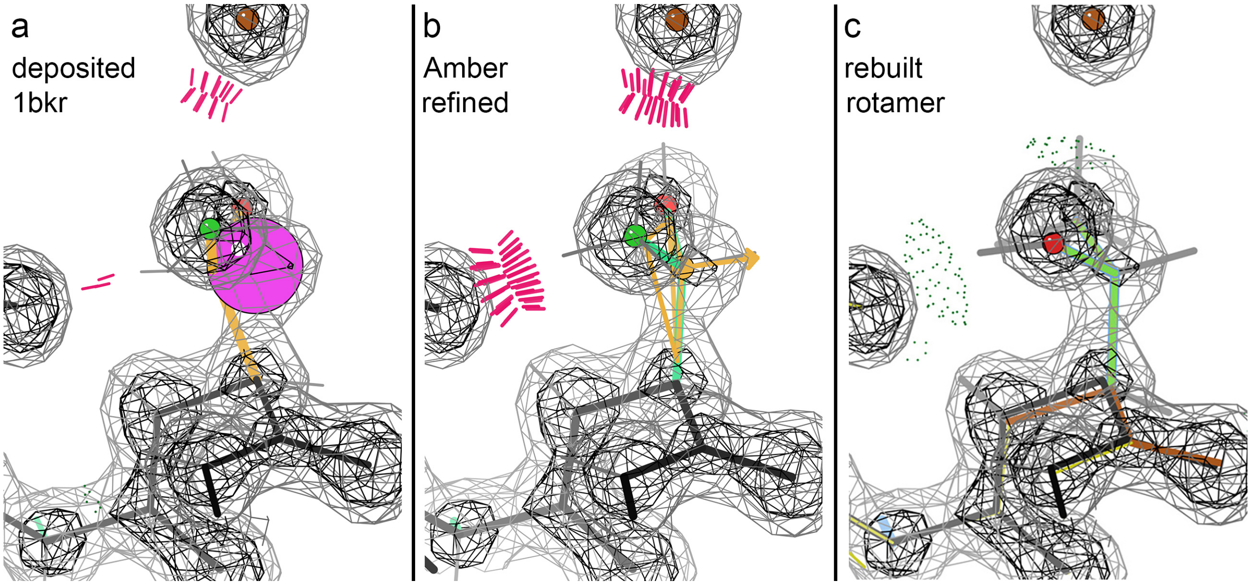


Figure 2: A real chiral outlier at a Thr. a) 1bkr as deposited, with a backward-fit rotamer and huge Cβ deviation (magenta ball, with ideal position at its center and observed position on its surface). b) Amber refinement moves all 3 sidechain non-H atoms into some density peak at the cost of reversed chirality at Cβ. c) The correct, outlier-free answer is just a different rotamer, with Oγ (red ball) in the higher peak and H-bonded.

backbone relationship from those for Thr and Ile.

- The nucleic acid P atom, with bonds to the backbone O5' and O3' and to the identical O1P and O2P atoms.

- Complexly connected het groups such as FeS clusters.

Categories of outlier cases

Chiral outliers are reported in three categories: chiral outliers, tetrahedral geometry outliers, and pseudo-chiral outliers, covering all chiral centers or tetrahedral centers defined in the Phenix geostd or monomer_library dictionaries.

True reversals of chirality can occur in software systems that do not include chirality among their geometrical restraints. This is true, for instance, in the otherwise-excellent Amber force-field refinement available in Phenix (Moriarty 2020). Figure 2a illustrates Thr 101 in 1bkr at 1.1Å resolution (Banuelos 1998), where the backward-fit sidechain places Oγ and methyl Cγ in the wrong density peaks and the Cβ far out of density; this produces clashes, a rotamer outlier, very bad covalent angles and a huge Cβ deviation (magenta sphere). Pure downhill refinement with the Amber force field moved all 3 non-H atoms into their density peaks by allowing the chirality around Cβ to reverse (Figure 2b), flagged as a yellow chirality outlier in current validation. The correct fix is to refit the sidechain rotamer, as shown in panel c, now with 2 H-bonds, no outliers and good density fit even before further refinement.

Apparent chiral outliers can occur because the group is misnamed in its 3-letter code. If an alanine D-amino-acid is called ALA rather than DAL, or a normal ALA is called DAL as in Figure 3, then MolProbity will also produce a graphical markup with an arrow to where the central tetrahedral atom of a DAL would be positioned, relative to the C, N, and Cβ atoms. In this case the fix is simple: just assign the correct residue name.

Tetrahedral-geometry outliers have chiral volumes more than 4σ different from the ideal chiral volume of the group involved. For

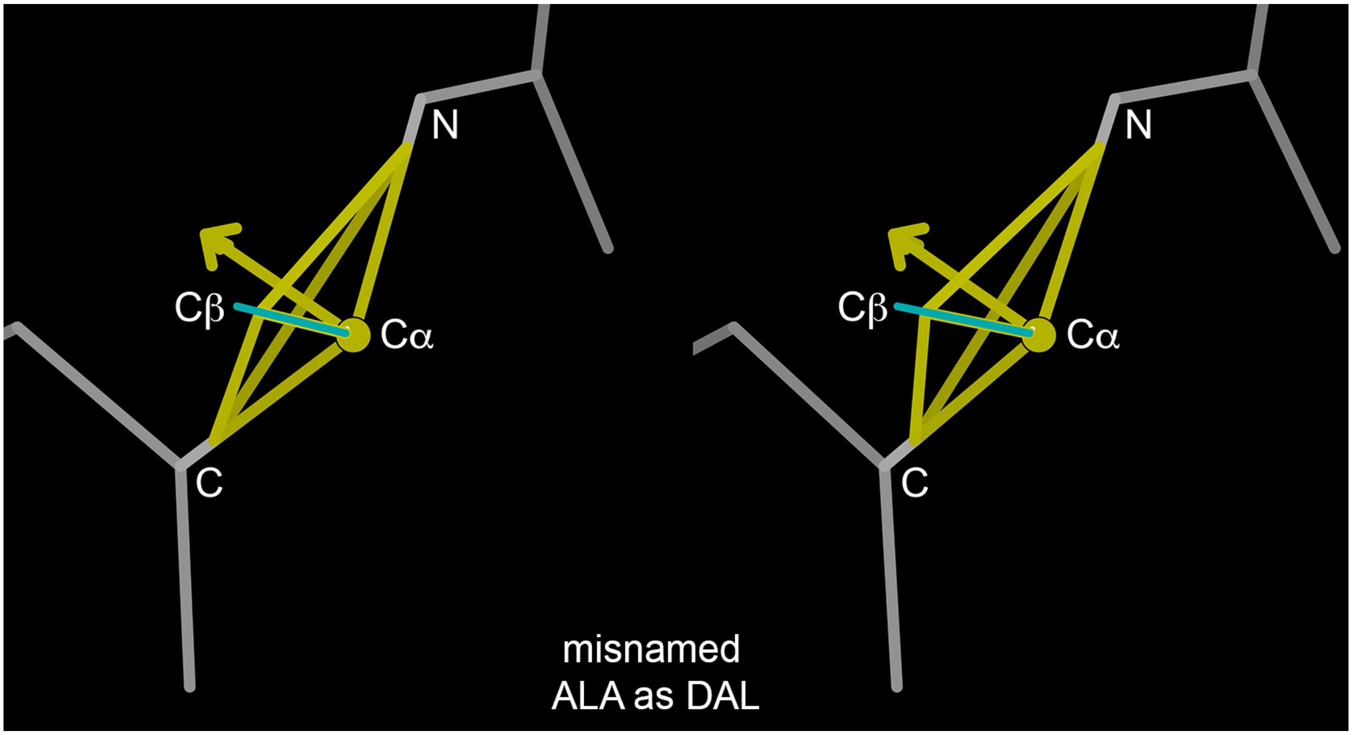


Figure 3: An apparent chiral outlier (yellow markup) caused by incorrect naming of the residue's 3-letter code. An actual ALA L-amino acid residue in a helix has been named as DAL (D-amino acid). In an actual DAL, the Cα would lie at the end of the arrow.

instance, Figure 4 shows Leu 995 in 3ogv at 1.4Å resolution, which is so far from tetrahedral that it is nearly planar. It is flagged with a similar yellow markup, but without the arrow. Bond-length and bond-angle outliers also show that there is a problem, but the chiral outlier more clearly indicates the problem: The Cβ should move left somewhat and the Cγ should move right, to fit the density better and provide stronger chirality, which could then be refined successfully. The Leu rotamer was presumably fit backward originally (with the Cγ back and left rather than forward and right), and the density pulled it almost flat during refinement.

Pseudo-chiral outliers are always caused by failure to name, or number, the two identical substituents in accordance with standard conventions. An easily understandable case would be switching Cg1 and Cg2 labels in a valine sidechain but with Cγ and Cβ atoms in the correct places and density peaks. An example is shown in Figure 5.

Atom naming issues do not matter in many ways, but they cause problems with identifying dihedral angles, superimposing related structures, and similar functionalities. Since they are trivial to fix, that should always be done. Look up the wwPDB naming conventions (by 3-letter code) for the particular group, and follow them.

These and the other chiral categories are reported individually by MolProbity in a chirals.txt report such as shown in Figure 6 for a file deliberately messed up to include all three types of chiral outliers.

Discussion

This chiral outlier validation does not flag naming errors among multiple H atoms (only between heavier atoms or between them and

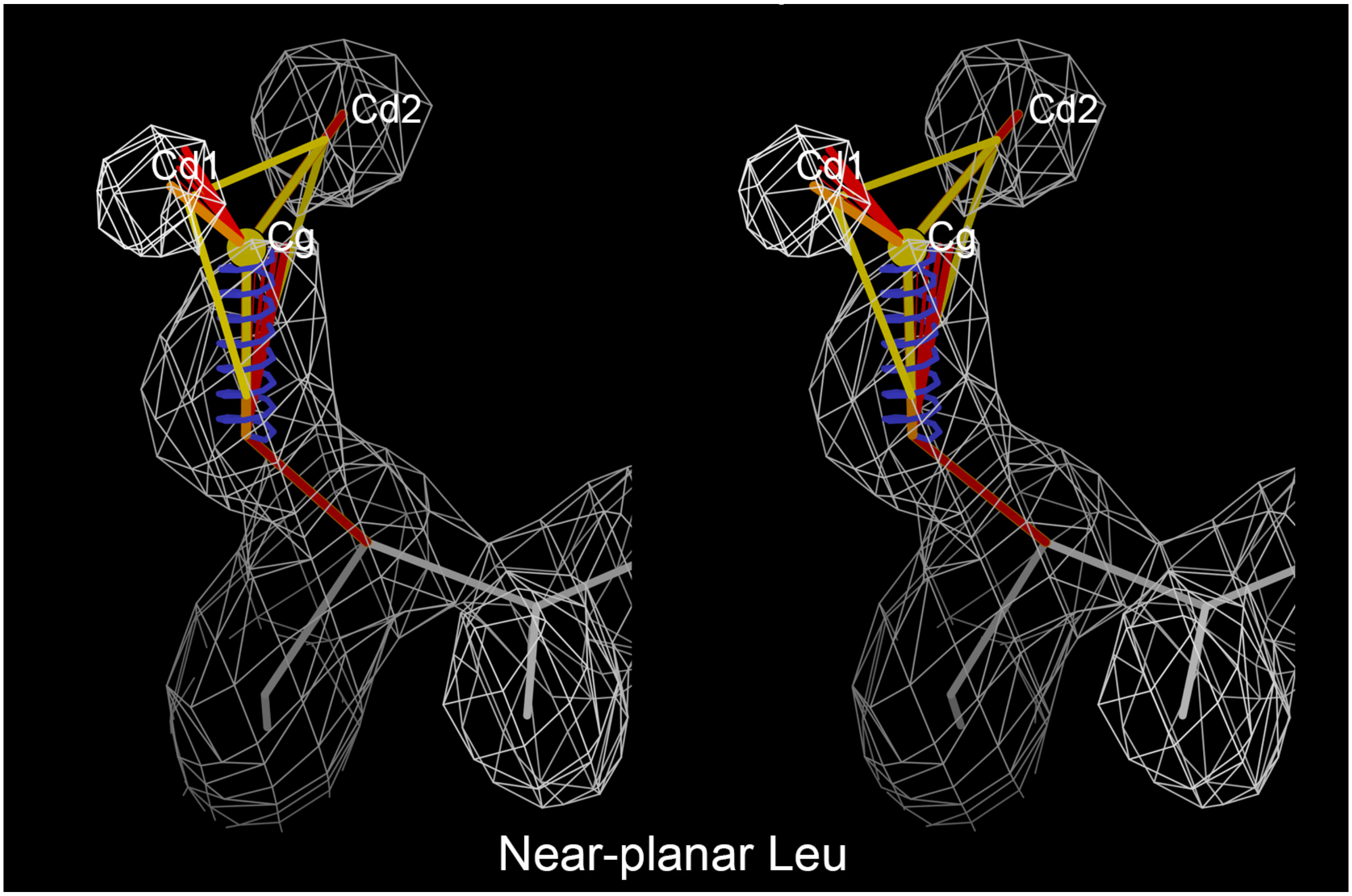


Figure 4: An extreme tetrahedral-geometry outlier at a Leu Cg. Presumably, the original fit in a 180°-opposite non-rotamer fights in refinement with fit to the density, producing this nearly flat tetrahedral group. This distortion also shows very large bond-angle outliers (red fans).

an H). Modern software should provide accurate H names, but you might encounter such errors in older files, since the conventions changed very thoroughly when the wwPDB moved from version 2.7 to v3.0 format nearly 10 years ago. For example, the hydrogens on a methylene were previously numbered 1 and 2, but are now 2 and 3. Specifically, at Cβ the continuing heavier-atom

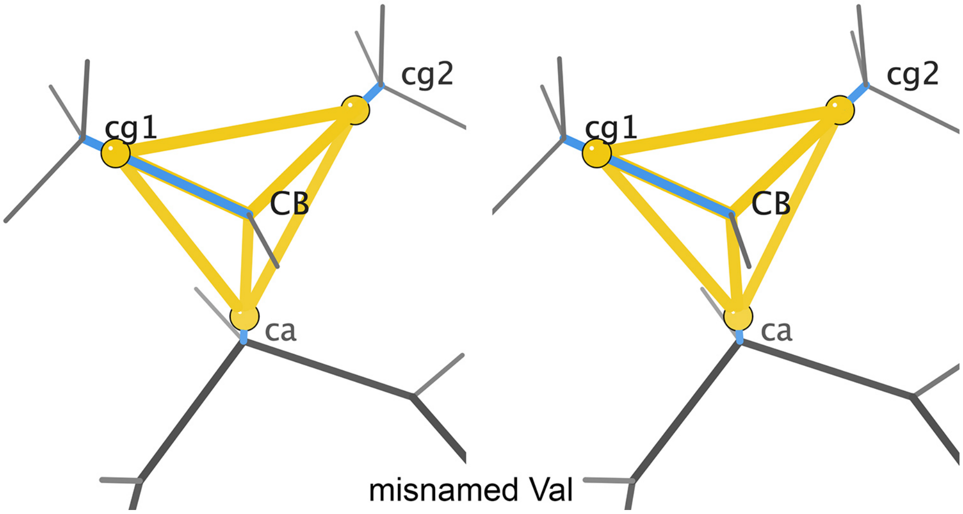
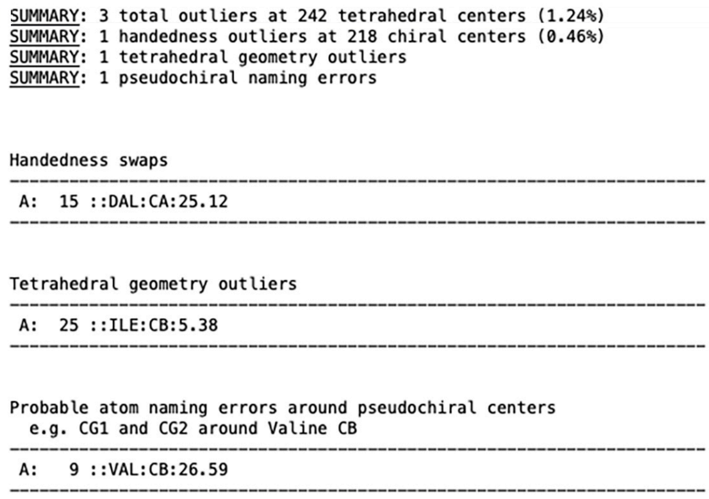


Figure 5: A pseudo-chiral atom-naming problem. The two branches of a Val sidechain are identical methyl groups, but the atoms need unique names. By pre-existing chemical convention, the righthand arm should be labeled as branch 1, but here the names cg1 and cg2 are assigned backward (cg1 on the left-hand branch) and therefore are flagged as a pseudo-chiral naming error.

branch (Cγ) is now considered #1, and Hb2 and Hb3 are named successively in the clockwise direction looking out the sidechain. If you ever need it, MolProbity still includes a utility for converting v2.7 to v3.0 format.



```
SUMMARY: 3 total outliers at 242 tetrahedral centers (1.24%)
SUMMARY: 1 handedness outliers at 218 chiral centers (0.46%)
SUMMARY: 1 tetrahedral geometry outliers
SUMMARY: 1 pseudochiral naming errors


Handedness swaps
-----------------------------------------------------------------
 A:  15 ::DAL:CA:25.12
-----------------------------------------------------------------


Tetrahedral geometry outliers
-----------------------------------------------------------------
 A:  25 ::ILE:CB:5.38
-----------------------------------------------------------------


Probable atom naming errors around pseudochiral centers
  e.g. CG1 and CG2 around Valine CB
-----------------------------------------------------------------
 A:   9 ::VAL:CB:26.59
-----------------------------------------------------------------
```

Figure 6: A chirals.txt report for an artificially constructed file with all 3 types of chiral outliers: handedness swaps, tetrahedral geometry, and pseudo-chiral naming.

Chiral problems in ligands, modified residues, and especially carbohydrates can happen in good, modern structures. Figure 7 shows a true chiral outlier at the C15 branchpoint of the YG 37 base in the anticodon loop of the 1ehz tRNA at 1.93Å resolution. It is in weak, patchy density, but it would be preferable to model the correct enantiomer. In complex carbohydrates, a chiral or pseudo-chiral outlier may often mean that either the wrong sugar or the wrong linkage type has been modeled. In Phenix, the carbohydrate libraries were recently re-analyzed and updated. The wwPDB now has much better carbohydrate validation available at deposition and has recently finished remediating carbohydrates in previous deposits.

As macromolecular structural biologists, we should all be grateful for the voluminous libraries of chemical and conformational

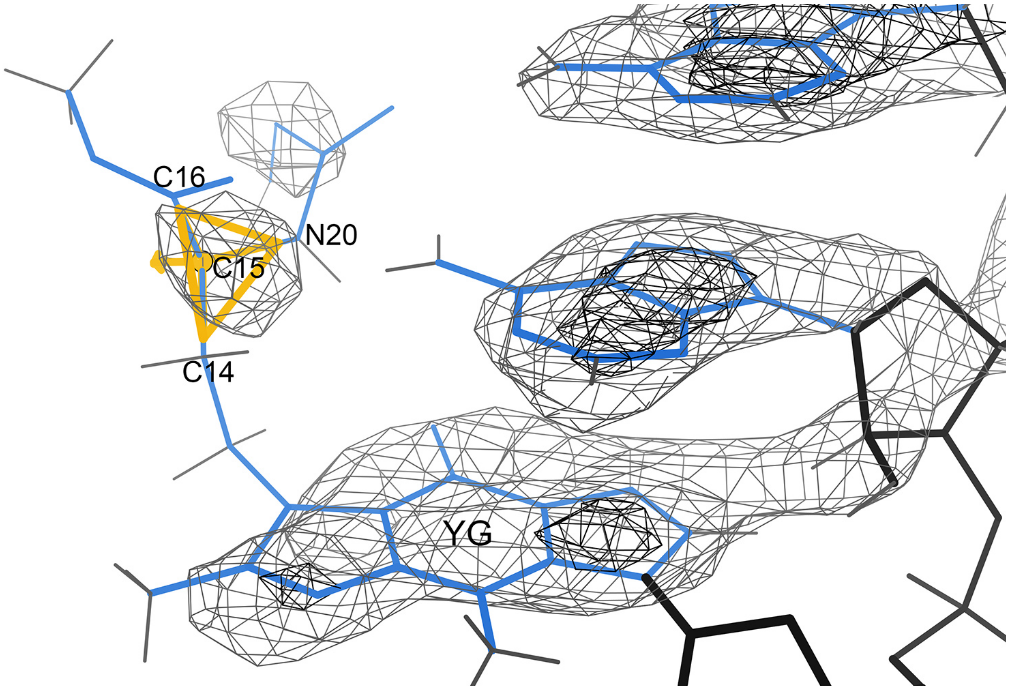


Figure 7: A true chiral outlier (yellow markup) at the C15 branchpoint in the YG 37 modified base of the 1.93Å 1ehz tRNA structure (Shi & Moore 2000). The local electron density suggests substantial disorder, but the two branches are different lengths with different atom types, so that in the other handedness there would be different possibilities for H-bonding with the neighboring bases.

constraints and especially to the work of chemistry, computation, and small-molecule crystallography that made those libraries possible. They are not infallible, but only rarely is the problem their fault.

The bottom line

Chiral outliers, and even pseudo-chiral outliers, occur very rarely when using modern software but are well worth flagging, understanding, and fixing when they do. Naming will of course be fixed by the annotators when you deposit your structure, but it is cleaner and more polite to do it yourself. Tetrahedral-geometry outliers are much more common and serious, and are flagged by the same chiral volume formalism. They almost always signal that the local group has been fit in the wrong conformation, which definitely should be rebuilt before final refinement.

References:

Banuelos S, Saraste M, Carugo KD (1998) Structural comparisons of calponin homology domains: implications for actin binding, *Structure* **6**: 1419-1431

Maksimainen M, Hakulinen N, Kallio JM, Timoharju T, Turunen O, Rouvinen J (2011) Crystal structures of Trichoderna reesii beta-galactosidase reveal conformational changes in the active site, *J Struct Biol* **174**: 156-163

Prisant MG, Williams CJ, Chen VB, Richardson JS, Richardson DC (2020) New tools in MolProbity validation: CaBLAM for cryoEM backbone, UnDowser to rethink "waters", and NGL Viewer to recapture online graphics *Prot Science* **29**: 315-329

Moriarty NW, Janowski PA, Swails JM, Nguyen H, Richardson JS, Case DA, Adams PD (2020) Improved chemistry restraints for crystallographic refinement by integrating the Amber force field into Phenix, *Acta Crystallogr* **D76**: 51-62

Shi H, Moore PB (2000) The crystal structure of yeast phenylalanine tRNA at 1.93Å: a classic structure revisited, *RNA* **6**: 1091-1105

FAQ

Can I have an angle restraint involving symmetry atoms?

The short answer is no. The main reason is that applying a symmetry operator is ambiguous in the case of three atoms compared to two atoms needed for a bond.

Updates from the Worldwide PDB: Celebrating PDB50 and PDBx/mmCIF news

Christine Zardecki
RCSB Protein Data Bank

Correspondence email: Zardecki@rcsb.rutgers.edu

Celebrating the 50th Anniversary of the Protein Data Bank

In 1971, the structural biology community established the single worldwide archive for macromolecular structure data–the Protein Data Bank (PDB). From its inception, the PDB has embraced a culture of open access, leading to its widespread use by the research community. PDB data are used by hundreds of data resources and millions of users exploring fundamental biology, energy, and biomedicine.

To commemorate and celebrate 50 years of the PDB, the wwPDB is organizing multiple events in 2021 (wwpdb.org/pdb50):

- The inaugural PDB50 event was held virtually in May 2021.
- Transactions Symposium 2021: Function Follows Form: Celebrating the 50th Anniversary of the Protein Data Bank (July 30-31, 2021).
  This virtual event is part of the Annual Meeting of the American Crystallographic Association.
- Bringing Molecular Structure to Life: 50 Years of the PDB (October 20-22, 2021) Virtual EMBL Conference
- Royal Society of Chemistry PDB Workshop (Nov 16 and 18, virtual)
- Learning from 50 years of the Protein Data Bank: A satellite symposium of the Biophysical Society of Japan (Nov 25-27, 2021)

Visit wwpdb.org/pdb50 for updates and related materials.

PDBx/mmCIF News

PDB users and related software developers should be aware of upcoming developments and plans related to the distribution of PDB data. Announcements are made at wwpdb.org.

Modifications to Support for SHEET and Ligand SITE records in June 2021

In 2014, PDBx/mmCIF became the PDB's archive format and the legacy PDB file format was frozen. In addition to PDBx/mmCIF files for all entries, wwPDB produces PDB format-formatted files for entries that can be represented in this legacy file format (e.g., entries with over 99,999 atoms or with multi-character chain IDs are only available in PDBx/mmCIF)

As the size and complexity of PDB structures increases, additional limitations of the legacy PDB format are becoming apparent and need to be addressed.

Defining complex SHEET records

Restrictions in the SHEET record fields in legacy the PDB file format do not allow for the generation of complex beta sheet topology. Complex beta sheet topologies include instances where beta strands are part of multiple beta sheets and other cases where the definition of the strands within a beta sheet cannot be presented in a linear description. For example, in PDB entry 5wln a large beta barrel structure is created from multiple copies of a single protein; within the beta sheet forming the barrel are instances of a single beta strand making contacts on one side with multiple other strands, even from different chains.

This limitation, however, is not an issue in the PDBx/mmCIF formatted file, where these complex beta sheet topology can be captured in *_struct_sheet*, *_struct_sheet_order*, *_struct_sheet_range*, and *_struct_sheet_hbond*.

Starting June 8th, 2021, legacy PDB format files will no longer be generated for PDB entries where the SHEET topology cannot be generated. For these structures, wwPDB will continue to provide secondary structure information with helix and sheet information in the PDBx/mmCIF formatted file.

Deprecation of _struct_site (SITE) records

wwPDB regularly reviews the software used during OneDep biocuration. The *_struct_site* and *_struct_site_gen* categories in PDBx/mmCIF (SITE records in the legacy PDB file format) are generated by in-house software and based purely upon distance calculations, and therefore may not reflect biological functional sites.

Starting in June 2021, the in-house legacy software which produces *_struct_site* and *_struct_site_gen* records will be retired and wwPDB will no longer generate these categories for newly-deposited PDB entries. Existing entries will be unaffected.

Consistent Format for Validation and Coordinate Data

wwPDB validation reports are now provided in PDBx/mmCIF format for all new depositions in OneDep. This change makes validation data more interoperable with the PDB archival format. Data are more logically and better organized in the PDBx/mmCIF reports, and therefore more "database-friendly" than the report in XML format. PDBx/mmCIF-format validation reports for newly released and modified entries will be distributed through the PDB and EMDB Core Archives.

The new PDBx/mmCIF reports are easier to interpret. They contain a high-level summary and offer easier access to residue-level information. Data are provided at multiple levels: entity, chain-specific, and even at the individual residues. For example, it is more straightforward to obtain the total number of clashes. The corresponding validation dictionary is available at mmcif.wwpdb.org/dictionaries/mmcif_pdbx_vrpt.dic/Index. Examples of PDBx/mmCIF validation reports for X-ray, 3DEM, and NMR are publicly available at GitHub.

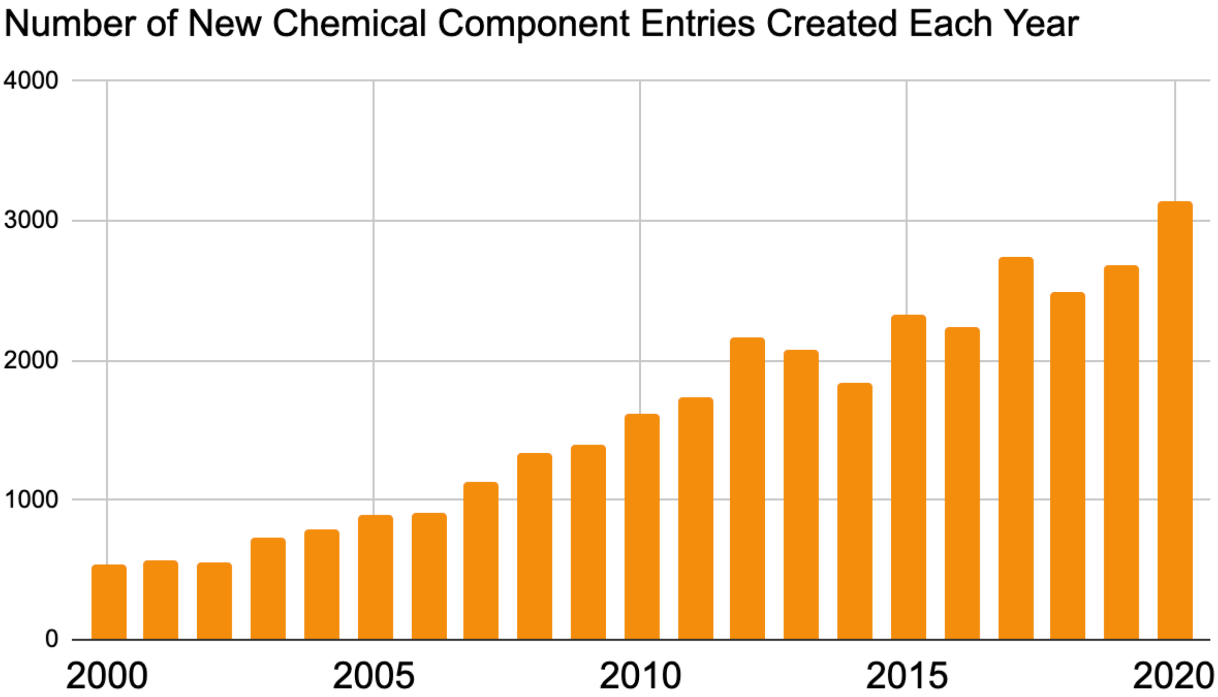
PDBx/mmCIF validation reports will be provided for the full PDB and EMDB archives once archival validation recalculation is performed.

wwPDB strongly recommends all PDB users and software developers adopt this format for future applications.

Future Planning: Entries with Extended PDB and CCD ID Codes will be Distributed in PDBx/mmCIF Format only

wwPDB, in collaboration with the PDBx/mmCIF Working Group, has set plans to extend the length of ID codes for PDB and Chemical Component Dictionary (CCD) ID entries in the future. Entries containing these extended IDs will not be supported by the legacy PDB file format.

CCD entries are currently identified by unique three-character alphanumeric codes. At current growth rates, we anticipate running out of



available new codes in the next three to four years. At this point, the wwPDB will issue four-character alphanumeric codes for CCD IDs in the OneDep system. Due to constraints of the legacy PDB file format, entries containing these new, four character ID codes will only be distributed in PDBx/mmCIF format. The wwPDB will begin implementation of extended CCD ID codes in 2022.

In addition, wwPDB also plans to extend PDB ID length to eight characters prefixed by ‘PDB’, e.g., pdb_00001abc. Each PDB ID has a corresponding Digital Object Identifier (DOI), often required for manuscript submission to journals and described in publications by the structure authors. Both extended PDB IDs and corresponding PDB DOIs, along with existing four character PDB IDs, will be included in the PDBx/mmCIF formatted files for all new entries by Fall 2021.

For example, PDB entry 1ABC will also have the extended PDB ID (pdb_00001abc) and the corresponding PDB DOI (10.2210/pdb1abc/pdb) listed in the _database_2 PDBx/mmCIF category.

```
loop_
_database_2.database_id
_database_2.database_code
_database_2.pdbx_database_accession
_database_2.pdbx_DOI
PDB 1abc pdb_00001abc
10.2210/pdb1abc/pdb
WWPDB D_1xxxxxxxxx ? ?
```

Once four-character PDB IDs are all consumed, newly-deposited PDB entries will only be issued extended PDB ID codes, and entries will only be distributed in PDBx/mmCIF format.

wwPDB is asking PDB users and related software developers to review code and begin to remove such limitations for the future.

Lessons from using the Cambridge Structure Database: III – Outlier rejection

Nigel W. Moriarty[a]*

[a]Molecular Biosciences and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

*Correspondence e-mail: nwmoriarty@lbl.gov

Preface

Continuing the series about lessons from using the Cambridge Structural Database (CSD), this work delves deeper into the nuances of data handling. More information about goals in the previous installment (Moriarty, 2020, 2021).

Introduction

The Cambridge Structural Database (CSD, Groom *et al.*, 2016) contains a wealth of small molecules that can be mined for geometry information. The tools in the CSD suite – Conquest (Bruno *et al.*, 2002), a structure based search tool, and Mercury (Macrae *et al.*, 2006, 2008), a data analysis tool – are flexible and highly featured making them ideal for their designated tasks.

The CSD is a curated data set leading to reliable entries. However, it is almost impossible to have consistent results from a particular search. Reasons for this may be user "error" as addressed in the first two editions of this series. More precise specification of the search structure will return the group of structures desired.

Another reason may be atypical or aberrant data in the specific entry. This could be an error that is corrupting the database or simply an example where a more nuanced search is required. Either way, it is more difficult to identify so filtering out these entries is desirable.

Outlier rejection

Thankfully, there is a technique that can help with database anomalies and user errors. Outlier rejection is the identification of outliers and removing them the analysed data. This is an active field of research with many techniques with various applications and effectiveness. In fact, the Mercury program has outlier identification.

One of the first signs of problems is an unusually large standard deviation of the bond lengths and/or bond angles. Theoretically, one can step though each entry (using Mercury) to "eyeball" for any issues but this gets tedious very quickly.

An example

Arginine is a charged essential amino acid containing the guanidinium moiety. Figure 1 shows the guanidinium group terminating the side chain with a positively charged central carbon atom and an electronic resonance bond structure. Note that the generally planar structure includes the central charged carbon atom, the three bound nitrogen atoms and the bound hydrogen atoms. A recent CSD structure search of the guanidinium group was reported as part of study (Moriarty *et al.*, 2020) into the planarity of the guanidinium group. The entries returned were analysed in a spreadsheet to enable more detailed study of features of the data. It should be noted that Mercury performs similar tasks but it also allows easy export in helpful formats for spreadsheet programs.

The standard deviations of several geometric features were considered too large so the tedious task and stepping though each entry was undertaken. Several incongruous entries were identified including GUACET shown in figure 2 produced by Mercury. The hydrogen atoms are not in the plane of the moiety. This is hypothesis not to happen but is evident in more than one case. Regardless of whether the planarity of the hydrogen atoms is correct or an error, the other geometric features are affected. This makes these entries outliers to the hypothesised geometry of the guanidinium.

One could remove them in this “eyeball” fashion but using an outlier rejection technique a uniform, defensible and efficient process. The selected technique was Tukey’s fences (Beyer, 1981) which removed all the examples discovered in the step through and a couple more that were also questionable. Based on the quadrature method, it was easy to program and gave similar results to the outlier identification in Mercury.

Conclusions

It has been a theme of this series to “Always verify that the results from a structure search are reasonable.” This installment provided insights into removing the “unreasonable.”

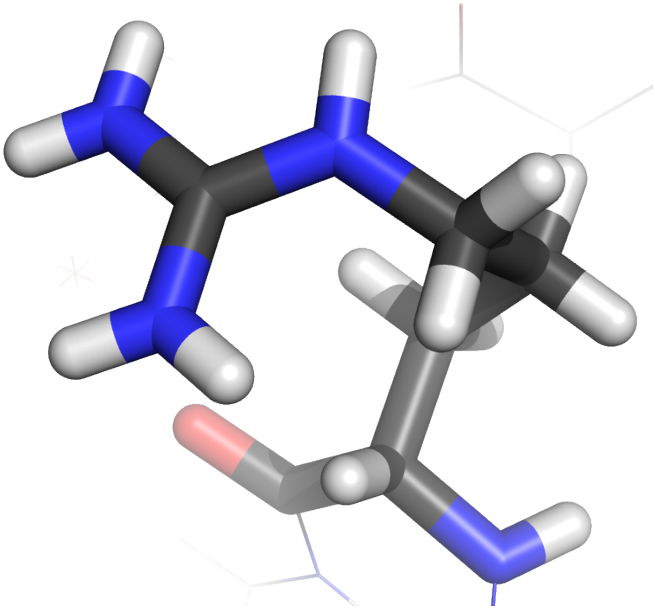


Figure 1: The arginine amino acid with the charged, planar guanidinium group in the upper left.

Coda

However, the inconsistencies of the entries removed are based on the hypothesis that the hydrogen atoms are planar. There is one entry, HOWHIK, that has out-of-plane hydrogen atoms but there is a SO_4^+ molecule that is attracting them to a far less non-planar positions than the example in figure 2. Clearly, the hydrogen atoms are affected by the nearby charge. This is an example of a more nuanced understanding of the guanidinium. Is it possible that the hydrogen atoms are more flexible? If so, by how much?

References

Beyer, H. (1981). *Biom. J.* **23**, 413–414.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Crystallogr. B*. **58**, 389–397.

Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179.

Macrae, C. F., Bruno, I. J., Chisholm, J. A., Edgington, P. R., McCabe, P., Pidcock, E., Rodriguez-Monge, L., Taylor, R., Streek, J. van de & Wood, P. A. (2008). *J. Appl. Crystallogr.* **41**, 466–470.

Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G. P., Taylor, R., Towler, M. & Streek, J. van de (2006). *J. Appl. Crystallogr.* **39**, 453–457.

Moriarty, N. W. (2020). *Comput. Crystallogr. Newsl.* **11**, 7–10.

Moriarty, N. W. (2021). *Comput. Crystallogr. Newsl.* **12**, 6–8.

Moriarty, N. W., Liebschner, D., Tronrud, D. E. & Adams, P. D. (2020). *Acta Crystallogr. Sect. Struct. Biol.* **76**, 1159–1166.

The effect of adding a single peptide bond class

Nigel W. Moriarty[1] and Paul D. Adams[1,2]

[1]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA

[2]Department of Bioengineering, University of California at Berkeley, Berkeley, CA

Introduction

Comprehensive restraints for refinement of protein structure were introduced by Engh & Huber (1991) for the standard amino acids. Gleaned from the Cambridge Structural Database (CSD, Groom *et al.*, 2016), the group of restraints (EH91) became the standard for crystallographic refinement forming the basis of the Monomer Library (Vagin *et al.*, 2004) used in *REFMAC* (Murshudov *et al.*, 2011) and BUSTER (Bricogne *et al.*, 2011) while also being available in the *Phenix* suite of programs (Liebschner *et al.*, 2019).

Briefly, the EH91 restraints provided ideal bond lengths and angles for each of the designated standard amino acids at that time. Generally, each geometry restraint's ideal was based on the identity of amino acid. Programmatically, the three-letter code of each amino acid was the key to a dictionary of bond and angle ideal values. That is, there is a single value for each bond and angle based on the amino acid that was used for each instance of that amino acid type in the macromolecule. This paradigm can be called a Single Value Library (SVL) as the bond and angle ideal values are set once.

Engh & Huber updated the restraints (2001) for the International Tables of Crystallography, Volume F, that has been designated EH99 elsewhere. One of the major changes in the EH99 restraints from the EH91 restraints was the recognition that *cis*-proline has different ideal values for most of the bonds and angles compared to the *trans* form. The largest difference is found in the linking angle C–N–Cα that increases from 122.6° in the *trans* form restraints (which were previously used for *cis*-PRO) to 127.0° in the *cis* (Fig. 1). The estimated standard deviation (e.s.d.) was reduced from 5° to 2.4°. This is quite a large change, effectively doubling the contribution of the restraint to the final target. Other PRO restraints were changed that are purely in the amino acid entity as shown in Fig. 1 alone with others not shown. This results in the PRO restraints being based on the peptide bond form in addition to the identity of the amino acid; a small step away from the SVL paradigm. Interestingly, in neither set of restraints do the sums of the angles around the nitrogen atom add to 360°. The EH91 restraints have a sum of 359.6° compared to 359.1° for EH99; arguably, negligible compared to the e.s.d values.

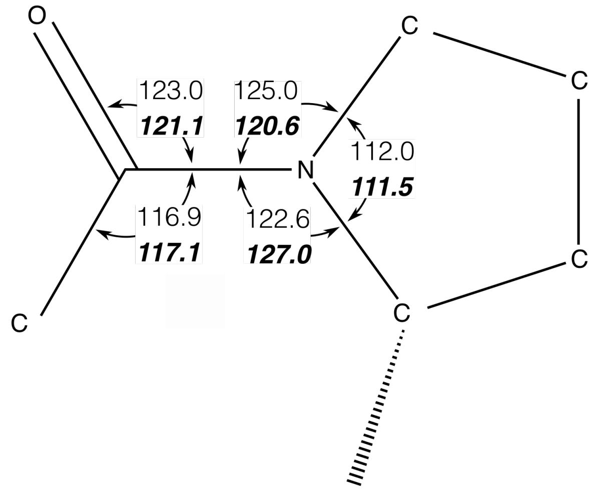


Figure 1: Diagram of a selected set of ideal angle values. EH99 values for *cis*-PRO are shown in bold italics with the EH91 values included for comparison.

More recent studies have investigated the influence of other factors on the ideal geometric values. One such study on the Conformation Dependent Library (CDL, Berkholz *et al.*, 2009) showed that the backbone geometry bond and angle ideal values depend on the ψ/ϕ angles of the backbone. The efficacy of the CDL was investigated (Moriarty *et al.*, 2014) by re-refining a large number of the entries available in the Protein Data Bank (PDB, Burley *et al.*, 2019) leading to the adoption of the CDL as the default (Moriarty *et al.*, 2016) in all *Phenix* packages. One caveat is that the CDL v1.2 is only for *trans*-peptides.

Despite the popularity of the EH91 restraints, the EH99 restraints were not implemented in the Monomer Library being absent in version 5.41. One can infer that no comprehension investigation of the influence of the EH99 *cis*-PRO restraints on protein refinement has been performed.

Approximately 5% of prolines are *cis*-PRO making the investigation into the addition of two sets of restraints for PRO nuanced.

Methods

To compare refinements using the EH99 *cis*-PRO restraints against EH91 restraints, the EH99 restraints were implemented in *Phenix* for use in all programs. Technically, the generic mechanism using the *cif_link* and *cif_mod* in the Monomer Library could have been used to add the EH99 *cis*-PRO restraints to *Phenix*, however, because of the CDL implementation there was an opportunity to implement a more flexible algorithm by using the CDL infrastructure.

To test the restraint libraries, structures were selected from the PDB using the following criteria. Entries must have untwinned experimental data available that are at least 90% complete. Each entry's R_{free} was limited to a maximum of 35%, R_{work} to 30% and the **Δ**R (R_{free}-R_{work}) to a minimum of 1.5%. Entries containing nucleic acids were excluded.

Each model was then subjected to 10 macrocyles of refinement using the default strategy in *phenix.refine* for reciprocal space coordinate refinement. Other options applied to both EH99 and EH91 refinements included optimization of the weight between the experimental data and the geometry restraints. This protocol was performed in parallel. The quality of the resulting models was assessed numerically using MolProbity (Williams *et al.*, 2018) available in *Phenix*. To avoid typographical ambiguity, PDB codes are given here with lower case for all letters except L (e.g., 1nLs). Post-refinement filtering removed refined models that exceeded a *clashscore* of 12.

Results & Discussion

As previously stated, the *cis* peptide link occurs in approximately 5% of prolines. This implies that the change will not be reflected in global measures like the R factors. This is indeed true. The same is true for many of the other validation metrics reported by Molprobity.

One metric reported by Molprobity and PDB alike is the root mean squared deviation (r.m.s.d.) values for the bond and angle restraints in the refined model compared to the ideal values of the restraints. Another similar metric is the r.m.s.Z values that use the e.s.d. values of the restraints to calculate the number of standard deviations from the mean – the Z-score.

Both the r.m.s.d. and r.m.s.Z values will be largely unaffected by the modified *cis*-PRO restraints if the entire model is compared. Focusing the scope of the metrics has been demonstrated to provide validation of new restraints for iron-sulfur clusters (Moriarty & Adams, 2019) and arginine (Moriarty *et al.*, 2020). The latter has a detailed discussion of the nuances of validating single amino acid restraints as well as applying the metrics to other internal coordinates such as torsion angles.

For this case, the focus is ever tighter – just the *cis*-PRO instances in the models. Figure 2 shows the comparison of the r.m.s.d. values for the EH91 restraints denoted “122” to indicate the approximate ideal angle for C–N–Cα and EH99 denoted “127” for the new angle ideal value. The results for the entire models are shown as dashed and dotted lines but have negligible differences. Notwithstanding, the *cis*-PRO restraints (denoted “PRO 122” and “PRO 127”) have significant differences. All bond r.m.s.d. values are similar at resolution worse than 2Å. At better than 2Å, the *cis*-PRO entities have smaller r.m.s.d. values. This change is not based on the new *cis*-PRO restraints as both the old and new restraints are very similar.

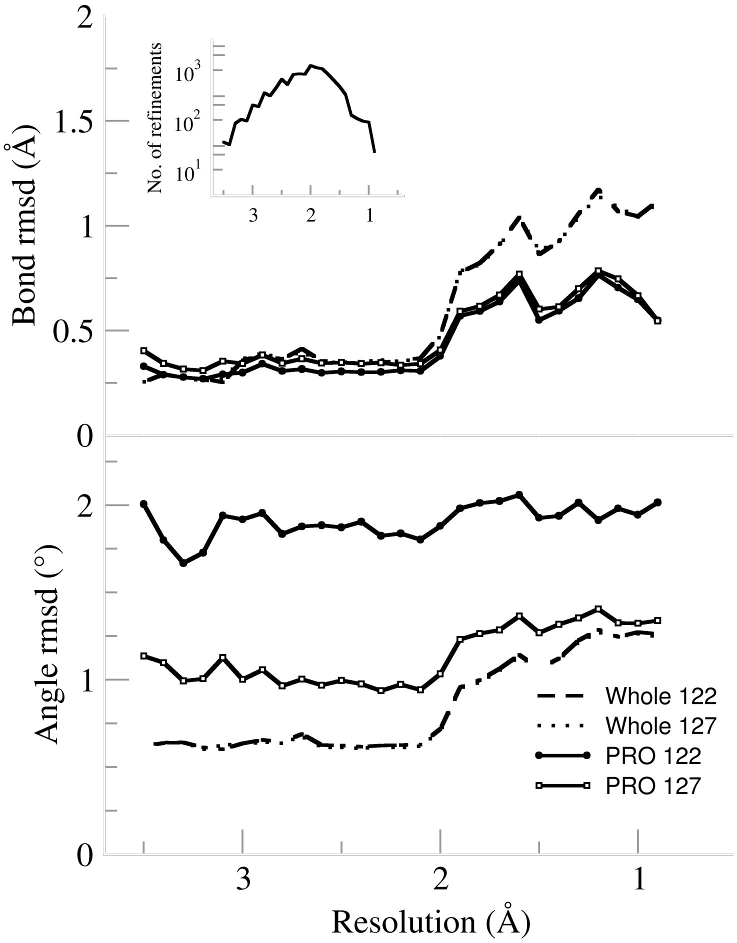


Figure 2: Bond and angle r.m.s.d. values averaged in 0.1Å bins. The r.m.s.d. values for the whole model are shown in dashed and dotted lines, while for the *cis*-PRO r.m.s.d. values are solid lines. Refinements with original EH91 restraints are denoted by solid circle markers (*cis*-PRO only) and EH99 restraints are denoted with open circle markers (*cis*-PRO only). Inset shows the number of refinements in each resolution bin.

Understandably, because the angle ideal values have a far greater change than the bond ideal values. the r.m.s.d. values for the angles are affected to a much greater extent. Not difference is detectable in the values for the whole models but the r.m.s.d. values for just the *cis*-PRO differ by approximately 1°. The EH91 values are uniformly approximately 2° across all resolutions. This uniformity indicates that the EH91 restraints are not suitable as the geometries do not approach the ideal values as the experimental data has less information (low resolution). For the EH99 (PRO 127), the r.m.s.d. values are approximately 1° at low resolution increasing to 1.3° at higher resolution reflecting the experimental data information. This trend is also inline with the values for entire model indicating a more balanced set of restraints.

Figure 3 shows the r.m.s.Z results in a similar format as Fig. 2. Similarly, the bond values have very little differentiating between the two sets of restraints. By contrast, the angle r.m.s.Z values for the angles are informative. At higher resolutions, the EH99 (PRO 127) restraints result in similar r.m.s.Z values for both the whole models and the *cis*-PRO indicating a balance. Tellingly, the r.m.s.Z values for the EH91 refinements are approximately 0.1 larger at all resolutions even though the e.s.d. for the angle was reduced by half in the EH99 restraints. This implies that the larger e.s.d. was necessary in the earlier restraints to cover the correct ideal angle value.

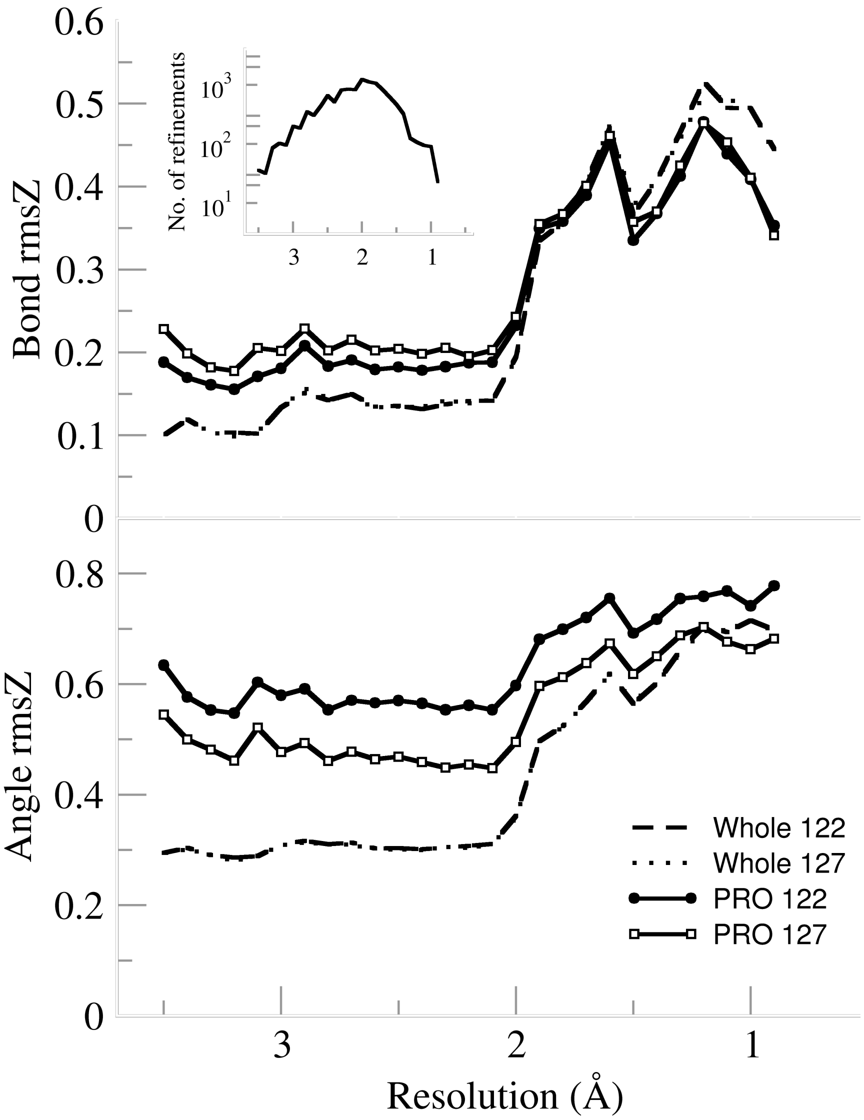


Figure 3: Bond and angle r.m.s.Z values averaged in 0.1Å bins. The r.m.s.Z values for the whole model are shown in dashed and dotted lines, while for the *cis*-PRO r.m.s.Z values are solid lines. Refinements with original EH91 restraints are denoted by solid circle markers (*cis*-PRO only) and EH99 restraints are denoted with open circle markers (*cis*-PRO only). Inset shows the number of refinements in each resolution bin.

A more focused view of the behavior of the restraints for the *cis*-PRO entities appears in Figure 4. The graph is a comparison of the deviations of the refined C–N–Cα angle values from the ideal specified by the restraints. Error bars are placed at the standard error of measurement values. As expected from the results shown in both the r.m.s.d. and r.m.s.Z figures, the r.m.s.d. values of the specific angle using the EH99 restraints are smaller than the earlier released restraints. Specifically, the EH99 values are less than 2° while the EH91 values hover around 5°. This is an affirmation that the latter restraints are an improvement. A counter argument is the increase in r.m.s.d. values at low resolution for both sets of restraints. There must be other forces (restraints) at play.

Conclusion

The subtle differences between the overall results using the EH91 and EH99 restraints hide the large improvement of the *cis*-PRO entities. The metrics indicate that the *cis*-PRO entities have better geometries (lower r.m.s.d. values) using the EH99 restraints. Even though only 5% of PRO are *cis*-peptides, clearly, any improvement in the restraints will help generate more accurate models but there appears to be room for improvement in the area of *cis*-PRO restraints.

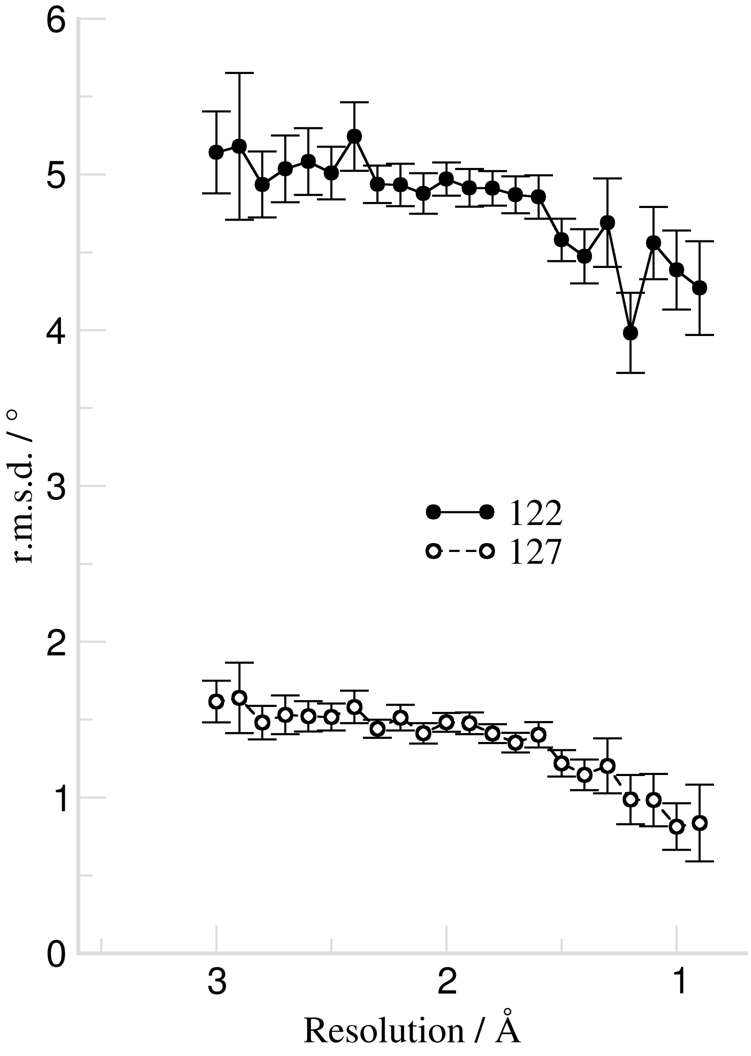


Figure 4: Deviations of the C–N–Cα angle values from the ideal value in 0.1Å bins.

References

Berkholz, D. S., Shapovalov, M. V., Dunbrack, Jr., R. L. & Karplus, P. A. (2009). *Structure*. **17**, 1316–1325.

Bricogne, G., Blanc, E., Brandi, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P., Sharff, A., Smart, O. S., Vonrhein, C. & Womack, T. O. (2011). BUSTER Cambridge, United Kingdom: Global Phasing Ltd.

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Costanzo, L. D., Christie, C., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranovic, V., Guzenko, D., Hudson, B. P., Liang, Y., Lowe, R., Peisach, E., Periskova, I., Randle, C., Rose, A., Sekharan, M., Shao, C., Tao, Y.-P., Valasatava, Y., Voigt, M., Westbrook, J., Young, J., Zardecki, C., Zhuravleva, M., Kurisu, G., Nakamura, H., Kengaku, Y., Cho, H., Sato, J., Kim, J. Y., Ikegawa, Y., Nakagawa, A., Yamashita, R., Kudou, T., Bekker, G.-J., Suzuki, H., Iwata, T., Yokochi, M., Kobayashi, N., Fujiwara, T., Velankar, S., Kleywegt, G. J., Anyango, S., Armstrong, D. R., Berrisford, J. M., Conroy, M. J., Dana, J. M., Deshpande, M., Gane, P., Gáborová, R., Gupta, D., Gutmanas, A., Koča, J., Mak, L., Mir, S., Mukhopadhyay, A., Nadzirin, N., Nair, S., Patwardhan, A., Paysan-Lafosse, T., Pravda, L., Salih, O., Sehnal, D., Varadi, M., Vařeková, R., Markley, J. L., Hoch, J. C., Romero, P. R., Baskaran, K., Maziuk, D., Ulrich, E. L., Wedell, J. R., Yao, H., Livny, M. & Ioannidis, Y. E. (2019). *Nucleic Acids Res.* **47**, D520–D528.

Engh, R. & Huber, R. (1991). *Acta Crystallogr. Sect. A*. **47**, 392–400.

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

Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Crystallogr. Sect. Struct. Biol.* **75**, 861–877.

Moriarty, N. W. & Adams, P. D. (2019). *Acta Crystallogr. Sect. Struct. Biol.* **75**, 16–20.

Moriarty, N. W., Liebschner, D., Tronrud, D. E. & Adams, P. D. (2020). *Acta Crystallogr. Sect. Struct. Biol.* **76**, 1159–1166.

Moriarty, N. W., Tronrud, D. E., Adams, P. D. & Karplus, P. A. (2014). *FEBS J.* **281**, 4061–4071.

Moriarty, N. W., Tronrud, D. E., Adams, P. D. & Karplus, P. A. (2016). *Acta Crystallogr. Sect. -Biol. Crystallogr.* **72**, 176–179.

Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Crystallogr. Sect. -Biol. Crystallogr.* **67**, 355–367.

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. & Murshudov, G. N. (2004). *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2184–2195.

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (2018). *Protein Sci.* **27**, 293–315.

The top2018 pre-filtered dataset of high-quality protein residues

Christopher J. Williams, David C. Richardson and Jane S. Richardson

Department of Biochemistry, Duke University, Durham, NC 27710

Correspondence email: dcrjsr@kinemage.biochem.duke.edu

Introduction

This article announces the recent release on Zenodo of a large, high-quality reference dataset of PDB-format coordinate files from which all residues with low model certainty have been removed. Each file is a single protein chain while the total set of files were selected for low redundancy, high resolution, good MolProbity score and other chain-level criteria. Residue-level validation is even more important than overall validation, but only recently has it become feasible to distribute reference datasets in this pre-filtered form.

Our laboratory has emphasized the importance of residue-level as well as chain-level quality filtering of reference datasets as a foundation for model validation and for further bioinformatic structural studies. We began such work in the late 1990s when we introduced our flagship validation of all-atom contact analysis based on the Top100 dataset of reference protein chains, which in our own use we filtered at the residue level on any atomic B-factor >40 (Word 1999). We made available the list for those 100 chains and for all our subsequent, increasingly larger reference datasets (8000 chains by 2013), but had to leave the application of B cutoffs to the user. After deposition of structure factors became required, our validations used explicit electron-density filters for map value and correlation coefficient at each atom, as well as B-factor, all-atom clash and covalent-geometry filters, but we still found no feasible mechanism for distributing all the coordinate files with residue-filter annotations.

Our residue-level quality filtering process relies on extensive infrastructure, especially our developer team's integration into the Phenix software project (Liebschner 2019). We also now manage the filtering information with a Neo4j graphical database (Yoon 2017; Webber 2020). We have switched to using a graphical database to store our reference data because sequence connectivity is modeled natively there (but cumbersome in relational databases), as are the cyclic graphs that define local structural motifs.

The recent breakthrough in our ability to distribute coordinate files in a residue-filtered mode has been enabled by two things. First is our realization that making residue-level quality filtering easily available is worth giving up user flexibility in setting filter thresholds. Second, even more important, is the Zenodo online service that hosts open access to very large, DOI-identified datasets (Sicilia 2017). We have now taken advantage of that venue to distribute our current residue-level pre-filtered datasets. This development allows other researchers to make full and proper use of our curated reference data without needing the expertise, infrastructure and effort required to perform residue-level quality-filtering themselves.

Here we outline the production of this high-quality Top2018 (~15,000-chain) protein dataset and announce the availability of two residue-level pre-filtered versions suitable for general use with little or no further modification. One set is residue-filtered on mainchain criteria and the other on both mainchain and sidechain criteria. Each set is available at 30%, 50%, 70% and 90% sequence-identity levels. The filtered-out residues leave gaps in the chain, but the remaining high-reliability fragments are surprisingly long –mostly 20-30 residues or more.

Chain selection

We assembled a set of high-quality, low-redundancy protein chains. Chains were selected for consideration from the Protein Data Bank on the following criteria:

- Chain is protein
- Sequence length ≥ 38 residues
- Parent structure solved with x-ray crystallography
- Parent structure solved at better than 2.0Å resolution
- Parent structure has deposited structure factors
- Parent structure deposited on or before December 31, 2018

These chains were analyzed with our validation statistics and chains that failed the following criteria were removed from consideration:

- MolProbity score < 2.0
- <3% of residues have Cβ deviations
- <2% of residues have covalent bond length outliers
- <2% of residues have covalent bond angle outliers

The remaining chains were treated within their PDB-defined sequence-identity clusters, which are calculated weekly with MMseqs2 (Steinegger & Soedling 2018). From each cluster, we selected the chain with the best (lowest) average of resolution and MolProbity score as the best-quality representative of that cluster.

The PDB provides homology clustering at several different levels of stringency. We prepared sets of chains at the 90%, 70%, 50% and 30% sequence-identity levels. (90% is the most permissive, allowing as much as 90% sequence homology between the representatives from different clusters. 30% is the most restrictive, grouping chains into fewer clusters with greater differences between clusters.)

Residue-level filtering

While the selected chains are of good overall quality, this does not guarantee that all residues in them are modeled at high quality with high confidence (Figure 1). Therefore, we applied a residue-level filtering process. Two different residue-filtered sets were created, one filtered just on the mainchain and one filtered on the full residue, including the sidechains. The mainchain filtering considered the atoms N, Cα, C, O and Cβ. Cβ is included with the mainchain atoms since its ideal position is determined solely from other mainchain atom positions. The full-residue filtering considered both mainchain and sidechain heavy atoms. Attached Hydrogen atoms were considered for all-atom contact analysis. Hydrogen atoms were not considered in fit-to-map analyses, as their signal in the map is generally weak or absent.

For a residue to be included in the final dataset, all atoms under consideration had to meet the following criteria:

- B-factor < 40
- Real-space correlation coefficient (rscc) > 0.7

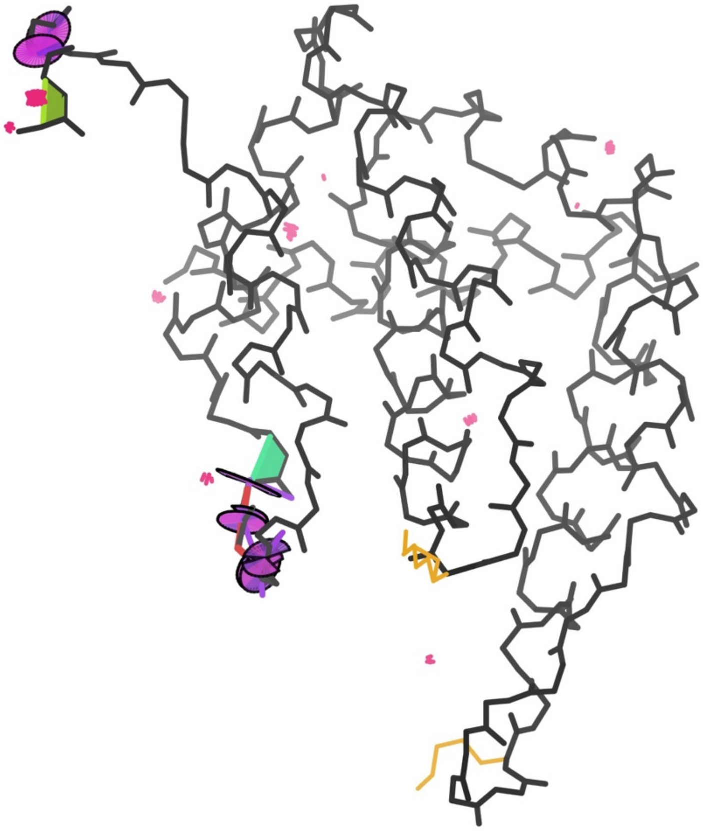


Figure 1: *3D distribution of quality in 5Lp0*

The structure 5Lp0 demonstrates a typical distribution of structure quality for models included in the Top2018. Most of the model is reliable and free from outliers, but two short regions contain a concentration of significant outliers. These problematic regions should not be blindly accepted with the rest of the structure.

- 2mFo-DFc map value at atom position > 1.2 sigma
- No covalent geometry outliers involving those atoms
- No steric clashes involving those atoms
- No alternate modeling conformations for those atoms

All atoms from residues that failed any of these criteria were removed from the PDB files.

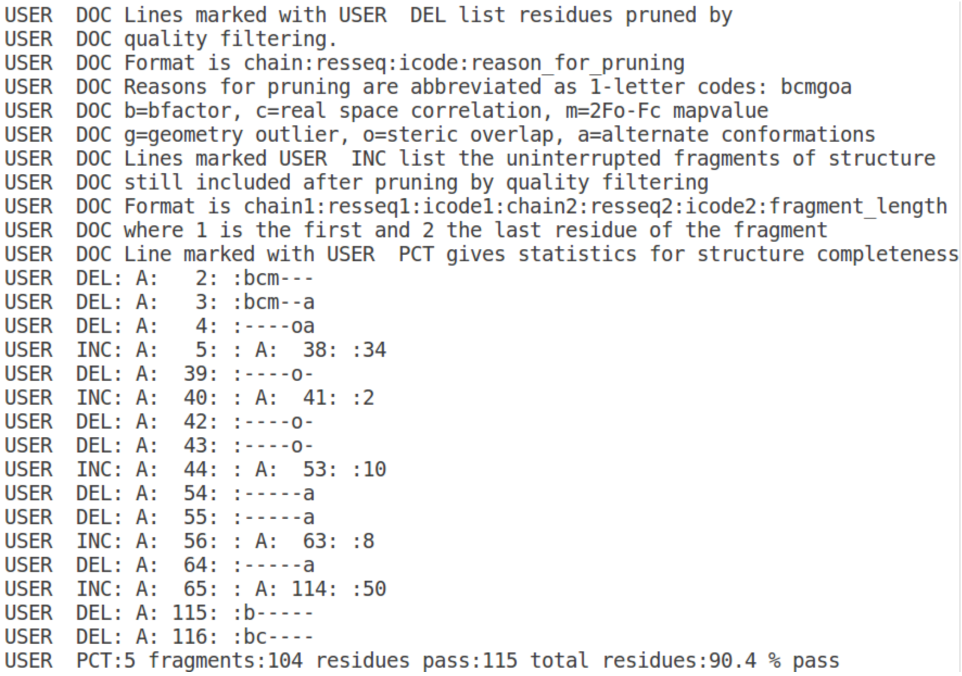
The fit-to-map criteria (B-factor, rscc and map value) were obtained using:

```
phenix.real_space_correlation
detail=atom
```

Fit-to-map assessment could not be performed for some structures due to bad MTRIX records or other data issues. Chains from those structures were discarded. The B-factor, rscc and map cutoffs were those developed during production of a rotamer library using our previous, top8000 dataset (Hintze, 2016).

Chains that were < 60% complete after residue filtering were discarded from the final dataset. This serves as a final check on overall structure quality and reduces the amount of chain fragmentation in the included chains.

Only protein residues were filtered. Individual filtering of ligands, ions and waters is beyond



```
USER  DOC Lines marked with USER  DEL list residues pruned by
USER  DOC quality filtering.
USER  DOC Format is chain:resseq:icode:reason_for_pruning
USER  DOC Reasons for pruning are abbreviated as 1-letter codes: bcmgoa
USER  DOC b=bfactor, c=real space correlation, m=2Fo-Fc mapvalue
USER  DOC g=geometry outlier, o=steric overlap, a=alternate conformations
USER  DOC Lines marked USER  INC list the uninterrupted fragments of structure
USER  DOC still included after pruning by quality filtering
USER  DOC Format is chain1:resseq1:icode1:chain2:resseq2:icode2:fragment_length
USER  DOC where 1 is the first and 2 the last residue of the fragment
USER  DOC Line marked with USER  PCT gives statistics for structure completeness
USER  DEL: A:   2: :bcm---
USER  DEL: A:   3: :bcm--a
USER  DEL: A:   4: :----oa
USER  INC: A:   5: : A:  38: :34
USER  DEL: A:  39: :----o-
USER  INC: A:  40: : A:  41: :2
USER  DEL: A:  42: :----o-
USER  DEL: A:  43: :----o-
USER  INC: A:  44: : A:  53: :10
USER  DEL: A:  54: :-----a
USER  DEL: A:  55: :-----a
USER  INC: A:  56: : A:  63: :8
USER  DEL: A:  64: :-----a
USER  INC: A:  65: : A: 114: :50
USER  DEL: A: 115: :b-----
USER  DEL: A: 116: :bc----
USER  PCT:5 fragments:104 residues pass:115 total residues:90.4 % pass
```

Figure 2: *In-file documentation of the residue-level quality filtering.*

Each filtered .pdb file ends with USER DEL records that document the residues that were removed and the reasons for removal as a 6-letter string, USER INC records for the residue stretches that remain and an explanation of the formatting for these records.

the current scope of this dataset. Ligands, ions and waters are included in these files in the interest of completeness, but no guarantee of their quality is implied.

In-file Documentation

The results of residue-level filtering are documented in each resulting .pdb file in USER records appended to the end of the file (Figure 2). These records report the residues that were removed and the reasons for their removal (as a string of 6 single-letter codes), the residues that remain and the lengths of the sequence fragments they form and the overall completeness statistics for the filtered file. See the self-documentation in these USER records for full details.

Importance of Residue Filtering

A key fact that motivated preparation of these datasets is that good average model quality across a whole structure is nevertheless compatible with extremely bad model quality in locally disordered regions with poor density. Familiar cases of this are mobile, unresolved sidechains on a protein's surface compared to well-packed sidechains in a protein's core and unseen backbone at chain termini or in disordered loops.

The CCTBX community may remember the crisis of overabundant *cis*-non-prolines some years ago (Croll, 2015). This phenomenon was pronounced at lower resolutions, but is present in poorly-resolved regions of even very high-resolution structures. Residue-level filtering guards against the inclusion of incorrectly-modeled *cis*-nonPro peptides, on both a statistical and an individual level.

Before filtering, the 70% homology set of the top2018 contained 1959 *cis*-nonPro out of

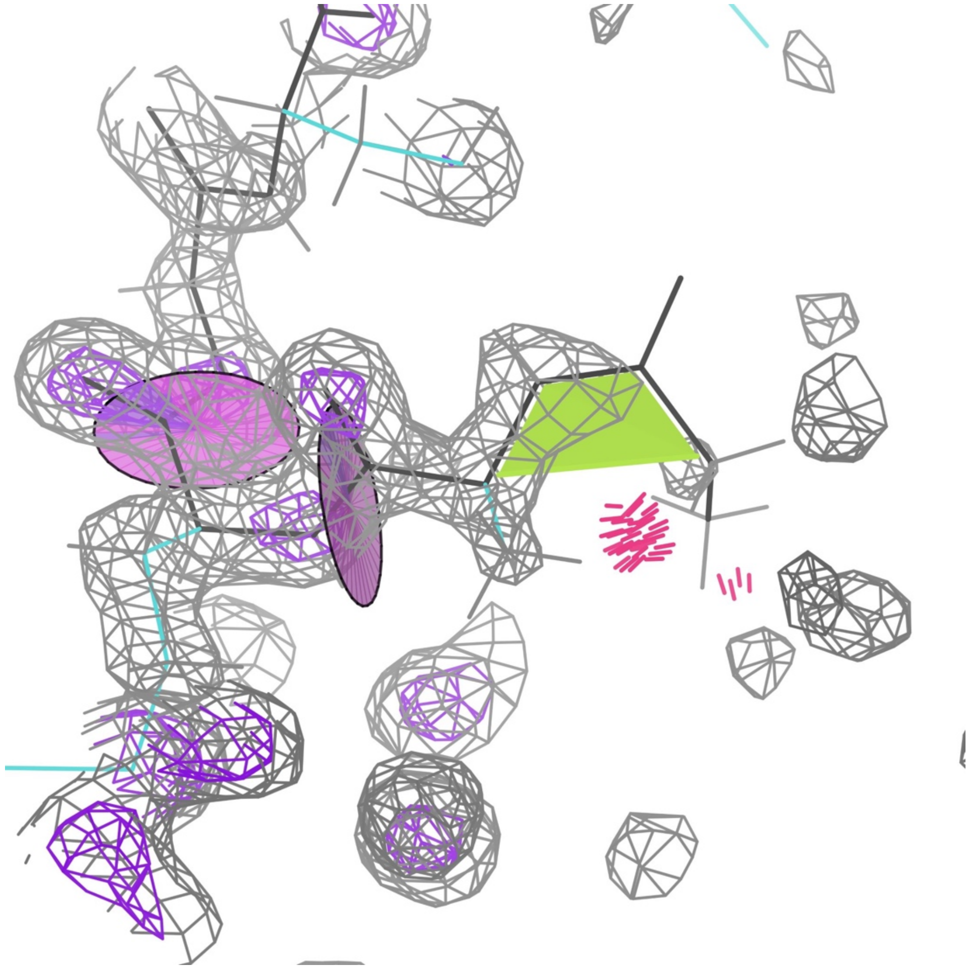


Figure 3: *Erroneous cis-non-proline in 5Lp0*

5Lp0 contains a *cis*-alanine modeled at its N-terminus. The sparse electron density at the terminus is misleading, creating a temptation to model this conformation, but providing no justification for it. Since there is nothing to hold them in place, *cis* conformations at termini are always modeling errors. This residue fails our fit-to-map criteria and has a steric clash. It is therefore removed from the file during filtering.

3,324,246 evaluable peptide bonds, for an occurrence rate of 0.048% or about 1 in 2000 (a rate often reported before any data-quality controls). After filtering, there remain 776 *cis*-nonPro out of 2,652,118, for an occurrence rate of 0.029% or about 1 in 3500. This lower rate agrees with recent observations of valid *cis*-nonPro occurrence (Williams 2018b).

More importantly than these general statistics, residue-level filtering removes many obviously incorrect *cis*-nonPro peptides from the dataset. These include some known, systematic patterns of incorrect *cis* modeling, such as building *cis*-peptides into the truncated density at chain termini (Figure 3). *Cis*-peptides are particularly valuable to filter out, as they tend to be modeled into regions of low certainty (Figure 4). The lack of strong electron density in such regions *allows* this and other modeling errors to occur. It is vital to the health of a statistical reference dataset, homology model, or fragment library to remove these regions of poor and/or unsupported model, as we do in this dataset.

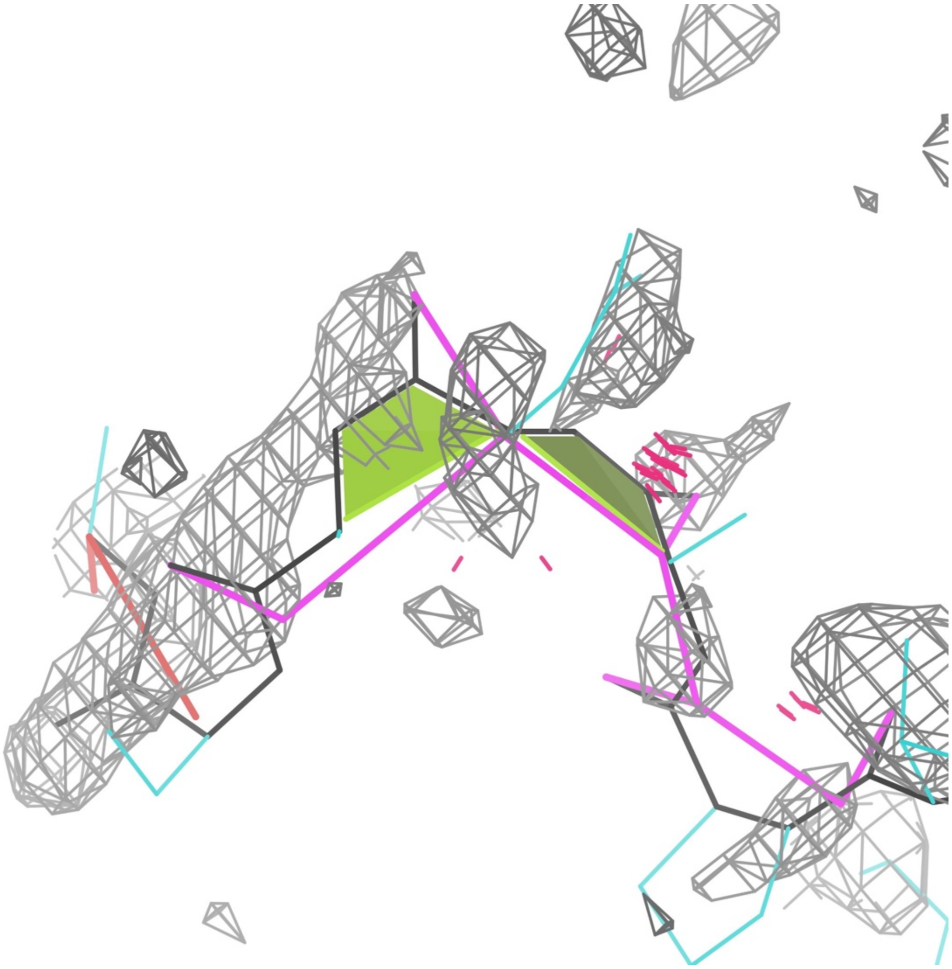


Figure 4: *Double* cis-*non-Pro in 4rm4*

Residue-level filtering also catches and removes this badly-resolved region of 4rm4, including residues 170-172, which form two successive unsupported *cis*-nonPro peptides. This region is clearly not a reasonable interpretation of even this minimal density and should not be allowed to influence future models or statistics. Multiple successive *cis*-nonPro are also a recognized systematic error never seen in genuine cases. Magenta lines show CaBLAM outliers (Williams 2018a) and a cluster of hotpink spikes shows a steric clash ≥4.0Å.

Residue-level filtering thus ensures that the population of *cis*-nonPro peptides is not statistically or locally overrepresented due to modelling errors. The *cis*-nonPro that remain in the dataset (Figure 5) do so based on a reasonable standard of map and model quality and can be used in fragment-based methods or the like with confidence (although we would still advise reasonable statistical weighting).

Conclusions

The full-coordinate, residue-filtered reference datasets described here omit all residues that fail the quality filters, so that they contain only coordinates for residues which are almost certainly correct. The full-residue quality-filtered reference dataset can be used to prepare protein sidechain rotamer libraries (Lovell 2000; Hintze 2016) or to study macromolecular structural motifs that span multiple residues and involve backbone-

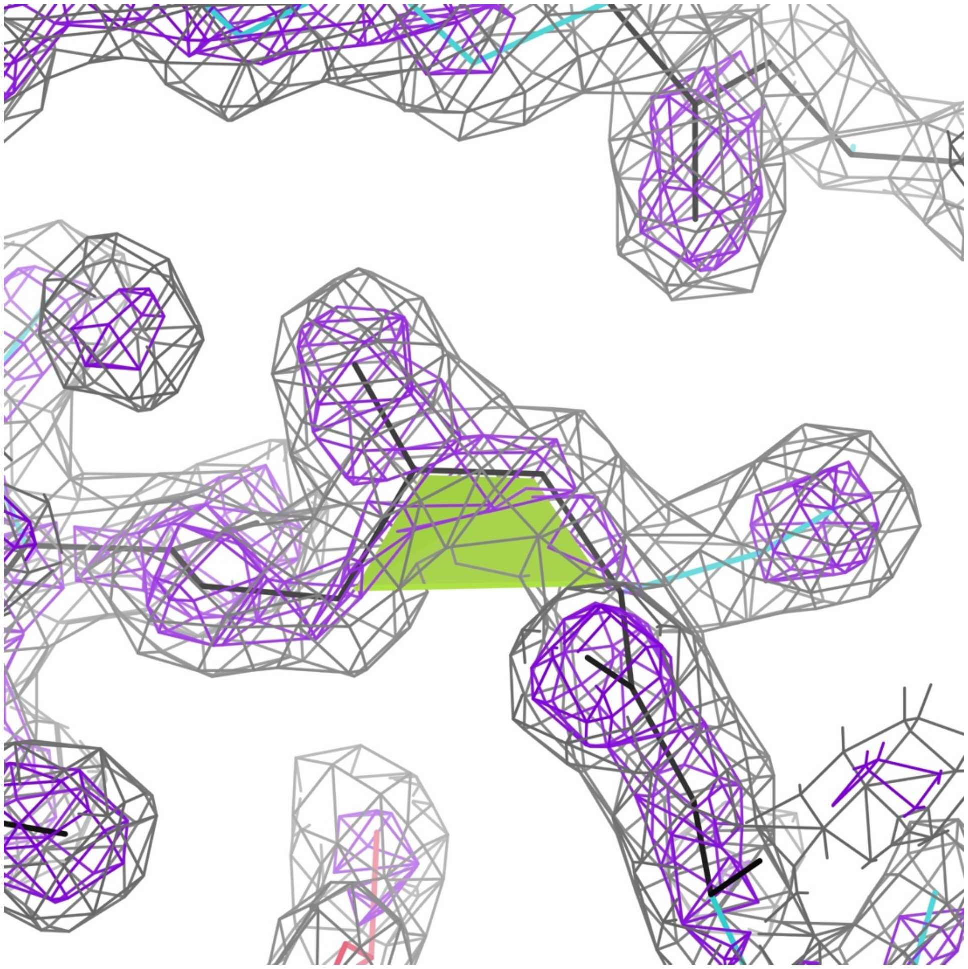


Figure 5: *Supported* cis-*nonPro in 6btf*

This *cis*-serine, residue 275 of 6bft, passes our quality criteria and is included in the structure after filtering. The 1.6Å density is persuasive and well fit by the model. Residues like this that pass our filtering are not guaranteed to be correct, but are guaranteed to meet acceptable quality standards.

sidechain interactions (Videau 2004; Richardson, 2013). The mainchain residue-filtered reference dataset can be used to define Ramachandran distributions (Lovell2003; Read2011) and to prepare curated fragment libraries for model-building or for protein design (Leaver-Fay 2013; Williams2015; Williams2018). In contrast, these gapped, residue-filtered datasets are not suitable for applications that require the full local context, such as Voronoi analyses or all-atom contacts.

Availability

These datasets are available on the Zenodo data repository, each at four levels of sequence redundancy.

The mainchain-filtered set is here: https://doi.org/10.5281/zenodo.4626149.

The full-residue-filtered set is here: https://doi.org/10.5281/zenodo.5115232.

Zenodo supports versioning and these links will resolve to the latest version of each dataset.

References

Croll TI (2015). The rate of *cis–trans* conformation errors is increasing in low-resolution crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, *71*(3), 706-709.

Hintze BJ, Lewis SM, Richardson JS, Richardson DC (2016). Molprobity's ultimate rotamer-library distributions for model validation. *Proteins: Structure, Function, and Bioinformatics*, *84*(9), 1177-1189.

Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lysko S, Gray JJ, Kortemme T, Richardson JS, Havranek JJ, Snoeyink J, Baker D, Kuhlman B (2013). Chapter Six - Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. *Methods in Enzymology*, 523. 109-143.

Liebschner D, Afonine PV, Baker ML, Bunkoczi G, Chen VB, Croll TI. Hintze BJ, Hung L-W, Jain S, McCoy AJ, Moriarty NW, Oeffner RD, Poon BK, Prisant MG, Read RJ, Richardson JS, Richardson DC, Sammito MD, Sobolev OV, Stockwell DH, Terwilliger TC, Urzhumtsev AG, Videau LL, Williams CJ, Adams PD (2019). Macromolecular structure determination using X-rays, neutrons, and electrons: Recent developments in Phenix, *Acta Cryst* **D75**: 861-877

Lovell SC, Word JM, Richardson JS, Richardson DC (2000). The penultimate rotamer library. *Proteins*, 40(3):389-408.

Read RJ, Adams PD, Arendall WB III, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissine EB, Lutteke T, Otwinowski Z, Perrakis A, Richardson JS, Sheffler WH, Smith JL, Tickle IJ, Vriend G, Zwart PH (2011). A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure*, 19(10) 12 October 1395-1412.

Sicilia MA, García-Barriocanal E, Sánchez-Alonso S (2017). Community Curation in Open Dataset Repositories: Insights from Zenodo*. Procedia Computer Science*. 106: 54-60.

Steinegger M, Soedling J (2018) Clustering huge protein datasets in linear time, *Nature Communications*, doi: 10.1038/s41467-018-04964-5.

Videau LL, Arendall WB III, Richardson JS (2004). The Cis Pro Touch-Turn: a Rare Motif Preferred at Functional Sites. *Proteins*, 56, 298-309.

Richardson JS, Keedy DA, Richardson DC (2013) "The Plot thickens: more data, more dimensions, more uses", pp. 46-61 in Biomolecular Forms and Functions: A Celebration of 50 Years of the Ramachandran Map, ed. Bansal M, Srinivasan N, World Scientific Publishing, Singapore.

Webber J, Van Bruggen R (2020) Graph Databases for Dummies, Neo4j Special Edition. John Wiley & Sons, ISBN: 978-1-119-74602-7.

Williams CJ "Using C-Alpha Geometry to Describe Protein Secondary Structure and Motifs" (2015) *Duke University*. ProQuest Dissertations Publishing.

Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, Verma V, Keedy DA, Hintze BJ, Chen VB, Jain S, Lewis SM, Arendall WB, Snoeyink J, Adams PD, Lovell SC, Richardson JS, Richardson DC (2018a). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*, *27*(1), 293-315.

Williams CJ, Videau LL, Hintze BJ, Richardson JS, Richardson DC (2018b) *Cis*-nonPro peptides: Genuine occurrences and their functional roles, *bioRxiv*, doi: 10.1101/324517

Word JM, Lovell SC, LaBean TH, Zalis ME, Presley BK, Richardson JS, Richardson DC (1999) "Visualizing and Quantitating Molecular Goodness-of-Fit: Small-probe Contact Dots with Explicit Hydrogen Atoms", *J Mol Biol* **285**: 1711-1733.

Yoon BH, Kim SK, Kim SY (2017) Use of Graph Database for the Integration of Heterogeneous Biological Data. *Genomics Inform*, 15(1) 19–27.