

The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues

Christopher J. Williams  | David C. Richardson  | Jane S. Richardson 

Department of Biochemistry, Duke University, Durham, North Carolina, USA

Correspondence

Jane S. Richardson, Department of Biochemistry, Duke University, Durham, NC 27710, USA.
Email: dcrjsr@kinemage.biochem.duke.edu

Funding information

National Institutes of Health, Grant/Award Numbers: P01-GM063210, R35-GM131883

Abstract

We have curated a high-quality, “best-parts” reference dataset of about 3 million protein residues in about 15,000 PDB-format coordinate files, each containing only residues with good electron density support for a physically acceptable model conformation. The resulting prefiltered data typically contain the entire core of each chain, in quite long continuous fragments. Each reference file is a single protein chain, and the total set of files were selected for low redundancy, high resolution, good MolProbity score, and other chain-level criteria. Then each residue was critically tested for adequate local map quality to firmly support its conformation, which must also be free of serious clashes or covalent-geometry outliers. The resulting Top2018 prefiltered datasets have been released on the Zenodo online web service and are freely available for all uses under a Creative Commons license. Currently, one dataset is residue filtered on main chain plus C β atoms, and a second dataset is full-residue filtered; each is available at four different sequence-identity levels. Here, we illustrate both statistics and examples that show the beneficial consequences of residue-level filtering. That process is necessary because even the best of structures contain a few highly disordered local regions with poor density and low-confidence conformations that should not be included in reference data. Therefore, the open distribution of these very large, prefiltered reference datasets constitutes a notable advance for structural bioinformatics and the fields that depend upon it.

KEYWORDS

protein library, reference data, structural bioinformatics, structure validation, Zenodo

1 | INTRODUCTION

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 > 40.¹ We made available the list for those 100 chains and for all our subsequent, increasingly large reference datasets (8,000 chains by 2013) but had to leave the application of B cutoffs to the user. Disappointingly, work that references our datasets essentially never applies residue-level filters, in spite of the fact that B-factors are in the coordinate file itself and many other useful residue-level properties have long been available

from the Electron Density Server.² Other groups who distribute reference datasets also provide only lists of PDBid and chain,^{3–6} although most of them use residue-level quality filtering in their own work. 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,⁷ but we still found no feasible mechanism for distributing all the coordinate files with residue-filter annotations.

Our current residue-level quality filtering process relies on extensive infrastructure, especially our developer team's integration into the Phenix software project.⁸ We also now manage the filtering information with a Neo4j graphical database.^{9,10} 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, and most important, is the Zenodo online service that hosts open access to very large, DOI-identified datasets.¹¹ We have now taken advantage of that venue to distribute our current best-parts 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 prefiltered versions suitable for general use with little or no further modification. One set is residue filtered on main chain criteria and the other on both main chain and side chain 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.

2 | METHODS

2.1 | Chain selection

We assembled a set of high-quality, low-redundancy protein chains.

Chains are 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 2.0 Å resolution
- Parent structure has deposited structure factors
- Parent structure deposited on or before December 31, 2018

These chains are analyzed with our validation statistics, and chains that fail the following criteria are removed from consideration:

- MolProbity score < 2.0 ¹²
- $< 3\%$ of residues have C β deviations¹³
- $< 2\%$ of residues have covalent bond length outliers $> 4\sigma$
- $< 2\%$ of residues have covalent bond angle outliers $> 4\sigma$

The remaining chains are treated within their PDB-defined sequence-identity clusters, which are calculated weekly with MMseqs2.¹⁴ From each cluster, we select 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 (and thus fewer one-per-cluster representatives), with greater differences between clusters. We recommend 70% for general use as a good balance point between coverage and nonredundancy.

To enable all-atom contact analysis¹ and produce the best quality chain possible, the parent structures have hydrogens added and N/Q/H flips performed by Reduce.¹⁵ This is done prior to splitting the structure into chains, so that hydrogen bonding networks and other contacts are kept intact within the asymmetric unit. Performing N/Q/H flips removes clashes, allowing more side chains to pass the subsequent residue-level filtering.

2.2 | Residue-level quality 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. Therefore, we apply a residue-level filtering process. Figure 1a shows 5Lp0¹⁶ a typical Top2018 chain. The main chain core is well

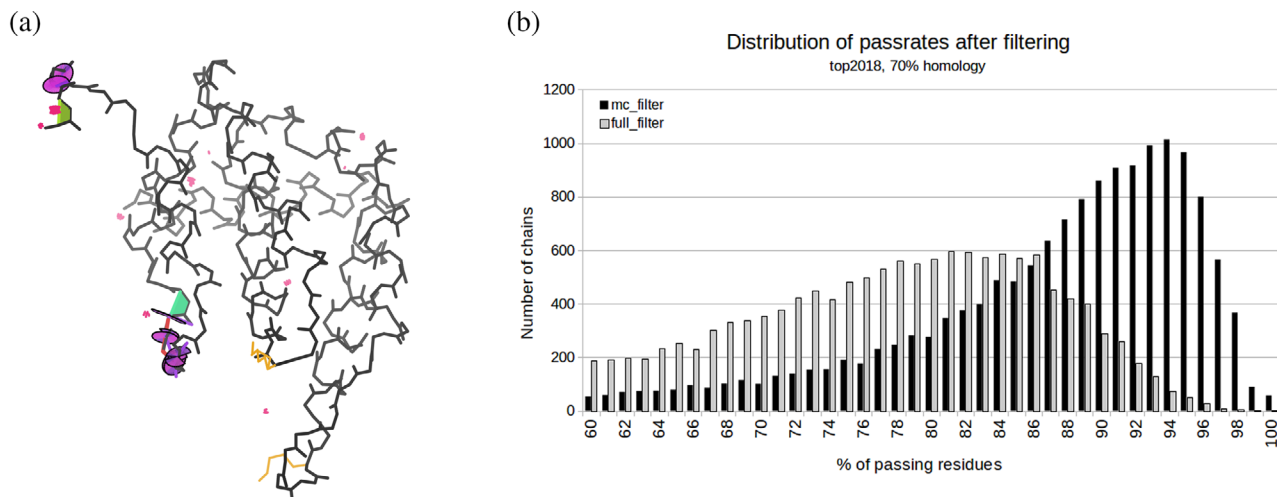


FIGURE 1 Distributions of model quality in high-resolution structure. (a) 5Lp0 demonstrates a typical distribution of structure quality for models included in the Top2018. Most of the model is reliable and free from outliers, but two short regions contain a concentration of significant errors, including the clashes (clusters of red spikes) and covalent-geometry outliers that trigger omissions, and the conformation outliers that also occur in these unreliable regions such as *cis*-peptides (green trapezoids) and CaBLAM outliers (magenta lines and purple wheels). (b) Distribution of % passing residues in Top2018 after main chain (solid bars) and full-residue filtering (open bars)

resolved and broken only by three individual clashes. The few serious errors are concentrated in just two short surface regions of poor density.

Two different residue-filtered sets were created, one filtered just on the main chain and one filtered on the full residue, including the side chain. Main chain filtering considers the atoms N, C α , C, O, and C β . C β is included with the main chain atoms since its ideal position is determined solely from other main chain atom positions. Full-residue filtering considers all main chain and side chain heavy atoms. Hydrogen atoms are used for all-atom contact analysis, but not for fit-to-map analyses, as their signal in the map is weak or absent. Nonstandard amino acids are considered in filtering if they are recognizably part of the protein chain.

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

- B-factor < 40
- Real-space correlation coefficient (RSCC) > 0.7
- 2mFo-DFc map value at atom position > 1.2 σ (This level is represented by gray contours in Figures 2 and 4–6.)
- No covalent geometry outliers > 4 σ involving those atoms
- No steric clashes (overlaps ≥ 0.4 Å) involving those atoms
- No alternate conformations for those atoms

Residues with atoms that fail any of these criteria are removed from the PDB files entirely.

The fit-to-map criteria (B-factor, RSCC, and map value) are obtained using `phenix.real_space_correlation_detail=atom`. This is done on the full parent structures before splitting into chains. Fit-to-map assessment could not be performed for 190 structures due either to crystal-symmetry MTRIX records that specify mathematically improper matrices or to other data issues. Chains from those structures were discarded. The B-factor, RSCC, and map cut-offs are those developed during production of a side chain rotamer library using our previous Top8000 dataset.⁷

Residues with alternate conformations are removed because the split in electron density that indicates an alternate necessarily lowers both those electron density peaks and the certainty of the model, especially when the alternates criss-cross each other, as frequently true. For applications that require the presence of alternates, we recommend that alternate conformations be validated and filtered as ensembles, taking care to spread alternate status along the backbone and to other interacting structure as needed to preserve acceptable geometry and avoid serious clashes.

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. Main-chain filtered chains typically retain a high level of completeness after filtering (Figure 1b), with almost half of the chains at least 90% complete. Full-residue-filtered chains are less complete, as expected, but over half of the chains are at least 75% complete after filtering.

At these high resolutions, the protein core is well resolved and the disordered, poor-density stretches that

fail the residue-level filtering are almost all at the surface and short. This means that the accepted continuous fragments between gaps are quite long and thus suitable as a basis of fragment libraries for model building, protein design or prediction, loop completion, and other uses. After filtering on main chain atoms, 70% of remaining residues are in continuous segments of at least 15 residues, and 47% are in very long segments of at least 30 residues. Full-residue filtering is more sensitive and the results are more fragmented, with 31% of residues in segments of at least 15, and only 10% in segments of at least 30.

Only protein residues were filtered (including modified amino acids with restraints available). Individual filtering of ligands, and especially ions and waters, requires development of additional filtering cutoffs and methods beyond 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.

This residue-filtering process sounds rather like the model validation that happens later, but the two procedures have distinct requirements. Residue-level filtering of reference data needs to be both stricter than, and independent of, the conformational model-validation criteria they will be used to define (such as rotamer, Ramachandran, or CaBLAM distributions). The overall chain-level filters include MolProbity score, which combines clashscore, Ramachandran, and rotamer quality measures. These overall, rather permissive, chain criteria lower the number of individual residue outliers but not the accepted residues or the future conformational distributions. Local map quality and fit are especially critical in reference data, because of the need to ensure that the experimental data are sufficient to specify the modeled conformation.

3 | RESULTS

3.1 | In-file documentation

The results of residue-level quality filtering are documented in each resulting PDB file by USER records appended to the end of the file. Our familiarity with USER records from documenting Reduce hydrogens and flips is part of why we use PDB format rather than the now-standard mmCIF format. These added records report: (a) the residues that were removed and the reasons for their removal (as a string of six single-letter codes), (b) the residues that remain and the lengths of the sequence fragments they form, and (c) the overall completeness statistics for the filtered file. These USER

records are self-documented, as in this small sample from 6git¹⁷:

```

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
USER  codes: bcmgoa
USER  DOC b=bfactor, c=real space correlation, m=2Fo-
USER  Fc mapvalue
USER  DOC g=geometry outlier, o=steric overlap,
USER  a=alternate conformations
USER  DOC Lines marked USER INC list the uninterrupted
USER  fragments of structure
USER  DOC still included after pruning by quality
USER  filtering
USER  DOC Format is chain1:resseq1:icode1:chain2:
USER  resseq2:icode2:fragment_length
USER  DOC where 1 is the first and 2 the last residue of
USER  the fragment
USER  DOC Line marked with USER PCT gives statistics for
USER  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
USER  residues:90.4 % pass

```

3.2 | Overall statistics and residue-level causes of filtering

Our previous dataset, the Top8000, contained 7,957 chains at the 70% sequence identity level and just over 1 million residues after full-residue filtering. Table 1 shows that the Top2018 contains about as many chains and more residues even at its most stringent level (30% sequence identity). At the equivalent 70%, there are now 1.5 times as many chains and 3 times as many residues.

TABLE 1 Chain and residue counts in the Top2018 datasets

	Main chain-filtered chains	Main chain-filtered residues	Full-residue-filtered chains	Full-residue-filtered residues
90% sequence identity	15,182	3,303,037	13,456	2,672,243
70% sequence identity	13,677	2,997,434	12,125	2,428,733
50% sequence identity	11,806	2,561,093	10,418	2,068,785
30% sequence identity	8,307	1,734,509	7,237	1,391,129

TABLE 2 Percent of filtered-out residues that fail each criterion. A residue can fail multiple criteria

	B factor (%)	RSCC (%)	Map value (%)	Geometry (%)	Clash (%)	Alternates (%)
Main chain filter	54	13	22	3	19	24
Full-residue filter	52	24	55	4	28	14

TABLE 3 Percent of pruned residues uniquely pruned by each criterion

	B factor (%)	RSCC (%)	Map value (%)	Geometry (%)	Clash (%)	Alternates (%)
Main chain filter	33	2	2	2	14	22
Full-residue filter	15	1	8	3	16	6

This relative increase in accepted residues per chain indicates an increase in structure size and in general structure quality¹⁸ since 2013.

Table 2 shows the total contributions of each residue-level filtering criterion to the overall filtering. At the selected cutoffs, B factor is the dominant criterion, followed by alternate conformations, 2mFo-DFc map value, and clashes. Map value is more important for full-residue filtering due to the relative abundance of poorly resolved sidechains. Covalent geometry contributes the fewest total filterings, since this geometry is often tightly restrained in refinement. Nevertheless, geometry outliers $>4\sigma$ usually diagnose severe modeling problems and are no less important for their rarity. See Supplement for filtering statistics by residue type.

Table 3 shows the unique contributions of each residue-level filtering criterion. B factor is again dominant at these cutoffs, with 33% of filtered-out residues being removed due to high B factor and no other causes. Alternate conformations and clashes are the other largest independent contributors. Residues that failed on RSCC coefficient or map value criteria usually also failed on at least one other measure.

The dominance of B-factor suggests an imbalance in our filtering criteria. However, this combination of density cutoffs was optimized to maximize passing residues with minimal false positives and negatives during previous work on side chains.⁷ Despite the high correlation

among B-factor, RSCC, and map-value, that work showed that using all three together was more effective than any one or two. The next evolution of our residue filtering will necessarily reevaluate this balance, especially for main chain only, but more importantly will treat different atom types separately.

3.3 | Limitations of residue filtering

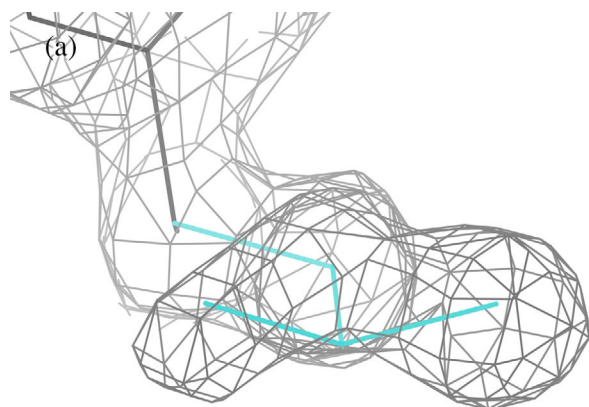
There is no substitute for visual inspection of structures. Our earliest dataset, the Top100, was hand curated. However, as the Protein Data Bank expanded, manual inspection of each structure became impractical. Automated quality assessments are required, but we still examine a sample of the results to ensure our automated code is doing what we intended.

Setting filtering cutoffs is a challenging balance. Overly aggressive selection against outliers on datasets that are then used to define outliers is circular reasoning. Preserving real irregularities, where they are justified, is also important to understanding the true variety of protein structure. Our goal is to set cutoffs where they permit most of the justified conformations and eliminate most of the unjustified ones, with minimal false identifications leaking through in either direction. To achieve this, cutoffs are necessarily set in the ambiguous region between clearly good and clearly bad structure. Our identification

of outliers for validation is set somewhat toward the forgiving side, to minimize “false alarms” for the structural biologist seeking to correct problems. The residue-level filtering cutoffs in these reference data are more toward the strict side to minimize the inclusion of incorrect local conformations. There is no magical, perfect cutoff within the ambiguous region.

An important commonality between reference-data filtering and later validation is that there are always at least a few genuine examples well supported by their experimental data but outside the prior-probability expectations from previous structures. Those examples can be correct, appropriate for reference data, but are still outliers (e.g., nonrotameric side chains). As well as having support in the map density, strained conformations need to be held in place. A couple of H-bonds can constrain one eclipsed dihedral for a polar side chain, and tight packing interactions can constrain an otherwise-unfavorable conformation of a nonpolar side chain, especially an aromatic ring.¹⁹

Our residue-level filtering relies heavily on fit-to-map criteria. This is appropriate for keeping our filtering independent of the conformation-based validations, we use these datasets to construct. However, as a result, some types of systematic errors that are not strongly marked by poor map fit are not fully excluded from this dataset, regardless of filtering.



Leucine has a systematic-error conformation (Figure 2a) that can evade these filters often enough to produce a datapoint cluster (Figure 2b, pink dots); that cluster is an “imposter” rotamer that should not be included in a rotamer library. In this modeling error, χ_2 is rotated by 140° – 150° and χ_1 by 30° – 40° from the very common **tp** (trans, plus) conformation, which keeps the C δ atoms in about the same peaks.¹⁹ This is clearly an attractive but incorrect fit to the density, and both branches are almost exactly eclipsed. But the side chain atoms sometimes fit within the density envelope just well enough to pass our filters. Fit-to-map filters aggressive enough to catch and remove all such tetrahedral-branch “imposters” would also remove real but marginal conformations that we wish to preserve.

3.4 | Documenting the need for and benefits of residue-level filtering

The key fact that motivates 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 side chains on a protein's surface compared to well-packed side chains in a protein's core, and unseen backbone at chain

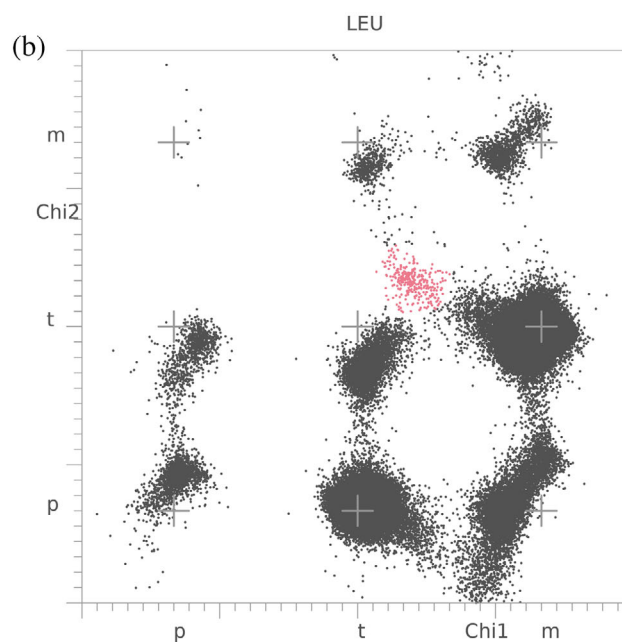


FIGURE 2 A systematic-error Leucine conformation. (a) Leucine has a common “imposter” conformation rotated 30° – 40° in χ_1 and 140° – 150° in χ_2 from the very common **tp** conformation, which keeps C δ atoms in about the same peaks. Despite the sidechain geometry not fitting the shape of the density, the atoms are often within the gray 1.2σ contour well enough to pass our filters. This example is 1ekq, leu B 62. The C γ has B 20.9, RSCC 0.8, and map value 1.75, despite the misfitting, and the C δ atoms are even better. (b) Even after filtering, a cluster of this outlier is clearly visible in pink to the upper right of the **tt** rotamer. The datapoints near χ_2 zero in the **mp** bin are a similar systematic error fit backwards from the very common **mt** rotamer

termini or in disordered loops. Chain-level filtering is an important first step, but it does not provide protection against residue-level modeling errors in those disordered regions of poor density.

The general-case Ramachandran distribution for residues in the Top2018 70% sequence identity dataset before residue-level filtering (Figure 3a) shows blurred edges and many outliers in the excluded regions. Filtering on main chain atoms (Figure 3b) renders the edges of the distribution cleaner and more interpretable and removes many outliers. The difference in the distribution without filtering is sufficient to change the 0.2% contour boundary between allowed and outlier conformations (Figure 3c), by a greater amount than does the omission of all residues in secondary structures.

Sidechain rotamer distributions are even more affected. For example, in isoleucine, significant populations of **pm** and **tm** highly strained conformations (which clash with their own backbone) are present in the unfiltered set (Figure 3d). Full-residue filtering removes

only 15–25% of the five most common Ile rotamers, while it removes 70–80% of these two disfavored conformations (Figure 3e), placing their remaining populations below the statistical threshold for being identified as true rotamers (Figure 3f).

We examined all 13 examples of Ile **tm** that passed the filters, finding 12 of them clearly supported by map density and constrained into this unfavorable conformation by tight packing that cannot accommodate any other rotamer. In one case (5nz4²⁰ Ile B 216) the packing allows a nonclashing **tp** rotamer, but it is not represented in the weak density even at a very low contour level. This check confirms that even for extremely rare conformations, examples in the filtered data are likely to be correct. Figure 4a shows one of these genuine conformations with completely unambiguous density peaks for each atom and very good all-atom contacts (Ile 103 in 4ezi²¹). The local necessity for this rare and strained conformation is explained by Figure 4b, which shows in orange the geometrically “ideal” **tm** conformation at -180° , -60° that

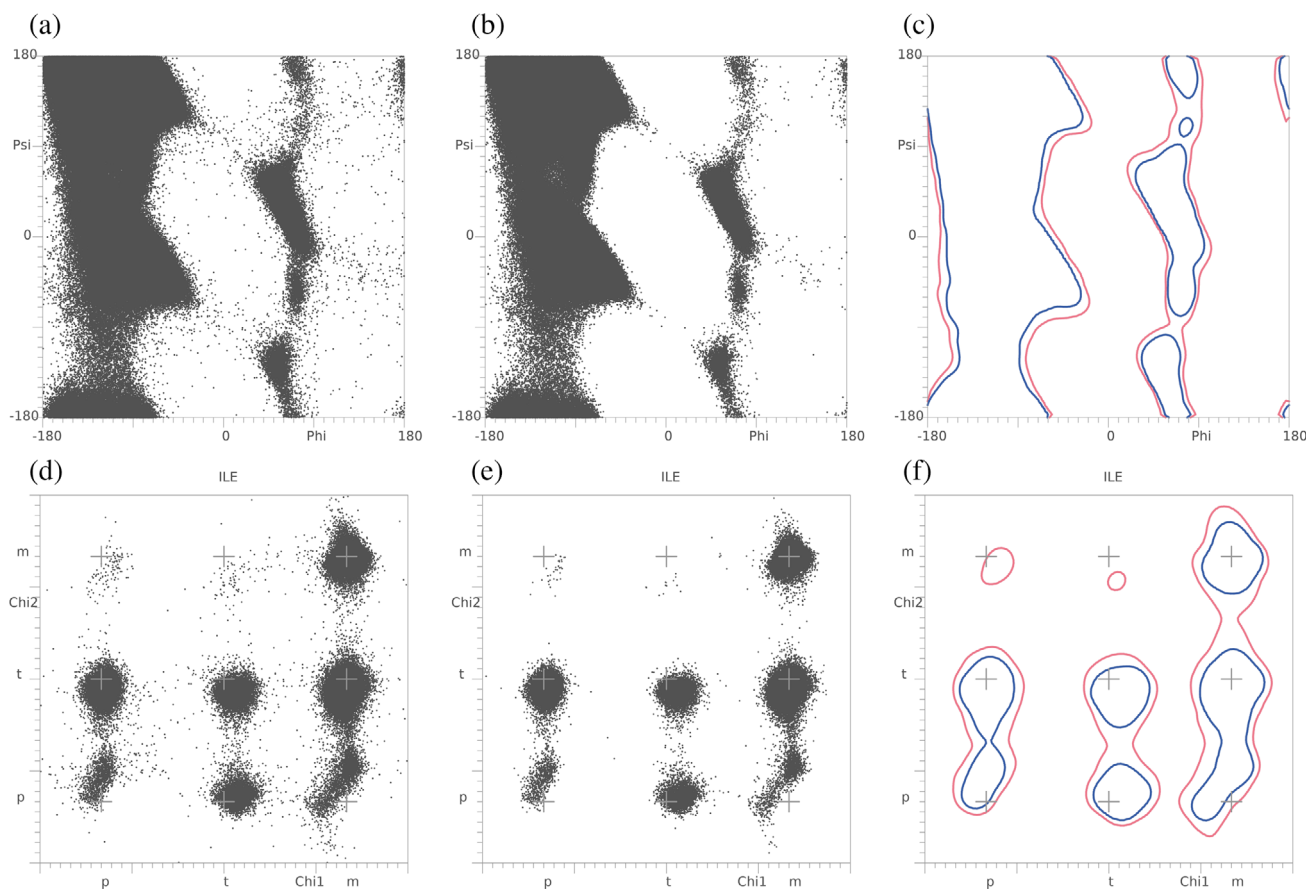


FIGURE 3 Ramachandran and rotamer distributions before and after residue filtering. (a) Ramachandran distribution for general-case residues in Top2018 70% sequence identity set before residue-level filtering. (b) Main chain filtering cleans up distribution edges and removes many outliers. (c) 0.2% allowed/outlier contour boundary before (red) and after (blue) filtering. (d) χ_1/χ_2 rotamer distribution for isoleucine before filtering. (e) Full-residue filtering cleans up distribution edges and prevents very rare, highly strained conformations from being defined as rotamers. (f) 0.3% allowed/outlier contour boundary before (red) and after (blue) filtering

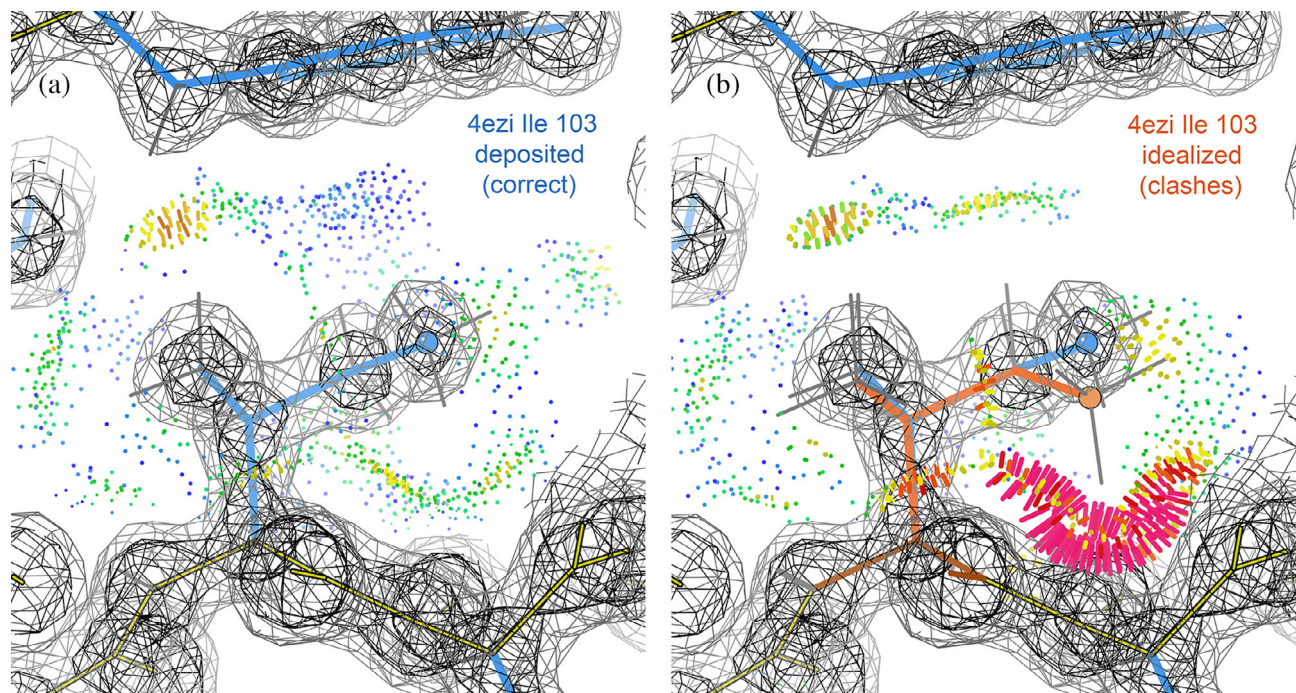


FIGURE 4 The rationale for a genuine but highly strained rare conformation. (a) The unambiguously correct **tm** model for 4ezi Ile 103 at 1.15 Å, with a separate, strong density peak at every atom and good van der Waals contacts all around the sidechain (green and blue dots) and one small overlap (yellow and orange). The Ile C δ atom is highlighted with a blue ball. (b) For comparison, the “ideal” **tm** conformation at $180^\circ \chi_1$, $-60^\circ \chi_2$ is overlaid (in orange, with orange ball on C δ), which has an extremely large clash (hotpink spikes). Gray contour represents 1.2 σ map level used by filters and black contour represents 3 σ

would have a very bad clash of C δ and its hydrogens with the following peptide. To avoid that clash, χ_2 has to change by 28° and each of the three covalent angles from the backbone out to C δ is opened up by about 2° , a large total amount of strain. No more favorable conformation is possible for this side chain because it is closely surrounded by other structure, especially the tyrosine packed tightly above it.

As another telling example of misfitting a rare, strained conformation, the structural biology community may remember the crisis of gross *cis*-non-proline overuse from about 2006 until after the problem was recognized.^{22,23} This phenomenon was most extreme at lower resolutions (often $>1\%$ of total residues, or 30 times too many), but it is also present in poorly resolved regions of high-resolution structures such as those in this dataset. The overuse was largely cured just by obvious flagging of *cis*-nonPro as probable outliers in graphics or validation reports.²⁴ Residue-level filtering of our reference data has always guarded against the inclusion of unsupported *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 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 analyses of valid *cis*-nonPro occurrence.²⁴

More importantly than these general statistics, residue-level filtering removes nearly all of the 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 weak, patchy, or truncated density at chain termini or in partly disordered loops (Figure 5a, 4rm4²⁵ 170–172). The lack of strong electron density in such regions *allows* this and other modeling errors to occur. Thus, it is vital to the health of a statistical reference dataset or fragment library to remove these regions of low certainty, as we do in this dataset. In contrast, real *cis*-nonPro are supported by clear electron density (Figure 5b, 6bt²⁶).

As a specific case, we considered *cis*-nonPro peptides at the very vulnerable position of chain ends (both the first and last residues in the chain, and the residues at the ends of unmodeled loops). The unfiltered Top2018 data contain 200 chain-terminal instances of *cis*-nonPro. Most of these 200 cases occur at the N- rather than C-terminus of a modeled stretch, where the lack of preceding structure makes the choice vulnerable to a misfit *cis* rather than *trans* peptide at the branch point between the carbonyl O and the N-terminal C α (Figure 6a, 5Lp0). A total of 197 (98.5%) of them fail the filters and are

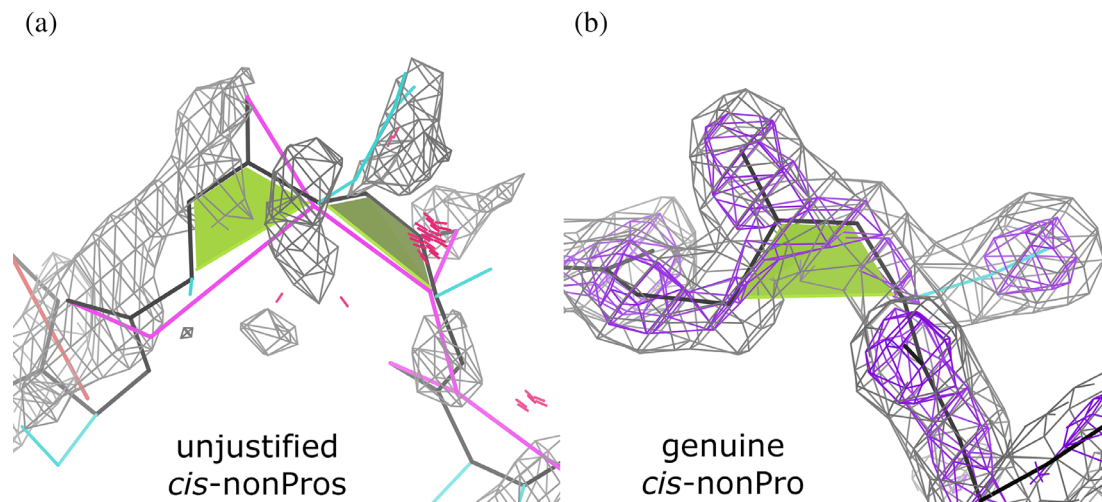


FIGURE 5 Non-proline *cis*-peptides (flagged by green trapezoids). (a) A double *cis*-nonPro modeled into a loop region of poor density in 4rm4 residues 170–172. Real *cis*-nonPro never occur sequentially, and this is not a reasonable interpretation of the map. These residues are removed in filtering. (b) A genuine *cis*-serine, residue 275 of 6btf. It passes our quality criteria and is included in the structure after filtering. The 1.6 Å density is persuasive and well fit by the model. Gray contour represents 1.2 σ map level used by filters and purple contour represents 3 σ

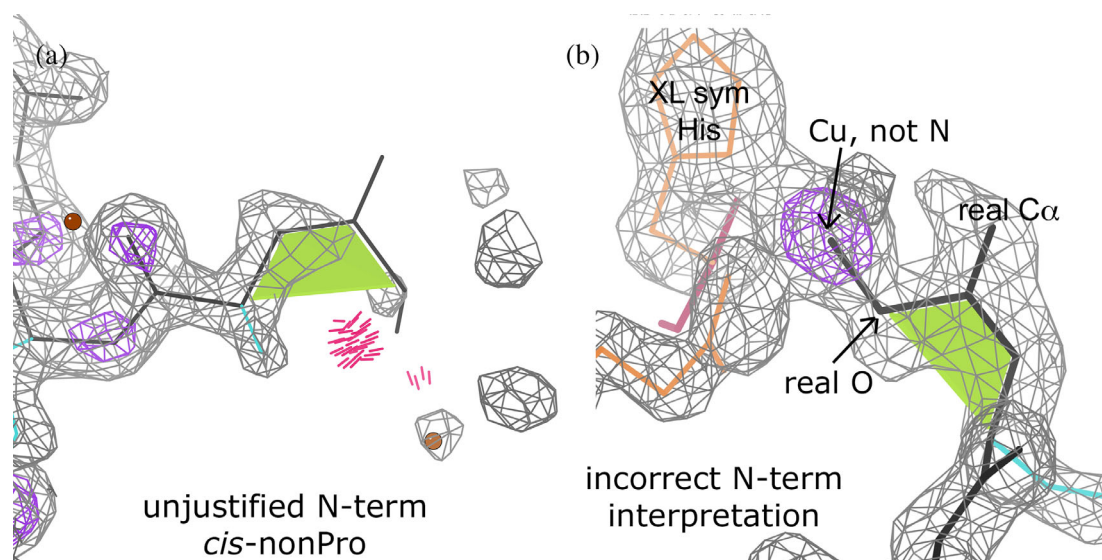


FIGURE 6 Non-proline cispeptides at chain termini. (a) A typical misfit *cis*-nonPro at a chain N-terminus in 5Lp0. The truncated density encourages but does not justify fitting a *cis* conformation here. This residue is removed from the dataset for poor fit to map and a clash. (b) One of three incorrect N-terminal *cis*-nonPro that passed the filters. This is a case of the systematic-error switch between the carbonyl O and the C α around the branch at the carbonyl C, but it is very unusual in having strong density for the actual CO that is a ligand to an adventitious, crystal-contact Cu site in this copper-export protein (4f2f). Gray contour represents 1.2 σ map level used by filters and purple contour represents 3 σ

removed. Of the 3 (1.5%) that passed the filters, all three are made difficult to avoid or validate by an N-terminal Gly that provides no side chain to help choose the correct alternative, and two are in rather weak, ambiguous density that only barely passes the filters. The third example is actually an adventitious crystal-contact Cu site (4f2f²⁷ Gly-Ser 1) in an expression tag, where the Cu is incorrectly modeled as the N-terminal N (Figure 6b). The density

peak is >16 σ and the B-factor of the modeled N is just 5 compared to B-factors of 20–30 for the other atoms of Gly 0, both indications that a much heavier atom should have been modeled there. This residue clashes across the crystal contact and would have been removed from the dataset if symmetry relationships were considered. It also motivates us in future to include maximum and minimum map-value cutoffs that are specific to the atom type.

Residue-level filtering thus ensures that the population of *cis*-nonPro peptides is not statistically or locally overrepresented due to modeling errors. The *cis*-nonPro that remain in the dataset (Figure 5b) do so based on a reasonable standard of map and model quality and provide genuine *cis* conformations and contexts that can be used in model building when the evidence in the experimental data is good enough to outweigh the very low prior probability of 1:3,500 (8 log likelihood units). If alternative *trans* and *cis* conformations are explicitly compared, the choice is usually clear at better than about 2.3 Å, and *trans* conformations should not be entirely disallowed, which has occasionally been done.²⁸ However, at resolution ≥ 3 Å or when the density is otherwise ambiguous, the much less probable *cis* alternative should never be chosen, unless it occurs in a related protein at higher resolution.

4 | CONCLUSIONS

Residue-level filtering is necessary even in otherwise excellent structures, in order to remove the significant populations of disordered residues, whose conformations have little or no experimental support. Conclusions based on unfiltered residues will be inaccurate representations of true protein behavior in both their statistics and their details. The easy availability of this prefiltered reference dataset on Zenodo is therefore an important step forward for structural bioinformatics.

The full-coordinate, residue-filtered reference datasets described here omit all residues that fail the quality filters, so that they contain only coordinates for amino acid residues that are almost certainly correct. These gapped, residue-filtered datasets are suitable for most uses, though not for applications that require the full, ungapped context, such as Voronoi analyses or molecular dynamics simulations. We, and other expert users, have used our own residue-filtered datasets for many purposes including protein side chain rotamer libraries,⁷ Ramachandran distributions,^{29,30} structural motifs that span multiple residues and involve backbone–side chain interactions,^{23,29} rare backbone conformations such as *cis*-nonPro, higher-dimensional conformation distributions²⁹ or other complex features,^{31,32} and to prepare curated fragment libraries for model building or for protein design.^{18,23,33} Now users not expert in this subfield have easy access to residue-filtered datasets.

Another important advantage of these new datasets is that they are the epitome of reusability and reproducibility, factors of increasing emphasis.

Future directions for this work will first involve releasing a version with the files in mmCIF format,

probably available by the time of publication. Developing advanced map-quality criteria to consider atom type and use both upper and lower cutoffs will increase the selectivity of our filters and will enable filtering of ligands and ions. We would appreciate and respond to user feedback for what expansions could enable other unanticipated uses. In the longer term, we plan to produce future editions of these datasets that incorporate new data, probably including high-resolution cryoEM structures, and to compile a similarly filtered reference dataset for RNA structures.

ACKNOWLEDGMENTS

This work was funded by NIH grants R35-GM131883 to David C. Richardson and P01-GM063210 Project IV to Jane S. Richardson.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Christopher J. Williams: conceptualization (equal); dataCuration (lead); formalAnalysis (equal); methodology (equal); software (lead); validation (equal); visualization (equal); writing, original draft (lead); writing, review & editing (equal). **Jane S. Richardson:** conceptualization (equal); dataCuration (supporting); formalAnalysis (equal); funding Acquisition (supporting); methodology (equal); resources (lead); supervision (lead); validation (equal); visualization (equal); writing, original draft (supporting); writing, review & editing (equal). **David C. Richardson:** conceptualization (supporting); formalAnalysis (supporting); funding Acquisition (lead); methodology (supporting); resources (supporting); supervision (supporting); validation (supporting); writing, review & editing (supporting).


DATA AVAILABILITY STATEMENT

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. If the dataset is updated for enhanced usability based on community feedback, these links will resolve to the latest version of each dataset. Distinct datasets such as files in mmCIF format or new editions with later primary data will be separate entries.

ORCID

Christopher J. Williams  <https://orcid.org/0000-0002-5808-8768>

David C. Richardson  <https://orcid.org/0000-0001-5069-343X>

Jane S. Richardson  <https://orcid.org/0000-0002-3311-2944>

REFERENCES

- Word JM, Lovell SC, Labean TH, et al. Visualizing and quantifying molecular goodness-of-fit: Small-probe contact dots with explicit hydrogen atoms. *J Mol Biol.* 1999;285:1711–1733.
- Kleywegt GJ, Harris MR, Zou J, Taylor TC, Wählby A, Jones TA. The Uppsala Electron-Density Server. *Acta Cryst.* 2004;D60:2240–2249.
- Hooft RWW, Sander C, Vriend G. Verification of protein structures: Side-chain planarity. *J Appl Cryst.* 1996;29:714–716.
- Noguchi T, Matsuda H, Akiyama Y. PDB-REPRDB: A database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res.* 2001;29:219–220.
- Wang G, Dunbrack RL. PISCES: a protein sequence culling server. *Bioinformatics.* 2003;19:1589–1591.
- Griep S, Hobohm U. PDBselect 1992–2009 and PDBfilter-select. *Nucleic Acids Res.* 2009;38:D318–D319.
- Hintze BJ, Lewis SM, Richardson JS, Richardson DC. MolProbity's ultimate rotamer-library distributions for model validation. *Proteins.* 2016;84:1177–1189.
- Liebschner D, Afonine PV, Baker ML, et al. Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallogr.* 2019;D75: 861–877.
- Yoon B-H, Kim S-K, Kim S-Y. Use of graph database for the integration of heterogeneous biological data. *Genomics Inform.* 2017;15:19–27.
- Webber J, Van Bruggan R. Graph databases for dummies, Neo4j special edition. Hoboken, NJ: John Wiley & Sons; 2020.
- Sicilia MA, García-Barricóanal E, Sánchez-Alonso S. Community curation in open dataset repositories: Insights from Zenodo. *Procedia Comput Sci.* 2017;106:54–60.
- Chen VB, Arendall WB, Headd JJ, et al. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr.* 2010;D66:12–21.
- Lovell SC, Davis IW, Arendall WB III, et al. Structure validation by $C\alpha$ geometry: φ , ψ and $C\beta$ deviation. *Proteins.* 2003;50: 437–450.
- Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun.* 2018;9:1–8.
- Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol.* 1999;285: 1735–1747.
- Levin-Kravets O, Tanner N, Shohat N, et al. A bacterial genetic selection system for ubiquitylation cascade discovery. *Nat Methods.* 2016;13:945–952.
- Faba-Rodríguez R, Dionisio G, Brinch-Pedersen H, Brearley CA, Hemmings A. Crystal structure of a cereal purple acid phytase provides insights to phytate degradation in plants. 2018; unpublished.
- Williams CJ, Headd JJ, Moriarty NW, et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* 2018;27:293–315.
- Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. *Proteins.* 2000;40:389–408.
- Pokorná J, Páchl P, Karlukova E, et al. Kinetic, thermodynamic, and structural analysis of drug resistance mutations in neuraminidase from the 2009 pandemic influenza virus. *Viruses.* 2018;10:339.
- The Joint Center for Structural Genomics Crystal structure of a hypothetical protein (lpg1103) from *Legionella pneumophila* subsp. *pneumophila* str. Philadelphia 1 at 1.15 Å resolution. 2012; unpublished.
- Croll TI. The rate of cis-trans conformation errors is increasing in low-resolution crystal structures. *Acta Crystallogr.* 2015;D71: 706–709.
- Williams CJ Using C-alpha geometry to describe protein secondary structure and motifs. 2015. PhD thesis, Duke University.
- Williams CJ, Videau LL, Richardson DC, Richardson JS. CisnonPro Peptides: Genuine occurrences and their functional roles. *bioRxiv.* 2018. <https://doi.org/10.1101/324517>
- Zhang A, Zhang T, Hall EA, et al. The crystal structure of the versatile cytochrome P450 enzyme CYP109B1 from *Bacillus subtilis*. *Mol Biosyst.* 2015;11:869–881.
- Liptak C, Mahmoud MM, Eckenroth BE, et al. I260Q DNA polymerase β highlights precatalytic conformational rearrangements critical for fidelity. *Nucleic Acids Res.* 2018;46: 10740–10756.
- Fu Y, Tsui HCT, Bruce KE, et al. A new structural paradigm in copper resistance in *Streptococcus pneumoniae*. *Nat Chem Biol.* 2013;9:177–183.
- Croll TI, Williams CJ, Chen VB, Richardson DC, Richardson JS. Improving SARS-CoV-2 structures: Peer review by early coordinate release. *Biophys J.* 2021;120:1085–1096.
- Richardson JS, Keedy DA, Richardson DC. “The plot” thickens: More data, more dimensions, more uses. Bansal M, Srinivasan N, editors. *Biomolecular forms and functions: A celebration of 50years of the Ramachandran Map.* Singapore: World Scientific Publishing, 2012; p. 46–61.
- Sobolev OV, Afonine PV, Moriarty NW, et al. A global Ramachandran score identifies protein structures with unlikely stereochemistry. *Structure.* 2020;28:1249–1258.
- Weichenberger CX, Pozharski E, Rupp B. Visualizing ligand molecules in twilight electron density. *Acta Crystallogr.* 2013; F69:195–200.
- van Beusekom B, Joosten K, Hekkelman ML, Joosten RP, Perrakis A. Homology-based loop modeling yields more complete crystallographic protein structures. *IUCrJ.* 2018;5: 585–594.
- Leaver-Fay A, O'Meara MJ, Tyka M, et al. Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol.* 2013;523:109–143.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Williams CJ, Richardson DC, Richardson JS. The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues. *Protein Science.* 2022;31:290–300. <https://doi.org/10.1002/pro.4239>