Published in final edited form as:
FEBS J. 2014 September ; 281(18): 4061–4071. doi:10.1111/febs.12860.

Conformation-dependent backbone geometry restraints set a new standard for protein crystallographic refinement

Nigel W. Moriarty[1], **Dale E. Tronrud**[2], **Paul D. Adams**[1,3], and **P. Andrew Karplus**[2,*]

[1]Lawrence Berkeley National Laboratory, One Cyclotron Road, MS64R0121, Berkeley, CA 94720 USA

[2]Department of Biochemistry and Biophysics, Oregon State University, 2011 Agriculture and Life Sciences Building, Corvallis, OR 97331 USA

[3]Department of Bioengineering, University of California Berkeley, Berkeley, CA, 94720 USA

Abstract

Ideal values of bond angles and lengths used as external restraints are crucial for the successful refinement of protein crystal structures at all but the highest of resolutions. The restraints in common usage today have been designed based on the assumption that each type of bond or angle has a single ideal value independent of context. However, recent work has shown that the ideal values are, in fact, sensitive to local conformation, and as a first step toward using such information to build more accurate models, ultra-high resolution protein crystal structures have been used to derive a conformation-dependent library (CDL) of restraints for the protein backbone (Berkholz *et al.* 2009. *Structure.* **17**, 1316). Here, we report the introduction of this CDL into the Phenix package and the results of test refinements of thousands of structures across a wide range of resolutions. These tests show that use of the conformation dependent library yields models that have substantially better agreement with ideal main-chain bond angles and lengths and, on average, a slightly enhanced fit to the X-ray data. No disadvantages of using the backbone CDL are apparent. In Phenix usage of the CDL can be selected by simply specifying the cdl=True option. This successful implementation paves the way for further aspects of the context-dependence of ideal geometry to be characterized and applied to improve experimental and predictive modelling accuracy.

Keywords

Protein structure; crystallographic refinement; geometry restraints; ideal geometry; structural genomics

[*]Corresponding author: P. Andrew Karplus, Department of Biochemistry & Biophysics, 2011 Ag & Life Sciences Bldg, Oregon State University, Corvallis, OR 97331, Ph. 541-737-3200, fax. 541-737-0481, karplusp@science.oregonstate.edu.

Author Contribution Statement

NWM, DET, PDA, and PAK planned experiments and analysed data; NWM performed experiments; and NWM, DET, and PAK wrote the paper with input from PDA.

Introduction

As the first protein crystal structures were solved, it was not clear if protein structures could be crystallographically refined the same way as small molecule structures. This concern was laid to rest by the successful refinement of the small protein rubredoxin at 1.5 Å resolution [1], although the geometric quality of the resulting model was recognized to be rather poor. A subsequent innovation overcome this problem was to alternate crystallographic refinement calculations with calculations that constrained or restrained the covalent geometry to take on reasonable values (e.g. [2–4]). Shortly thereafter, an approach was developed for simultaneously refining a model to agree with both diffraction data and covalent geometry restraints [5]. In this strategy, covalent geometry features were 'restrained' concurrently with the fit to the diffraction data by the inclusion of penalties for deviations from target values that were provided in the form of ideal bond lengths and angles. The standard deviations associated with the target values could then be used for weighting of the least-squares residuals (e.g. [6]). Many improvements in algorithms have ensued in the subsequent decades, but this concept of "restrained reciprocal space refinement" is still the basis of current refinement programs.

The long-standing paradigm guiding the design of the restraint libraries has been that each bond length or angle has a single ideal value that represents its minimum energy value independent of context (e.g. [7,8]). The ideal target values present in such single-value libraries, here called SVLs, have typically been derived from highly accurate small molecule (amino acid and oligopeptide) crystal structures (e.g. [9,10]). These target values have evolved slightly over the years as more and more such reference structures have been determined and allowed the average values to be more accurately known [11–13]. The current widely used standard is the Engh & Huber 'CSDX' library published in the International Tables for Crystallography [13]. Although it had long been known that "small variations" around the ideal values should be expected to occur "due to secondary structural and tertiary structural features" [3] it was not until much later that a tangible shortcoming of these libraries became apparent when it was discovered that the most accurately determined ultra-high resolution (≤ 1.2 Å resolution) protein crystal structures scattered widely about the ideal values (e.g. [14,15]), leading to debate about how to set optimal target values and weights [16–19].

Karplus *et al.* [20] suggested that the origin of this problem was not the target values themselves, but rather the inadequacy of the SVL paradigm (i.e. the view that there exists a single ideal value), which is consistent neither with existing quantum mechanical calculations [21,22] nor with empirical analyses of high and ultra-high resolution protein structures [23,24]. The special value of ultra-high resolution protein crystal structures is that the available X-ray diffraction data are of sufficiently high information content that the geometry restraints play such a small role in defining the final atomic positions that the true geometric properties of the protein molecules can be discovered. Recognizing this, Berkholz *et al.* [25] used the Protein Geometry Database [26] to gather the relevant information from protein crystal structures determined at 1 Å resolution or better, and developed the first empirical conformation-dependent library (CDL) for the protein backbone.

This CDL defined the relevant bond length and angle target values and their standard deviations as functions of the ϕ,ψ angles and residue types (Figure 1). In well-populated ϕ,ψ regions, the expected values for a backbone bond angle could vary by up to 7° (see figure 3 in Berkholz *et al.* [25]), and as is visible in the left hand panel of Figure 1A, the central N-Cα-C angle tends to lower than average values in the β-strand regions, near average values in the α-helix and polyproline II regions, and above average values in the bridge regions (ψ-values near 0°). Berkholz *et al.* [25] further showed that the variations in bond angles seen near the edges of allowed regions ϕ,ψ space made sense in terms of being variations that relieved close contacts between atoms (see Figure 6 of that paper). More recently, Esposito *et al.* [27] showed that similar ϕ,ψ-dependent variations in covalent geometry occur in short peptides and non-globular proteins, proving that the conformation-dependence seen is a fundamental property of peptide units rather than being a special property of folded proteins or their secondary structures. Also, it has been recently shown that in predictive protein modelling using Rosetta, allowing backbone bond geometry to relax from fixed ideal values improves convergence properties [28].

For testing to what extent using this CDL could improve crystallographic protein structure refinements, we took advantage of the design of the TNT [11] and SHELXL [29] refinement programs that made it possible to use the CDL without actually altering the programs. Tests using these programs and a handful of examples showed that use of the CDL improved refinement behavior at all resolutions in the tested range of 2.5 to 0.7 Å [30,31].

These proof-of-principle successes led us to incorporate the CDL into the Phenix refinement program *phenix.refine* [32–34]. The incorporation into Phenix both allows us to assess the impact of using the backbone CDL in the context of modern maximum likelihood refinement strategies, and in the context of the refinements of large numbers of diverse structures. Also, this implementation makes the CDL approach readily available so it can be broadly used in future protein structure refinements. We report here that tests involving the re-refinement of ~25,000 structures in the Protein Data Bank (PDB) [35] show that the use of the backbone CDL has no apparent drawbacks and leads to substantial reductions of geometric residuals at all resolutions. Based on the success of these tests, beginning with the Phenix 1.8 release (June 2012), the backbone CDL has been included as an option.

Results and Discussion

Implementation concept

The internal workings of programs in Phenix that process geometry restraints are especially well-suited for adaptation to the CDL, because each program accesses a single central object that contains all of the restraint information in terms of a list of atoms that are involved in each restraint along with the ideal values and their estimated standard deviations (ESDs). The new CDL module operates on this central object at the beginning of each refinement macro-cycle keeping the restraints in step with any changes in the ϕ,ψ-angles. Also, atoms in alternate conformations are treated independently and so are able to have their restraint targets appropriately customized if the ϕ,ψ-angles make that appropriate. The centralisation of the CDL module allows use of the CDL in other Phenix programs such as geometry minimization and real space refinement [36]. Selecting the backbone CDL for use in

refinement is done using the cdl=True option or by choosing the equivalent option in the graphical user interface [37].

Test cases

Our first approach to test the correctness of the Phenix implementation involved refining a subset of the test cases that had been used in the previous TNT and SHELXL implementations [30,31]. The four test cases were structures determined at 1.7, 1.3, 0.95 and 0.65 Å resolution, and each of these behaved comparably to the previous studies with the CDL refinements leading to much smaller backbone bond angle and bond length residuals with little change in the R-factors (Table 1). Important to note is that the overall rms bond length and angle deviations from ideality that would normally be reported in a paper are improved to a lesser degree, because they are an average of the backbone values that improve substantially and the values for the side chain restraints that stay the same (Table 1). These tests further showed that just as we have described in our previous work using SHELXL [31], the refinements performed in Phenix had difficulty appropriately restraining the geometry of alternate conformations (data not shown). For this reason, in the subsequent studies here, the reported statistics are derived only from the fully occupied residues in the structures.

Assessing the backbone CDL by re-refining Protein Data Bank entries

Having shown that the Phenix implementation of the backbone CDL behaved as expected in test cases, we were in a position to use *phenix.refine* to carry out a much broader comparison of the behavior of refinements using this CDL as opposed to the conventional SVL values. In an initial survey, we re-refined a group of 25976 protein structures from PDB that were solved at 3.55 Å or better and had diffraction data deposited that were not twinned, were ≥90% complete, and could be successfully converted to an MTZ file format using *phenix.cif_as_mtz*. Also, they had to have starting calculated R_{work} and R_{free} values that were less than 30% and 35%, respectively, and an R_{free}-R_{work} difference of ≥1.5%, with the latter criterion serving to filter out structures that may not have a correctly labelled R_{free} test set. Additional high resolution structures that were solved at better than 1.05 Å and had R_{free}-R_{work} difference of ≥0.5% were included.

In evaluating the performance of the SVL and CDL restraint libraries, we assessed the geometric residuals of models at three levels (Figure 2): first, we compared the residuals of the "as deposited" PDB entries against the SVL and CDL libraries (dashed lines); second, we refined the entries using either the SVL or CDL, and compared the residuals of the resulting models against the library with which refinement was performed (solid lines); and third, as a positive control, we regularized the entries using either the SVL or CDL (i.e. geometric minimization without reference to the X-ray diffraction data), and compared the residuals of the resulting models against the library with which refinement was performed (thick lines). The regularization shows that the minimizer routines function well with both libraries, as the residuals for bond lengths and angles can be reduced to the very low values of ~0.001 Å and ~0.3°, respectively (Figure 2A–C).

A first point we note is that even though all of the deposited structures in this study had been generated by refinement against an SVL library, those structures refined at ~2 Å resolution or better have backbone bonds that agree more closely with the CDL targets (Figure 2B blue *vs.* red dashed). Looking at the N-Cα-C bond angle alone, which is the backbone angle with the most conformation dependent variation [25], this behavior is even more pronounced with the crossover point happening at ~3 Å resolution and the agreement with the CDL becoming as much as 40% better for the highest resolution structures (Figure 2C). These observations are consistent with our expectations that the CDL targets better capture the properties of the true bond angles that are increasingly achieved in the more accurate higher resolution models. The robust ability of even medium resolution data to guide the N-Cα-C angle toward its true value despite SVL restraints can be understood in the that the N-Cα-C angle does not just define the relative positions of three local atoms, but defines a larger feature which is the relative orientations of two peptide planes (see Fig 1B,C).

A second observation we make is that automated refinements with Phenix even using the conventional SVL geometry leads to a substantial lowering of the geometry residuals (Figure 2A blue dashed *vs*. solid curves) compared with those that were present in the deposited structures. As most of the deposited structures would have originally been refined against these same restraints, we take the changes as being indicative of unique features of the phenix refinement algorithms and weighting strategy. This underscores that for the sake of consistency, it is crucial in assessing the impact of the CDL that the same refinement algorithm and protocol be used for both libraries.

Comparing results from Phenix refinements using the CDL *versus* the SVL, the first observation of note is that for bond lengths, the CDL does lead to 20–30% improvement in the residuals (Figure 2A). This is consistent with what was seen for the previous refinement tests performed in the programs TNT and SHELXL [30,31], in which the improved residuals were attributed to the smaller standard deviations (i.e. higher weights) associated with the CDL, rather than a necessarily signifying an improvement in the target values themselves. As was pointed out by Berkholz *et al.* [25], since the structures refined at near 1 Å resolution that were used in developing the CDL have a coordinate uncertainty on par with the variations in bond lengths (~0.01–0.02 Å), we do not consider the bond length targets as highly reliably defined. Nevertheless, this behavior of the CDL for bond lengths proves that the protein-derived conformation-dependent bond length targets are certainly not worse than the existing SVL targets, and may in fact be better.

Further, comparing the CDL and SVL performance with regard to backbone bond angle residuals, it can be seen that at all resolutions the CDL yields geometry deviations from ideality that are improved (i.e. lowered) by about 33% (Figure 2B red *vs.* blue solid curves). For the N-Cα-C bond angle (Figure 2C), the differential is even larger, in excess of 50% improvement at most resolutions. Consideration of R-factors shows that these improvements in residuals take place with no degradation of the fit of the model to the diffractions data (Figure 2B inset). In terms of the bond angle residuals, it is also gratifying that at medium to low resolutions where in theory tighter restraints are associated with more parsimonious, i.e. better models [19], the residuals using the CDL actually approach closely to those that result from regularization alone. Nevertheless, two surprises in this regard are (1) that at

resolutions between about 2 and 3 Å the plots are rather flat rather than continuing to give decreased residuals as the resolution lessens, and (2) that at resolutions poorer than 3.0 Å, the residuals for both the SVL and CDL libraries actually rise slightly rather than remaining low or even decreasing further. We attribute the first behavior to the Phenix refinement strategy (see methods) that defines resolution specific maximal rms deviation cut-offs, and that once the residual is below the cutoff the weight is not adjusted in such a manner to reduce the residual further. Regarding the second behavior, we hypothesized that this increase is caused by the increasing occurrence in lower resolution models of regions of the backbone that are sufficiently misfit so that they cannot achieve ideal geometry while also maintaining good agreement with the electron density. Such a connection between model quality and bond angle residuals has been shown to exist for the geometry around the Cα atom [38], where fundamentally misfit parts of a structure tend to have distorted bond angles.

Clashscore filtering of the Protein Data Bank

To test this hypothesis, we applied an additional filter to limit the structures included in our refinements based on an independent measure of their potential for having misfit segments. We chose the Molprobity 'clashscore' [39,40] which quantifies non-bonded atomic clashes such as tend to be present in fundamentally misfit regions and ran this analysis on the coordinate sets after modelled water molecules had been removed (see methods). As a confirmation of our hypothesis, applying increasingly stringent filters remarkably had virtually no impact on the residuals for the sensitive N-Cα-C bond angle in structures refined at 2.5 Å resolution or better, yet substantially decreased the residuals of the lowest resolution structures (Figure 3A). Using the most stringent clashscore ≤ 3 filter we applied, the low resolution behavior is completely changed so that it becomes exactly as expected, with the rms deviations from ideality continuously decreasing as resolution worsens.

However, that the clashscore ≤ 3 filter cutoff is too stringent can be seen in that it filters out over half of the structures determined at all resolutions (Figure 3B) even though the many structures filtered at resolutions better than 2.5 Å apparently had no seriously misfit regions since their removal leads to little or no improvement in the geometry residuals at these resolutions (Figure 3A). For our further analyses we chose a clashscore < 6 cutoff, as it appears to remove the large majority of models containing misfit regions, while still retaining in most resolution ranges the majority of models not having misfit regions (Figure 3). We note that the dramatic increase at poorer resolutions of rejected models (rising from ~10% at 2.3 Å resolution to ~80% at 3.5 Å; Figure 3B) highlights the importance of efforts now being invested in developing tools that improve our ability to build high quality models in these medium-low resolution ranges [41–43].

Assessing the backbone-CDL performance on clashscore filtered structures

Using the additional criterion that PDB entries must have a Molprobity clashscore < 6 allows 22052 models to be included in a final set of entries for which a complete set of assessments, like those shown in Figure 2, were carried out. For these results, in addition to reporting the mean rms deviations as a function of resolution, we also indicate the spread of behaviors by denoting the 25th to the 75th percentile range (Figure 4). The average results in

terms of bond angle and bond length deviations are largely the same as was seen for the whole PDB, except there is no anomalous rise in deviations at very low resolution. For bond lengths, the deviations are lowered by 20–30% with the larger improvements occurring at better resolutions (Fig. 4A); and for bond angles, the CDL reduces the residual deviations by roughly one-third across the whole range of resolutions (Fig. 4B), and for the N-Cα-C angle the deviations are cut in half (Fig. 4C).

The changes in R-factor between the SVL and CDL based refinements (Figure 4B inset) also match well the behaviors seen for the re-refinements of the PDB as a whole (Figure 2B inset). Encouragingly, with the use of the CDL, at the same time as the deviations from ideality are dramatically decreased, the R-factors not only do not get worse, but on the whole actually improve slightly. On average the R_{free} decreases by 0.12%, the R_{work} decreases by 0.05% and the R_{free}-R_{work} differential decreases by 0.15%.

Analysis of the range of bond angle and length deviations associated with the central 50 percent of the structures in each resolution bin shows that the distributions are for the most part non-overlapping between the SVL and CDL refinements. This implies that the improved behavior due to use of the CDL is a robust phenomenon applicable to structures in general. Especially striking is the large improvement seen in the residuals associated with the N-Cα-C bond angle (Figure 4C), which for structures refined at 1 Å or better decrease from ~2.25° (with the SVL) to ~1° (with the CDL). What is conceptually gratifying about this improvement is that the ~1° rms deviations associated with this angle when using the CDL library matches well the ~1° rms deviations seen for backbone angles in general suggesting that once conformation is accounted for, all angles have a similar level of intrinsic variability. This implies that rather than the N-Cα-C having a special higher intrinsic level of variation as appeared to be the case based on SVL refinements, it can be seen as having a higher sensitivity to conformation, but a similar intrinsic level of variation.

Outlook

A highlight article published alongside the Berkholz *et al.* [25] description of the backbone CDL, noted that “hopefully the structural biology community will soon adopt these ideas” [44]. By building the CDL into the widely used Phenix package for crystallographic computing, the work here provides such an adoption that will allow the CDL to be used widely in generating more accurate protein crystal structures. The dramatic improvement in the geometric residuals that derive from using the CDL comes with no drawbacks and this reinforces the conclusion that conformation-dependent ideal geometry functions truly are a more accurate representation of reality than are the conventional single value ideal geometry targets. Whereas this work does set a new standard for restraints to be used in crystallographic refinement, it is crucial to note that this backbone CDL (i.e. CDL-v1.2) is not the ultimate library, but is only a first step in this direction. As further ultrahigh resolution protein structures are determined, it will be possible to improve both the accuracy of the library values and the extent of conformational space for which conformation-dependent values can be obtained, rather than reverting to a global average value because too few observations exist to define it well. Further natural extensions will be to include conformation-dependent variations in the ω-torsion angle [45] as well as side chain bond

geometries, such as have already been documented to exist for proline residues [46]. Another possible direction for development would be to account for additional contextual factors (besides conformation) that might systematically influence geometry (e.g.[47]). Also as a parallel development, it will be important that the tools used to validate protein structures (e.g. [48]) be updated to incorporate the CDL concept so that the improved CDL-refined structures will not be flagged as defective.

Experimental Procedures

Implementation of the CDL in Phenix

Until now, the Phenix geometry restraints have been derived from the CCP4 monomer library [49] and converted by the module pdb_interpretation into a central restraints list 'object' that specifies for each restraint all the atoms involved along with the ideal value and ESD. This single, global restraints list is accessed by the individual Phenix programs. To implement the CDL in Phenix, first a CDL-v1.2 data object file was created that contains complete restraints for each of 10,368 possible circumstances corresponding to 8 residue classes (Gly, Pro, Ile/Val, other, and residues in each of these four classes preceding a Pro residue) multiplied by the 1,296 possible 10°×10° ϕ,ψ-bins (see Figure 1D and references [25]and [30] for further details on the contents of the library). Then, a python script 'CDL Module' was written that updates the central restraints list with the CDL specific values. This code first retrieves the needed residue type and ϕ,ψ-angle information from the model object and then looks up the conformation-dependent backbone bond length and angle target values and their ESDs in the CDL data object. Because the ϕ,ψ-angles change as the model shifts during refinement, the CDL module operates at the beginning of each refinement macro-cycle to update the restraints. The additional time taken by the CDL step is minimal; for instance, for the ribosome it takes 11.5 s to update the restraints, which is ~0.3% of the roughly 70 minutes for the complete macro-cycle.

Explicitly selecting the CDL for use in Phenix is done by adding cdl=True to the command-line or the input phil file. It is also available as an option in the graphical user interface [37]. The conventional SVL values are selected at the command line using cdl=False.

Phenix test refinements and analyses

The 25976 structures from the PDB that were re-refined were selected from a set of 42247 structures that were derived from X-ray and neutron studies and had associated diffraction data. For this study we chose a subset of this group by filtering out structures that were based on neutron diffraction (62 entries) or were not protein (849 entries), were solved at worse than 3.55 Å resolution (323 entries), were twinned (781 entries), had <90% completeness of data (5,005 entries), were missing reported R_{work} or R_{free} values (7,666 entries), or had R_{work} > 30% (211 entries), R_{free} > 35% (203 entries), or an R_{free}-R_{work} differential of <1.5% (1,670 entries). (A given structure may occur in more than one of these categories.) Also some structures were not used for miscellaneous reasons such as a large size or difficulty properly handling a ligand that was present. The required R_{free}-R_{work} difference of ≥1.5% was designed to filter out structures that may not have a correctly assigned R_{free} test set. This left 25939 entries, and to these were added back 37 entries that

were at resolutions better than 1.05 Å with an R_{free}-R_{work} differential of <0.5%, because at these resolutions even with a correct R_{free} test set, overfitting may be minimal enough such that the differential may truly be this small. This procedure yielded the final 25976 structures used in the initial re-refinement tests.

For the clashscore filtering, preliminary tests showed that clashscore not only filtered out many low resolution models, but surprisingly, also removed many ultrahigh resolution models. We thought that the higher clashscores of some ultra-high resolution models could be the result of alternate conformations of ordered water molecules, and found that the loss of ultra-high resolution structures could be substantially lessened if water molecules were removed from the structures before the clashscore calculations were done. Thus, for all analyses presented here, clashscores were calculated ignoring water clashes.

The structures in this study were refined using *phenix.refine* with and without the use of the CDL. Each refinement was performed for ten macro-cycles with the weight between the x-ray and geometry terms optimized at each step. The optimization of the weights are described in an article by Afonine *et al.* [50]. Briefly, a primary optimization criterion is a maximum cut-off for bond length and bond angle root-mean-square (rms) deviations. These cutoffs are resolution dependent with current bond length/angle values set to 0.025Å/3.0° for resolutions better than 1.5 Å, 0.02Å/2.5° for resolutions between 1.5 and 2.0 Å, and 0.015Å/2.0° for resolutions poorer than 2.0 Å. If the rms deviations are above these limits, weighting of the geometry terms is increased until the values drop below the cut-offs. Once this has been achieved, other criteria such as minimizing R_{work} and R_{free} are pursued.

For refinements carried out at resolutions better than 1.55 Å anisotropic atomic displacements parameters were enabled for all atoms except hydrogen atoms and water molecules. Ligand restraints were generated by eLBOW [51]. For 21 deposited data sets that otherwise qualified, but had no R_{free} test set selected, a test set comprising 10% of the data was chosen automatically by *phenix.refine*. All refinements were performed with Phenix version dev-1021.

Acknowledgments

This work was supported in part by National Institutes of Health (NIH) grant R01-GM083136 (to PAK), by the NIH Project 1P01 GM063210 (to PDA), and the Phenix Industrial Consortium. This work was further supported in part by the US Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Watenpaugh K, Sieker L, Herriott J, Jensen L. Refinement of the model of a protein: rubredoxin at 1.5 Å resolution. Acta Crystallogr Sect B. 1973; 29:943–956.
2. Deisenhofer J, Steigemann W. Crystallographic Refinement of Structure of Bovine Pancreatic Trypsin-Inhibitor at 1.5 a Resolution. Acta Crystallogr Sect B. 1975; 31:238–250.
3. Dodson E, Isaacs N, Rollett J. A Method for Fitting Satisfactory Models to Sets of Atomic Positions in Protein Structure Refinements. Acta Crystallogr A. 1976; 32:311–315.
4. Ten Eyck LF, Weaver LH, Matthews BW. A method of obtaining a stereochemically acceptable protein model which fits a set of atomic coordinates. Acta Crystallogr A. 1976; 32:349–350.
5. Konnert JH. A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units. Acta Crystallogr A. 1976; 32:614–17.

6. Hendrickson W. Stereochemically Restrained Refinement of Macromolecular Structures. Methods Enzymol. 1985; 115:252–270. [PubMed: 3841182]
7. Pauling L, Corey R, Branson H. The Structure of Proteins - 2 Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. Proc Natl Acad Sci U S A. 1951; 37:205–211. [PubMed: 14816373]
8. Diamond R. A Mathematical Model-Building Procedure for Proteins. Acta Crystallogr. 1966; 21:253–266.
9. Bowen, H.; Donohue, J.; Jenkin, D.; Kennard, O.; Wheatley, P.; Whiffen, D. Tables of Interatomic Distances and Configuration in Molecules and Ions. Mitchell, A.; Cross, L., editors. The Chemical Society; London: 1958.
10. Vijayan, M. CRC Handbook of Biochemistry and Molecular Biology, Proteins. 3. Fastman, G., editor. Cleveland: CRC Press; 1976. p. 742-749.
11. Tronrud DE, Ten Eyck LF, Matthews BW. An efficient general-purpose least-squares refinement program for macromolecular structures. Acta Crystallogr A. 1987; 43:489–501.
12. Engh R, Huber R. Accurate Bond and Angle Parameters for X-Ray Protein Structure Refinement. Acta Crystallogr A. 1991; 47:392–400.
13. Engh, R.; Huber, R. Structure Quality and Target Parameters. In: Rossmann, M.; Arnold, E., editors. International Tables for Crystallography. Dordrecht: Kluwer Academic Publishers; 2001. p. 382-392.
14. Stec B, Zhou R, Teeter M. Full-Matrix Refinement of the Protein Crambin at 0.83-Angstrom and 130-K. Acta Crystallogr D. 1995; 51:663–681. [PubMed: 15299796]
15. Vlassi M, Dauter Z, Wilson KS, Kokkinidis M. Structural parameters for proteins derived from the atomic resolution (1.09 angstrom) structure of a designed variant of the ColE1 ROP protein. Acta Crystallogr D. 1998; 54:1245–1260. [PubMed: 10089502]
16. Jaskolski M, Gilski M, Dauter Z, Wlodawer A. Numerology versus reality: a voice in a recent dispute. Acta Crystallogr D. 2007; 63:1282–1283. [PubMed: 18084076]
17. Jaskolski M, Gilski M, Dauter Z, Wlodawer A. Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? Acta Crystallogr D. 2007; 63:611–620. [PubMed: 17452786]
18. Stec B. Comment on - Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? by Jaskolski, Gilski, Dauter & Wlodawer (2007). Acta Crystallogr D. 2007; 63:1113–1114. [PubMed: 17881830]
19. Tickle IJ. Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values. Acta Crystallogr D. 2007; 63:1274–1281. [PubMed: 18084075]
20. Karplus PA, Shapovalov MV, Dunbrack RL Jr, Berkholz DS. A forward-looking suggestion for resolving the stereochemical restraints debate: ideal geometry functions. ACTA Crystallogr D. 2008; 64:335–336. [PubMed: 18323629]
21. Schafer L, Ewbank J, Klimkowski V, Siam K, Van Alsenoy C. Predictions of Relative Structural Trends from Abinitio Derived Standard Geometry Functions. Theochem J Mol Struct. 1986; 28:141–158.
22. Schafer L, Cao M. Predictions of Protein Backbone Bond Distances and Angles from First Principles. Theochem J Mol Struct. 1995; 333:201–208.
23. Karplus P. Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci. 1996; 5:1406–1420. [PubMed: 8819173]
24. Jiang X, Cao M, Teppen B, Newton S, Schāfer L. Predictions of protein backbone structural parameters from first principles: Systematic comparisons of calculated N-Ca-C′ angles with high-resolution protein crystallographic results. J Phys Chem. 99:10521–10525.
25. Berkholz DS, Shapovalov MV, Dunbrack RL Jr, Karplus PA. Conformation Dependence of Backbone Geometry in Proteins. Structure. 2009; 17:1316–1325. [PubMed: 19836332]
26. Berkholz DS, Krenesky PB, Davidson JR, Karplus PA. Protein Geometry Database: a flexible engine to explore backbone conformations and their relationships to covalent geometry. Nucleic Acids Res. 2010; 38:D320–D325. [PubMed: 19906726]

27. Esposito L, Balasco N, De Simone A, Berisio R, Vitagliano L. Interplay between Peptide Bond Geometrical Parameters in Nonglobular Structural Contexts. Biomed Res Int. 2013:326914. [PubMed: 24455689]
28. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D. Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. 2014; 23:47–55. [PubMed: 24265211]
29. Sheldrick GM. A short history of SHELX. Acta Crystallogr A. 2008; 64:112–122. [PubMed: 18156677]
30. Tronrud DE, Berkholz DS, Karplus PA. Using a conformation-dependent stereochemical library improves crystallographic refinement of proteins. ACTA Crystallogr D. 2010; 66:834–842. [PubMed: 20606264]
31. Tronrud DE, Karplus PA. A conformation-dependent stereochemical library improves crystallographic refinement even at atomic resolution. Acta Crystallogr D. 2011; 67:699–706. [PubMed: 21795811]
32. Zwart PH, Afonine PV, Grosse-Kunstleve RW, Hung L-W, Ioerger TR, McCoy AJ, McKee E, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Storoni LC, Terwilliger TC, Adams PD. Automated structure solution with the PHENIX suite. Methods Mol Biol Clifton NJ. 2008; 426:419–435.
33. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D. 2010; 66:213–221. [PubMed: 20124702]
34. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Echols N, Headd JJ, Hung L-W, Jain S, Kapral GJ, Kunstleve RWG, McCoy AJ, Moriarty NW, Oeffner RD, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH. The Phenix software for automated determination of macromolecular structures. Methods. 2011; 55:94–106. [PubMed: 21821126]
35. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
36. Afonine PV, Headd JJ, Terwilliger TC, Adams PD. New tool: phenix.real_space_refine. Comput Crystallogr Newsl. 2013; 4:43–44.
37. Echols N, Grosse-Kunstleve RW, Afonine PV, Bunkoczi G, Chen VB, Headd JJ, McCoy AJ, Moriarty NW, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Adams PD. Graphical tools for macromolecular crystallography in PHENIX. J Appl Crystallogr. 2012; 45:581–586. [PubMed: 22675231]
38. Lovell SC, Davis IW, Adrendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by C alpha geometry: phi, psi and C beta deviation. Proteins Struct Funct Genet. 2003; 50:437–450. [PubMed: 12557186]
39. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. ACTA Crystallogr D. 2010; 66:12–21. [PubMed: 20057044]
40. Davis IW, Murray LW, Richardson JS, Richardson DC. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res. 2004; 32:615–619.
41. Karmali AM, Blundell TL, Furnham N. Model-building strategies for low-resolution X-ray crystallographic data. Acta Crystallogr D. 2009; 65:121–127. [PubMed: 19171966]
42. Brunger AT, Das D, Deacon AM, Grant J, Terwilliger TC, Read RJ, Adams PD, Levitt M, Schroeder GF. Application of DEN refinement and automated model building to a difficult case of molecular-replacement phasing: the structure of a putative succinyl-diaminopimelate desuccinylase from Corynebacterium glutamicum. Acta Crystallogr D. 2012; 68:391–403. [PubMed: 22505259]
43. Headd JJ, Echols N, Afonine PV, Grosse-Kunstleve RW, Chen VB, Moriarty NW, Richardson DC, Richardson JS, Adams PD. Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. Acta Crystallogr D. 2012; 68:381–390. [PubMed: 22505258]

44. Dauter Z, Wlodawer A. Proteins Do Not Have Strong Spines After All. Structure. 2009; 17:1278–1279. [PubMed: 19836327]
45. Berkholz DS, Driggers CM, Shapovalov MV, Dunbrack RL, Karplus PA. Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. Proc Natl Acad Sci U S A. 2012; 109:449–453. [PubMed: 22198840]
46. Ho BK, Coutsias EA, Seok C, Dill KA. The flexibility in the proline ring couples to the protein backbone. Protein Sci. 2005; 14:1011–1018. [PubMed: 15772308]
47. Touw WG, Vriend G. On the complexity of Engh and Huber refinement restraints: the angle tau as example. Acta Crystallogr D. 2010; 66:1341–1350. [PubMed: 21123875]
48. Read RJ, Adams PD, Arendall WB, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Luetteke T, Otwinowski Z, Perrakis A, Richardson JS, Sheffler WH, Smith JL, Tickle IJ, Vriend G, Zwart PH. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. Structure. 2011; 19:1395–1412. [PubMed: 22000512]
49. Vagin AA, Steiner RA, Lebedev AA, Potterton L, McNicholas S, Long F, Murshudov GN. REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Crystallogr D. 2004; 60:2184–2195. [PubMed: 15572771]
50. Afonine PV, Echols N, Grosse-Kunstleve RW, Moriarty NW, Adams PD. Improved target weight optimization in phenix.refine. Comput Crystallogr Newsl. 2011; 2:99–103.
51. Moriarty NW, Grosse-Kunstleve RW, Adams PD. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. Acta Crystallogr D. 2009; 65:1074–1080. [PubMed: 19770504]

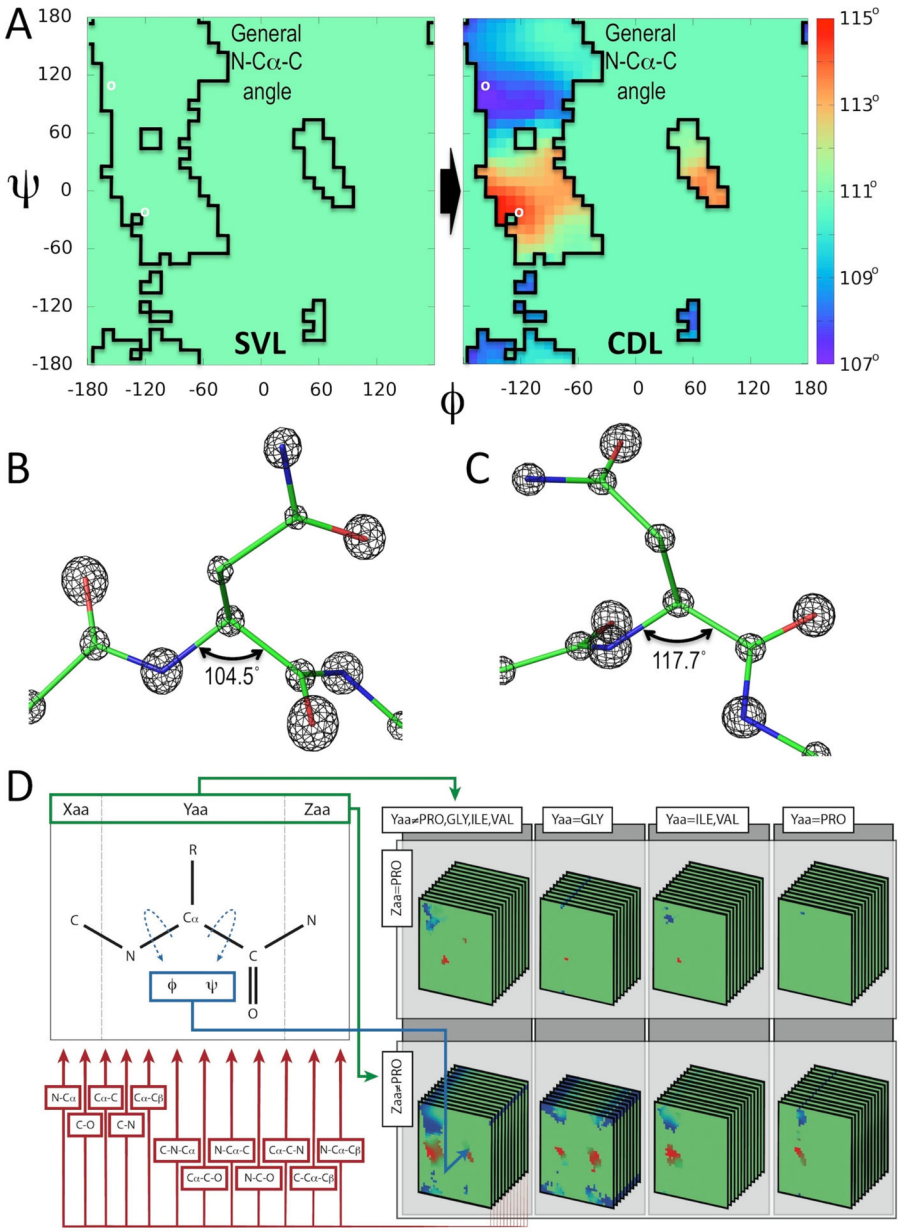

Figure 1. The conformation-dependent library (CDL) concept

(A) Ramachandran plots emphasizing the large increase in information content that is associated with shifting from the conventional SVL library in current use ([13]; left hand plot) to the CDL library ([25]; right hand plot) that we have here incorporated into Phenix. Using the color scheme indicated to the right, each plot shows the N-Cα-C bond angle targets for general residues (the 18 non-Gly, non-Pro residues in the case of the SVL and the 16 non-Gly, non-Pro, non-Ile/Val, non-PrePro residues in the case of the CDL). For the CDL, conformation-dependent the N-Cα-C bond angle targets are defined for 10×10° bins of φ and ψ whereas for the SVL all are given the single value of 111.0°. Shown in both panels are small white circles marking the φ,ψ-angles of the residues shown in panels B and C, and black outlines indicating the regions sufficiently populated so that the CDL library

provides actual conformation-dependent values rather than defaulting to a global average value. The global average value for the right hand panel is 110.8°, which can be perceived as having a slightly different hue than the left hand panel color that represents 111.0°. It is of interest to note that the previous adjustments in SVL target values over the last 60 years are equivalent to making such a slight change in hue, while switching from the SVL to the CDL paradigm introduces a rainbow of greater information. (B) The model and 0.86 Å resolution electron density map contoured at 7 ρ_{rms} showing the evidence for the N-Cα-C bond angle of residue Asn44 in PDB entry 1g6x (with φ,ψ-angles=−162°,+106°) that is observed to be 104.5°. (C) Same as B but for residue Asn108 of PDB entry 4ayo (with φ,ψ-angles=−122°, −26°) with its 0.85 Å resolution map contoured at 7 ρ_{rms} and an observed N-Cα-C bond angle of 117.7°. The examples in panels B and C were found using the Protein Geometry Database [26]. (D) Schematic of information content of the backbone CDL showing how a central residue (Yaa) and its C-terminal neighbour (Zaa) define one of 8 residues classes (green lines), and the φ,ψ-angles of the residue specify which restraint values to obtain from that class of residue (blue line) for each of the up to 7 backbone bond angles and 5 backbone bond lengths (red lines). The coloring scheme for the CDL plots of each residue type is similar to those in panel A, but with a common background color for simplicity.

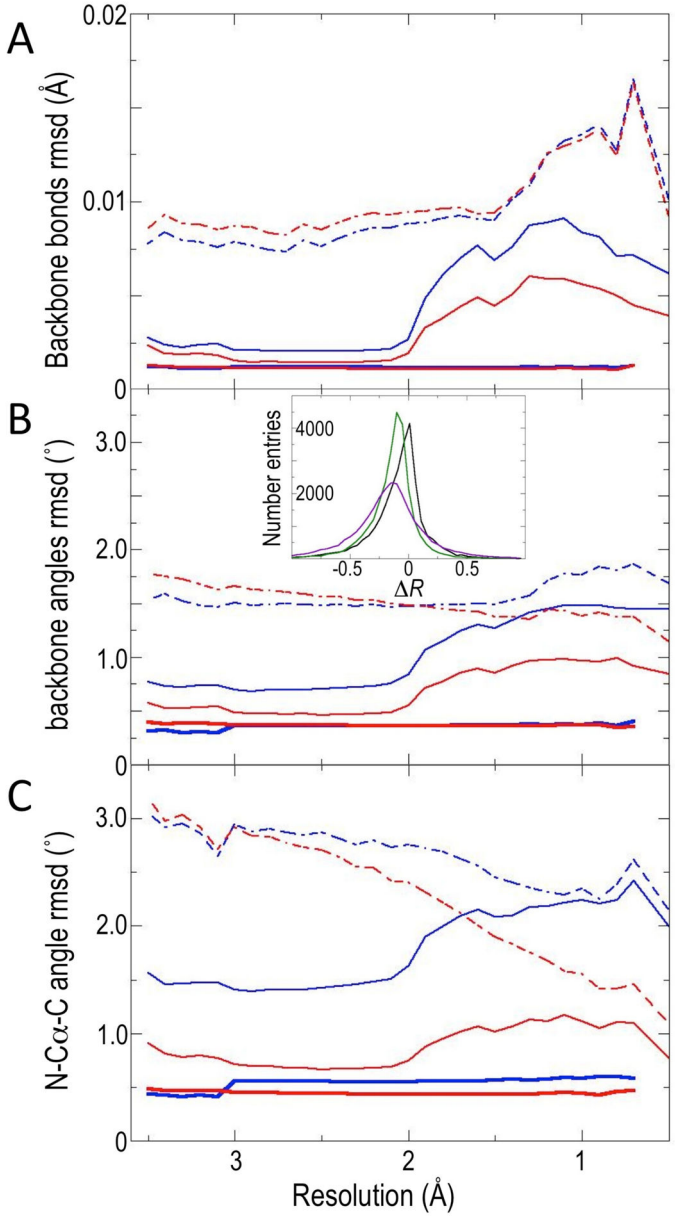

Figure 2. Comparing outcomes of the SVL- *vs.* CDL-based re-refinements of 25976 PDB entries

(A) Shown are the average rms deviations of the backbone bond lengths for structures grouped in 0.1 Å resolution bins. Colors distinguish the results derived from SVL (blue) and the CDL-based (red) calculations. The three pairs of curves are the rms deviations from the library values of the PDB entries as deposited (dashed lines), after re-refinement using Phenix (solid lines), and after regularisation (thick lines). Dotted lines indicate the values obtained if 'alternate location' atoms are included in the rmsd values. (B) same as 'A' but for backbone angles. (C) same as 'A' but for the N-Cα-C bond angle. An inset in panel B shows the absolute changes in *R*-factors (as measured in percent) for CDL re-refined minus SVL re-refined structures. Changes are shown for R_{free}(green), R_{work} (black), and R_{free}+R_{gap} (purple) where R_{gap} is R_{free}-R_{work}.

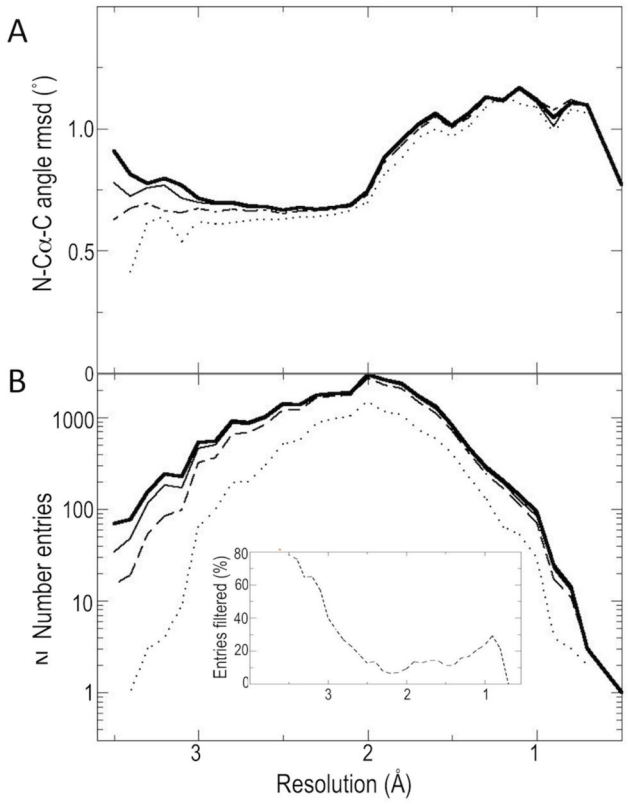

Figure 3. Clashscore filtering improves behavior of lower resolution structures during re-refinement

(A) The rms deviations for the N-Cα-C angle, as a representative indicator of model geometry quality, are plotted as a function of resolution for pdb entries re-refined using the CDL library. The four curves are shown are based on all structures (thick solid), or subsets of the structures that after re-refinement using the SVL had Molprobity clashscore of <9 (thin solid), <6 (dashed) or <3 (dotted). The number of structures in each of the four groups are 25976, 24867, 22052, and 10871, respectively. The clashscore filtering was carried out on PDB entries after discrete modelled water molecules were deleted. (B) A log-scale plot of the number of protein models as a function of resolution in each of the groups shown in figure 3A, with matching line types. The inset shows as a function of resolution what percent of the models at each resolution range were removed by the clashscore <6 filtering that was used for selecting files for the following refinement tests.

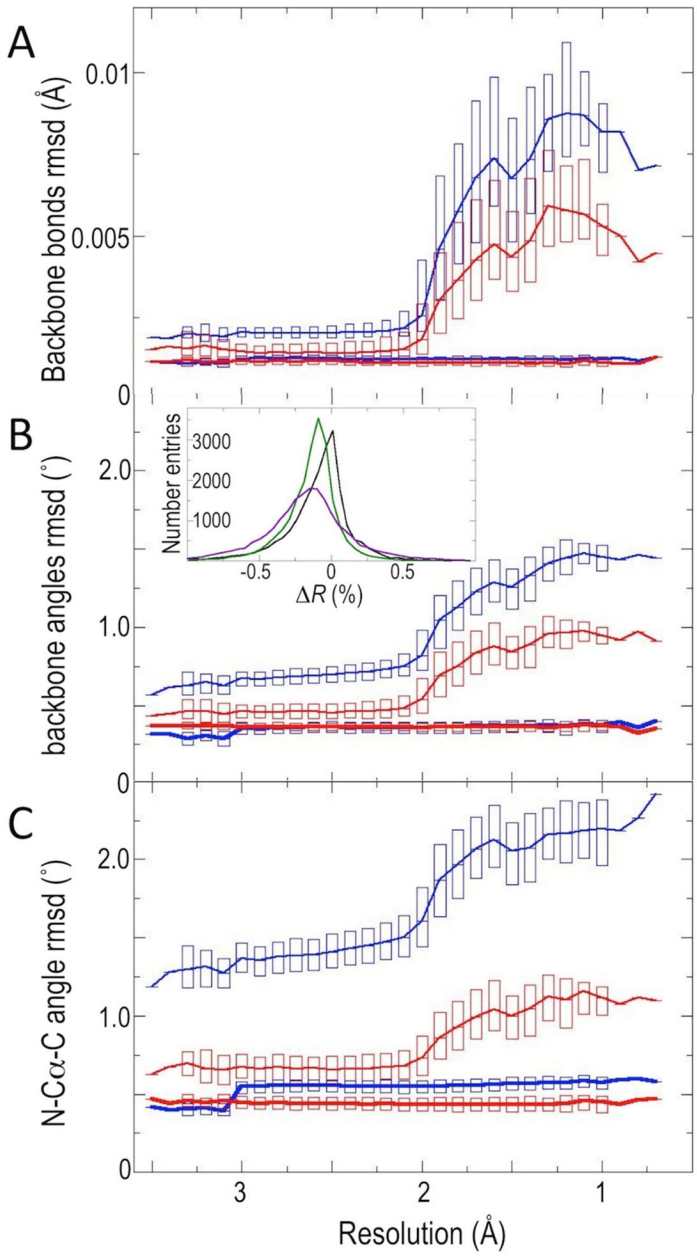

Figure 4. Comparing outcomes of the SVL- *vs.* CDL-based re-refinements of 22052 PDB entries surviving the Clashscore filter

Panels (A) – (C) show results as in Figure 2, comparing the SVL- (blue traces) and CDL-based (red traces) results, except that for each plot only two pairs of curves are shown: the rms deviations from the library values of the PDB entries after re-refinement using Phenix (solid lines), and after regularisation (thick lines). In addition to presenting the average rms deviations as a function of resolution, for panels (A) through (C), boxes are included that represent the range covered from the 25th to the 75th percentile values within each resolution bin. Also, as in Figure 2, panel B contains an inset showing the small changes in R_{free}

(green), R_{work} (black), and R_{free}+R_{gap} (purple) that occur upon changing from the SVL to the backbone CDL.

Table 1

Phenix refinement results for 4 previously studied test-cases using the SVL and CDL

Residuals are presented for the backbone (bkb) parameters only, as these are the only ones different in the CDL compared with the SVL, and for all geometric terms (i.e. including side chain) as this is the parameter normally reported in papers. As in our previous studies [30,31], the deviations for the important N-Cα-C angles are also reported. Refinement protocols were not optimized for these test cases, but used the standard Phenix weighting.

PDB entry	Resol (Å)	Restraint library	R_{work}/R_{free} (%)	Bkb bonds (Å)	All bonds (Å)	Bkb angles (°)	All angles (°)	N-Cα-C (°)
1jb9	1.70	SVL	14.99/18.17	0.013	0.016	1.51	1.52	2.18
		CDL	15.18/18.07	0.005	0.009	0.74	1.01	0.98
3eoj	1.30	SVL	13.53/16.13	0.008	0.012	1.60	1.49	2.54
		CDL	13.69/16.10	0.005	0.010	0.98	1.15	1.27
3dk9	0.95	SVL	13.67/15.25	0.009	0.010	1.40	1.36	2.06
		CDL	13.70/15.30	0.006	0.009	0.79	1.02	1.00
2vb1	0.65	SVL	9.78/9.99	0.007	0.008	1.37	1.35	1.98
		CDL	9.80/10.00	0.005	0.007	0.85	1.07	1.18