STRUCTURAL
BIOLOGY

ISSN: 2059-7983
journals.iucr.org/d

A complete Fourier-synthesis-based backbone-conformation-dependent library for proteins

Dale E. Tronrud and P. Andrew Karplus

Acta Cryst. (2021). D**77**, 249–266

Acta Cryst D
STRUCTURAL BIOLOGY

ISSN 2059-7983

A complete Fourier-synthesis-based backbone-conformation-dependent library for proteins

Dale E. Tronrud and P. Andrew Karplus*

Department of Biochemistry and Biophysics, College of Science, Oregon State University, Corvallis, OR 97331, USA. *Correspondence e-mail: karplusp@oregonstate.edu

Received 23 June 2020
Accepted 16 December 2020

Edited by K. Diederichs, University of Konstanz, Germany

Keywords: conformation-dependent stereochemical library; refinement; ideal geometry; restraints; Fourier representation.

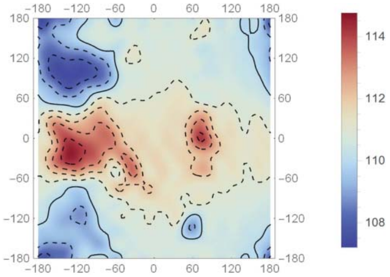
Supporting information: this article has supporting information at journals.iucr.org/d

While broadening the applicability of (φ/ψ)-dependent target values for the bond angles in the peptide backbone, sequence/conformation categories with too few residues to analyze via previous methods were encountered. Here, a method of describing a conformation-dependent library (CDL) using two-dimensional Fourier coefficients is reported where the number of coefficients for individual categories is determined via complete cross-validation. Sample sizes are increased further by selective blending of categories with similar patterns of conformational dependence. An additional advantage of the Fourier-synthesis-based CDL is that it uses continuous functions and has no artifactual steps near the edges of populated regions of φ/ψ space. A set of libraries for the seven main-chain bond angles, along with the ω and ζ angles, was created based on a set of Fourier analyses of 48 368 residues selected from high-resolution models in the wwPDB. This new library encompasses both *trans*- and *cis*-peptide bonds and outperforms currently used discrete CDLs.

1. Introduction

The knowledge of standard values for bond lengths and angles in proteins has been and continues to be a crucial foundation for both predictive and experimental model building as well as validation. Since the pioneering work of Pauling, Corey and Branson (Pauling *et al.*, 1951), the dominant paradigm has been that a single ideal value can be used for each chemical type of bond, and sets of these values have been updated over time (Bowen *et al.*, 1958; Vijayan, 1976; Engh & Huber, 1991, 2001). This is now changing because, as was anticipated by earlier theoretical (see, for example, Dodson *et al.*, 1976; Jiang *et al.*, 1995) and empirical (Karplus, 1996; Jiang *et al.*, 1995) studies, ultrahigh-resolution, mostly globular, protein crystal structures (Berkholz *et al.*, 2009), along with membrane proteins and collagen and amyloid-forming peptides (Esposito *et al.*, 2013), show that the main-chain bond angles of polypeptides do not have single fixed ideal values. Instead, they vary as a function of the conformational angles φ and ψ. It was seen that backbone bond lengths also vary (Berkholz *et al.*, 2009; Improta *et al.*, 2015), but these variations are less reliably determined and are small enough to be of little practical significance for modeling accuracy.

So that this complexity of ideal backbone bond angles could be accounted for during modeling and validation, a conformation-dependent library (CDL) for the geometry of *trans*-peptide units was developed and described as CDL-v1.2 (Berkholz, Shapovalov *et al.*, 2010; Berkholz *et al.*, 2009; Tronrud *et al.*, 2010). It was then shown that using this library during crystallographic refinement, as opposed to a conventional single-value library (SVL), results in 30–50% improvements in how well the modeled backbone bond angles



agree with the library values, along with a very small improvement in the fit to the crystallographic data (Tronrud *et al.*, 2010; Tronrud & Karplus, 2011; Moriarty *et al.*, 2014, 2016). It has also been shown that predictive modeling using *Rosetta* (Das & Baker, 2008) gives improved results if the backbone bond angles and the ω torsion angles are allowed to vary from the conventional standard target values (Conway *et al.*, 2014; Stein & Kortemme, 2013). Of course, target values such as these, derived from crystal structures, should not be mixed with molecular-mechanics force-field terms (for example van der Waals and electrostatics) because crystal structures do not represent energy-minimal models, but instead time and space averages of huge numbers of molecules, and their atomic locations represent the equilibrium values that result from the interplay of the various forces present.

In this first-generation backbone CDL (Berkholz *et al.*, 2009; Tronrud *et al.*, 2010), residues were grouped into eight classes: the four residue-based categories were proline, glycine, isoleucine or valine, and 'other', and each of these categories appeared either in a general context or in a context where the following residue was a proline, since residues preceding proline have a distinct behavior (see, for example, Nicholson *et al.*, 1992; Karplus, 1996; Ho & Brasseur, 2005). Introducing a shorthand nomenclature that uses *t* for *trans*-peptides, *c* for *cis*-peptides and **X** for 'other' residues (and is more fully discussed in Section 2.1), these eight categories are $t\mathbf{P}t$, $t\mathbf{G}t$, $t\mathbf{IV}t$, $t\mathbf{X}t$, $t\mathbf{P}t\mathrm{P}$, $t\mathbf{G}t\mathrm{P}$, $t\mathbf{IV}t\mathrm{P}$ and $t\mathbf{X}t\mathrm{P}$. For each of seven backbone angles in each residue category, the empirical patterns derived from a search of the Protein Geometry Database (PGD; Berkholz, Krenesky *et al.*, 2010) were smoothed using kernel density averaging (Berkholz *et al.*, 2009) and the resulting function was discretized into a 36 × 36 (= 1296) element lookup table that provided the empirical target value for each 10 × 10° box of φ/ψ space (Fig. 1*a*). For lookup-table elements with fewer than three observations, the local average was deemed to be unreliable and was replaced by the global average value.

While this CDL was a conceptual and practical step forward, it has certain limitations. A first set of limitations derives from the discrete nature of the library. This means that the CDL is not differentiable and cannot be used for energy or molecular dynamics (see, for example, Bernauer *et al.*, 2011). Even worse, the discrete binning of the library leads to unnatural steps, including rather dramatic 'cliffs' that often occur at the edges of populated regions because fewer than three observations exist in the adjacent bins [see, for example, Fig. 1*a* near $(\varphi, \psi) = (-120°, -30°)$]. As these edges typically occur at conformations involving atomic clashes that are lessened through bond-angle distortions (see, for example, Berkholz *et al.*, 2009) the abrupt transition back to global average values is completely inappropriate. In reality, the adjacent entry would be expected to have even more extreme bond angles to alleviate the more extreme potential collision (see, for example, Brereton & Karplus, 2015).

This expectation is shown to be true by comparing the original CDL with one based on a larger data set (compare Figs. 1*a* and 1*b*). While the target values for the well sampled regions change little, the updated CDL provides values for many of the regions around the edges of the old CDL, and these newly included regions do indeed continue the trends that were apparent in the conformations that were inside the old boundaries.

A second limitation relates to the poor ability of this type of CDL to handle small populations. For such populations, a large majority of the φ/ψ bins will have

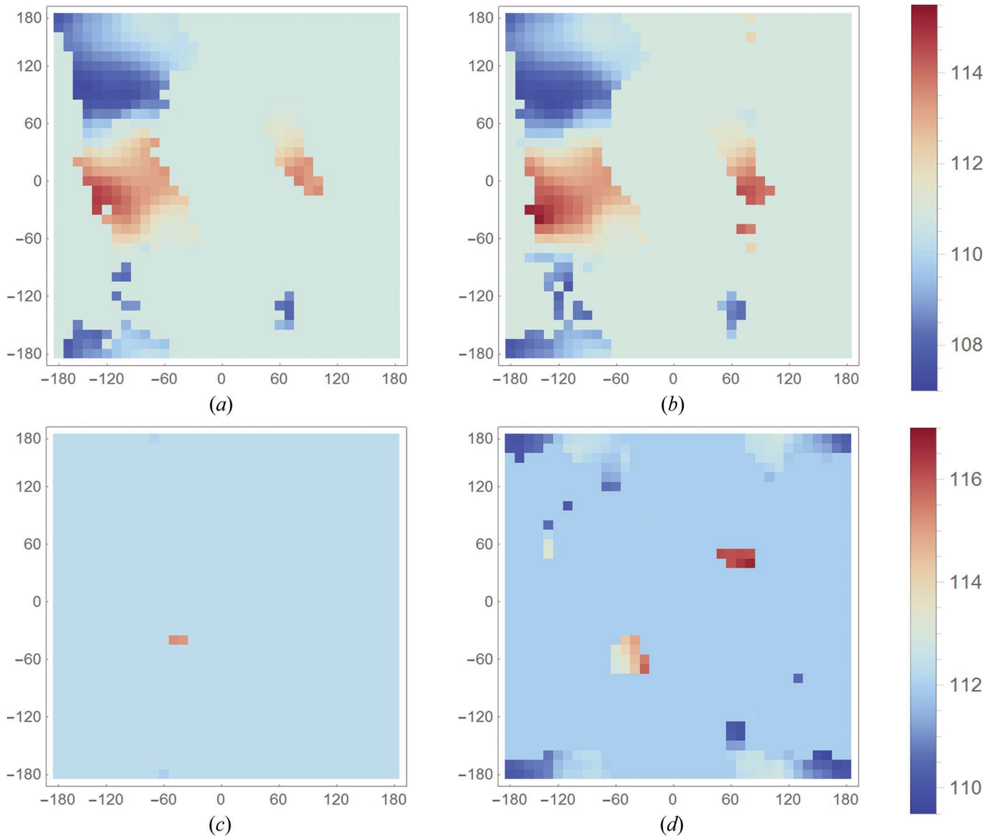


Figure 1
Discrete NCaC CDL for a large and a small residue category. (*a*) Discrete CDL for $t\mathbf{X}t$ residues based on 10 921 residues from a PGD survey performed on 9 October 2009. (*b*) CDL produced by an equivalent analysis using 28 944 residues derived from a PGD search on 4 April 2016. (*c*) Discrete CDL based on 44 residues from the 2009 data for $t\mathbf{G}t\mathrm{P}$ residues. (*d*) Equivalent CDL based on 1267 residues from the 2016 data. The number of residues for this category increased primarily owing to the relaxation of the model resolution filter from 1 Å to 1.5 Å. For each pair of plots the color scheme is defined by the bar on the right. The φ, ψ and NCaC angles are all measured in degrees.

Table 1
Abbreviations.

For a more detailed description of the angles, see Section 2.3.

CaCO	$C^{\alpha}-C-O$ bond angle
CaCN	$C^{\alpha}-C-N_{+1}$ bond angle
CbCaC	$C^{\beta}-C^{\alpha}-C$ bond angle
CNCa	$C_{-1}-N-C^{\alpha}$ bond angle
NCaC	$N-C^{\alpha}-C$ bond angle (also known as τ)
NCaCb	$N-C^{\alpha}-C^{\beta}$ bond angle
OCN	$O-C-N_{+1}$ bond angle
ζ	(Zeta) dihedral angle which describes the volume enclosed by the C^{α} atom and its ligands. It is the angle between the planes defined by (C^{α}, N, C) and (N, C, C^{β}).
ω	$C^{\alpha}-C-N_{+1}-C^{\alpha}_{+1}$ torsion angle of the peptide bond that follows the C^{α} atom
CCV	Complete cross-validation
CDL	Conformation-dependent library
FS	Fourier synthesis
PGD	Protein Geometry Database; currently unavailable
R.m.s.d.	Root-mean-square deviation
SVL	Single-value library
wwPDB	Worldwide Protein Data Bank
$h_{\max}$	Largest Fourier coefficient index (in any direction) in a given FS
$\sigma_{\rm ccv}$	Estimate of the standard deviation of a target value arrived at through CCV
t	*trans*
c	*cis*

fewer than three residues, resulting in a library that is not very different from a conventional SVL. For instance, the original CDL for the *t***G***t*P class of residues only had four φ/ψ bins with conformation-dependent values (Fig. 1*c*). With the larger data set, both the area of space covered and the range of target values has widened substantially (Fig. 1*d*). The increase in the diversity of target values occurs because in this case the new CDL has enough residues in the corners of the Ramachandran plot and in the α_L region (*i.e.* φ, $\psi \simeq +60°, +50°$) to justify using specific target values rather than global averages, and these regions contain considerably lower and higher values for the NCaC angle, respectively (see Table 1 for angle abbreviations).

A third limitation is simply that the CDL as developed was restricted to residues linked by *trans*-peptide bonds, and thus is not appropriate for residues with a *cis*-peptide link either before or after them. Although *cis* linkages are rare, occurring in only 0.3% of peptide bonds, with 6% of proline resides being preceded by a *cis*-peptide and 0.04% of all other amino acids (Weiss *et al.*, 1998), they have been shown to have distinct geometric features (Jabs *et al.*, 1999). Despite these differences, they are currently not appropriately handled in the SVLs used in crystallographic refinement. To our knowledge, the only *cis*-peptide bond angle treated as a special case in current SVL restraint libraries is the CNCa bond angle for a *cis*-proline residue (Engh & Huber, 2001), even though other bond angles are also known to differ for *cis*-peptide bonds whether prior to a proline residue or not (Jabs *et al.*, 1999).

Our goal in this work was to create a new CDL which overcomes these limitations. The primary obstacle to success was the absence of specific structural examples to define our library. This lack of data exists both in the sequence/conformation space, where some amino-acid types are rarely found next to a *cis*-peptide bond, and in every (φ/ψ) plot, where large regions are unoccupied. Despite the extreme rarity of these instances, a library of target values must always return a value to its calling program. The fact that something has not been seen before does not indicate that such a model will never be seen, if only by mistake. While we wish for a library with great accuracy in situations where there are data to define that accuracy, we do not believe that accuracy is necessary for those which are exceedingly rare. In those cases a library need only provide plausible target values which the modeling program can use while it tries to repair the likely faulty model that it has been given.

In seeking to extend the CDL concept to include residues linked by *cis*-peptide bonds, we realized that it was essential that we develop an approach that could capture real trends in bond-angle variations from small populations. We conceived of representing the conformational dependence of bond angles as a Fourier series (FS), which is well known as an effective way to model periodic functions with varying levels of detail. We found that this not only helped with capturing trends for small populations, but at the same time, because they are continuous functions, solved the problems associated with the discontinuous nature of our first-generation CDL. Also, through developing a complete cross-validation (CCV) strategy to optimize the level of detail included in the Fourier analysis, we are able to assess the quality of any library. Such an analysis shows that the FS-based CDLs are an improvement over SVLs and over the previous, discrete CDL.

Here, we present this new FS-based approach and use it to generate CDLs for the conformational dependence of the backbone bond angles for all 32 classes of residues which include all *cis*- and *trans*-peptide possibilities. Using new capabilities of the PGD, we were also able to create conformation-dependent target values for the peptide plane torsion angle (ω) and a dihedral angle (called by some ζ; Laskowski *et al.*, 1993) designed to reflect the chirality (in its sign) and tetrahedral nature of the C^{α} atom by assessing how far the C^{α} atom is from the plane formed by N, C^{β} and C. The definitions of the names used for all of the angles in the library can be found in Table 1.

We designate this FS-based CDL library CDL-v2.0.

2. Methods

2.1. Nomenclature for the 32 categories

For this study, residues were categorized into the same eight sequence types that were used in our previous work (Tronrud *et al.*, 2010). This set of categories was embellished by adding special cases for the *cis*/*trans* configuration of the peptide bonds before and after the residue. This additional complexity required that we develop a novel nomenclature.

Each of the 20 canonical amino acids is assigned to one of four classes: **G** for glycine, **P** for proline, **IV** for either isoleucine or valine, and **X** for the others. If the amino acid is followed in the protein by a proline residue, we add the letter P to the end of the category name. We indicate a *cis*-peptide bond with the letter *c* and a trans-peptide bond with a *t*. Since every nonterminal amino-acid residue within a protein has a peptide bond both before and after it, we place either a *c* or a *t* on either side of the central designator in the category name.

A specific example of this nomenclature would be an arginine residue that has *trans*-peptide bonds on either side and is followed by a histidine residue. This residue would be in category *t***X***t*. If this residue were instead followed by a proline residue, it would be in category *t***X***t*P. A further example would be the most frequently occurring category containing a *cis*-peptide bond, which is a *cis*-proline residue followed by a *trans*-peptide bond and something other than a proline residue. This category is named *c***P***t*.

While this notation is focused on residues, many papers that consider *cis*/*trans* conformations instead focus on the peptide bonds. It is important to note that the presence of a single *cis*-peptide bond will result in two neighboring residues with a *cis* notation. For example, if the structure is . . . Ala(*t*)His(*c*) Pro(*t*)Ser(*t*) . . . then the central residues will be categorized as *t***X***c*P and *c***P***t*. Each *cis*-peptide will potentially be found associated with two residues, although either instance may not be present in the search results if it fails one of the filters.

2.2. Protein Geometry Database searches

The data sets were created using the Protein Geometry Database (PGD; Berkholz, Krenesky *et al.*, 2010) web service, which sadly is currently not operational. To avoid N- and C-terminal residues, the search length was set to 4, with the residue of principal interest being in position 2. All searches were performed using the default settings for *R*-factor and $R_{\rm free}$ upper limits (0.25 and 0.3, respectively) and a 25% sequence-identity threshold (Wang & Dunbrack, 2003). The PGD implemented a residue-quality filter based on three quantities derived from the *B* factors of the model: $B_{\rm mc}$, $B_{\rm sc}$ and $B_{\rm g}$ (the average main-chain *B* factor, the average side-chain *B* factor and the *B* factor of the first γ atom, respectively). We used the default cutoff of 25 Å^2 for each of these filters, except that the filters on the *B* factors for side-chain atoms in the first and last residues were removed.

Separate searches were performed for each of the 32 conformation/sequence categories. A *trans*-peptide bond was defined as $140^\circ \le \omega \le 220^\circ$ and a *cis*-peptide bond as $-40^\circ \le \omega \le 40^\circ$. Residues with ω angles outside these ranges were individually examined and all were rejected based on the fit to their electron-density maps (Kleywegt *et al.*, 2004).

The searches should be limited to models based on diffraction data of sufficiently high resolution that they are minimally affected by prior geometry libraries. Fig. 1(*a*) of Tronrud & Karplus (2011) shows that models in the wwPDB determined from X-ray data of resolution higher than 1.5 Å begin to deviate from the nearly universally used CSD-X library and that the r.m.s.d. levels out at resolutions higher than 1 Å. This behavior implies that X-ray data begin to overwhelm the geometry restraints around 1.5 Å resolution and dominate refinement at resolutions higher than 1 Å. All searches would ideally be limited to models of 1 Å resolution or better, but for residue types that occur rarely 1.5 Å is a good compromise.

For each category we performed a search limited to models based on data of 1 Å resolution or better. If this search returned fewer than 500 residues, however, we broadened the cutoff for the category to 1.5 Å.

As even high-resolution models in the wwPDB may contain residues which are simply not correct, we sought to avoid the skewing of the library by incorrect models with outlier angles by individually examining all residues which contained bond lengths or angles deviating by more than eight standard deviations from the mean. We identified 65 such residues and examined each one using *Coot* (Emsley *et al.*, 2010). We removed 38 of these from our data set. Nearly all of the rejected residues had multiple conformations, but too few atoms were split to allow the model to fit the density and have reasonable geometry. Some of the models did not have a map on the Electron Density Server (Kleywegt *et al.*, 2004) and were rejected based solely on the poor modeling of disorder and/or bad geometry.

One of the reviewers of this paper pointed out that *cis*-peptide bonds following N-terminal residues are most likely to be erroneous. Since our PGD searches looked for four-residue fragments, it is possible that the first residue could be an N-terminus and that the leading *cis*-peptide bond could indicate an unreliable model. This led us to review each of the 174 *cis*-nonproline residues in our data set. We found that five were N-terminal; two of these were clearly incorrectly modeled and the remaining three were ambiguous. All of the rest of these members of our data set were of acceptable quality. It is entirely reasonable to expect that there is an equivalent fraction of N-terminal *trans*-peptides that are poorly modeled. Since none of the other 169 fragments exhibited modeling errors, the frequency of low-quality, non-N-terminal residues in the data set is expected to be much smaller.

We recognize that our data set, which contains over 40 000 residues, will contain some that were incorrectly built. To minimize the distortions to the CDL that these errors could introduce, we used strategies to ensure that each φ/ψ function was created based on a large number of residues. These involved both lowering the search cutoff to 1.5 Å resolution for smaller groups and also blending sparsely sampled categories with more populated ones.

2.3. Domain of the library

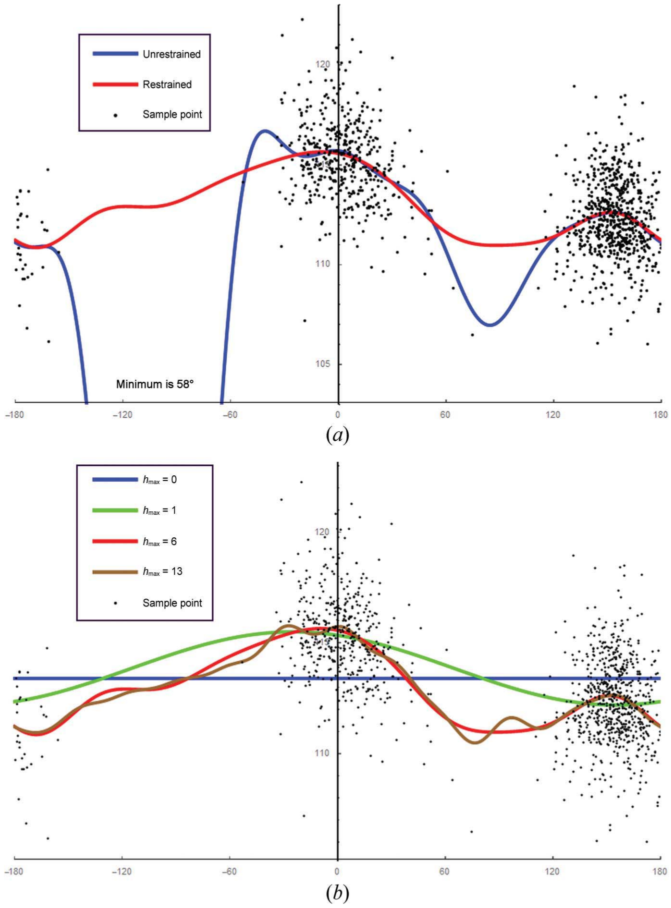
The previously described conformation-dependent library (Berkholz *et al.*, 2009; Moriarty *et al.*, 2014) provided target values for the five protein main-chain bond lengths and seven bond angles. Subsequent to this work, the PGD was enhanced to add support for both the ω and ζ angles (see Table 1 for definitions of the angle names). A discrete CDL was created for the ω angle and added to the *Phenix* refinement program (Liebschner *et al.*, 2019; Moriarty *et al.*, 2016).

At the start of this work, we decided to pursue CDLs for all of the main-chain lengths and angles supplied by the PGD.

With the development of the tools described later, we learned that any statistically significant bond-length variation with φ/ψ was too small to have any practical significance. This left us with the seven bond angles CNCa, NCaCb, NCaC, CbCaC, CaCO, CaCN, the torsion angle ω of the peptide following the residue and the dihedral angle ζ (Table 1).

A protein is a repeating sequence of residues, and therefore each of these angles repeat. While it seems likely that the angles centered on the C^{α} atom will vary most strongly in response to the φ/ψ angles of the very same residue, the same conclusion is not as certain for the angles defined by atoms in the linker between residues. In our previous, discrete, CDL the target values for the leading CNCa and the trailing CaCO, CaCN and OCN angles were chosen (Berkholz *et al.*, 2009; Tronrud *et al.*, 2010), and we adopted this convention here.

Consistent with earlier observations (Esposito *et al.*, 2005), Berkholz *et al.* (2012) concluded that of the two ω angles that bracket the residue, the ω angle that follows the residue is better predicted by a function of the φ/ψ angles of the residue. Therefore, in this paper, it is only the dependence of this trailing ω angle on conformation that is considered and modeled. [This is also consistent with the IUPAC nomenclature, which associates each residue with the peptide bond that follows it (Hoffmann-Ostenhof *et al.*, 1974), but is different from the common usage that associates a residue with the *cis* or *trans* conformation of the peptide preceding it owing to the high probability that a *cis*-peptide will be followed by a proline.] However, the notable impact on the expected bond angles of the *cis* versus *trans* nature of the peptide bonds both preceding and following the central residue are accounted for through our 32 categories of residues. Accounting for these impacts was, in fact, the main motivation for this project.



(*a*)

(*b*)

Figure 2
One-dimensional fits of the variation of the NCaC angle for the **cP***t* category as a function of ψ. The black dots indicate the 1140 observed values. (*a*) The best least-squares fits using $h_{\max}$ = 6 are shown for the unrestrained and the restrained Fourier coefficients. (*b*) Four restrained fits with differing values of $h_{\max}$.

2.4. Least-squares determination of Fourier coefficients to model φ/ψ dependence

The Fourier coefficients of any function, sampled at discrete points, can be calculated with least squares. The function defining the target values for a restraint is defined as

$$\theta(\varphi,\psi|C)=\sum_{h=-h_{\max}}^{h_{\max}}\sum_{k=-h_{\max}}^{h_{\max}}C(h,k)\exp 2\pi i(h\varphi+k\psi), \qquad (1)$$

where the Fourier coefficients $C(h, k)$ are a set of complex numbers and $\theta(\varphi, \psi|C)$ is the modeled value as a function of φ and ψ given that particular set of Fourier coefficients. In this study, θ is the restraint target value. Since such angles must be real-valued, the Fourier coefficients must have Hermitian symmetry [*i.e.* $C(h,k)=C^{*}(\overline{h},\overline{k})$] and $C(0, 0)$ must be real. The amount of detail that is possible for $\theta(\varphi, \psi|C)$ is limited by the parameter $h_{\max}$, which we optimize using complete cross-validation as described below.

We initially tried the simple least-squares residual function

$$\sum_{i=1}^{n}[\theta_{\rm o}(\varphi_i,\psi_i)-\theta(\varphi_i,\psi_i|C)]^2, \qquad (2)$$

to estimate the optimal set of Fourier coefficients. $\theta_{\rm o}(\varphi_i, \psi_i)$ is the list of values to be fitted (here a series of angles associated with particular φ/ψ angles). While this equation is quadratic in $C(h, k)$, its normal matrix is often singular owing to the large regions of φ/ψ space which are not sampled. When the matrix is singular there are an infinite number of sets of Fourier coefficients which fit the sample points equivalently but have very different behaviors in the unsampled regions, often fluctuating wildly (see, for example, the blue trace in Fig. 2*a*).

This problem was addressed in Rowicka & Otwinowski (2004), where it occurs in the similar problem of representing the probability distribution of protein conformational angles in a Ramachandran plot. Their solution is not directly applicable here since we are not working with probability distributions. In addition, it is very difficult to implement.

In our method for preventing this singularity, and damping the fluctuations, we followed the common procedure of adding a restraint. Assuming that the smallest amount of variability consistent with the data is best, we sought to minimize the amplitudes of the Fourier coefficients $C(h, k)$ [except for $C(0, 0)$]. Among a few weighting schemes that we tried, we found that a weight that increased linearly with frequency

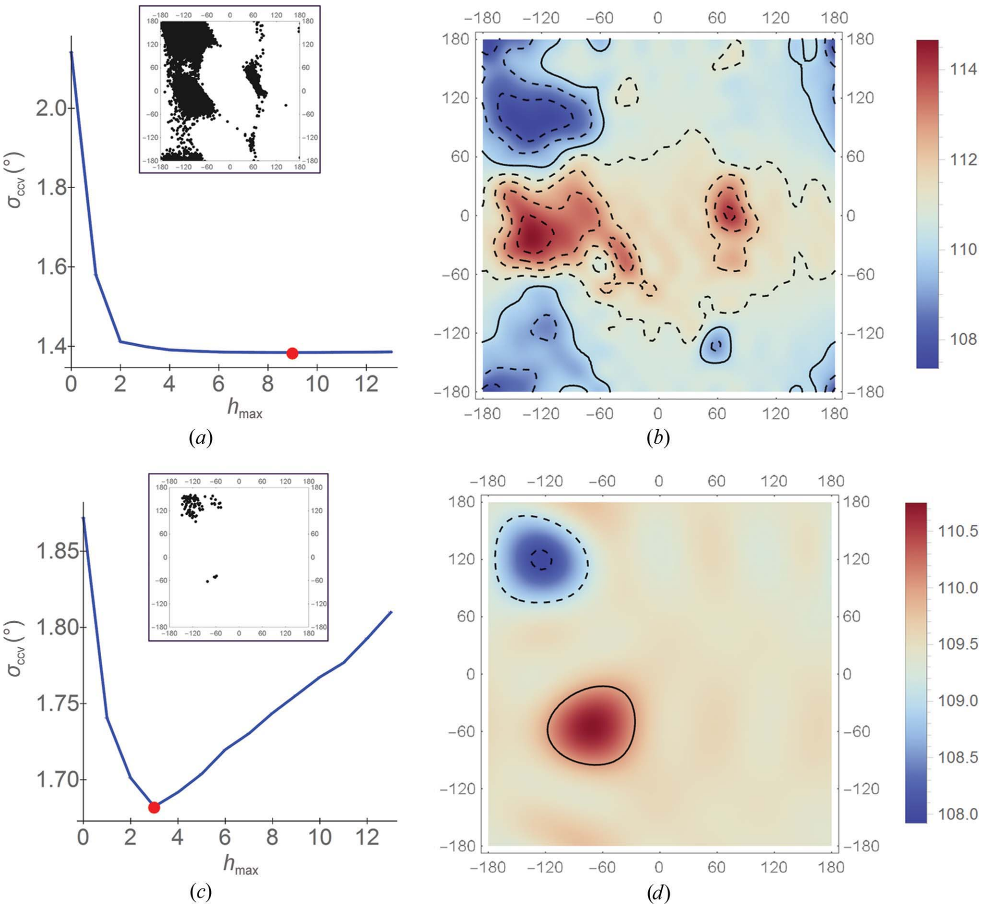


Figure 3
Complete cross-validation to determine the optimal $h_{\rm max}$ for two CDLs. (*a*) Plot of $\sigma_{\rm ccv}$ for the NCaC angle of category *t***X***t*, with 28 944 residues, as a function of $h_{\rm max}$. The optimal value for $h_{\rm max}$ is marked with a red dot. (*b*) φ/ψ plot of the corresponding FS-CDL calculated at the optimal $h_{\rm max}$ of 9. A solid contour line is drawn at multiples of 5°, with dotted contours every 1°. (*c*) The same as in (*a*) but for category *t***IV***c*P, with 97 residues. (*d*) The FS-CDL for category *t***IV***c*P calculated at the optimal $h_{\rm max}$ of 3. The insets in (*a*) and (*c*) are plots showing the location in φ/ψ space of the sample points in each category. For each plotted CDL the color scheme is defined by the bar on the right. The φ, ψ and bond angles are all measured in degrees.

gave good results. The function we minimize to determine $C(h, k)$ is

$$\sum_{i=1}^{n}[\theta_o(\varphi_i, \psi_i) - \theta(\varphi_i, \psi_i|C)]^2 + W \sum_{h,k\neq 0,0}(h^2+k^2)^{1/2}|C(h,k)|^2. \quad (3)$$

The first term is a sum over the data, and its value becomes larger when the data set is larger. The overall weight W is used to ensure that the restraint term can grow to match. This factor is similar to a Lagrange multiplier, and one would like to set it as large as possible, imposing the strongest constraint, without degrading the fit to the data. We found, empirically, that the fit to the data is not terribly sensitive to the exact value of W, but that a larger value could be applied to the very large category *t***X***t*. For the fits that we report, we set W = 1000 when analyzing the *t***X***t* category (~29 000 residues) and W = 250 for all other categories (all $\lesssim$ 5000 residues).

As seen in Fig. 2(*a*), the Fourier summation of the restrained coefficients (red line) has the desired property of fitting the data points similarly to the unrestrained fit (blue line), while showing little variation in the unsampled regions. While this smoothing restraint meets all our requirements, it should not be considered to be 'correct' in the regions with no data since we have no knowledge of what is 'correct'. For instance, there is no reason to suspect that the bump near $\psi = -120°$ is a real feature. We do not claim this to be the best smoothing restraint, but simply one that works adequately.

As illustrated in Fig. 2(*b*), the least-squares filtering can be carried out for any given value of $h_{\rm max}$, providing more detailed fits to the data. The program that we used to calculate an FS-CDL – given a set of data points, the chosen $h_{\rm max}$ and weight – is written in Fortran 2003 and is available as supporting information.

2.5. Complete cross-validation to determine the optimal $h_{\rm max}$

Complete cross-validation (CCV) can be used to choose between several different models of how a data set behaves (Mosteller & Tukey, 1968). With this method one assesses the quality of a particular model by repeatedly determining the optimal parameters using data sets with each individual sample left out, and using these parameters to predict the omitted point. The r.m.s.d. of the discrepancies from all of these tests, which we call $\sigma_{\rm ccv}$, is an unbiased estimate of the amount of variability that is not fitted by this model. The model that fits the most variation in the data (*i.e.* the lowest $\sigma_{\rm ccv}$) is superior.

To find the optimal $h_{\rm max}$, we considered each potential value of $h_{\rm max}$ to be a different model. $\sigma_{\rm ccv}$ was calculated for all values of $h_{\rm max}$ between zero and 13.[1] The typical behavior, for the analysis of main-chain geometry data, was that $\sigma_{\rm ccv}$ decreased rapidly and at some point started increasing (see Figs. 3*a* and 3*c*), revealing the optimal $h_{\rm max}$. When the number of residues was greater than 500 this minimum was broad and the precise value of $h_{\rm max}$ was not critical. For some residue categories with smaller sample sizes the curve did not follow this simple pattern, making the choice of $h_{\rm max}$ unclear. Fortunately for these cases the problem was eliminated by the increase in sample size afforded by category blending (see Section 2.6).

A second problem that can impact small categories is that complete cross-validation will overestimate the value of $h_{\rm max}$ if the data set contains observations that are not truly independent. To maximize the independence of the observations used to build the library, we used the *PISCES* server (Wang & Dunbrack, 2003) option in the PGD to limit the proteins in the data set to those with no more than 25% sequence identity. Despite this filter, however, some homologous proteins with very similar structures remained, and their presence artificially increases the 'optimal' values of $h_{\rm max}$ for some categories.

2.6. Blending categories

To overcome our inability to create meaningful CDLs for categories with few to no residues, we look for sets of categories that can be merged. The compatibility of categories will differ depending on the angle under consideration, requiring a separate analysis for each angle. To avoid imposing an arbitrary size cutoff we look for blending opportunities in all categories.

To decide which categories could benefit by being blended with a larger category, we use the FS-CDL of the larger category to predict values for the residues of the smaller category and calculate the r.m.s.d. between these predictions and the observations. Blending is clearly justified if this r.m.s.d. is equal to or smaller than the $\sigma_{\rm ccv}$ of the CDL of the smaller group, since this proves that the CDL of the larger category has a predictive power at least as great as the smaller category's own library.

For simplicity, we represent both this r.m.s.d. and the CDL $\sigma_{\rm ccv}$ by the number of percentage points decrease relative to the smaller group's SVL $\sigma_{\rm ccv}$. This distillation allows us to relate all three libraries at once. For example, consider a case where the smaller category's CDL $\sigma_{\rm ccv}$ is 90% of its SVL $\sigma_{\rm ccv}$ and the larger category's prediction has an r.m.s.d. of 85% of the same SVL $\sigma_{\rm ccv}$. The 10% decrease of the former is its 'modeled variation', while the modeled variation of the latter is 15%. The five percentage points of additional modeled variation is the 'percent improvement' achieved by the larger category. A percentage improvement of zero would indicate that the larger category predicts the variation of the smaller category as well as the CDL derived from that category itself, and the two categories can be blended. A positive value indicates that the larger category predicts the variation better than the smaller category's CDL, and blending will improve the quality of the CDL.

On the other hand, a negative percent improvement indicates that the larger category is not predicting the smaller category's variation as well as an individualized CDL. How we handle this depends on the size of the smaller category's data set. When it is large we tolerate very little degradation in the CDL and will blend only if the difference is better than −1 percentage points. When the smaller category has few residues we will blend in the presence of greater degradation.

Following this procedure, one is left with a list of categories that have too few residues for our 'percent improvement' tests to function. The categories that make up each of the existing blending groups are examined for patterns in their amino-acid types and peptide conformations. The remaining 'tiny' categories are placed in the blending group whose pattern they best match. This, again, is a subjective choice.

Since residues fall into these 'tiny' categories only rarely, few residues in protein models will be affected by poor choices and, in any case, the CDL applied to these residues will at least contain plausible target values since they will be based on actual protein models.

Both stages of category classification required decisions for which we could not define a rigorous set of rules. Some examples of such decisions are as follows. How many residues are necessary for the 'percent improvement' test to be valid? How much worse should we allow the blending group's prediction be relative to the category's own $\sigma_{\rm ccv}$ and still be acceptable? How distinct must a pattern be and how large must a group be to create a new category? The admittedly subjective choices we made in these cases described above were based on careful consideration of various structural commonalities between groups and an experience-guided intuition. Assessing the wisdom of these inherently subjective choices will only be possible when there are many more models in the wwPDB. To make such a future re-evaluation possible, we have included in the supporting information the *Mathematica* notebooks (Wolfram Research) that we used to assist our blending choices.

2.6.1. Nomenclature for blended groups. We indicate a blended group by enclosing the name of its most sampled category in angle brackets (*i.e.* '⟨ ' and '⟩'). For example, the CDL for the NCaC angle that results from blending the residues of categories *t***X***t*, *c***X***t* and *t***P***c*P is named NCaC/⟨*t***X***t*⟩.

2.7. Utilizing the CDL-v2.0

In the supporting information, we have supplied an implementation of our library written in *Mathematica* script (Wolfram Research). The Fourier coefficients for each angle/category combination are stored in the file. The target value is generated by identifying the category of the residue (based on

[1] Going beyond 13 greatly increased the computation time and did not prove necessary in this application.

Table 2
Number of residues found in each residue category.

The number is shown in bold for those categories where searches were limited to structures determined at 1 Å resolution and better.

Category	N	Category	N	Category	N	Category	N
*t***X***t*	**28944**	*t***X***c*	117	*c***X***t*	132	*c***X***c*	0
*t***IV***t*	**5042**	*t***IV***c*	6	*c***IV***t*	14	*c***IV***c*	0
*t***G***t*	**3523**	*t***G***c*	39	*c***G***t*	19	*c***G***c*	0
*t***P***t*	**1743**	*t***P***c*	14	*c***P***t*	1140	*c***P***c*	0
*t***X***t*P	**1382**	*t***X***c*P	859	*c***X***t*P	7	*c***X***c*P	1
*t***IV***t*P	3026	*t***IV***c*P	97	*c***IV***t*P	0	*c***IV***c*P	0
*t***G***t*P	1267	*t***G***c*P	148	*c***G***t*P	2	*c***G***c*P	1
*t***P***t*P	711	*t***P***c*P	68	*c***P***t*P	63	*c***P***c*P	3

its amino-acid type, whether or not it is followed by a proline and the conformation of the leading and trailing peptide bonds) and evaluating the Fourier series at the φ/ψ location. The standard deviation is the $\sigma_{\rm ccv}$ of that category. The complexity of the blending mapping is hidden from the application by duplicating the blended coefficients for each component individual group.

3. Results and discussion

3.1. Selecting residues from the wwPDB

Drawing on a representative subset of high-resolution structures of proteins from the Worldwide Protein Data Bank (wwPDB Consortium, 2018), we used the Protein Geometry Database (PGD) web service (Berkholz, Krenesky *et al.*, 2010) to create our data set. At the time, the PGD was loaded with data taken from the wwPDB on April 4, 2016. An inventory of all of the protein chains then in the PGD is listed in the supporting information file `20160404-selection.txt`. For each of the 32 categories, we performed an initial search allowing only models based on X-ray diffraction data of 1.0 Å resolution or higher. Only five categories contained more than 500 residues (Table 2), and for the others we performed a second search using the more relaxed filter of 1.5 Å resolution, which typically yielded about nine times more residues. This relaxation, for example, increased the number of hits for *t***IV***t*P from 260 to 3026 and for *t***G***t*P from 138 to 1267. Even with the large increases that occurred for some small categories, the large categories searched at 1 Å resolution, *t***X***t*, *t***IV***t*, *t***G***t*, *t***P***t* and *t***X***t*P, still dominate the total number of residues in our data set.

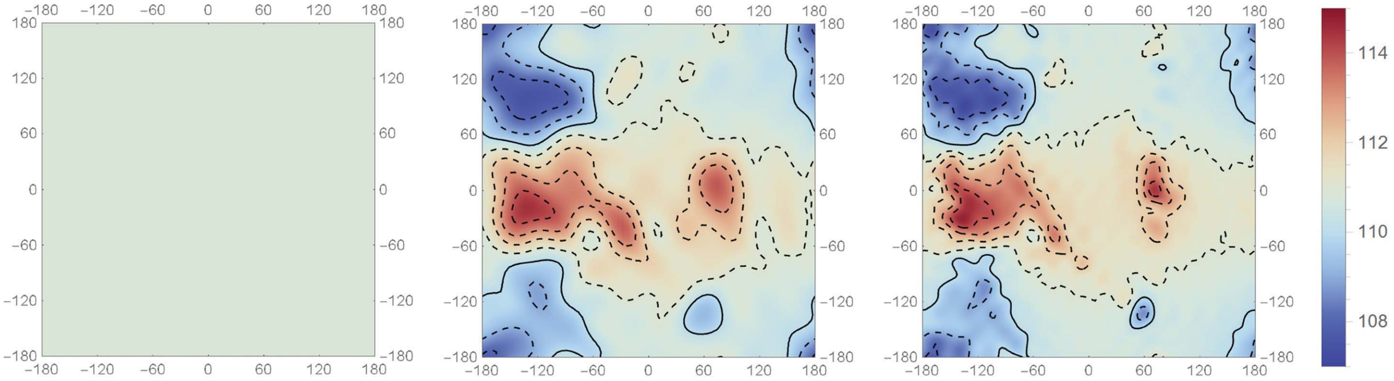
We felt this relaxation was acceptable because the analysis of refinement tests in both Tronrud & Karplus (2011) and Moriarty *et al.* (2016) showed that refinement restraints do not dominate model bond-angle values at resolutions of 1.5 Å resolution and better, although they do still influence the model. Even using the 1.5 Å resolution limit, we found only 2730 residues bounded on one side and/or the other by *cis*-peptide bonds out of a total of 48 368.

Numbers of residues in each category ranged from 28 944 to zero (Table 2), and even using the 1.5 Å resolution limit many of the 32 categories have few observations: 15 have less than 20 and 22 have less than 500. Six categories have zero observations, and these include all four categories with two *cis*-peptides in a row with the second *cis*-peptide not followed by a proline (Table 2). For these categories with zero observations, even though no residues were found in these categories and we have only indirect information to base our library on, we still have chosen to include them in the CDL because completeness is one of our primary goals in the library design. As noted in Section 1, these very sparsely sampled categories were handled by 'blending', which will be described shortly.

The results of our searches have been deposited as supporting information in both *Mathematica* (Wolfram Research) and CSV text formats.

3.2. Fourier series representation of the CDL

To represent the conformational dependence of a target value, we have chosen to use a two-dimensional Fourier



$h_{\rm max} = 0 \quad N = 1$
$\sigma_{\rm ccv} = 2.141°$

$h_{\rm max} = 6 \quad N = 169$
$\sigma_{\rm ccv} = 1.385°$

$h_{\rm max} = 13 \quad N = 729$
$\sigma_{\rm ccv} = 1.386°$

Figure 4
FS-CDLs for the *t***X***t* category for three $h_{\rm max}$ values. These plots are graphical representations of the CDLs for the NCaC angle of the *t***X***t* category calculated with a range of values for $h_{\rm max}$. Of these three, the CDL with $h_{\rm max}$ = 6 has the lowest $\sigma_{\rm ccv}$. Of all values for $h_{\rm max}$, 9 is optimal, with a $\sigma_{\rm ccv}$ of 1.384° and a modeled variation of 35%. For each plotted CDL the color scheme is defined by the bar on the right. The φ, ψ and bond angles are all measured in degrees. A solid contour line is drawn at multiples of 5°, with dotted contours every 1°.

synthesis (FS). Since φ/ψ space 'wraps around' just as electron density in a crystal, the values of the reciprocal-space indices (h and k, corresponding to φ and ψ, respectively) must likewise be integral. The level of detail of each FS-CDL can be controlled by changing the limit on the number of high-frequency Fourier coefficients. To simplify matters, we set a common upper limit on both indices and call it $h_{\rm max}$. An FS-CDL with $h_{\rm max} = 0$ will have no variation and a value equal to the average value of the angles of the residues; this is a single-value library (SVL). Three example FS-CDL fits for a particular data set with $h_{\rm max} = 0$, 6 and 13 are shown in Fig. 4. Each additional Fourier coefficient requires the addition of two more independent parameters to the CDL, with the total number of parameters equaling $(2h_{\rm max} + 1)^2$. One would, of course, prefer to limit the number of parameters to the lowest number which still model the data.

The choice of sine waves as a basis set is not only familiar but is tied to the underlying physics of bond-angle variation with conformation. Bond angles are deformed away from the values that they would adopt in small molecules by nonbonded interactions with surrounding atoms, and these contacts appear and disappear as conformational angles turn. For a torsion angle describing the orientation of two sp^3-hybridized atoms, one would expect the strongest contacts to vary with a periodicity of three, resulting in a strong $h = 3$ Fourier coefficient. The relationships between protein main-chain atoms are more complex, resulting in a more complex Fourier analysis, but still we expect, and find, that the CDL can be represented principally by low-resolution Fourier coefficients.

This raises the question of the value that should be chosen for $h_{\rm max}$. As expected, the r.m.s.d. between any given FS and the raw data used to generate its Fourier coefficients continues to become lower as $h_{\rm max}$ increases, making it useless for defining $h_{\rm max}$. This is a classic problem in modeling and is successfully solved by cross-validation (Mosteller & Tukey, 1968). Two common approaches to cross-validation are to leave out a discrete subset of data, say 5–10%, as has been used for many years in X-ray structure refinement (Brünger, 1997) and far longer in many other fields (see, for example, Clark, 1975), or to carry out complete cross-validation (CCV), also known as 'leave-one-out' cross-validation (Mosteller & Tukey, 1968).

CCV is the ultimate extension of the cross-validation principle. Each observation is left out of the analysis in turn, and the model is applied to the remainder to predict that observation. The r.m.s.d. of the resulting discrepancies is calculated and is here called $\sigma_{\rm ccv}$. This is an unbiased estimate based on the largest possible amount of data while at the same time treating all data points equally. The $\sigma_{\rm ccv}$ is a good estimate of the standard uncertainty of the prediction of the model.

CCV can be computationally intensive since it requires a full analysis of the data for each data point. While for large data sets simple cross-validation performs adequately, it is unreliable for small data sets. CCV gives superior results for data sets of all sizes, but is especially useful when assessing models based on small data sets because the amount of computation is less and the value of using as much data as possible is greater.

Table 3
Modeled variation of the FS-CDLs for the NCaC angle in *tt* sequence categories relative to their SVLs.

The $\sigma_{\rm ccv}$ of the SVL and the best CDL are compared for the NCaC angle of each *trans–trans* category. The values for all backbone angles of all categories are listed in Supplementary Tables S4–S6.

		SVL	FS-CDL		
Category	No. of observations	$\sigma_{\rm ccv}$ (°)	$h_{\rm max}$	$\sigma_{\rm ccv}$ (°)	Modeled variation (%)
$t\mathbf{X}t$	28944	2.14	9	1.38	35
$t\mathbf{IV}t$	5042	2.10	8	1.36	35
$t\mathbf{G}t$	3523	2.28	4	1.52	33
$t\mathbf{P}t$	1743	1.99	7	1.48	26
$t\mathbf{X}t\mathrm{P}$	1382	2.18	7	1.57	28
$t\mathbf{IV}t\mathrm{P}$	3026	2.42	8	1.87	23
$t\mathbf{G}t\mathrm{P}$	1267	2.11	5	2.08	14
$t\mathbf{P}t\mathrm{P}$	711	2.40	5	2.05	14

The $\sigma_{\rm ccv}$ is a powerful tool for comparing the utility of various models applied to the same data set. The best model for predicting a new observation will be the one with the lowest $\sigma_{\rm ccv}$. One use we made of $\sigma_{\rm ccv}$ is to determine the value of $h_{\rm max}$ which best models the variation in the data set. When one calculates $\sigma_{\rm ccv}$ for a range of values of $h_{\rm max}$, generally one finds that $\sigma_{\rm ccv}$ will decrease as $h_{\rm max}$ increases, but at some point $\sigma_{\rm ccv}$ will begin to rise (see Fig. 3). This point of minimal standard uncertainty is the optimal value for $h_{\rm max}$.

In addition, we used $\sigma_{\rm ccv}$ to assess the utility of any particular CDL relative to an SVL. When the $\sigma_{\rm ccv}$ for a CDL is lower than that of the SVL, we conclude that the modeled variation in that CDL is meaningful. We find it convenient to quantify this as the difference between the SVL $\sigma_{\rm ccv}$ and that of the CDL, expressed as a percentage of the SVL $\sigma_{\rm ccv}$. We call this metric the 'modeled variation' of the CDL. For instance, the SVL for NCaC/$c\mathbf{P}t$ has a $\sigma_{\rm ccv}$ of 2.45° and the optimal FS-CDL has a $\sigma_{\rm ccv}$ of 1.86°, meaning that the modeled variation is $100(2.45 - 1.86)/2.45 = 24\%$.

As is typical for the FS-CDLs developed here, the vast majority of the power of the CDL relative to the SVL is achieved by the first few Fourier synthesis terms (see, for example, Figs. 3*a* and 3*c*). Also, typically, the smaller the data set, the smaller the optimal $h_{\rm max}$ and the smaller the modeled variation (Table 3 and Supplementary Tables S4, S5 and S6). In some cases the optimal $h_{\rm max}$ is zero, meaning that the SVL is best and the existing data do not justify using a CDL.

Table 3 shows the $\sigma_{\rm ccv}$ values for the SVL and FS-CDL libraries for the NCaC angles of the well represented categories with *trans–trans* conformations. For the larger categories, the CDLs are able to account for about 20–35% of the observed variability. CDLs for categories containing *cis*-peptide bonds (other than $c\mathbf{P}t$ and $t\mathbf{X}c\mathrm{P}$) have few residues in our data set and perform less well, but usually are superior to the SVL (Supplementary Tables S5 and S6).

Since the $\sigma_{\rm ccv}$ of a CDL is an estimate of the standard uncertainty of that restraint, it is the proper value to use when

Table 4
Comparison of the performance of discrete CDLs and FS-CDLs based on the same data.

This table compares the performance (as measured by $\sigma_{\rm ccv}$ in units of degrees) of three styles of libraries based on the same 2016 data set for a selection of main-chain angles and residue categories. The first is an SVL whose target values are simply the mean of the angles in the data set. The second is a discrete CDL of the same type as CDL-v1.2 and the third is an FS-CDL without blending so that it is directly comparable. The best of the three is rendered in bold text (with one tie). Percent improvement is the number of percentage points difference between the modeled variation of the FS-CDL and the discrete CDL.

		CNCa				NCaC			
Category	*N*	SVL (°)	Discrete CDL (°)	FS-CDL (°)	Percent improvement	SVL (°)	Discrete CDL (°)	FS-CDL (°)	Percent improvement
*t*X*t*	28944	1.77	1.49	**1.46**	2	2.14	1.41	**1.38**	1
*t*G*t*	3523	1.65	1.45	**1.43**	1	2.28	1.61	**1.52**	4
*c*P*t*	1140	1.83	**1.69**	**1.69**	0	2.45	1.92	**1.86**	2
*c*X*t*	132	1.73	1.65	**1.62**	2	3.05	2.95	**2.25**	23

Table 5
Blending groups for the NCaC angle.

Each row describes an individual residue category (ordered from the largest sample size to the smallest) with its number of residues and the modeled variation listed. Each principal column describes a blending category with the total number of residues it contains, and its $h_{\rm max}$, $\sigma_{\rm ccv}$ and percent modeled variation. In bold is the percentage improvement that the blended CDL provides over the individual CDL for the categories it contains. (A residue category can be in only one blending category.) Horizontal lines separate the categories with more than 500 residues and those with less than ten. We consider the former to have very good individual CDLs, while the latter have too few residues to draw any conclusions. Those categories in the middle have CDLs that may or may not be reliable, but at least have sufficient residues to make some judgments about blending.

			Blended groups									
	N	%	⟨*t*X*t*⟩	⟨*t*IV*t*⟩	⟨*t*G*t*⟩	⟨*t*IV*t*P⟩	⟨*t*P*t*⟩	⟨*t*X*t*P⟩	⟨*t*G*t*P⟩	⟨*c*P*t*⟩	⟨*t*X*c*P⟩	⟨*t*P*t*P⟩
N			29144	5056	3542	3129	1743	1506	1471	1140	860	777
$h_{\rm max}$			9	8	4	8	7	7	5	6	8	5
CDL $\sigma_{\rm ccv}$ (°)			1.39	1.36	1.53	1.86	1.48	1.63	2.11	1.86	1.98	2.02
Modeled variation (%)			35	35	33	23	26	27	13	24	11	14
*t*X*t*	28944	35	**0**									
*t*IV*t*	5042	35		**0**								
*t*G*t*	3523	33			**0**							
*t*IV*t*P	3026	23				**0**						
*t*P*t*	1743	26					**0**					
*t*X*t*P	1382	28						**0**				
*t*G*t*P	1267	14							**+1**			
*c*P*t*	1140	24								**0**		
*t*X*c*P	859	11									**0**	
*t*P*t*P	711	14										**0**
*t*G*c*P	148	0							**+1**			
*c*X*t*	132	26	**−2**									
*t*X*c*	117	6						**+7**				
*t*IV*c*P	97	10				**+5**						
*t*P*c*P	68	0	**+2**									
*c*P*t*P	63	1										**+3**
*t*G*c*	39	1							**+8**			
*c*G*t*	19	13			**+6**							
*t*P*c*	14	0							**+5**			
*c*IV*t*	14	7		**+33**								
*c*X*t*P	7	?						**?**				
*t*IV*c*	6	?				**?**						
*c*P*c*P	3	?										**?**
*c*G*t*P	2	?							**?**			
*c*X*c*P	1	?									**?**	
*c*G*c*P	1	?							**?**			
*c*IV*t*P	0	?				**?**						
*c*X*c*	0	?						**?**				
*c*IV*c*	0	?				**?**						
*c*IV*c*P	0	?				**?**						
*c*G*c*	0	?							**?**			
*c*P*c*	0	?							**?**			

weighting restraints relative to each other. A restraint based on a CDL with a small $\sigma_{\rm ccv}$ should be given more importance than one with a large $\sigma_{\rm ccv}$. In most cases the restraints will be weighted by $1/\sigma^2_{\rm ccv}$.

3.3. Comparing discrete and Fourier synthesis CDLs

While we would like to directly compare the performance of CDL-v1.2 and CDL-v2.0, the actual list of residues used to construct CDL-v1.2 was not preserved and it is not possible to calculate $\sigma_{\rm ccv}$ for its various categories. We can, however, perform a more stringent comparison with a library of identical construction based on our 2016 data set. Table 4 shows the results of CCV assessments for a conventional SVL, a discrete CDL library (like CDL-v1.2) and the FS-CDLs described here, all derived from our 2016 data set. (For all categories and angles, see Supplementary Tables S1–S6.)

For each CDL, whether of the discrete or Fourier synthesis variety, one can calculate the percentage of the variation of the angle not modeled by the SVL but modeled by the CDL (*i.e.* modeled variation). We call the number of percentage points by which the modeled variation of an alternative CDL is smaller than the modeled variation of the original CDL the 'percent improvement'.

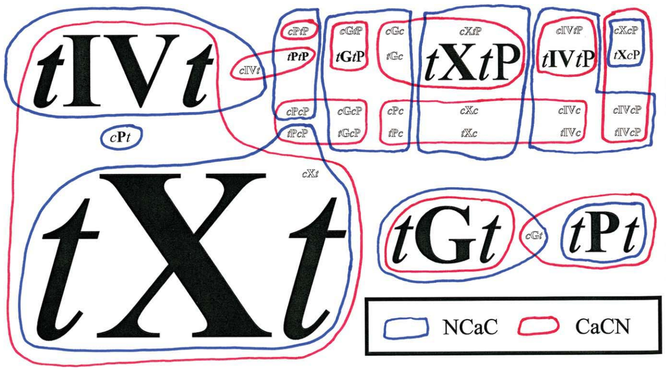


Figure 5
Blending groups vary for different angles. Shown are the cluster patterns for the NCaC angle and the CaCN angle, using a common placement of categories that was crafted specifically to allow these two sets of blendings to be displayed most clearly. The text size of the category names is proportional to the sample size, with the exception that categories with few residues (<150) are printed at a fixed size with hollow letters.

For the heavily sampled categories (*t***X***t*, *t***G***t* and *c***P***t*) the percent improvement is near zero. In contrast, the FS-CDL for the NCaC angle in *c***X***t* performs 20 percentage points better than the discrete CDL (increasing from 3% modeled variation to 23%), an improvement of nearly a factor of eight. In this category the 132 residues are scattered over the φ/ψ plot and rarely reach a density high enough to trigger a specialized block in the discrete CDL. These residues are able to support an $h_{\max}$ of 3 in the FS-CDL, which allows the allows the target value for this angle to be ∼108° near $(\varphi, \psi) = (-160^\circ, -160^\circ)$ and ∼113° near $(-110^\circ, -10^\circ)$, with a smooth transition between.

The $\sigma_{\rm ccv}$ values for the FS-CDL are equal to or lower than the $\sigma_{\rm ccv}$ values for the discrete CDLs based on the same data set (Table 4 and Supplementary Tables S4–S6), demonstrating the overall superiority of the new CDL form.

3.4. Blending groups enhances the modeling of categories with few to no residues

In general, categories that are represented by more than 500 residues produce quality FS-CDLs for all of the backbone angles examined in this study. However, the remaining 22 categories had lower quality FS-CDLs, and of course for those categories with no observations no FS-CDLs can be made. Even though the ten of 32 categories with over 500 residues represent 99.8% of all residues in the wwPDB, our goal is to create a complete backbone CDL.

Fortunately, we found that the angles of rarely occurring categories could be predicted well by the FS-CDLs from one of the highly populated categories. This means that it is possible to create FS-CDLs based on combined or blended data sets from multiple categories. Such an analysis would generally be justified only when the blended FS-CDL predicts the residues in the individual categories roughly as well or better than their individual FS-CDLs. Each angle of each category must be treated as a special case since the characteristics of a category (for example the conformation of the two peptide bonds) will affect each angle differently.

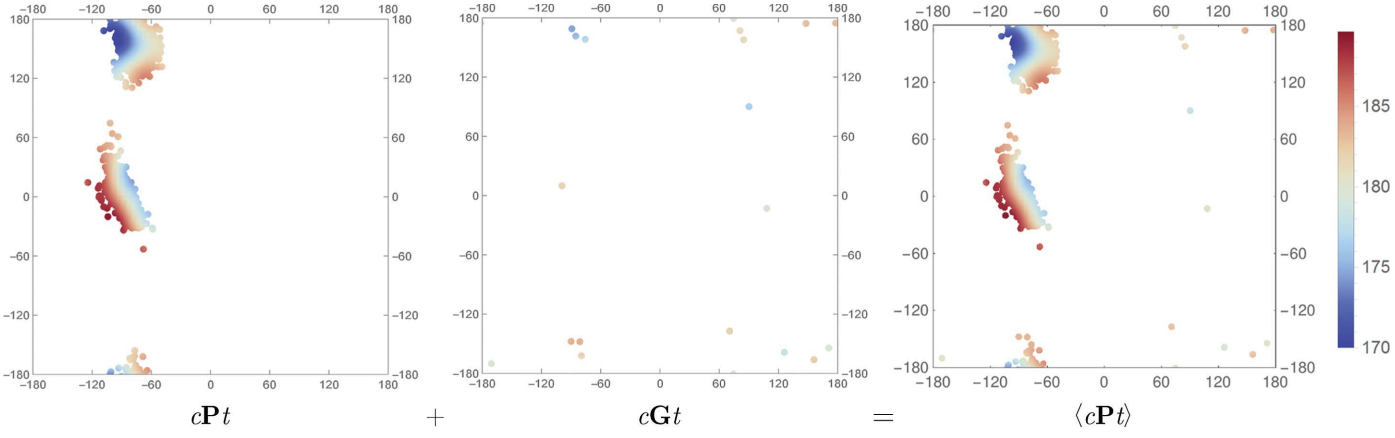


Figure 6
The blending of *c***P***t* and *c***G***t* categories to create $\omega/\langle c\mathbf{P}t\rangle$. These plots show the individual CDLs, with the target values colored only in the vicinity of the observed residues used to construct the CDL. While all of the residues of *c***P***t* lie in a narrow line near $\varphi = -90^\circ$, only seven of the 19 *c***G***t* residues do the same. The two individual CDLs are very similar where they overlap, and this compatibility makes them excellent candidates for blending.

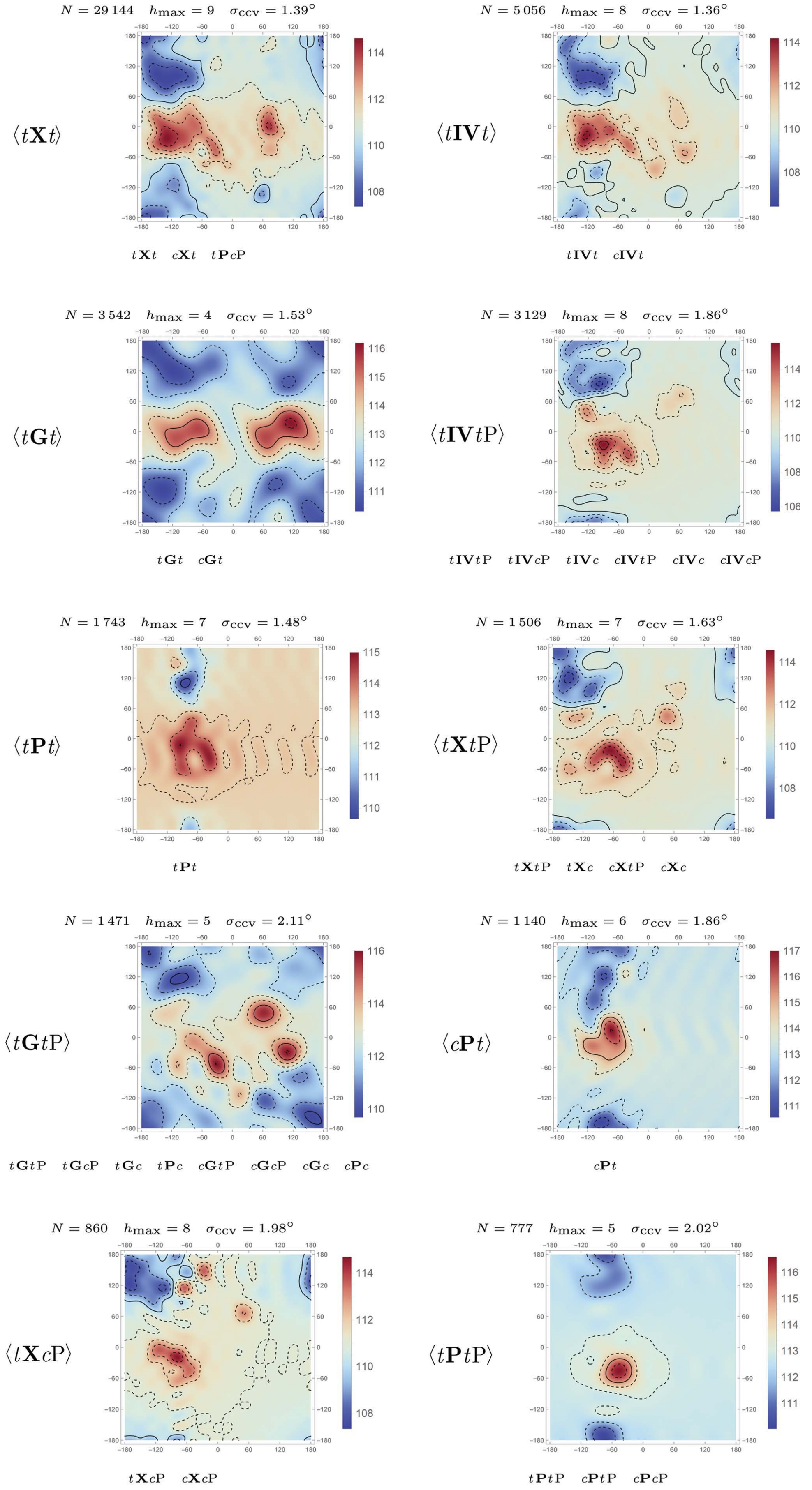


Figure 7
Ten blended groups for the NCaC angle. Each panel describes a CDL for one of the blended groups. The name of the group is on the left. Across the top is listed the number of residues used to create the CDL, the $h_{\max}$ used in the calculation and $\sigma_{\rm ccv}$. A list of the individual categories blended to form this group is displayed at the bottom. For each plotted CDL the color scheme is defined by the bar on the right. The φ, ψ and bond angles are all measured in degrees. A solid contour line is drawn at multiples of 5°, with dotted contours every 1°.

We did not try every possible combination of category blending, but instead used a shortcut (see Section 2.6 for details). We used the larger, and not already blended, categories to predict each angle and each category. If one or more of these categories predicted the angles at least as well as the individual CDL we chose to blend the target category with the best performing of the larger categories. We had less faith in inferences based on our statistical tests as the number of residues in the target category became smaller. In recognition of this problem, we increasingly biased our decisions towards blending even in the presence of some degradation in the performance of the blended CDL. Ultimately, we reached categories with so few (or no) residues that we did not trust any measure of prediction accuracy.

The members of this final set of categories were assigned to one of the established blending groups by a rather *ad hoc* process of examining the structural similarity in each blending group and placing the leftover category in the group that we considered to be the most similar. For example, the established blending groups suggested that the *cis*/*trans* status of the preceding peptide bond was of less import to the angles in the trailing peptide bond. Since there were only three residues in the category *c***P***c*P, we used this observation to justify the assumption that the FS-CDLs from *t***P***c*P for the CaCO, CaCN, OCN and ω angles will also predict those in the *c***P***c*P category.

The smallest of the blended groups are ω/⟨*t***P***c*P⟩ with $N = 71$ and CaCN/⟨*c***P***t*P⟩ with N = 77. While these are very small sample sizes with which to uncover conformational dependence, all the residues for these categories occur in a small region of φ/ψ space, resulting in a high population density.

Table 5 shows the blending groups and modeled variation for the NCaC angle. For this angle, we found that the 32 individual categories could be reduced to ten blended groups, with the smallest being based on 777 residues (NCaC/⟨*t***P***t*P⟩). The vast majority of the smaller categories are predicted more successfully by one of these blending groups than by their individual CDL. *c***IV***t* is particularly well served, with an improvement of 33 percentage points (rising from 7% better than its individual SVL to 40% better). Category *c***X***t* is predicted slightly less well by the blending group we chose. Since its individual CDL modeled 26% of the variability of its SVL, this loss of two percentage points still leaves *c***X***t* ranking among the best-modeled categories.

While none of the NCaC categories with more than 500 residues are blended together, some are quite similar. The FS-CDL from *t***X***t* can predict the NCaC angles for *t***IV***t* residues 28% better than the *t***IV***t* SVL. While this is six percentage points less than the FS-CDL of *t***IV***t* itself, it is a respectable success. For the CNCa, CaCO, CaCN, OCN and ω angles, the FS-CDL from *t***X***t* predicted the *t***IV***t* examples only one percentage point worse than *t***IV***t* itself, and our protocol led us to blend these categories (Supplementary Tables S7, S11, S12, S13 and S15).

To demonstrate the kinds of differences that occur in blending groups, the groups for the NCaC and CaCN angles are compared in Fig. 5. Easy to notice in this figure, and remarkable, is that none of the blending categories are completely the same for these two angles. As one theme, the side chain plays a larger role for the NCaC angle than for the CaCN angle. For instance, for the CaCN angle, but not the NCaC angle, the *t***IV***t* category is well predicted by the *t***X***t* CDL (see the previous paragraph). In fact, these two categories were never blended when the C$^{\alpha}$ atom is at the vertex of the angle, but are otherwise always blended. This result is not surprising when one considers the β-branched nature of **IV** amino acids; however, structural considerations were not used in making these blending choices.

Similarly, for the NCaC angle many of the glycine-containing categories blend together well, but for the CaCN angle the conformation of the following peptide bond and whether or not the next residue is proline becomes more important, leading to a fracturing of the glycine-containing blending groups.

Surprisingly, proline residues can often be blended successfully with glycine residues for both of these angles. Even though glycine residues often exist in conformations that are impossible for proline residues, when glycine does adopt a compatible conformation, certain bond angles are not much different from those of proline.

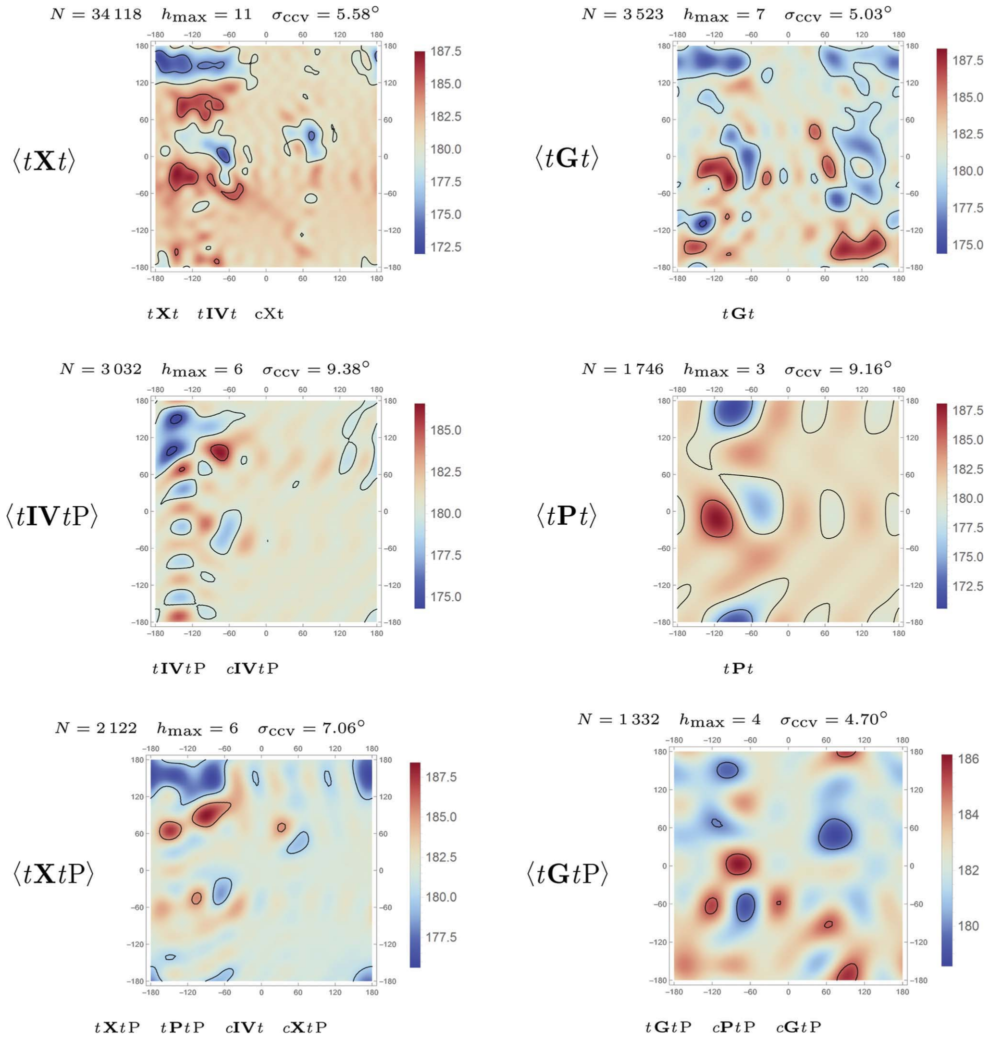


Figure 8
The first six of 12 blended CDLs for the ω angle. The data representation is described in the caption to Fig. 7 with the exception that no dashed contours are drawn. Fig. 9 shows the other 6 blended CDLs for this angle.

Fig. 6 demonstrates the blending of *c***P***t* and *c***G***t* to create the blended group $\omega/\langle c\mathbf{P}t\rangle$. The $c\mathbf{P}t$ CDL predicts the ω angles observed for the $c\mathbf{G}t$ category with an r.m.s.d. of 7.42° when the CDL based directly on the $c\mathbf{G}t$ hits has an unbiased prediction quality of $\sigma_{\rm ccv}$ = 8.46°, demonstrating that these categories can be blended. This particular blending is successful because the variation of the ω angle for the seven $c\mathbf{G}t$ residues that fall in the same region as the proline residues have similar variation. The other 12 residues are widely dispersed in φ/ψ space and create a poor CDL by themselves, but the $c\mathbf{P}t$ CDL does well with its prediction based primarily on the average of their ω angles.

Clearly, a proline residue mistakenly built in a conformation with a positive φ angle would be unlikely to have an ω angle similar to a glycine residue with the same φ/ψ angles. We do not consider this to be a problem because our goal is to provide the best target value for all residues that are likely to exist, and we are not concerned about the accuracy of our target values for 'impossible' residues. We also note that this particular blending is actually not conceptually surprising because the ω angle being modeled occurs after the residue, where the unique features of proline residues do not have substantial impact.

As noted above, the blending protocol was carried out independently for all seven backbone angles, the ζ dihedral angle and the ω torsion angle, and the results for each angle (equivalent to Table 5) are given in Supplementary Tables S7 through S15.

3.5. The new backbone CDL-v2.0

Using this process, we have created a set of FS-CDLs which cover all 20 amino acids and all *cis*/*trans* possibilities for the peptide on either side, *i.e.* all 32 categories. As examples, plots for each CDL for the blended groups of the NCaC and ω angles are shown in Figs. 7, 8 and 9. Plots for the complete set of backbone angles are shown in Supplementary Figs. S1–S14. These plots are repeated in Supplementary Figs. S15–S28, where the CDL is displayed in a 'paintball plot' representation like those shown in Fig. 6. Recalling that blended groups were independently constructed for each angle, we note that CaCO and ω have the most groups with 12 each, while NCaCb has the fewest with six (in part because glycine residues do not contain this angle).

The performance of each blended group is measured by the amount by which the $\sigma_{\rm ccv}$ of the CDL decreases compared with that of the SVL, which we express as a percentage of the SVL value. If the CDL encompasses all of the variability of the angle the 'modeled variation' will be 100%, but none of the CDLs based on φ/ψ alone reach this ideal. Fig. 10 displays the modeled variation for the blending groups in each angle. The NCaC angle is handled the best by the heavily sampled group NCaC/$\langle t\mathbf{X}t\rangle$, modeling 35% of its observed variability. CNCa, CaCO and CaCN all have their $\langle t\mathbf{X}t\rangle$ groups coming in around 20%. The angles which involve the C^{β} atom are generally modeled less well, we suspect owing to the bundling of 16 residue types in $\mathbf{X}$. The ω angle is also modeled weakly. The angle with the poorest performing CDL is OCN. While its CDL does not improve much upon the performance of an SVL, this angle

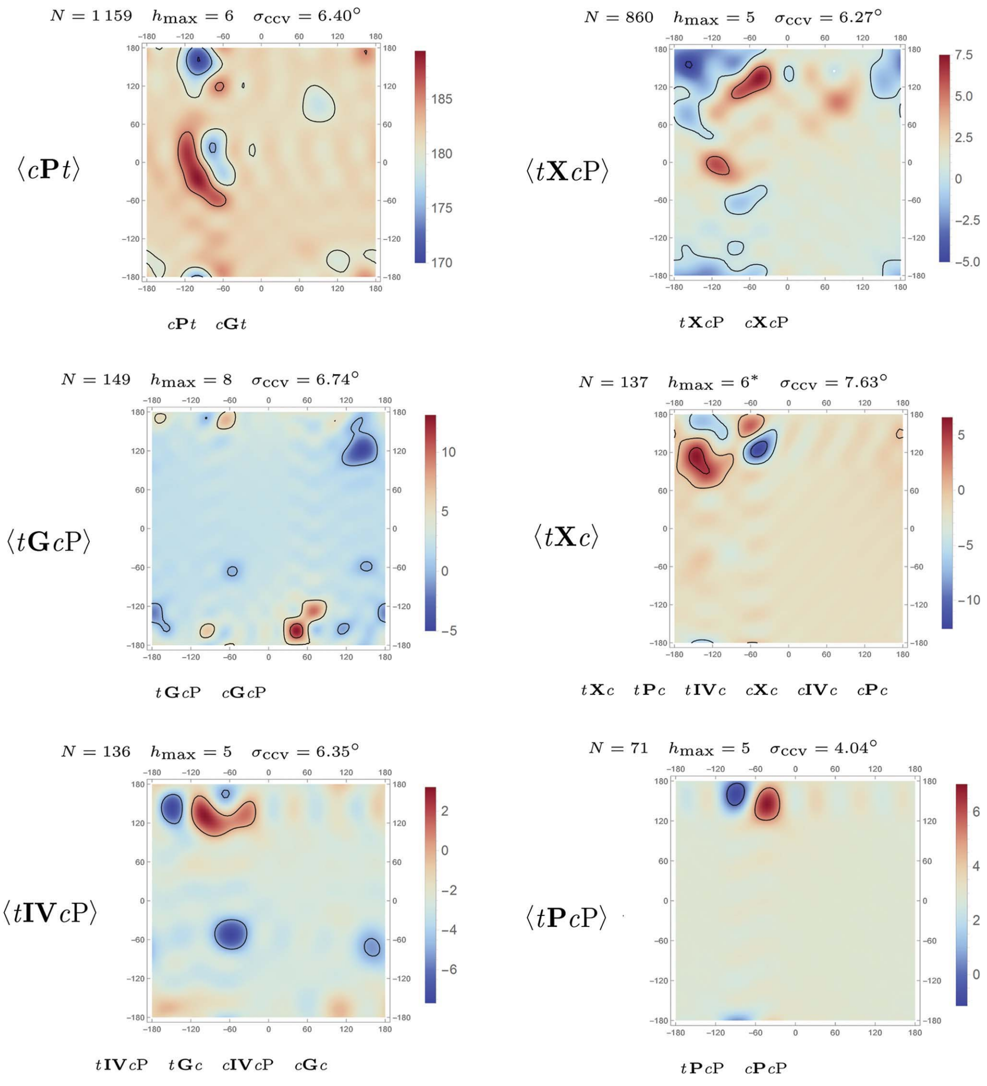


Figure 9
The last six of 12 blended CDLs for the ω angle. The data representation is described in the caption to Fig. 7 with the exception that no dashed contours are drawn.

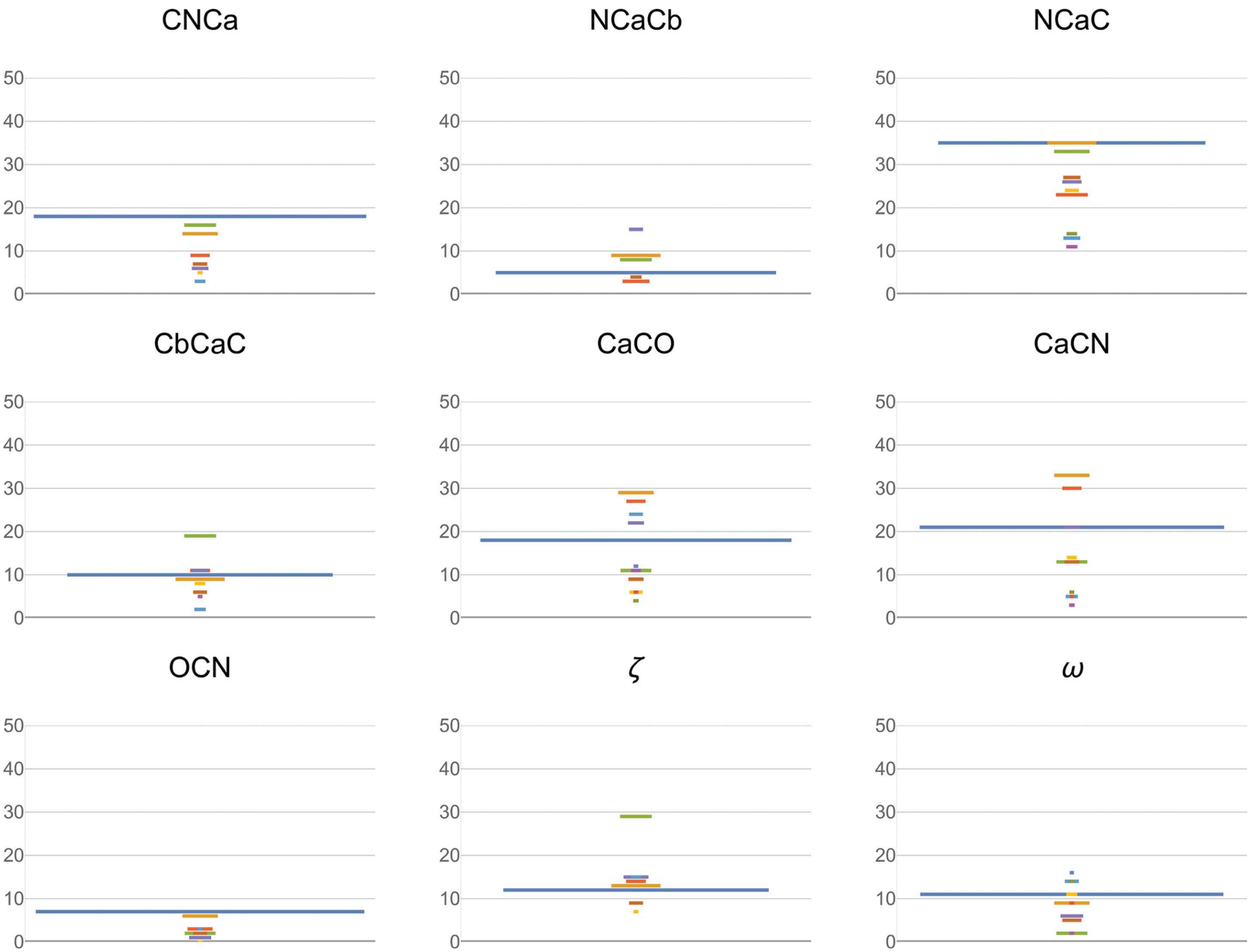


Figure 10
Percentage of modeled variation for every blended group. The amount of modeled variation by each blending group, grouped by angle, is plotted. The lengths of the lines are proportional to the number of residues in that group. As a consequence, the summed length of all lines for each angle is proportional to the total number of residues in the data set containing that angle. The colors of the lines are varied to allow the visualization of overlapping lines.

varies little and has very low modeled variations anyway (see Supplementary Fig. S11). Blending groups based on fewer residues than the $t\mathbf{X}t$ group usually perform less well, but they also will be used less often in applications, resulting in fewer consequences.

Some of the CDLs are quite complex, such as $\omega/\langle t\mathbf{X}t\rangle$ (Fig. 8) which has $h_{\max} = 11$ and a range of 15°. As has been discussed before (see, for example, Karplus, 1996; Jiang *et al.*, 1997; Esposito *et al.*, 2005; Berkholz *et al.*, 2009; Brereton & Karplus, 2015), the major trends seen in CDLs are consistent with quantum- and molecular-mechanics calculations and make sense in terms of simple steric and electrostatic considerations. While we suspect there are important things to learn from particular details of such complex CDLs, such considerations are beyond the scope of this paper. An important advantage of an empirical library such as this one is that it is useful for building more accurate structures even without understanding the origins of all of its features. The fact that the most accurately defined residues in the Protein Data Bank behave this way is sufficient justification.

3.6. Comparison of the performance of the Engh and Huber 2001 SVL and CDL-v2.0

The most widely used protein geometry library is that published by Engh & Huber (1991). This SVL was updated in 2001 (Engh & Huber, 2001) by making generally small changes to most target values and introducing specialized target values for the CNCa angle of *cis*- and *trans*-proline (with no consideration of *cis*-peptide bonds in other residue types nor of any residues followed by a *cis*-peptide bond).

Table 6 compares, for seven residue categories, the performance of the Engh and Huber 2001 (E&H 2001) SVL with our CDL-v2.0 for the CNCa and CaCN angles, as representative

Table 6
Comparison of the performance of the E&H 2001 values, the SVL values based on the current data set and CDL-v2.0.

Comparisons are shown between the quality of the predictions of the E&H 2001 library and CDL-v2.0. Also shown as 'SVL' are the SVL targets that result when the CDLs in CDL-v2.0 are evaluated with $h_{\max} = 0$ (*i.e.* the global averages). The target values for SVLs are enclosed in square brackets, with their standard deviations bounded by parentheses and written with units of the least significant digit of the target value. For the SVL and CDL comparisons target values were taken from the indicated blended groups. The categories above the horizontal line are for groups where the E&H library provides specific targets. For the rest, the E&H category used by most programs was chosen. Only comparisons for the CNCa and CaCN angles are shown since these angles are strongly influenced by the *cis*/*trans* status of their nearest peptide bond.

	CNCa			CaCN		
	E&H 2001	SVL	CDL-v2.0	E&H 2001	SVL	CDL-v2.0
	R.m.s.d. (°)	$\sigma_{\rm ccv}$ (°)	$\sigma_{\rm ccv}$ (°)	R.m.s.d. (°)	$\sigma_{\rm ccv}$ (°)	$\sigma_{\rm ccv}$ (°)
$t\mathbf{X}t$	'Peptide' 1.78 [121.7 (25)]	$\langle t\mathbf{X}t\rangle$ 1.77 [121.6 (15)]	$\langle t\mathbf{X}t\rangle$ 1.46	'Peptide' 1.62 [117.2 (22)]	$\langle t\mathbf{X}t\rangle$ 1.55 [116.7 (16)]	$\langle t\mathbf{X}t\rangle$ 1.21
$t\mathbf{G}t$	'Glycine' 1.89 [122.3 (21)]	$\langle t\mathbf{G}t\rangle$ 1.65 [121.4 (16)]	$\langle t\mathbf{G}t\rangle$ 1.43	'Glycine' 1.89 [116.2 (20)]	$\langle t\mathbf{G}t\rangle$ 1.83 [116.7 (18)]	$\langle t\mathbf{G}t\rangle$ 1.22
$t\mathbf{P}t$	't-Proline' 1.39 [119.3 (15)]	$\langle t\mathbf{P}t\rangle$ 1.35 [119.6 (14)]	$\langle t\mathbf{P}t\rangle$ 1.23	'Proline' 1.92 [117.1 (28)]	$\langle t\mathbf{P}t\rangle$ 1.86 [116.6 (19)]	$\langle t\mathbf{P}t\rangle$ 1.30
$c\mathbf{P}t$	'c-Proline' 1.84 [127.0 (24)]	$\langle c\mathbf{P}\mathrm{t}t\rangle$ 1.82 [127.3 (18)]	$\langle c\mathbf{P}t\rangle$ 1.69	'Proline' 1.68 [117.1 (28)]	$\langle\langle t\mathbf{X}t\rangle$ 1.61 [116.7 (16)]	$\langle t\mathbf{X}t\rangle$ 1.16
$c\mathbf{X}t$	'Peptide' 5.43 [121.7 (25)]	$\langle c\mathbf{X}t\rangle$ 1.72 [126.8 (19)]	$\langle c\mathbf{X}t\rangle$ 1.61	'Peptide' 1.50 [117.2 (22)]	$\langle t\mathbf{X}t\rangle$ 1.34 [116.7 (16)]	$\langle t\mathbf{X}t\rangle$ 1.24
$t\mathbf{X}c\mathrm{P}$	'Peptide' 1.73 [121.7 (25)]	$\langle t\mathbf{X}t\rangle$ 1.74 [121.6 (15)]	$\langle t\mathbf{X}t\rangle$ 1.61	'Peptide' 1.81 [117.2 (22)]	$\langle t\mathbf{X}t\mathrm{P}\rangle$ 1.31 [118.5 (13)]	$\langle t\mathbf{X}t\mathrm{P}\rangle$ 1.25
$t\mathbf{X}c$	'Peptide' 1.69 [121.7 (25)	$\langle t\mathbf{X}t\rangle$ 1.71 [121.6 (15)]	$\langle t\mathbf{X}t\rangle$ 1.48	'Peptide' 2.96 [117.2 (22)]	$\langle t\mathbf{X}c\rangle$ 1.33 [119.8 (13)]	$\langle t\mathbf{X}c\rangle$ 1.24

angles on the N- and C-terminal sides of the residue. The libraries are assessed by their ability to predict the values in our data set, and in every case CDL-v2.0 outperforms E&H 2001.

Since our assessment of CDL-v2.0 is cross-validated, its improvement is not simply owing to overfitting. The leaves two possible contributors to why CDL-v2.0 might be superior: the mean values of the new library could be more accurate and/or there really is a significant dependence on conformation. To distinguish these factors, we also determined the performance (Table 6) of an SVL based on the current data set and created using the same blending groups as CDL-v2.0 but with $h_{\max}$ set to zero.

The seven residue categories shown include four for which E&H 2001 has specific values and three for which it does not. For the first four categories, the new SVL performs somewhat better (∼13%) than E&H 2001 owing to its altered target values, which change between 0.1° and 0.9°. In every case the conformational dependence contributes a larger improvement, with the CDL-v2.0 library performing an additional 9–33% better than even the new SVL.

The largest improvement occurs in CaCN/$\langle t\mathbf{G}t\rangle$. The CDL is superior for this angle because it accounts for the 4° difference between glycine residues with $\psi \simeq 0°$ and those with $\psi \simeq -180°$, which by definition cannot be captured by any SVL. The least impressive example is the CNCa/$\langle t\mathbf{P}t\rangle$ blended group. For this angle the CDL only provides an additional 9% (or 0.12°) decrease. It is known that this angle depends strongly on the φ torsion angle (Berkholz *et al.*, 2009), but φ is very restricted in proline residues.

The total improvement can be even greater in categories for which E&H 2001 has no specific target values. The three categories shown in Table 6 are general residues preceded by a *cis*-peptide ($\langle c\mathbf{X}t\rangle$) and general residues followed by a *cis*-peptide ($\langle t\mathbf{X}c\rangle$ and $\langle t\mathbf{X}c\mathrm{P}\rangle$). As expected, for these categories very large improvements are obtained for the bond angle that is part of the *cis*-peptide (31–70%) and more modest improvements are seen for the bond angles in the *trans*-peptides (7–17%). Also, as should be expected, the angle served best is the CNCa/$\langle c\mathbf{X}t\rangle$ blended group, as this angle is much larger in a *cis*-peptide bond, causing the target value of the E&H 2001 SVL to predict these residues with a terrible r.m.s.d. of 5.43°. Changing the target value to that of an updated SVL dramatically improves the r.m.s.d. to 1.71° and the CDL (using the CNCa/$\langle c\mathbf{X}t\rangle$ blended group) achieves a slight further improvement to 1.61°.

3.7. Comparing CDL-v2.0 and CDL-v1.2

Figs. 11(*a*) and 11(*b*) compare the CDLs for the NCaC/$\langle t\mathbf{X}t\rangle$ category between CDL-v1.2 and the new CDL-v2.0. For the well sampled regions of Ramachandran space the two libraries are very similar, with a notable difference being the continuous nature of the FS-CDL. In contrast, the two libraries differ substantially in those regions of φ/ψ space where there are few residues. In frequently occurring categories such as $t\mathbf{X}t$, this occurs along the 'shores' of the Ramachandran occupied regions, where the new analysis eliminates the sharp discontinuities.

This difference is even more dramatic in the rarer categories, as is shown for NCaC/$\langle t\mathbf{G}t\mathrm{P}\rangle$ (Figs. 11*c* and 11*d*). The CDL-v1.2 targets for this category were deduced from just 44 residues, which mostly clustered around (−60°, −50°) and (−60°, −180°) in φ/ψ space. In CDL-v2.0 the equivalent entry was based on 1471 residues, achieved by the expanded data set of the new PGD search and the blending of eight categories of residues (see Supplementary Fig. S3). This CDL exhibits much more variability in the target value for the NCaC angle owing to the presence of residues in regions such as φ/ψ = (60°, 50°) and (−120°, 100°). It is able to make good use even of the residues located in thinly sampled regions because of the Fourier representation.

We cannot make a direct comparison of the two libraries using $\sigma_{\rm ccv}$ since, as mentioned in Section 3.3, we did not preserve the list of residues used to construct CDL-v1.2 and cannot perform a cross-validation analysis. We can, however, use our new data set to make two indirect comparisons: generating a discrete CDL, such as CDL-v1.2, and an unblended FS-CDL, such as CDL-v2.0, and comparing them (Table 4 and Supplementary Tables S1–S6). The results show

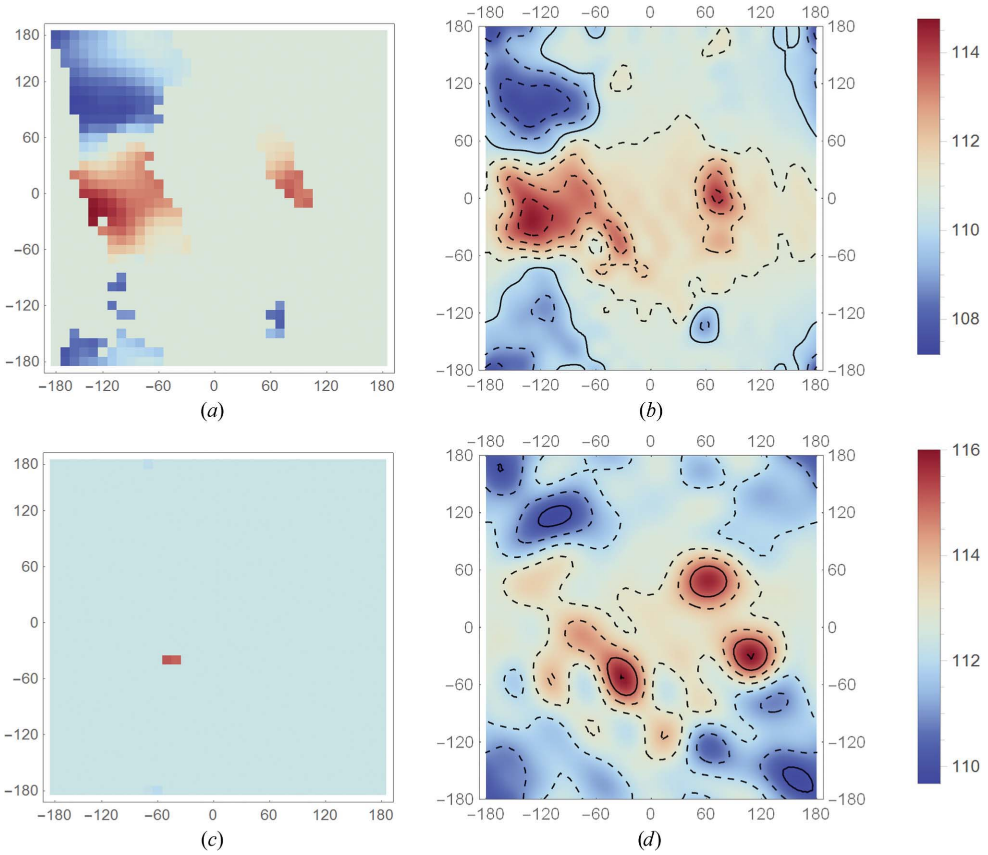


Figure 11

Comparison of CDL-v1.2 and CDL-v2.0 for the NCaC angle in categories $t\mathbf{X}t$ and $t\mathbf{G}t$P. (*a*) CDL-v1.2 $t\mathbf{X}t$. (*b*) CDL-v2.0 NCaC/$\langle t\mathbf{X}t\rangle$. (*c*) CDL-v1.2 $t\mathbf{G}t$P. (*d*) CDL-v2.0 NCaC/$t\mathbf{G}t$P. For each plotted CDL the color scheme is defined by the bar on the right. The φ, ψ and bond angles are all measured in degrees. A solid contour line is drawn at multiples of 5°, with dotted contours every 1°.

that the FS-CDL does as well or better in every category than the discrete CDL based on the same data. This result, combined with the knowledge that the new data set is nearly three times larger, and that a blended FS-CDL will outperform an unblended one, leaves no doubt that CDL-v2.0 is superior to CDL-v1.2.

4. Summary and outlook

The new protein main-chain library that we have presented here has specific entries for all backbone conformations of all 20 amino acids and in each case performs better than CDL-v1.2, which itself is efficient, effective and performed better than conventional SVLs (Tronrud *et al.*, 2010; Tronrud & Karplus, 2011; Moriarty *et al.*, 2016). Thus, CDL-v2.0 represents a step forward in the quality of crystallographic restraints. It should also find use in other protein-modeling applications, including those that could not make use of CDL-v1.2 because it was not a continuous, derivatizable function.

In addition, we emphasize that the Fourier summation method that we have introduced here for describing bond-angle, ζ and ω distributions over φ/ψ space will be applicable for modeling a wide variety of conformation-dependent properties. Such applications could include side-chain CDLs, which would not only allow differing target values for different rotomers, but would allow variations in the target values within rotomers. Also, CDLs could be developed for non-protein molecules such as nucleic acids. The method is not limited to geometric target values, but could also be used to model the behavior of many other properties of molecules, such as the chemical shifts seen in protein NMR experiments, which appear to have a strong, systematic dependence on backbone conformation (Ozenne *et al.*, 2012).

While a library of bond angles parameterized by the φ and ψ angles of a residue is an advance over current stereochemical libraries, there are many avenues for additional improvement, especially as the size of the Protein Data Bank grows. An obvious extension would be to add more independent parameters, such as the conformation angles of neighboring residues and the side-chain torsion angles. Another direction would be to add more categories, such as separating out some of the 16 residues in the **X** category and/or separating out residues in helices and sheets or even in more specific tripeptide or higher order conformational motifs (see, for example, Hollingsworth *et al.*, 2012). Indeed, it has been reported that the NCaC bond angle and the ω torsion angle of residues in regular secondary structures do differ from residues in the same φ/ψ region but not in a helix or sheet [see, for example, Fig. 7 of Touw & Vriend (2010) and Fig. 3(*a*) of Berkholz *et al.* (2012)]. An effect of α-helix formation on geometry is also consistent with the presence of a narrowly focused ~2° dip near (φ, ψ) = (−60°, −40°) seen for the NCaC angle in CDL-v2.0 (Fig. 11*b*). In any of these extensions, group sizes will become smaller, so the strategies described in this paper, of FS modeling to overcome poorly sampled populations, cross-validation to reliably assess the quality of empirical target value functions and blending similarly behaving categories, should all be very helpful.

Acknowledgements

Much of the analysis of our data was performed on the computing resources of the Center for Genome Research and Biocomputing, Oregon State University, Corvallis, Oregon, USA. Also of great help to us was a computer donated to the Karplus laboratory by Art Robbins. We would like to thank Nigel W. Moriarty for many discussions.

Funding information

This work was supported in part by NIH grant GM083136 to PAK.

References

Berkholz, D. S., Driggers, C. M., Shapovalov, M. V., Dunbrack, R. L. & Karplus, P. A. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 449–453.

Berkholz, D. S., Krenesky, P. B., Davidson, J. R. & Karplus, P. A. (2010). *Nucleic Acids Res.* **38**, D320–D325.

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

Berkholz, D. S., Shapovalov, M. V., Karplus, P. A. & Dunbrack, R. L. (2010). *Conformation-Dependent Library for Backbone Geometry in Proteins.* http://dunbrack.fccc.edu/cdl/.

Bernauer, J., Huang, X., Sim, A. Y. I. & Levitt, M. (2011). *RNA*, **17**, 1066–1075.

Bowen, H. J. M., Donohue, J., Jenkin, D. G., Kennard, O., Wheatley, P. J. & Whiffen, D. H. (1958). In *Tables of Interatomic Distances and Configuration in Molecules and Ions*, edited by A. D. Mitchell & L. C. Cross. London: The Chemical Society.

Brereton, A. E. & Karplus, P. A. (2015). *Sci. Adv.* **1**, e1501188.

Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.

Clark, R. M. (1975). *Antiquity*, **49**, 251–266.

Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. (2014). *Protein Sci.* **23**, 47–55.

Das, R. & Baker, D. (2008). *Annu. Rev. Biochem.* **77**, 363–382.

Dodson, E. J., Isaacs, N. W. & Rollett, J. S. (1976). *Acta Cryst.* A**32**, 311–315.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

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

Engh, R. A. & Huber, R. (2001). *International Tables for Crystallography*, Vol. F, pp. 382–392. Dordrecht: Kluwer Academic Publishers.

Esposito, L., Balasco, N., De Simone, A., Berisio, R. & Vitagliano, L. (2013). *Biomed. Res. Int.* **2013**, 326914.

Esposito, L., De Simone, A., Zagari, A. & Vitagliano, L. (2005). *J. Mol. Biol.* **347**, 483–487.

Ho, B. K. & Brasseur, R. (2005). *BMC Struct. Biol.* **5**, 14.

Hoffmann-Ostenhof, O., Cohn, W. E., Braunstein, A. E., Horecker, B. L., Jakoby, W. B., Karlson, P., Keil, B., Klyne, W., Liébecq, C. & Webb, E. C. (1974). *Pure Appl. Chem.* **40**, 291–308.

Hollingsworth, S. A., Lewis, M. C., Berkholz, D. S., Wong, P. & Karplus, P. A. (2012). *J. Mol. Biol.* **416**, 78–93.

Improta, R., Vitagliano, L. & Esposito, L. (2015). *Proteins*, **83**, 1973–1986.

Jabs, A., Weiss, M. S. & Hilgenfeld, R. (1999). *J. Mol. Biol.* **286**, 291–304.

Jiang, X., Cao, M., Teppen, B., Newton, S. Q. & Schaefer, L. (1995). *J. Phys. Chem.* **99**, 10521–10525.

Jiang, X., Yu, C.-H., Cao, M., Newton, S. Q., Paulus, E. F. & Schäfer, L. (1997). *J. Mol. Struct.* **403**, 83–93.

Karplus, P. A. (1996). *Protein Sci.* **5**, 1406–1420.

Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

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

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

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

Mosteller, F. & Tukey, J. W. (1968). *Handbook of Social Psychology*, 2nd ed., edited by G. Lindzey & E. Aronson, Vol. 2, pp. 109–112. Reading: Addison–Wesley.

Nicholson, H., Tronrud, D. E., Becktel, W. J. & Matthews, B. W. (1992). *Biopolymers*, **32**, 1431–1441.

Ozenne, V. A., Schneider, R., Yao, M., Huang, J. R., Salmon, L., Zweckstetter, M., Jensen, M. R. & Blackledge, M. (2012). *J. Am. Chem. Soc.* **134**, 15138–15148.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). *Proc. Natl Acad. Sci. USA*, **37**, 205–211.

Rowicka, M. & Otwinowski, Z. (2004). *AIP Conf. Proc.* **707**, 359–370.

Stein, A. & Kortemme, T. (2013). *PLoS One*, **8**, e63090.

Touw, W. G. & Vriend, G. (2010). *Acta Cryst.* D**66**, 1341–1350.

Tronrud, D. E., Berkholz, D. S. & Karplus, P. A. (2010). *Acta Cryst.* D**66**, 834–842.

Tronrud, D. E. & Karplus, P. A. (2011). *Acta Cryst.* D**67**, 699–706.

Vijayan, M. (1976). *CRC Handbook of Biochemistry and Molecular Biology*, 3rd ed., edited by G. D. Fasman, Vol. II, pp. 742–759. Boca Raton: CRC Press.

Wang, G. & Dunbrack, R. L. (2003). *Bioinformatics*, **19**, 1589–1591.

Weiss, M. S., Jabs, A. & Hilgenfeld, R. (1998). *Nat. Struct. Mol. Biol.* **5**, 676.

wwPDB Consortium (2018). *Nucleic Acids Res.* **47**, D520–D528.