Structure
Article

A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions

Maxim V. Shapovalov[1] and Roland L. Dunbrack, Jr.[1,*]
[1]Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA
*Correspondence: roland.dunbrack@fccc.edu
DOI 10.1016/j.str.2011.03.019

SUMMARY

Rotamer libraries are used in protein structure determination, prediction, and design. The backbone-dependent rotamer library consists of rotamer frequencies, mean dihedral angles, and variances as a function of the backbone dihedral angles. Structure prediction and design methods that employ backbone flexibility would strongly benefit from smoothly varying probabilities and angles. A new version of the backbone-dependent rotamer library has been developed using adaptive kernel density estimates for the rotamer frequencies and adaptive kernel regression for the mean dihedral angles and variances. This formulation allows for evaluation of the rotamer probabilities, mean angles, and variances as a smooth and continuous function of phi and psi. Continuous probability density estimates for the nonrotameric degrees of freedom of amides, carboxylates, and aromatic side chains have been modeled as a function of the backbone dihedrals and rotamers of the remaining degrees of freedom. New backbone-dependent rotamer libraries at varying levels of smoothing are available from http://dunbrack.fccc.edu.

INTRODUCTION

Rotamers are discrete conformations of organic molecules arising from large barriers to rotation about single bonds. Protein side-chain rotamer libraries, which contain frequencies, mean dihedral angles, and standard deviations of common conformations (Dunbrack and Cohen, 1997; Dunbrack and Karplus, 1993; Lovell et al., 2000), are used extensively in structure determination, structure prediction, and protein design. The subdivision of dihedral angle space into rotamers for the sp^3-sp^3 hybridized degrees of freedom enables fast enumeration over all possible conformers. In structure determination they are used as a search space in the process of fitting side-chain conformations to electron density (Adams et al., 2002; Headd et al., 2009) as well as in a number of structure validation methods (Davis et al., 2004). In structure prediction they are used as a discrete search space of conformations (Desmet et al., 1992; Dunbrack and Karplus, 1993), and log rotamer probabilities are sometimes used as a term in scoring functions (Canutescu et al., 2003; Krivov et al., 2009; Liang and Grishin, 2002; Rohl et al., 2004b). In protein design the sequence is altered by substituting in rotamers of different residue types and scoring these conformations in the environment of the side chain, including the rest of the protein and ligands and/or protein partners (Gordon et al., 2003; Kuhlman and Baker, 2004). Thus, rotamer libraries form a critical element in much of computational structural biology, and their ongoing development remains an important task.

We have previously developed backbone-dependent rotamer libraries in which the rotamer frequencies and mean dihedral angles and their standard deviations are given on a 10° × 10° grid of the backbone dihedral angles ϕ and ψ (Dunbrack, 2002; Dunbrack and Cohen, 1997). These libraries were developed using a Bayesian formalism by combining a prior estimate of the probabilities for each (ϕ, ψ) bin with raw counts of the rotamers in overlapping 20° × 20° bins (Dunbrack and Cohen, 1997). The prior estimates came from modeling the observed (ϕ, ψ)-dependent frequencies as the product of ϕ and ψ dependencies. The mean dihedral angles and their variances were determined with a Bayesian normal model that combined separate ϕ- and ψ-dependent estimates with data points around each (ϕ, ψ) grid point.

In attempting to optimize the most recent version of this rotamer library (Dunbrack, 2002) in the program Rosetta (Rohl et al., 2004b), we found that both the rotamer probabilities and the mean dihedral angles and their standard deviations were quite bumpy in their variation with ϕ and ψ, a result of using raw counts in the probability estimates and calculation of simple averages. Rosetta uses the first derivatives of the rotamer probabilities, $-\partial \log P(r|\phi,\psi)/\partial\phi$ and $-\partial \log P(r|\phi,\psi)/\partial\psi$, in the local minimization of its scoring function (Leaver-Fay et al., 2011). The jaggedness of the rotamer library is likely to cause artifacts in any structure determination, prediction, or design program that models backbone flexibility and utilizes local minimization of scoring terms based on the backbone-dependent rotamer library. Backbone flexibility is increasingly incorporated into comparative modeling and protein design (Friedland et al., 2008; Smith and Kortemme, 2008).

Another shortcoming of the previous libraries was the treatment of nonrotameric degrees of freedom, in particular the amide, carboxylate, and aromatic dihedral angle degrees of freedom (the terminal χ angles of Asn, Asp, Gln, Glu, Phe, Trp, His, and Tyr). These degrees of freedom, connecting sp^3 to sp^2 hybridized groups, are difficult to describe as "rotamers" with mean dihedral angles and variances about these means.

Instead, they are usually quite broadly distributed with asymmetric density distributions (Lovell et al., 1999). These distributions may vary with the backbone conformation because the polar side chains interact electrostatically with the local backbone, and the aromatic side chains encounter large steric clashes dependent on ϕ and ψ. Therefore, it is desirable to calculate a full density distribution of these dihedral angles for each (ϕ, ψ) grid point and χ_1 rotamer (or χ_1,χ_2 rotamer for Gln and Glu). This is a complex estimation problem involving the regression of a density onto two angular degrees of freedom.

In this paper our aims in deriving a new backbone-dependent rotamer library are several: (1) to take advantage of the much larger data set that is available now than at the time of the last library (2002); (2) to use electron density calculations to remove highly dynamic side chains (or protein segments) that have uncertain conformations or coordinates (Shapovalov and Dunbrack, 2007); (3) to derive accurate and smooth density estimates of rotamer populations and their relative frequencies, including rare rotamers, as a continuous function of backbone dihedral angles; (4) to derive smooth estimates of the mean values and variances of rotameric side-chain dihedral angles; (5) to improve the treatment of nonrotameric degrees of freedom, i.e., those that are not well described by the rotamer model; and (6) to employ methods producing meaningful estimates of rotamer frequencies, dihedral angles means, and variances in the Ramachandran areas lacking experimental data.

In order to produce smooth and continuous estimates of the rotamer probabilities in this work, we use *kernel density estimation*. A kernel is a nonnegative symmetric function, such as a Gaussian, that integrates to 1.0 and is centered on each data point. Density estimates at specific query points are determined by summing the values of the kernel functions centered on the data points. The smoothness of the density estimate is determined by the form of the kernel, in particular its *bandwidth*. Wider kernels produce smoother density estimates, whereas narrow kernels produce bumpier estimates. For each rotamer, *r*, of a given residue type, we determine a probability density estimate, $\rho(\phi,\psi|r)$, essentially a Ramachandran distribution for each rotamer, and then use Bayes' rule to invert this density to produce an estimate of the rotamer probability, $P(r|\phi,\psi)$:

$$P(r|\phi,\psi) = \frac{\rho(\phi,\psi|r)P(r)}{\sum_{r'}\rho(\phi,\psi|r')P(r')}, \quad (1)$$

where *P*(*r*) is the backbone-independent probability of rotamer (*r*).

Density estimates for angles are more appropriately modeled using the von Mises probability density function (PDF) as the kernel rather than Gaussian or other nonperiodic kernels (Mardia and Jupp, 2000). The von Mises distribution has the form: $\rho(x) = \exp(\kappa \cos x)/I_0(\kappa)$, where *x* is an angle on the circle, and I_0 is the modified Bessel function of the first kind of order zero. The concentration parameter, κ, is inversely proportional to the squared width of the von Mises kernel, with larger values of κ producing narrower kernels. In order to deal with the large variation in density of data points on the Ramachandran map, we use *adaptive kernel density estimation* (Abramson, 1982; Breiman et al., 1977), in which the bandwidth is allowed to vary with the local density of data points. In this way, in sparse regions the kernels placed on each data point are wider, whereas in dense regions the kernels are narrower.

An important feature of our rotamer libraries is the ϕ,ψ-dependence of the means and variances of the dihedral angles for each rotamer, especially for χ_1. Due to interactions with the local backbone, both steric and electrostatic, these average angles have strong and systematic variation with ϕ and ψ for each rotamer (Dunbrack and Cohen, 1997). For this purpose in the new rotamer library, we use adaptive kernel regression (KR) estimators (Brockmann et al., 1993) to determine $\overline{\chi}|\phi,\psi,r$ as smoothly varying functions of the backbone dihedrals. For the KRs we make the concentration parameters of all kernels, κ, adaptive to the same local density of data around the *query* point, rather than the data point as in the kernel density estimates. We also make the variance heteroscedastic, such that it is dependent on the backbone dihedral angles ϕ and ψ.

In our earlier libraries all dihedral angle degrees of freedom were treated as "rotameric." That is, the entire dihedral angle space was broken up into bins and conformations counted. For asparagine, for instance, in 1997 we divided χ_2 into three bins, (−90°,−30°), (−30°,+30°), and (+30°,+90°), by considering OD1 and ND2 atoms as indistinguishable. Later in 2002, we used the *reduce* program of Word et al. (1999) to orient OD1 and ND2 of Asn as well as possible, considering hydrogen-bonding patterns. We then divided χ_2 in the range (−180°,180°) into six bins, with different offsets depending on the χ_1 rotamer. In each of these bins, we calculated mean dihedral angles and standard deviations. This is a poor model for the density, which is broadly distributed and asymmetric. In this work we produce probability density estimates for the nonrotameric degrees of freedom: $\rho(\chi_n|r_{-n},\phi,\psi)$, where r_{-n} in this case represents the rotameric degrees of freedom. This is accomplished by combining the techniques of adaptive kernel density estimation and adaptive KR. These probability distributions will be useful in minimizing the conformational energies of flexible degrees of freedom on smooth potential energy surfaces in the form of $U = -\log \rho(\chi_n|r_{-n},\phi,\psi)$.

The rotamer libraries described here are evaluated on a 10° × 10° grid of ϕ and ψ, but it should be noted that the use of kernels with support from $-\pi$ to π allows us to develop functions that can be evaluated as continuous functions of ϕ and ψ, i.e., at any value of ϕ and ψ, not just those on a predefined grid. This is in contrast to our previous rotamer library formulation using multinomial functions, which required integer counts of each rotamer type within square bins of ϕ,ψ space.

RESULTS

Data Set

The data set used in the new rotamer library was prepared through a series of steps. We first determined the full list of protein-containing PDB entries for which we could obtain electron densities from the Uppsala Electron Density Server (EDS) (Kleywegt et al., 2004). We have shown previously that side chains with sp^3-sp^3 hybridized bonds with nonrotameric dihedral angles, those far from the typical mean values for (60°, 180°, 300°), have much lower electron density than average (Shapovalov and Dunbrack, 2007). This list was then filtered by the PISCES server (Wang and Dunbrack, 2005) and run through

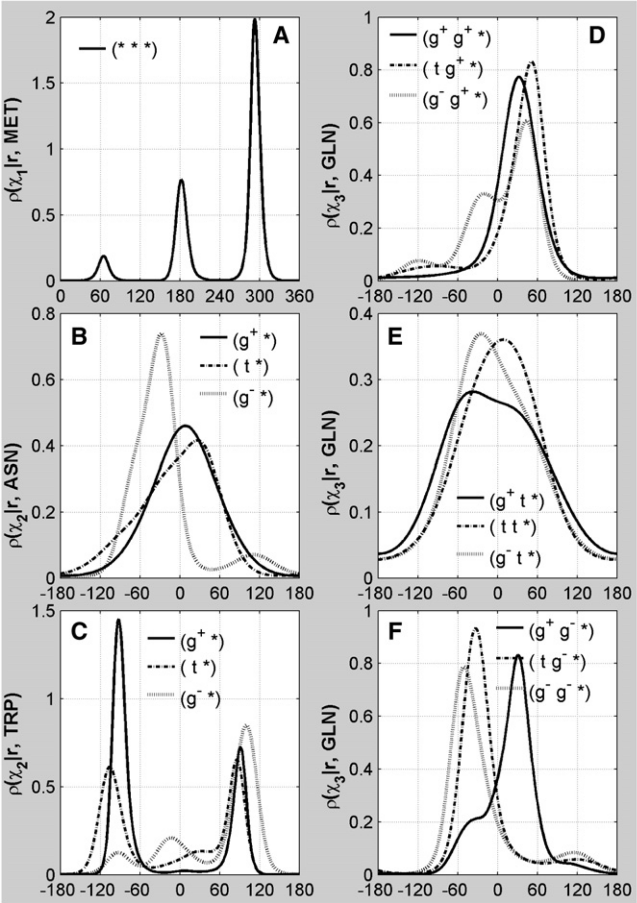

Figure 1. Backbone-Independent Distribution of Rotameric and Nonrotameric χ

(A) A probability density distribution of dihedral angles for a rotameric degree of freedom tightly and symmetrically clustered near the canonical values of 60°, 180°, and 300° based on Met χ_1 data, regardless of χ_1, χ_2, or χ_3 rotamer (*, *, *).

(B and C) Distribution of the nonrotameric χ_2 degree of freedom of Asn and Trp, respectively, for each of their χ_1 rotamers: g^+, t, and g^-.

(D–F) The backbone-independent distribution of nonrotameric χ_3 of Gln for each of its (χ_1, χ_2) rotamers. Nonrotameric χ_3 distributions for Gln are dependent on both the χ_1 and χ_2 rotamers. The distributions of the nonrotameric degrees of freedom are very broad and asymmetric and cannot be modeled with a rotameric model.

the SIOCS program to flip Asn, Gln, and His terminal dihedral angles to account for hydrogen bonding. Finally, we obtained a list of 3985 protein chains from 3845 entries with resolution better than or equal to 1.8 Å, an R factor cutoff of 0.22, and mutual sequence identity of the chains of 50% or less.

We distinguish between rotameric and nonrotameric degrees of freedom based on the hybridization state of the atoms involved in the dihedral angle. Dihedral degrees of freedom are centered on sp^3-sp^3 hybridized bonds and exhibit three narrow, approximately symmetric peaks in their probability density distributions. As an example, the χ_1 density for methionine is shown in Figure 1A, with *gauche*$^+$ {g^+}, *trans* {t}, and *gauche*$^-$ {g^-} peaks at approximately 60°, 180°, and 300°, respectively. Nonrotameric degrees of freedom in protein side chains, by contrast, are centered on sp^3-sp^2 bonds, and exhibit broad and often asymmetric probability density distributions. As examples, the χ_2 probability densities for asparagine and tryptophan are shown in Figures 1B and 1C for each of the three χ_1 rotamers of these residue types. The χ_3 densities for Gln depend on both the χ_1 and χ_2 rotamers, as shown in Figures 1D–1F.

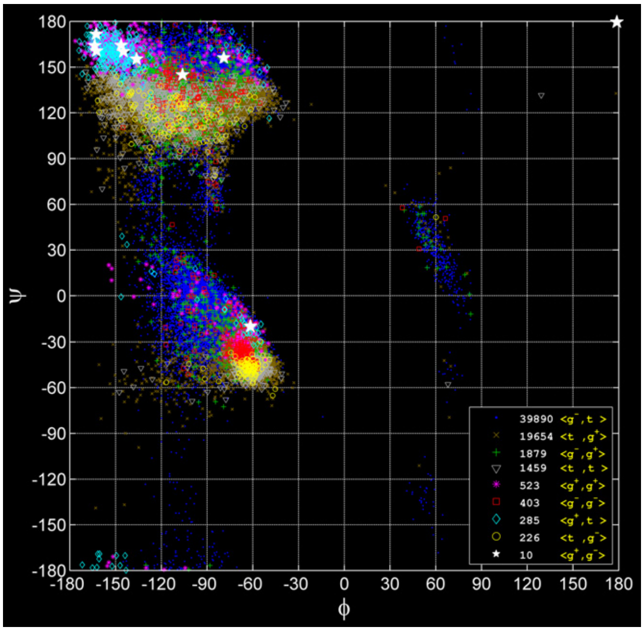

Figure 2. The Backbone-Dependent Rotamer Library Problem

ϕ-ψ Scatter plot of nine leucine rotamers and statistics of the total number of rotamers of each type. The scatter plot has larger and brighter markers for rare rotamers and smaller and darker markers for abundant rotamers. The total number of rotamers differs significantly among the nine types. The relative distributions of each rotamer depend strongly on backbone conformation.

We calculated the electron density at the atom coordinates of 3985 chains using methods described earlier (Shapovalov and Dunbrack, 2007) and calculated the geometric mean of the electron density at the atomic positions in each residue as a quality filter to remove disordered residues—those with electron densities in the bottom 25th percentile for each residue type. For the rotamer library calculations, the resulting number of residues totaled 581,128, and their individual counts are given in Table S1 (available online), along with the degrees of freedom defined for each side-chain type. We also accounted for incorrectly modeled leucine residues (see Figure S1 and Table S2), and we analyzed *trans* and *cis* proline separately, as well as disulfide-bonded and nondisulfide-bonded cysteines.

Deriving Backbone-Dependent Rotamer Probabilities from Kernel Density Estimates

The challenging statistical problem that the backbone-dependent rotamer library presents is shown in Figure 2, a scatter plot of the nine leucine rotamer types on the Ramachandran map. The goal is to calculate $P(r|\phi,\psi)$, the probability of each rotamer as a function of the backbone dihedrals ϕ and ψ. The nonuniform distribution in ϕ,ψ and the large differences in overall populations and distributions of the different rotamers must all be accounted for. Our solution to this problem is to use adaptive

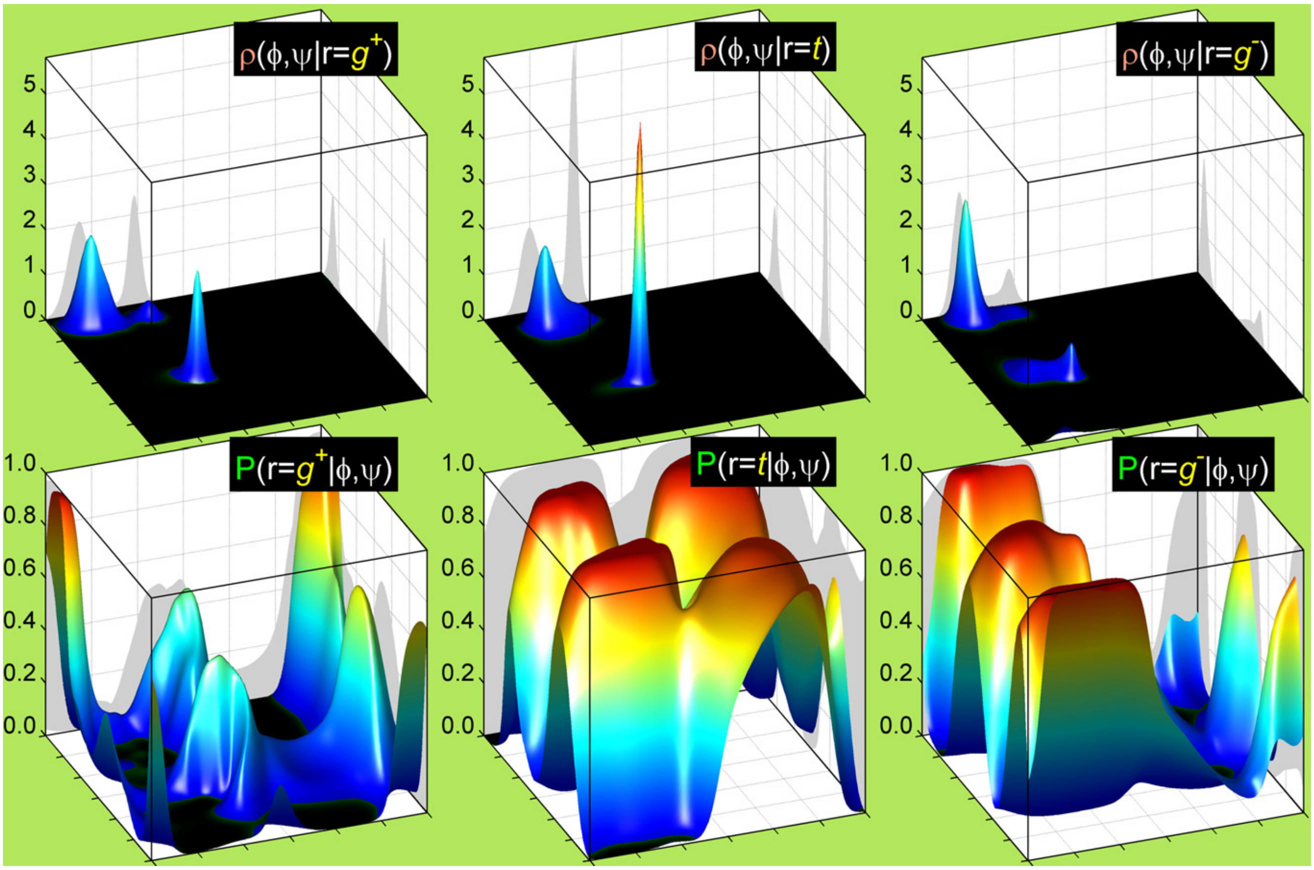

Figure 3. Rotamer Ramachandran Densities and Their Corresponding Backbone-Dependent Rotamer Probabilities from the New 2010 Rotamer Library

The top view shows smoothed Ramachandran PDFs of the backbone conformation (ϕ, ψ) for g^+, *t*, and g^- rotamers (left to right) of Val computed with adaptive kernel density estimation. ϕ and ψ are plotted along x axis and y axis, respectively, within their standard limits of (−180°, 180°). The PDFs are plotted along the z axis and scaled in 1/radian^2. For every rotamer the density integrates to 1 over the whole Ramachandran area. The bottom view illustrates corresponding 2010 smooth backbone-dependent rotamer probabilities, calculated by inverting the Ramachandran densities in the top row with Bayes' rule. The probabilities of all three g^+, *t*, and g^- rotamers sum up to 1 for every (ϕ, ψ). The Val bandwidth radius is 5°, and the concentration parameter, κ, is 120. These values match the 5% step down from the optimal log-likelihood score for additional smoothness with the best SCWRL4 prediction rates (see Results).

kernel density estimates (AKDEs) to obtain rotamer PDFs, $\rho(\phi,\psi|r)$ from the input data set $\{\phi_i,\psi_i,r_i\}$, and to use Bayes' rule to invert these densities to obtain the rotamer probabilities, $P(r|\phi,\psi)$.

As an example, in the top row of Figure 3, we show the PDFs $\rho(\phi,\psi|r)$ of the g^+, *t*, and g^- rotamers of valine above their resulting backbone-dependent probabilities, $P(r|\phi,\psi)$, shown in the bottom row. The three rotamers have notably different ϕ,ψ probability densities that in turn produce quite different relative frequencies of the three rotamers as a function of ϕ and ψ. These estimates match conformational analysis of syn-pentane interactions (Wiberg and Murcko, 1988) of the side-chain Cγ_1 and Cγ_2 atoms with atoms of the backbone whose positions are dependent on ϕ and ψ (Dunbrack and Cohen, 1997; Dunbrack and Karplus, 1994).

To reach the results shown in Figure 3 for the new backbone-dependent rotamer library, we investigated and compared the results from a number of different methods. These are shown together in Figure 4 for the g^+ rotamer of serine, $P(r=g^+|\phi,\psi, aa=\mathrm{Ser})$. In the straightforward histogram approach (Figure 4A), the number of data points with a particular rotamer in every nonoverlapping (ϕ,ψ) bin is counted and divided by the total number of data points of any rotamer type in the same bin. This approach produces crude estimates of the rotamer probabilities. The prevailing majority of the 10° × 10° histogram bins have "unknown" values (set to zero in the figure), produced by division of zero points by zero points. A large proportion of the bins have very spiky and extremely unreliable probability estimates.

The Bayesian approach used in our 1997 and 2002 rotamer libraries used 2-fold periodic kernels (although we did not call them as such at the time) to produce separate ϕ-dependent and ψ-dependent counts as a prior in the form of a Dirichlet function, which were combined with integer data counts in a multinomial likelihood to produce posterior estimates also in the form of Dirichlet functions (Dunbrack and Cohen, 1997). As shown in Figure 4B, this approach produced reasonable estimates for all values of ϕ and ψ, but because of the integer counts in the Dirichlet function, the posterior estimates were very bumpy as a function of ϕ and ψ.

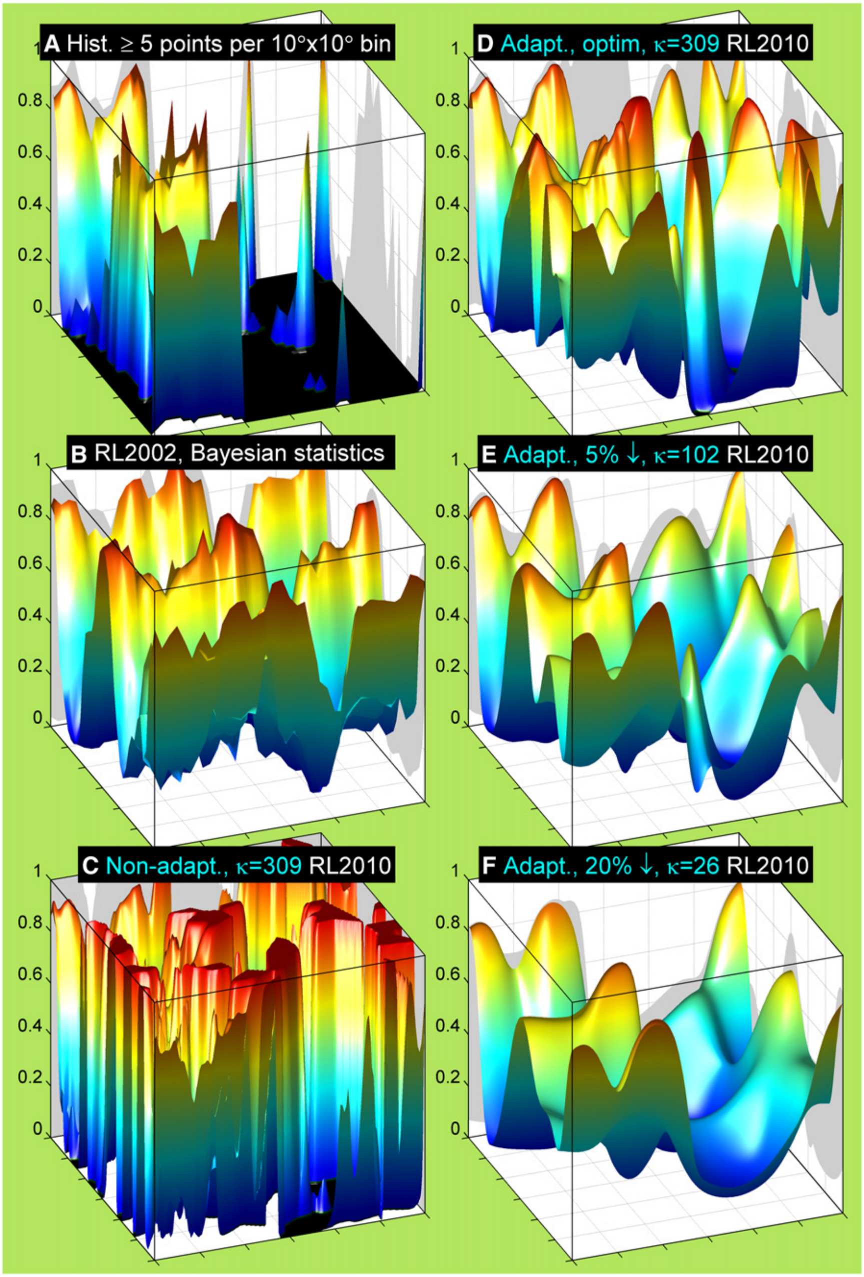

Figure 4. Rotamer Probability Estimates Produced by Several Methods and Smoothing Effect of Adaptive Kernel Density with Narrower or Wider Bandwidths

Nonoverlapping 10° × 10° bin histogram (A), 2002 Bayesian (B), nonadaptive kernel density (C), and adaptive kernel density (D–F) estimates are shown for $P(r = g^{+} \mid \phi, \psi, aa = Ser)$. The histogram estimate (A) depicts only the bins with at least five points of any rotamer per bin. The non-AKDE (C) has a fixed bandwidth (κ = 309, bandwidth radius, R = 3.3°), the same as for (D). The AKDEs with widening geometric-mean kernel bandwidth are ordered from (D)–(F). The maximum log-likelihood (κ = 309, R = 3.3°), 5% step-down (κ = 102, R = 6°), and 20% step-down (κ = 29, R = 11°) bandwidths are shown in (D), (E), and (F), respectively.

In this work we are also using kernels to estimate the ϕ,ψ-dependent densities of each rotamer, but instead of combining them with data counts, we use the kernels directly to determine density estimates for each rotamer and Bayes' rule to determine the rotamer probabilities. In our first attempt we used kernel density estimates with fixed and constant kernel widths for all data points. The resulting rotamer probability for the serine g^{+} rotamer, calculated with a concentration parameter in the von Mises kernel function of κ = 309 (a bandwidth radius of 3.3°) is shown in Figure 4C. It reproduces the form of the Bayesian estimates, but the transitions are rather sharp, and it is very sensitive to outlier data in the ϕ,ψ space. A wider radius for the non-AKDE data produces smoother estimates than shown in Figure 4C, but such a radius flattens out the rotamer probabilities too much, leading to inaccurate probabilities even when data are plentiful (not shown).

To reduce the effect of outliers, we then employed AKDEs in which the kernel widths vary with the local density of data points. At higher densities, the kernels are narrower, and at lower densities, such as in the vicinity of outliers, the kernels are wider, thus spreading out and minimizing their effect on the density estimates. The widths of the kernels are determined by a concentration parameter scaled with the square root of the local density of points, $\widehat{f}(\phi,\psi)$, obtained from some pilot estimate (in this case the nonadaptive kernel density). With the base kernel concentration parameter κ optimized to maximize the log likelihood of $P(r|\phi,\psi)$ using 10-fold cross-validation, we calculated the rotamer probabilities shown in Figure 4D. The optimized value for serine is κ = 309, so that the nonadaptive and adaptive rotamer probabilities in Figures 4C and 4D use the same value of κ. The adaptive version is much smoother than the nonadaptive version.

While eliminating the effects of the outliers, the changes in rotamer probability in Figure 4D may be sharper than optimal for programs like Rosetta that depend on the first derivatives of log $P(r|\phi,\psi)$. In order to increase the smoothness, we employed a *penalized* maximum likelihood procedure for optimizing the concentration parameter κ. This is a common procedure in density estimations (Eggermont and LaRiccia, 2001). The total log-likelihood expression can be modified in a number of ways. We use a simple approach that penalizes the *average* log likelihood by a fixed percentage of the range from its maximum value to its minimum value. In Figures 4E and 4F, we show the g^{+} rotamer of serine calculated with concentration parameters such that the average log likelihood is 5% and 20% less than

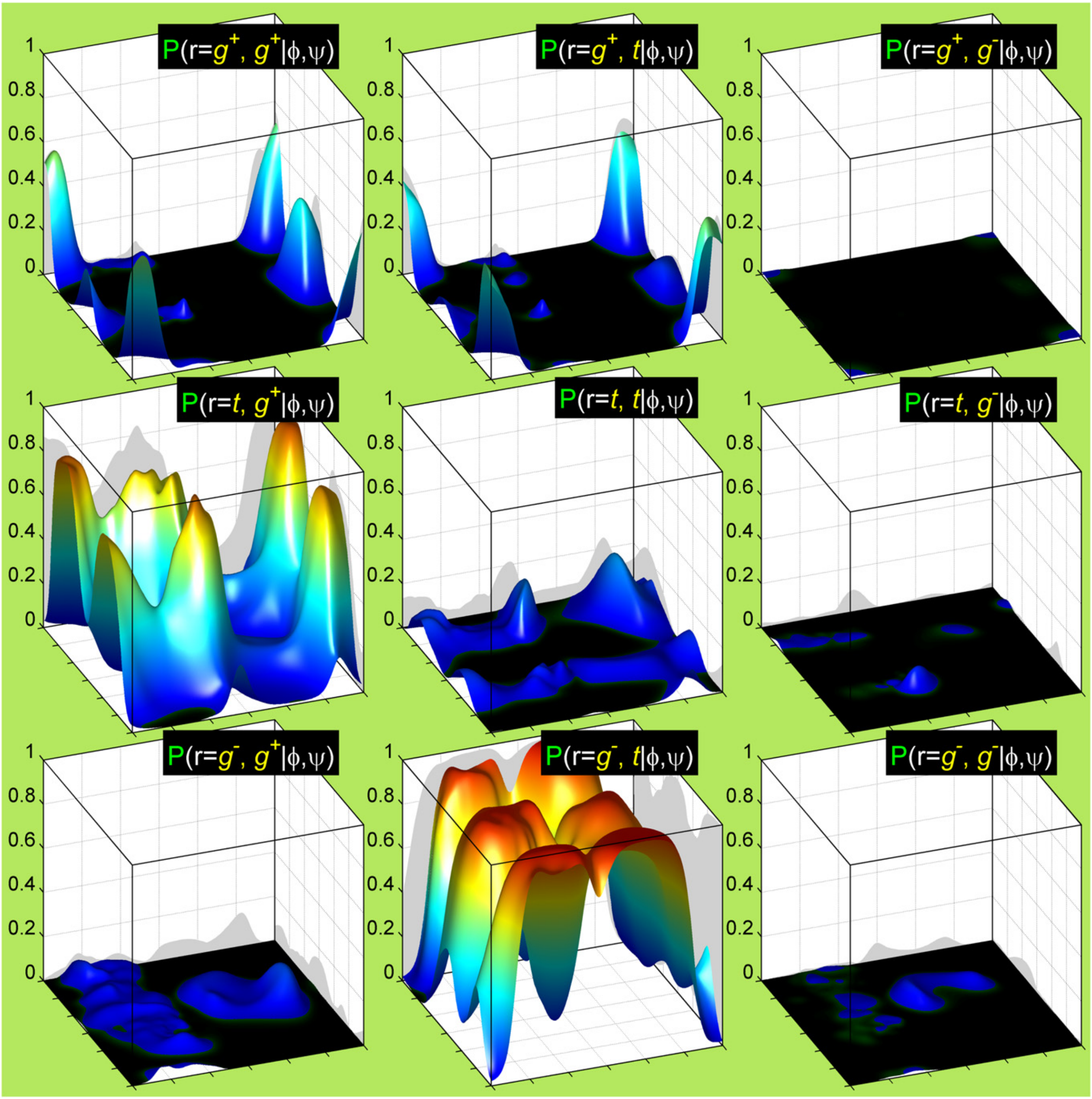

Figure 5. A Complete Set of Backbone-Dependent Rotamer Probabilities for Leucine Derived from AKDEs of the New 2010 Rotamer Library

Leu demonstrates strong variation in its rotamer preferences both in the backbone-dependent and backbone-independent rotamer libraries. Some of its rotamers are restricted everywhere on the (ϕ, ψ) map, due to strong clashes of the side-chain conformations with its own backbone. The $\{g^+, g^-\}$ rotamer has only 10 data points in our data set, whereas the total number of leucines is 64,329. The rare rotamer fix is used to calculate the Ramachandran probability density for the $\{g^+, g^-\}$ rotamer using only the $\{g^+\}$ data.

its full range shown in Figure 4D. Thus, Figures 4D–4F illustrate the smoothing effect of the widening bandwidth radius (2°, 5°, and 11°) of the AKDEs on the rotamer probability estimates. The methods for choosing the optimized κ and the step-down values of κ are illustrated in Figure S2. The optimized values of κ and the bandwidth radius and the same values for the 5% step down in the average log likelihood are given in Table S3. The appropriate choice of smoothing level may depend on the application for which the rotamer library is intended. We explore this further below.

For the rarer rotamers (those with less than 25 examples in the data set), we approximated the rotamer probability density $\rho(\phi, \psi|r)$ with rotamer data of the same side-chain type with one or more fewer degrees of freedom. In Figure 5, we present

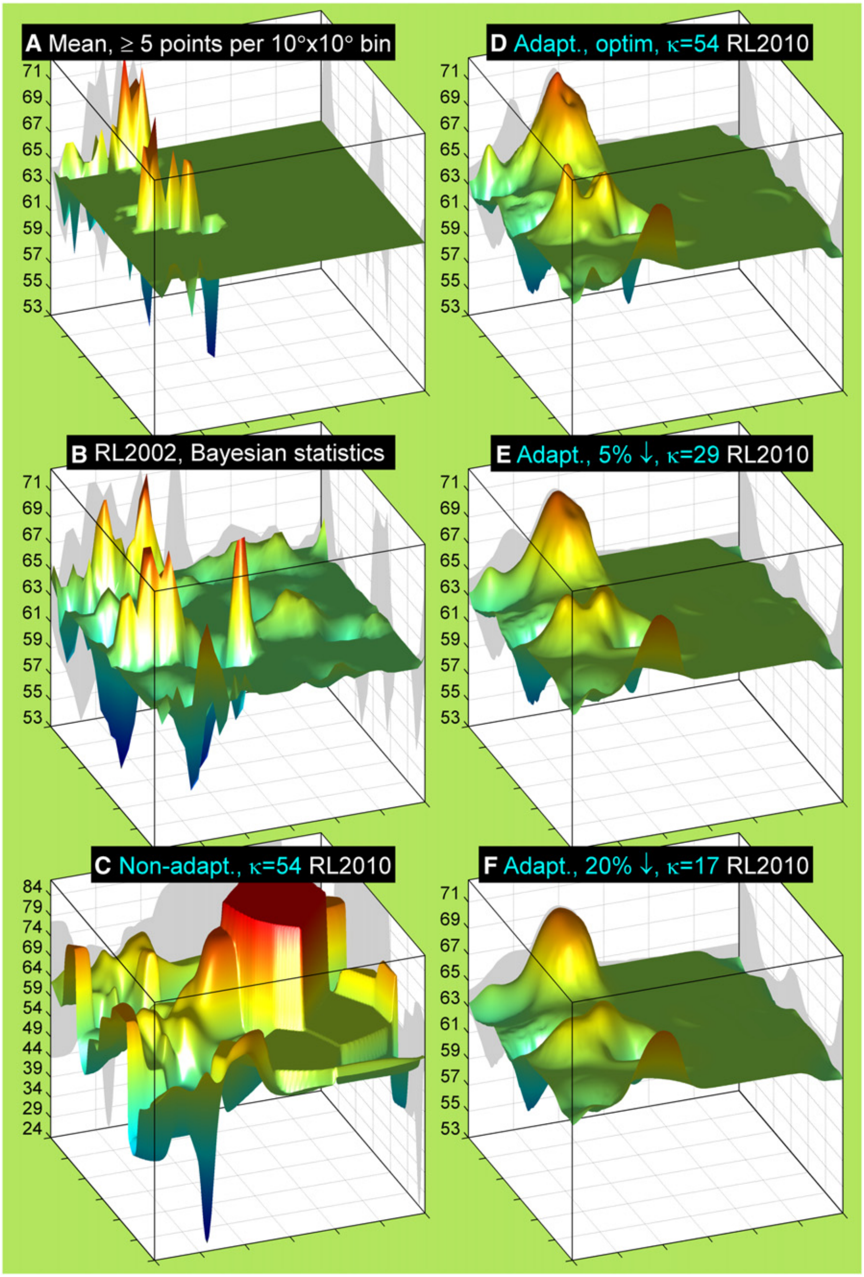

Figure 6. Rotameric χ Mean Estimates Calculated with Several Methods and Smoothing Effects of Query-Adaptive KRs

Nonoverlapping 10° × 10° bin average (A), 2002 Bayesian (B), nonadaptive KR (C), and query-adaptive KR (D–F) estimates are shown for μ (χ | ϕ, ψ, *r* = g^+, *aa* = *Cys*). The 10° × 10° bin average has only the bins with at least five g^+ rotamers per bin. The nonadaptive KR (C) has a fixed bandwidth (κ = 54, bandwidth radius, *R* = 8°), the same as for (D). The query-adaptive KR estimates with widening geometric-mean kernel bandwidth are ordered from (D)–(F). The maximum log-likelihood (κ = 54, *R* = 8°), 5% step-down (κ = 29, *R* = 11°), and 20% step-down (κ = 17, *R* = 14°) bandwidths are in (D), (E), and (F), respectively.

the rotamer probability estimates for the nine rotamers of leucine. For leucine, the {g^+, g^-} probability density was calculated with the {g^+, X} data of leucine. The factor *P*(*r*) in Equation 1 is calculated based on the actual counts of the {g^+, g^-} rotamer, whereas $\rho(\phi,\psi|r)$ is calculated with the *r* = g^+ data, producing a reasonable estimate of $P(r = g^+, g^- | \phi, \psi)$.

Rotameric Side-Chain Degrees of Freedom: Backbone-Dependent KR of χ Means and Variances

As with the backbone-dependent rotamer probabilities, we investigated a number of approaches in calculating the backbone-dependent means and standard deviations of side-chain dihedral angles for the rotameric degrees of freedom. In Figures 6 and 7, respectively, we show the results of several different ways of calculating the $\mu\chi_1$ and $\sigma\chi_1$ estimates for g^+ rotamer of cysteine: $\mu(\chi_1|\phi,\psi,r=g^+)$ and $\sigma(\chi_1|\phi,\psi,r=g^+)$. The simplest way is to average χ_1 points and also calculate their standard deviation within nonoverlapping 10° × 10° bins. As with the histogram approach to rotamer probabilities described above, this method produces very crude and spiky estimates of $\mu\chi_1$ and its $\sigma\chi_1$, as observed in Figures 6A and 7A. In the bins with few data points, their means and deviations are statistically unreliable.

In the 1997 and 2002 rotamer libraries, we combined ϕ-dependent and ψ-dependent estimates of the mean angles and their variances with the data in overlapping 20° × 20° bins in a Bayesian estimation procedure. The 2002 rotamer library estimates are shown in Figures 6B and 7B. These estimates are extremely bumpy due to the large effect of a small number of side chains when the data are sparse. A nonadaptive KR scheme also produces bumpy and extreme estimates, as shown for a bandwidth of 8° in Figures 6C and 7C. This kernel captures very few data points at most query points and produces unreliable estimates of mean and standard deviation. The nonadaptive KR with a much wider bandwidth (not shown) is not as noisy but loses valuable features in the populated areas of (ϕ, ψ).

Thus, we moved to an adaptive scheme, applying *query-adaptive* KR to estimate the rotameric χ means and their variances. The bandwidth varies as a function of the density local to the query point, rather than by the density around the data points, as used in the density estimates described earlier. We found that query-adaptive kernels provided regression curves and surfaces that more accurately modeled the observable variations in the χ angles as a function of ϕ and ψ than data-adaptive kernels.

For rotameric backbone-dependent χ mean and variance, we utilized the sum of the squared residuals between the

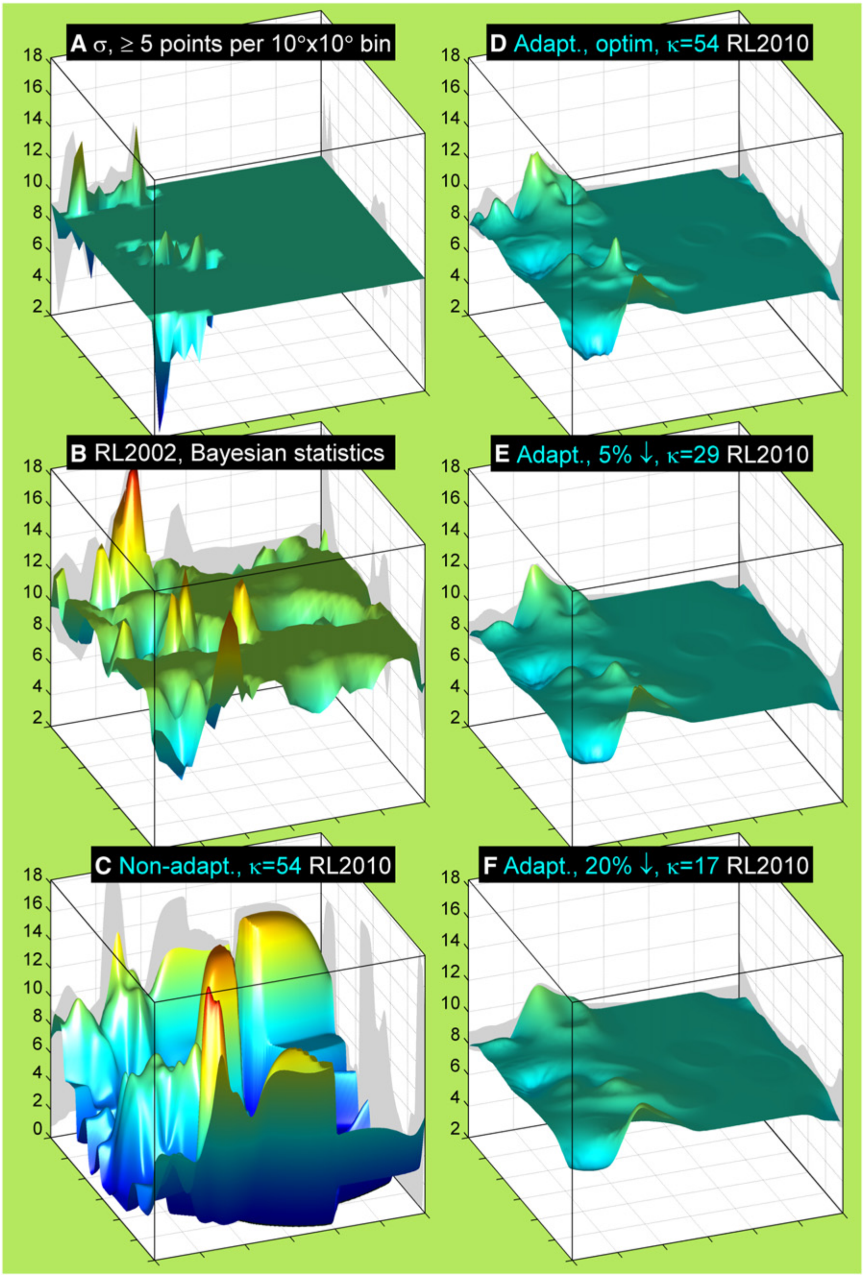

Figure 7. Rotameric χ Standard Deviation Estimates Calculated with Several Methods and Smoothing Effects of Query-Adaptive KRs

Nonoverlapping 10° × 10° bin (A), 2002 Bayesian (B), nonadaptive KR (C), and query-adaptive KR (D–F) estimates are plotted for σ (χ | ϕ, ψ, $r = g^+$, aa = Cys). The 10° × 10° bin estimate is shown only in the bins with at least five g^+ rotamers per bin. Other information and parameters are the same as in Figure 6.

experimental χ points and the surface of the mean estimate as the objective function for minimization. The minimization was carried out for each χ angle of each rotamer separately. The optimal concentration parameters and their corresponding bandwidths used in the KR can be found in Table S3.

As with the kernel density estimates, we also applied a simple form of penalized KR, by stepping down the objective function by 2%, 5%, 10%, and 20%. The values of κ that result from the 5% step down are also given in Table S3. Figures 6D–6F and 7D–7F reveal the smoothing effect of the widening bandwidth radius (7°, 10°, and 13°) of the query-adaptive KR of $\mu\chi_1$ and $\sigma\chi_1$, respectively. Higher values of κ produce bumpier regression surfaces, and lower κ values produce flatter, smoother surfaces. As in the case with the rotamer probabilities, the appropriate level of smoothing may depend on the application.

For some (ϕ, ψ) values, clashes between the side-chain Xγ atom and backbone atoms whose positions are dependent on ϕ and ψ push the χ_1 means away from their canonical values in order to relieve the clash (Dunbrack, 2002; Dunbrack and Cohen, 1997). For instance the g^+ rotamer shown in Figures 6 and 7 has steric clashes with backbone atoms O_i and N_{i+1} when ψ is near 120° and −60°, respectively, and these interactions lead to a deviation in the χ_1 dihedral angle means. In the unpopulated regions of the Ramachandran map, the query-dependent KRs return to the backbone-independent mean value, which is a reasonable estimate because the angles do not usually vary more than about 15° from these values in any case. These are the flat areas in the $\mu\chi_1$ and $\sigma\chi_1$ KR surfaces in Figures 6 and 7. The $\sigma\chi_1$ estimates are also larger when the side-chain and backbone atoms clash.

Nonrotameric Side-Chain Degrees of Freedom: Backbone-Dependent KR of χ Angle Densities

The terminal dihedral angles of Asn, Asp, Gln, and Glu have very broad distributions, when considered independent of ϕ and ψ, as shown for Asn in Figure 1B and Gln in Figures 1D–1F. The terminal dihedral angles of the aromatic amino acids have distributions broader than typical rotameric degrees of freedom, and these are somewhat asymmetric, as shown for Trp in Figure 1C. Therefore, the normal model used for the rotameric degrees of freedom as for Met χ_1 in Figure 1A (regression to a mean and standard deviation) is inappropriate for these degrees of freedom, and therefore, we refer to them as "nonrotameric." The distributions of these nonrotameric angles vary significantly with ϕ and ψ. However, because they cannot be modeled parametrically, they must be modeled with nonparametric density estimates. Therefore, we seek a method to determine a regression of the density of an angle onto the explanatory variables ϕ and ψ.

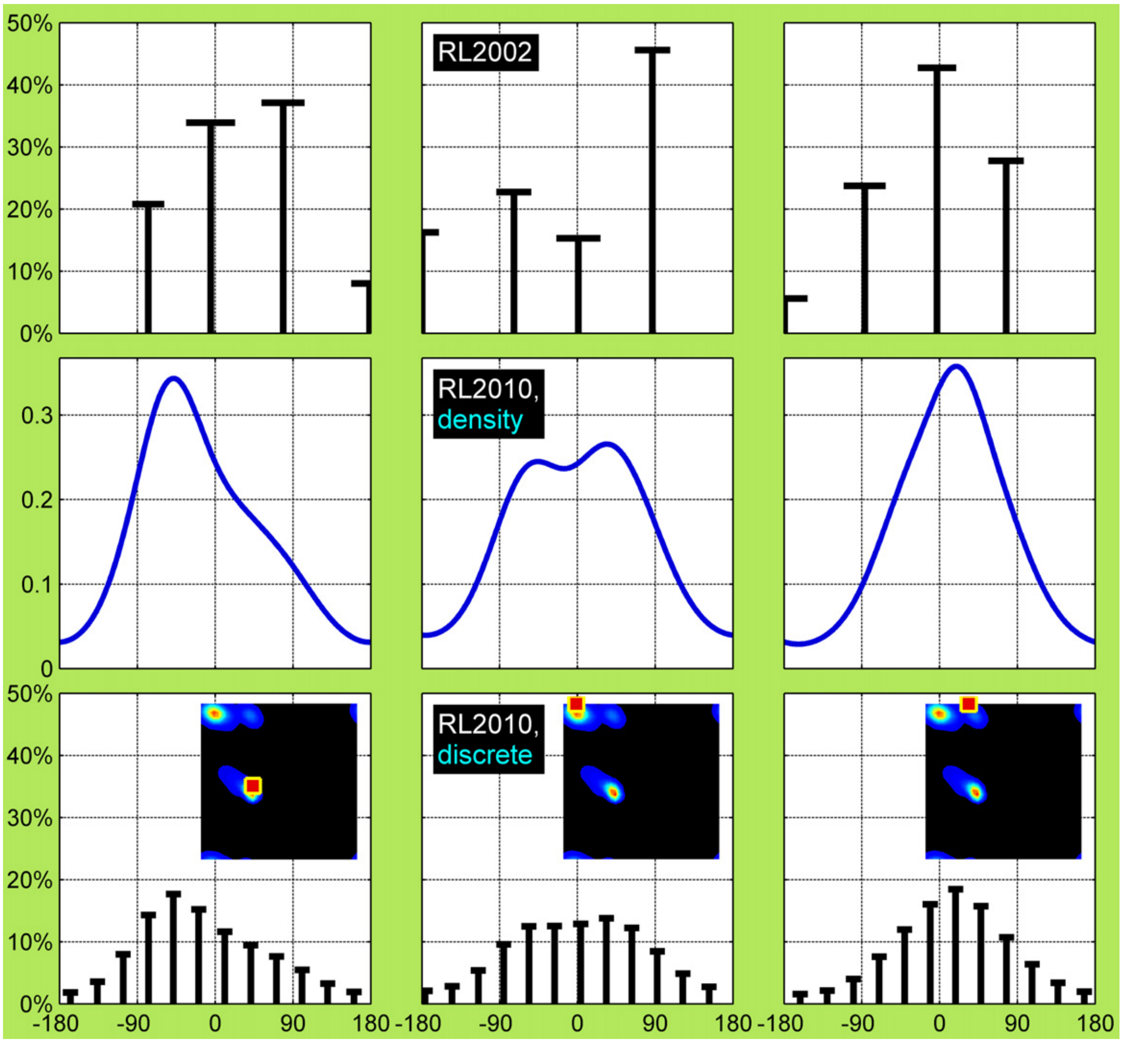

Figure 8. Backbone-Dependent Treatment of Nonrotameric Side-Chain χ: 2002 Rotamer Library, 2010 Density Model, and 2010 Discrete Model

Backbone-dependent modeling of nonrotameric χ_3 of χ_1,χ_2 rotamer = { g^+,t } of Gln using Bayesian formalism of the 2002 rotamer library (top), 2010 query-adaptive KR of densities (middle), and 2010 binned "rotameric" model (bottom). These three models are provided at three different selected (ϕ, ψ) locations: (−60°, −10°), (−150°, 180°), and (−80°, 180°), indicated on the Ramachandran { g^+, t } density insets in the bottom row. The top and bottom models are binned or "rotameric," whereas the middle model is continuous density. The "rotameric modeling" of the nonrotameric χ_3 includes: r_3 probabilities (heights of the bars); $P(r_3 \,|\phi, \psi, r_{12} = \{g^+, t\})$ summing up to 1; χ_3 means (positions of the bars); $\mu(\chi_3 \,|\phi, \psi, r_{123})$; and χ_3 standard deviations (lengths of the horizontal bars at the tip of the bars), $\sigma(\chi_3 \,|\phi, \psi, r_{123})$.

In the 1997 and 2002 rotamer libraries, nonrotameric χ angles were modeled in a manner very similar to the rotameric degrees of freedom despite the deficiencies of such modeling. This was accomplished by defining bins for each "rotamer," establishing prior estimates formed from a product of individual ϕ-dependent and ψ-dependent distributions, and adding counts of χ_2 in each bin from the neighborhood around each ϕ,ψ grid point. In the 2002 library, Asn had six χ_2 bins for each χ_1 rotamer over 360°, whereas Gln had four bins for χ_3. Asp and Glu had three bins over 180°, whereas Phe and Tyr had two bins. His and Trp each had three bins of 120° each. For each bin we calculated mean dihedral angles and their variances as well as relative populations. This is shown in the first row of Figure 8 for the {g^+,t} rotamer of Gln for three different ϕ,ψ positions: near the α helix region (−60°, −10°); near the β sheet region (−150°, 180°); and near the polyproline II region frequently occupied in loops (−80°, 180°). Each bar is located at the mean value of each bin, and the horizontal bars indicate the standard deviation of the data in that bin, which is proportional to the bin widths.

In the new 2010 rotamer library, we take a different approach and model the nonrotameric χ as continuous distributions as a function of (ϕ, ψ) for every rotamer combination of the rotameric degrees of freedom of the residue. For example, Gln has three side-chain degrees of freedom: rotameric χ_1, χ_2, and the terminal nonrotameric χ_3. Therefore, we calculate backbone-dependent χ_3 density distributions for each of the nine χ_1,χ_2 rotamers of Gln. We accomplish this by applying query-adaptive kernels to ϕ and ψ and data-adaptive kernels to the nonrotameric

Table 1. 2002 versus Best Smooth 2010 Rotamer Libraries: Benchmarking Based on SCWRL4 Side-Chain Conformation Prediction Accuracy

	χ Angles	TRP	PHE	GLN	GLU	TYR	SER	ARG	HIS	LEU	MET	CYS	THR	ASP	ILE	VAL	LYS	ASN	PRO	ALL
Best ’10	χ_1	94.1	98.1	85.0	81.0	97.1	75.4	83.1	93.5	96.4	90.2	93.2	94.3	90.5	98.5	96.9	82.8	91.7	87.1	90.15
Old ’02		92.9	97.6	84.6	80.1	96.5	74.3	83.3	93.9	95.9	90.4	92.8	94.0	90.6	98.4	96.7	82.6	91.7	87.3	89.83
Δ(Best,Old)		**+1.2**	**+0.5**	**+0.4**	**+1.0**	**+0.6**	**+1.1**	*−0.2*	*−0.4*	**+0.5**	*−0.2*	**+0.4**	**+0.4**	*−0.1*	**+0.1**	**+0.2**	**+0.3**	0.0	*−0.2*	**+0.32**
Best ’10	χ_{1+2}	84.6	96.6	71.1	68.0	94.8		72.9	66.4	91.9	81.9			84.7	91.0		72.3	76.7	83.9	81.73
Old ’02		78.9	93.7	71.0	67.5	92.9		72.5	64.6	91.2	81.9			83.8	90.6		72.5	77.0	84.3	81.01
Δ(Best,Old)		**+5.7**	**+2.9**	**+0.1**	**+0.6**	**+1.9**		**+0.5**	**+1.8**	**+0.8**	0.0			**+0.9**	**+0.4**		*−0.2*	*−0.3*	*−0.4*	**+0.72**
Best ’10	χ_{1+2+3}			48.8	52.4			51.0			64.2						58.4			54.01
Old ’02				44.5	49.3			49.6			62.5						58.7			52.05
Δ(Best,Old)				**+4.2**	**+3.1**			**+1.5**			**+1.7**						*−0.3*			**+1.96**
Best ’10	$\chi_{1+2+3+4}$							38.1									39.9			38.99
Old ’02								36.3									39.6			38.01
Δ(Best,Old)								**+1.8**									**+0.2**			**+0.98**
Best ’10	χ_{all}	89.3	97.4	68.3	67.1	96.0	75.4	61.3	80.0	94.1	78.8	93.2	94.3	87.6	94.8	96.9	63.4	84.2	85.5	83.72
Old ’02		85.9	95.7	66.7	65.6	94.7	74.3	60.4	79.2	93.5	78.3	92.8	94.0	87.2	94.5	96.7	63.4	84.3	85.8	83.04
Δ(Best,Old)		**+3.4**	**+1.7**	**+1.6**	**+1.5**	**+1.2**	**+1.1**	**+0.9**	**+0.7**	**+0.6**	**+0.5**	**+0.4**	**+0.4**	**+0.4**	**+0.3**	**+0.2**	**0.0**	*−0.2*	*−0.3*	**+0.67**

The performances of the new 2010 rotamer libraries were compared with the 2002 rotamer library. SCWRL4 was run on a set of 379 high-resolution proteins used previously (Krivov et al., 2009). The FRM of SCWRL4 was used, and crystal symmetry was used in the calculations (all side chains in all copies of the asymmetric unit were calculated simultaneously). Accuracy was evaluated on all side chains in the proteins excluding those with electron density in the bottom 25th percentile for each residue type. A predicted side-chain χ is considered correct if its value lies within 40° from its experimental value. For each residue type the 2002 and 2010 accuracies are provided for each individual χ angle. χ_{all} is an absolute average of all degrees of freedom for each residue (see Supplemental Experimental Procedures). ALL is an average accuracy among all 18 standard residue types. Percentages in bold type show improvements in prediction rate; those in italics are decreases in prediction rate.

χ_n to estimate $\rho(\chi_n|\phi,\psi,r_{-n})$, where n indicates the terminal dihedral angle, and r_{-n} indicates the rotamer of the nonterminal degrees of freedom. In the middle panel of Figure 8, the Gln χ_3 density of the $r_{-n}=\{g^+,t\}$ rotamer is evaluated every 1° for the same three (ϕ, ψ)s as in the first row for the 2002 library. The distributions show that the modes are located at different positions for each ϕ,ψ point, the peaks are asymmetric, and in one case the distribution is bimodal. The curves roughly parallel the 2002 rotamer library, if the curves are integrated over 90° regions. For practical applications we report backbone-dependent nonrotameric χ_n density every 10°.

To support existing applications such as SCWRL, which rely on our older 1997/2002 libraries and their format, for the new rotamer library, we also create a new more detailed “rotameric” model for nonrotameric χ. To meet this goal and to accommodate a more complex distribution structure, we increased the number of bins for the nonrotameric χ (Table S1). The rotamer bin width is decreased to 30°. The backbone-dependent probabilities are estimated by the product of the integrated continuous density over each bin and the corresponding backbone-dependent probabilities *of* r_{-n} (see Equations S39 and S40). The vertical bars are centered at the means, and their horizontal bars specify the standard deviations of each of the 12 χ_3 rotamers. These are estimated by integrating a product of the χ_3 density and corresponding function over each of 12 bins (Equation S39). Thus, Figure 8 illustrates binned and continuous models of nonrotameric χ angles and how the binned modeling has been changed since the 2002 analysis.

We also provide a movie of the probability density of χ_2 for the g^+ rotamer of Asn as a function of (ϕ,ψ) (Movie S1). Additional figures and movies are available at http://dunbrack.fccc.edu.

Using the Backbone-Dependent Rotamer Library in Structure Prediction

The methods we have developed using kernel density estimates and KRs have allowed us to develop smooth and statistically reliable backbone-dependent rotamer libraries. We can adjust the level of smoothing for different applications by adjusting the penalties in the objective functions for the rotamer probabilities and regressions. To choose a reasonable set of values, we tested a number of different libraries with our side-chain prediction program SCWRL4 (Krivov et al., 2009) and Rosetta (Rohl et al., 2004b). For SCWRL4 benchmarking we used the same testing set of 379 high-resolution protein monomers as in the original SCWRL4 work with a resolution cutoff of 1.8 Å and maximum mutual sequence identity of 30%. For Rosetta we used a set of 50 monomeric, ligand-free proteins without disulfides and with resolution of 1.6 Å or better and less than 20% mutual sequence identity.

In the side-chain prediction literature, a side-chain torsion angle is considered correctly predicted if its value is within 40° from the experimental one. Using this traditional definition, in Table 1, we compare the best 2010 library versus the older 2002 library in SCWRL4 prediction rates based on the flexible rotamer model (FRM) for each individual degree of freedom (χ_1, χ_2, χ_3, and χ_4) and the overall χ accuracy. The best 2010 rotamer library gives an overall increase of +0.67% in χ angle predictions on a test set of 379 proteins. This is a weighted average over χ_1, χ_{1+2}, χ_{1+2+3}, and $\chi_{1+2+3+4}$ accuracies (see Equations S42 and S43). Although this is a modest increase, many highly populated side-chain types are already at very high accuracies and cannot be improved much further. Except for Pro (−0.3%) and Asn (−0.2%), the best 2010 library has

Table 2. Effects of 2010 Rotamer Library Smoothing in SCWRL4 and Rosetta

		2002	Optim	2%↓	5%↓	10%↓	20%↓	25%↓	2009it10
Side-Chain Prediction									
SCWRL4	D('10,'02), asymm. ED25-100%	**83.04%**	**+0.57%**	**+0.61%**	**+0.67%**	**+0.40%**	**+0.11%**	*−0.08%*	N.D.
SCWRL4	D('10,'02), symm., ED0-100%	**79.33%**	**+0.55%**	**+0.59%**	**+0.64%**	**+0.39%**	**+0.11%**	*−0.11%*	N.D.
Rosetta *FastRelax*	D('10,'02), symm., ED0-100%	**72.82%**	**+0.48%**	**+0.21%**	**+0.10%**	*−0.09%*	*−1.04%*	*−1.44%*	*−1.45%*
Rosetta *ClassicRelax*	D('10,'02), symm., ED0-100%	**76.12%**	**+0.21%**	**+0.13%**	**0.00%**	*−0.12%*	*−0.81%*	*−1.12%*	*−1.57%*
Rmsd Differences									
Rosetta FastRelax:	D('10,'02)/'02 (backbone)	**1.112**	**–2.23%**	**–0.37%**	**–0.19%**	*+0.04%*	*+0.67%*	*+1.36%*	*+0.63%*
	D('10,'02)/'02 (all atoms)	**1.596**	**–1.83%**	**–0.49%**	**–0.58%**	**–0.19%**	*+0.76%*	*+1.22%*	*+1.01%*
Rosetta ClassicRelax:	D('10,'02)/'02 (backbone)	**1.081**	**–1.21%**	**–0.67%**	**–2.06%**	**–1.35%**	*+0.64%*	*+2.41%*	*+0.18%*
	D('10,'02)/'02 (all atoms)	**1.517**	**–0.02%**	*+0.28%*	**–0.76%**	**–0.21%**	*+1.36%*	*+2.36%*	*+1.91%*
TotalScoreMinusDun									
Rosetta FastRelax:	D('10,'02)	–382.28	1.783	–0.004	–1.035	–2.436	–4.884	–5.645	–1.965
Rosetta ClassicRelax:	D('10,'02)	–379.00	0.871	–0.692	–1.548	–2.685	–5.068	–5.504	–1.970

2010 library names are listed in the first row. 2009it10 is a modified version of a developmental rotamer library created by using similar methods (with some important differences) in 2008. It is distributed with Rosetta3 and was recently described by Song et al. (2011). For side-chain accuracy the absolute average percent accuracy is given for the 2002 library, and the differences from those values are given for the other libraries (2010 library–2002). For rmsd differences the mean rmsd in angstroms (Å) is given for the 2002 library, and the percent differences from 2002 are given for the 2010 libraries. For TotalScoreMinusDun, the mean values are given for the 2002 library, and the differences (in Rosetta score units) are given for the 2010 libraries. For side-chain accuracy, "symm" indicates that Asn, His, and Gln terminal dihedrals were treated as symmetric, whereas "asymm" indicates that they are treated like other dihedral angles. ED25-100% indicates that only side chains with electron density in the 25th–100th percentile were included in the accuracy assessment. ED0-100% means all side chains were included. Better numbers are in bold type (higher side-chain accuracy, lower rmsd values), whereas worse numbers are in italic type. N.D., not done. See also Figure S3.

performance better than 2002 for all residues types. Several dihedral angles have strong improvements in prediction rates, for example Trp χ_2 +6%, Gln χ_3 +4%, Phe χ_2 +3%, Glu χ_3 +3%, Ser χ_1 +1%, Met χ_3 +2%, Arg χ_3 and χ_4 +2%, and Tyr χ_2 +2% and Trp χ_1 +1%.

To create smoother rotamer libraries from the 2010 data set, we determined lower κ's (smoother functions) by finding the κ that had a lower value of the objective function by some percentage of its range (i.e., the maximum value minus the minimum value over all κ; see Figure S2 for an example). For SCWRL4 the best 2010 library is the one with the 5% step down in the objective functions from the optimal κ values. Increased smoothness (step downs of 10%, 20%, 25%) or reduced smoothness (2% or fully optimized) produces slightly lower prediction rates as shown in Table 1. For a more stringent definition of correct χ angles, within 10°, SCWRL4 demonstrates more improvement for 2010 versus 2002, a total of +1.1% (data not shown).

Because the new rotamer libraries were developed in part to improve Rosetta performance when backbone flexibility is modeled, we tested Rosetta's energy minimization protocols with the various rotamer libraries. After fitting the structures with standard bond lengths and bond angles, we separately ran two types of minimization tests on: FastRelax and ClassicRelax on the idealized structures generating 100 decoys for each. The FastRelax protocol (Tyka et al., 2011) consists of five rounds of the following: multiplying the repulsive van der Waals parameters by a scale factor C ($0 < C \leq 1$); Monte Carlo simulated-annealing repacking of side chains using the rotamer library (replacing all side chains with random rotamers, several times over, with Metropolis criterion acceptance); and then continuous energy minimization of the backbone and side chains. The factor is ramped up from 0.02 to 1.0 over four steps in each round. The lowest energy structure when the scale factor is 1.0 is saved as a decoy. The ClassicRelax protocol (Bradley et al., 2005) consists of many rounds of small backbone perturbation moves (2°–3° in ϕ and ψ) and complete side-chain repacking, followed by backbone and side-chain continuous energy minimization. The *FastRelax* protocol is the one currently recommended for high-resolution refinement in Rosetta, but we decided to test the older protocol as well to see if it behaved differently. The results are shown in Table 2.

The goal of these calculations is to perturb the backbone and side chains from the native structure and to determine whether the energy function minimization is able to bring or keep the structure as close to native as possible, as measured by backbone and full-atom rmsd values. For Rosetta, FastRelax we gained a 2.2% and 1.8% improvement for the optimized 2010 library relative to 2002 for the average backbone and full-atom rmsd values, respectively. For ClassicRelax, we achieved the best results with the smoother 5% step-down rotamer library. For this library the backbone and full-atom rmsd values from native are 2.1% and 0.8% lower than the results with the 2002 rotamer library, respectively.

The FastRelax decoys achieved the best side-chain accuracies with the optimized 2010 library compared to the 2010 libraries with additional smoothing. For cutoffs for correct predictions of 40° and 10°, the absolute average accuracies over all dihedral angles were 73.3% and 56.4%, which is an improvement of 0.5% and 1.0% when comparing to the 2002 library, respectively. The ClassicRelax decoys also achieve the best side-chain accuracies with the optimized 2010 library, with average absolute accuracies of 76.3% (40°) and 58.9% (10°). SCWRL4 with crystal symmetry but without removing side chains in the bottom 25th percentile achieves an average absolute accuracy of 80.0% (40°) and 57.9% (10°) with the 5% step-down library. The crystal symmetry is responsible for about a 2% increase in average absolute accuracy (Krivov et al., 2009).

Note that in these calculations, neither SCWRL4 nor Rosetta has been optimized to work with the 2010 libraries. The SCWRL4 calculations used constant parameters for all residue types and all rotamer libraries. The distributed version of SCWRL4, by contrast, has optimized values for several parameters for each residue type. The Rosetta calculations used the standard ''score12'' scoring function, except for the different rotamer libraries. Song et al. (2011) have recently reported an optimization of Rosetta's energy function for an earlier version of the rotamer libraries described here. They modified the rotamer library to compensate for doubly counted interactions such as side-chain/backbone hydrogen bonding and steric interactions. We tested one version of this rotamer library distributed with Rosetta3.1, ''2009it10''; the results are shown in the last column of Table 2. Its side-chain and rmsd performances are worse than both the 2002 library and the fully optimized 5% and 10% step-down libraries presented here.

The backbone-dependent rotamer library is one component (designated ''fa_dun'' in Rosetta output) of several in the Rosetta scoring function, which includes repulsive and attractive van der Waals interactions, Ramachandran energies, solvation terms, and hydrogen bonding. We analyzed the scoring function values for the decoys generated with the two relax protocols and the various rotamer libraries, shown at the bottom of Table 2. As the smoothness is increased, the nonside-chain energy terms (''TotalScoreMinusDun'') optimized to lower values. This may be due to flatness of the smoother rotamer libraries, although the dynamic range of the smoother libraries is not significantly less than the fully optimized rotamer library.

One feature of the new rotamer libraries that improves the results of Rosetta is the nature of the nonrotameric degrees of freedom. For the 2002 library (''dun02'' in Rosetta protocols), the nonrotameric degrees of freedom had between two (Phe, Tyr) and six (Asn) bins for rotamer probabilities, means, and standard deviations. In Rosetta, when the 2002 library is used, a harmonic energy term is applied to these mean values with a force constant inversely related to the standard deviation. When the developmental version of the rotamer libraries described here was implemented in Rosetta3 (''dun08'' flag in Rosetta) (Leaver-Fay et al., 2011), Rosetta was modified to use the continuous probability estimates for the nonrotameric degrees of freedom. Thus, these dihedral angles are free to change over a wide range in the smooth, backbone-dependent potentials, as shown for Gln in Figure 8. As a result, the output distributions of χ angles for these degrees of freedom are much closer to native structures than the results of the 2002 library, which are discretely distributed The distributions of χ_2 for the decoys generated by *FastRelax* for the 2002 and optimized 2010 libraries are shown in Figure S3. The results may be compared to the backbone-independent χ_2 distributions for Asn shown in Figure 1B.

Further testing is needed of the different rotamer libraries in various protocols (ab initio structure prediction, comparative modeling, docking, protein design, etc.) to determine which is most suitable for each application. On our website, http://dunbrack.fccc.edu, we provide access to the full range of rotamer libraries described here, as well as images and movies of the distributions. For most purposes the 5% step-down library may be most appropriate because it provides a good trade-off between appropriate details and smoothness of the probability distributions.

DISCUSSION

The backbone-dependent rotamer libraries we have developed previously have found uses in many different applications in protein structure prediction (Andrusier et al., 2007; Bower et al., 1997; Hartmann et al., 2007; Krieger et al., 2009; Krivov et al., 2009; Liang and Grishin, 2002; Mendes et al., 2001; Rohl et al., 2004a; Smith et al., 2007; Zhang et al., 2004) and protein design (Calhoun et al., 2003; Dahiyat and Mayo, 1997; Kuhlman and Baker, 2000; Pokala and Handel, 2005; Saraf et al., 2006; Stiebritz and Muller, 2006). In these applications both the backbone-dependent probabilities and the backbone-dependent dihedral angles have made important contributions. Therefore, we have taken great care in producing a new backbone-dependent rotamer library, testing many different ways of estimating the probabilities and regression functions that make up the library.

A number of different technical obstacles have been overcome in developing the new rotamer library. In our previous libraries we did not use methods that reliably produced smoothly varying estimates of the rotamer probabilities and dihedral angles with backbone ϕ and ψ. The kernel density estimates and regressions used here coupled with the penalized maximum likelihood optimization of the smoothing parameters have produced smooth, reliable estimates of the library values. Filtering by electron density and AKDEs and regressions reduced the effects of outliers in Ramachandran space.

An important innovation in this rotamer library is the treatment of nonrotameric degrees of freedom. The previous model of a small number of χ angle bins for these dihedrals sometimes resulted in likely artifacts in structure prediction and design. For instance, Rosetta previously placed harmonic energy functions on each of the ''rotamers'' of χ_n, which for the amides and carboxylates in particular created potential functions with four or six minima with large energy barriers in between. However, these degrees of freedom do not fit a rotamer model of discrete side-chain conformations with relatively small dihedral angle variances. Instead, they have widely distributed densities and, especially in the case of Asp and Asn, strong backbone-dependence. In the new rotamer library, smooth densities are achieved with a novel combination of query-dependent adaptive kernels on ϕ,ψ and data-dependent adaptive kernels on the χ angles, effectively the regression of an angular density onto two angular explanatory variables.

Two other studies have presented analyses similar to that of the backbone-dependent rotamer library. Amir et al. (2008) used the data from our 2002 library (850 proteins) and cubic splines to produce both joint and conditional probability distributions of ϕ,ψ, and χ angles. Such an analysis does emphasize smoothness of the probability distributions. Harder et al. (2010) have recently developed a generative model of protein side-chain conformations called BASILISK. It generates samples of side-chain dihedral angles for given input backbone dihedral angles. It is also capable of returning a log-likelihood value for any query side-chain conformation (χ_1, χ_2, χ_3, χ_4) given a backbone conformation. Because it ties together χ angle probabilities of different residue types, it does have incorrect ordering of rotamer probabilities for some residues, such as serine, for which the g^- χ_1 rotamer is not the most common.

Neither of these programs uses the rotamer model and, thus, may not be easily incorporated into programs that utilize such models to enumerate all possible rotamers in structure prediction and design. It should be noted that our methods for determining the nonrotameric χ angle densities can be used for any of the side-chain degrees of freedom, not just the nonrotameric ones. So, for instance it is possible to create estimates (including multidimensional estimates) for $\rho(\chi|\phi,\psi)$ for rotameric degrees of freedom independent of rotamer state. Such a model would include changes in probability of the rotamers, the positions of modes in the density, as well as covariance of the dihedral angles with respect to each other and the backbone dihedral angles. We are currently exploring the utility of such probability density estimates.

We believe the new backbone-dependent rotamer library has a number of useful characteristics that will make it useful in a variety of applications in protein structure determination, prediction, and design.

EXPERIMENTAL PROCEDURES

The full methods are given in the Supplemental Experimental Procedures.

Deriving Backbone-Dependent Rotamer Probabilities from Ramachandran Densities of Each Rotamer from AKDEs

We want to determine the rotamer probabilities, $P(r|\phi,\psi,aa)$, for each amino acid type, *aa*, and each rotamer (*r*), so that:

$$\sum_r P(r|\phi,\psi,aa) = 1 \tag{2}$$

for any values of (ϕ, ψ). Using Bayes' rule (see Equation 1), these probabilities can be derived from the Ramachandran PDFs of each rotamer, $\rho(\phi,\psi|r,aa)$ and the backbone-independent frequencies of each rotamer, $P(r|aa)$. The sum in the denominator of Equation 1 is over all rotamers of a given residue type. $P(r|aa)$ can be calculated easily from the observed frequencies of each rotamer in the data set. However, to calculate accurate and smooth estimates of $P(r|\phi,\psi,aa)$, we require accurate and smooth estimates of $\rho(\phi,\psi|r,aa)$. We drop "*aa*" from the formulas below. Also, we denote probabilities with *P* and probability densities with ρ.

Smooth estimates of $\rho(\phi,\psi|r)$ can be calculated from kernel density estimates. A kernel is a nonnegative function that integrates to 1. In one dimension a kernel density estimate may be written:

$$\widehat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N} K_h(\|x - x_i\|), \tag{3}$$

where *K* is the kernel function, *N* is the number of data points, and *h* is the kernel bandwidth. For instance if the kernel is Gaussian, *h* is the square root of the variance, or σ.

Because Ramachandran probability density is defined for the backbone torsion angles ϕ and ψ as two arguments, we use a two-dimensional kernel density estimate using the von Mises distribution as the kernel. The nonadaptive or fixed-bandwidth KDE in two dimensions for Ramachandran data can be written as the sum over products of ϕ- and ψ- von Mises kernels for N_r data points of rotamer type, *r*:

$$\begin{aligned}\rho(\phi,\psi|r) &= \frac{1}{N_r}\sum_{i=1}^{N_r} K_h(\|\phi-\phi_i\|)K_h(\|\psi-\psi_i\|)\\ &= \frac{1}{4\pi^2 N_r}\sum_{i=1}^{N_r}\frac{1}{(I_0(\kappa))^2}\exp\Big(\kappa\big(\cos(\phi-\phi_i)+\cos(\psi-\psi_i)\big)\Big).\end{aligned} \tag{4}$$

In this case, $\sqrt{1/\kappa}$ defines a radius of the two-dimensional hump covering 67% of the kernel density. I_0 is the Bessel function of the first kind of order 0; it normalizes the kernels to 1. For simplicity we do not place a caret on top of kernel density or KR estimates.

To reduce the effect of outliers, we use AKDEs in which the bandwidth parameter (κ) varies across the sample data points, depending on the local density of the data (Abramson, 1982; Breiman et al., 1977). For the Ramachandran density the AKDE is:

$$\rho(\phi,\psi|r) = \frac{1}{4\pi^2 N_r}\sum_{i=1}^{N_r}\frac{1}{(I_0(\kappa/\lambda_i))^2}\exp\left(\frac{\kappa}{\lambda_i}\Big(\cos(\phi-\phi_i)+\cos(\psi-\psi_i)\Big)\right). \tag{5}$$

The adaptive parameters λ_i are based on a pilot estimate of the Ramachandran density for the residue type as a whole:

$$\lambda_i = \left(\frac{\left(\prod_{j=1}^{N}\widehat{f}(\phi_j,\psi_j)\right)^{\frac{1}{N}}}{\widehat{f}(\phi_i,\psi_i)}\right)^{\alpha} = \left(\frac{g}{\widehat{f}(\phi_i,\psi_i)}\right)^{\alpha}. \tag{6}$$

For the pilot estimate, we use the non-AKDE given in Equation 4. The factor *g* is simply the geometric mean of the pilot density estimates at the *N* data points. We use α = 1/2, a value that is commonly used to regulate the magnitude of how much sample points from the sparsely populated regions have their bandwidths expanded and how much those in the populated regions have their bandwidths shrunk relative to the geometric mean sample point (Abramson, 1982; Silverman, 1986).

We chose the parameter κ for each residue type using cross-validation of the average log likelihood of the rotamers as described in the Supplemental Experimental Procedures.

Adaptive KR for the Rotameric χ Angles and Variances

The second major component of the rotamer library consists of the backbone-dependent population means, μ, and standard deviations, σ, of the available side-chain dihedral angles (χ_1, χ_2, χ_3, and χ_4) for each rotamer of the 22 residue types. We model the regression relation between the response variable, χ and the explanatory variables (ϕ, ψ):

$$\chi_i = m(\phi_i,\psi_i|r) + \nu^{\frac{1}{2}}(\phi_i,\psi_i)\varepsilon_i, \tag{7}$$

where $m(\phi_i,\psi_i|r)$ is the unknown regression function, $\nu(\phi_i,\psi_i)$ is the variance, and ε_i are random observation errors normally distributed with a mean of zero and variance 1. Given that side chains in backbone-constrained conformations experience greater uncertainty in their χ angles, we assume the standard deviation of the observation errors varies as a function of ϕ and ψ; that is, the model is *heteroscedastic*. In this case the regression function is the conditional expectation or population mean of χ given the backbone conformation:

$$m(x,y|r) = E(\chi|\phi=x,\psi=y,r) = \mu(\chi|\phi=x,\psi=y,r) \tag{8}$$

$$\nu(x,y|r) = \mathrm{Var}(\chi|\phi=x,\psi=y,r) = \sigma^2(\chi|\phi=x,\psi=y,r) \tag{9}$$

Because we do not expect $\mu(\chi|\phi,\psi,r)$ and $\sigma^2(\chi|\phi,\psi,r)$ to vary rapidly with ϕ and ψ, we use the Nadaraya-Watson or local constant KR estimator to model them. It corresponds to a local constant or zero-order polynomial, *kernel-weighted* least-squares fit:

$$\begin{aligned}\mu(\chi|\phi,\psi,r) &= \frac{\sum_{i=1}^{N_r} K_h(\phi-\phi_i,\psi-\psi_i)\chi_i}{\sum_{i=1}^{N_r} K_h(\phi-\phi_i,\psi-\psi_i)}\\ \sigma^2(\chi|\phi,\psi,r) &= \frac{\sum_{i=1}^{N_r} K_h(\phi-\phi_i,\psi-\psi_i)(\mu(\chi|\phi_i,\psi_i,r)-\chi_i)^2}{\sum_{i=1}^{N_r} K_h(\phi-\phi_i,\psi-\psi_i)}\end{aligned}. \tag{10}$$

The appropriate adaptive kernel for regression onto the angles ϕ and ψ is again a symmetric two-dimensional von Mises kernel:

$$K_h(\phi-\phi_i,\psi-\psi_i) = \frac{1}{4\pi^2\left(I_0\left(\frac{\kappa}{\lambda_{\phi\psi}}\right)\right)^2}\exp\left(\frac{\kappa}{\lambda_{\phi\psi}}(\cos(\phi-\phi_i)+\cos(\psi-\psi_i))\right). \tag{11}$$

However, in this case we use a kernel that is adaptive based on the query point rather than the data point:

$$\lambda_{\phi\psi} = \left(\frac{\left(\prod_{j=1}^{N_r}\widehat{f}(\phi_j,\psi_j|r)\right)^{\frac{1}{N_r}}}{\widehat{f}(\phi,\psi|r)}\right)^{\frac{1}{2}} = \left(\frac{g_r}{\widehat{f}(\phi,\psi|r)}\right)^{\frac{1}{2}}. \tag{12}$$

This estimator can adapt to the density of sample points, taking a larger bandwidth where points are sparse. It can adapt to changes in residual variance in case of heteroscedacity, smoothing more where residual variance is high. The estimator can adapt to the structure of the regression function, smoothing more in flat parts of the surface and less in steeper parts. This leads to improved smoothness, which is expected to lead to better side-chain modeling.

Backbone-Dependent Modeling of Nonrotameric Degrees of Freedom

The terminal dihedral angle for certain side-chain types is not well described as a rotamer. These include the terminal degrees of freedom of Asn, Asp, Glu, and Gln. The aromatic residues, Phe, Tyr, His, and Trp, also have more broadly distributed χ_2 angles than rotameric degrees of freedom, although not to the same extent as the amide and carboxylate groups. We model the terminal dihedral angle of side chains with nonrotameric degrees of freedom, χ_n, as continuous PDFs as a function of the backbone conformation, (ϕ, ψ), $\rho(\chi_n|\phi,\psi,r_{-n})$, where r_{-n} denotes the rotamer of the rotameric degrees of freedom (χ_1 for Asn, Asp, and the aromatics; χ_1, χ_2 for Gln and Glu), such that:

$$\int_{\chi_n}\rho(\chi_n'|\phi,\psi,r_{-n})d\chi_n' = 1. \tag{13}$$

With $\rho(\chi_n|\phi,\psi,r_{-n})$ in hand on a fine grid of χ_n values, we can calculate binned probabilities at any desired resolution, 5°, 10°, or 30° for instance.

Modeling $\rho(\chi_n|\phi,\psi,r_{-n})$ is effectively the regression of a PDF onto the explanatory variables ϕ,ψ; that is, we want a separate $\rho(\chi_n)$ for any ϕ,ψ. We have calculated Ramachandran map PDFs with data point-adaptive kernels, while we have found that regressions were better produced using query point-adaptive kernels. We achieve the backbone-dependent nonrotameric χ_n density modeling by computing the backbone-dependent KR of the χ_n densities, each of which is based on an individual χ_n data point taken from the input sample:

$$\rho(\chi_n|\phi,\psi,r_{-n}) = \frac{\sum_{i=1}^{N_r}K_{h(\phi,\psi)}(\phi-\phi_i,\psi-\psi_i)K_{h(\chi_i)}(\chi_n-\chi_i)}{\sum_{i=1}^{N_r}K_{h(\phi,\psi)}(\phi-\phi_i,\psi-\psi_i)}, \tag{14}$$

where χ_i are the data points of χ_n, and $K_{\phi,\psi}(\phi-\phi_i,\psi-\psi_i)$ is the query-adaptive kernel with the same expression as in Equation 11, and its κ is the von Mises concentration parameter in the (ϕ, ψ) space. We take the kernels on χ to be one-dimensional von Mises functions (Equation 6) centered on χ_i taken from the data sample:

$$K_{h(\chi_i)}(\chi_n-\chi_i) = \frac{1}{2\pi I_0(\kappa_{1d}/\lambda_i)}\exp\left(\frac{\kappa_{1d}}{\lambda_i}\cos(\chi_n-\chi_i)\right). \tag{15}$$

The concentration parameter, κ_{1d}, sets the overall bandwidth in the χ_n space and is chosen independently from its counterpart, the (ϕ, ψ)-space κ. λ_i are the scaling parameters calculated in the data-adaptive fashion in accordance with the one-dimensional χ_i backbone-independent density:

$$\lambda_i = \left(\frac{\left(\prod_{j=1}^{N_r}\widehat{f}_\chi(\chi_j|r_{-n})\right)^{\frac{1}{N_r}}}{\widehat{f}_\chi(\chi_i|r_{-n})}\right)^{\alpha} = \left(\frac{g_r^{1d}}{\widehat{f}_\chi(\chi_i|r_{-n})}\right)^{\alpha}, \tag{16}$$

where $\widehat{f}_\chi(\chi_n|r_{-n})$ is a χ_n pilot density estimate and $\alpha = 1/2$. The pilot density is modeled with a non-AKDE with the same concentration parameter, κ_{1d}:

$$\widehat{f}_\chi(\chi_n|r_{-n}) = \frac{1}{2\pi I_0(\kappa_{1d})N_r}\sum_{j=1}^{N_r}\exp\left(\kappa_{1d}\cos(\chi_n-\chi_j)\right) \tag{17}$$

The χ_n concentration parameters, κ_{1d}/λ_i (Equation 15) are data adaptive in order to produce a true PDF that integrates to 1. If κ_{1d}/λ_i is query adaptive, the resulting function would not integrate to 1 and would not meet the definition of a PDF (Sain, 1994).

Note that κ and κ_{1d} have different and specific values for each rotamer, r_{-n}. It is also worth pointing out that in very empty parts of the (ϕ, ψ) map where $\kappa/\lambda_{\phi\psi}\rightarrow 0$, the KR of the χ_n densities defaults to the backbone-independent density:

$$\begin{aligned}\rho(\chi_n|\phi,\psi,r_{-n}) &= \frac{\sum_{i=1}^{N_r}K_{h(\phi,\psi)}(\phi-\phi_i,\psi-\psi_i)K_{h(\chi_i)}(\chi_n-\chi_i)}{\sum_{i=1}^{N_r}K_{h(\phi,\psi)}(\phi-\phi_i,\psi-\psi_i)}\\ &= \frac{\sum_{i=1}^{N_r}Const\cdot K_{h(\chi_i)}(\chi_n-\chi_i)}{\sum_{i=1}^{N_r}Const} = \frac{1}{N_r}\sum_{i=1}^{N_r}K_{h(\chi_i)}(\chi_n-\chi_i)\equiv\rho_\chi(\chi_n|r_{-n}).\end{aligned} \tag{18}$$

Further details on optimizing the bandwidths and converting nonrotameric density into rotamer probabilities for the nonrotameric degrees of freedom are given in the Supplemental Experimental Procedures.

Availability

The 2010 rotamer libraries are available from http://dunbrack.fccc.edu. The website also presents additional images of the backbone-dependent probabilities, dihedral angle means, and movies of the nonrotameric probability densities.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, three tables, and one movie and can be found with this article online at doi:10.1016/j.str.2011.03.019.

ACKNOWLEDGMENTS

We thank Georgy Krivov, Benjamin Blum, and Professor Michael I. Jordan (University of California Berkeley) for helpful discussions. Brian Weitzner provided invaluable help in setting up the Rosetta calculations. This work was funded under NIH Grants P20 GM76222 and R01 GM84453 to R.L.D.

Received: October 23, 2010
Revised: March 19, 2011
Accepted: March 22, 2011
Published: June 7, 2011

REFERENCES

Abramson, I.S. (1982). On bandwidth variation in kernel estimates—a square root law. Ann. Stat. *10*, 1217–1223.

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K., and Terwilliger, T.C. (2002). PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr. D Biol. Crystallogr. *58*, 1948–1954.

Amir, E.D., Kalisman, N., and Keasar, C. (2008). Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. Proteins *72*, 62–73.

Andrusier, N., Nussinov, R., and Wolfson, H.J. (2007). FireDock: fast interaction refinement in molecular docking. Proteins *69*, 139–159.

Bower, M.J., Cohen, F.E., and Dunbrack, R.L., Jr. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J. Mol. Biol. *267*, 1268–1282.

Bradley, P., Misura, K.M., and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868–1871.

Breiman, L., Friedman, J.H., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. Technometrics *19*, 135–144.

Brockmann, M., Gasser, T., and Herrmann, E. (1993). Locally adaptive bandwidth choice for kernel regression-estimators. J. Am. Stat. Assoc. *88*, 1302–1309.

Calhoun, J.R., Kono, H., Lahr, S., Wang, W., DeGrado, W.F., and Saven, J.G. (2003). Computational design and characterization of a monomeric helical dinuclear metalloprotein. J. Mol. Biol. *334*, 1101–1115.

Canutescu, A.A., Shelenkov, A.A., and Dunbrack, R.L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. *12*, 2001–2014.

Dahiyat, B.I., and Mayo, S.L. (1997). De novo protein design: fully automated sequence selection. Science *278*, 82–87.

Davis, I.W., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2004). MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res. *32*, W615–W619.

Desmet, J., De Maeyer, M., Hazes, B., and Lasters, I. (1992). The dead-end elimination theorem and its use in protein sidechain positioning. Nature *356*, 539–542.

Dunbrack, R.L., Jr. (2002). Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. *12*, 431–440.

Dunbrack, R.L., Jr., and Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J. Mol. Biol. *230*, 543–574.

Dunbrack, R.L., Jr., and Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat. Struct. Biol. *1*, 334–340.

Dunbrack, R.L., Jr., and Cohen, F.E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. *6*, 1661–1681.

Eggermont, P.P.B., and LaRiccia, V.N. (2001). Maximum Penalized Likelihood Estimation. Volume I: Density Estimation (New York: Springer-Verlag).

Friedland, G.D., Linares, A.J., Smith, C.A., and Kortemme, T. (2008). A simple model of backbone flexibility improves modeling of side-chain conformational variability. J. Mol. Biol. *380*, 757–774.

Gordon, D.B., Hom, G.K., Mayo, S.L., and Pierce, N.A. (2003). Exact rotamer optimization for protein design. J. Comput. Chem. *24*, 232–243.

Harder, T., Boomsma, W., Paluszewski, M., Frellsen, J., Johansson, K.E., and Hamelryck, T. (2010). Beyond rotamers: a generative, probabilistic model of side chains in proteins. BMC Bioinformatics *11*, 306.

Hartmann, C., Antes, I., and Lengauer, T. (2007). IRECS: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. Protein Sci. *16*, 1294–1307.

Headd, J.J., Immormino, R.M., Keedy, D.A., Emsley, P., Richardson, D.C., and Richardson, J.S. (2009). Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place. J. Struct. Funct. Genomics *10*, 83–93.

Kleywegt, G.J., Harris, M.R., Zou, J.Y., Taylor, T.C., Wahlby, A., and Jones, T.A. (2004). The Uppsala Electron-Density Server. Acta Crystallogr. D Biol. Crystallogr. *60*, 2240–2249.

Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D., and Karplus, K. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins *77* (*Suppl 9*), 114–122.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795.

Kuhlman, B., and Baker, D. (2000). Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. USA *97*, 10383–10388.

Kuhlman, B., and Baker, D. (2004). Exploring folding free energy landscapes using computational protein design. Curr. Opin. Struct. Biol. *14*, 89–95.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. *487*, 545–574.

Liang, S., and Grishin, N.V. (2002). Side-chain modeling with an optimized scoring function. Protein Sci. *11*, 322–331.

Lovell, S.C., Word, J.M., Richardson, J.S., and Richardson, D.C. (1999). Asparagine and glutamine rotamers: B-factor cutoff and correction of amide flips yield distinct clustering. Proc. Natl. Acad. Sci. USA *96*, 400–405.

Lovell, S.C., Word, J.M., Richardson, J.S., and Richardson, D.C. (2000). The penultimate rotamer library. Proteins *40*, 389–408.

Mardia, K.V., and Jupp, P.E. (2000). Directional Statistics (London: Wiley).

Mendes, J., Nagarajaram, H.A., Soares, C.M., Blundell, T.L., and Carrondo, M.A. (2001). Incorporating knowledge-based biases into an energy-based side-chain modeling method: application to comparative modeling of protein structure. Biopolymers *59*, 72–86.

Pokala, N., and Handel, T.M. (2005). Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J. Mol. Biol. *347*, 203–227.

Rohl, C.A., Strauss, C.E., Chivian, D., and Baker, D. (2004a). Modeling structurally variable regions in homologous proteins with Rosetta. Proteins *55*, 656–677.

Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. (2004b). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66–93.

Sain, S.R. (1994). Adaptive kernel density estimation. PhD dissertation, Department of Statistics, Rice University, Houston, Texas.

Saraf, M.C., Moore, G.L., Goodey, N.M., Cao, V.Y., Benkovic, S.J., and Maranas, C.D. (2006). IPRO: an iterative computational protein library redesign and optimization procedure. Biophys. J. *90*, 4167–4180.

Shapovalov, M.V., and Dunbrack, R.L., Jr. (2007). Statistical and conformational analysis of the electron density of protein side chains. Proteins *66*, 279–303.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis (New York: Chapman & Hall).

Smith, C.A., and Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J. Mol. Biol. *380*, 742–756.

Smith, R.E., Lovell, S.C., Burke, D.F., Montalvao, R.W., and Blundell, T.L. (2007). Andante: reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities. Bioinformatics *23*, 1099–1105.

Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J., and Baker, D. (2011). Structure-guided forcefield optimization. Proteins, in press. Published online February 15, 2011.

Stiebritz, M.T., and Muller, Y.A. (2006). MUMBO: a protein-design approach to crystallographic model building and refinement. Acta Crystallogr. D Biol. Crystallogr. *62*, 648–658.

Tyka, M.D., Keedy, D.A., Andre, I., Dimaio, F., Song, Y., Richardson, D.C., Richardson, J.S., and Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol. *405*, 607–618.

Wang, G., and Dunbrack, R.L., Jr. (2005). PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res. *33*, W94–W98.

Wiberg, K.B., and Murcko, M.A. (1988). Rotational barriers. 2. Energies of alkane rotamers. An examination of gauche interactions. J. Am. Chem. Soc. *110*, 8029–8038.

Word, J.M., Lovell, S.C., Richardson, J.S., and Richardson, D.C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. *285*, 1735–1747.

Zhang, C., Liu, S., and Zhou, Y. (2004). Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci. *13*, 391–399.

SUPPLEMENTARY MATERIAL

A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions

M. V. Shapovalov and R. L. Dunbrack, Jr.

Institute for Cancer Research
Fox Chase Cancer Center
333 Cottman Avenue
Philadelphia PA 19111
USA

CONTENTS

Table S1, related to Figure 1. 2010 Rotamer library data and rotamer definitions

N	Res type	Count	%	# χ angles	{rotam. χ}	# rotamers (rotam. χ)	Non-rotameric χ: interval	# rotamer (non-rotam χ)
1	**PRO**	33,053	5.7	1	$\{\chi_1\}^\dagger$	2	–	–
1A	**TPR**	31,317	*(95)	1	$\{\chi_1\}^\dagger$	2	–	–
1B	**CPR**	1,736	*(5)	1	$\{\chi_1\}^\dagger$	2	–	–
2	**SER**	41,237	7.1	1	$\{\chi_1\}$	3	–	–
3	**VAL**	50,999	8.8	1	$\{\chi_1\}$	3	–	–
4	**THR**	38,679	6.7	1	$\{\chi_1\}$	3	–	–
5	**CYS**	9,086	1.6	1	$\{\chi_1\}$	3	–	–
5A	**CYH**	6,818	*(75)	1	$\{\chi_1\}$	3	–	–
5B	**CYD**	2,268	*(25)	1	$\{\chi_1\}$	3	–	–
6	**ILE**	40,117	6.9	2	$\{\chi_1, \chi_2\}$	9	–	–
7	**LEU**	64,329	11.1	2	$\{\chi_1, \chi_2\}$	9	–	–
8	**MET**	12,240	2.1	3	$\{\chi_1, \chi_2, \chi_3\}$	27	–	–
9	**ARG**	34,910	6.0	4	$\{\chi_1, \chi_2, \chi_3, \chi_4\}$	81	–	–
10	**LYS**	37,268	6.4	4	$\{\chi_1, \chi_2, \chi_3, \chi_4\}$	81	–	–
11	**ASN**	28,607	4.9	2	$\{\chi_1\}$	3	χ_2: [-180°, 180°)	12
12	**TRP**	10,571	1.8	2	$\{\chi_1\}$	3	χ_2: [-180°, 180°)	12
13	**HIS**	15,014	2.6	2	$\{\chi_1\}$	3	χ_2: [-180°, 180°)	12
14	**ASP**	41,769	7.2	2	$\{\chi_1\}$	3	χ_2: [-90°, 90°)	6
15	**PHE**	28,900	5.0	2	$\{\chi_1\}$	3	χ_2: [-30°, 150°)	6
16	**TYR**	25,490	4.4	2	$\{\chi_1\}$	3	χ_2: [-30°, 150°)	6
17	**GLU**	45,206	7.8	3	$\{\chi_1, \chi_2\}$	9	χ_3: [-90°, 90°)	6
18	**GLN**	23,653	4.1	3	$\{\chi_1, \chi_2\}$	9	χ_3: [-180°, 180°)	12
	Total	**581,128**	**100**					

*TPR (trans proline), CPR (cis proline) and CYH (non-disulfide-bonded Cys), CYD (disulfide-bonded Cys) percentages are calculated relative to the total number of PRO and CYS respectively. Each rotameric degree of freedom, χ has the rotamer definitions: g^+ = [0°,120°), t = [120°,240°), g^- = [240°,360°). †PRO, CPR, and TPR have only g^+ and g^- rotamers.

Table S2. Leucine rotamer counts before and after reassigning flipped χ_2 values

Rotamer	# before fix	# after fix	Diff	P_bef(*r*) %	P_aft(*r*) %	P_aft / P_bef (%)
g+ g+	523	523	0	0.81	0.81	100.0
g+ t	285	285	0	0.44	0.44	100.0
g+ g-	10	10	0	0.015	0.015	100.0
t g+	19,654	19,906	**+**252	30.32	30.71	101.3
t t	1,711	1,459	**-**252	2.64	2.25	85.3
t g-	226	226	0	0.35	0.35	100.0
g- g+	2,118	1,879	**-**239	3.27	2.90	88.7
g- t	39,890	40,129	**+**239	61.54	61.91	100.6
g- g-	403	403	0	0.62	0.62	100.0
Total	64,820	64,820	0	100.0	100.0	

Leucine <*t t*> and <*g*- *g*+> rotamers with certain dihedral angle combinations of χ_1 and χ_2 are reassigned to the <*t g*+> and <*g*- *t*> rotamers, which result by rotating χ_2 around by 180°. The residues that are "fixed" are removed from the data set and only used to change the values of P(*r*) used in Equation S2, where P_aft(*r*) is used. The populations of <*t t*> and <*g*- *g*+> decrease by 14.7% and 11.3% of their original populations respectively. The populations of the <*t g*+> and <*g*- *t*> rotamers increase only by 1.3% and 0.6%.

Table S3, related to Figures 4, 6, and 7. Optimal and 5%↓ κ values for rotamer probabilities, 5%↓ bandwidth radius for rotamer probabilities and regressions of rotameric χ and non-rotameric χ_n densities.

Res	5%↓κ $P(r)$	5%↓-rad $P(r)$	Opt-κ $P(r)$	Opt-rad $P(r)$	5%↓-rad for $\rho_\chi(\chi^{nonRot} \mid r)$	5%↓-rad χ_1 regr	5%↓-rad χ_2 regr	5%↓-rad χ_3 regr	5%↓-rad χ_4 regr
CYD	28	11°	61	7°	–	11°-31°			
MET	39	9°	76	7°	–	9°-∞	7°-∞	10°-∞	
ARG	40	9°	80	6°	–	5°-∞	5°-∞	6°-∞	6°-∞
LYS	43	9°	83	6°	–	5°-∞	6°-∞	6°-∞	7°-∞
ASP	45	9°	282	3°	22°-27°	6°-9°	7°-9°		
CYH	46	8°	150	5°	–	7°-10°			
CYS	47	8°	150	5°	–	7°-11°			
TRP	57	8°	176	4°	6°-11°	7°-11°	7°-14°		
ASN	64	7°	239	4°	14°-27°	8°-9°	10°-11°		
CPR	67	7°	136	5°	-	8°-12°	9°-13°	7-∞	
GLN	67	7°	184	4°	11°-31°	5°-24°	5°-∞	7°-46°	
THR	85	6°	308	3°	–	4°-7°			
PHE	95	6°	301	3°	15°-23°	4°-10°	6°-13°		
HIS	95	6°	203	4°	11°-13°	7°-12°	8°-12°		
SER	102	6°	309	3°	–	4°-10°			
TYR	103	6°	294	3°	16°-23°	4°-10°	6°-13°		
GLU	118	5°	249	4°	25°-35°	5°-28°	5°-36°	5°-36°	
VAL	120	5°	576	2°	–	4°-6°			
ILE	122	5°	402	3°	–	4°-∞	5°-∞		
LEU	125	5°	309	3°	–	4°-∞	4°-∞		
PRO	139	5°	868	2°	–	3°-4°	3°-4	4°-4°	
TPR	142	5°	893	2°	–	3°-4°	3°-4	4°-4°	

"rad" indicates radius about ϕ,ψ or χ point encompassing 67% of density of von Mises kernel function. Residue types are arranged according to their ascending κ values for rotamer probabilities.

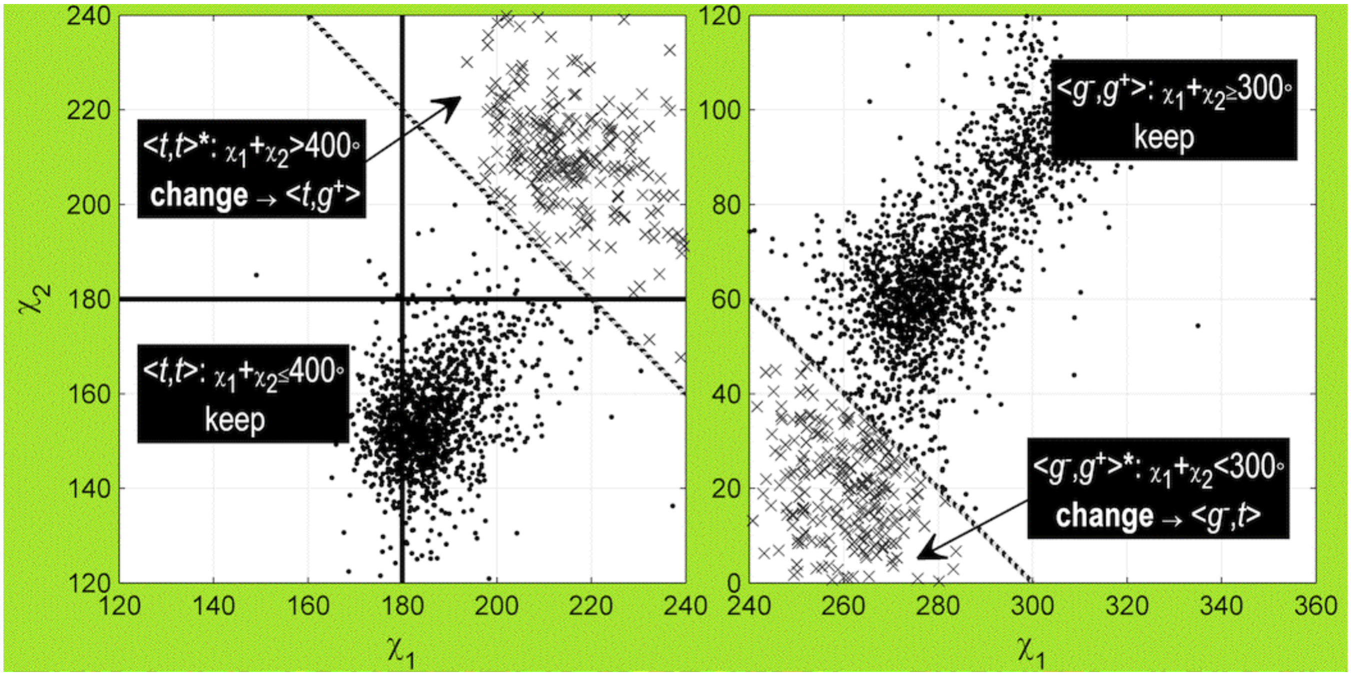

Figure S1. Criteria for identifying Leu residues that are probably incorrectly modeled in X-ray structures.

On the left, Leu <*t*, *t*>* rotamers having $\chi_1 + \chi_2 > 400°$ and located above the line are likely to belong to <*t*, g^+> rotamer. On the right, < g^-, g^+>* rotamers having $\chi_1 + \chi_2 < 300°$ and below the line are likely to belong to the <g^-, *t*> rotamer.

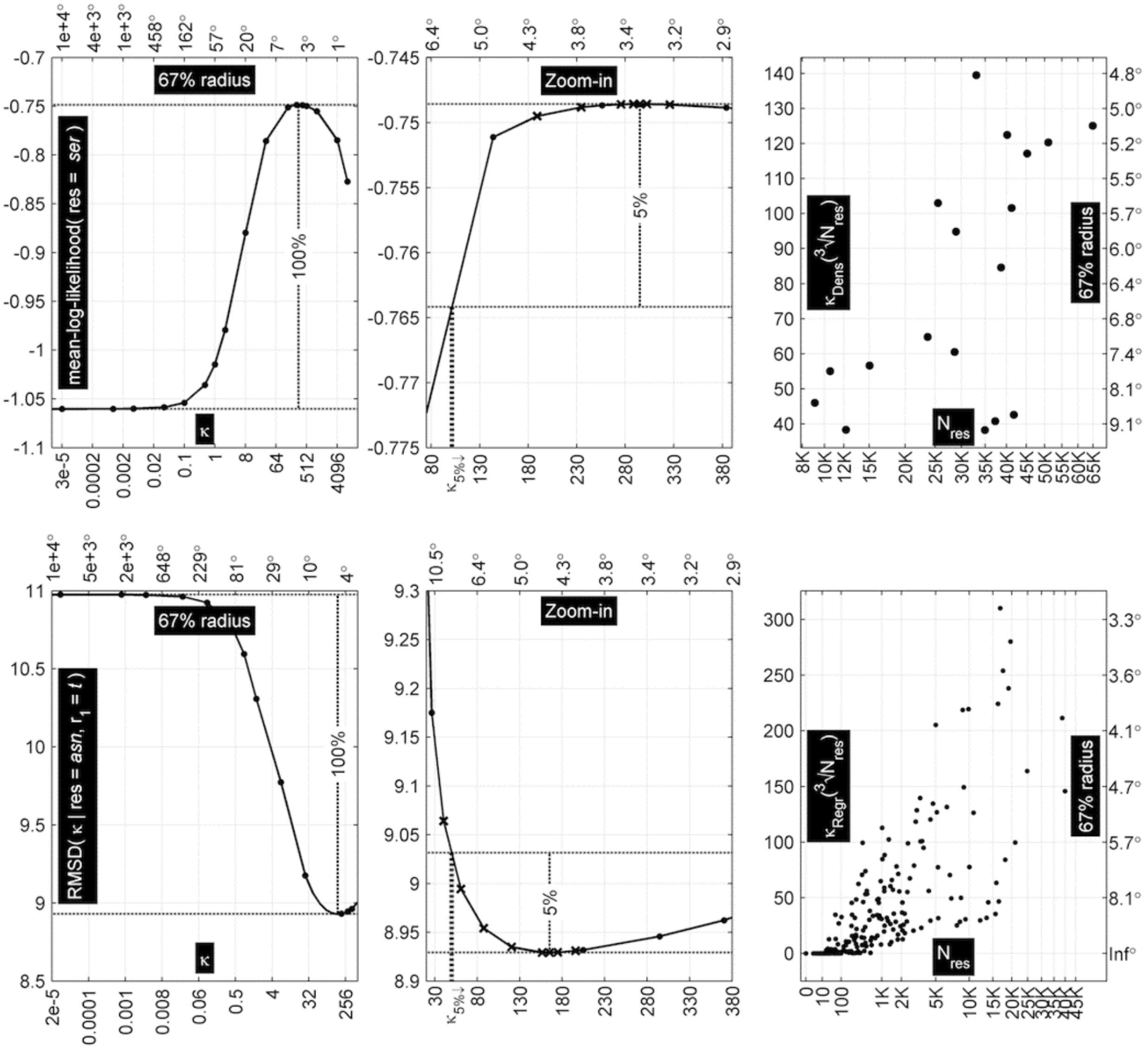

Figure S2, related to Figures 4 and 6. Optimization of adaptive kernel bandwidths for rotamer probability and rotamer χ means.

The *top* row shows the optimization of the 2010 rotamer probabilities. The *bottom* row addresses the 2010 rotameric χ mean and standard deviation optimization. *Top-left* and *top-middle*: mean log-likelihood of Ser rotamers vs. geometric-mean κ. *Bottom-left* and *bottom-middle*: RMSD of Asn *trans* χ_1 dihedral angles and their query-adaptive kernel regression of the mean vs. geometric-mean κ. On these four figures there are two sets of tick marks for the x-axis, one for the concentration parameter, κ (at the bottom of each plot) and one for the 67% bandwidth radius in degrees, *R* (at the top of each plot). The optimal κ values are indicated with a vertical dashed line with a 100% label next to it. The 5%-stepdown-from-optimum κ values are also shown. The period markers on the optimization curves indicate iterations from the first, initial stage of the optimization algorithm tailored to find an optimal value with a minimum number of iterations. The cross markers are iterations from the second, refinement optimization stage. *Right*: scatter plot of the 5%-stepdown-from-optimum κ values vs. the number of data points. *Right-top* is for rotamer probability optimization; a single point corresponds to an individual residue type. The number of data points is the number of residues of this type in the set. *Right-bottom* is for rotameric χ_1 means and standard deviations; each point depicts each rotamer type of each residue type. For example, for Arg there are 81 χ_1 κ's on this subplot.

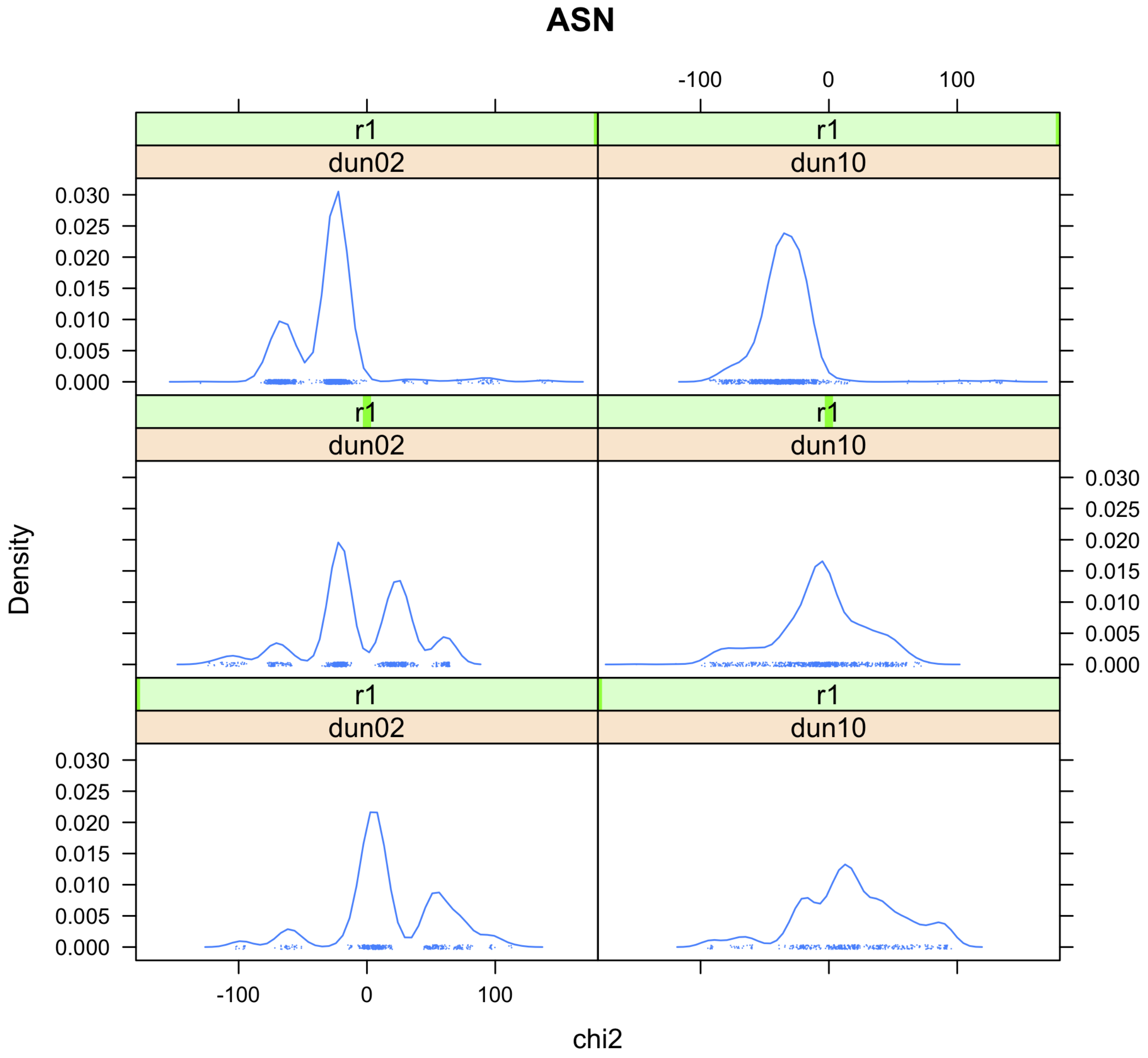

Figure S3. Comparison of Asn χ_2 distributions from Rosetta *FastRelax* decoys for the 2002 and 2010 libraries.

Density probability estimates for Asn χ_2 for each χ_1 rotamer (g-, trans, g+ from top to bottom) in FastRelax decoys generated by the 2002 rotamer library (left column) and the 2010 optimized library (left column). The 2002 libraries are implemented with a harmonic force constant at the χ_2 dihedral values with the force constants based on the standard deviations given in the library, which contains six bins for Asn χ_2. The 2010 library contains a continuous probability distribution (evaluated in 10° intervals), and the dihedral angles are allowed to minimize freely on these potential surfaces in the *FastRelax* protocol. The χ_2 distributions over the 2010-library decoys for the 50-protein test set thus resemble the backbone-independent distributions shown in Figure 1B of the main text, while the 2002-library decoys do not.

SUPPLEMENTARY METHODS

Dataset preparation

Step 1: List of proteins with available electron density maps. We first determined the full list of PDB entries for which we could obtain electron densities from the Uppsala Electron Density Server (EDS) (Kleywegt et al., 2004). We downloaded 36,125 *sigmaA*-weighted $2mF_{obs} - DF_{calc}$ electron density maps on January 15, 2010. Among them 98% (35,394) contained protein structures.

Step 2: Selecting unique chains. This list of 35,394 entries containing 85,344 chains was then filtered down by the PISCES server (Wang and Dunbrack, 2003, 2005) to obtain a list of 4,018 protein chains from 3,877 entries with resolution better than or equal to 1.8 Å, maximum R value of 0.22, chain sequence length of at least 50 residues, and mutual sequence identity of the chains of 50% or less. The remediated entries were downloaded with *BioDownloader* (Shapovalov et al., 2007) from the PDB (Berman et al., 2000). To choose a threshold for the R-factor, we collected statistics on R-factor and R-free vs resolution for all X-ray structures in the PDB. We removed obvious outliers due to reporting errors. With the remaining data we performed a seven-order polynomial regression of R-factor and R-free on resolution. On the plotted curves (not shown) at 1.8 Å resolution the mean R-factor and R-free are 0.19 and 0.23 respectively. We added one standard deviation to each of these values to get R-factor and F-free cutoffs of 0.22 and 0.27 respectively. PISCES only uses the R-factor cutoff.

Step 3: Fixing of Asn, His and Gln. We used the software program, *SIOCS* (Heisen and Sheldrick, unpublished) available with SHELX (Sheldrick, 2008). *SIOCS* relies on hydrogen bonding and crystal contacts to determine whether Asn, His and Gln side chains are correctly placed in crystal structures. It leaves the correct side chains unmodified ("*Kept*"). *SIOCS* flips the terminal dihedrals by 180° of the side chains it identifies as misplaced ("*Flipped*"). It also provides a confidence level for either of these two actions. It failed with a segmentation error on 32 (0.8%) of the 3,877 PDB entries, reducing the data set to 3,845 entries with 3,985 unique chains and 936,489 residues.

Step 4: Filtering residues with defined backbone and side chain. We used only residues without missing Cartesian coordinates. We removed the first and last residue of each chain in order to have both backbone torsion angles (ϕ, ψ) defined and any residues next to residues with missing coordinates. We observed several bond length artifacts in the initial dataset, such as a serine having a 2.1 Å CB-OG bond. We calculated side-chain bond lengths and eliminated residues having bond lengths longer than 2 Å. After applying these quality-control filters, the number of residues decreased by 3% to 909,576 from 936,489.

Step 5: Applying electron density filter. We computed electron densities at the atom positions, $\rho_{point}\left(\vec{r}_{atom}\right)$, using the methods described earlier (Shapovalov and Dunbrack, 2007), and used the geometric mean of the atom electron densities of each residue (residue confidence level, $\rho_{point}^{residue}$ (Shapovalov and Dunbrack, 2007), as a quality control to remove disordered residues from further analysis. We individually calculated percentiles of electron density for each residue type from all the chains in the 3,845 entries prior to the filters applied in Step 4, and used the 25th percentile value for each residue type to remove disordered side chains. The electron density filter reduced the total number of the residues by an additional 206,294 or 22.0% to 703,282. This filter would have eliminated most but not all of the 3% of residues flagged by other filters in Step 4.

Step 6: Selecting *Flipped* and *Kept* Asn, His and Gln with high confidence level only. We resolved ambiguity in the flip state of Asn and His χ_2 and Gln χ_3 by leaving *Kept* and *Flipped* conformations only with a high confidence level assigned by *SIOCS*. For each action SIOCS provides a numerical confidence level and one of three labels: *unsure*, *probable* and *clear*. We only included Asn, His and Gln with the *clear* confidence level, whether *Kept* or *Flipped*; others were discarded. 96% of Asn, 94% of Gln, and 98% of His were *clear*. Among the *clear* conformations, 15% of Asn, 17% of Gln, and 15% of His had *Flipped* states.

Step 7: Removing some Leu rotamers which are physically impossible. Some Leu <*t*, *t*> and <g^-, g^+> rotamers are likely to be incorrectly modeled into the density by shifting Cγ slightly and reversing the labels on Cδ1 and Cδ2 (Lovell et al., 2000). We identified such probable flips as those Leu rotamers (Figure S1), denoted <*t*, *t*>*, having $\chi_1 + \chi_2 > 400°$ (χ_1 and χ_2 in the range 120°-240°) and those Leu rotamers, denoted < g^-, g^+>*, having $\chi_1 + \chi_2 < 300°$ (χ_1 in the range 240°-360° and χ_2 in the range 0°-120°). The <*t*, *t*>* are likely to belong to the populated <*t*, g^+> rotamer, and the <g^-, g^+>* are likely to belong to the populated <g^-, *t*> rotamer. We excluded these unreliable Leu conformations from the dataset, so that they were not used for rotamer probability or side-chain dihedral angle calculations. However we adjusted the Leu backbone-independent rotamer probabilities, *P*(*r*) (used in Eq. S2) to account for the misfit $<t,t>^* \rightarrow <t,g^+>$ and $<g^-,g^+>^* \rightarrow <g^-,t>$ rotamers. The values of *P*(*r*) before and after the fix are given Table S1. A total of 14.7% and 11.3% of the original $<t,t>$ and $<g^-,g^+>$ residues were moved to $<t,g^+>$ and $<g^-,t>$ rotamer counts respectively. Because the latter are highly populated their populations change by only 1.3% and 0.6% respectively. These corrections should lead to a more reliable rotamer library for Leu.

Step 8: Splitting cysteine into disulfide-bonded and non-disulfide-bonded. We identified cysteines in disulfide bonds from the SSBOND records in the PDB files. We divided CYS into two categories: disulfide-bonded (*CYD*) and *not* disulfide-bonded (*CYH*). The *CYH* and *CYD* proportions are 75% and 25% (Table S2). In the 2002 rotamer library we had a single residue type, *CYS* including both *CYH* and *CYD*. In the 2010 rotamer library, we calculate statistical results for both CYH and CYD separately as well as all CYS.

Step 9: Distinguishing between *trans*- and *cis*- prolines. In protein structures about 5% of proline residues are preceded by a *cis* peptide bond. We computed a rotamer library for prolines in three categories (Table S2): cis-proline or *CPR* (-90° < ω < 90°), trans proline or *TPR* (90°≤ ω ≤ 270°) and all prolines, *PRO* (*TPR* + *CPR*).

Step 10: Excluding Mse, Ala and Gly. There are 16% of selenium methionine residues (Mse) in the dataset. Having a reasonable total number of data points for the unmodified methionine (Met) we excised the selenium methionines from our dataset. Ala and Gly do not have any degrees of freedom in their side chains and therefore were removed from the dataset for the rotamer library calculations.

We provide rotamer libraries for 22 amino acid types: *ARG*, *ASN*, *ASP*, *CPR*, *CYD*, *CYH*, *CYS*, *GLN*, *GLU*, *HIS*, *ILE*, *LEU*, *LYS*, *MET*, *PHE*, *PRO*, *SER*, *THR*, *TPR*, *TRP*, *TYR*, *VAL*. The total number of unique residues in these sets is 581,128. The individual counts of residues used in the 2010 rotamer library analysis for each of 22 residue types is reported in Table S2.

Deriving backbone-dependent rotamer probabilities from Ramachandran densities of each rotamer using Bayes' rule

We want to determine the rotamer probabilities, $P\left(r\middle|\phi,\psi,aa\right)$, for each amino acid type, *aa*, and each rotamer *r*, so that:

$$\sum_r P\left(r\middle|\phi,\psi,aa\right)=1 \qquad \text{(S1)}$$

for values of (ϕ, ψ) on a 10°x10° grid. Using Bayes' rule, these probabilities can be derived from the Ramachandran probability density functions of each rotamer, $\rho\left(\phi,\psi\middle|r,aa\right)$ and the backbone-independent frequencies of each rotamer, $P\left(r\middle|aa\right)$:

$$P\left(r\middle|\phi,\psi,aa\right)=\frac{\rho\left(\phi,\psi\middle|r,aa\right)P\left(r\middle|aa\right)}{\sum_{r'}\rho\left(\phi,\psi\middle|r',aa\right)P\left(r'\middle|aa\right)} \qquad \text{(S2)}$$

The sum in the denominator of Equation S2 is over all rotamers of a given residue type. $P\left(r\middle|aa\right)$ can be calculated easily from the observed frequencies of each rotamer in the dataset. However, to calculate accurate and smooth estimates of $P\left(r\middle|\phi,\psi,aa\right)$, we require accurate and smooth estimates of $\rho\left(\phi,\psi\middle|r,aa\right)$. In order to make the subsequent formulas easier to read we drop "*aa*" from the formulas. Also we denote probabilities with *P* and probability densities with ρ.

Kernel Density Estimates (KDE) and adaptive kernel density estimates (AKDE)

In the simplest case, a one-dimensional distribution of some variable can be modeled from a random data sample $\{x_i\}$, where *i* = 1 .. *N* with a fixed-bandwidth or non-adaptive kernel density estimator (non-adaptive KDE) (Parzen, 1962):

$$\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N} K_h\left(\left\|x - x_i\right\|\right) \qquad \text{(S3)}$$

where $\left\|x - x_i\right\|$ is the metric distance between the estimation or query point *x* and the data point x_i. K_h is a symmetric, nonnegative function, centered at 0 that integrates to 1. This function is referred to as a kernel. It meets the definition of a probability density function (PDF).

The kernel function has an important parameter, *h*, which acts to control the level of smoothing. It establishes the width of the kernel: the larger *h* is the wider the kernel is; the smaller *h* is the narrower the kernel. For instance, the Gaussian kernel with parameter *h* is:

$$K_h\left(x - x_i\right) = \frac{1}{\sqrt{2\pi}h}\exp\left(-\frac{\left(x - x_i\right)^2}{2h^2}\right) \qquad \text{(S4)}$$

When *h* is held constant across all sample data points as in Equation S3, the estimator is referred to as a fixed-bandwidth or non-adaptive kernel density estimate. Visually interpreting the formula, we can imagine that a separate kernel curve with the same width and of the same shape is placed on top of each sample data point. When we determine a value of the density estimator at a location, *x*, we sum the heights of the kernel curve tails of all sample points at this location and then average them. There are a number of different kernels for data on the line. Fix and Hodges (Fix and Hodges, 1951; Silverman and Jones, 1989) first introduced this estimator with the uniform kernel, *U*(-1, 1). A histogram with fixed and nonoverlapping bin widths is a particular case of the non-adaptive kernel density estimate when a uniform distribution is selected as the kernel.

Silverman (Silverman, 1986) and Sain (Sain, 2002) pointed out that there are many situations where the fixed-bandwidth or non-adaptive estimators do not perform well. The drawbacks of the non-adaptive estimators are more evident when there are large differences in the density of points in different regions of the variable space. Minnotte (Minnotte, 1992) demonstrated that the non-adaptive estimators experience difficulty with multi-modal distributions. Sain (Sain, 2002) noted that for the distributions with multiple modes it is difficult to find a single bandwidth that adequately differentiates between distinct peaks and valleys between the peaks. Ramachandran distributions are inherently multi-modal and have both highly populated and large sparsely populated or even empty regions. Sain (Sain, 1994) also notes that in higher dimensional settings, due to the scarcity of data over much of the effective space, a fixed bandwidth procedure is likely to fail unless the sample size is extremely large. Cacoullos (Cacoullos, 1966) and Terrell and Scott (Terrell and Scott, 1992) showed that distributions with a high local curvature have significant reductions in the bias when an adaptive procedure is used. Ramachandran densities and their resulting rotamer probabilities have such large curvature.

Breiman, Abramson, and others developed *adaptive kernel density estimates* (*AKDE*) also referred as sample point density estimators in which the bandwidth parameter varies across the sample data points, depending on the local density of the data (Abramson, 1982b; Breiman et al., 1977). In this case, *h* can be replaced with $h_i = \lambda_i h$, where λ_i is a scaling parameter. Abramson provided the following as an expression for λ_I:

$$\lambda_i = \left(\frac{\left(\prod_{j=1,n} \hat{f}\left(x_j\right) \right)^{\frac{1}{N}}}{\hat{f}\left(x_i\right)} \right)^{\alpha} = \left(\frac{g}{\hat{f}\left(x_i\right)} \right)^{\alpha} \qquad \text{(S5)}$$

$\hat{f}\left(x\right)$ is any pilot estimate of the density, for instance one calculated from a non-adaptive KDE with a reasonable value of the bandwidth. The factor *g* is simply the geometric mean of the pilot density estimates at the *N* data points. The result is not strongly dependent on the form of the pilot estimate (Abramson, 1982a). The power α ($\alpha \geq 0$) regulates the magnitude of how much sample points from the sparsely populated regions have their bandwidths expanded and how much those in the populated regions have their bandwidths shrunk relative to the “geometric mean sample point.” The special case of $\alpha = 0$ degenerates AKDE into the non-adaptive one. A value of 1/2 is commonly used (Abramson, 1982b; Silverman, 1986).

Calculating Ramachandran densities of each rotamer with 2D KDE

We apply kernel density estimates to model Ramachandran densities for each rotamer of each residue type. Since Ramachandran probability density is defined for the backbone torsion angles ϕ and ψ as two arguments, we use a two-dimensional kernel density estimate. One of the commonly used circular probability densities is the von Mises distribution (von Mises, 1918). It is also often called the Gaussian distribution analogue for circular data:

$$\rho_{vonMises}\left(\theta \middle| \theta_0, \kappa\right) = \frac{1}{2\pi I_0\left(\kappa\right)} \exp\left(\kappa \cos\left(\theta - \theta_0\right)\right) \qquad \text{(S6)}$$

where θ_0 is a location of the mode of the distribution, $I_0(\kappa)$ is the modified Bessel function of the first kind of order zero, and κ is the von Mises concentration parameter or an inverse measure of dispersion. When κ >> 1, the von Mises distribution becomes very concentrated about the angle θ_0 with κ being a measure of concentration, and it approaches the traditional normal distribution with mean θ_0

and standard deviation, $\sigma = \sqrt{1/\kappa}$. For this reason $\sqrt{1/\kappa}$ can be regarded as a bandwidth, *h*. As with the normal distribution, approximately 67% of the density is contained within $h \equiv \sqrt{1/\kappa}$ radians from the mean value of a von Mises distribution.

The non-adaptive or fixed-bandwidth KDE in two dimensions for Ramachandran data can be written as the sum over products of ϕ- and ψ- von Mises kernels for N_r data points of rotamer type, *r*:

$$\begin{aligned}\hat{\rho}(\phi,\psi|r) &= \frac{1}{N_r}\sum_{i=1}^{N_r} K_h\left(\|\phi-\phi_i\|\right)K_h\left(\|\psi-\psi_i\|\right)\\ &= \frac{1}{4\pi^2 N_r}\sum_{i=1}^{N_r}\frac{1}{\left(I_0\left(\kappa\right)\right)^2}\exp\left(\kappa\left(\cos\left(\phi-\phi_i\right)+\cos\left(\psi-\psi_i\right)\right)\right)\end{aligned} \qquad \text{(S7)}$$

In this case $\sqrt{1/\kappa}$ defines a radius of the two-dimensional hump covering 67% of the kernel density. For simplicity of further formulas we do not place a cap on top of kernel density or kernel regression estimates whether it is density, $\hat{\rho}$ or probability, $\hat{P}$ or χ mean, $\hat{\mu}(\chi)$ or χ variance, $\hat{\sigma}^2(\chi)$. These estimates can be easily identified by the sum expressions of kernels or products of a kernel and some object. The only exception where we still preserve the cap is a pilot estimate, $\hat{f}\left(\bullet\right)$.

Data-adaptive KDE for Ramachandran density modeling

Rewriting Equation S7, we generate the data-adaptive kernel density estimate (AKDE) of the Ramachandran density of a rotamer, *r*:

$$\rho(\phi,\psi\,|\,r) = \frac{1}{4\pi^2 N_r}\sum_{i=1}^{N_r}\frac{1}{\left(I_0\left(\kappa/\lambda_i\right)\right)^2}\exp\left(\frac{\kappa}{\lambda_i}\left(\cos\left(\phi-\phi_i\right)+\cos\left(\psi-\psi_i\right)\right)\right) \qquad \text{(S8)}$$

where $\lambda_i = \lambda\left(\phi_i,\psi_i\right)$ scales the width of the von Mises kernels. Large values of λ_i produce wide kernels and small values produce narrow kernels. For the Ramachandran densities for the rotamer library, we can make the scaling parameter, λ_i, data-adaptive in the following two ways

1) Separately adaptive within each rotamer, *r*:

$$\lambda_i = \left(\frac{\left(\prod_{j=1}^{N_r}\hat{f}\left(\phi_j,\psi_j\,|\,r\right)\right)^{\frac{1}{N_r}}}{\hat{f}\left(\phi_i,\psi_i\,|\,r\right)}\right)^{\alpha} = \left(\frac{g_r}{\hat{f}\left(\phi_i,\psi_i\,|\,r\right)}\right)^{\alpha} \qquad \text{(S9)}$$

2) Adaptive within all conformations of each residue type:

$$\lambda_i = \left(\frac{\left(\prod_{j=1}^{N} \hat{f}\left(\phi_j, \psi_j\right) \right)^{\frac{1}{N}}}{\hat{f}\left(\phi_i, \psi_i\right)} \right)^{\alpha} = \left(\frac{g}{\hat{f}\left(\phi_i, \psi_i\right)} \right)^{\alpha} \qquad \text{(S10)}$$

where N_r is a total number of conformations per residue type, $\hat{f}\left(\phi, \psi \mid r\right)$ and $\hat{f}\left(\phi, \psi\right)$ are respectively the pilot estimates of the Ramachandran density for a rotamer r and for a residue type as a whole. g_r and g are the geometric means of the corresponding densities. For the pilot estimate at the point $\left(\phi_i, \psi_i\right)$, we use the non-adaptive kernel density estimate with the same value of κ using Equation S7.

We first used Eq. S9 with several different values of α to produce the scaling parameters for each rotamer. However, we found that this produced undesirable results, particularly for rare rotamers. For instance, we might have two groups of points of one rotamer type in proximity to each other in (ϕ, ψ)-space with one additional point of a rare rotamer type in their midst. Since Eq. S9 produces different kernel widths for the different rotamers, the data points for the common rotamer will have relatively narrow kernels due to the higher density of these data points, while the rare rotamer will have a very wide kernel due to its low density in that region. Some distance away from these two clusters of points, the values of the kernels for the popular rotamer have dropped orders of magnitude, while the kernel of the rare rotamer may still have a non-negligible value. The inversion of the rotamer probability density estimates using Bayes rule (Eq. S2) will produce high probabilities for these rare rotamers in sparsely populated parts of the map. Using a rotamer-independent scaling of κ with Equation S10 avoids this problem by making the scaling parameter dependent only on $\left(\phi_i, \psi_i\right)$. In general, we found a value of α=1/2 produced reasonable results.

Treatment of rare rotamers

Some very rare rotamers for longer side chains such as Arg and Lys and even for shorter side chains like Leu have only a few examples in the data set, sometimes less than 10 examples (Leu < g^+, g^- > has 10 data points, Figure 2), and it is therefore quite difficult to estimate their Ramachandran densities with any accuracy. Some of the rarest ones are likely very high in energy and may be misfit to the density. In any case, we want to prevent them from having significant probability in any part of the (ϕ, ψ) space. We therefore implemented a “rare rotamer fix” by which if there were fewer than 25 data points for a rotamer, we estimated the Ramachandran PDF with the data having one less degree of freedom, starting from the last χ angle of the side chain. For example, if a Lys or Arg rotamer has less than 25 instances, we use the following equation to express the numerator of Equation S2:

$$P\left(r\middle|\phi,\psi,aa\right)\propto\rho\left(\phi,\psi\middle|r_1,r_2,r_3,aa\right)P\left(r_1,r_2,r_3,r_4\middle|aa\right) \tag{S11}$$

where $r=\left\{r_1,r_2,r_3,r_4\right\}$ and r_i is the rotamer for the χ_i degree of freedom. The denominator of Equation S2 must also be altered so that when r' represents a rotamer treated in this way, the term in the sum is also calculated using Equation S11. If there are not enough data points in $\left\{r_1,r_2,r_3\right\}$ then $\rho\left(\phi,\psi\middle|r_1,r_2\right)$ can be used, etc.

Maximum-likelihood cross validation for κ parameter

The extent of smoothing is determined by the concentration parameter, κ in Eq. S8 used to determine $\rho\left(\phi,\psi\middle|r\right)$. Oversmoothed PDFs may lose important details in the variation of rotamer probabilities, while undersmoothed PDFs may produce results that are too bumpy. We calculate one value of κ for each residue type by maximizing the likelihood of the data points using cross validation. Since we are trying to determine the best κ to determine the rotamer probabilities, $P\left(r\middle|\phi,\psi\right)$, we chose to maximize the likelihood function of these probabilities, rather than the Ramachandran probability densities, $\rho\left(\phi,\psi\right)$ or $\rho\left(\phi,\psi\middle|r\right)$. That is, we used the likelihood function,

$$\mathcal{L}\left(\kappa\right)=\prod_{i=1}^{N}P\left(r_i\mid\phi_i,\psi_i\right) \tag{S12}$$

(here r_i means the rotamer of the i-th side chain). In practice, we use a ten-fold cross validation of the log likelihood:

$$\kappa_{Optim}=\arg\max_{\kappa}\mathcal{L}^*\left(\kappa\right)=\arg\max_{\kappa}\sum_{i=1}^{N}\log\left(P\left(r_i\mid\phi_i,\psi_i\right)\right) \tag{S13}$$

Adaptive kernel regression (KR) for the rotameric χ angles and variances

The second major component of the rotamer library is the backbone-dependent population means, μ and standard deviations, σ of the available side-chain dihedral angles (χ_1, χ_2, χ_3 and χ_4) for each rotamer of the 22 residue types (Table S1). In contrast to the previous versions of the rotamer library, our goal is to model the χ means and their variances as smoothly varying functions and to achieve a good trade-off between the amount of smoothing and the accuracy of finer details in the fit to the experimental data.

For the side chains, i = 1..N_r, of the residue type *aa* and rotamer r, we model the regression relation between the response variable, χ (which can be χ_1, χ_2, χ_3, or χ_4), and the explanatory variables (ϕ, ψ) (Härdle et al., 2004):

$$\chi_i = m(\phi_i,\psi_i \mid r) + v^{\frac{1}{2}}\left(\phi_i,\psi_i\right)\varepsilon_i \tag{S14}$$

where $m(\phi_i,\psi_i \mid r)$ is the unknown regression function, $v\left(\phi_i,\psi_i\right)$ is the variance, and ε_i are random observation errors normally distributed with a mean of zero and variance 1. Given that side chains in backbone-constrained conformations experience greater uncertainty in their χ angles, we assume the standard deviation of the observation errors vary as a function of ϕ and ψ; that is, the model is *heteroscedastic.* In this case the regression function is the conditional expectation or population mean of χ given the backbone conformation:

$$m\left(x,y \mid r\right) = E\left(\chi \mid \phi = x,\psi = y,r\right) = \mu(\chi \mid \phi = x,\psi = y,r) \tag{S15}$$

$$v\left(x,y \mid r\right) = \mathrm{Var}\left(\chi \mid \phi = x,\psi = y,r\right) = \sigma^2\left(\chi \mid \phi = x,\psi = y,r\right) \tag{S16}$$

Since we do not expect $\mu(\chi \mid \phi,\psi,r)$ and $\sigma^2\left(\chi \mid \phi,\psi,r\right)$ to vary rapidly with ϕ and ψ, we use the Nadaraya-Watson or local constant kernel regression (KR) estimator to model them. The Nadaraya-Watson estimator can be seen as a special case of a larger class of KR estimators. It corresponds to a local constant or zero-order polynomial, *kernel-weighted* least squares fit:

$$\begin{aligned}
\mu\left(\chi \mid \phi,\psi,r\right) &= \frac{\sum_{i=1}^{N_r} K_h\left(\phi-\phi_i,\psi-\psi_i\right)\chi_i}{\sum_{i=1}^{N_r} K_h\left(\phi-\phi_i,\psi-\psi_i\right)} \\
\sigma^2\left(\chi \mid \phi,\psi,r\right) &= \frac{\sum_{i=1}^{N_r} K_h\left(\phi-\phi_i,\psi-\psi_i\right)\left(\mu\left(\chi \mid \phi_i,\psi_i,r\right)-\chi_i\right)^2}{\sum_{i=1}^{N_r} K_h\left(\phi-\phi_i,\psi-\psi_i\right)}
\end{aligned} \tag{S17}$$

Rewriting Equation S17:

$$\mu\left(\chi \mid \phi,\psi,r\right) = \frac{1}{N_r}\sum_{i=1}^{N_r}\left(\frac{K_h\left(\phi-\phi_i,\psi-\psi_i\right)}{\frac{1}{N_r}\sum_{i=1}^{N_r} K_h\left(\phi-\phi_i,\psi-\psi_i\right)}\right)\chi_i = \frac{1}{N_r}\sum_{i=1}^{N_r} W_i(\phi,\psi)\chi_i = \sum_{i=1}^{N_r} W_i^*(\phi,\psi)\chi_i$$

$$\begin{aligned}
\sigma^2\left(\chi \mid \phi,\psi,r\right) &= \frac{1}{N_r}\sum_{i=1}^{N_r}\left(\frac{K_h\left(\phi-\phi_i,\psi-\psi_i\right)}{\frac{1}{N_r}\sum_{i=1}^{N_r} K_h\left(\phi-\phi_i,\psi-\psi_i\right)}\right)\left(\mu\left(\chi \mid \phi_i,\psi_i,r\right)-\chi_i\right)^2 \\
&= \sum_{i=1}^{N_r} W_i^*(\phi,\psi)\left(\mu\left(\chi \mid \phi_i,\psi_i,r\right)-\chi_i\right)^2
\end{aligned}$$

$$\sum_{i=1}^{N_r} W_i^*(\phi,\psi) = \sum_{i=1}^{N_r} \frac{1}{N_r} W_i(\phi,\psi) = 1 \tag{S18}$$

reveals that the Nadaraya-Watson estimator can be seen as a weighted, local average of the response variables. $\sigma^2(\phi,\psi \mid r)$ is an estimate of the data variance as a function of ϕ and ψ which is what we want to be a part of the rotamer library, not the uncertainty in $\mu(\chi \mid \phi,\psi,r)$ *per se*. The appropriate kernel for regression onto the angles ϕ and ψ is again a symmetric two-dimensional von Mises kernel:

$$K_h(\phi-\phi_i,\psi-\psi_i) = \frac{1}{4\pi^2 (I_0(\kappa))^2} \exp\left(\kappa\left(\cos(\phi-\phi_i)+\cos(\psi-\psi_i)\right)\right) \tag{S19}$$

The non-adaptive KR estimate (Eq. S17 and S19) encounters problems similar to those of non-adaptive kernel density estimate described above. Using a locally adaptive instead of a fixed bandwidth may be advantageous for several reasons (Brockmann et al., 1993). The estimator can adapt to the density of sample points, taking a larger bandwidth where points are sparse. It can adapt to changes in residual variance in case of heteroscedacity, smoothing more where residual variance is high. The estimator can adapt to the structure of the regression function, smoothing more in flat parts of the surface and less in steeper parts. This leads to improved smoothness that is one of our goals of better side-chain modeling.

For the KR we compared two ways of adapting the local bandwidth by replacing κ with κ/λ in Eq. S19. The first is data-adaptive or also called sample-point adaptive as in Equation S9 or S10:

$$\lambda_i = \left(\frac{\left(\prod_{j=1}^{N_r} \hat{f}(\phi_j,\psi_j \mid r) \right)^{\frac{1}{N_r}}}{\hat{f}(\phi_i,\psi_i \mid r)} \right)^{\alpha} = \left(\frac{g_r}{\hat{f}(\phi_i,\psi_i \mid r)} \right)^{\alpha} \tag{S20}$$

Unfortunately, the data-adaptive scaling leads to highly unreliable results in sparsely populated (ϕ, ψ) areas. It creates an effect such that even one or two points in an otherwise unpopulated region can determine the values of KR in a large area.

The second way is the query-point adaptive kernel regression, also called a balloon estimator (Breiman et al., 1977; Sain, 2002). In our case,

$$\lambda(\phi,\psi) = \left(\frac{g_r}{\hat{f}(\phi,\psi \mid r)} \right)^{\alpha} \equiv \lambda_{\phi,\psi} \tag{S21}$$

The major difference here is that the bandwidths of each kernel are adapted relative to the density at the *query* or estimation point, (ϕ, ψ) rather than the densities at the sample points. The balloon estimator is a locally fixed-bandwidth estimator with different bandwidth values for different query points. In trial calculations, the query-adaptive approach with α=1/2 proved to be better suited for our regression problem.

Bandwidth estimation for kernel regression

We found two drawbacks in directly applying the query-adaptive scaling to the kernel regression (Equation S21). First, even for common rotamers, a few points in a sparse region might cause rather sharp changes in the mean dihedral angles. Second, and more problematic, rare rotamers might have high relative densities due to a very small number of points, and we wish to determine the mean from these points. For instance, a rotamer having only 25 examples might have 20 points in the α-helical region and 5 points in the β-sheet region. A mean from these 5 points would be very unreliable. We therefore developed the following modification to the query-point adaptive KR scaling parameters, $\lambda_{\phi,\psi}$ of Equation S21. We want kernel functions such that at least 25 points are contained within one bandwidth radius of the query point, that is within a circle enclosing 67% of the kernel density. Starting from the query point, we count the number of query points within the radius of the kernel calculated from Eq. S21. If this number is greater than or equal to 25, then we use Eq. S21. If not, we increase λ until the circle (or bin) about the query point includes exactly 25 data points.

$$\lambda_{\phi,\psi}^{Bin} = \begin{cases} \lambda_{\phi,\psi} \;\; if \; \sum_{i=1}^{N_r} \delta\left\{ \sqrt{\left(\phi_i-\phi\right)^2+\left(\psi_i-\psi\right)^2} \le R\left(\dfrac{\kappa}{\lambda_{\phi,\psi}}\right)\right\} \ge 25 \\ \lambda^{*}_{\phi,\psi} \;\; : \; \sum_{i=1}^{N_r} \delta\left\{ \sqrt{\left(\phi_i-\phi\right)^2+\left(\psi_i-\psi\right)^2} \le R\left(\dfrac{\kappa}{\lambda^{*}_{\phi,\psi}}\right)\right\} = 25 \;\; otherwise \end{cases} \tag{S22}$$

$$K_{h(\phi,\psi)}\left(\phi-\phi_i,\psi-\psi_i\right) = \frac{1}{4\pi^2\left(I_0\left(\kappa/\lambda_{\phi,\psi}^{Bin}\right)\right)^2}\exp\left(\frac{\kappa}{\lambda_{\phi,\psi}^{Bin}}\left(\cos\left(\phi-\phi_i\right)+\cos\left(\psi-\psi_i\right)\right)\right) \tag{S23}$$

where $R\left(x\right)=1/\sqrt{x}$ is the bandwidth radius of the 2D von Mises distribution with concentration parameter x that contains 67% of the density. The δ-function is 1 if its argument is true, and 0 otherwise. While the conditional equation may seem complicated, it has a simple interpretation. $\lambda^{*}_{\phi,\psi}$ is the value of λ required to enclose 25 data points, given κ. The query-point adaptive scheme in Equation S22, $\lambda_{\phi,\psi}^{Bin}$, guarantees that at each query point, the local χ regression is based on at least 25 sample points with non-negligible values of the kernel function. These values will dominate the kernel regression value, although points further away will also contribute. This modification of the query-adaptive kernel regression leads to a more statistically significant estimates of $\mu\chi$ and $\sigma\chi$, and therefore underlying variation in $\mu(\chi \mid \phi, \psi, r)$ surface. The threshold of 25 data points was chosen as a trade-off between increasing statistical significance and decreasing backbone-dependent information of these estimates. In very sparsely populated regions of the (ϕ, ψ) space, $\kappa \big/ \lambda_{\phi,\psi}^{Bin} \to 0$, the estimates will approach the backbone-independent means and variances:

$$\mu\left(\chi \mid \phi, \psi, r\right) = \frac{\sum_{i=1}^{N_r} K_{h(\phi,\psi)}\left(\phi - \phi_i, \psi - \psi_i\right)}{\sum_{i=1}^{N_r} K_{h(\phi,\psi)}\left(\phi - \phi_i, \psi - \psi_i\right)} = \frac{\sum_{i=1}^{N_r} Const \cdot \chi_i}{\sum_{i=1}^{N_r} Const} = \frac{1}{N_r}\sum_{i=1}^{N_r} \chi_i = \overline{\chi}^{(r)}$$

$$\sigma^2\left(\chi \mid \phi, \psi, r\right) = \frac{\sum_{i=1}^{N_r} Const \cdot \left(\mu\left(\chi \mid \phi_i, \psi_i, r\right) - \chi_i\right)^2}{\sum_{i=1}^{N_r} Const} = \frac{1}{N_r}\sum_{i=1}^{N_r}\left(\chi_i - \overline{\chi}^{(r)}\right)^2 = \left(\sigma\chi^{(r)}\right)^2 \tag{S24}$$

For each of 22 residue types, *aa*, of each rotamer type *r* and for each rotameric χ, the von Mises concentration parameter, κ, sets the level of details in the regression model (Eq. S22, S23 and S17) as in the case of the rotamer probabilities. We individually optimize each concentration parameter, κ_{Optim} by the method of ten-fold least-squares cross validation. The objective function is the sum of the squared residuals of the sample:

$$\kappa_{Optim} = \arg\min_{\kappa} SR(\kappa) = \arg\min_{\kappa} \sum_{i=1}^{N_r}\left(\mu\left(\chi \mid \phi_i, \psi_i, r\right) - \chi_i\right)^2 \tag{S25}$$

That is, for each κ, 90% of the data are used to calculate the terms of the remaining 10% in Eq. S25, and this procedure is repeated 10 times, leaving out a different 10% each time. Once κ_{Optim} is found that minimizes the objective function, this value is used in Eq. S22, S23 and S17 to resolve $\mu\left(\chi \mid \phi, \psi, r\right)$ and $\sigma^2\left(\chi \mid \phi, \psi, r\right)$ for the complete input data sample. In *Results* we report standard deviations of χ instead of the squared residuals as in Eq. S25 in order to be able to compare and interpret optimal standard deviations for different rotamers and different residue types.

Backbone-dependent modeling of non-rotameric degrees of freedom

The terminal dihedral angle for certain side chain types is not well described as a rotamer. These include the terminal degrees of freedom of Asn, Asp, Glu, and Gln. The aromatic residues, Phe, Tyr, His, and Trp, also have more broadly distributed χ_2 angles than rotameric degrees of freedom, although not to the same extent as the amide and carboxylate groups. We propose to model the terminal dihedral angle of side chains with non-rotameric degrees of freedom, χ_n, as continuous probability density functions as a function of the backbone conformation, (ϕ, ψ), $\rho\left(\chi_n \middle| \phi, \psi, r_{-n}\right)$, where r_{-n} denotes the rotamer of the rotameric degrees of freedom (χ_1 for Asn, Asp, and the aromatics; χ_1, χ_2 for Gln and Glu), such that:

$$\int_{\chi_n} \rho\left(\chi_n' \mid \phi, \psi, r_{-n}\right) d\chi_n' = 1 \tag{S26}$$

With $\rho\left(\chi_n \mid \phi, \psi, r_{-n}\right)$ in hand on a fine grid of χ_n values, we can calculate binned probabilities at any desired resolution, 5°, 10°, or 30° for instance.

Modeling $\rho\left(\chi_n \mid \phi, \psi, r_{-n}\right)$ is effectively the regression of a probability density function (PDF) onto the explanatory variables ϕ,ψ; that is, we want a separate $\rho\left(\chi_n\right)$ for every ϕ, ψ on a 10°x10° grid (or any grid spacing). We have calculated Ramachandran map PDFs with data-point adaptive kernels, while we have found that regressions were better produced using query-point adaptive kernels. We achieve the backbone-dependent non-rotameric χ_n density modeling by computing the backbone-dependent KR of the χ_n densities, each of which is based on an individual χ_n data point taken from the input sample:

$$\rho\left(\chi_n \middle| \phi, \psi, r_{-n}\right) = \frac{\sum_{i=1}^{N_r} K_{h(\phi,\psi)}\left(\phi - \phi_i, \psi - \psi_i\right) K_{h(\chi_i)}\left(\chi_n - \chi_i\right)}{\sum_{i=1}^{N_r} K_{h(\phi,\psi)}\left(\phi - \phi_i, \psi - \psi_i\right)} \tag{S27}$$

where χ_i are the data points of χ_n and $K_{\phi,\psi}\left(\phi - \phi_i, \psi - \psi_i\right)$ is the query-adaptive kernel with the same expression as in Eq. S23 and its κ is the von Mises concentration parameter in the (ϕ,ψ) space. We take the kernels on χ to be one-dimensional von Mises functions (Eq. S6) centered on χ_i taken from the data sample:

$$K_{h(\chi_i)}\left(\chi_n - \chi_i\right) = \frac{1}{2\pi I_0\left(\kappa_{1d}/\lambda_i\right)} \exp\left(\frac{\kappa_{1d}}{\lambda_i} \cos(\chi_n - \chi_i)\right) \tag{S28}$$

The concentration parameter, κ_{1d} sets the overall bandwidth in the χ_n space and is chosen independently from its counterpart, the (ϕ,ψ)-space κ. λ_i are the scaling parameters calculated in the data-adaptive fashion in accordance with the one-dimensional χ_i backbone-independent density:

$$\lambda_i = \left(\frac{\left(\prod_{j=1}^{N_r} \hat{f}_\chi\left(\chi_j \middle| r_{-n}\right) \right)^{\frac{1}{N_r}}}{\hat{f}_\chi\left(\chi_i \middle| r_{-n}\right)} \right)^\alpha = \left(\frac{g_r^{1d}}{\hat{f}_\chi\left(\chi_i \middle| r_{-n}\right)} \right)^\alpha \tag{S29}$$

where $\hat{f}_\chi\left(\chi_n \mid r_{-n}\right)$ is a χ_n pilot density estimate and α=1/2. The pilot density is modeled with a non-adaptive KDE with the same concentration parameter, $\kappa_{1\mathrm{D}}$:

$$\hat{f}_\chi\left(\chi_n \mid r_{-n}\right) = \frac{1}{2\pi I_0\left(\kappa_{1d}\right) N_r} \sum_{j=1}^{N_r} \exp\left(\kappa_{1d} \cos\left(\chi_n - \chi_j\right)\right) \tag{S30}$$

The χ_n concentration parameters, κ_{1d}/λ_i (Eq. 28) are data-adaptive in order to produce a true PDF that integrates to 1. If they were forced to be χ_n query-adaptive, i.e. $\kappa_{1\mathrm{D}}/\lambda_i$, the resulting function would not integrated to 1 and would not meet the definition of a PDF (Sain, 1994).

Note that κ and κ_{1d} have different and specific values for each rotamer, r_{-n}. It is also worth pointing out that in very empty parts of the (ϕ,ψ) map where $\kappa/\lambda_{\phi\psi} \to 0$, the KR of the χ_n densities defaults to the backbone-independent density:

$$\begin{aligned}
\rho\left(\chi_n \middle| \phi,\psi,r_{-n}\right) &= \frac{\sum_{i=1}^{N_r} K_{h(\phi,\psi)}\left(\phi-\phi_i,\psi-\psi_i\right) K_{h(\chi_i)}\left(\chi_n-\chi_i\right)}{\sum_{i=1}^{N_r} K_{h(\phi,\psi)}\left(\phi-\phi_i,\psi-\psi_i\right)} \\
&= \frac{\sum_{i=1}^{N_r} Const \cdot K_{h(\chi_i)}\left(\chi_n-\chi_i\right)}{\sum_{i=1}^{N_r} Const} = \frac{1}{N_r}\sum_{i=1}^{N_r} K_{h(\chi_i)}\left(\chi_n-\chi_i\right) \equiv \rho_\chi\left(\chi_n \middle| r_{-n}\right)
\end{aligned} \tag{S31}$$

Cross-validated optimization for non-rotameric κ and $\kappa_{1\mathrm{D}}$

In the equations for the backbone-dependent KR of the non-rotameric χ_n density (Eq. S27, S23, 28, S29 and S30) there are two von Mises concentration parameters, κ and κ_{1d} regulating the soundness and smoothness of $\rho\left(\chi_n \mid \phi,\psi,r_{-n}\right)$. Both of these parameters need to be optimized by cross validation. In the best scenario they should be optimized together in the mutual 2D space,

$\kappa \times \kappa_{1d}$. However, the overhead of the KR of densities is more computationally expensive than the KR of a single response variable. In order to produce accurate probability estimates over various intervals by integrating the χ_n density, we calculated it every 1° in a 360° range. This produces a 360-element response vector instead of a single-element response variable. Despite different optimizations in the source code, it takes 12~24 hours to perform cross-validated kernel regression for one residue type with χ_n in contrast to 0.5~2 hours for the rotameric residue type. Fortunately, according to our trials there is no strong interconnection between these two in terms of influence on the objective function score. We believe it is a good approximation to separately optimize them, at first κ_{1d} and then κ.

We maximize the 10-fold cross-validated likelihood of the backbone-independent χ^{nonRot} and find κ_{1d} for each rotamer, r_{-n}:

$$\begin{cases} \kappa_{1\mathrm{D}}^{Optim} = \arg\max_{\kappa_{1\mathrm{D}}} \mathcal{L}^*\left(\kappa_{1\mathrm{D}}\right) = \arg\max_{\kappa_{1\mathrm{D}}} \sum_{i=1}^{N_r} \log \rho\left(\chi_i \middle| r_{-n}\right) \\ \rho\left(\chi_n \middle| r_{-n}\right) = \dfrac{1}{2\pi N_r} \sum_{i=1}^{N_r} \dfrac{1}{I_0\left(\kappa_{1\mathrm{D}}/\lambda_i\right)} \exp\left(\dfrac{\kappa_{1\mathrm{D}}}{\lambda_i}\cos\left(\chi_n - \chi_i\right)\right) \end{cases} \tag{S32}$$

Once $\kappa_{1\mathrm{D}}^{Optim}$ is determined, we keep it fixed in the 10-fold maximization of the log-likelihood of the backbone-dependent non-rotameric χ_n and find κ_{Optim} for each rotamer, r_{-n}:

$$\kappa_{Optim} = \arg\max_{\kappa} \mathcal{L}^*\left(\kappa\right) = \arg\max_{\kappa} \sum_{i=1}^{N_r} \log\left(\rho\left(\chi_i \mid \phi_i, \psi_i, r_{-n}\right)\right) \tag{S33}$$

Optimization of kernel bandwidths in new 2010 rotamer library: rotamer probabilities, rotameric χ means and variances and non-rotameric χ densities

For the problems of rotamer probabilities, KR of rotameric means and KR of non-rotameric densities, we maximize a score function, $S(\kappa)$:

$$\kappa_{Optim} = \arg\max_{\kappa} S\left(\kappa\right) \tag{S34}$$

We do not assume $S(\kappa)$ to be unimodal. We search for the global maximum in a wide κ range. Since we apply a 10-fold cross validation and $S(\kappa)$ itself involves a lot of computation, we require an optimization method with a minimal number of iterations and evaluations of $S(\kappa)$. We used the following method to find the optimal κ.

We first cover a vast κ-interval, $\kappa_{scale} \cdot \left[0, 2^{12}\right] = \kappa_{scale} \cdot \left[0, 4096\right]$ with 16 unevenly arranged κ_i at

$$\begin{aligned}\kappa_{Initial} &= \left\{0,\frac{1}{2^{15}},\frac{1}{2^{10}},\frac{1}{2^{8}},\frac{1}{2^{5}},\frac{1}{2^{3}},\frac{1}{2},1,2,2^3,2^5,2^8,3\cdot 2^7,2^9,2^{10},2^{12}\right\}\cdot\kappa_{scale} \\ &= \left\{0,\frac{1}{32768},\frac{1}{1024},\frac{1}{256},\frac{1}{32},\frac{1}{8},\frac{1}{2},1,2,8,32,256,384,512,1024,4096\right\}\cdot\kappa_{scale}\end{aligned} \qquad \text{(S35)}$$

where κ_{scale} is an estimate of the scale of the κ-interval where we expect to locate the maximum.

For the rotamer probability problem we calculate κ_{Optim} for each residue type. The number of data points has the same order of $(10\sim 100)\cdot 10^3$ per residue type (Table S2). The κ-interval of $\left[0,4096\right]$ corresponds to the bandwidth radius interval of $\left[1^\circ,\infty^\circ\right)$. A bandwidth corresponds to an averaging window. We choose the same $\kappa_{scale}=1$ for each residue type.

For the KR problems where we determine κ_{Optim} for each rotamer and for each degree of freedom, χ_n of a residue type, *aa*. We have varying amounts of sample data, 0~10,000. For that reason we set κ_{scale} in a relation to the number of data points per rotamer, *r*:

$$\kappa_{scale} = \frac{1}{\left(\max_i\left(N_i\right)/N_i\right)^d} \qquad \text{(S36)}$$

where *d* is 1/3 for a two-dimensional von Mises kernels and 2/5 for one-dimensional von Mises kernels (Taylor, 2008). The regression of rotameric means and non-rotamer density as a function of ϕ and ψ is a two-dimensional problem, while KDE of the backbone-independent non-rotameric χ_n density for each r_{-n} is a one-dimensional problem.

We compute $S(\kappa_i)$ at each of 16 values of $\kappa_{Initial}$ (Figure S2). We find κ_m with a maximum value, $S(\kappa_m)$. Then we locate a refinement interval of $\left(\kappa_{m-1},\kappa_{m+1}\right)$ and perform a second, refinement stage of optimization. For this purpose we utilize an algorithm which is based on golden section search and parabolic interpolation by calling a function, *fminbnd* from Matlab package. To decrease the number of iterations and score function calls, we set a relative error of 5% for κ_{Optim} to stop the optimization, so that the termination tolerance on κ_{Optim} is $0.05\cdot\left(\kappa_{m-1}+\kappa_{m+1}\right)/2$. The refinement requires about 5-10 score function evaluations to determine κ_{Optim} with the specified accuracy. It totals to ~25 iterations for the whole two-stage optimization.

We plotted the optimization curves (Figure S2) and found that $S(\kappa)$ is a unimodal function with a very broad location of κ_{Optim}. In all cases we succeeded in locating κ_{Optim} within the $\kappa_{scale}\cdot\left[0,4096\right]$

interval. We noticed that for many residue types and/or rotamers the curve rose rapidly from κ= 0 to κ = 50 to 100 and then rose very slowly at higher values of κ, sometimes reaching a maximum at very high values of κ, in the range of 500-1000. At κ= 33, the circle that encloses 67% of the von Mises density has a radius of 10º. The radius is 5º at κ = 131, 2.5° at κ=525, and 1.5° at κ=1459. Since smoothness is a desired quality in the new rotamer library, we wanted to favor lower values of κ, as long as we would not sacrifice much in the likelihood or squared residuals and the *SCWRL4* prediction rates improve, remain the same or have negligible reduction.

To avoid large κ values for little improvement in the score function of the held-out data, we tolerate a certain percentage decrease of the score range away from its maximum value. We define the range as

$$\Delta S = S(\kappa_{Optim}) - S(0) \tag{S37}$$

We can achieve smoother functions at the cost of a small decrease in *S* by choosing $\kappa_{\downarrow}$ such that

$$S\left(\kappa_{Optim}\right) - S\left(\kappa_{\downarrow}\right) = p\Delta S \tag{S38}$$

where *p* is some small percentage. We conclude that the 5% stepdown (*p* = 0.05) for the rotamer probability and regression problems produces the best results in *SCWRL4* tests (Table 1 and 2 of main text), the 5% stepdown (*p* = 0.05) is the best for *Rosetta* RMSD tests with *ClassicRelax* (Table 2), and the optimal κ values are best for the Rosetta RMSD tests with FastRelax as well as for the Rosetta side-chain conformations with both protocols. The optimal κ_{Optim} and 5%-stepdown $\kappa_{\downarrow}$ for each of these rotamer library problems can be found in Table S3. The scatter plots of $\kappa_{\downarrow}$ vs. the number of data points per residue, N_{res} or rotamer N_r are shown in Figure S2 in the right panel. It is interesting to note that with a good precision $\kappa_{\downarrow}(N)$ follow the least mean integrated squared error asymptotic (Eq. S36) for both the one-dimensional (not shown) and two-dimensional problems.

Converting 2010 non-rotameric χ density model to "rotamer" model

We can split the non-rotameric χ_n interval (Table S2) into "rotamers" for each rotamer, r_{-n}. Then we locate the main mode, $\widehat{\chi}_n$ of the backbone-independent $\rho\left(\chi_n \middle| r_{-n}\right)$ from Eq. S32. We place the first 30° χ_n bin centered on $\widehat{\chi}_n$. Then every 30° we add the remaining non-overlapping bins of the same width starting from the first bin and moving rightward. These bins designate the borders on the non-rotameric "rotamers", $\left[\theta_j^1, \theta_j^2\right)$ where $j = 1..J$ and *J* is the number of such bins.

Once the backbone-independent definitions of χ_n “rotamers” are established, we can calculate their backbone-dependent probabilities, means and standard deviations from the backbone-dependent χ_n probability density estimates (Eq. S27) by integrating over the above non-overlapping bins:

$$\begin{cases} P\left(r_j \middle| \phi,\psi,r_{-n}\right) = P\left(\theta_j^1 \le \chi_n \le \theta_j^2 \middle| \phi,\psi,r_{-n}\right) = \int\limits_{\theta_j^1}^{\theta_j^2} \rho\left(\chi_n' \middle| \phi,\psi,r_{-n}\right) d\chi_n' \\ \mu\left(\chi_n \mid \phi,\psi,r_{-n},r_j\right) = \int\limits_{\theta_j^1}^{\theta_j^2} \chi_n' \rho\left(\chi_n' \middle| \phi,\psi,r_{-n}\right) d\chi_n' \\ \sigma^2\left(\chi_n \mid \phi,\psi,r_{-n},r_j\right) = \int\limits_{\theta_j^1}^{\theta_j^2} \left[\mu\left(\chi_n \mid \phi,\psi,r_{-n},r_j\right) - \chi_n'\right]^2 \cdot \rho\left(\chi_n' \middle| \phi,\psi,r_{-n}\right) d\chi_n' \end{cases} \tag{S39}$$

The r_{-n}-conditional individual probabilities for r_j sum to 1.0 for each r_{-n}, $\sum\limits_1^J P(r_j \mid \phi,\psi,r_{-n}) = 1$. In the rotamer library, we report the joint probability of these 30° bins and the probability of rotamer r_{-n}, i.e., the probability of the conformation of the entire side chain. We accomplish this by:

$$P(< r_{-n}, r_n >| \phi,\psi) = P(r_n \mid \phi,\psi,r_{-n}) P(r_{-n} \mid \phi,\psi) \tag{S40}$$

In Eq. S39 we used the backbone-independent χ_n “rotamer” definitions for the discrete model. While it may be helpful in some rotamer-library applications, such as Rosetta, other programs including SCWRL4 may benefit from the *backbone-dependent* χ_n “rotamer” definitions. Such definitions guarantee an exact centering of the first non-rotameric “rotamer”, $\chi^0(\phi,\psi \mid r_{-n})$ on the main mode of $\rho_\chi\left(\chi_n \middle| \phi,\psi,r_{-n}\right)$ density. It may lead to higher accuracy of side-chain modeling:

$$\begin{aligned} &\left[\theta_1^1(\phi,\psi \mid r_{-n}), \theta_1^2(\phi,\psi \mid r_{-n})\right) = \left[\chi^0(\phi,\psi \mid r_{-n}) - 15°, \chi^0(\phi,\psi \mid r_{-n}) + 15°\right) \\ &\left[\theta_2^1(\phi,\psi \mid r_{-n}), \theta_2^2(\phi,\psi \mid r_{-n})\right) = \left[\chi^0(\phi,\psi \mid r_{-n}) - 15° + 30°, \chi^0(\phi,\psi \mid r_{-n}) + 15° + 30°\right) \\ &\ldots \end{aligned} \tag{S41}$$

We provide both types of the discrete rotamer libraries based on the *backbone-independent* and *backbone-dependent* χ_n “rotamer” definitions.

SCWRL4 and Rosetta calculations

To reduce bias in SCWRL4 benchmarking we specifically built a new set of rotamer libraries based on a smaller set of chains than in the original dataset. From the 4,018 chains we removed chains having sequence identity more than 50% with any of the 379 chains in the previously published SCWRL4 testing set (Krivov et al., 2009). The 4,018 chain list shrank by 366 producing a list of 3,652 proteins. We found it impractical to recalculate the older 2002 rotamer library using Bayesian statistical methods based on a smaller than its original high-resolution 850 chain list available in 2002. Therefore in our benchmarking results the 2002 rotamer library may have some bias toward higher prediction rates then it genuinely has, diminishing the difference between the 2002 and 2010 libraries.

SCWRL4 was run in its default flexible-rotamer-model mode (FRM) and crystal symmetry was enabled. In this mode, residues in the asymmetric unit of the crystal may have contacts with side chains in crystal neighbors, and these interactions are added as edges in the interaction graph. Thus, a *bona fide* prediction of the side-chain conformations in the crystal is performed.

SCWRL4 was run without the residue-specific parameters described by Krivov et al. (2009). In that work, three parameters were optimized for each residue type: the constant before the rotamer log probability term in the scoring function, the temperature in the flexible rotamer model free energy calculation, and a scaling factor in front of the standard deviations for sampling subrotamers. For the calculations in this paper on many different rotamer libraries, we set these values to 3.0, 2.0, and 1.0 respectively for all residue types.

The results reported in Table 1 are for side chains in the test set with electron density in the top 75th percentile, i.e., discarding potentially disordered side chains in the accuracy evaluation, although the predictions were performed on all residues. The proteins in the test set were analyzed with the program SIOCS (Heisen and Sheldrick, unpublished), which flips Asn, His and Gln residues if better hydrogen bonding can be formed. The results in Table 1 compare the actual value predicted by SCWRL4 with the SIOCS-processed X-ray structures.

Table 1 reports the *average absolute* accuracy. For a side-chain type such as Lys, this is an average of percent χ_1, percent χ_{1+2}, percent χ_{1+2+3}, and percent $\chi_{1+2+3+4}$ correct:

$$PC_{Lys} = 100\frac{N_1 + N_{12} + N_{123} + N_{1234}}{4N_{Lys}} \tag{S42}$$

where N_{12} for instance is the number of lysine side chains with both χ_1 and χ_2 correct within 40°. This value gives added weight to the more reliably determined degrees of freedom closer to the backbone. To obtain an accuracy across all side-chain types, we weight *PC* for each amino acid type by its frequency:

$$PC = \frac{\sum_{Res} N_{Res} PC_{Res}}{\sum_{Res} N_{Res}} \tag{S43}$$

The Rosetta calculations were performed with Rosetta3.1 {Leaver-Fay, 2011 #7968} with the 2002 library and six 2010 libraries (optimized, 2%, 5%, 10%, 20% and 25% stepdown) and one developmental library version (2009it10, distributed with Rosetta3.1) as modified by Song et al. (2011). *ClassicRelax* was run with 25 Stage 1 outer cycles and 25 Stage 1 inner cycles. *FastRelax* was run with its default settings. Using the maximum-likelihood superposition software program, *Theseus* (Theobald and Wuttke, 2006), each of the 100 resulting decoys was translated and rotated closest to the idealized structure. The maximum-likelihood backbone and full-atom RMSD and χ prediction accuracy between each decoy and original idealized structure were calculated and averaged for each rotamer library. The average RMSD over the 50 proteins was calculated for each library, and then the difference in percent was calculated from the average for the 2002 library. For side-chain accuracy, we did not optimize the flip state of Asn, Gln, and His for the 50-protein benchmark. Side-chain accuracies are therefore reported with Asn and His χ_2 correct if either χ_2 or χ_2+180° was within 40° of the X-ray structure. The same holds for Gln χ_3. To compare side-chain results of SCWRL4 with those of Rosetta, we also calculated SCWRL4 accuracies treating these residues as symmetric about their terminal dihedral angles and for all side chains regardless of electron density values.

SUPPLEMENTARY REFERENCES

Abramson, I.S. (1982a). Arbitrariness of the Pilot Estimator in Adaptive Kernel Methods. J. Multivariate Anal. *12*, 562-567.

Abramson, I.S. (1982b). On bandwidth variation in kernel estimates - a square root law. Ann. Statist. *10*, 1217-1223.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Breiman, L., Friedman, J.H., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. Technometrics *19*, 135-144.

Brockmann, M., Gasser, T., and Herrmann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. J. Am. Stat. Assoc. *88*, 1302-1309.

Cacoullos, T. (1966). Estimation of a multivariate density. Ann. Inst. Statist. Math. *18*, 178-189.

Fix, E., and Hodges, J.L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties, Technical Report 4, Project Number 21-49-004. (Randolph Field, Texas, USAF School of Aviation Medicine).

Härdle, W.K., Müller, M., Sperlich, S., and Werwatz, A. (2004). Nonparametric and Semiparametric Models (Berlin: Springer-Verlag).

Kleywegt, G.J., Harris, M.R., Zou, J.Y., Taylor, T.C., Wahlby, A., and Jones, T.A. (2004). The Uppsala Electron-Density Server. Acta Crystallogr D Biol Crystallogr *60*, 2240-2249.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778-795.

Lovell, S.C., Word, J.M., Richardson, J.S., and Richardson, D.C. (2000). The penultimate rotamer library. Proteins *40*, 389-408.

Minnotte, M.C. (1992). A test of mode existence with applications to multimodality. In Dept. of Statistics (Houston, Texas, Rice University).

Parzen, E. (1962). On estimation of a probability density function and mode. Annals Math. Stat. *33*, 1065-1076.

Sain, S.R. (1994). Adaptive kernel density estimation. Ph. D. Dissertation. In Dept. of Statistics (Houston, Texas, Rice University).

Sain, S.R. (2002). Multivariate locally adaptive density estimation. Comp. Stat. and Data Analysis *39*, 165-186.

Shapovalov, M.V., Canutescu, A.A., and Dunbrack, R.L., Jr. (2007). BioDownloader: Bioinformatics downloads and updates in a few clicks. Bioinformatics.

Shapovalov, M.V., and Dunbrack, R.L., Jr. (2007). Statistical and conformational analysis of the electron density of protein side chains. Proteins *66*, 279-303.

Sheldrick, G.M. (2008). A short history of SHELX. Acta Crystallogr A *64*, 112-122.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis (New York: Chapman & Hall).

Silverman, B.W., and Jones, M.C. (1989). E. Fix and J.L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951). International Stat. Review *57*, 233-238.

Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J., and Baker, D. (2011). Structure-guided forcefield optimization. Proteins *In press*.

Taylor, C.C. (2008). Automatic bandwidth selection for circular density estimation. Comp. Stat. and Data Analysis *52*, 3493-3500.

Terrell, G.R., and Scott, D.W. (1992). Variable kernel density estimation. Ann. Statist. *20*, 1236-1265.

Theobald, D.L., and Wuttke, D.S. (2006). THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics *22*, 2171-2172.

von Mises, R. (1918). Über die "Ganzzahligkeit" der Atomgewicht und verwandte Fragen. Phys. Z. *19*, 490-500.

Wang, G., and Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589-1591.

Wang, G., and Dunbrack, R.L., Jr. (2005). PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res *33*, W94-98.