HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2024 February 06.

Published in final edited form as:

J Chem Theory Comput. 2019 February 12; 15(2): 1355–1366. doi:10.1021/acs.jctc.8b00572.

Experimentally derived and computationally optimized backbone conformational statistics for blocked amino acids

Jeong-Mo Choi,

Rohit V. Pappu*

Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, One Brookings Drive, Campus Box 1097, St. Louis, Missouri 63130

Abstract

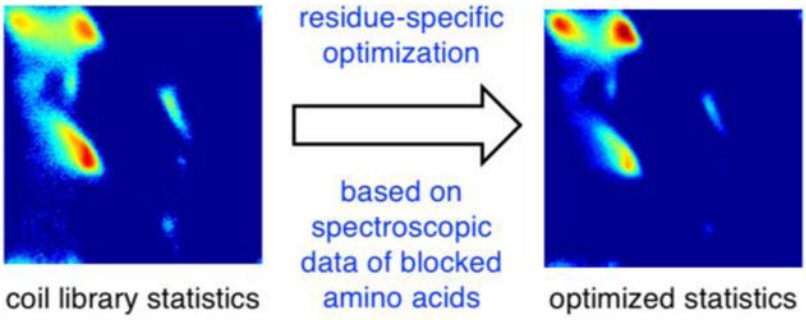
Experimentally derived, amino acid specific backbone dihedral angle distributions are invaluable for modeling data-driven conformational equilibria of proteins and for enabling quantitative assessments of the accuracies of molecular mechanics forcefields. The *protein coil library* that is extracted from analysis of high-resolution structures of proteins, has served as a useful proxy for quantifying intrinsic and context-dependent conformational distributions of amino acids. However, data that go into coil libraries will have hidden biases, and *ad hoc* procedures must be used to remove these biases. Here, we combine high-resolution biased information from protein structural databases with unbiased low-resolution information from spectroscopic measurements of blocked amino acids to obtain experimentally derived and computationally optimized coil library landscapes for each of the twenty naturally occurring amino acids. Quantitative descriptions of conformational distributions require parsing of data into conformational basins with defined envelopes, centers, and statistical weights. We develop and deploy a numerical method to extract conformational basins. The weights of conformational basins are optimized to reproduce quantitative inferences drawn from spectroscopic experiments for blocked amino acids. The optimized distributions serve as touchstones for assessments of intrinsic conformational preferences and for quantitative comparisons of molecular mechanics forcefields.

Graphical Abstract

* pappu@wustl.edu .

SUPPORTING INFORMATION

The supplementary material provides supporting information including nine figures S1–S9 that are referenced in the main text, Tables S1 and S2 in the form of spreadsheets and an archive of the optimized coil library statistics for each of the twenty naturally occurring amino acids. This information is available free of charge via the Internet at http://pubs.acs.org.



1. INTRODUCTION

Proteins and peptides consist of amino acid residues linked by peptide bonds. Information encoded in amino acid sequences and their interplay with solution conditions determines the ensemble of conformations that proteins adopt [1]. Proteins can fold into ordered three-dimensional structures either autonomously or upon binding to suitable ligands [2]. They can also display significant conformational heterogeneity, as is the case with intrinsically disordered proteins [3]. Details of the conformational ensembles accessible to a protein, be it an intrinsically foldable or intrinsically disordered protein, are governed by a combination of the amino-acid specific distributions of backbone dihedral angles, namely ϕ and ψ, and the through-space interactions amongst the amino acid residues.

The *intrinsic*, amino-acid specific backbone conformational preferences are typically presented using Ramachandran plots [4]. Amino-acid specific ϕ and ψ preferences arise from local steric considerations and the interplay between intra-peptide electrostatics and peptide-solvent interactions [5]. The intrinsic amino-acid specific ϕ and ψ preferences combined with local and long-range interactions contribute, in part, to the formation of well-defined stable structures for autonomously foldable proteins. These intrinsic backbone ϕ and ψ angle preferences also determine the conformational preferences of intrinsically disordered proteins and unfolded states of autonomously foldable proteins [6]. Context-independent backbone conformational statistics are directly relevant for two reasons: First, they provide a quantitative assessment of the range and unbiased weights of conformations that are accessible to an individual amino acid. This is useful because these intrinsic conformational statistics for blocked amino acids in water will change in different sequence and structural contexts. The intrinsic statistics serve as suitable priors and help one calibrate how sequence- and structure-specific contexts alter these preferences, thereby providing a quantitative assessment of how the interactions that go beyond the dipeptide help overcome the intrinsic conformational entropy. Second, intrinsic conformational statistics provide a useful touchstone for calibrating the accuracies of molecular mechanics forcefields for capturing residue-specific conformational preferences. This is essential because the backbone ϕ and ψ angle preferences are governed by chemical details of the interplay between backbone and sidechain atoms.

The protein coil library provides a widely accepted, quantitative description of ϕ and ψ preferences for each of the twenty naturally occurring amino acids [7]. These libraries

are assembled using backbone conformational statistics extracted from protein structures deposited in the protein data bank. Specifically, the coil library for each amino acid represents the conformational preferences for the amino acid when it is not part of regular secondary or ordered tertiary structures. Secondary structures result from the interplay between intrinsic dihedral angle preferences and intermediate / long-range inter-residue interactions. A residue that is part of a regular secondary structure will be biased toward specific values for the backbone ϕ and ψ angles. Accordingly, occurrences of residues in regular secondary structures are removed prior to assembling coil libraries, providing information on the intrinsic ϕ and ψ preferences of each amino acid sans biases imposed by secondary structure.

Since the original work of Swindells *et al.* [8], many research groups have compiled coil libraries using different data sets and filtering approaches [6c, 7, 9]. Coil library populations show distinct, residue-specific conformational basins when projected onto the two-dimensional (ϕ, ψ)-space [10]. The statistical weights assigned to each of the basins are derived from conformational populations of the basins [11]. The basin-based topography, which includes delineation of basins, identification of basin centers, and assignment of basic weights, can be converted into an amino-acid specific conformational free energy landscape. Some of the basins are given names corresponding to secondary structure elements because repetitions of the ϕ and ψ angles corresponding to the basin centers will yield these canonical secondary structures. Accordingly, the α_R basin centered on $\phi \sim -60^\circ$ and $\psi \sim -40^\circ$ and the β basin centered on $\phi \sim -150^\circ$ and $\psi \sim 150^\circ$ are the basins associated with the right-handed α-helix and the β-strand, respectively.

To zeroth order, the eighteen non-glycine and non-proline amino acids may be viewed as having similar basin structures. This is true if one uses the hard-sphere model for steric interactions [12]. However, this view becomes considerably more nuanced even for purely steric interactions, providing one uses soft-sphere repulsions [11]. Residue-specific basin topographies and basin weights emerge when considering the detailed stereochemistries of amino acids. These nuances, which are prevalent in many of the extant coil libraries, distinguish β- and γ-branched sidechains from those with straight-chain hydrocarbons or sidechains such as asparagine and glutamine from one another [9d, 13]. Therefore, coil libraries provide an experimentally derived high-resolution description of backbone dihedral angle preferences. However, these libraries do not reflect the unbiased intrinsic conformational preferences of amino acids because they are extracted from structures of folded proteins. As a result, the presence of additional structural context can introduce non-negligible biases. This limits the use of the coil library as a representation of intrinsic conformational preferences of amino acids.

Spectroscopic data on short peptides provide useful information to complement and refine the conformational statistics extracted from coil libraries. Oligopeptides afford the advantage of eliminating confounding contributions from long-range interactions, thus uncovering intrinsic conformational preferences of amino acids. The shortest peptide constructs that are ideal targets for spectroscopic investigations are blocked amino acids of the form Ace-Xaa-Nme. Here, Ace is the N-acetyl group, Nme is the N′-methylamide group, and

Xaa is the amino acid of interest. Nuclear magnetic resonance (NMR) experiments report scalar coupling constants that depend on the dihedral angle distribution of the given structural ensemble [14], while data from infrared and Raman spectroscopy can be used to extract relative populations of distinct conformational classes [15]. Far ultraviolet spectra from circular dichroism measurements also have been used to determine relative populations of different conformations. These spectroscopic investigations have been used to study various lengths and sequences of short peptides [16]. The resultant data provide insights regarding the complex interplay between specific residues and their chain contexts as determinants of the backbone dihedral angle distribution. Information from various experiments has been collated together to reconstruct Ramachandran plots, typically based on certain assumption of basin shapes [16b, 16f, 16g, 16i]. These inferences suggest that the dihedral angle distributions of short peptides are different from their counterparts extracted from coil libraries. For example, short peptides show higher populations for the polyproline II (P_{II}) basin ($\phi \sim -60^{\circ}$ and $\psi \sim 150^{\circ}$), when compared to statistics from coil libraries. High-resolution data for short peptides, specifically blocked amino acids, are essential for quantifying the intrinsic, residue-specific (ϕ, ψ)-preferences.

In this work, we present a computational method to derive high-resolution backbone dihedral angle distributions of blocked amino acids, by combining biased high-resolution data from a coil library and unbiased low-resolution data from spectroscopic measurements. Our method employs the approach of representing conformational statistics using basin structures that are defined by basin envelopes, basin centers, and basin weights [17]. We extract basins using a facsimile of the steepest descent configurational mapping method pioneered by Stillinger and Weber for describing potential energy surfaces in condensed phases [18]. We adapt this approach to extract basin structures from effective free energy landscapes whereby conformational statistics are first converted to potentials of mean force (PMFs) written in terms of backbone ϕ and ψ angles. The PMFs are used to identify basin structures.

The basin structures derived from the default coil library are optimized vis-à-vis spectroscopic data and tested for accuracy against scalar coupling constants derived from NMR experiments. The final result is an experimentally derived and computationally optimized, high-resolution, quantitative description of intrinsic conformational statistics for all twenty amino acids. These statistics, presented in the form of refined coil library landscapes, will be of direct use for modeling conformational equilibria of intrinsically disordered proteins and unfolded states of proteins [6d–g, 7a, 10a, 19]. They will also be of use in improving molecular mechanics forcefields [9d, 20], and serve as a high-resolution touchstone for comparing different forcefields.

2. GENERAL FEATURES EXTRACTED FROM COIL-LIBRARY STATISTICS

To investigate general features of coil-library statistics, we focused on the Protein Coil Library that is maintained, updated, and distributed by the Fitzkee group [7b]. As a reminder, the coil library consists solely of the structures of amino acid residues that are classified as existing outside of regular secondary structure. Specifically, we employed a pre-compiled

list (retrieved on June 5, 2017) that fits the following criteria. The conformations were excised from proteins sharing sequence identities below 90%. All of the data were extracted from protein crystal structures with an overall resolution of 1.6 Å or higher and a crystallographic R-factor below 0.25. Using structures that meet these criteria, we collected the distribution of ϕ and ψ for each of the twenty amino acid residues. Glycine has the largest number of data points (288,146), while cysteine data form the smallest set (39,067; see Supplementary Figure S1).

To generate an effective free energy landscape, *i.e.*, a PMF written in terms of ϕ and ψ angles, we binned the ϕ and ψ angles using an appropriate grid size. We applied a criterion derived from information theory to justify the choice of grid size that would maximize information content without overfitting. We chose two amino acids with non-canonical Ramachandran distributions, namely glycine (Gly) and proline (Pro), and two other amino acids with the largest and smallest data sizes among non-glycine, non-proline amino acids: leucine (Leu) and cysteine (Cys), respectively. For each of the four amino acids, we quantified the information content as a function of grid size by calculating the Shannon entropy, $S = -\sum_{i=1}^{n_\phi}\sum_{j=1}^{n_\psi} p_{ij} \log p_{ij}$ where the ϕ and ψ axes are tiled into n_ϕ and n_ψ bins and the p_{ij} values are the normalized frequencies associated with the grid with row index i and column index j.

As grids become finer, the Shannon entropy increases, until it saturates to a value governed by the logarithm of the size of the dataset (Supplementary Figure S2a). Although the saturation value has the highest Shannon entropy, this yields a sparsely populated grid that introduces additional issues in later stages of our workflow. Therefore, we computed the numerical derivative of the Shannon entropy with respect to grid size. This quantity represents the information loss associated with the increase of grid size, or equivalently, the information gained by reducing the grid size. All curves for the numerical derivatives peak around similar regions (Supplementary Figure S2b). This implies that we obtain less additional information by reducing a grid size beyond this region. Accordingly, a grid size in this interval will provide an optimal resolution where we do not lose details by coarse-graining or fine-graining. Therefore, for constructing residue-specific PMFs we used 2.5° × 2.5° grids.

The formal connection between information such as conformational statistics and free energies has been well established in statistical physics. Accordingly, instead of using the raw statistics, we converted the grid-base probabilities into PMFs. For each amino acid, the probability density $P(\varphi, \psi)$ is a marginal density that quantifies the likelihood that specific (φ, ψ)-values or intervals are realizable. Accordingly, the PMF $W(\varphi, \psi) = -kT \log[P(\varphi, \psi)]$ is an effective free energy landscape that can be used to quantify the relative free energy preference associated with one (ϕ, ψ)-pair over another, as reflected by the information content of the coil library.

Figure 1 shows the (ϕ, ψ)-dependent PMFs for all twenty amino acids using 2.5° × 2.5° grids. As noted by Perskie *et al.*, [10b] the coil-library data show distinctive, amino-acid specific (ϕ, ψ) preferences. The basin structures look significantly different between amino

acids with topologically different sidechains such as alanine (Ala) versus aspartate (Asp) or valine (Val), while topologically similar amino acids show similar basin structures. The β-branched amino acids, Val, isoleucine (Ile), and threonine (Thr), show strong preferences for the β basin. In contrast, Asp and asparagine (Asn) have higher populations in the P_{II} basin and left-handed alpha-helical basin. Note that while the Ramachandran map of Gly is required to be symmetric with respect to the origin (due to the achiral nature of its α-carbon), the coil-library landscape for Gly violates this symmetry because the Protein Coil Library imposes a filter to remove signals from regular secondary structures [7b]. Accordingly, for all of the subsequent analysis, we used a symmetrized landscape for Gly by adding a population reflected through to the origin, unless otherwise stated.

Next, we assessed the accuracy of PMFs extracted from the coil library by comparing calculated parameters to experimentally measured parameters for blocked amino acids. Specifically, we compared $^3J(H_N, H_\alpha)$ scalar coupling constants. For each amino acid, we converted all coil-library data points to their corresponding coupling constants using the Karplus equation [21] that is based on optimized parameters [16d, 22]:

$$J(\phi) = 7.09\cos^2(\phi - 60^\circ) - 1.42\cos(\phi - 60^\circ) + 1.55$$

The units of $J(\varphi)$ are hertz (Hz). The average value of $J(\varphi)$ over all data points of the specific amino acid was used as a calculated coupling constant. Comparisons of the coupling constant values obtained using the residue-specific coil-library landscapes and the experimental values reported by Avbelj *et al.* [16c] show a consistent bias towards larger J values in the coil-library data (Figure 2; mean absolute error = 0.33 Hz). This is presumably due to inconsistencies in basin distributions of coil-region residues and blocked amino acids. For example, P_{II} and β basins have different populations in the case of blocked amino acids [16h].

In order to calibrate basin structures, we need a systematic approach to quantify and refine basin weights. However, the coil library by itself provides no prior knowledge regarding basin centers, basin envelopes, and basin weights. Many previous studies have used arbitrarily drawn boundaries to quantify basin weights from PMF landscapes (for example, see [23]). We present an alternative systematic approach that is based on the concept of inherent structures developed by Stillinger and Weber [18].

3. INHERENT BASIN STRUCTURE ANALYSIS

Motivated by a previous study that demonstrated the power of configurational mapping to extract basin structures for blocked amino acids [17], we developed a method to extract basin structures from a given (ϕ, ψ)-dependent PMF. The PMFs associated with each grid value (Figure 3a) are first averaged using a 4×4 grid (which corresponds to a 10° × 10° window; see the next paragraph for a justification of the chosen grid size) to smoothen fluctuations. This provided us with coarse-grained PMF surfaces (Figure 3b). Next, we neglect windows where more than 75% of the constituent grids are empty, leaving us with only well populated regions of dihedral space (Figure 3c). For each coarse-grained grid,

the average PMF and those of immediate neighbor grids are compared to determine the direction where the PMF difference is largest (Figure 3c, arrows). If the current grid has the lowest PMF, it is considered a local minimum (Figure 3c, crosses). After determining all the steepest-gradient directions, we follow the directions to reach the nearest local minimum. The grids that eventually reach the same local minimum are defined to be part of the same basin, and the local minimum is defined as the center of the basin (Figure 3d). This approach identifies residue-specific numbers of basins that reflect the underlying PMF and are defined in an unbiased manner.

We tested different grid sizes and calculated the number of basins found for each grid size. The number of basins is greater for finer grids, which is to be expected (Supplementary Figure S3a). The data contain statistical and innate fluctuations and coarse-graining smoothens these fluctuations. We computed the number of *major basins*, which we define as the basins whose weights are above a threshold value. As shown in Supplementary Figure S3b, the numbers of basins show a non-monotonic behavior as the grid size increases, for all the threshold values we tested. This demonstrates that very fine grids make it difficult to extract accurate basin structures because the underlying structure is masked by fluctuations in the data. Accordingly, we chose 4 × 4 grids (10° × 10° window), since it is the finest grid system that provides simple and easy-to-interpret basin structures for the region $\phi \in [-180^{\circ}, -60^{\circ}]$ and $\psi \in [60^{\circ}, 180^{\circ}]$, where β and P_{II} basins lie (compare Figure 4, Supplementary Figures S4, S5, and S6).

Our formalism does not make any *a priori* assumptions regarding distributions for basins. It is computationally tractable and provides quantitative estimates for basin centers and basin envelopes. Hence, it can be co-opted to quantify residue-specific basin structures from conformational statistics either from a coil library or from simulations. In Figure 4, we show basin structures as derived from coil-library landscapes for all 20 amino acids (see also Supplementary Table S1). As shown in Figure 1, topologically similar amino acids have similar basin structures. Non-canonical basin structures of Gly and Pro are shown clearly. There are three groups that show “equivocal” basin structures. This refers to the fact that it is unclear which basin corresponds to β or P_{II} (see also Figure 1). The first group consists of Asp and Asn, where we find two significant basins for $\psi > 60^{\circ}$, whose local minima lie around $\phi = -60^{\circ}$. The second group includes only Thr, where two equally significant basins are revealed with minima at $\phi \sim -130^{\circ}$ and $\psi > 120^{\circ}$. The last group includes Ile and Val, for which β and P_{II} basins are apparently merged to one single basin with the new local minimum at $\phi \sim -115^{\circ}$ and $\psi \sim 125^{\circ}$.

Using the extracted basin structures, we next calculated the weights associated with each of the basins. To determine correct weights, we first compared to the coil-library derived weights for β, P_{II}, and α_R basins with experimental data taken from spectroscopic approaches. From the obtained basin structures, we calculated basin weights (normalized by the total data size of each amino acid) and collected basins with normalized weights > 15 %. For the 13 amino acids with canonical and “unequivocal” basin structures, we could unambiguously assign β, P_{II}, and α_R basins (Supplementary Table S1). We compared these basin weights with experimental data, which were reported as relative basin populations of the three

basins for blocked amino acids as determined by infrared and Raman spectroscopies [16h]. We find that the coil-library weights are significantly different from the experimentally derived populations (Figure 5). This implies that the coil-library statistics do not correctly reproduce backbone dihedral angle distributions of blocked amino acids.

We also employed a sub-sampling method to assess the sensitivities of basin statistics to sampling and noise. We randomly sampled 90% of the data from each data set and performed the basin analysis on the data subset as described above. We repeated this procedure 100 times for each amino acid, which allowed us to collect average values and standard deviations of basin statistics. To represent basin topographies, we quantified: (1) the total number of basins; (2) the number of major basins, *i.e.*, basins whose weights are greater than 15%; (3) the basin weights; (4) the depths of the basin centers, which refer to the locations of local minima; (5) and the basin areas (Supplementary Figure S7). For items 3 through 5, we employed the values for the α_R basin, since it has a well defined and nearly invariant position of basin center for all 20 amino acids (even in the case of Gly and Pro) and accordingly, it provides a useful way to compare statistics across different amino acids. Also, the depth of each local minimum (item 4) was corrected by adding -log(0.9), which is the compensation factor for 90% sub-sampling.

As expected, the total number of basins is more sensitive than the number of major basins (Supplementary Figure S7a and S7b). The number of major basins does not change upon sub-sampling, unless the basin weight is close to the threshold value for determining major basins (see data for phenylalanine (Phe), tyrosine (Tyr) and Gly in Supplementary Table 1). At the level of individual basins, the most sensitive statistic is the basin area (Supplementary Figure S7e), while the depths of basin centers are almost invariant (Supplementary Figure S7d). Also, the positions of basin centers are invariant; all amino acids (except Thr) show no change of basin center positions. In the case of Thr, the basin analysis sometimes found the grid right below the original one as the basin center. This is because the real basin center is close to the boundary of two neighboring grids that small randomness in sampling may lead to change of basin center position.

4. COMPUTATIONAL OPTIMIZATION OF LANDSCAPES

We used quantitative inferences from spectroscopic data to refine the coil-library derived PMFs. The refinement / optimization was designed to reproduce accurate backbone (ϕ, ψ) statistics for blocked amino acids, based on what we refer to as the *coil library ansatz*. According to this ansatz, the coil library provides a high-resolution and accurate basin structure, albeit with incorrect basin weights for blocked amino acids. Inaccuracies in basin weights can be corrected by reweighting against extant experimental data leading to optimized and accurate effective free energy landscapes for blocked amino acids.

To arrive at optimized and accurate effective free energy landscapes for blocked amino acids we employed a Monte Carlo (MC) based iterative method for basin reweighting (Figure 6). The target values were the basin weights inferred from spectroscopic data for blocked amino acids [16h]. For each basin X, we denote the basin population as $P(X)$. However, the spectroscopic experiments only provide population data for the three basins normalized

to the sum of the three basins. Therefore, we use the symbol $\bar{P}(X)$ for the normalized population in basin X such that:

$$\bar{P}(X) = \frac{P(X)}{P(\beta) + P(\mathrm{P_{II}}) + P(\alpha_R)};$$

Here, X is one of β, $\mathrm{P_{II}}$, or α_R. The cost-function E that we use for the MC optimization is the sum of absolute errors from the target populations of the three basins:

$$E = \left|\bar{P}(\beta) - \bar{P}_{\mathrm{target}}(\beta)\right| + \left|\bar{P}(\mathrm{P_{II}}) - \bar{P}_{\mathrm{target}}(\mathrm{P_{II}})\right| + \left|\bar{P}(\alpha_R) - \bar{P}_{\mathrm{target}}(\alpha_R)\right|;$$

Here, $\bar{P}_{\mathrm{target}}(X)$ is the experimentally derived target population for basin X. We also accounted for the contributions from populations corresponding to minor basins that are not the three basins considered in analysis of the experimental data. Based on physicochemical considerations it follows that regions outside the three standard basins should not contribute much to the overall landscapes. Accordingly, we assumed that their sum is close to a certain value designated as $1 - c$ within tolerance d (set to 0.01 in this work), *i.e.*, $P(\beta) + P(\mathrm{P_{II}}) + P(\alpha_R) = c$. The value of c is determined below.

For each MC step, one of the three basins is randomly chosen and is populated or depopulated. After this trial move, the cost function is calculated, and based on the Metropolis criterion (MC temperature = 1.0), the proposed move is either accepted or rejected by comparing the value of the cost function to the previous value. If the cost function becomes smaller than a tolerance (set to 0.01 in this work) and the second criterion on other basins is satisfied, the MC process is terminated.

To populate or depopulate each basin, we use a combination of basin structure and steepest descent information (Figure 7a). First, we extract the boundary pixels of a basin, which we define as pixels that are in immediate contact with non-basin pixels. Each boundary pixel has its own trajectory, which refers to a collection of pixels following steepest descent directions from the boundary pixel to the basin center. All pixels along each trajectory are assigned their relative positions from the boundary with respect to the number of pixels on the trajectory (Figure 7b). The relative positions are used to determine the scaling factor of each pixel: the scaling factor for a pixel is the maximum of all the relative position values for the pixel (Figure 7c). For a pixel missing its scaling factor, we assigned an average value of non-zero scaling factors of its neighboring pixels. Lastly, we converted this information on a coarse grid system (10° × 10°) into the original fine grid system (2.5° × 2.5°; see Figure 3a) by simple linear interpolation (Figure 7d). With these scaling factors, a basin is populated or depopulated according to following equation:

$$p_{\mathrm{new}}(i,\ j) = p_{\mathrm{old}}(i,\ j)^{\pm k s(i,j)};$$

Here, (i, j) are pixel indices, $p_{\mathrm{old}}(i,\ j)$ is the old population of pixel (i, j), $p_{\mathrm{new}}(i,\ j)$ is the new adjusted population of pixel (i, j), $s(i,\ j)$ is the scaling factor calculated above, and k

is a global coefficient set to 0.005. The plus and minus signs correspond to populating and depopulating of pixel (i, j), respectively. The parameter k serves a role of "step size."

As noted above, there are certain amino acids with "equivocal" basin structures that do not yield a unique candidate for the β or P_{II} basin. For example, the coil-library landscapes of Asp and Asn show two candidates for the P_{II} basin (Figures 1 and 4). To determine the identities of basins that are to be populated, we employed the $J(\varphi)$ coupling constant values reported by Avbelj *et al.* as an additional constraint [16c]. We tested all possible combinations of different candidate basins, and for each combination, we computed $J(\varphi)$ coupling constants from reweighted populations. We then selected the optimal combination of basins that provided the closest match between the calculated and experimentally measured $J(\varphi)$ coupling constants (Supplementary Table 2). Additionally, the coil library populations of Ile and Val indicate the absence of a P_{II} basin. Instead, the shallowness of the basins in this region suggests that the P_{II} basin is a shoulder of the β basin for Ile and Val (Figures 1 and 4). Since Ile and Val are β-branched amino acids, we mixed their populations with the population of Thr, the third and last β-branched amino acid, in 1:1 ratio, to provide a starting basin structure.

The Parameter c is another factor that may affect the optimized basin structure. We tested three different values: 1, 0.975, and 0.95. The optimized PMFs are respectively given in Figure 8, Supplementary Figure S8, and Supplementary Figure S9. As shown, $c = 0.975$ (Supplementary Figure S8) and $c = 0.95$ (Supplementary Figure S9) provide an unexpected basin structure for Ile, which is expected to be consistent with that for Val, given that they share similar side chain topologies. Similarly, Asp and Asn show different basin structures in the case of $c = 0.975$ or 0.95. Based on this rationale, we chose 1 for the value of c (meaning that the populations of the three major basins constitute more than 99 % of the total population).

5. RESULTS AND DISCUSSIONS

Figure 8 shows the optimized PMFs in terms of ϕ and ψ angles. Here the total data size for each amino acid is normalized to 5×10^5 for a statistically equivalent comparison amongst all amino acids. When compared to the original populations (Figure 1), the basins become sharper and the background is less noisy. Although the three major basins become more significant, other basins are not removed. Since the basis of basin structures comes from the coil-library data, the unique features of basin structures are preserved. The amino acids with equivocal basin structures now have well-defined basins. Finally, in accord with the experimental data [16h], the populations show strong preference for β and P_{II} basins over α_R, except for the case of Gly. It should be noted that since we do not have spectroscopic data [16h] for Pro (an imino acid with no α_L basin due to its unique secondary amine structure), the Pro data are presented directly from the coil library without any reweighting.

The residue-specific shifts that have occurred upon optimization can be quantified in several ways. We calculated the overlap coefficients (OC) and the Kullback-Leibler (KL) divergences (D_{KL}), which are useful ways to compare pairs of distinct probability

distributions. Accordingly, if P_{ij} denotes the probability of grid cell (i, j) in the coil-library landscape and Q_{ij} denotes the corresponding probability in the optimized landscape, then the overlap coefficient, which refers to the area shared under two probability distributions, is defined as:

$$OC = \sum_i \sum_j \min(P_{ij}, Q_{ij});$$

The KL divergence is defined as:

$$D_{KL} = \sum_i \sum_j P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right).$$

If the two distributions are identical, the overlap coefficient is unity and the KL divergence is zero. As the two distributions diverge from each other, the *OC* value decreases, and the D_{KL} value increases. Figure 9a shows the overlap coefficients and KL divergences for nineteen amino acids except proline. The two measures show qualitatively same trends. The shift upon optimization is smallest for Gly and largest for Asn.

To quantify the improvement of the optimized effective free energy landscapes, we used the refined statistics and basin descriptions to calculate the scalar coupling constants (Figure 9b; mean absolute error = 0.24 Hz). The absolute errors are reduced vis-à-vis the experimental data. More importantly, the bias toward positive errors is now corrected. Since the *J* coupling constant employed here depends only on ϕ, this correction mainly comes from change in relative populations amongst β and $P_{II} + \alpha_R$ basins.

6. CONCLUSIONS

Our primary goal was to obtain high-resolution, experimentally derived conformational statistics for blocked amino acids. These could serve as inputs for modeling conformational equilibria of sequences with significant conformational heterogeneity. They could also serve as a touchstone for the calibration of molecular mechanics forcefields. We developed an automated way of describing effective free energy landscapes as PMFs on the (ϕ, ψ) space. We find that the unrefined statistics drawn from the coil library are inconsistent with experimental data for blocked amino acids. To remedy this, we developed a simple basin analysis method to extract basin structures from raw data for backbone dihedral angles without imposing any *a priori* structure on the distributions. This method provides quantitative information regarding basins, including the quantification of basin centers, widths, and statistical weights. Using this information, we optimized the effective free energy landscapes, written as PMFs in terms of backbone ϕ and ψ angles that conform to quantitative inferences from spectroscopic data for blocked amino acids. The optimized landscapes yield improved estimates of scalar coupling constants and represent a useable library of conformational statistics for blocked amino acids.

Our work presents two distinct innovations: (1) the basin topography extraction method representing generalization of the configurational mapping approach pioneered by Stillinger

and Weber [18] and (2) the optimized PMFs for each of the twenty amino acids. The latter provide an important touchstone for quantitative analysis and comparisons of intrinsic, amino acid specific conformational statistics. The basin analysis method presented here is simple yet quantitative, so it can be employed to compare basin structures from different sources, such as different force fields or different structural libraries. Additionally, the optimized basin landscapes can serve as reference landscapes for backbone dihedral distributions of blocked amino acids, since currently there are no available high-resolution landscapes directly inferred from experiments. For example, as part of the development of molecular mechanics forcefields, one can compare the performance of the forcefield by computing the full basin topography for the forcefield in question and comparing this to the reference landscapes provided here.

In an ideal scenario, one would obtain the effective free energy landscapes as PMFs from high-level quantum mechanical simulations of blocked amino acids in condensed phases. The basin mapping approach developed here will be a useful way to obtain comparative descriptions of conformational statistics for different amino acids because the approach does not rely on *ad hoc* tiling of conformational space. However, pending the availability of accurate results from high-level quantum mechanical simulations, the optimized basin-based descriptions we provide here are intended to serve as experimentally derived and computationally optimized touchstones for calibrating the accuracy of forcefields. The availability of high resolution unbiased data will help improve the accuracy of the experimentally derived and computationally optimized amino acid specific conformational distributions. It may also be possible to employ state-of-the-art machine learning techniques to optimize basin landscapes that simultaneously satisfy different constraints from various experimental sources. In the accompanying manuscript[24], we demonstrate the utility of the optimized conformational distributions by deploying it for refining potential functions that describe local conformational equilibria within the ABSINTH implicit solvation model and forcefield paradigm [25].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by grant 5R01-NS05611410 from the National Institutes of Health. We are grateful to Martin Fossat, Alex Holehouse, Jared Lalmansingh, and Kiersten Ruff for useful discussions.

REFERENCES

1. Holehouse AS; Pappu RV, Collapse Transitions of Proteins and the Interplay Among Backbone, Sidechain, and Solvent Interactions. Annual Review of Biophysics 2018, 47 (1), 19–39.
2. Sosnick TR; Barrick D, The folding of single domain proteins--have we reached a consensus? Current opinion in structural biology 2011, 21 (1), 12–24. [PubMed: 21144739]
3. Wright PE; Dyson HJ, Intrinsically disordered proteins in cellular signalling and regulation. Nature Reviews in Molecular and Cell Biology 2015, 16 (1), 18–29. [PubMed: 25531225]
4. Ramachandran GN; Ramakrishnan C; Sasisekharan V, Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology 1963, 7 (1), 95–99. [PubMed: 13990617]

5. Drozdov AN; Grossfield A; Pappu RV, Role of Solvent in Determining Conformational Preferences of Alanine Dipeptide in Water. Journal of the American Chemical Society 2004, 126 (8), 2574–2581. [PubMed: 14982467]

6. (a)Brant DA; Flory PJ, The Configuration of Random Polypeptide Chains. II. Theory. Journal of the American Chemical Society 1965, 87 (13), 2791–2800;(b)Toal S; Schweitzer-Stenner R, Local Order in the Unfolded State: Conformational Biases and Nearest Neighbor Interactions. Biomolecules 2014, 4 (3), 725; [PubMed: 25062017] (c)Bernadó P; Blanchard L; Timmins P; Marion D; Ruigrok RWH; Blackledge M, A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. Proceedings of the National Academy of Sciences of the United States of America 2005, 102 (47), 17002–17007; [PubMed: 16284250] (d)Ozenne V; Bauer F; Salmon L; Huang JR; Jensen MR; Segard S; Bernado P; Charavay C; Blackledge M, Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics (Oxford, England) 2012, 28 (11), 1463–70; [PubMed: 22613562] (e)Schneider R; Huang JR; Yao M; Communie G; Ozenne V; Mollica L; Salmon L; Jensen MR; Blackledge M, Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. Mol Biosyst 2012, 8 (1), 58–68; [PubMed: 21874206] (f)Sziegat F; Silvers R; Hahnke M; Jensen MR; Blackledge M; Wirmer-Bartoschek J; Schwalbe H, Disentangling the coil: modulation of conformational and dynamic properties by site-directed mutation in the non-native state of hen egg white lysozyme. Biochemistry 2012, 51 (16), 3361–72; [PubMed: 22468860] (g)Schwalbe M; Ozenne V; Bibow S; Jaremko M; Jaremko L; Gajda M; Jensen MR; Biernat J; Becker S; Mandelkow E; Zweckstetter M; Blackledge M, Predictive atomic resolution descriptions of intrinsically disordered hTau40 and alpha-synuclein in solution from NMR and small angle scattering. Structure (London, England : 1993) 2014, 22 (2), 238–49; [PubMed: 24361273] (h)Avbelj F; Baldwin RL, Origin of the neighboring residue effect on peptide backbone conformation. Proceedings of the National Academy of Sciences USA 2004, 101 (30), 10967–72;(i)Chen K; Liu Z; Zhou C; Shi Z; Kallenbach NR, Neighbor effect on PPII conformation in alanine peptides. Journal of the American Chemical Society 2005, 127 (29), 10146–7; [PubMed: 16028907] (j)Fitzkee NC; Rose GD, Sterics and solvation winnow accessible conformational space for unfolded proteins. Journal of Molecular Biology 2005, 353 (4), 873–87; [PubMed: 16185713] (k)Panasik N Jr.; Fleming PJ; Rose GD, Hydrogen-bonded turns in proteins: the case for a recount. Protein Science 2005, 14 (11), 2910–4; [PubMed: 16251367] (l)Beck DA; Alonso DO; Inoyama D; Daggett V, The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. Proceedings of the National Academy of Sciences USA 2008, 105 (34), 12259–64;(m)Perskie LL; Rose GD, Physical-chemical determinants of coil conformations in globular proteins. Protein Science 2010, 19 (6), 1127–36. [PubMed: 20512968]

7. (a)Jha AK; Colubri A; Zaman MH; Koide S; Sosnick TR; Freed KF, Helix, Sheet, and Polyproline II Frequencies and Strong Nearest Neighbor Effects in a Restricted Coil Library. Biochemistry 2005, 44 (28), 9691–9702; [PubMed: 16008354] (b)Fitzkee NC; Fleming PJ; Rose GD, The Protein Coil Library: A structural database of nonhelix, nonstrand fragments derived from the PDB. Proteins: Structure, Function, and Bioinformatics 2005, 58 (4), 852–854.

8. Swindells MB; MacArthur MW; Thornton JM, Intrinsic φ, ψ propensities of amino acids, derived from the coil regions of known structures. Nature Structural Biology 1995, 2, 596. [PubMed: 7664128]

9. (a)Lovell SC; Davis IW; Arendall WB; de Bakker PIW; Word JM; Prisant MG; Richardson JS; Richardson DC, Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins: Structure, Function, and Bioinformatics 2003, 50 (3), 437–450;(b)Avbelj F; Baldwin RL, Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: Distributions of phi. Proceedings of the National Academy of Sciences 2003, 100 (10), 5742–5747;(c)Ormeci L; Gursoy A; Tunca G; Erman B, Computational basis of knowledge-based conformational probabilities derived from local- and long-range interactions in proteins. Proteins: Structure, Function, and Bioinformatics 2007, 66 (1), 29–40;(d)Jiang F; Han W; Wu Y-D, The intrinsic conformational features of amino acids from a protein coil library and their applications in force field development. Physical Chemistry Chemical Physics 2013, 15 (10), 3413–3428; [PubMed: 23385383] (e)Shen Y; Roche J; Grishaev A; Bax A, Prediction of nearest neighbor effects on

backbone torsion angles and NMR scalar coupling constants in disordered proteins. Protein Science 2018, 27 (1), 146–158. [PubMed: 28884933]

10. (a)Zaman MH; Shen M-Y; Berry RS; Freed KF; Sosnick TR, Investigations into Sequence and Conformational Dependence of Backbone Entropy, Inter-basin Dynamics and the Flory Isolated-pair Hypothesis for Peptides. Journal of Molecular Biology 2003, 331 (3), 693–711; [PubMed: 12899838] (b)Perskie LL; Street TO; Rose GD, Structures, basins, and energies: A deconstruction of the Protein Coil Library. Protein Science 2008, 17 (7), 1151–1161; [PubMed: 18434497] (c)Perskie LL; Street TO; Rose GD, Structures, basins, and energies: a deconstruction of the Protein Coil Library. Protein Science 2008, 17 (7), 1151–61. [PubMed: 18434497]
11. Pappu RV; Rose GD, A simple model for polyproline II structure in unfolded states of alanine-based peptides. Protein Science 2002, 11 (10), 2437–2455. [PubMed: 12237465]
12. Pappu RV; Srinivasan R; Rose GD, The Flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding. Proceedings of the National Academy of Sciences 2000, 97 (23), 12565–12570.
13. Dahl DB; Bohannan Z; Mo Q; Vannucci M; Tsai J, Assessing Side-Chain Perturbations of the Protein Backbone: A Knowledge-Based Classification of Residue Ramachandran Space. Journal of Molecular Biology 2008, 378 (3), 749–758. [PubMed: 18377931]
14. Wirmer J; Schlörb C; Schwalbe H, Conformation and Dynamics of Nonnative States of Proteins studied by NMR Spectroscopy. In Protein Folding Handbook, Wiley-VCH Verlag GmbH: 2008; pp 737–808.
15. Bandekar J, Amide modes and protein conformation. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology 1992, 1120 (2), 123–143. [PubMed: 1373323]
16. (a)Shi Z; Chen K; Liu Z; Ng A; Bracken WC; Kallenbach NR, Polyproline II propensities from GGXGG peptides reveal an anticorrelation with β-sheet scales. Proceedings of the National Academy of Sciences of the United States of America 2005, 102 (50), 17964–17968; [PubMed: 16330763] (b)Hagarman A; Measey T; Doddasomayajula RS; Dragomir I; Eker F; Griebenow K; Schweitzer-Stenner R, Conformational Analysis of XA and AX Dipeptides in Water by Electronic Circular Dichroism and 1H NMR Spectroscopy. The Journal of Physical Chemistry B 2006, 110 (13), 6979–6986; [PubMed: 16571011] (c)Avbelj F; Grdadolnik SG; Grdadolnik J; Baldwin RL, Intrinsic backbone preferences are fully present in blocked amino acids. Proceedings of the National Academy of Sciences USA 2006, 103 (5), 1272–1277;(d)Graf J; Nguyen PH; Stock G; Schwalbe H, Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study. Journal of the American Chemical Society 2007, 129 (5), 1179–1189; [PubMed: 17263399] (e)Grdadolnik J; Golič Grdadolnik S; Avbelj F, Determination of Conformational Preferences of Dipeptides Using Vibrational Spectroscopy. The Journal of Physical Chemistry B 2008, 112 (9), 2712–2718; [PubMed: 18260662] (f)Hagarman A; Measey TJ; Mathieu D; Schwalbe H; Schweitzer-Stenner R, Intrinsic Propensities of Amino Acid Residues in GxG Peptides Inferred from Amide I′ Band Profiles and NMR Scalar Coupling Constants. Journal of the American Chemical Society 2010, 132 (2), 540–551; [PubMed: 20014772] (g)Hagarman A; Mathieu D; Toal S; Measey TJ; Schwalbe H; Schweitzer-Stenner R, Amino Acids with Hydrogen-Bonding Side Chains have an Intrinsic Tendency to Sample Various Turn Conformations in Aqueous Solution. Chemistry – A European Journal 2011, 17 (24), 6789–6797; [PubMed: 21547966] (h)Grdadolnik J; Mohacek-Grosev V; Baldwin RL; Avbelj F, Populations of the three major backbone conformations in 19 amino acid dipeptides. Proceedings of the National Academy of Sciences 2011, 108 (5), 1794–1798;(i)Schweitzer-Stenner R; Hagarman A; Toal S; Mathieu D; Schwalbe H, Disorder and order in unfolded and disordered peptides and proteins: A view derived from tripeptide conformational analysis. I. Tripeptides with long and predominantly hydrophobic side chains. Proteins: Structure, Function, and Bioinformatics 2013, 81 (6), 955–967;(j)Rybka K; Toal SE; Verbaro DJ; Mathieu D; Schwalbe H; Schweitzer-Stenner R, Disorder and order in unfolded and disordered peptides and proteins: A view derived from tripeptide conformational analysis. II. Tripeptides with short side chains populating asx and β-type like turn conformations. Proteins: Structure, Function, and Bioinformatics 2013, 81 (6), 968–983;(k)DiGuiseppi D; Schweitzer-Stenner R, Probing conformational propensities of histidine in different protonation states of the unblocked glycyl-histidyl-glycine peptide by vibrational and NMR spectroscopy. Journal of Raman Spectroscopy 2016, 47 (9), 1063–1072;(l)Schweitzer-Stenner R; Toal SE, Construction and comparison of the statistical coil states of unfolded and

intrinsically disordered proteins from nearest-neighbor corrected conformational propensities of short peptides. Molecular BioSystems 2016, 12 (11), 3294–3306. [PubMed: 27545097]

17. Tran HT; Wang X; Pappu RV, Reconciling Observations of Sequence-Specific Conformational Propensities with the Generic Polymeric Behavior of Denatured Proteins. Biochemistry 2005, 44 (34), 11369–11380. [PubMed: 16114874]
18. (a)Stillinger FH; Weber TA, Packing structures and transitions in liquids and solids. Science 1984, 225 (4666), 983–9; [PubMed: 17783020] (b)Weber TA; Stillinger FH, Interactions, local order, and atomic-rearrangement kinetics in amorphous nickel-phosphorous alloys. Physical Review B 1985, 32 (8), 5402–5411.
19. Meng W; Lyle N; Luan B; Raleigh DP; Pappu RV, Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. Proceedings of the National Academy of Sciences 2013, 110 (6), 2123–2128.
20. (a)Han W; Wan CK; Jiang F; Wu YD, PACE Force Field for Protein Simulations. 1. Full Parameterization of Version 1 and Verification. Journal of Chemical Theory and Computation 2010, 6 (11), 3373–89; [PubMed: 26617092] (b)Jiang F; Han W; Wu YD, Influence of side chain conformations on local conformational features of amino acids and implication for force field development. Journal of Physical Chemistry B 2010, 114 (17), 5840–50; [PubMed: 20392111] (c)Jiang F; Han W; Wu YD, The intrinsic conformational features of amino acids from a protein coil library and their applications in force field development. Physical Chemistry Chemical Physics 2013, 15 (10), 3413–28; [PubMed: 23385383] (d)Jiang F; Zhou CY; Wu YD, Residue-specific force field based on the protein coil library. RSFF1: modification of OPLS-AA/L. Journal of Physical Chemistry B 2014, 118 (25), 6983–98; [PubMed: 24815738] (e)Zhou CY; Jiang F; Wu YD, Residue-specific force field based on protein coil library. RSFF2: modification of AMBER ff99SB. Journal of Physical Chemistry B 2015, 119 (3), 1035–47. [PubMed: 25358113]
21. (a)Karplus M, Contact Electron-Spin Coupling of Nuclear Magnetic Moments. The Journal of Chemical Physics 1959, 30 (1), 11–15;(b)Karplus M, Vicinal Proton Coupling in Nuclear Magnetic Resonance. Journal of the American Chemical Society 1963, 85 (18), 2870–2871.
22. Hu J-S; Bax A, Determination of φ and χ1 Angles in Proteins from 13C−13C Three-Bond J Couplings Measured by Three-Dimensional Heteronuclear NMR. How Planar Is the Peptide Bond? Journal of the American Chemical Society 1997, 119 (27), 6360–6368.
23. (a)Beauchamp Kyle A.; Pande Vijay S.; Das R, Bayesian Energy Landscape Tilting: Towards Concordant Models of Molecular Ensembles. Biophysical Journal 2014, 106 (6), 1381–1390; [PubMed: 24655513] (b)Huang J; Rauscher S; Nawrocki G; Ran T; Feig M; de Groot BL; Grubmuller H; MacKerell AD Jr, CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nature Methods 2017, 14 (1), 71–73. [PubMed: 27819658]
24. Choi J-M; Pappu Rohit V., Improvements to the ABSINTH forcefield for proteins based on experimentally derived amino-acid specific backbone conformational statistics. Journal of Chemical Theory and Computation 2018, Submitted.
25. Vitalis A; Pappu RV, ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. Journal of Computational Chemistry 2009, 30 (5), 673–699. [PubMed: 18506808]

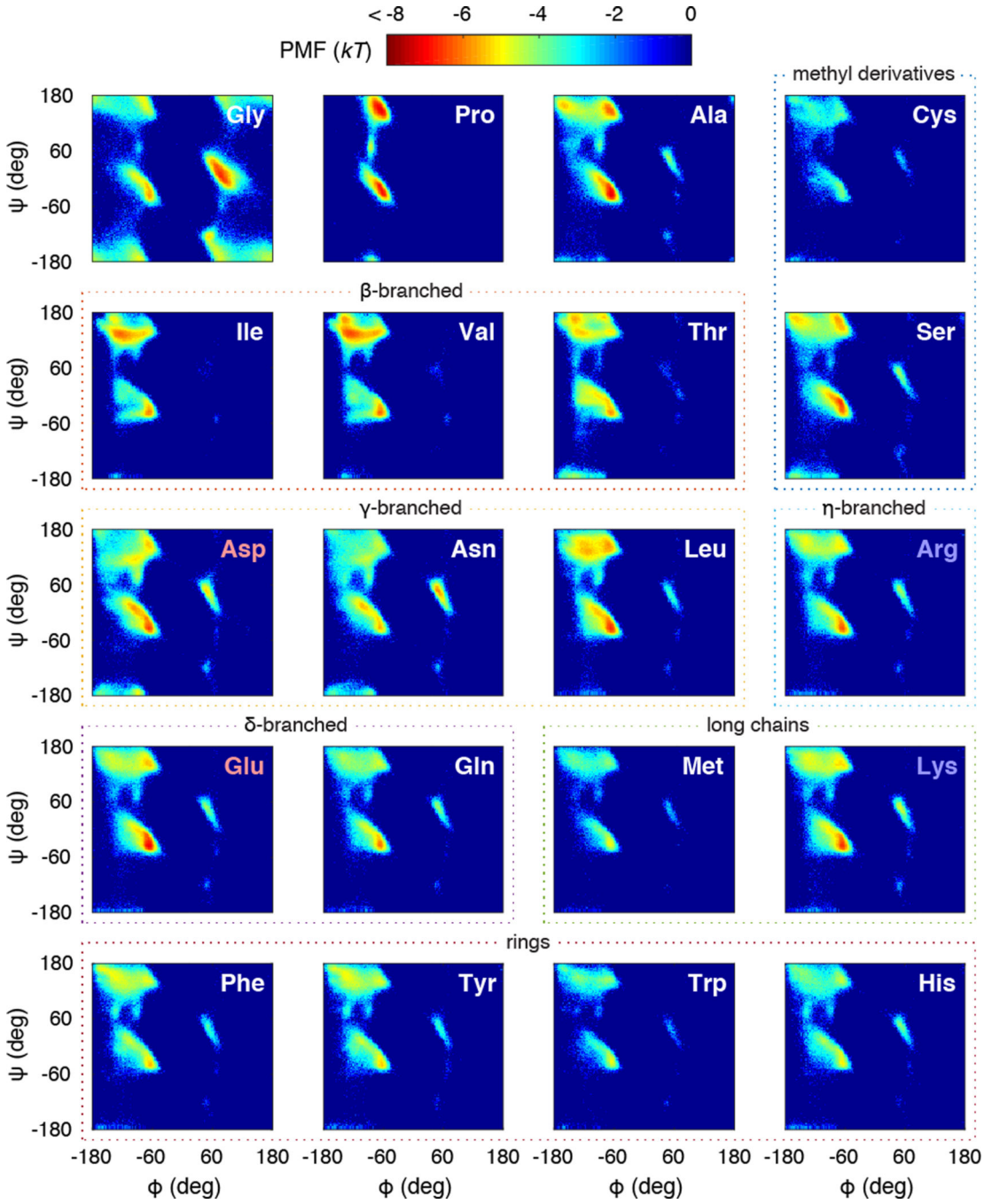


Figure 1: Coil-library landscapes for all twenty amino acids with color scale in units of thermal energy (kT) with $T = 300\text{K}$ (top).

Each panel shows a PMF in terms of the backbone dihedral angles. The amino acid label is located on the upper right corner of each panel. Red labels indicate acidic amino acids, and blue labels indicate basic ones. The different amino acids are grouped according to their stereochemistry (dotted lines).

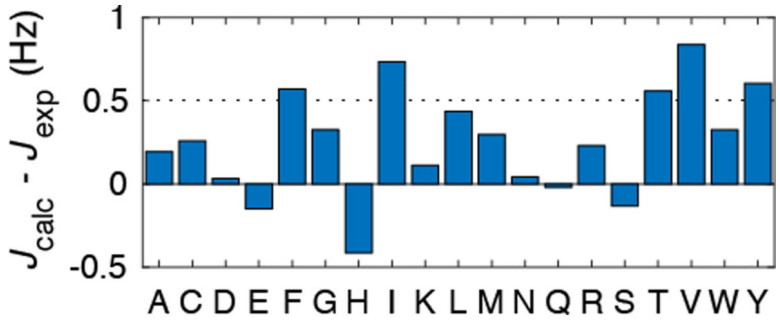


Figure 2: Differences between NMR scalar coupling constants calculated by applying the Karplus equation to the coil-library data (J_{calc}) and those determined experimentally (J_{exp})

[16c]. Note the bias toward positive errors. Absolute error = 0.25 Hz, standard deviation = 0.32 Hz, and mean absolute error = 0.33 Hz.

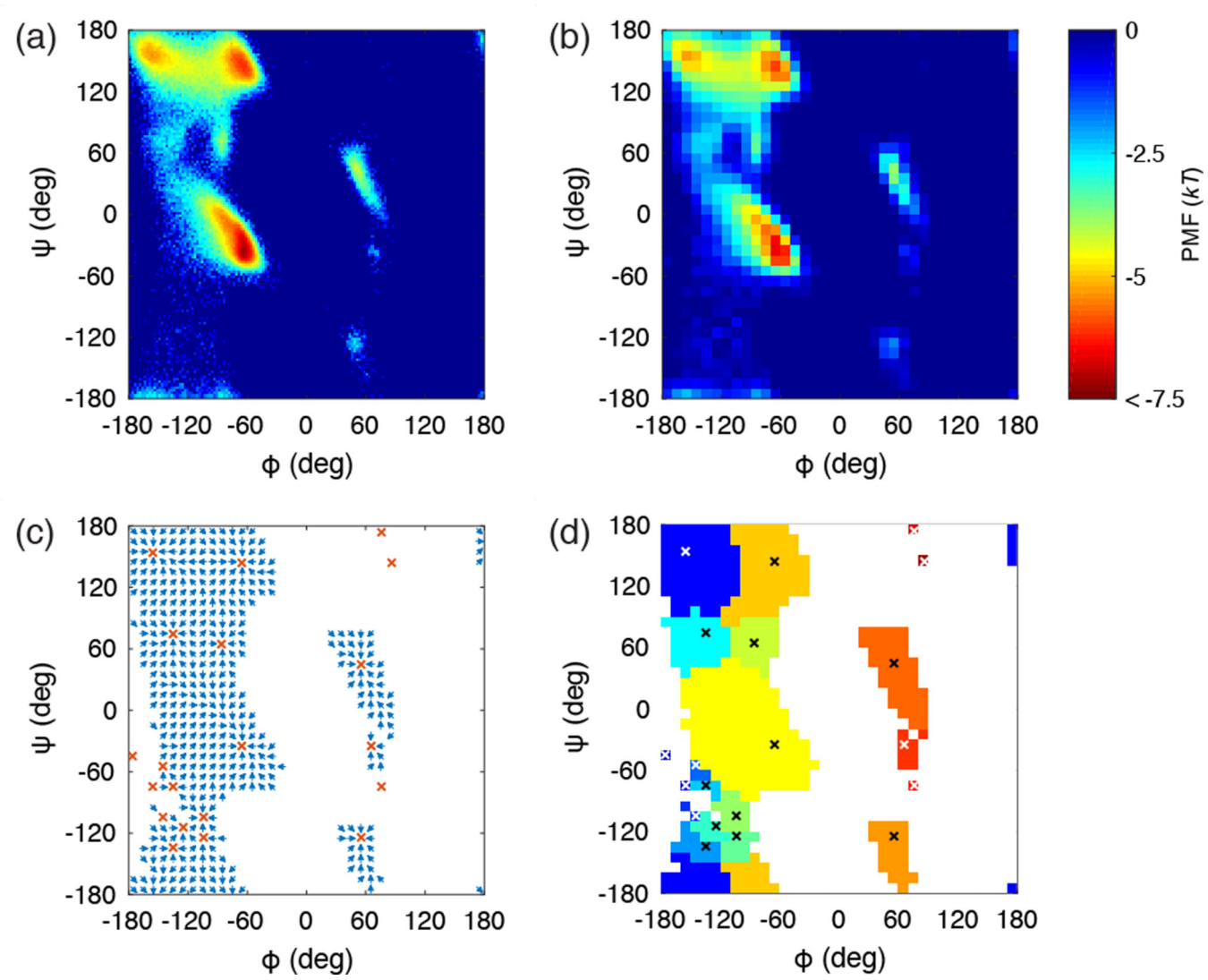


Figure 3: Steps to extract basin structures.

The method is illustrated using data for Ala from the Protein Coil Library. (a) The original potential of mean force (PMF) in backbone dihedral angles with color scale in kT. (b) The PMF averaged over a 10° × 10° window. (c) Directions of steepest gradients based on the averaged PMF. Crosses indicate local minima. (d) Final inherent basin structure. Different colors indicate different basins, each defined by a basin center – the local minimum – marked by crosses.

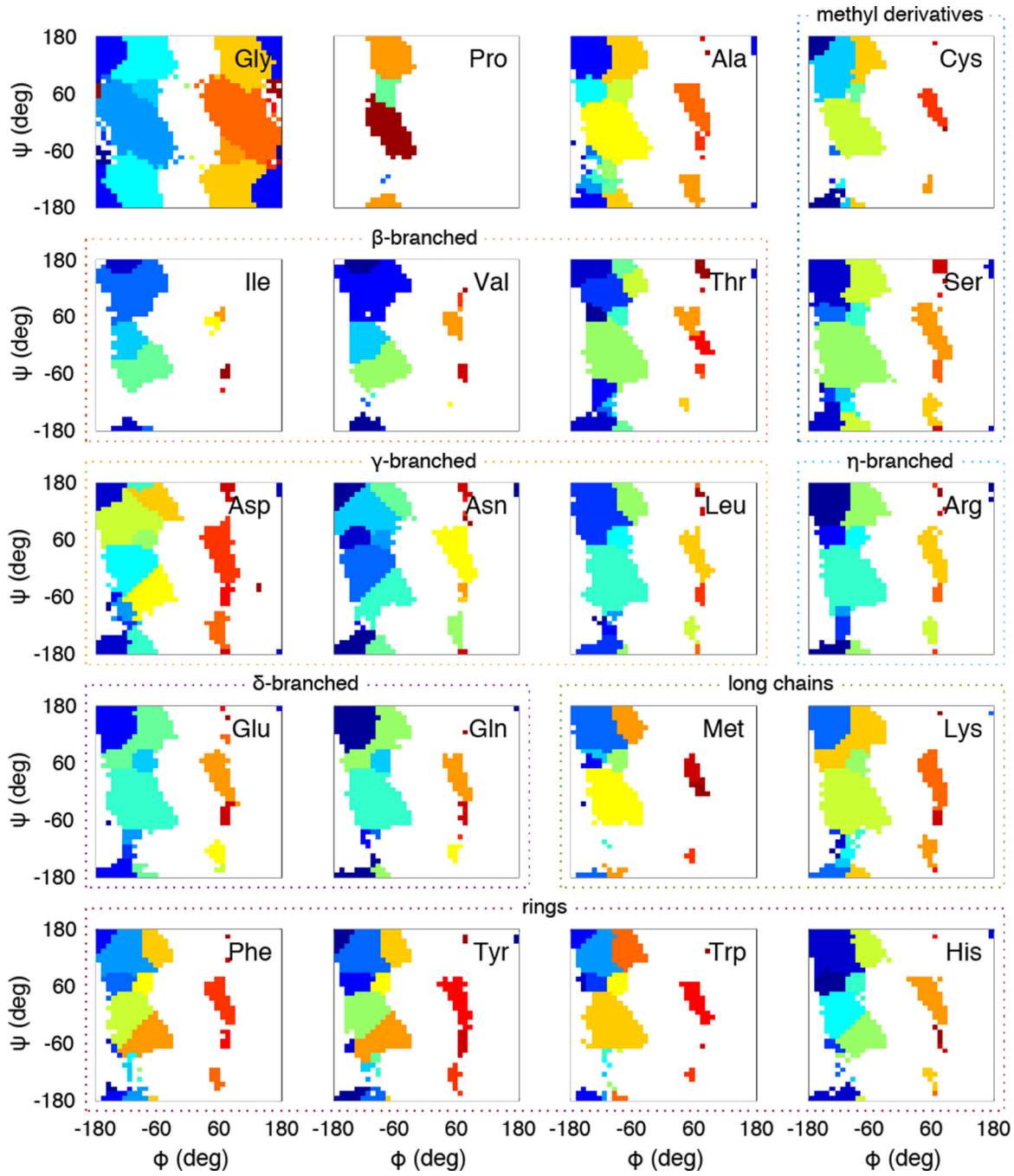


Figure 4: Basin structures derived from Protein Coil Library data for each of the twenty amino acids.

Each panel shows a basin structure of each amino acid on the Ramachandran space with different colors to indicate different basins. The grouping and order of amino acids follows that of Figure 1. Statistics for each of the basins and the coordinates of basin centers are shown in Table S1 of the supplementary material.

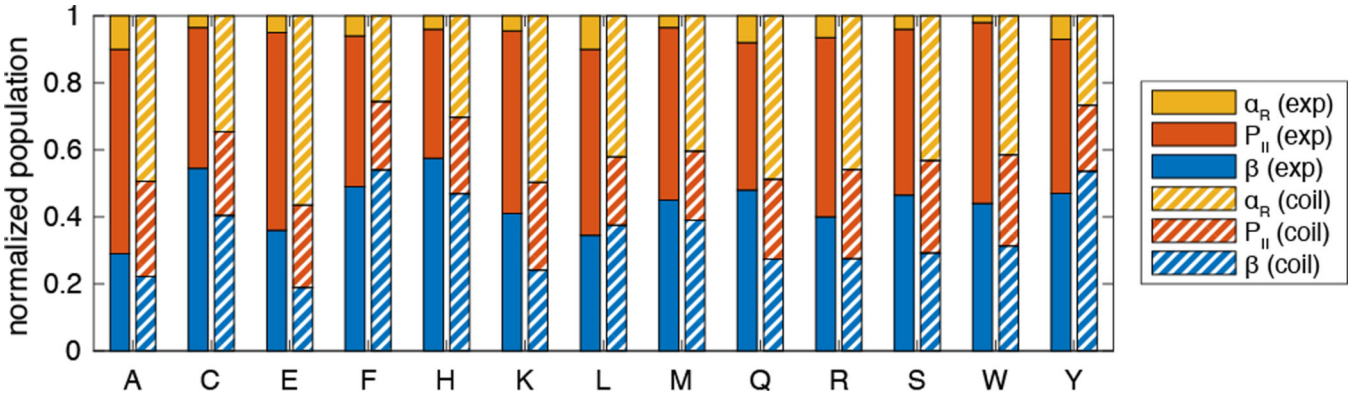


Figure 5: Relative populations of three major basins (α_R, P_{II}, and β) for blocked amino acids. The populations are derived from IR/Raman data (filled bars, normalized by the sum of the three basin populations) [16e]. These data are compared to populations derived from the unoptimized coil-library statistics (striped bars). Data are shown for 13 amino acids with canonical and unequivocal basin structures (see text).

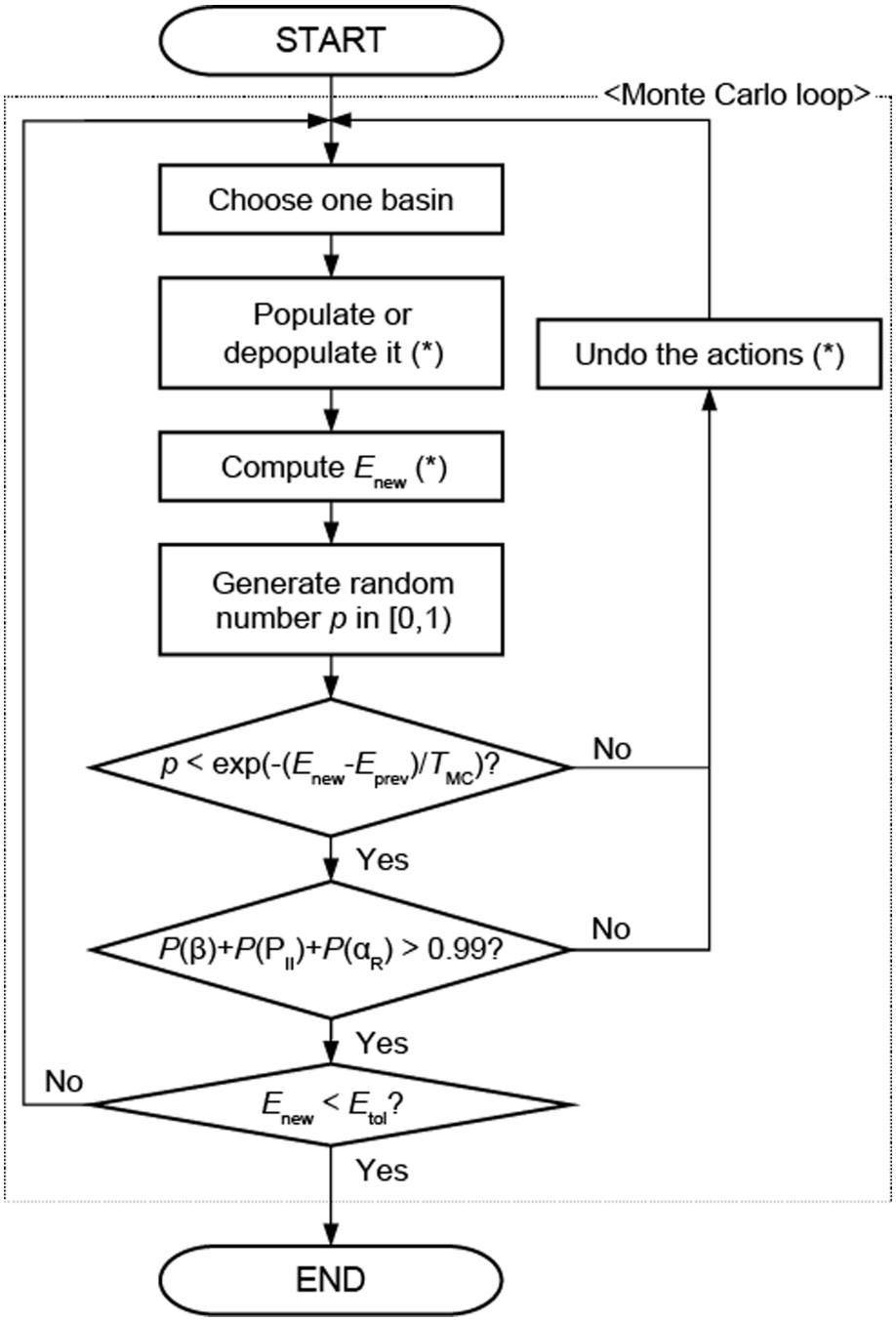


Figure 6:
Flowchart of the Metropolis Monte Carlo algorithm for optimizing the basin weights to be concordant with spectroscopic data for blocked amino acids.

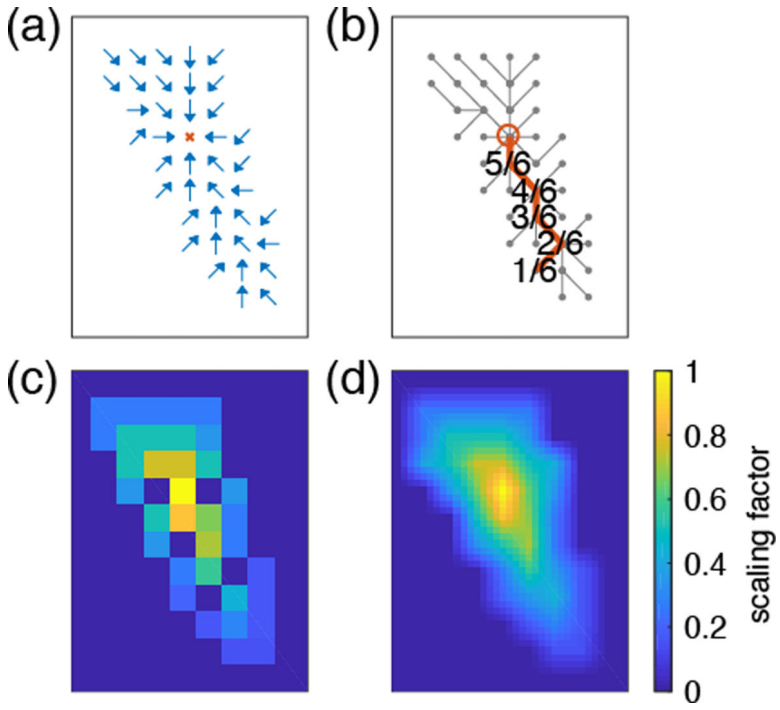


Figure 7: Steps to calculate the scaling factor for the cost function used in the Monte Carlo optimization protocol.

(a) Illustrative quiver map showing directions of steepest gradients. (b) An example of a trajectory, colored in red. The open circle indicates the basin minimum, and the numbers indicate the relative distances of grid points on the trajectory from the boundary. (c) The maximum of all the relative position values for all pixels on any trajectory. Note some pixels that do not have affiliated trajectories. (d) The final scaling factor values on a fine grid system.

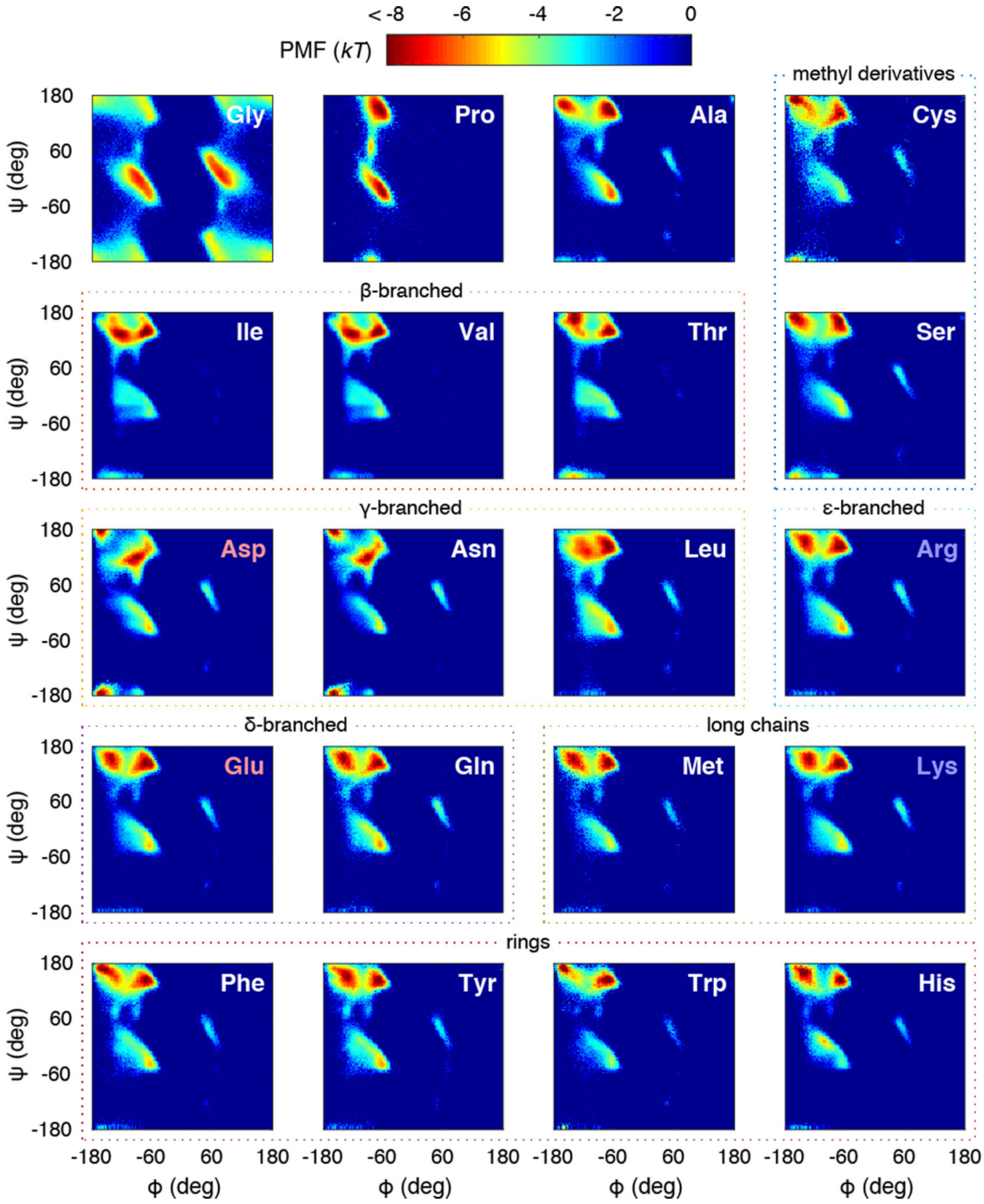


Figure 8: Optimized effective free energy landscapes for all twenty amino acids with color scale in kT (top).

Each panel shows a PMF distribution on the Ramachandran space of an amino acid labeled at the upper right corner in the three-letter notation. The grouping and order of amino acids follows that of Figure 1. The total data size for each amino acid is normalized to 5×10^5.

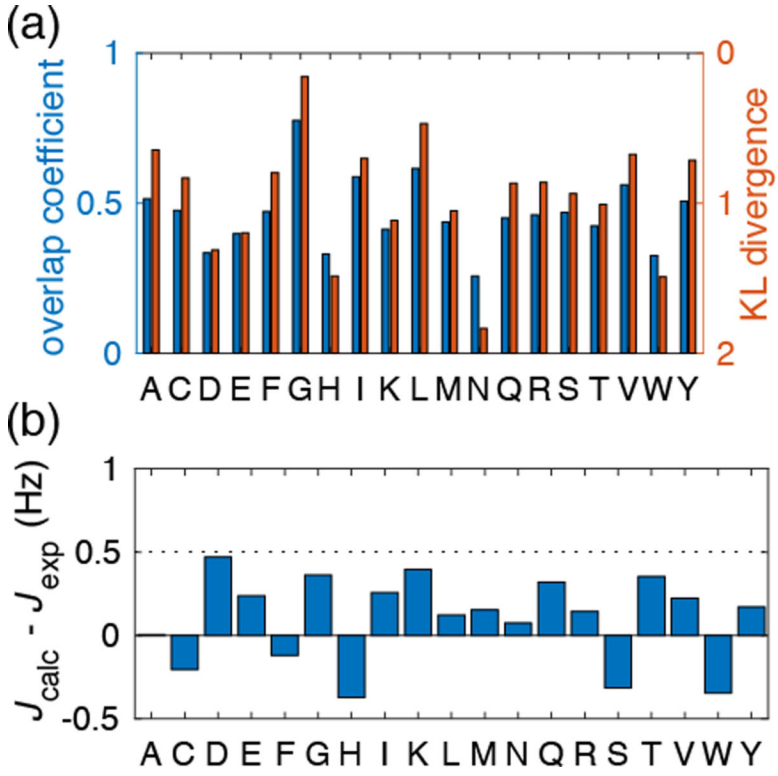


Figure 9: Evaluation of optimized landscapes.

(a) Overlap coefficient *OC* and Kullback-Leibler divergence D_{KL} (see text for definitions) to quantify the (dis)similarities between the unoptimized coil-library and optimized landscapes for nineteen of the twenty amino acids (except Pro). (b) Differences between NMR scalar coupling constants calculated by applying the Karplus equation to the optimized landscape data (J_{calc}) and those determined experimentally (J_{exp}) [16c]. Absolute error = 0.10 Hz, standard deviation = 0.26 Hz, and mean absolute error = 0.24 Hz. These results are to be contrasted with the results shown in Figure 2 for J_{calc} obtained from the unoptimized coil library.