Bioinformatics, 2023, **39**, i357–i367
https://doi.org/10.1093/bioinformatics/btad251
ISMB/ECCB 2023

Getting '$\phi\psi\chi$al' with proteins: minimum message length inference of joint distributions of backbone and sidechain dihedral angles

Piyumi R. Amarasinghe[1], Lloyd Allison[1], Peter J. Stuckey[1,2], Maria Garcia de la Banda[1,2], Arthur M. Lesk[3], Arun S. Konagurthu[1,*]

[1]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia
[2]OPTIMA ARC Industrial Training and Transformation Centre, Carlton, VIC 3053, Australia
[3]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, United States

*Corresponding author. Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia. E-mail: arun.konagurthu@monash.edu

Abstract

The tendency of an amino acid to adopt certain configurations in folded proteins is treated here as a statistical estimation problem. We model the joint distribution of the observed mainchain and sidechain dihedral angles ($\langle\phi, \psi, \chi_1, \chi_2, \ldots\rangle$) of any amino acid by a mixture of a product of von Mises probability distributions. This mixture model maps any vector of dihedral angles to a point on a multi-dimensional torus. The continuous space it uses to specify the dihedral angles provides an alternative to the commonly used rotamer libraries. These rotamer libraries discretize the space of dihedral angles into coarse angular bins, and cluster combinations of sidechain dihedral angles ($\langle\chi_1, \chi_2, \ldots\rangle$) as a function of backbone $\langle\phi, \psi\rangle$ conformations. A 'good' model is one that is both concise and explains (compresses) observed data. Competing models can be compared directly and in particular our model is shown to outperform the Dunbrack rotamer library in terms of model complexity (by three orders of magnitude) and its fidelity (on average 20% more compression) when losslessly explaining the observed dihedral angle data across experimental resolutions of structures. Our method is unsupervised (with parameters estimated automatically) and uses information theory to determine the optimal complexity of the statistical model, thus avoiding under/over-fitting, a common pitfall in model selection problems. Our models are computationally inexpensive to sample from and are geared to support a number of downstream studies, ranging from experimental structure refinement, *de novo* protein design, and protein structure prediction. We call our collection of mixture models as `PhiSiCal` ($\phi\psi\chi$al).

Availability and implementation: PhiSiCal mixture models and programs to sample from them are available for download at http://lcb.infotech.monash.edu.au/phisical.

1 Introduction

The 20 naturally occurring amino acids form the nature's part list from which proteins are made within the cells of organisms. In all amino acids a central carbon atom (the α-carbon) binds an amino group (-NH_2), a carboxylic acid (-COOH) group, and a hydrogen atom, but differ in the fourth group attached, a sidechain (R).

Protein polypeptide chains of amino acids fold into compact three-dimensional shapes stabilized by inter-atomic interactions between the amino acids. The resultant amino acid conformations are determined by the varying degrees of rotations ('torsions') around the atomic bonds, subject to the physics and chemistry of protein folding.

Any torsion can be mathematically calculated as a 'dihedral angle'—the angle between two planes—defined by four points (here, the coordinates of successively bonded atoms) sharing a common basis vector (here, the central bond around which the torsion is being measured) (IUPAC-IUB Commission, 1970). Thus, any amino acid conformation can be described as a vector of dihedral angles, conventionally denoted by the sequence of symbols, $\langle\phi, \psi, \omega, \chi_1, \chi_2, \ldots\rangle$ (see Fig. 1).

Across all amino acids, the symbols $\langle\phi, \psi, \omega\rangle$ are used to denote the dihedral angles around the backbone bonds, whereas $\langle\chi_1, \chi_2, \ldots\rangle$ are used to denote exclusively the torsions around the sidechain bonds. Note that the number of sidechain dihedral angles depends on the sidechain (R) groups, and hence varies with the amino acid type.

Analysis of the observed distributions of backbone and sidechain dihedral angles has been an object of intense interest since the early protein structural and biophysical studies: Ramachandran et al. (1963), Janin and Wodak (1978), McGregor et al. (1987), Dunbrack and Karplus (1993), Dunbrack and Cohen (1997), Dunbrack (2002), and Shapovalov and Dunbrack (2007, 2011). This interest is fuelled by the need for accurate statistical models that can effectively characterize the observed dihedral angle distributions of proteins, as these models are used by techniques for protein experimental structure determination, computational prediction, rational design, and many other protein structural analyses.

One of the results has been the creation of rotamer libraries. A 'rotamer' is any rotational preference of the set of dihedral angles along the sidechain bonds within amino acids. These libraries are compiled from the statistical clustering of sidechain conformations of known protein structures (Dunbrack 2002). Rotamer libraries are 2-fold: backbone independent and backbone dependent. Backbone-dependent rotamer libraries contain rotameric preferences conditioned

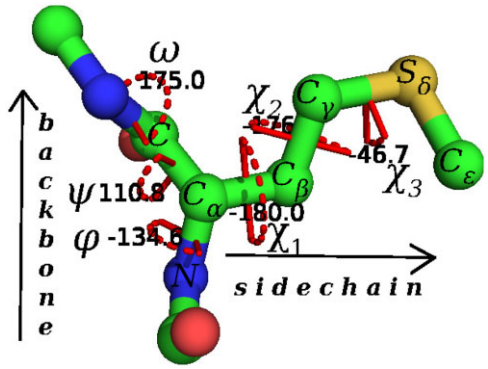


Figure 1. The amino acid Methionine (MET) has its conformation specified by six dihedral angles: $\langle \phi, \psi, \omega, \chi_1, \chi_2, \chi_3 \rangle$, where each angle is in the range $(-180^\circ, 180^\circ]$. (The angles shown above are those observed for MET67 in the fibroblast growth factor protein, 1BAR. Note that the value of $\chi_1 = -180^\circ$ for the C_α–C_β bond corresponds to the *trans* conformation.) For MET, the sidechain, or R group, is -C_β-C_γ-S_δ-C_ϵ.

on any observed backbone dihedral angles (Dunbrack and Karplus 1993; Dunbrack and Cohen 1997; Shapovalov and Dunbrack 2011), and differ from the backbone-independent libraries which simply cluster sidechain conformations agnostic to the backbone conformation of amino acids (Ponder and Richards 1987; Lovell et al. 2000).

Rotamer libraries derive sidechain conformation statistics using coarse quantization of the observed rotation space for each sidechain dihedral angle. This discretization often uses an angular interval of 120° regions, yielding a $(-60^\circ, 60^\circ, 180^\circ)$ trisection of the rotational space, that corresponds to the staggered conformation of two sp^3-hybridized atoms (Dunbrack 2002). Under such a discretization, each rotamer clusters around a mean conformational preference over a discretized interval. Such rotameric descriptions of sidechain torsions have the advantage of yielding a computationally tractable conformation space when inferring rotational preferences of individual amino acids and fitting them in several protein modelling tasks [e.g. in *de novo* protein design (Desmet et al. 1992)].

However, such discretizations can also bias downstream studies, e.g. leading to inaccurate modelling of the details of inter-atomic interactions for protein docking (Wang et al. 2005), and to imprecise protein conformational energy landscapes (Grigoryan et al. 2007), among others (Lassila 2010). Further, several of the outermost dihedral angles of certain amino acids – χ_3 of glutamic acid (GLU) and glutamine (GLN), χ_2 of aspartic acid (ASP), and asparagine (ASN) – flout the three-way discretization of its rotational space and hence lead to broad and visually featureless distributions that have resisted attempts to characterize the observed spread accurately (Lovell et al. 1999; Shapovalov and Dunbrack 2011). As discussed by Schrauber et al. (1993), in these instances the rotameric representation of sidechain conformations is limited and large deviations of χ angles from the canonical values can be observed. The existence of such 'non-rotameric' conformations was also discussed in detail by Heringa and Argos (1999).

An approach employed to mitigate this issue is to calculate distribution frequencies on a finer grid (Schrauber et al. 1993). A more accurate approach is to model the distribution over a continuous space, as this would result in a finer representation minimizing information loss. This is the approach taken by BASILISK (Harder et al. 2010) which formulates a probabilistic model that represents the torsion angles in a continuous space. However, it uses a single probabilistic model for all the amino acids.

The Dunbrack rotamer library (Dunbrack and Karplus 1993; Dunbrack and Cohen 1997; Dunbrack 2002; Shapovalov and Dunbrack 2007, 2011) is a continually maintained and improved rotamer library. It defines the state of the art and is among the most widely used rotamer libraries across many downstream applications that employ them. While this library is backbone dependent, it uses the same supervised-discretized choices. This discretization renders their resultant models both overly complex as well as inaccurate in capturing the observed distributions of dihedral angles when sampled from its libraries (see Section 3).

In this work, we take a different approach by modelling the joint distributions of the observed mainchain and sidechain dihedral angles of individual amino acids by a mixture of a product of von Mises probability distributions. To infer these mixture models, we use the Bayesian and information-theoretic criterion of minimum message length (MML) (Wallace and Boulton 1968; Wallace and Freeman 1987; Wallace 2005). In the theory of learning and generalization, this unsupervised model selection framework falls under the class of statistical inductive inference (Wallace 2005). Among other notable and well-established statistical properties, MML allows an objective trade-off between model complexity and fit—these form two opposing criteria that all model selection problems contend with, but for which MML provides an intuitive, objective, and rigorous reconciliation.

We compared our mixture models inferred for each amino acid with the Dunbrack rotamer library on large datasets containing structures that are non-redundant in sequence and filtered based on high-resolution, B-factor, and R-factor cutoffs. Our results clearly demonstrate that the mixture models we infer outperform the Dunbrack rotamer library both in its model complexity (by three orders of magnitude) and its fidelity (yielding on average 20% more lossless compression) when explaining the observed dihedral angle data. Our MML mixture model library, termed '$\phi\psi\chi$al' supports fast sampling of joint and conditionally distributed dihedral angle vectors to support their use in many downstream studies involving protein structures.

2 Methods

2.1 Mixture model overview

We present a systematic method of 'unsupervised' estimation of a statistical model that can effectively explain any given observations of 'vectors' (of any dimension) of dihedral angles using the statistical inductive inference framework of MML (Wallace and Boulton 1968; Wallace 2005; Allison 2018).

Specifically, this work infers a 'mixture model' under the Bayesian and information-theoretic criterion of MML, where each component of the mixture defines a 'product' of a series of von Mises distributions (Mardia et al. 2000), one for each dihedral angle observed in the specified amino acid. We note that the number of components, their probabilities, and corresponding parameters are all unknown and are inferred unsupervised by our method.

Formally, for a specified amino acid 'aa' (i.e. any of the 20 naturally occurring amino acids in proteins), $X = \{x_1, x_2, \ldots, x_N\}$ represents an input set of N observations of the conformational

states of that amino acid. Each $x_i \in X$ defines a vector of the d dihedral angles (whose terms are specified in some canonical order) as observed in the i-th instance of 'aa'. For example, each instance of the amino acid methionine (see Fig. 1) is defined by a $d = 6$-dimensional vector containing its dihedral angles $\langle \phi, \psi, \omega, \chi_1, \chi_2, \chi_3 \rangle$. In this case, X captures the set of observed instances of various conformational states of methionine derived from a non-redundant set of experimental coordinates in the world-wide protein data bank (Berman et al. 2000).

A 'mixture model' is any convex combination of 'component' probability density functions used to explain some observed data containing a number of subpopulations (often unknown in advance) within an overall population (Figueiredo and Jain 2002; McLachlan et al. 2019). Specifically, in this work, we consider a mixture model that takes the general form:

$$\mathcal{M}(\Lambda) = \sum_{j=1}^{|\mathcal{M}|} w_j f(\Theta_j) \text{ such that } \sum_{j=1}^{|\mathcal{M}|} w_j = 1. \quad (1)$$

This defines a continuous probability distribution for a d-dimensional random vector

$$x_i = \langle x_{i_1},\ x_{i_2},\ \ldots,\ x_{i_d} \rangle$$

such that $x_{i_p} \in (-\pi, \pi], \forall 1 \leq p \leq d$. Thus, the support for x_i defines a surface of a d-Torus (denoted as $\mathbb{T}^d$). $|\mathcal{M}| \in \mathbb{Z}^+$ denotes the size of the mixture model given by the number of 'components' it defines. Each component function $f(\Theta_j)$ denotes the joint probability distribution of the random vector $x_i \in \mathbb{T}^d$. In this work, each mixture component takes the form of a product of d von Mises circular distributions, $f(\Theta_j) \propto \prod_{p=1}^{d} \exp(\kappa_{j_p} \cos(x_{i_p} - \mu_{j_p}))$, where each $\langle \mu_{j_p}, \kappa_{j_p} \rangle$ represent the $\langle \text{mean}, \text{concentration} \rangle$ parameters of each von Mises term in the product and $\Theta_j = \{\langle \mu_{j_p}, \kappa_{j_p} \rangle\}_{\forall 1 \leq p \leq d}$ denotes the collection of all von Mises' parameters of the j-th mixture component. Each w_j denotes a mixture components' respective 'weight' which, over all $|\mathcal{M}|$ terms in the mixture, add up to 1. Finally, we use Λ as a shorthand to collectively denote all mixture model's parameters:

1) the 'number' of mixture components $|\mathcal{M}|$,
2) the set of 'weights' of mixture components $\{w_j\}_{\forall 1 \leq j \leq |\mathcal{M}|}$, and
3) the set of all parameters defining the mixture 'components' $\{\Theta_j\}_{\forall 1 \leq j \leq |\mathcal{M}|} \equiv \{\{\langle \mu_{j_p}, \kappa_{j_p} \rangle\}_{\forall 1 \leq p \leq d}\}_{\forall 1 \leq j \leq |\mathcal{M}|}$.

Thus, for any specified amino acid 'aa' with its given set of dihedral angle tuples X, the goal of this work is to infer a mixture model $\mathcal{M}$ that best explains all the observations in X. The key challenge in doing so is to estimate the mixture parameters Λ unsupervised. To address this unsupervised estimation problem, we employ the Bayesian and information-theoretic criterion of MML, as follows.

2.2 MML inference foundations

2.2.1 MML and model selection

MML is a Bayesian method for hypothesis/model selection. In general terms, if X is some given data and M is some statistical model describing that data, the joint probability of the model M and data X is given by the product rule of probability: $\Pr(M, X) = \Pr(M)\Pr(X|M)$. This can be recast in terms of Shannon information based on the observation that the optimal code length to represent any event E (with a probability $\Pr(E)$) is given by the measure of Shannon information content quantified (say in bits of information) as $I(E) = -\log_2(\Pr(E))$ (Shannon 1948). Expressing the above product rule of probability in terms of Shannon information content, we get:

$$\underbrace{I(M, X)}_{\text{Total Message Length}} = \underbrace{I(M)}_{\text{first part}} + \underbrace{I(X|M)}_{\text{second part}}. \quad (2)$$

In the above equation, the amount of information required to losslessly explain the observed data X with a hypothesis/model M can be seen as the length of a two-part message: the 'first part' contains the information required to state the model M losslessly (quantifying the model's descriptive 'complexity'), whereas the 'second part' contains the information required to state the data X 'given' the model M (quantifying the model's 'fit' with the data). It is easy to see that, in this information-theoretic view, the best model M^* is the one whose total two-part message is minimum (optimally trading-off the model's complexity and fit): $M^* = \arg\min_{\forall M} I(M, X)$. This is equivalent to maximizing the joint probability $\arg\max_{\forall M} \Pr(M, X)$. Thus, under the MML framework, any pair of competing models explaining the same data can be compared based on their respective total lengths: the difference in total message lengths derived using any two models gives their log-odds posterior ratio, making this method of model selection Bayesian (Wallace 2005; Allison 2018).

2.2.2 Wallace–Freeman method of parameter estimation using MML

Let $M(\alpha)$ denote a twice-differentiable statistical model with a parameter vector α (with $|\alpha|$ number of free parameters) and X denote some observed data (containing $|X|$ number of observations). Wallace and Freeman (1987) showed that the total message length of any general model M with a vector of parameters α can be approximated as

$$I(M(\alpha), X) \approx \underbrace{\log\left(\frac{\sqrt{\det(\mathcal{F}(\alpha))}\ \sqrt[\frac{2}{|\alpha|}]{q_{|\alpha|}}}{h(\alpha)}\right)}_{\text{First part: } I(M(\alpha))} + \underbrace{\mathcal{L}(\alpha) - |X||\alpha|\log(\epsilon) + \frac{|\alpha|}{2}}_{\text{Second part: } I(X|M(\alpha))}, \quad (3)$$

where $h(\alpha)$ is the prior probability density of the parameters α, $\det(\mathcal{F}(\alpha))$ is the determinant of the 'expected' Fisher information matrix, $\mathcal{L}(\alpha)$ is the negative log-likelihood function of X given α, $q_{|\alpha|}$ represents the Conway–Sloane (Conway and Sloane 1984) lattice quantization constant in $|\alpha|$-dimensional space, and ϵ is the uncertainty of each datum in the set X of size $|X|$. Refer to Wallace (2005) and Allison (2018) for details of this method of estimation.

This Wallace and Freeman (1987) method informs the computation of various message length terms in the work presented here.

2.3 Message length of a mixture model

Applying the general MML framework to the mixture models introduced in Section 2.1 allows us to characterize the length of the message needed to explain jointly any observed set of dihedral angle vectors X using a mixture model $\mathcal{M}$ with parameter vector Λ analogously to Equation (2) as

$$I(\mathcal{M}(\Lambda), X) = I(\mathcal{M}(\Lambda)) + I(X|\mathcal{M}(\Lambda)). \tag{4}$$

This in turn is used to define the objective function we use to estimate an optimal set of mixture model parameters that can losslessly explain itself ($\mathcal{M}(\Lambda)$) and the observations X in the most succinct way in terms of Shannon information: $\Lambda_{\text{MML}} = \arg\min_{\forall\Lambda} I(\mathcal{M}(\Lambda), X)$.

2.3.1 Computing $I(\mathcal{M}(\Lambda))$ term of Equation (4)

As described in Section 2.1, Λ denotes the combined set of mixture model parameters $(|\mathcal{M}|, \{w_j\}_{\forall 1 \le j \le |\mathcal{M}|}, \{\Theta_j\}_{\forall 1 \le j \le |\mathcal{M}|})$. Thus, the Shannon information content in a mixture model can be expressed as the summation of the message lengths terms required to state all its parameters losslessly:

$$I(\mathcal{M}(\Lambda)) = \underbrace{I(|\mathcal{M}|)}_{\text{term 1}} + \underbrace{\sum_{j=1}^{|\mathcal{M}|} I(w_j)}_{\text{term 2}} + \underbrace{\sum_{j=1}^{|\mathcal{M}|} I(\Theta_j)}_{\text{term 3}}. \tag{5}$$

Computation of each of the message length terms on the right-hand side of Equation (5) is described below.

Computation of Term 1 of Equation (5)

$|\mathcal{M}| \in \mathbb{Z}^+$ is a countable positive integer and thus can be stated using an universal prior for integers over a variable-length integer code (Allison et al. 2019). We employ the Wallace Tree Code (Wallace and Patrick 1993; Allison et al. 2019) to compute $I(|\mathcal{M}|)$ in Equation (5).

Computation of Term 2 of Equation (5)

The set of L_1 normalized weight vector $\{w_j\}_{\forall 1 \le j \le |\mathcal{M}|}$ can be viewed as a parameter of a multinomial distribution, whose support defines a unit $(|\mathcal{M}| - 1)$ simplex (Wallace 2005; Allison 2018). Using the Wallace–Freeman method of estimation described in Section 2.2.2, assuming a uniform prior for the weights as a point in a unit $(|\mathcal{M}| - 1)$ simplex, i.e. the prior $h = (|\mathcal{M}| - 1)!/\sqrt{|\mathcal{M}|}$, and computing the determinant of the Fisher information matrix for a multinomial distribution (with parameters $\{w_j\}$) as $N^{|\mathcal{M}|-1}/\Pi_{j=1}^{|\mathcal{M}|} w_j$, it can be shown [as per the first part of Equation (3)] that the message length of Term 2 is given by (Allison 2018):

$$\sum_{j=1}^{|\mathcal{M}|} I(w_j) = \frac{(|\mathcal{M}| - 1)}{2}\log(q_{(|\mathcal{M}|-1)}) - \log\left(\frac{(|\mathcal{M}| - 1)!}{\sqrt{|\mathcal{M}|}}\right) + \frac{(|\mathcal{M}| - 1)}{2}\log(N) - \frac{1}{2}\sum_{j=1}^{|\mathcal{M}|}\log(w_j)$$

Computation of Term 3 of Equation (5)

Recall (from Section 2.1) that each $\Theta_j = \{\langle \mu_{j_p}, \kappa_{j_p} \rangle\}_{\forall 1 \le p \le d}$. Thus, $I(\Theta_j) = \sum_{p=1}^{d} I(\langle \mu_{j_p}, \kappa_{j_p} \rangle\})$. Each $I(\langle \mu_{j_p}, \kappa_{j_p} \rangle\})$ term in the summation is estimated by again applying the Wallace–Freeman method (Section 2.2.2), this time for a von Mises circular distribution. A von Mises distribution defines a probability distribution of a random variable x on a circle (i.e. $x \in (-\pi, \pi]$) as a function of its two free parameters, mean $\mu \in (-\pi, \pi]$ and concentration $\kappa > 0$: $f(x; \langle \mu, \kappa \rangle) = \frac{\exp^{\kappa\cos(x-\mu)}}{2\pi B_0(\kappa)}$, where the denominator on the right-hand side gives the normalization constant of the distribution in terms of the modified Bessel function (of order 0), denoted here as $B_0(\kappa)$. More commonly, modified Bessel functions of order r are denoted as $I_r(\cdot)$. We use B_r here only to avoid confusion with the Shannon information content notation, $I(\cdot)$.

In applying the Wallace–Freeman method, the assumed priors for the two parameters are [as per Kasarapu and Allison (2015)]: $h(\mu) = \frac{1}{2\pi}$ and $h(\kappa) = \frac{\kappa}{(1+\kappa^2)^{\frac{3}{2}}}$. Thus, $h(\langle \mu, \kappa \rangle) = h(\mu)h(\kappa)$. We note that the rationale and behaviour of these priors for von Mises has been previously studied (Wallace 2005). The chosen prior on μ is uniform (and hence uninformative/flat), giving only general information about the variable being estimated, which makes it suitable. On the other hand, no truly uninformative prior exists for κ. The chosen prior ensures the function is smooth (without singularities) and commonly preferred when the data concentration is expected to arise from physical interactions (Wallace 2005).

Further, for some N observations of circular angles in the range $(-\pi, \pi]$ defined by (say) the set $X = \{x_1, x_2, \ldots, x_N\}$, it can be shown that the 'determinant' of the expected Fisher information matrix for a von Mises distribution can be characterized as $\det(\mathcal{F}(\langle \mu, \kappa \rangle)) = \kappa N A(\kappa) A'(\kappa)$, where $A(\kappa) = \frac{B_1(\kappa)}{B_0(\kappa)}$ and $A'(\kappa) = \frac{\mathrm{d}}{\mathrm{d}\kappa} A(\kappa)$. Using this prior and determinant, the message length term to state the pair of $\langle \mu, \kappa \rangle$ parameters of any single von Mises circular distribution [as per the first part of Equation (3)] can be written as

$$I(\langle \mu, \kappa \rangle) = \log(q_2) - \log(h(\langle \mu, \kappa \rangle)) + \frac{1}{2}\log(\det(\mathcal{F}(\langle \mu, \kappa \rangle))). \tag{6}$$

2.3.2 Computing $I(X|\mathcal{M}(\Lambda))$ term of Equation (4)

The second part of Equation (4) deals with explaining the observations of the vectors of dihedral angles X using the mixture model parameters that have been stated losslessly via the first part (Section 2.3.1). Using the relationship between Shannon information and probability (Section 2.1), that is, $I(\cdot) = -\log(\Pr(\cdot))$, $I(X|\mathcal{M}(\Lambda))$ can be decomposed using the likelihood of each d-dimensional dihedral angle $x_{i_p} \in x_i \in X$ (assuming independent and identically distributed datum) using the mixture model parameters as

$$I\Big(X|\mathcal{M}\Big(\Lambda\Big)\Big) = \sum_{i=1}^{N} -\log\left(\sum_{j=1}^{|\mathcal{M}|}\left(w_j \Pi_{p=1}^{d} f(x_{i_p}|\langle \mu_{j_p}, \kappa_{j_p} \rangle)\epsilon^d\right)\right),$$

where ϵ in the above expression denotes the degree of uncertainty of each dihedral angle x_{i_p} to estimate its component likelihood over a von Mises distribution. This work sets $\epsilon = 0.0873$ radians, based on the observation that the effective precision of 3D atomic coordinate is not better than 0.1Å (Konagurthu et al. 2014).

2.4 Search for optimal mixture model parameters

2.4.1 Expectation–maximization (EM)

To search for an optimal mixture model $\mathcal{M}(\Lambda_{\text{MML}})$ that minimizes Equation (4), we employ a deterministic EM algorithm commonly employed for statistical parameter estimation problems (Dempster et al. 1977; McLachlan and Basford 1988; McLachlan et al. 2019). EM is an iterative algorithm which, in each iteration, explores local updates to the current parameter estimates to be able to generate new parameter estimates that yield progressively shorter message lengths [in this work, the evaluation of Equation (4)] until convergence.

Let $\Lambda(t)$ denote the state of the mixture parameters at an iteration indexed by $t \geq 0$. Then at each iteration indexed as $\{1, 2, \ldots, t, t+1, \ldots\}$ the EM performs an *E(xpectation)*-step followed by a *M(aximization)*-step, as described below.

E-step

Using the current state of parameter estimates after iteration t, i.e. $\Lambda(t)$, the E-step calculates the (probabilistic) 'responsibilities' $r_{ij}(t+1) \forall 1 \leq i \leq N, 1 \leq j \leq |\mathcal{M}|$ in the next iteration $t+1$ as

$$r_{ij}(t+1) = \frac{w_j(t) f(x_i|\Theta_j(t))}{\sum_{j'=1}^{|\mathcal{M}|} w_{j'}(t) f(x_i|\Theta_{j'}(t))}. \tag{7}$$

Formally responsibility r_{ij} is the posterior probability that x_i belonging to j and it quantifies the degree to which a component j 'explains' the data point x_i (McLachlan et al. 2019). From these responsibilities, given N observations of dihedral angles, any j-th component's membership in iteration $t+1$ is calculated as

$$n_j(t+1) = \sum_{i=1}^{N} r_{ij}(t+1) \quad \text{and} \quad \sum_{j=1}^{|\mathcal{M}|} n_j(t+1) = N.$$

M-step

In the M-step, the mixture parameters are updated as follows. The set of weights for $t+1$ are derived as the MML estimates of parameters of a multistate distribution (Allison 2018) with N observations over $|\mathcal{M}|$ distinct states while treating $\{n_j(t+1)\}_{\forall 1 \leq j \leq |\mathcal{M}|}$ as each component/state's number of observed instances (out of N):

$$w_j(t+1) = \frac{n_j(t+1) + \frac{1}{2}}{N + \frac{|\mathcal{M}|}{2}}. \tag{8}$$

Further, the update to each mean parameter of a von Mises distribution ($\forall 1 \leq j \leq |\mathcal{M}|, 1 \leq p \leq d$) is given by

$$\mu_{j_p}(t+1) = \frac{R_{j_p}}{\|R_{j_p}\|}, \tag{9}$$

where R_{j_p} is the 'vector sum' of each x_{i_p}th dihedral angle in the tuple $x_i \in X$, weighted by its corresponding responsibility $r_{ij}(t+1)$. We note that this vector sum arises because each dihedral angle is written as a 2D trigonometric coordinate $(\cos x_{i_p}, \sin x_{i_p})$ on a unit circle. $\|R_{j_p}\|$ is the vector norm of the resultant vector R_{j_p}.

Finally, the update to the concentration parameter κ_{j_p} of von Mises distribution ($\forall 1 \leq j \leq |\mathcal{M}|, 1 \leq p \leq d$) follows a numerical approach, as solving for the roots of $\frac{\partial}{\partial \kappa} I(\langle \mu, \kappa \rangle, X_p) = 0$ has no closed form (see Supplementary Section S1).

2.4.2 Search for the optimal number of mixture components, $|\mathcal{M}|$

A priori, the number of mixture components $|\mathcal{M}|$ is unknown, along with other mixture parameters. Thus, the EM algorithm starts with a single component mixture model at iteration $t = 0$ (i.e. $|\mathcal{M}| = 1$). It then follows similar mechanics to that described by Kasarapu and Allison (2015), albeit with some improvements.

Starting from a single-component mixture at $t = 0$, during each iteration $(t+1)$, a set of perturbations, `Split`, `Merge`, and `Delete` are systematically executed on each component of the mixture model $\Lambda(t)$. We note that each `Split` of a component increases the number of components $|\mathcal{M}|$ by $+1$, whereas `Merge` and `Delete` decrease it by -1. After each such perturbation, the parameters of the resulting new mixture (with increased/decreased number of components) are reestimated using EM updates described in Section 2.4.1 starting with initial parameters assigned deterministically at the E-step. After systematically exploring all of the above perturbations on each component, the perturbation that yields the best improvement to the message length [as per Equation (4)] is chosen going into the next iteration, and so on, until convergence.

The rationale of each `Split`, `Merge`, and `Delete` operations together with the full details of their mechanics are provided in Supplementary Section S2. Furthermore, Supplementary Section S11 demonstrates the stability and convergence of this search process.

3 Results and discussion

3.1 Datasets and benchmarks

3.1.1 Curating the dihedral angle datasets

Atomic coordinates of 38,895 protein structures with non-redundant amino acid sequences ($\leq$ 50% sequence identity) were derived from the Protein Data Bank (Berman et al. 2000), considering only structures with an R-factor cut-off at 0.3 and resolution cut-off at 3.5 Å or better. We call this collection PDB50. Further, as a way to test the effect that precision of input data has on the inferred models, we also consider another ($\leq$ 50% sequence identity) dataset containing 9568 high-resolution ($\leq$ 1.8 Å) X-ray structures with a B-factor cut-off of 40 and R-factor cut-off of 0.22. We call this collection PDB50HighRes.

For a complete atomic coordinate record of each amino acid observed in any considered structure, we calculate a vector of backbone and sidechain dihedral angles: $\{\phi, \psi, \omega, \chi_1, \chi_2, \ldots\}$. (We note that the partial double-bond characteristic of peptide bond makes ω typically $\sim 180^\circ$ and rarely $\sim 0^\circ$. Thus, for our inference, ω dihedrals were ignored from the input set.) Overall, this resulted in 22,177,093 observations (vectors of dihedral angles) from PDB50 and 3,774,207 observations for PDB50HighRes, considering only the atomic coordinates of 20 natural amino acids within proteins. We then partitioned these observations into 20 sets of amino acid specific dihedral angle vectors ($X^{(\text{aa})}$), one for each distinct amino acid (aa).

Table 1. PDB50 dataset statistics: amino acid type (aa), number of observations of that amino acid in PDB50 ($N^{(aa)}$), and the total number of (backbone + sidechain) dihedral angles in that amino acid ($d^{(aa)}$).

aa	$N^{(aa)}$	$d^{(aa)}$	aa	$N^{(aa)}$	$d^{(aa)}$	aa	$N^{(aa)}$	$d^{(aa)}$
LEU	2,171,630	4	ASP	1,279,567	4	GLN	820,871	5
ALA	1,861,359	2	THR	1,221,604	3	TYR	788,176	4
VAL	1,601,058	3	LYS	1,176,395	6	HIS	515,611	4
GLY	1,588,115	2	ARG	1,130,448	7	MET	417,170	5
GLU	1,446,860	5	PRO	1,004,859	4	TRP	310,470	4
SER	1,337,273	3	ASN	948,274	4	CYS	296,547	3
ILE	1,333,508	4	PHE	927,298	4			

The counts in $d^{(aa)}$ ignore the ω dihedral angle.

Table 1 gives the breakdown of the number of observations per amino acid type, along with their corresponding number of (backbone + sidechain) dihedral angles. For each of these amino acid specific input sets $X^{(aa)}$, its corresponding mixture model $\mathcal{M}(\Lambda^{(aa)})$ (one for PDB50 dataset and another for PDB50HighRes dataset) was inferred and their parameters estimated automatically using the MML methodology (described in Section 2).

3.1.2 Dunbrack backbone-dependent rotamer libraries

We benchmark the performance and fidelity of our inferred mixture models against the latest version of the Dunbrack 'backbone-dependent' rotamer (sidechain conformation) libraries (Shapovalov and Dunbrack 2011), across varying degrees of smoothing [2%, 5% (default), 10% and 20%] that those libraries provide. The Dunbrack libraries define the state of the art for modelling and sampling sidechain conformations, 'conditioned' on any stated backbone dihedral angles $\langle\phi,\psi\rangle$. Specifically, the Dunbrack rotamer library discretizes each amino acid's backbone dihedral angles $\langle\phi,\psi\rangle$ into $36^2 = 1296$ bins (of $10^\circ \times 10^\circ$ granularity). For each $\langle\phi,\psi\rangle$ bin, there are commonly 3^m models. Here, 3 arises from the three-way discretization of each sidechain dihedral angle into {gauche+ (g+), trans (t), gauche- (g-)} states, whereas m denotes the number of 'sidechain' dihedral angles $\langle\chi_1,\chi_2,\ldots\rangle$ in that amino acid. For example, amino acid, methionine has $m=3$ and the Dunbrack rotamer library lists $36\times 36\times 3^3 = 34,992$ models across its 1296 possible $\langle\phi,\psi\rangle$ bins. The Dunbrack rotamer library divides the set of amino acid types into 'rotameric' and 'non-rotameric' categories. The use of the closed-form computation of 3^m models holds for all 'rotameric' amino acids, whereas the 'non-rotameric' amino acids (glutamic acid, glutamine, aspartic acid, asparagine, tryptophan, histadine, tyrosine, and phenylalanine) have more components, as some of their sidechain dihedrals do not conform to three-way discretizations.

3.2 Information-theoretic complexity versus fidelity/fit of the inferred models

In almost all model selection problems, one seeks answers to two key questions: (i) What is the fidelity of the model in its ability to explain observed data? (ii) How complex is the selected model?. The second question is necessary for when there is a simpler model (in complexity terms) that can explain/fit the same data equivalently or better than a more complex model, then the simpler model is preferred not only due to Ockham's razor, but also made rigorous by the Bayes theorem (Allison 2018).

The information-theoretic framework of MML provides a direct way to quantify model complexity and fit in terms of bits. For any proposed model, the total two-part message length combines (i) the lossless encoding of the model, the length (bits) of which yields the model's (descriptive) complexity, and (ii) the lossless encoding of the observed data given that model, the length (bits) of which yields its fidelity by quantifying how well the model fits the data (see Section 2.2).

Table 2 gives the complexity and fidelity statistics of our inferred models and compares it directly with the state-of-the-art Dunbrack rotamer library at 5% ('default') smoothing level (see Supplementary Section S12 for results on other smoothing levels). Before we discuss these quantitative results, let us explore how/why they can be evaluated fairly, and on an equal footing.

For each of the 1296 bins in the Dunbrack library, the information in their library can be directly translated as a bin-wise mixture model with a fixed number of mixture components, where each component contains a product of m von Mises circular distributions, and m is the number of sidechain dihedral angles for the specified amino acid (aa). [We note that amino acids alanine (ALA) and glycine (GLY) have no sidechain dihedral angles, so the Dunbrack library do not have any models for ALA and GLY.] However, as mentioned above, the number of components of the each of those 1296 mixture models related to an amino acid is static/fixed and corresponds to the number of discrete states over m sidechain angles (often three-way for each sidechain dihedral angle χ, as discussed earlier). Thus, the number of mixture components for each of the $\langle\phi,\psi\rangle$ bin is usually 3^m which yield a large number of models across all bins (e.g. 34,992 for methionine as shown in Table 2). This number matters, as it is proportional to the number of von Mises parameters (and respective mixtures' weights) that informs the complexity of the statistical model being proposed. In contrast, the MML mixture model infers only one mixture model for any amino acid, jointly over all (backbone + sidechain) dihedral angles with all of its mixture parameters estimated unsupervised, including the number of mixture components $|\mathcal{M}^{(aa)}|$.

Comparing the model fit/fidelity is more involved: while our work models the joint distributions over all (backbone + sidechain) dihedral angles, Dunbrack's only deals with sidechain dihedrals conditioned on discretized states of the backbone. With this difference in the models, there are two possible directions to take to ensure the comparison of fidelity between the two is on the same footing. For any set of observations of all dihedral angles for a specified amino acid $X^{(aa)}$:

1) The ϕ and ψ under Dunbrack model are stated over a uniform distribution—for this is precisely their underlying model—so that the message length of stating each vector of dihedrals using both models can be objectively compared. We show these results for PDB50 in the main text (see Table 2). Results for PDB50HighRes are included in Supplementary Section S4.
2) From each MML-inferred mixture model, we drop/omit the von Mises circular terms corresponding to backbone dihedral angles when estimating the length, yielding the second part of the message for only the sidechain dihedral angles of the observations. These results are presented in Supplementary Sections S3 (for PDB50) and S5 (for PDB50HighRes).

The above two ways of comparing the fidelity of the two models yield a similar conclusion: the MML-inferred mixture

Table 2. Quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(aa)}$) and that of the Dunbrack rotamer library ($\mathcal{D}^{(aa)}_{rotamer}$).

		MML mixture model ($\mathcal{M}^{(aa)}$) message length statistics in bits (rounded)					Dunbrack rotamer library ($\mathcal{D}^{(aa)}_{rotamer}$) message length statistics in bits (rounded)					Null model (raw) in bits	
(aa)	$N^{(aa)}$	($\lvert\mathcal{M}^{(aa)}\rvert$; $\lvert\Lambda^{(aa)}\rvert$)	First part (complexity)	Second part (fit)	Total (complexity + fit)	$\frac{\text{Total}}{N^{(aa)}}$	($\lvert\mathcal{D}^{(aa)}_{rotamer}\rvert$; #Params)	First part (complexity)	Second part (fit)	Total (complexity + fit)	$\frac{\text{Total}}{N^{(aa)}}$	Null($X^{(aa)}$)	$\frac{\text{Null}(X^{(aa)})}{N^{(aa)}}$
LEU	2,171,630	(165; 1484)	7017	34,540,650	34,547,667	15.9	(11,664; 57,024)	1,079,722	46,109,408	47,189,130	21.7	53,595,177	24.7
ALA	1,861,359	(25; 124)	701	14,847,660	14,848,361	8.0	(N/A; N/A)	N/A	N/A	N/A	N/A	22,968,891	12.3
VAL	1,601,058	(96; 671)	3389	18,795,871	18,799,260	11.7	(3888; 10,368)	217,209	26,750,651	26,967,860	16.8	29,635,223	18.5
GLY	1,588,115	(30; 149)	746	15,965,309	15,966,055	10.1	(N/A; N/A)	N/A	N/A	N/A	N/A	19,597,101	12.3
GLU	1,446,860	(262; 2881)	12,205	33,234,644	33,246,849	23.0	(69,984; 488,592)	9,696,033	39,933,578	49,629,612	34.3	44,635,088	30.8
SER	1,337,273	(114; 797)	3825	18,289,465	18,293,291	13.7	(3888; 10,368)	210,730	23,624,303	23,835,033	17.8	24,752,622	18.5
ILE	1,333,508	(172; 1547)	7356	20,475,170	20,482,526	15.4	(11,664; 57,024)	964,619	27,688,670	28,653,289	21.5	32,910,577	24.7
ASP	1,279,567	(170; 1529)	6524	23,223,302	23,229,826	18.2	(23,328; 115,344)	2,336,817	27,634,793	29,971,610	23.4	31,579,330	24.7
THR	1,221,604	(90; 629)	3057	15,740,512	15,743,569	12.9	(3888; 10,368)	211,687	20,733,566	20,945,253	17.1	22,611,615	18.5
LYS	1,176,395	(266; 3457)	13,691	32,006,948	32,020,639	27.2	(104,976; 943,488)	14,337,386	37,818,245	52,155,632	44.3	43,549,614	37.0
ARG	1,130,448	(250; 3749)	15,898	32,987,603	33,003,501	29.2	(104,976; 943,488)	15,442,702	37,252,663	52,695,365	46.6	48,823,456	43.2
PRO	1,004,859	(231; 2078)	13,779	11,810,146	11,823,926	11.8	(2592; 11,664)	254,495	18,318,268	18,572,763	18.5	24,799,619	24.7
ASN	948,274	(180; 1619)	6793	17,855,829	17,862,622	18.8	(46,656; 231,984)	4,586,850	21,232,141	25,818,991	27.2	23,403,118	24.7
PHE	927,298	(226; 2033)	9365	15,950,596	15,959,961	17.2	(23,328; 115,344)	2,216,337	19,089,817	21,306,154	23.0	22,885,436	24.7
GLN	820,871	(239; 2628)	10,868	18,921,120	18,931,988	23.1	(139,968; 978,480)	18,417,683	23,167,291	41,584,974	50.7	25,323,563	30.8
TYR	788,176	(192; 1727)	7830	13,596,728	13,604,557	17.3	(23,328; 115,344)	2,248,951	16,184,209	18,433,160	23.4	19,451,947	24.7
HIS	515,611	(163; 1466)	6227	9,602,801	9,609,028	18.6	(46,656; 231,984)	4,373,651	11,419,682	15,793,334	30.6	12,725,125	24.7
MET	417,170	(270; 2969)	12,440	9,306,924	9,319,365	22.3	(34,992; 243,648)	4,222,664	11,504,102	15,726,767	37.7	12,869,538	30.8
TRP	310,470	(212; 1907)	8591	5,397,385	5,405,976	17.4	(46,656; 231,984)	4,062,897	6,659,922	10,722,819	34.5	7,662,306	24.7
CYS	296,547	(96; 671)	3148	3,943,308	3,946,457	13.3	(3,888; 10,368)	190,183	5,025,548	5,215,731	17.6	5,489,018	18.5

For each of the 20 naturally occurring amino acids (aa), $N^{(aa)}$ gives the size of the input set ($X^{(aa)}$) on which the comparison is based. $|\mathcal{M}^{(aa)}|$ gives the number of components of the mixture model, and $|\Lambda^{(aa)}|$ gives the number of parameters across all components of the mixture model, inferred unsupervised. $|\mathcal{D}^{(aa)}_{rotamer}|$ is the cumulative sum of all components described by the Dunbrack rotamer library, whereas #Params gives the corresponding total number of parameters implicit in their library. Across both models, the complexity (first part length in bits), fidelity (second part length in bits), and their two-part total are shown. The number of bits-per-residue for each of the models is also shown (the respective total message length by $N^{(aa)}$). Finally, to measure the extent of lossless compression each model provides, the null model message length of stating the vector of dihedral angles encoded under a uniform distribution is shown as a bottom-line. Note the 'N/A' terms across alanine (ALA) and glycine (GLY) arise because those amino acids do not have sidechain dihedral angles. While we model the joint distributions of dihedral including the backbone, Dunbrack on the other hand only provide sidechain distributions conditional on the backbone. Hence for ALA and GLY, Dunbrack library estimates are necessarily empty.

models (across all amino acid) are not only significantly more concise, but also explain the observed data better than the Dunbrack rotamer library (across the levels of smoothing they provide). Supplementary Section S9 provides a detailed explanation of how the lossless message length terms for Dunbrack's model are calculated.

Comparing the model complexity, Table 2 clearly shows that MML-inferred models are three orders of magnitude (in bits) more concise than those of the Dunbrack rotamer library. This is mainly due to the proliferation of the number of parameters in the Dunbrack model (see the eighth column of Table 2 under #nParams) compared with the lower number in the MML mixture model (third column under $|\Lambda^{(aa)}|$).

Further, comparing the model fidelity, all MML mixture models yield a better (lossless) explanation of the observed data than the corresponding Dunbrack models. The improvement varies with amino acids with most improvement observed for proline (PRO) where the second-part message length from MML mixture model is ∼35% shorter than Dunbrack. On the other end, for arginine (ARG) the improvement is ∼11%. The median improvement is ∼ 18% for glutamine (GLN). The mean sits at 20.1% improvement on PDB50 and 19.3% on PDBHighRes (Supplementary Table S2). Thus, from the results, it can be unambiguously concluded that the MML mixture models from this work outperform the state of the art in an objective quantitative comparison. Supplementary Sections S3 and S5 provide the alternative comparison between complexity and fit of the two models, involving the lossless comparison of sidechain dihedral angles and ignoring the backbone for PDB50 and PDB50HighRes.

Finally, we also assess how similar/different the inferred MML mixture models are across individual amino acids on the two datasets we have considered: PDB50 and PDB50HighRes. We use the measure of Kullback–Leibler (KL) relative entropy divergence that provides a direct way to compare two probability distributions. Supplementary Table S4 provides the KL-divergence values. The small KL-divergence across all amino acids indicates the proximity/similarity of the two inferred distributions. More generally, it has been demonstrated that the MML estimator is statistically robust to detect signal reliably even when the precision of input data varies (Wallace 2005).

3.3 Visualization of fidelity of the models

Here, we compare the fidelity of MML mixture models and Dunbrack rotamer library by randomly sampling 100,000 data points (vectors of dihedral angles) and contrasting the resultant distributions from the two models against the observed (empirical) distribution. The method of sampling from any MML-inferred mixture model and (for comparison) Dunbrack's library is described in Supplementary Section S10.

To be able to assess similarities and differences visually, we examine two specific amino acids, methionine (MET) and glutamine (GLN). We choose these pairs because (i) they both have three sidechain angles $\langle \chi_1, \chi_2, \chi_3 \rangle$, thus allowing their joint visualizations in 3D and (ii) MET falls into the 'rotameric' class of amino acids, whereas GLN falls into the 'non-rotameric' class (Shapovalov and Dunbrack 2011), hence providing a representation from those two classes for inspection.

Below we show these qualitative comparisons for the models inferred on the PDB50 dataset. The corresponding ones for PDB50HighRes are included in Supplementary Section S6.

Figure 2 clearly shows that the sampled points/vectors from the MML-inferred mixture model for both these cases are significantly closer to the empirical distribution of those respective amino acids than the points/vectors randomly sampled from the Dunbrack library, which are comparatively sparser. Although the sampled points cover the main rotameric preferences, they do fall short in modelling the details of the spread seen in the empirical distribution, which the MML mixture model does well in explaining. This visualization is a qualitative demonstration of the clear quantitative difference we observed in their second part message length terms (which quantifies fidelity/fit in bits of information) shown earlier in Table 2: MET (19.1% difference) and GLN (18.3%). We already saw that the complexity (first) part of these models are orders of magnitude different (in bits), again in favour of the MML mixture model. This in itself demonstrates the power of inference made under the MML framework, and the natural trade-off between complexity and fit the framework permits. It is also a demonstration of the effectiveness of the EM method employed to infer these mixtures.

Finally, to give an overall view of the qualitative differences across all amino acids, we plot the probability distribution for each sidechain angle for which the MML mixture model can project onto the respective dihedral angle dimension, and compare it against the empirical (observed) distribution of that angle. For each amino acid, we randomly sample data points (vector of dihedral angles) from mixture models and plot against the corresponding empirical distribution. Figure 3 shows these plots across all amino acids, with the mixture model shown as a red curve, and the empirical distribution shown in yellow. For comparison, we include the distribution of sidechain dihedral angles by randomly sampling from the Dunbrack library across amino acids, shown in the same figure (in blue). The plots show that our mixture models fit better the empirical distribution than the Dunbrack models. (The visualization for PDB50HighRes is provided in Supplementary Section S7, and follows the same conclusions as above.)

4 Conclusion

We have successfully modelled the joint distribution of main-chain and sidechain dihedral angles of amino acids using mixture models. By measuring the Shannon information content, we showed that our mixture models outperform the models implied by the Dunbrack rotamer libraries (across levels of smoothing), both in terms of its model complexity (by three orders of magnitude) and its fidelity (yielding on average 20% more lossless compression) when explaining the observed dihedral angle datasets with varying resolution and filtering thresholds. We also demonstrated the robustness of the MML method of estimation, and show that the inferred mixture models are not prone to the pitfalls of under/over-fitting and other inconsistencies common to many statistical model selection exercises. The brevity of our mixture models also provide computationally cheap and reliable way to sample jointly $\langle \phi, \psi, \chi_1, \chi_2, \ldots \rangle$ dihedral angles (and also conditionally given $\langle \phi, \psi \rangle$) and are ready for use in downstream studies: experimental structure refinement, *de novo* protein design,

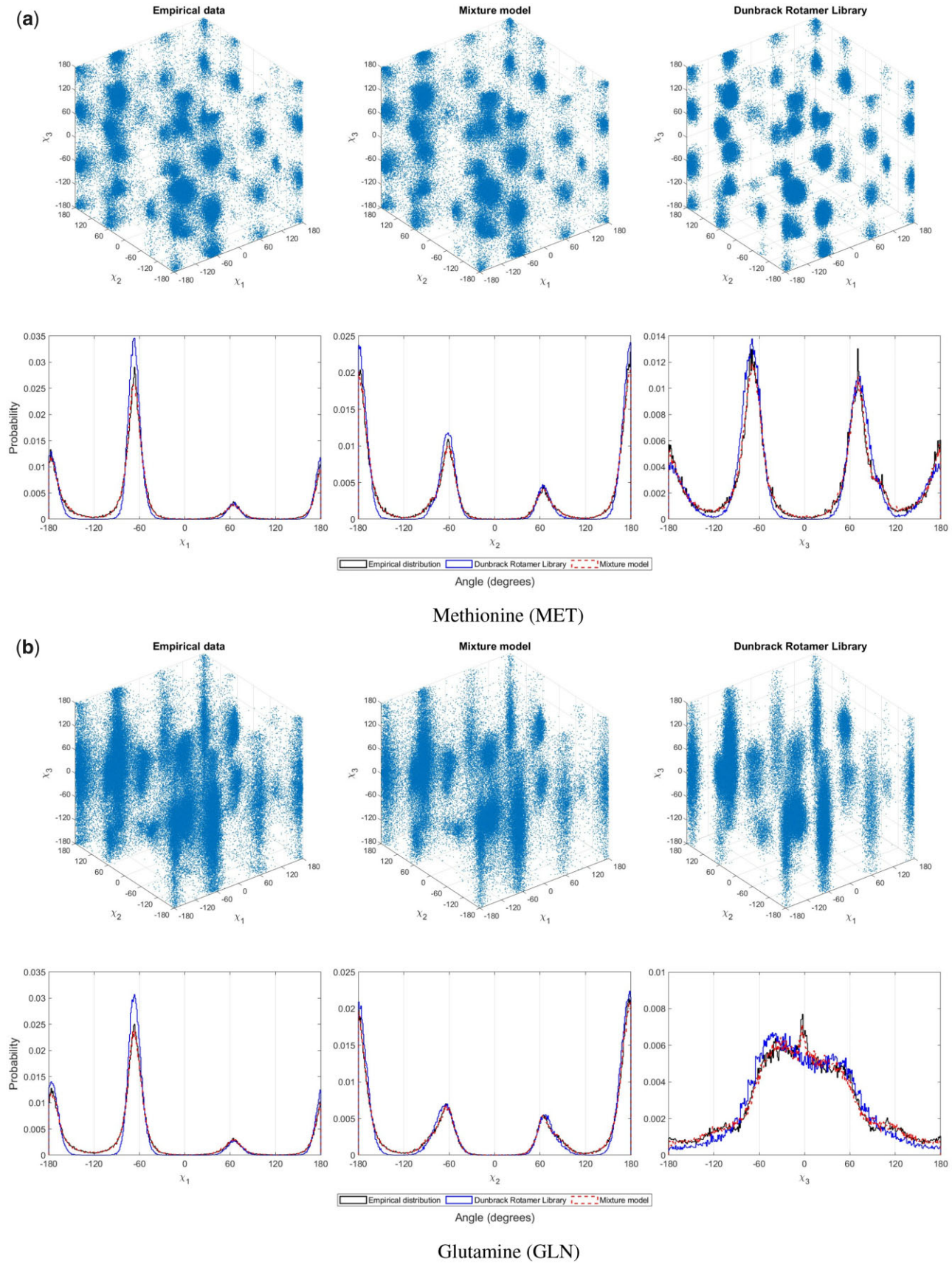


Figure 2. (a) The projection, into the sidechain (χ_1, χ_2, χ_3) space (unwrapped), of 100,000 randomly sampled points (vector of dihedral angles) for the amino acid methionine (MET) from MML mixture model (first row, center), of the same number of points from the Dunbrack model (first row, right), and of the observed (empirical) distribution of the same angles (first row, left). In the plots of the second row, the same data are visualized differently over three separate plots, with each of the three sidechain dihedral angles as *x*-axis (unwrapped), with *y*-axis showing the corresponding relative probabilities (in a 1° intervals). (b) The third and fourth rows plots are similar to first and second, respectively, but for the 'non-rotameric' amino acid, glutamine (GLN).

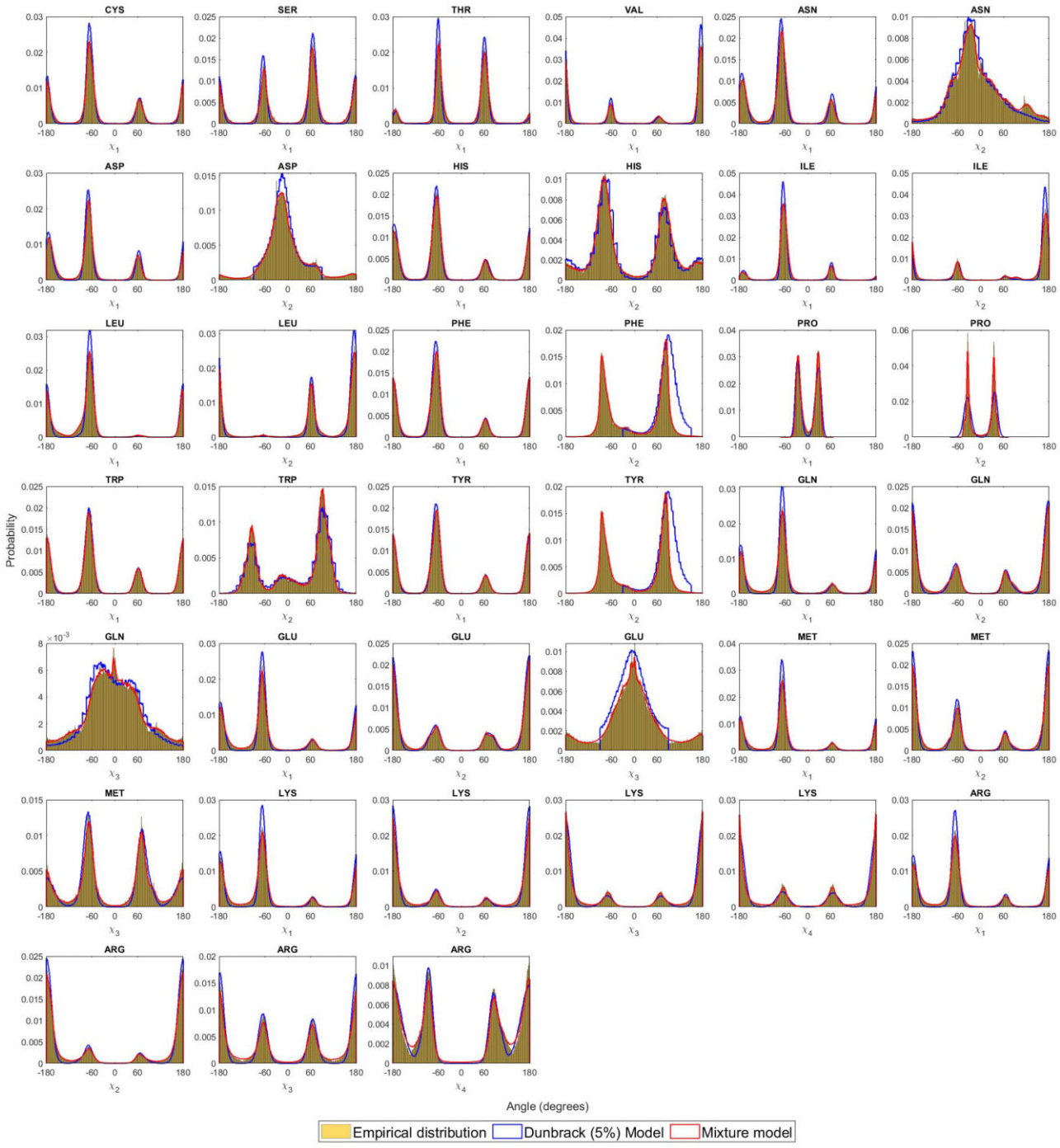


Figure 3. Fidelity of the inferred MML mixture models: the projected distribution of individual sidechain dihedral angles across all amino acids derived by randomly sampling $N^{(aa)}$ datapoints (see Table 1) from MML derived mixture models and Dunbrack (5% smoothed) library, and compared with the empirical distribution.

protein structure prediction, among others. Our mixture models, `PhiSiCal` ($\phi\psi\chi$al), are available for download from http://lcb.infotech.monash.edu.au/phisical. Also available from this link are programs to sample from the mixture models and report descriptive statistics (probability, log-odds ratios between pairs of models, null probability to estimate statistical significance, etc.) for use in modelling and simulation exercises.

We foresee several applications of candidate samples of amino acid conformations generated from `PhiSical` models. These include computational support to model amino acid 3D coordinates into electron density maps, predicting sidechain conformations given backbone states of amino acids, assessing protein structures to detect conformation-outliers, driving perturbations in molecular dynamic simulations, among others. We aim to address these as future work.

Acknowledgements

The authors thank Monash eResearch Centre and eServices for special job allocations on Monash HPC clusters that facilitated this work.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

PJS and MG's research is partially supported by OPTIMA ARC Industrial Transformation Training Centre [Project ID IC200100009].

References

Allison L. *Coding Ockham's Razor*. Cham, Switzerland: Springer, 2018.

Allison L *et al.* On universal codes for integers: Wallace tree, elias omega and variations. *arXiv preprint, arXiv:1906.05004*, 2019.

Berman HM, Westbrook J, Feng Z *et al.* The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.

Conway JH, Sloane NJ. On the Voronoi regions of certain lattices. *SIAM J Algebraic Discrete Methods* 1984;**5**:294–305.

Dempster AP, Laird NM, Rubin DB *et al.* Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol* 1977;**39**: 1–22.

Desmet J, De Maeyer M, Hazes B *et al.* The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 1992;**356**:539–42.

Dunbrack RL Jr. Rotamer libraries in the 21st century. *Curr Opin Struct Biol* 2002;**12**:431–40.

Dunbrack RL Jr, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* 1997;**6**:1661–81.

Dunbrack RL Jr, Karplus M. Backbone-dependent rotamer library for proteins application to side-chain prediction. *J Mol Biol* 1993;**230**:543–74.

Figueiredo MAT, Jain AK. Unsupervised learning of finite mixture models. *IEEE Trans Pattern Anal Mach Intell* 2002;**24**:381–96.

Grigoryan G, Ochoa A, Keating AE *et al.* Computing van der Waals energies in the context of the rotamer approximation. *Proteins Struct Funct Bioinformatics* 2007;**68**:863–78.

Harder T, Boomsma W, Paluszewski M *et al.* Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics* 2010;**11**:1–13.

Heringa J, Argos P. Strain in protein structures as viewed through non-rotameric side chains: I. their position and interaction. *Proteins* 1999;**37**:30–43.

IUPAC-IUB Commission On Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Journal of Biological Chemistry 1970;**245**:6489–97. https://doi.org/10.1016/S0021-9258(18)62561-X.

Janin J, Wodak S. Conformation of amino acid side-chains in proteins. *J Mol Biol* 1978;**125**:357–86.

Kasarapu P, Allison L. Minimum message length estimation of mixtures of multivariate Gaussian and von Mises-Fisher distributions. *Mach Learn* 2015;**100**:333–78.

Konagurthu AS, Allison L, Abramson D *et al.* How precise are reported protein coordinate data? *Acta Crystallogr D Biol Crystallogr* 2014; **70**:904–6.

Lassila JK. Conformational diversity and computational enzyme design. *Curr Opin Chem Biol* 2010;**14**:676–82.

Lovell SC, Word JM, Richardson JS *et al.* Asparagine and glutamine rotamers: B-factor cutoff and correction of amide flips yield distinct clustering. *Proc Natl Acad Sci USA* 1999;**96**: 400–5.

Lovell SC, Word JM, Richardson JS *et al.* The penultimate rotamer library. *Proteins Struct Funct Bioinformatics* 2000;**40**: 389–408.

Mardia KV *et al.* 2000. *Directional Statistics*, vol. **2**. West Sussex England: Wiley Online Library.

McGregor MJ, Islam SA, Sternberg MJ *et al.* Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J Mol Biol* 1987;**198**:295–310.

McLachlan GJ, Basford KE. *Mixture Models: Inference and Applications to Clustering*, vol. **38**. New York: Marcel Dekker, 1988.

McLachlan GJ, Lee SX, Rathnayake SI *et al.* Finite mixture models. *Annu Rev Stat Appl* 2019;**6**:355–78.

Ponder JW, Richards FM. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* 1987;**193**:775–91.

Ramachandran GN, Ramakrishnan C, Sasisekharan V *et al.* Stereochemistry of polypeptide chain configurations. *J Mol Biol* 1963;**7**:95–9.

Schrauber H, Eisenhaber F, Argos P *et al.* Rotamers: to be or not to be?: an analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol* 1993;**230**:592–612.

Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* 1948;**27**:379–423.

Shapovalov MV, Dunbrack RL Jr. Statistical and conformational analysis of the electron density of protein side chains. *Proteins Struct Funct Bioinformatics* 2007;**66**:279–303.

Shapovalov MV, Dunbrack RL Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 2011;**19**:844–58.

Wallace CS. 2005. *Statistical and Inductive Inference by Minimum Message Length*. New York, USA: Springer.

Wallace CS, Boulton DM. An information measure for classification. *Comput J* 1968;**11**:185–94.

Wallace CS, Freeman PR. Estimation and inference by compact coding. *J R Stat Soc Ser B Methodol* 1987;**49**:240–52.

Wallace CS, Patrick JD. Coding decision trees. *Mach Learn* 1993;**11**: 7–22.

Wang C, Schueler-Furman O, Baker D *et al.* Improved side-chain modeling for protein–protein docking. *Protein Sci* 2005;**14**: 1328–39.

Supplementary Material for

Getting '$\phi\psi\chi$al' with proteins: minimum message length inference of joint distributions of backbone and sidechain dihedral angles

Piyumi R. Amarasinghe[1], Lloyd Allison[1], Peter J. Stuckey[1,2], Maria Garcia de la Banda[1,2], Arthur M. Lesk[3], and Arun S. Konagurthu[1,*]

[1]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia
[2]OPTIMA ARC Industrial Training and Transformation Centre, Carlton, VIC 3053, Australia
[3]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA

Supplementary data is available to download from:
`PhiSiCal` ($\phi\psi\chi$al), `http://lcb.infotech.monash.edu.au/phisical`

S1 MML estimation for concentration parameter

We have seen in the main text (section 2.3.2, equation 9) how the mean (μ) parameter is estimated for each von Mises component in the mixture. Here, we will provide details of the numerical estimation of the concentration parameter (κ). By using the Wallace-Freeman method (Wallace and Freeman, 1987), the total message length for stating N observations of circular random variables x, where $x \in (-\pi, \pi]$ using a von Mises distribution can be derived as follows. Let $x \in X$ and $X = \{x_1, x_2, ..., x_N\}$ are the N observations of x. Further consider the von Mises distribution $f(x; \langle \mu, \kappa \rangle)$ with mean $\mu \in (-\pi, \pi)$ and concentration $\kappa > 0$. The associated net message $I(\langle \mu, \kappa \rangle, X)$ can be approximated as,

$$I(\langle \mu, \kappa \rangle, X) \approx \log(q_2) + \log\left(\frac{\sqrt{\det(\mathcal{F}(\langle \mu, \kappa \rangle))}}{h(\langle \mu, \kappa \rangle)}\right) + \mathcal{L}(\langle \mu, \kappa \rangle) - 2N\log(\epsilon) + 1 \tag{1}$$

Therefore, optimal concentration parameter κ_{MML} that minimizes Equation 1, can be derived by

$$\kappa_{MML} = \underset{\kappa}{\operatorname{argmin}}\, I(\langle \mu, \kappa \rangle, X) \tag{2}$$

As $\frac{\partial I(\langle \mu, \kappa \rangle, X)}{\partial \kappa} = 0$ results in a non-linear equation without a closed-form solution, we use the Newton–Raphson method as a reasonable approximation for finding roots of Equation 2 (Kasarapu and Allison, 2015).

Let $G(\kappa) = \frac{\partial I(\langle \mu, \kappa \rangle, X)}{\partial \kappa}$, then κ_{MML} can be approximated by twice iteration of Newton-Rapson method and using initial guess of roots $\kappa_0 = \kappa_B$. Here κ_B is the *Banerjee's approximation* (Banerjee *et al.*, 2005) for the concentration parameter which can be used as a feasible starting point for approximating roots of $G(\kappa) = 0$. Let κ_1, κ_{MML} correspond to the roots approximated at the first two iterations of the Newton-Rapson method,

$$\kappa_1 = \kappa_B - \frac{G(\kappa_B)}{G'(\kappa_B)} \quad \text{and} \quad \kappa_{MML} = \kappa_1 - \frac{G(\kappa)}{G'(\kappa)} \tag{3}$$

Here κ_B is evaluated by,

$$\kappa_B = \frac{\bar{R}(2 - \bar{R}^2)}{(1 - \bar{R}^2)} \text{ where } \bar{R} = \frac{||R||}{N} \tag{4}$$

R is the vector sum of each x circular variable and $||R||$ is the vector norm of the resultant vector R. We use κ_{MML} as the approximation for concentration parameter minimizing $I(\langle \mu, \kappa \rangle, X)$.

S2 Searching for optimal mixture

This algorithm employs the MML paradigm to quantify the fitness of competing statistical models (see main text for a detailed explanation of the MML model selection paradigm). In a nutshell, under MML, an optimal model is the one that yields the minimum two-part message length over all possible competing models. This remains a hard optimization problem. We employ a Expectation-Maximization (EM) based approach, which is commonly used for unsupervised statistical parameter estimation problems.

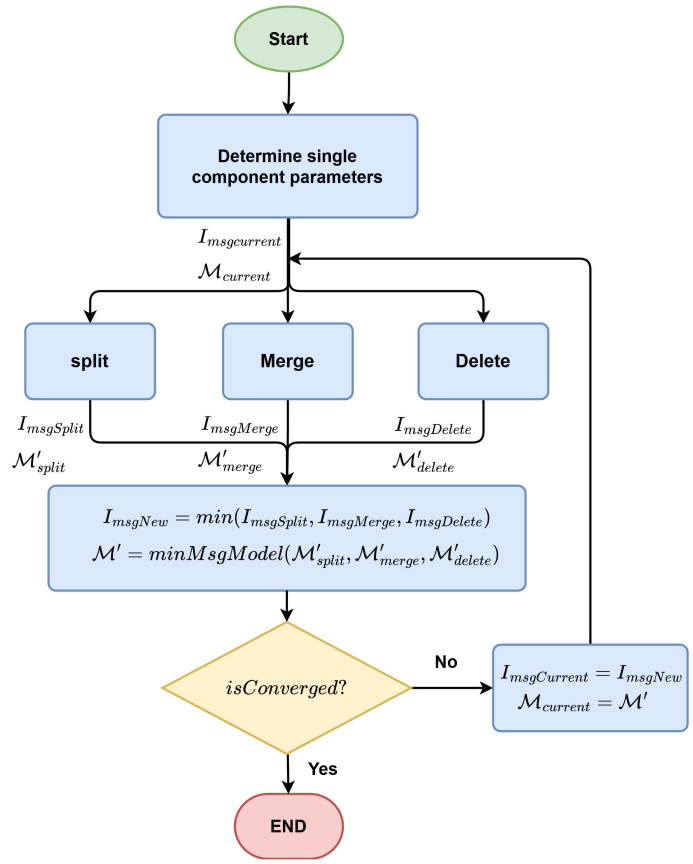


Figure SF 1. Flow chart representing the functional flow of the search algorithm

The conceptual flow of the EM search is shown in Fig SF 1. The EM starts with a single component mixture model ($|\mathcal{M}| = 1$) at iteration $t = 0$, whose parameters are estimated as described in section S1. Starting from this single component mixture, the model undergoes a series of *Split*, *Merge* and *Delete* perturbations, chosen deterministically and greedily to improve the total message length objective. Each operation and their rationale is described below. For this, assume that at an arbitrary iteration t, we have a K component mixture model ($|\mathcal{M}| = K$).

S2.1 Split operation:

Given a K component mixture, the primary goal of this operation is to find two distinct sub-populations in the current mixture such that the new mixture with $K + 1$ components is capable of better explaining the data. Consider an arbitrary $\mathcal{M}_j$ component ready for split operation. This $\mathcal{M}_j$ component (*parent*) is split into two new components (*children*). The resulting sub-mixture is then optimized using the EM algorithm while the other $(K - 1)$ components are untouched. Let $\mathcal{M}_j^a, \mathcal{M}_j^b$ be child components of $\mathcal{M}_j$.

S2.1.1 Initialize child components.

EM algorithm is sensitive to initial parameter values. As the message length (or likelihood) function of a mixture model is not unimodal, depending on the starting parameters of the mixture, it could get trapped in a sub-optimal solution. One could execute the EM algorithm starting from a variety of initial parameter values chosen randomly to avoid a poor approximation to the true minimum (for message length as the objective function). However, this does not ensure the accuracy of the resulting estimates. In order to select the initial parameters of the sub-mixture, we consider the distribution of membership among the probable child components.

Given $\Theta_j = \{\langle \mu_{j_p}, \kappa_{j_p} \rangle\}_{\forall 1 \leq p \leq d}$ of the parent component, we can locate two starting mean values for each dihedral-angle p by selecting two points (in $(-\pi, \pi]$ space) which are one standard deviation away on either side of Θ_j. These new mean values in the vicinity of the parent's mean can serve as starting parameters of the child components resulting from *splitting* the parent component. For a d-dimensional datum, there are (2^d) potential split combinations. But the split could be required in only a few directions while other mean directions need unchanged. Hence altogether there are $(3^d - 1)$ split combinations that can be assigned as initial mean values. Executing the EM algorithm on all the combinations to select the optimal sub-mixture is computationally expensive. Due to that, we calculate the membership (refer to Equation (7) in the main text) under each plausible split combination and select the two combinations producing the highest membership. The corresponding mean values serve as good starting values of the mean direction of the two-component sub-mixture. Furthermore, the concentration parameters of the parent component are used as starting concentration values and weights are equally distributed among children.

Once child components are initialized, the EM algorithm is executed on the two-component sub-mixture.

S2.1.2 E-step

Let r_{ij}^a, r_{ij}^b be responsibilities of each child on datum $x_i \in X$ and n_j^a, n_j^b be memberships of $\mathcal{M}_j^a, \mathcal{M}_j^b$ child components. These values are calculated from Equation (7) in the main text.

S2.1.3 M-step

Given the responsibilities and memberships calculated at E-step, new parameters of the child components are calculated as follows. Let r_{ij} be the responsibility of parent component $\mathcal{M}_j$ on x_i.

$$w_j^a(t+1) = \frac{n_j^a + \frac{1}{2}}{N + \frac{1}{2}} \tag{5}$$

$$\mu_{j_p}^a = \frac{R_{j_p}^a}{||R_{j_p}^a||} \tag{6}$$

where $R_{j_p}^a$ is the vector sum of each $x_{i_p} \in X$ weighted by corresponding responsibilities r_{ij} and r_{ij}^a, similarly $||R_{j_p}^a||$ is the vector norm of the resultant vector $R_{j_p}^a$. Finally, concentration parameter $\kappa_{j_p}^a$,$(\forall 1 \leq p \leq d)$ is calculated by,

$$\bar{R}_{j_p}^a = \frac{||R_{j_p}^a||}{\sum_{i=1}^{N} r_{ij} r_{ij}^a} \tag{7}$$

The E-step and M-step are executed repeatedly until there is no *sufficient* gain in the message lengths between consecutive iterations. Once the sub-mixture is optimized, it is integrated into the $(K - 1)$ mixture such that the $\mathcal{M}_j$ component is replaced by its successors. The resulting $(K + 1)$ mixture is then optimized again by the EM algorithm to tune the parameters of the new mixture. The split operation is performed on every K element in the mixture, and only the $(K + 1)$ component combination having the minimum two-part message length is selected to proceed (see Figure SF 1). Let $\mathcal{M}'_{split}$ be the selected new mixture at split operation.

S2.2 Merge operation:

Primary goal of Merge perturbation is to join components in a $K(K > 1)$ component mixture and explore how the resulting $(K - 1)$ mixture performs. We consider Kullback-Leibler Distance (Kullback and Leibler, 1951) as a potential heuristic to identify components that are plausible to join. A component M_j is merged with another component with the minimum KL distance among the remaining $(K - 1)$ components. During the merging, the responsibilities of one component are handed over to the other component resulting in a $(K-1)$ component mixture. The resulting $(K-1)$ component mixture is then tuned from the EM algorithm to readjust the component parameters. The merge operation is performed on every K element in the mixture exhaustively such that, only the $(K - 1)$ component combination having the minimum two-part message length is selected to proceed. Let $\mathcal{M}'_{merge}$ be the selected new mixture at merge operation.

S2.2.1 KL Distance of a von Mises Distribution

Let $Q(x; \langle \mu_q, \kappa_q \rangle), R(x; \langle \mu_r, \kappa_r \rangle)$ be two von Mises distributions then,

$$\begin{aligned} D_{KL}(Q \parallel R) &= \int_x Q(x) \log\left(\frac{Q(x)}{R(x)}\right) dx \\ &= \log\left(\frac{B_0(\kappa_r)}{B_0(\kappa_q)}\right) + A(\kappa_q)(\kappa_q - \kappa_r \cos(\mu_q - \mu_r)) \end{aligned} \tag{8}$$

Similar to usage in the main text, the modified Bessel function of order 0, I_0 is stated as B_0 in Equation 8.

By using Equation 8, the KL distance between two joint von Mises distributions $U(x; \Theta_u), V(x; \Theta_v)$ where $\Theta_u = \{\langle \mu_{u_p}, \kappa_{u_p} \rangle\}$ and $\Theta_v = \{\langle \mu_{v_p}, \kappa_{v_p} \rangle\}, (\forall 1 \leq p \leq d)$ is

$$D_{KL}(U \parallel V) = \sum_{p=1}^{d} \log\left(\frac{B_0(\kappa_{v_p})}{B_0(\kappa_{u_p})}\right) + A(\kappa_{u_p})(\kappa_{u_p} - \kappa_{v_p} \cos(\mu_{u_p} - \mu_{v_p})) \tag{9}$$

S2.3 Delete operation:

Given the greedy nature of the merge (which, for every component, always explores merging with the closest-neighbouring component in the current mixture), it becomes necessary to include a delete operation to remove any component that although being redundant escapes the greedy merge. Thus the delete operation aims to remove components from a $K(K > 1)$ mixture, one at a time and redistribute the responsibilities of the deleted component to the remaining components. The resulting $(K - 1)$ mixture is then optimized using the EM algorithm to check whether an improved model could be achieved. The delete operation is performed on every K element in the mixture exhaustively such that, only the $(K - 1)$ component combination having the minimum two-part message length is selected to proceed. Let M'_{delete} be the selected new mixture at the delete operation.

On completion of all perturbations, now we have $\mathcal{M}'_{split}, \mathcal{M}'_{merge}, \mathcal{M}'_{delete}$ mixtures which are the best choices from each perturbation. Finally, the model with the minimum two-part message length is selected as the starting mixture ($\mathcal{M}'$) of the next iteration. This process is continued repeatedly until the gains are minimal.

Refer (Kasarapu and Allison, 2015) for a detailed explanation of the perturbations.

S3 Quantitative comparison of message lengths for stating amino acid sidechain dihedral angles from PDB50 dataset

Table ST 1. This table provides a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) when explaining the PDB50 dataset. We emphasize that the reported message length terms are for losslessly stating only the sidechain dihedral angles of individual amino acids (aa) and do not consider the backbone dihedral angles $\langle \phi, \psi \rangle$. The 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles.

		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	Null($X^{(\text{aa})}$)	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	2,171,630	(165;1,484)	4,095	17,578,246	17,582,342	8.1	(11,664;57,024)	1,079,717	41,766,148	42,845,865	19.7	26,797,588	12.3
ALA	1,861,359	(25;124)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
VAL	1,601,058	(96;671)	1,625	6,607,950	6,609,575	4.1	(3,888;10,368)	217,204	23,548,535	23,765,739	14.8	9,878,408	6.2
GLY	1,588,115	(30;149)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
GLU	1,446,860	(262;2,881)	7,754	21,695,912	21,703,666	15.0	(69,984;488,592)	9,696,028	37,039,858	46,735,886	32.3	26,781,053	18.5
SER	1,337,273	(114;797)	1,757	6,592,674	6,594,431	4.9	(3,888;10,368)	210,725	20,949,757	21,160,482	15.8	8,250,874	6.2
ILE	1,333,508	(172;1,547)	4,346	10,688,100	10,692,445	8.0	(11,664;57,024)	964,613	25,021,654	25,986,267	19.5	16,455,289	12.3
ASP	1,279,567	(170;1,529)	3,640	12,462,677	12,466,317	9.7	(23,328;115,344)	2,336,812	25,075,659	27,412,471	21.4	15,789,665	12.3
THR	1,221,604	(90;629)	1,445	5,531,460	5,532,905	4.5	(3,888;10,368)	211,682	18,290,358	18,502,040	15.1	7,537,205	6.2
LYS	1,176,395	(266;3,457)	9,833	22,078,096	22,087,929	18.8	(104,976;943,488)	14,337,381	35,465,455	49,802,836	42.3	29,033,076	24.7
ARG	1,130,448	(250;3,749)	12,164	23,504,690	23,516,854	20.8	(104,976;943,488)	15,442,697	34,991,767	50,434,464	44.6	34,873,897	30.8
PRO	1,004,859	(231;2,078)	8,956	5,257,024	5,265,980	5.2	(2,592;11,664)	254,490	16,308,550	16,563,040	16.5	12,399,809	12.3
ASN	948,274	(180;1,619)	3,703	9,582,207	9,585,910	10.1	(46,656;231,984)	4,586,845	19,335,593	23,922,438	25.2	11,701,559	12.3
PHE	927,298	(226;2,033)	5,401	8,460,779	8,466,181	9.1	(23,328;115,344)	2,216,332	17,235,221	19,451,553	21.0	11,442,718	12.3
GLN	820,871	(239;2,628)	6,804	12,270,999	12,277,803	15.0	(139,968;978,480)	18,417,678	21,525,549	39,943,227	48.7	15,194,138	18.5
TYR	788,176	(192;1,727)	4,480	7,193,098	7,197,578	9.1	(23,328;115,344)	2,248,946	14,607,857	16,856,803	21.4	9,725,974	12.3
HIS	515,611	(163;1,466)	3,443	5,175,762	5,179,204	10.0	(46,656;231,984)	4,373,646	10,388,460	14,762,106	28.6	6,362,562	12.3
MET	417,170	(270;2,969)	7,919	5,954,095	5,962,013	14.3	(34,992;243,648)	4,222,659	10,669,762	14,892,422	35.7	7,721,723	18.5
TRP	310,470	(212;1,907)	4,816	2,958,040	2,962,856	9.5	(46,656;231,984)	4,062,892	6,038,982	10,101,874	32.5	3,831,153	12.3
CYS	296,547	(96;671)	1,433	1,389,186	1,390,619	4.7	(3,888;10,368)	190,178	4,432,454	4,622,632	15.6	1,829,673	6.2

S4 Quantitative comparison of message lengths for stating amino acid (backbone + sidechain) dihedral angles from PDB50HighRes dataset using PDB50-inferred mixture models

Corrigendum: The table below is an updated version to the one that appears in the SM linked to the published version, updated due to a clerical error in the table production (where a few cells containing message length terms on the MML mixture models side were accidentally permuted during production).

Table ST 2. This table provides a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) for stating dihedral angles (backbone + sidechain) of each of the twenty naturally occurring amino acids (aa). The the 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles. While we model the joint distributions of dihedral including the backbone, Dunbrack on the other hand only provides sidechain distributions conditional on the backbone. Hence ALA and GLY Dunbrack libraries are necessarily empty. Tables corresponding to PDB50HighRes-inferred mixture models (instead of PDB50 models used below) can be found at https://lcb.infotech.monash.edu.au/phisical/.

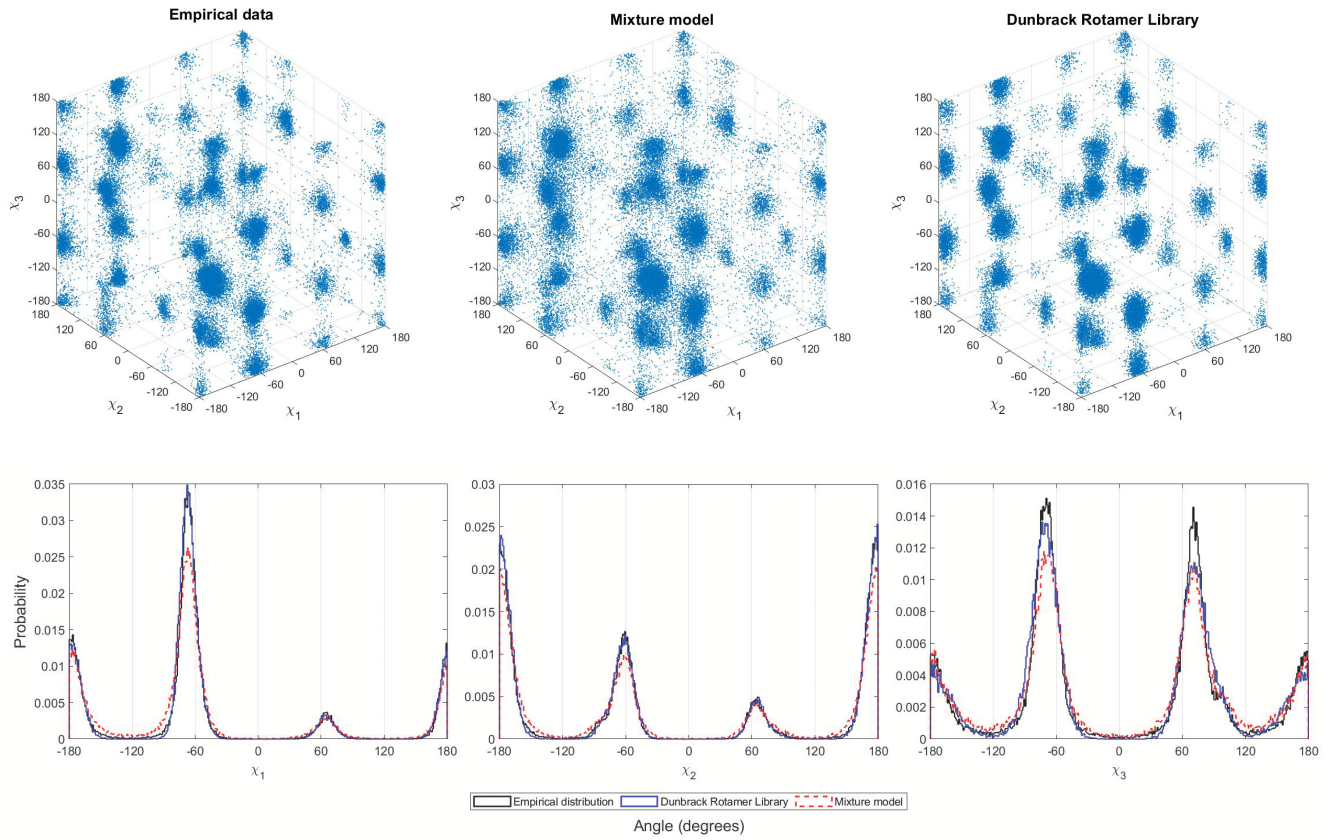
		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	Null($X^{(\text{aa})}$)	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	343,752	(165;1,484)	6,788	5,012,392	5,019,181	14.6	(11,664;57,024)	950,779	6,587,654	7,538,433	21.9	8,483,696	24.7
ALA	334,111	(25;124)	666	2,533,968	2,534,634	7.6	(N/A;N/A)	N/A	N/A	N/A	N/A	4,122,880	12.3
GLY	294,278	(30;149)	706	2,851,845	2,852,551	9.7	(N/A;N/A)	N/A	N/A	N/A	N/A	3,631,346	12.3
VAL	274,596	(96;671)	3,261	2,971,624	2,974,885	10.8	(3,888;10,368)	192,834	4,163,843	4,356,677	15.9	5,082,710	18.5
GLU	238,682	(262;2,881)	11,853	5,161,336	5,173,189	21.7	(69,984;488,592)	8,509,192	6,189,754	14,698,946	61.6	7,363,250	30.8
ASP	227,558	(170;1,529)	6,304	3,867,073	3,873,377	17.0	(23,328;115,344)	2,062,259	4,632,995	6,695,254	29.4	5,616,063	24.7
SER	222,721	(114;797)	3,671	2,862,306	2,865,978	12.9	(3,888;10,368)	185,948	3,657,879	3,843,828	17.3	4,122,516	18.5
ILE	215,684	(172;1,547)	7,121	3,037,346	3,044,467	14.1	(11,664;57,024)	848,230	4,018,755	4,866,985	22.6	5,323,016	24.7
THR	212,562	(90;629)	2,937	2,534,288	2,537,226	11.9	(3,888;10,368)	187,511	3,295,033	3,482,543	16.4	3,934,475	18.5
LYS	195,868	(266;3,457)	13,332	5,049,521	5,062,854	25.8	(104,976;943,488)	12,526,749	5,878,978	18,405,727	94.0	7,250,945	37.0
ARG	188,400	(250;3,749)	15,558	5,307,530	5,323,088	28.3	(104,976;943,488)	13,518,806	5,758,992	19,277,798	102.3	8,136,897	43.2
PRO	177,534	(231;2,078)	13,482	2,014,498	2,027,980	11.4	(2,592;11,664)	225,394	3,195,740	3,421,135	19.3	4,381,486	24.7
ASN	162,196	(180;1,619)	6,554	2,908,746	2,915,300	18.0	(46,656;231,984)	4,025,549	3,472,766	7,498,315	46.2	4,002,949	24.7
PHE	153,192	(226;2,033)	9,063	2,540,627	2,549,690	16.6	(23,328;115,344)	1,940,884	3,077,402	5,018,286	32.8	3,780,733	24.7
GLN	136,703	(239;2,628)	10,547	2,986,890	2,997,437	21.9	(139,968;978,480)	16,100,149	3,615,527	19,715,676	144.2	4,217,236	30.8
TYR	134,950	(192;1,727)	7,576	2,254,154	2,261,731	16.8	(23,328;115,344)	1,970,652	2,718,877	4,689,528	34.8	3,330,526	24.7
HIS	89,382	(163;1,466)	6,013	1,609,829	1,615,841	18.1	(46,656;231,984)	3,818,112	1,928,561	5,746,674	64.3	2,205,921	24.7
MET	68,907	(270;2,969)	12,078	1,445,826	1,457,903	21.2	(34,992;243,648)	3,657,582	1,752,132	5,409,714	78.5	2,125,755	30.8
TRP	56,696	(212;1,907)	8,322	961,834	970,157	17.1	(46,656;231,984)	3,547,218	1,197,638	4,744,855	83.7	1,399,240	24.7
CYS	46,435	(96;671)	3,013	584,959	587,972	12.7	(3,888;10,368)	164,549	747,043	911,592	19.6	859,501	18.5

S5 Quantitative comparison of message lengths for stating amino acid sidechain dihedral angles from PDB50HighRes dataset using PDB50-inferred mixture models

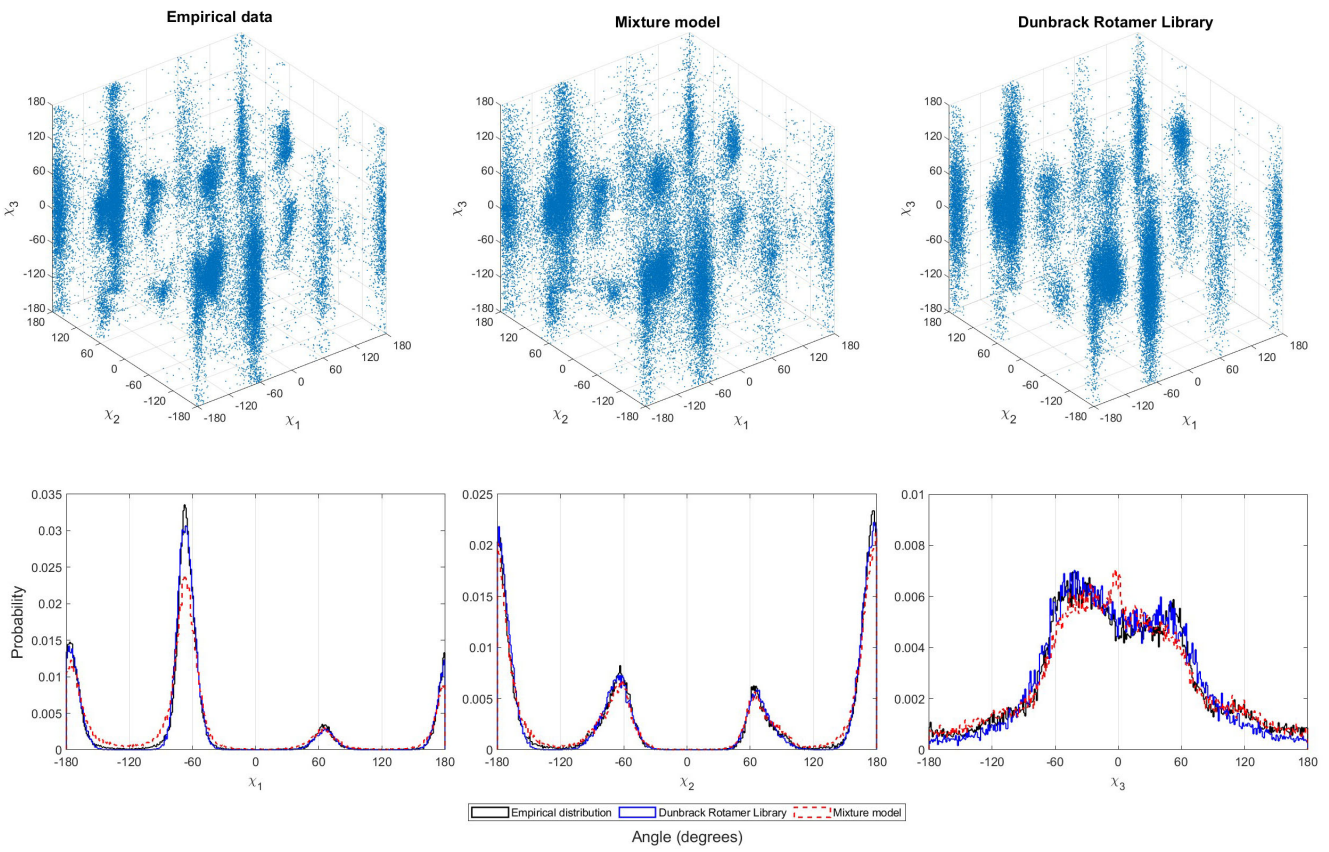
Table ST 3. This table illustrates a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) to state only sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa). Here we have only considered the cost of stating the sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa) and omitted the backbone $\langle \phi, \psi \rangle$. The 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles. Tables corresponding to PDB50HighRes-inferred mixture models (instead of PDB50 models used below) can be found at https://lcb.infotech.monash.edu.au/phisical/.

		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	Null($X^{(\text{aa})}$)	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	343,752	(165;1,484)	3,872	2,460,581	2,464,453	7.2	(11,664;57,024)	950,774	5,900,150	6,850,924	19.9	4,241,848	12.3
ALA	334,111	(25;124)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
GLY	294,278	(30;149)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
VAL	274,596	(96;671)	1,502	1,001,833	1,003,334	3.7	(3,888;10,368)	192,829	3,614,651	3,807,480	13.9	1,694,237	6.2
GLU	238,682	(262;2,881)	7,407	3,387,681	3,395,088	14.2	(69,984;488,592)	8,509,187	5,712,390	14,221,577	59.6	4,417,950	18.5
ASP	227,558	(170;1,529)	3,424	2,088,202	2,091,626	9.2	(23,328;115,344)	2,062,254	4,177,879	6,240,133	27.4	2,808,032	12.3
SER	222,721	(114;797)	1,608	1,012,207	1,013,816	4.6	(3,888;10,368)	185,943	3,212,437	3,398,380	15.3	1,374,172	6.2
ILE	215,684	(172;1,547)	4,116	1,556,994	1,561,110	7.2	(11,664;57,024)	848,225	3,587,387	4,435,612	20.6	2,661,508	12.3
THR	212,562	(90;629)	1,331	861,198	862,528	4.1	(3,888;10,368)	187,506	2,869,909	3,057,414	14.4	1,311,492	6.2
LYS	195,868	(266;3,457)	9,480	3,482,352	3,491,831	17.8	(104,976;943,488)	12,526,744	5,487,242	18,013,986	92.0	4,833,963	24.7
ARG	188,400	(250;3,749)	11,829	3,806,227	3,818,056	20.3	(104,976;943,488)	13,518,801	5,382,192	18,900,993	100.3	5,812,069	30.8
PRO	177,534	(231;2,078)	8,664	932,851	941,515	5.3	(2,592;11,664)	225,389	2,840,672	3,066,061	17.3	2,190,743	12.3
ASN	162,196	(180;1,619)	3,470	1,575,926	1,579,395	9.7	(46,656;231,984)	4,025,544	3,148,374	7,173,918	44.2	2,001,474	12.3
PHE	153,192	(226;2,033)	5,104	1,356,899	1,362,003	8.9	(23,328;115,344)	1,940,879	2,771,018	4,711,896	30.8	1,890,366	12.3
GLN	136,703	(239;2,628)	6,489	1,944,633	1,951,122	14.3	(139,968;978,480)	16,100,144	3,342,121	19,442,265	142.2	2,530,342	18.5
TYR	134,950	(192;1,727)	4,232	1,201,580	1,205,812	8.9	(23,328;115,344)	1,970,647	2,448,977	4,419,623	32.8	1,665,263	12.3
HIS	89,382	(163;1,466)	3,233	871,873	875,106	9.8	(46,656;231,984)	3,818,107	1,749,797	5,567,904	62.3	1,102,960	12.3
MET	68,907	(270;2,969)	7,561	916,757	924,319	13.4	(34,992;243,648)	3,657,577	1,614,318	5,271,895	76.5	1,275,453	18.5
TRP	56,696	(212;1,907)	4,552	531,702	536,255	9.5	(46,656;231,984)	3,547,212	1,084,246	4,631,458	81.7	699,620	12.3
CYS	46,435	(96;671)	1,303	203,080	204,383	4.4	(3,888;10,368)	164,544	654,173	818,717	17.6	286,500	6.2

S6 Qualitative comparison of model fit for methionine(MET) and glutamine(GLN) sidechain dihedral angles from PDB50HighRes dataset using PDB50-inferred mixture models



(a) Methionine (MET)



(b) Glutamine (GLN)

Figure SF 2. (a) The projection, into the sidechain (χ_1, χ_2, χ_3) space (unwrapped), of 50,000 randomly sampled points (vector of dihedral angles) for the amino acid Methionine (MET) from MML mixture model (first row, center), of the same number of points from the Dunbrack model (first row, right), and of the observed (empirical) distribution of the same angles (first row, left) from PDBHighRes. In the plots of the second row, the same data is visualized differently over three separate plots, with each of the three sidechain dihedral angles as x-axis (unwrapped), with y-axis showing the corresponding relative probabilities (in a 1° intervals). (b) The third and fourth rows plots are similar to first and second, respectively, but for the non-rotameric amino acid, Glutamine (GLN). Plots corresponding to PDB50HighRes-inferred mixture models (instead of PDB50 models used above) can be found at https://lcb.infotech.monash.edu.au/phisical/.

S7 Qualitative comparison of model fit across all amino acid sidechain dihedral angles from PDB50HighRes dataset using PDB50-inferred mixture models

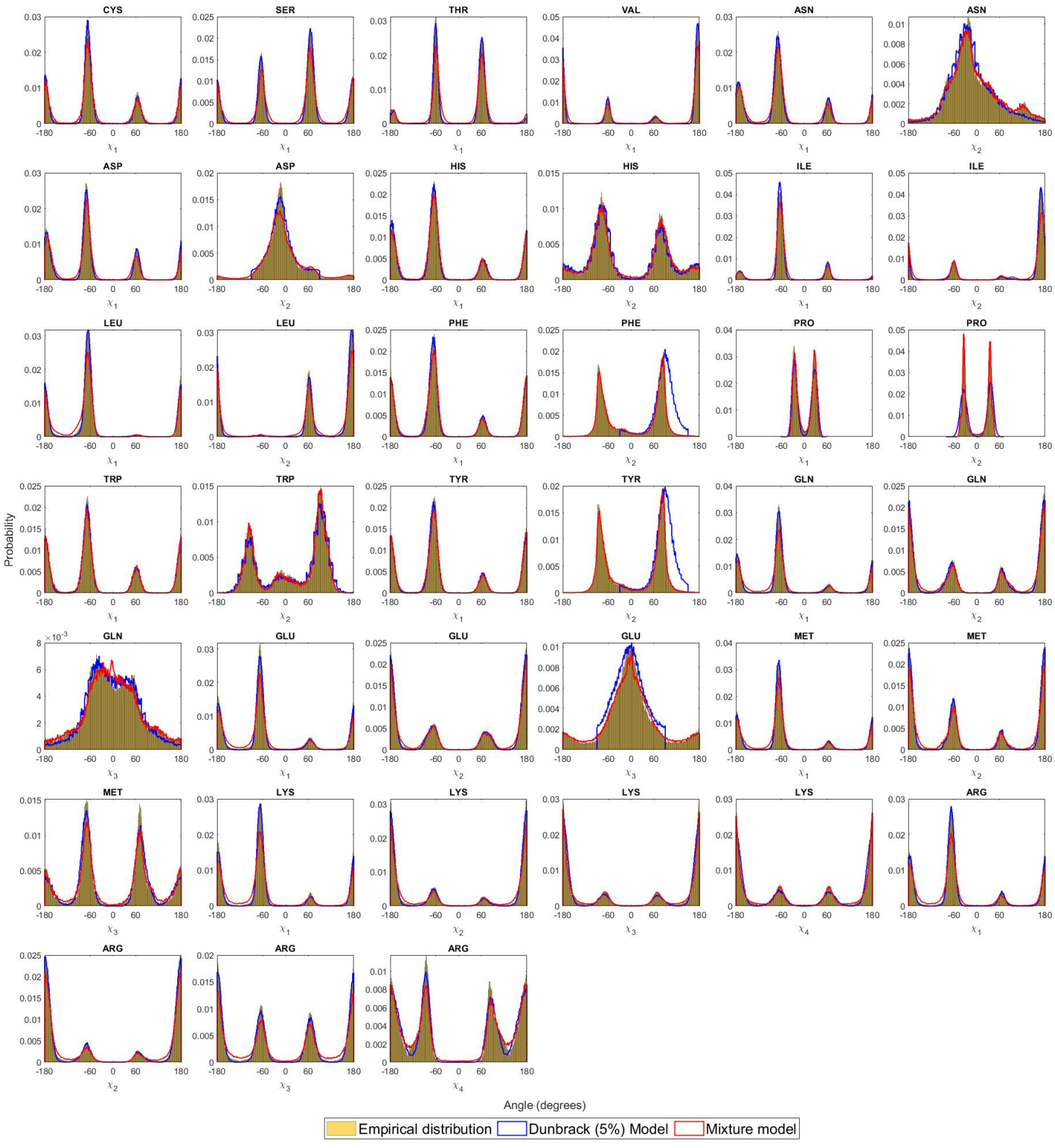


Figure SF 3. Fidelity of the inferred MML mixure models: the projected distribution of individual sidechain dihedral angles across all amino acids derived by randomly sampling $N^{(aa)}$ datapoints (see Table ST 2) from MML-derived mixture models and Dunbrack (5% smoothed) library, and compared to the empirical distribution. Plots corresponding to PDB50HighRes-inferred mixture models (instead of PDB50 models used above) can be found at https://lcb.infotech.monash.edu.au/phisical/.

S8 KL-divergence between mixture models inferred on PDB50 and PDB50HighRes dataset

Kullback-Leibler (KL) divergence informs a measure of relative entropy needed for encoding data over a probability model relative to another. We used this measure to evaluate mixture models derived for PDB50 and PDB50HighRes datasets under an empirical method to estimate KL divergence. Consider a set of data points $x_i; \forall 1 \leq i \leq N$ sampled from a true mixture distribution $\mathcal{M}^t$. Then the *empirical* KL-divergence $D_{KL}(\mathcal{M}^t \parallel \mathcal{M}))$ between $\mathcal{M}^t$ (true distribution) and another mixture model $\mathcal{M}$ can be approximated as (Kasarapu and Allison, 2015),

$$D_{KL}(\mathcal{M}^t \parallel \mathcal{M}) = E_{\mathcal{M}^t}\left[\log\left(\frac{\Pr(x_i;\mathcal{M}^t)}{\Pr(x_i;\mathcal{M})}\right)\right] \approx \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\Pr(x_i;\mathcal{M}^t)}{\Pr(x_i;\mathcal{M})}\right) \tag{10}$$

Table ST 4 shows KL-divergence between the mixture models inferred from the two datasets. The low KL-divergence values provide evidence that PDB50 and PDB50HighRes models are practically similar to sample representative amino acid conformations in proteins.

Table ST 4. This table illustrates the KL-divergence between mixture models inferred from PDB50 and PDB50HighRes datasets for each amino acid (a.a)

a.a	KL-divergence	a.a	KL-divergence
ALA	0.02675	MET	0.31734
CYS	0.05314	ASN	0.15406
ASP	0.26640	PRO	0.42636
GLU	0.20948	GLN	0.24193
PHE	0.16896	ARG	0.18094
GLY	0.03034	SER	0.06501
HIS	0.15723	THR	0.05209
ILE	0.62360	VAL	0.03940
LYS	0.20449	TRP	0.15949
LEU	0.22799	TYR	0.06479

S9 Message length of Dunbrack backbone-dependant rotamer library

For a fair and objective comparison in terms of Shannon information content, we need to translate any Dunbrack model ($\mathcal{D}^{aa}_{rotamer}$) to estimate their equivalent first part and second part message length terms. This is achieved as follows. Dunbrack report their latest backbone-dependent libraries in $10^\circ \times 10^\circ$ bins of $\langle\phi,\psi\rangle$ values. In each bin, they report a statistical model with a fixed number of components determined by the number of discrete rotamer states of the residue being considered. For example, consider the rotamer library for methionine amino acid. Methionine has 3 sidechain angles and each angle is considered to have 3 distinct rotameric states, yielding a total number of $3 \times 3 \times 3 = 27$ possibilities over all the three angles. Hence each bin of the backbone grid of methionine can be directly interpreted as a mixture model containing 27 components. For consistency, we followed the rotamer categorization used in their reported work describing latest libraries (Shapovalov and Dunbrack Jr, 2011). For example, Proline has only 2 rotameric states for χ_1.

Each component of this bin-wise mixture model is considered a product of von Mises distributions (since they independently model each dihedral angle by a von Mises circular distribution) where the weight parameter of each component is their defined conditional probability of each discrete rotameric state. Specifically, each component of a selected bin-wise mixture of methionine will be a product of 3 von Mises distributions where the parameters of each von Mises distribution are directly mapped from the parameters they report in their library.

Consider a d-dimensional dihedral angle datum (backbone + sidechain) $x_i \in X$ where X denotes the input set of N observations from a non-redundant protein dataset. Under this bin-wise mixture representation, we compared the complexity and fidelity of MML-derived mixture models and the Dunbrack rotamer library for stating N observations in 2 possible ways.

1. The backbone dihedral angles ϕ and ψ under the Dunbrack model are stated over a uniform distribution.
2. For each MML-inferred mixture model drop/ignore the von Mises terms corresponding to backbone dihedral angles when calculating the second part.

Approach 1

Calculating second part. We first select the bin in the Dunbrack model into which any specfic ϕ_i and ψ_i of x_i falls. Then the mixture model of sidechain dihedral angles specified by the $\langle\phi_i,\psi_i\rangle$-bin is used to encode the sidechain angles observed for x_i, identical to how we encode the same using the MML model. For the methionine example, the message length associated with stating sidechain dihedral angles of x_i is,

$$I(\langle x_{i_3}, x_{i_4}, x_{i_5}\rangle | \mathcal{D}^{\text{MET}}_{\text{rotamer}}) = -\log\left(\sum_{j=1}^{27}\left(w_j \prod_{p=3}^{5} f(x_{i_p}|\left\langle\mu_{j_p},\kappa_{j_p}\right\rangle)\right)\epsilon^3\right) \tag{11}$$

Here mixture component, $\prod_{p=3}^{5} f(x_{i_p}|\left\langle\mu_{j_p},\kappa_{j_p}\right\rangle)$ denote the j^{th} rotameric state of $x_{i_3}, x_{i_4}, x_{i_5}$ angles in ϕ_i, ψ_i bin. $f(x_{i_p}|\left\langle\mu_{j_p},\kappa_{j_p}\right\rangle$ represent von Mises distribution of sidechain dihedral angle x_{i_p} with parameters μ_{j_p}(mean) and κ_{j_p} (concentration). As with our models, the ϵ is set to 0.0873 radians (see main text for details).

In the first approach, we state the backbone dihedral angles of x_i under a uniform distribution. Hence the message length of nominating a $10^\circ \times 10^\circ$ $\langle\phi,\psi\rangle$-bin losslessly takes $\log(36^2) = 2 \times \log 36$ bits (assume logarithms are all base-2). Further any statement of the observed ϕ and ψ uniformly distributed within each 10° interval defines a probability of $\frac{\epsilon^\circ}{10^\circ}$, and negative logarithm of that probability yields $\log(10^\circ/\epsilon^\circ)$ bits. Thus, the amount of information to state the two backbone dihedrals under this approach losslessly takes $2 \times (\log 36 + \log(10^\circ/\epsilon^\circ))$ bits. Hence the second part of stating datum x_i of methionine under $\mathcal{D}^{\text{MET}}_{\text{rotamer}}$ can be quantified as

$$I\left(x_i|\mathcal{D}^{\text{MET}}_{\text{rotamer}}\right) = I(\langle x_{i_3}, x_{i_4}, x_{i_5}\rangle | \mathcal{D}^{\text{MET}}_{\text{rotamer}}) + 2(\log(36) + \log(10^\circ/\epsilon^\circ)). \tag{12}$$

Hence, the total message length of encoding N observations is the summation of individual message length terms $I\left(x_i|\mathcal{D}^{\text{MET}}_{\text{rotamer}}\right), \forall \leq i \leq N$

Calculating first part. The first part term of the $\mathcal{D}^{\text{aa}}_{\text{rotamer}}$ contains the message length of stating all the mixture parameters (across $36^2 = 1,296$) bins and the message length of stating the backbone dihedral angle parameters. For a single ϕ,ψ bin, the first part term of the mixture model's message length is the summation of 3 message length terms (see Section 2.3 in the main text) associated with stating mixture components, weights of the mixture and component parameters. We calculate these three terms using exactly the same method we use to calculate the MML mixture model's first part (as described in the main text). The total message of stating all mixture parameters is calculated by summing the first part terms over $1,296$ mixtures.

Approach 2

In the second approach, we entirely ignore stating the backbone dihedral angles, by ignoring the corresponding von Mises terms from the derived mixture models and this allows us to compare with Dunbrack's models on an equal footing to explain only sidechain dihedral angles, which they are geared to explain.

Calculating second part. Similar to the first approach, we calculate the message length of stating sidechain dihedral angles of a d-dimensional datum (x_i) by selecting the mixture model associated with the bin into which any observed ϕ_i, ψ_i falls, and using that implied mixture model in the Dunbrack's library to determine the message length of stating sidechain angles of x_i. Compared to the first approach, in this method, we are not sending the backbone dihedral angles. But for Dunbrack's model, we still need to include the corresponding bin information in the second part of the message so that a receiver is able to recover the data losslessly by selecting the correct mixture model. Without that information, the message will not be lossless. Hence, the message length of stating a backbone bin is simply the information content associated with selecting a bin out of 1296 possible bins (assuming each bin is equally probable). As before, as in illustrative example we can apply this to the methionine example to calculate the message length of x_i,

$$I(\langle x_{i_3}, x_{i_4}, x_{i_5}\rangle | \mathcal{D}^{\text{MET}}_{\text{rotamer}}) = -\log\left(\sum_{j=1}^{27}\left(w_j \prod_{p=3}^{5} f(x_{i_p}|\langle \mu_{j_p}, \kappa_{j_p}\rangle)\right)\epsilon^3\right) + \log(1296) \quad (13)$$

Equation 13 states the message length of stating a single datum only. The message length of stating all the data can be calculated by summing over individual message length terms $I\left(\langle x_{i_3}, x_{i_4}, x_{i_5}\rangle | \mathcal{D}^{\text{MET}}_{\text{rotamer}}\right), \forall \leq i \leq N$.

Calculating first part. Similar to calculating the first-part term in the first approach, we calculate the message length required to state parameters of 1296 mixture models in Dunbrack's library, and on the MML side, we ignore the message length associated with stating the parameters associated with backbone dihedral angles.

S10 Sampling from `PhiSiCal` mixture models

The collection of mixture models and conformation sampling methods are accessible from http://lcb.infotech.monash.edu.au/phisical. The formal statistical distributions provide direct ways to sample under their implied distributions. Specifically, each mixture model reports a set of weight parameters of mixture components (product of von Mises distributions) along with their associated parameter estimates. For individual amino acids, the inferred mixture model can be used to randomly sample its dihedral angles in two distinct ways: (1) jointly sample backbone and sidechain dihedral angles $\langle \phi, \psi, \chi_1, \chi_2, ...\rangle$, and (2) conditionally sample only sidechain dihedral angles $\langle \chi_1, \chi_2, ...\rangle$ given any specified backbone dihedral angles $\langle \phi, \psi\rangle$.

Sampling joint (mainchain and sidechain) dihedral angles

Sampling any (backbone and sidechain) dihedral angle vector for any specified amino acid involves these operations:

1. First, identify the inferred PhiSiCal mixture model associated with the specified amino acid.
2. Probabilistically select a component from that identified mixture model. That is, randomly select a component based on the inferred weight parameters of the mixture model. These weights give the probability of selecting a component in the mixture.
3. Once a component is identified, for each von Mises term (in the product of such terms) of that component, randomly sample a dihedral angle from that von Mises distribution given its (μ, κ) parameters.

The resultant dihedral angle vector is a sample from the joint distribution.

Sampling sidechain dihedral angles conditional on backbone dihedral angles

Importantly, given the Bayesian framework on which these models stand, a simple technique of *posterior reweighting* can be employed on any of the inferred mixture models to transform them into conditional distributions, and sample conditionally sidechain dihedrals given any specified backbone dihedral angles. In this method of posterior reweighting, for a mixture model $\mathcal{M}(\Lambda) = \sum_{j=1}^{|\mathcal{M}|} w_i f_i(\Theta_j)$ (see Equation 1 in the main text), only its component-weights $w_1, w_2, \ldots w_{|\mathcal{M}|}$ are updated to $w'_1, w'_2, \ldots w'_{|\mathcal{M}|}$, such that each w'_i is the corresponding posterior probability of the component given $\langle \phi, \psi\rangle$. (Note, the other parameters of the original mixture model are left intact and only the original inferred weights of the mixture model are updated). The resultant $\mathcal{M}(\Lambda|\langle \phi, \psi\rangle) = \sum_{j=1}^{|\mathcal{M}|} w'_i f_i(\Theta_j)$ is now a conditional mixture model (given any $\langle \phi, \psi\rangle$). Since this is yet another mixture model, it can be sampled using the same approach as the one described in the section above. The elegance of this approach is that only the originally inferred 'beliefs' of the component probabilities (i.e., the component weights as given by the original joint mixture model) have been updated based on evidence of some observed $\langle \phi, \psi\rangle$, thus yielding a mixture model conditioned on those observations to sample from.

S11 Assessing the stability of the search algorithm

The employed search algorithm is a core component of our inference process to determine the optimal mixture model *explaining* underlying dihedral angle distributions *unsupervised*. Hence we evaluated the stability of the solutions of this search process thoroughly to assess its applicability to this modelling challenge. The following discusses the method that we employed to evaluate the validity of the solutions.

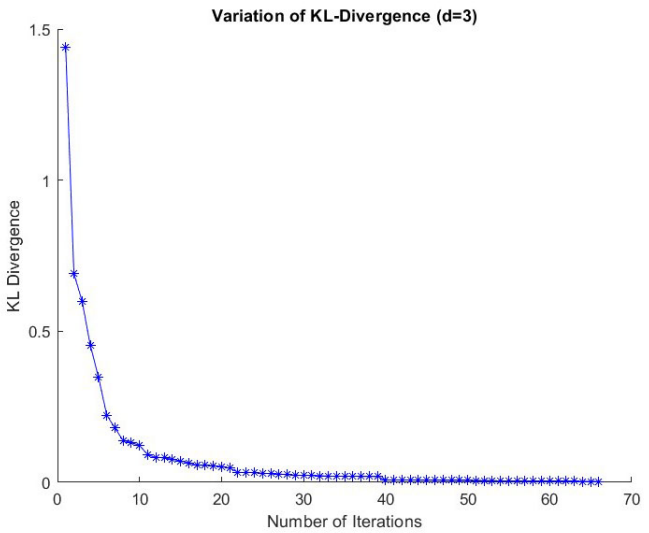


Figure SF 4. Variation of KL-divergence of a mixture inferred from a 3-dimension dihedral angle dataset. During 66 iterations this mixture settled in to 66 components.

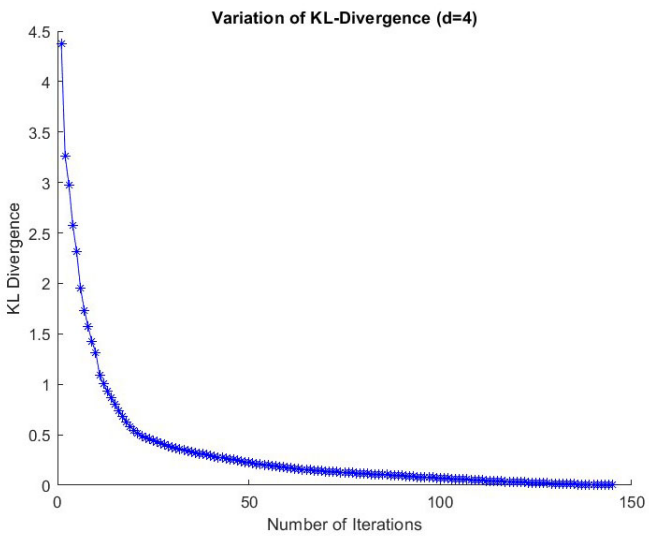


Figure SF 5. Variation of KL-divergence of a mixture inferred from a 4-dimension dihedral angle dataset. During 145 iterations this mixture settled in to 143 components.

S11.0.1 Method

For this task, synthetic data was generated by randomly sampling from mixture models with component parameters and the number of components known a priori. Then the sampled data was utilized to infer new mixture models. In order to assess the reliability of the underlying search process, the derived new mixture models were compared against their original mixture models (true distributions), under the KL-divergence metric (see Equation 10). We considered original mixture distributions having similar dimensions observed in amino acid dihedral angles ($d = 2, .., 7$) and having number of components between 20 - 300 range.

Figures SF 4, SF 5 and SF 6 show variation of KL-divergence between original distribution and inferred model during different iterations of the search process. Starting from a single component and increase(decrease) component count from split(merge,delete) operations and until convergence. From these illustrations it is clearly evident that employed search process converges *near* to the ground truth.

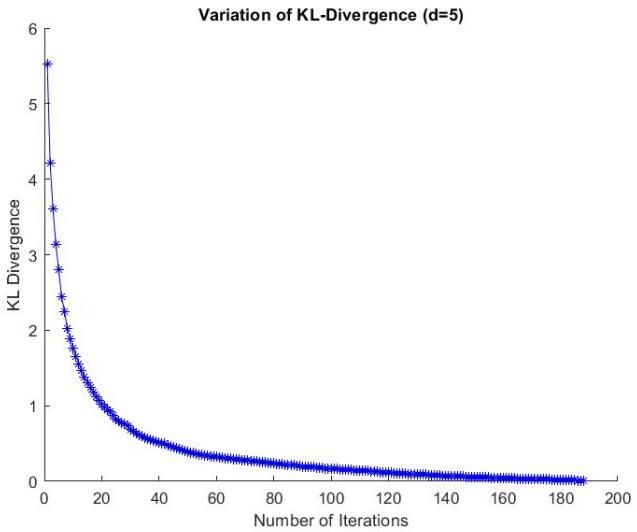


Figure SF 6. Variation of KL-divergence of a mixture inferred from a 5-dimension dihedral angle dataset. During 188 iterations this mixture settled in to 180 components.

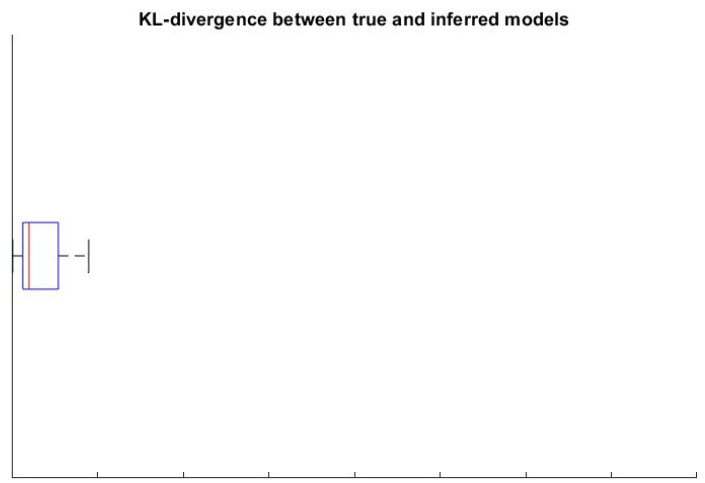


Figure SF 7. Quartile statistics of KL-divergence between true and inferred distributions (converged) from 17 experiments.

Further we observe at convergence KL-divergence quartile statistics, Q_1=0.0076, Q_2 = 0.0100 and Q_3=0.0259 from 17 experiments with varying number of components and dihedral angle dimensions (see Figure SF 7). These statistics further support the stability of the solutions of the underlying EM search process (true and the inferred models are *nearly* similar).

S12 Comparing different smoothing levels of Dunbrack's rotamer library.

In the main text, we presented the comparison between MML-derived mixture models and the Dunbrack rotamer library at 5% smoothing, which were the default libraries that Shapovalov and Dunbrack Jr (2011) earmarked as the best performing. To complete the comparison, we present below additional tables of comparison between the MML-derived mixture models and Dunbrack's rotamer library at varying (2%, 10%, 20%, and 25%) smoothing levels. See supplementary tables ST 5, ST 6, ST 7, ST 8 below.
Further, Figures SF 8, SF 9, SF 10 and SF 11, show a qualitative comparison between the MML-derived mixture model against the Dunbrack's library at those different smoothing levels.
These tables and figures clearly illustrate that the accuracy of Dunbrack's models starts to decline when the smoothing level increases. For example, one can see the decline in the accuracy of the χ_2 of TRP, χ_3 of GLU, χ_3 of GLN when smoothing value increases from 2% to 25% (see Figures SF 8 to SF 11).

S12.1 Message lengths for stating amino acid sidechain dihedral angles from PDB50 dataset: Quantitative comparison for 2% smoothing level of Dunbrack library

Table ST 5. This table illustrates a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) with 2% smoothing level to state sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa). Here we have only considered the cost of stating the sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa) and omitted the backbone $\langle \phi, \psi \rangle$. The 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles.

		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	Null($X^{(\text{aa})}$)	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	2,171,630	(165;1,484)	4,095	17,578,246	17,582,342	8.1	(11,664;57,024)	1,038,297	41,823,773	42,862,070	19.7	26,797,588	12.3
ALA	1,861,359	(25;124)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
VAL	1,601,058	(96;671)	1,625	6,607,950	6,609,575	4.1	(3,888;10,368)	217,597	23,549,959	23,767,556	14.8	9,878,408	6.2
GLY	1,588,115	(30;149)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
GLU	1,446,860	(262;2,881)	7,754	21,695,912	21,703,666	15.0	(69,984;488,592)	9,557,315	37,051,108	46,608,423	32.2	26,781,053	18.5
SER	1,337,273	(114;797)	1,757	6,592,674	6,594,431	4.9	(3,888;10,368)	210,873	20,961,278	21,172,151	15.8	8,250,874	6.2
ILE	1,333,508	(172;1,547)	4,346	10,688,100	10,692,445	8.0	(11,664;57,024)	929,222	25,042,446	25,971,668	19.5	16,455,289	12.3
ASP	1,279,567	(170;1,529)	3,640	12,462,677	12,466,317	9.7	(23,328;115,344)	2,340,479	25,105,109	27,445,589	21.4	15,789,665	12.3
THR	1,221,604	(90;629)	1,445	5,531,460	5,532,905	4.5	(3,888;10,368)	212,264	18,316,720	18,528,984	15.2	7,537,205	6.2
LYS	1,176,395	(266;3,457)	9,833	22,078,096	22,087,929	18.8	(104,976;943,488)	13,524,209	35,501,351	49,025,560	41.7	29,033,076	24.7
ARG	1,130,448	(250;3,749)	12,164	23,504,690	23,516,854	20.8	(104,976;943,488)	14,637,302	35,020,688	49,657,990	43.9	34,873,897	30.8
PRO	1,004,859	(231;2,078)	8,956	5,257,024	5,265,980	5.2	(2,592;11,664)	255,399	16,318,837	16,574,237	16.5	12,399,809	12.3
ASN	948,274	(180;1,619)	3,703	9,582,207	9,585,910	10.1	(46,656;231,984)	4,513,404	19,373,159	23,886,563	25.2	11,701,559	12.3
PHE	927,298	(226;2,033)	5,401	8,460,779	8,466,181	9.1	(23,328;115,344)	2,168,432	17,228,831	19,397,263	20.9	11,442,718	12.3
GLN	820,871	(239;2,628)	6,804	12,270,999	12,277,803	15.0	(139,968;978,480)	17,500,129	21,548,829	39,048,958	47.6	15,194,138	18.5
TYR	788,176	(192;1,727)	4,480	7,193,098	7,197,578	9.1	(23,328;115,344)	2,234,390	14,604,830	16,839,220	21.4	9,725,974	12.3
HIS	515,611	(163;1,466)	3,443	5,175,762	5,179,204	10.0	(46,656;231,984)	4,317,912	10,395,540	14,713,451	28.5	6,362,562	12.3
MET	417,170	(270;2,969)	7,919	5,954,095	5,962,013	14.3	(34,992;243,648)	4,114,107	10,682,963	14,797,070	35.5	7,721,723	18.5
TRP	310,470	(212;1,907)	4,816	2,958,040	2,962,856	9.5	(46,656;231,984)	3,981,481	6,031,196	10,012,677	32.3	3,831,153	12.3
CYS	296,547	(96;671)	1,433	1,389,186	1,390,619	4.7	(3,888;10,368)	191,072	4,438,901	4,629,973	15.6	1,829,673	6.2

S12.2 Message lengths for stating amino acid sidechain dihedral angles from PDB50 dataset: Quantitative comparison for 10% smoothing level of Dunbrack library

Table ST 6. This table illustrates a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) with 10% smoothing level to state sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa). Here we have only considered the cost of stating the sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa) and omitted the backbone $\langle \phi, \psi \rangle$. The 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles.

		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	Null($X^{(\text{aa})}$)	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	2,171,630	(165;1,484)	4,095	17,578,246	17,582,342	8.1	(11,664;57,024)	1,137,226	41,710,760	42,847,985	19.7	26,797,588	12.3
ALA	1,861,359	(25;124)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
VAL	1,601,058	(96;671)	1,625	6,607,950	6,609,575	4.1	(3,888;10,368)	216,601	23,551,213	23,767,814	14.8	9,878,408	6.2
GLY	1,588,115	(30;149)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
GLU	1,446,860	(262;2,881)	7,754	21,695,912	21,703,666	15.0	(69,984;488,592)	9,902,176	37,023,602	46,925,778	32.4	26,781,053	18.5
SER	1,337,273	(114;797)	1,757	6,592,674	6,594,431	4.9	(3,888;10,368)	210,474	20,922,682	21,133,156	15.8	8,250,874	6.2
ILE	1,333,508	(172;1,547)	4,346	10,688,100	10,692,445	8.0	(11,664;57,024)	1,011,840	24,987,375	25,999,216	19.5	16,455,289	12.3
ASP	1,279,567	(170;1,529)	3,640	12,462,677	12,466,317	9.7	(23,328;115,344)	2,334,307	25,077,909	27,412,216	21.4	15,789,665	12.3
THR	1,221,604	(90;629)	1,445	5,531,460	5,532,905	4.5	(3,888;10,368)	211,193	18,261,872	18,473,065	15.1	7,537,205	6.2
LYS	1,176,395	(266;3,457)	9,833	22,078,096	22,087,929	18.8	(104,976;943,488)	14,982,586	35,432,372	50,414,958	42.9	29,033,076	24.7
ARG	1,130,448	(250;3,749)	12,164	23,504,690	23,516,854	20.8	(104,976;943,488)	16,182,120	34,966,919	51,149,039	45.2	34,873,897	30.8
PRO	1,004,859	(231;2,078)	8,956	5,257,024	5,265,980	5.2	(2,592;11,664)	254,275	16,307,481	16,561,756	16.5	12,399,809	12.3
ASN	948,274	(180;1,619)	3,703	9,582,207	9,585,910	10.1	(46,656;231,984)	4,590,827	19,367,626	23,958,453	25.3	11,701,559	12.3
PHE	927,298	(226;2,033)	5,401	8,460,779	8,466,181	9.1	(23,328;115,344)	2,296,537	17,252,420	19,548,957	21.1	11,442,718	12.3
GLN	820,871	(239;2,628)	6,804	12,270,999	12,277,803	15.0	(139,968;978,480)	19,052,891	21,526,082	40,578,973	49.4	15,194,138	18.5
TYR	788,176	(192;1,727)	4,480	7,193,098	7,197,578	9.1	(23,328;115,344)	2,260,947	14,620,212	16,881,159	21.4	9,725,974	12.3
HIS	515,611	(163;1,466)	3,443	5,175,762	5,179,204	10.0	(46,656;231,984)	4,384,777	10,386,932	14,771,709	28.6	6,362,562	12.3
MET	417,170	(270;2,969)	7,919	5,954,095	5,962,013	14.3	(34,992;243,648)	4,336,874	10,655,090	14,991,964	35.9	7,721,723	18.5
TRP	310,470	(212;1,907)	4,816	2,958,040	2,962,856	9.5	(46,656;231,984)	4,148,829	6,055,285	10,204,114	32.9	3,831,153	12.3
CYS	296,547	(96;671)	1,433	1,389,186	1,390,619	4.7	(3,888;10,368)	189,723	4,428,936	4,618,660	15.6	1,829,673	6.2

S12.3 Message lengths for stating amino acid sidechain dihedral angles from PDB50 dataset: Quantitative comparison for 20% smoothing level of Dunbrack library

Table ST 7. This table illustrates a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) with 20% smoothing level to state sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa). Here we have only considered the cost of stating the sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa) and omitted the backbone $\langle \phi, \psi \rangle$. The 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles.

		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	Null($X^{(\text{aa})}$)	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	2,171,630	(165;1,484)	4,095	17,578,246	17,582,342	8.1	(11,664;57,024)	1,204,083	41,623,038	42,827,121	19.7	26,797,588	12.3
ALA	1,861,359	(25;124)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
VAL	1,601,058	(96;671)	1,625	6,607,950	6,609,575	4.1	(3,888;10,368)	215,569	23,595,581	23,811,150	14.9	9,878,408	6.2
GLY	1,588,115	(30;149)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
GLU	1,446,860	(262;2,881)	7,754	21,695,912	21,703,666	15.0	(69,984;488,592)	10,109,636	37,022,486	47,132,122	32.6	26,781,053	18.5
SER	1,337,273	(114;797)	1,757	6,592,674	6,594,431	4.9	(3,888;10,368)	210,388	20,902,387	21,112,775	15.8	8,250,874	6.2
ILE	1,333,508	(172;1,547)	4,346	10,688,100	10,692,445	8.0	(11,664;57,024)	1,119,018	24,951,525	26,070,543	19.6	16,455,289	12.3
ASP	1,279,567	(170;1,529)	3,640	12,462,677	12,466,317	9.7	(23,328;115,344)	2,325,771	25,168,174	27,493,945	21.5	15,789,665	12.3
THR	1,221,604	(90;629)	1,445	5,531,460	5,532,905	4.5	(3,888;10,368)	211,124	18,273,176	18,484,300	15.1	7,537,205	6.2
LYS	1,176,395	(266;3,457)	9,833	22,078,096	22,087,929	18.8	(104,976;943,488)	15,458,386	35,415,191	50,873,578	43.2	29,033,076	24.7
ARG	1,130,448	(250;3,749)	12,164	23,504,690	23,516,854	20.8	(104,976;943,488)	16,740,256	34,949,127	51,689,383	45.7	34,873,897	30.8
PRO	1,004,859	(231;2,078)	8,956	5,257,024	5,265,980	5.2	(2,592;11,664)	253,869	16,319,192	16,573,062	16.5	12,399,809	12.3
ASN	948,274	(180;1,619)	3,703	9,582,207	9,585,910	10.1	(46,656;231,984)	4,585,079	19,443,062	24,028,141	25.3	11,701,559	12.3
PHE	927,298	(226;2,033)	5,401	8,460,779	8,466,181	9.1	(23,328;115,344)	2,290,175	17,377,675	19,667,850	21.2	11,442,718	12.3
GLN	820,871	(239;2,628)	6,804	12,270,999	12,277,803	15.0	(139,968;978,480)	19,446,584	21,542,593	40,989,176	49.9	15,194,138	18.5
TYR	788,176	(192;1,727)	4,480	7,193,098	7,197,578	9.1	(23,328;115,344)	2,253,206	14,733,786	16,986,991	21.6	9,725,974	12.3
HIS	515,611	(163;1,466)	3,443	5,175,762	5,179,204	10.0	(46,656;231,984)	4,374,205	10,421,993	14,796,198	28.7	6,362,562	12.3
MET	417,170	(270;2,969)	7,919	5,954,095	5,962,013	14.3	(34,992;243,648)	4,449,503	10,637,850	15,087,353	36.2	7,721,723	18.5
TRP	310,470	(212;1,907)	4,816	2,958,040	2,962,856	9.5	(46,656;231,984)	4,222,971	6,112,991	10,335,962	33.3	3,831,153	12.3
CYS	296,547	(96;671)	1,433	1,389,186	1,390,619	4.7	(3,888;10,368)	189,461	4,427,876	4,617,337	15.6	1,829,673	6.2

S12.4 Message lengths for stating amino acid sidechain dihedral angles from PDB50 dataset: Quantitative comparison for 25% smoothing level of Dunbrack library

Table ST 8. This table illustrates a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) with 25% smoothing level to state sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa). Here we have only considered the cost of stating the sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa) and omitted the backbone $\langle \phi, \psi \rangle$. The 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles.

		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	Null($X^{(\text{aa})}$)	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	2,171,630	(165;1,484)	4,095	17,578,246	17,582,342	8.1	(11,664;57,024)	1,202,034	41,619,581	42,821,614	19.7	26,797,588	12.3
ALA	1,861,359	(25;124)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
VAL	1,601,058	(96;671)	1,625	6,607,950	6,609,575	4.1	(3,888;10,368)	215,428	23,603,397	23,818,825	14.9	9,878,408	6.2
GLY	1,588,115	(30;149)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
GLU	1,446,860	(262;2,881)	7,754	21,695,912	21,703,666	15.0	(69,984;488,592)	10,117,196	37,040,961	47,158,157	32.6	26,781,053	18.5
SER	1,337,273	(114;797)	1,757	6,592,674	6,594,431	4.9	(3,888;10,368)	210,361	20,901,312	21,111,674	15.8	8,250,874	6.2
ILE	1,333,508	(172;1,547)	4,346	10,688,100	10,692,445	8.0	(11,664;57,024)	1,131,142	24,904,652	26,035,795	19.5	16,455,289	12.3
ASP	1,279,567	(170;1,529)	3,640	12,462,677	12,466,317	9.7	(23,328;115,344)	2,323,875	25,227,990	27,551,865	21.5	15,789,665	12.3
THR	1,221,604	(90;629)	1,445	5,531,460	5,532,905	4.5	(3,888;10,368)	211,013	18,291,157	18,502,169	15.1	7,537,205	6.2
LYS	1,176,395	(266;3,457)	9,833	22,078,096	22,087,929	18.8	(104,976;943,488)	15,751,256	35,410,141	51,161,397	43.5	29,033,076	24.7
ARG	1,130,448	(250;3,749)	12,164	23,504,690	23,516,854	20.8	(104,976;943,488)	17,019,872	34,948,061	51,967,933	46.0	34,873,897	30.8
PRO	1,004,859	(231;2,078)	8,956	5,257,024	5,265,980	5.2	(2,592;11,664)	253,628	16,345,948	16,599,576	16.5	12,399,809	12.3
ASN	948,274	(180;1,619)	3,703	9,582,207	9,585,910	10.1	(46,656;231,984)	4,578,420	19,510,335	24,088,755	25.4	11,701,559	12.3
PHE	927,298	(226;2,033)	5,401	8,460,779	8,466,181	9.1	(23,328;115,344)	2,285,453	17,463,173	19,748,626	21.3	11,442,718	12.3
GLN	820,871	(239;2,628)	6,804	12,270,999	12,277,803	15.0	(139,968;978,480)	19,540,094	21,556,082	41,096,176	50.1	15,194,138	18.5
TYR	788,176	(192;1,727)	4,480	7,193,098	7,197,578	9.1	(23,328;115,344)	2,248,767	14,808,452	17,057,218	21.6	9,725,974	12.3
HIS	515,611	(163;1,466)	3,443	5,175,762	5,179,204	10.0	(46,656;231,984)	4,369,499	10,444,307	14,813,806	28.7	6,362,562	12.3
MET	417,170	(270;2,969)	7,919	5,954,095	5,962,013	14.3	(34,992;243,648)	4,498,337	10,632,880	15,131,217	36.3	7,721,723	18.5
TRP	310,470	(212;1,907)	4,816	2,958,040	2,962,856	9.5	(46,656;231,984)	4,224,051	6,136,965	10,361,015	33.4	3,831,153	12.3
CYS	296,547	(96;671)	1,433	1,389,186	1,390,619	4.7	(3,888;10,368)	189,285	4,429,420	4,618,705	15.6	1,829,673	6.2

S12.5 Model fit across all amino acid sidechain dihedral angles for PDB50 dataset: Qualitative comparison for 2% smoothing level of Dunbrack library

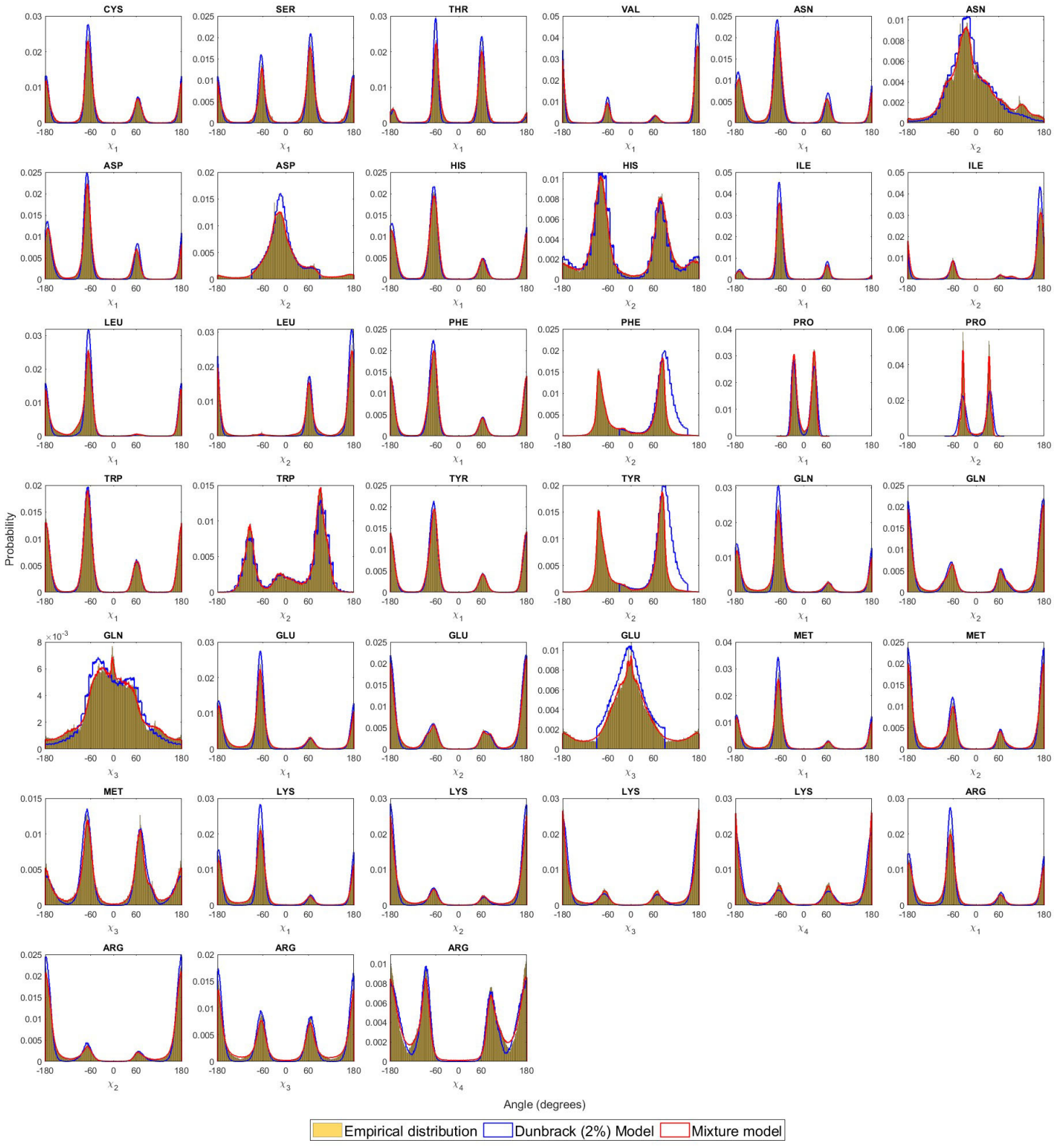


Figure SF 8. Fidelity of the inferred MML mixture models: the projected distribution of individual sidechain dihedral angles across all amino acids derived by randomly sampling $N^{(aa)}$ datapoints (see Table ST 5) from MML-derived mixture models and Dunbrack (2% smoothed) library, and compared to the empirical distribution.

S12.6 Model fit across all amino acid sidechain dihedral angles for PDB50 dataset: Qualitative comparison for 10% smoothing level of Dunbrack library

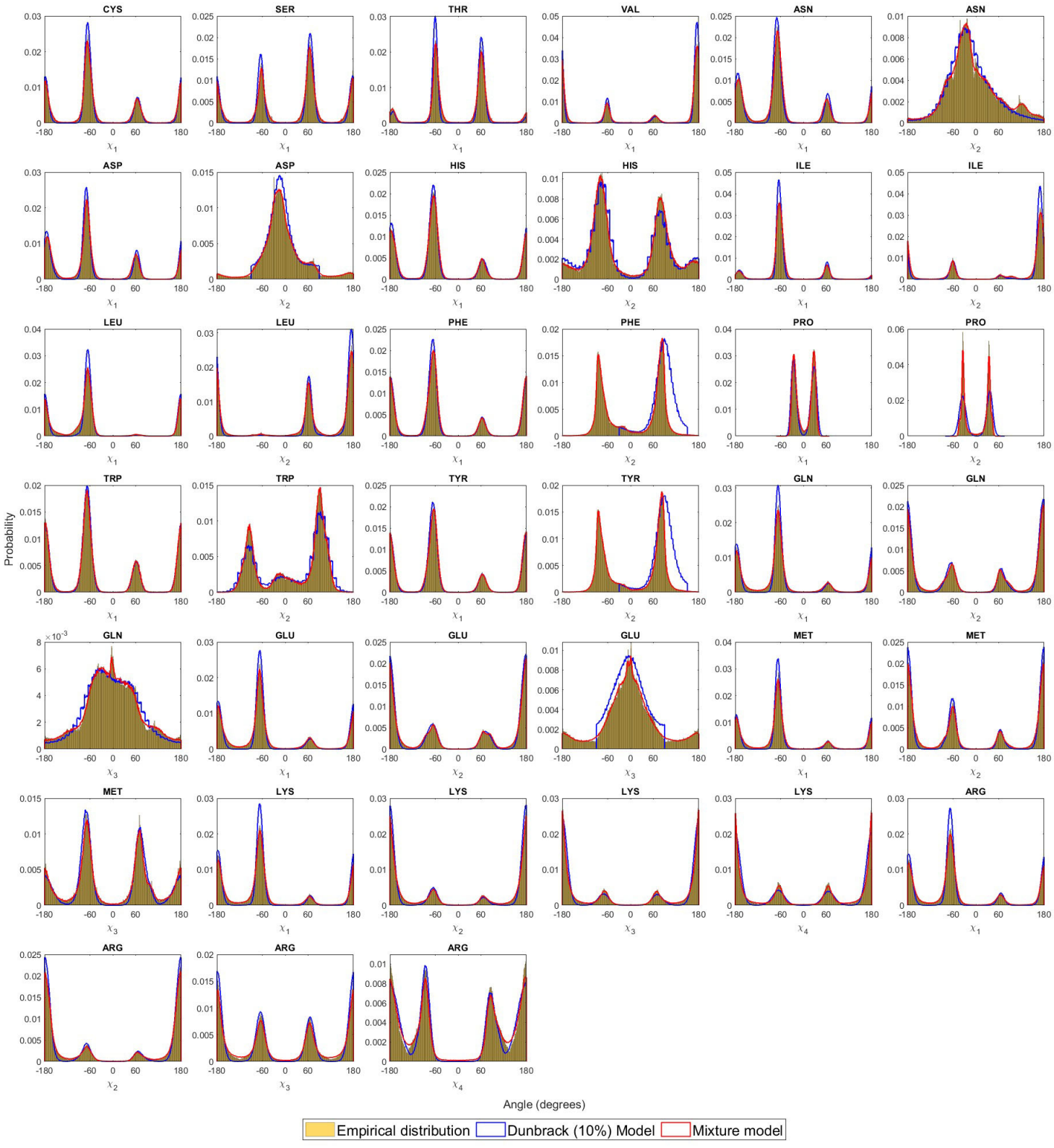


Figure SF 9. Fidelity of the inferred MML mixture models: the projected distribution of individual sidechain dihedral angles across all amino acids derived by randomly sampling $N^{(aa)}$ datapoints (see Table ST 6) from MML derived mixture models and Dunbrack (10% smoothed) library, and compared to the empirical distribution.

S12.7 Model fit across all amino acid sidechain dihedral angles for PDB50 dataset: Qualitative comparison for 20% smoothing level of Dunbrack library

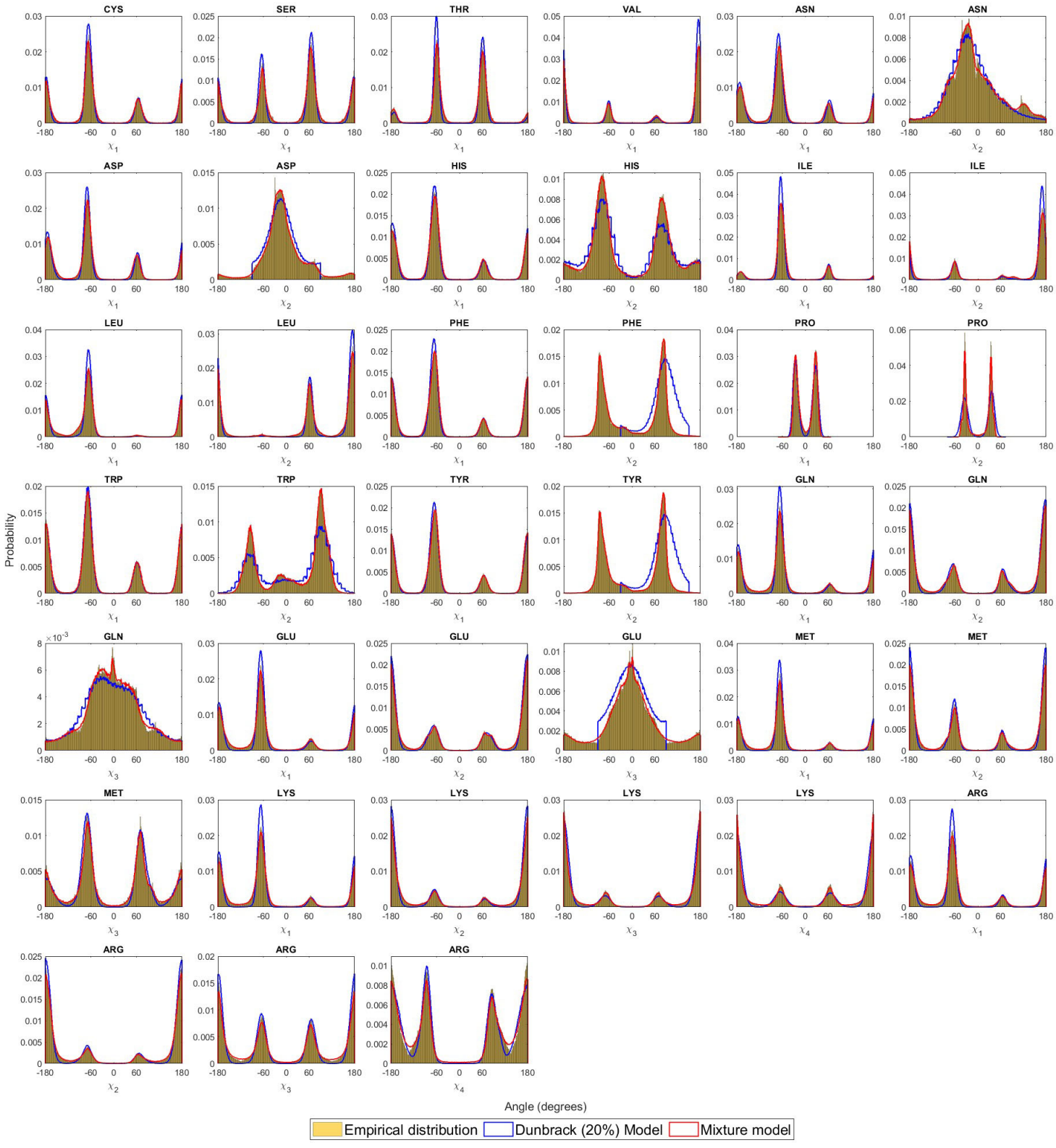


Figure SF 10. Fidelity of the inferred MML mixure models: the projected distribution of individual sidechain dihedral angles across all amino acids derived by randomly sampling $N^{(aa)}$ datapoints (see Table ST 7) from MML derived mixture models and Dunbrack (20% smoothed) library, and compared to the empirical distribution.

S12.8 Model fit across all amino acid sidechain dihedral angles for PDB50 dataset: Qualitative comparison for 25% smoothing level of Dunbrack library

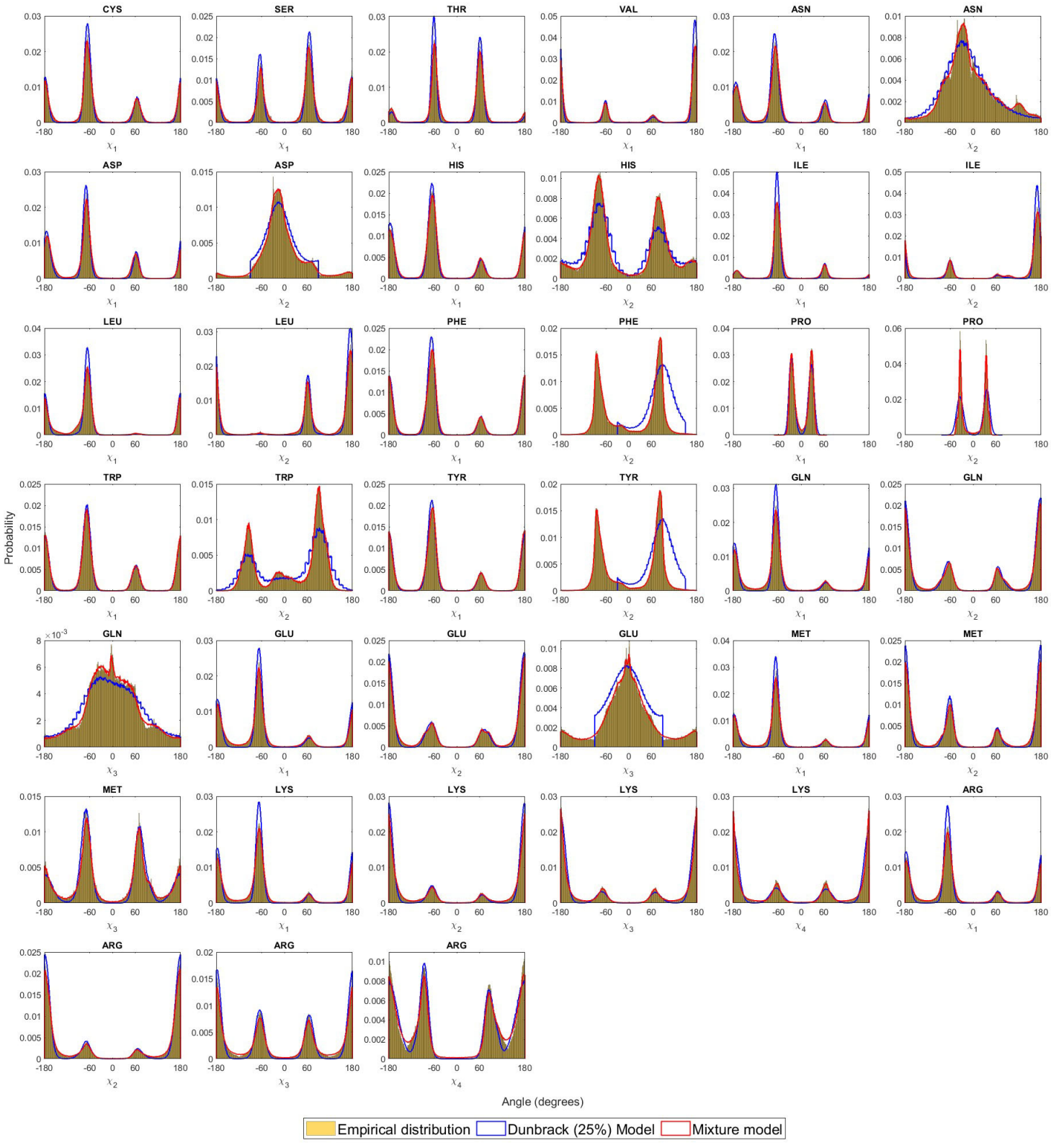


Figure SF 11. Fidelity of the inferred MML mixure models: the projected distribution of individual sidechain dihedral angles across all amino acids derived by randomly sampling $N^{(aa)}$ datapoints (see Table ST 8) from MML derived mixture models and Dunbrack (25% smoothed) library, and compared to the empirical distribution.

S13 Testing Mixture models for overfitting

The MML framework provides a trade-off between model complexity and model fit when inferring a model (Allison, 2018). In order to assess the MML-derived mixture models' ability to capture the underlying dihedral angle distribution without overfitting to the data. We conducted a test to access the fidelity of the derived mixture models to explain an unforeseen dataset. For this test, we considered a collection of 2,238 protein structures that is part of the 2010 Dunbrack Rotamer Library but was not included in our dataset. We compared the dihedral angles calculated from these structures against data sampled from mixture models. Figure SF 12 illustrate how well mixture models capture the underlying distribution, even when the empirical distribution comprises dihedral angles that were not a part of the initial training set. Additionally, Dunbrack's model (with 5%) is included in the same figure for further comparison.

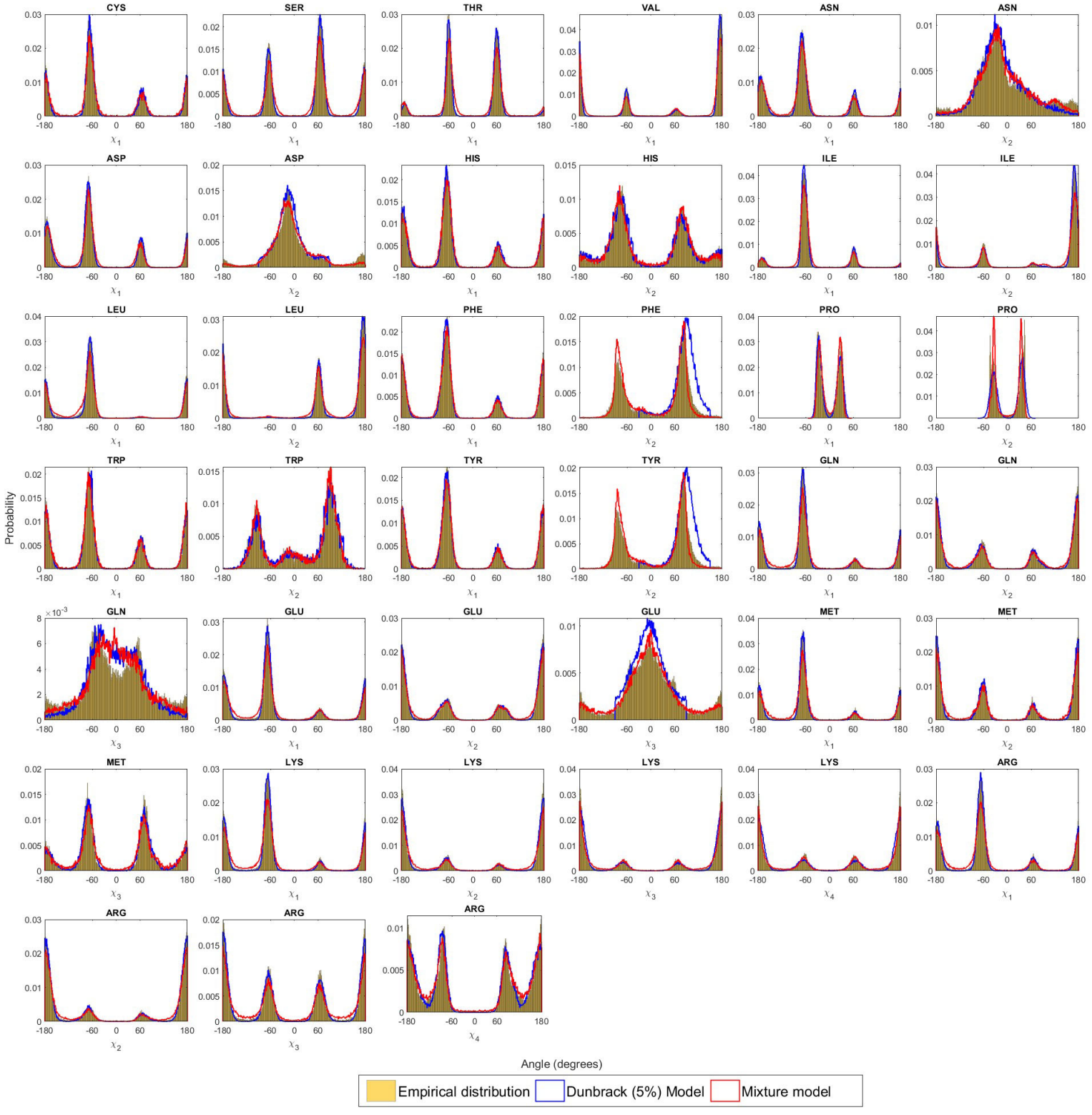


Figure SF 12. the projected distribution of individual sidechain dihedral angles across all amino acids derived by randomly sampling data points from MML-derived mixture models compared against an empirical dataset that is not a part of the mixture models training set.

S14 Assessing the validity of the PDB50HighRes dataset

We assessed the quality of the PDB50HighRes dataset curated from the Protein Data Bank with the corresponding refined protein structures from the PDB-REDO databank (van Beusekom *et al.*, 2018). For each amino acid, a collection of d-dimensional deviation (δ) vectors was created by calculating the deviation of each dihedral angle of the corresponding residue in the PDB-REDO refined structure and in the PDB50HighRes structure. The resulting collection of δ dihedral angle vectors was then statistically analyzed to assess the validity of the dataset used for this research work. The mean and standard deviations of the deviations are presented in Table ST 9, demonstrating that the differences are insignificant.

Table ST 9. This table illustrates mean and standard deviation (SD) statistics (in radians) of dihedral angle deviations δ between PDB-REDO refined structures for the PDB50HighRes dataset. Statistics are presented per each amino acid (a.a) and per each dihedral angle $\langle \phi, \psi, \chi_1, \chi_2, ... \rangle$.

	ϕ_δ (rad)		ψ_δ (rad)		χ_{1_δ} (rad)		χ_{2_δ} (rad)		χ_{3_δ} (rad)		χ_{4_δ} (rad)		χ_{5_δ} (rad)	
a.a	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ALA	-0.00650	0.05218	0.00590	0.04708										
GLY	-0.00029	0.06921	-0.00162	0.06539										
VAL	-0.00301	0.04698	0.00500	0.03940	-0.00145	0.04974								
SER	-0.00808	0.05958	0.00670	0.05394	-0.00486	0.10115								
THR	-0.00615	0.05375	0.00673	0.04564	-0.00734	0.05934								
CYS	-0.00581	0.04966	0.00694	0.04315	-0.00590	0.07286								
ILE	-0.00424	0.04689	0.00581	0.03853	-0.00568	0.04758	0.00321	0.09040						
LEU	-0.00588	0.04780	0.00600	0.04109	-0.00374	0.05630	0.00193	0.06668						
PRO	-0.00459	0.05462	0.00138	0.05284	-0.00636	0.11185	0.00589	0.16209						
PHE	-0.00505	0.04714	0.00645	0.04240	-0.00499	0.03757	-0.00060	0.05874						
TRP	-0.00525	0.04744	0.00681	0.04000	-0.00516	0.03472	-0.00033	0.04347						
TYR	-0.00613	0.04606	0.00985	0.04041	-0.00469	0.03492	-0.00192	0.05799						
ASP	-0.00431	0.05880	0.00566	0.05419	-0.00507	0.07146	-0.00214	0.13428						
HIS	-0.00618	0.05235	0.00733	0.04772	-0.00348	0.04822	-0.00209	0.09082						
ASN	-0.00330	0.05627	0.00517	0.05292	-0.00461	0.06436	-0.00053	0.11478						
GLU	-0.00764	0.05770	0.00739	0.05310	-0.00261	0.10601	0.00164	0.10872	-0.00180	-0.00180				
MET	-0.00652	0.05297	0.00729	0.04562	-0.00223	0.08656	0.00091	0.08828	-0.00329	-0.00329				
GLN	-0.00708	0.05416	0.00682	0.04951	-0.00363	0.09195	0.00052	0.09428	-0.00124	-0.00124				
LYS	-0.00651	0.05830	0.00642	0.05240	-0.00282	0.10009	0.00096	0.13190	-0.00053	-0.00053	-0.00077	-0.00077		
ARG	-0.00621	0.05386	0.00902	0.04920	-0.00100	0.09482	-0.00085	0.10300	-0.00224	-0.00224	-0.00026	-0.00026	0.00025	0.08824

References

Allison, L. (2018). *Coding Ockham's Razor*. Springer.

Banerjee, A. *et al.* (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, **6**(9).

Kasarapu, P. and Allison, L. (2015). Minimum message length estimation of mixtures of multivariate gaussian and von mises-fisher distributions. *Machine Learning*, **100**(2), 333–378.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, **22**(1), 79–86.

Shapovalov, M. V. and Dunbrack Jr, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, **19**(6), 844–858.

van Beusekom, B. *et al.* (2018). Homology-based hydrogen bond information improves crystallographic structures in the pdb. *Protein Science*, **27**(3), 798–808.

Wallace, C. S. and Freeman, P. R. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society: Series B (Methodological)*, **49**(3), 240–252.

Table complementing supplementary S4: Quantitative comparison of message lengths for stating amino acid (backbone + sidechain) dihedral angles from PDB50HighRes dataset using PDB50HighRes-inferred mixture models

This table provides a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) for stating dihedral angles (backbone + sidechain) of each of the twenty naturally occurring amino acids (aa). The the 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles. While we model the joint distributions of dihedral including the backbone, Dunbrack on the other hand only provides sidechain distributions conditional on the backbone. Hence ALA and GLY Dunbrack libraries are necessarily empty.

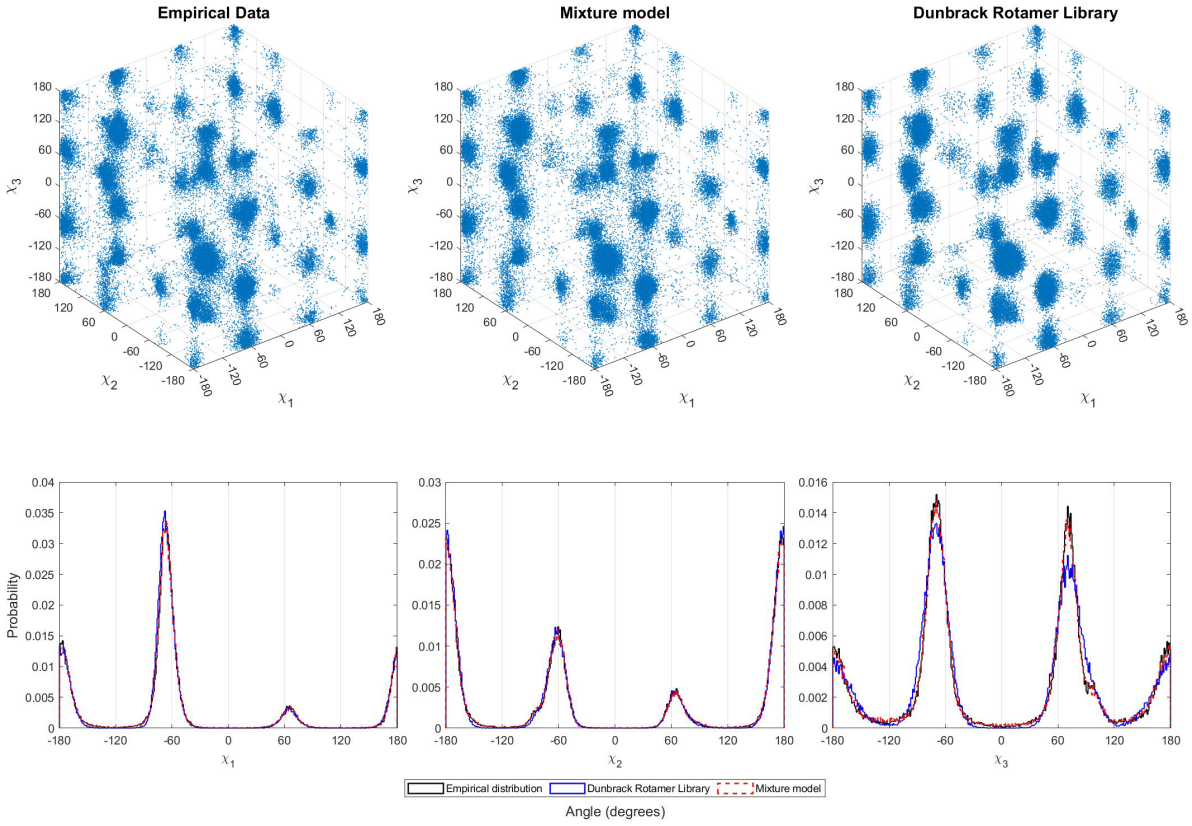
		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	$\text{Null}(X^{(\text{aa})})$	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	343,752	(152;1,367)	6,939	4,926,340	4,933,279	14.4	(11,664;57,024)	950,779	6,587,654	7,538,433	21.9	8,483,696	24.7
ALA	334,111	(26;129)	779	2,509,435	2,510,214	7.5	(N/A;N/A)	N/A	N/A	N/A	N/A	4,122,880	12.3
GLY	294,278	(35;174)	900	2,829,700	2,830,600	9.6	(N/A;N/A)	N/A	N/A	N/A	N/A	3,631,346	12.3
VAL	274,596	(97;678)	3,706	2,918,937	2,922,643	10.6	(3,888;10,368)	192,834	4,163,843	4,356,677	15.9	5,082,710	18.5
GLU	238,682	(240;2,639)	11,801	5,095,773	5,107,574	21.4	(69,984;488,592)	8,509,192	6,189,754	14,698,946	61.6	7,363,250	30.8
ASP	227,558	(219;1,970)	9,252	3,801,033	3,810,284	16.7	(23,328;115,344)	2,062,259	4,632,995	6,695,254	29.4	5,616,063	24.7
SER	222,721	(110;769)	3,859	2,816,672	2,820,530	12.7	(3,888;10,368)	185,948	3,657,879	3,843,828	17.3	4,122,516	18.5
ILE	215,684	(130;1,169)	6,073	2,978,210	2,984,282	13.8	(11,664;57,024)	848,230	4,018,755	4,866,985	22.6	5,323,016	24.7
THR	212,562	(85;594)	3,076	2,488,418	2,491,493	11.7	(3,888;10,368)	187,511	3,295,033	3,482,543	16.4	3,934,475	18.5
LYS	195,868	(368;4,783)	19,991	4,962,216	4,982,208	25.4	(104,976;943,488)	12,526,749	5,878,978	18,405,727	94.0	7,250,945	37.0
ARG	188,400	(353;5,294)	24,041	5,201,624	5,225,666	27.7	(104,976;943,488)	13,518,806	5,758,992	19,277,798	102.3	8,136,897	43.2
PRO	177,534	(146;1,313)	8,925	1,974,713	1,983,637	11.2	(2,592;11,664)	225,394	3,195,740	3,421,135	19.3	4,381,486	24.7
ASN	162,196	(224;2,015)	9,194	2,865,575	2,874,770	17.7	(46,656;231,984)	4,025,549	3,472,766	7,498,315	46.2	4,002,949	24.7
PHE	153,192	(199;1,790)	8,552	2,516,999	2,525,552	16.5	(23,328;115,344)	1,940,884	3,077,402	5,018,286	32.8	3,780,733	24.7
GLN	136,703	(225;2,474)	10,776	2,947,251	2,958,026	21.6	(139,968;978,480)	16,100,149	3,615,527	19,715,676	144.2	4,217,236	30.8
TYR	134,950	(164;1,475)	6,884	2,237,744	2,244,627	16.6	(23,328;115,344)	1,970,652	2,718,877	4,689,528	34.8	3,330,526	24.7
HIS	89,382	(188;1,691)	7,605	1,593,555	1,601,160	17.9	(46,656;231,984)	3,818,112	1,928,561	5,746,674	64.3	2,205,921	24.7
MET	68,907	(209;2,298)	10,259	1,425,449	1,435,708	20.8	(34,992;243,648)	3,657,582	1,752,132	5,409,714	78.5	2,125,755	30.8
TRP	56,696	(154;1,385)	6,406	954,385	960,791	16.9	(46,656;231,984)	3,547,218	1,197,638	4,744,855	83.7	1,399,240	24.7
CYS	46,435	(72;503)	2,436	577,807	580,243	12.5	(3,888;10,368)	164,549	747,043	911,592	19.6	859,501	18.5

Table complementing supplementary S5: Quantitative comparison of message lengths for stating amino acid sidechain dihedral angles from PDB50HighRes dataset using PDB50HighRes-inferred mixture models

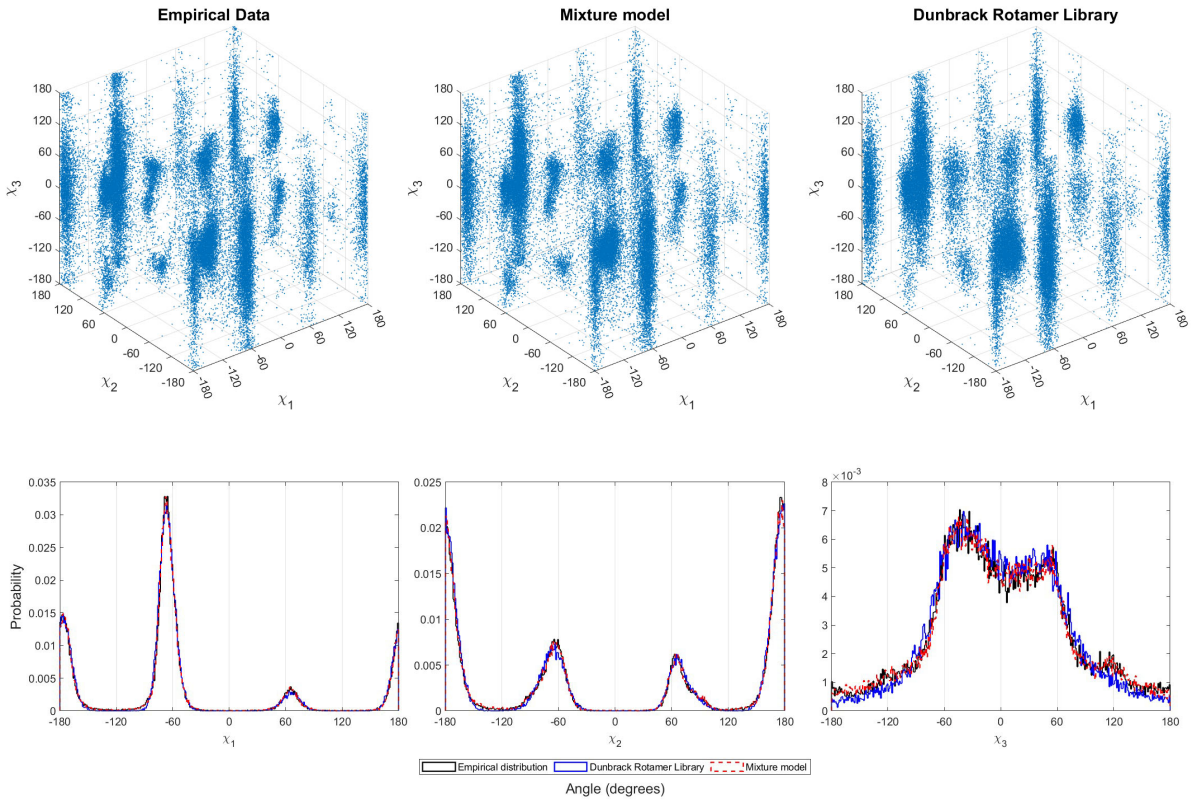
This table illustrates a quantitative comparison between the MML-inferred mixture model ($\mathcal{M}^{(\text{aa})}$) and that of Dunbrack rotamer library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) to state only sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa). Here we have only considered the cost of stating the sidechain dihedral angles of each of the twenty naturally occurring amino acids (aa) and omitted the backbone $\langle\phi, \psi\rangle$. The 'N/A' terms across Alanine (ALA) and Glycine (GLY) arise because those amino acids do not have sidechain dihedral angles.

		MML Mixture Model ($\mathcal{M}^{(\text{aa})}$) message length statistics in bits (rounded)					Dunbrack Rotamer Library ($\mathcal{D}^{(\text{aa})}_{\text{rotamer}}$) message length statistics in bits (rounded)					Null Model (Raw) in bits	
(aa)	$N^{(\text{aa})}$	($\lvert\mathcal{M}^{(\text{aa})}\rvert$, $\lvert\Lambda^{(\text{aa})}\rvert$)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	($\lvert\mathcal{D}^{(\text{aa})}_{\text{rotamer}}\rvert$; #Params)	first-part (complexity)	second-part (fit)	Total (complexity+fit)	$\frac{\text{Total}}{N^{(\text{aa})}}$	$\text{Null}(X^{(\text{aa})})$	$\frac{\text{Null}(X^{(\text{aa})})}{N^{(\text{aa})}}$
LEU	343,752	(152;1,367)	3,957	2,405,711	2,409,669	7.0	(11,664;57,024)	950,774	5,900,150	6,850,924	19.9	4,241,848	12.3
ALA	334,111	(26;129)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
GLY	294,278	(35;174)	N/A	N/A	N/A	N/A	(N/A;N/A)	N/A	N/A	N/A	N/A	N/A	N/A
VAL	274,596	(97;678)	1,692	975,607	977,299	3.6	(3,888;10,368)	192,829	3,614,651	3,807,480	13.9	1,694,237	6.2
GLU	238,682	(240;2,639)	7,241	3,347,665	3,354,906	14.1	(69,984;488,592)	8,509,187	5,712,390	14,221,577	59.6	4,417,950	18.5
ASP	227,558	(219;1,970)	5,003	2,056,298	2,061,301	9.1	(23,328;115,344)	2,062,254	4,177,879	6,240,133	27.4	2,808,032	12.3
SER	222,721	(110;769)	1,696	994,203	995,899	4.5	(3,888;10,368)	185,943	3,212,437	3,398,380	15.3	1,374,172	6.2
ILE	215,684	(130;1,169)	3,481	1,519,622	1,523,103	7.1	(11,664;57,024)	848,225	3,587,387	4,435,612	20.6	2,661,508	12.3
THR	212,562	(85;594)	1,418	838,398	839,817	4.0	(3,888;10,368)	187,506	2,869,909	3,057,414	14.4	1,311,492	6.2
LYS	195,868	(368;4,783)	13,295	3,442,340	3,455,635	17.6	(104,976;943,488)	12,526,744	5,487,242	18,013,986	92.0	4,833,963	24.7
ARG	188,400	(353;5,294)	18,048	3,739,597	3,757,645	19.9	(104,976;943,488)	13,518,801	5,382,192	18,900,993	100.3	5,812,069	30.8
PRO	177,534	(146;1,313)	5,524	910,641	916,165	5.2	(2,592;11,664)	225,389	2,840,672	3,066,061	17.3	2,190,743	12.3
ASN	162,196	(224;2,015)	4,843	1,559,593	1,564,435	9.6	(46,656;231,984)	4,025,544	3,148,374	7,173,918	44.2	2,001,474	12.3
PHE	153,192	(199;1,790)	4,767	1,349,699	1,354,466	8.8	(23,328;115,344)	1,940,879	2,771,018	4,711,896	30.8	1,890,366	12.3
GLN	136,703	(225;2,474)	6,544	1,921,179	1,927,724	14.1	(139,968;978,480)	16,100,144	3,342,121	19,442,265	142.2	2,530,342	18.5
TYR	134,950	(164;1,475)	3,750	1,197,881	1,201,631	8.9	(23,328;115,344)	1,970,647	2,448,977	4,419,623	32.8	1,665,263	12.3
HIS	89,382	(188;1,691)	4,011	867,532	871,543	9.8	(46,656;231,984)	3,818,107	1,749,797	5,567,904	62.3	1,102,960	12.3
MET	68,907	(209;2,298)	6,489	905,273	911,762	13.2	(34,992;243,648)	3,657,577	1,614,318	5,271,895	76.5	1,275,453	18.5
TRP	56,696	(154;1,385)	3,507	529,675	533,183	9.4	(46,656;231,984)	3,547,212	1,084,246	4,631,458	81.7	699,620	12.3
CYS	46,435	(72;503)	1,065	200,506	201,571	4.3	(3,888;10,368)	164,544	654,173	818,717	17.6	286,500	6.2

Figure complementing supplementary S6: Qualitative comparison of model fit for methionine(MET) and glutamine(GLN) sidechain dihedral angles from PDB50HighRes dataset using PDB50HighRes-inferred mixture models



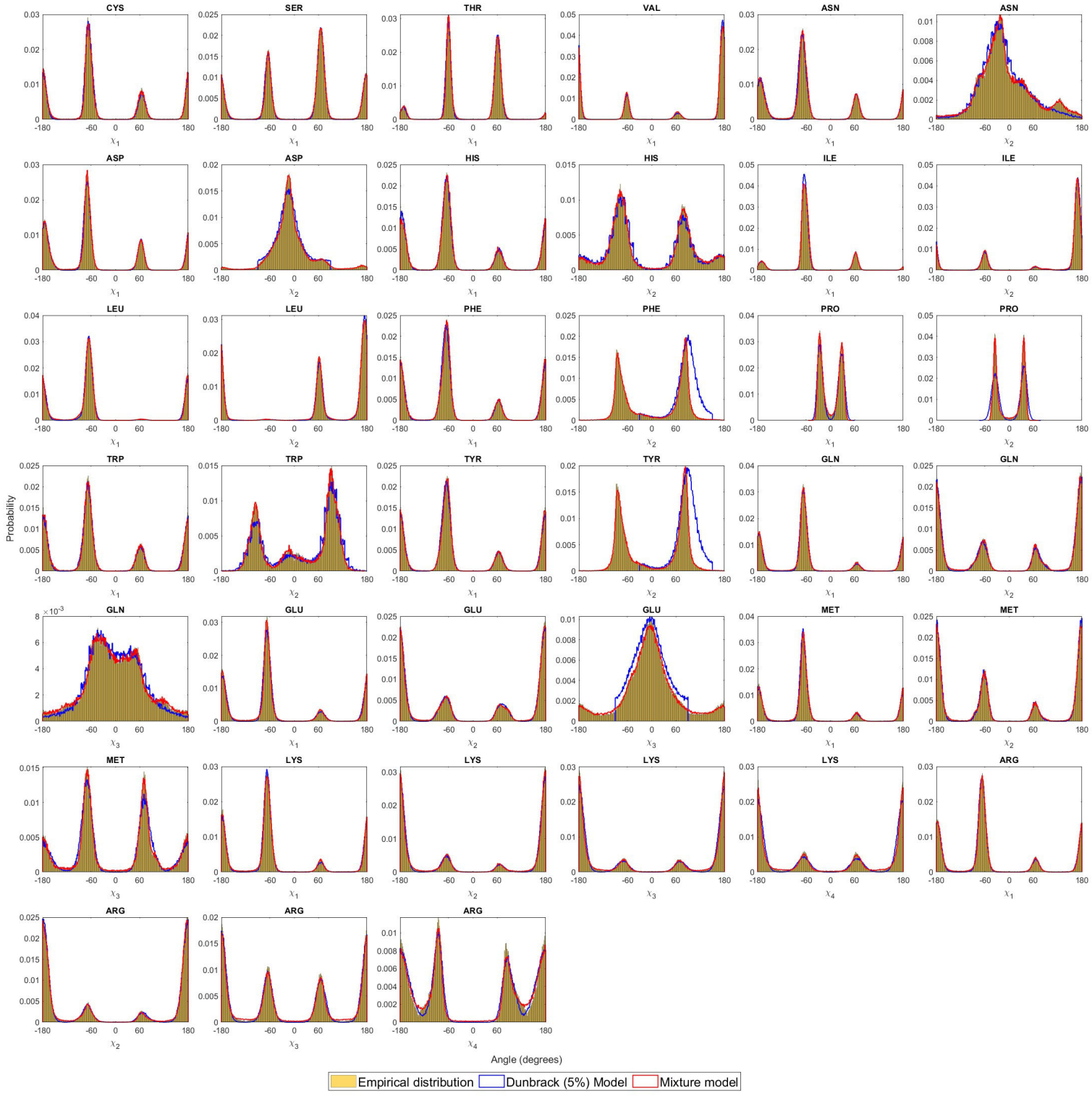
(a) Methionine (MET)



(b) Glutamine (GLN)

(a) The projection, into the sidechain (χ_1, χ_2, χ_3) space (unwrapped), of 50,000 randomly sampled points (vector of dihedral angles) for the amino acid Methionine (MET) from MML mixture model (first row, center), of the same number of points from the Dunbrack model (first row, right), and of the observed (empirical) distribution of the same angles (first row, left) from PDBHighRes. In the plots of the second row, the same data is visualized differently over three separate plots, with each of the three sidechain dihedral angles as x-axis (unwrapped), with y-axis showing the corresponding relative probabilities (in a 1° intervals). (b) The third and fourth rows plots are similar to first and second, respectively, but for the *non-rotameric* amino acid, Glutamine (GLN).

Figure complementing supplementary S7:Qualitative comparison of model fit across all amino acid sidechain dihedral angles from PDB50HighRes dataset using PDB50HighRes-inferred mixture models



Fidelity of the inferred MML mixure models: the projected distribution of individual sidechain dihedral angles across all amino acids derived by randomly sampling $N^{(aa)}$ datapoints (see Table ST 1) from MML-derived mixture models and Dunbrack (5% smoothed) library, and compared to the empirical distribution.