

# MolProbity's Ultimate Rotamer-Library Distributions for Model Validation

Bradley J. Hintze, Steven M. Lewis, Jane S. Richardson, and David C.  
Richardson

Department of Biochemistry, Duke University

February 16, 2016

Short title: MolProbity's Ultimate Rotamer-Library

Keywords: sidechain rotamer library, rare sidechain conformations, structural bioinformatics, structure validation, Phenix, high-quality dataset, protein conformation

All work performed at Duke University, Durham, North Carolina.

Correspondence to: David C. Richardson, 132 Nanaline Duke Building, Duke University, Durham NC 27710. E-mail: dcrjsr@kinemage.biochem.duke.edu

# Contents

<b>1 Abstract</b>	<b>3</b>
<b>2 Introduction</b>	<b>4</b>
<b>3 Methods</b>	<b>7</b>
3.1 Chain-level Dataset Filters . . . . .	7
3.2 Residue-level Dataset Filters . . . . .	8
3.3 Determination of Distributions, Scores, and Contours . . . . .	10
3.4 $\chi$ and Covalent Bond Angle Statistics . . . . .	11
3.5 Rotamer Assignment and Score . . . . .	12
<b>4 Results</b>	<b>13</b>
4.1 Residue Filter Effects . . . . .	14
4.2 Rotamer Evaluation . . . . .	14
4.3 Dihedral and Bond Angle Deviations . . . . .	16
4.4 Interesting Rotamer Anecdotes . . . . .	17
4.4.1 Glutamate <b>pm20</b> & <b>mp0</b> . . . . .	17
4.4.2 Isoleucine <b>pp</b> . . . . .	18
4.4.3 Methionine <b>ppm</b> . . . . .	18
4.5 Conformations Between Tetrahedral and Planar Atoms . . . . .	19
<b>5 Discussion</b>	<b>20</b>
5.1 In what sense is this ultimate? . . . . .	20
5.2 Use of the Top8000 rotamer distributions in model building . . . . .	21
5.3 I have a rotamer outlier; is it wrong? . . . . .	22
<b>6 References</b>	<b>24</b>
<b>7 Figure Legends</b>	<b>28</b>
<b>8 Tables</b>	<b>29</b>
<b>9 Figures</b>	<b>31</b>
<b>10 Acknowledgments</b>	<b>35</b>

# 1 Abstract

Here we describe the updated MolProbity rotamer-library distributions derived from an order-of-magnitude larger and more stringently quality-filtered dataset of about 8000 (vs. 500) protein chains, and we explain the resulting changes and improvements to model validation as seen by users. To include only sidechains with satisfactory justification for their given conformation, we added residue-specific filters for electron-density value and model-to-density fit. The combined new protocol retains a million residues of data, while cleaning up **false-positive** noise in the multi- $\chi$  **datapoint** distributions. It enables clean characterization of conformational clusters nearly 1000-fold less frequent than the most common ones. We describe examples of local interactions that favor these rare conformations, including the role of authentic covalent bond-angle deviations in enabling presumably strained sidechain conformations. Further, along with *favored* and *outlier*, an *allowed* category (0.3% to 2.0% occurrence in reference data) has been added, analogous to Ramachandran validation categories. The new rotamer distributions are used for current rotamer validation in MolProbity and *PHENIX*, and for rotamer choice in *PHENIX* model-building and refinement. The multi-dimensional  $\chi$  distributions and Top8000 reference dataset are available on GitHub. These rotamers are termed "ultimate" because data sampling and quality are now fully adequate for this task, and also because we believe the future of conformational validation should integrate sidechain and backbone **criteria**.

## 2 Introduction

Protein sidechains take on preferred conformations which fall into distinct local energy minima known as rotamers, defined by the set of sidechain dihedral ( $\chi$ ) angles. For tetrahedral geometry,  $\chi$  values fall into three discrete ranges: **p** (plus, centered near  $+60^\circ$ ), **t** (trans, centered near  $180^\circ$ ), and **m** (minus, centered near  $-60^\circ$ ), as named in[1]<sup>1</sup>. These correspond to low-energy staggered conformations expected between  $sp^3$  hybridized atoms [3].  $sp^3$  to  $sp^2$  bonds (tetrahedral to planar geometries) have more complex and much broader distributions. Overall rotamer conformations, however, are not simply the product of their individual  $\chi$  distributions, since wider steric and other atomic interactions influence the preferred, or even the possible, combinations.

Rotamers have been studied extensively since the concept was introduced by Ponder and Richards in 1987 [4], and they are important tools in structural biology [5]. Rotamer libraries classically catalog favored sidechain conformations by the mean  $\chi$  values and standard deviations for each rotamer. They are created by performing statistical analysis on a selected dataset of experimentally-determined models, usually crystal structures archived in the Protein Data Bank (PDB) [6]. Along with Ramachandran backbone  $\phi$ ,  $\psi$  analysis [7, 8], these libraries form the conformational criteria used in a variety of applications including crystallographic model building and refinement [9, 10, 11, 12, 13], protein structure prediction and design [14, 15, 16], and protein model validation [17, 18, 19].

Different rotamer libraries represent the allowed variability around each central conformation in one of three ways. Early rotamer libraries simply provided the mean value and some estimate of allowable range for all  $\chi$  angles (or, often, just for  $\chi_1$  and  $\chi_2$ ) of each identified rotamer for each side-chain type [4, 20, 21]. A user would either simply use the mean-value conformation, or else would optimize it manually or computationally within the allowable range. Because modeling the lowest-energy conformation fails to capture allowed

---

<sup>1</sup>The p, t, m nomenclature was adopted in [1] and in MolProbity [2], to give a single letter for use in rotamer strings, and because the more common g+, t, g- terminology was, in 2000, still being assigned to opposite conformations (see discussion in [1]).

variation and further minimization is computationally expensive, design methods expanded to employ "grid" libraries of arbitrarily-spaced, discrete sample points in  $\chi$  space around the low-energy mean of each rotamer, which allowed development of the influential Dead-End Elimination method in protein design [22, 23]. A third type of rotamer software [24, 19] evaluates a given sidechain conformation by its position in a multi-dimensional probability distribution. Early such distributions were binned (often at  $\geq 10^\circ$ ) [17], but recent ones use smooth contour surfaces, scored by what percent of the reference data lies outside that contour [1, 25]. Design libraries and validation libraries focus on two distinct areas of the reference data distributions. While design and prediction are primarily concerned with statistics and cluster shapes inside the low-energy wells [24], validation is primarily concerned with robustly identifying the outliers beyond the edges of those wells. Such outliers are usually wrong but sometimes valid and interesting, so they provide a short list always worth examining [Richardson 2013 COSB validation].

Because rotamer libraries or distributions are an integral component of modern structural biology, it is imperative that they provide only authentic, low-energy sidechain conformations so that errors are not propagated. Such errors of circular reasoning can be documented [1] for many early cases, such as high-energy eclipsed  $\chi$  angles added for "completeness", or real empirical data clusters caused by incorrect backward-fit branched sidechains. The accuracy of these libraries depends on including only reliably modeled sidechains in the reference dataset. All rotamer libraries filter their datasets at the file level by resolution and redundancy. As the PDB grew in size, it became increasingly practical to filter also at the residue level. This process very effectively lowers noise and sharpens clustering, because even at high resolution the poorly ordered regions are susceptible to misfitting and are often worse than the good parts at low resolution.

Previously our group developed the "penultimate" rotamer library, using the quality-filtered Top240 PDBs [1]. Soon afterward we updated the library using our Top500 dataset [26]. This library used many file-level filters such as requiring  $\leq 1.8$  resolution, a clashscore

< 30 [27], and few backbone bond-angle outliers from Engh and Huber standards [28], and when the file contained multiple identical changes, used only the best-ordered one. The residue-level filters required a B factor < 40 for all atoms in the residue and the absence of serious steric clashes, defined as hydrogen-aware atomic overlaps  $\geq 0.4 \text{ \AA}$  [27]. These filters were used in order to eliminate residues with questionable justification for their given conformation, thereby increasing reliability. Specifically, the B-factor filter (the best metric available before mandatory data deposition) was meant to eliminate residues with poor electron density or other local uncertainty. However, as we realized and as Shapovalov and Dunbrack later clearly demonstrated [29], in many structures the B factors are a poor indicator of good density, primarily because they are often restrained to change only modestly between adjacent, covalently bonded atoms. Since 2008, when the wwPDB started requiring structure-factor data with all depositions, it has become feasible to do routine analysis of electron density directly.

In the current work we use a new quality-filtered dataset curated by our lab, the Top8000, to develop a new rotamer library. The main differences between this dataset and our previous dataset are sheer size (8000 vs. 500 chains) and improved quality filters, which are stricter and better balanced at the overall file (or chain) level and especially at the residue level. We now use a three-component residue filter that adds real-space correlation coefficient (RSCC) and local map value to B factor, effectively eliminating all residues with poor electron density. We explain how these strict filters reveal authentic, rare, and interesting rotameric states, including cases where bond angles must open up significantly. The improved statistics enable a three-level rotamer classification for validation of *favored*, *allowed*, and *outlier* (analogous to classic Ramachandran measures), where now only 0.3%, rather than the previous 1.0%, of the high-quality, filtered reference data lies outside both the *allowed* and the *favored* regions.

## 3 Methods

### 3.1 Chain-level Dataset Filters

Our previous rotamer library [1] used the Top500 quality-filtered data [26] as its reference dataset. Since then the number of high-resolution structures in the PDB has skyrocketed, allowing us to create a new quality-filtered, X-ray model database, the Top8000. The Top8000 was curated by assessing all PDB crystal structures as of March 29, 2011 with a protein chain of  $\geq 38$  residues at  $< 2.0 \text{ \AA}$  resolution. Hydrogens were added to each PDB file using Reduce, including Asn, Gln, and His flip corrections [30]. MolProbity analysis was performed on each chain [19] and the results entered into a MySQL database. The chains were additionally filtered on the following criteria: chain MolProbity score  $< 2.0$ ,  $\leq 5\%$  of residues with bond length outliers ( $> 4\sigma$ ),  $\leq 5\%$  of residues with bond angle outliers ( $> 4\sigma$ ), and  $\leq 5\%$  of residues with C $\beta$  deviation outliers ( $> 0.25\text{\AA}$ ).

In order to control redundancy in the dataset, we made use of the PDB homology clusters, taken separately for 50% sequence identity (most stringent similarity filtering), 70%, 90%, and 95%. For each homology cluster, we selected the best chain based on the average of resolution and chain MolProbity score. This scoring scheme produced ties within some clusters (for  $< 1\%$  of the final chain tallies); these were resolved, arbitrarily but reproducibly, by alphabetical order of PDB ID + single-character chain ID. At each homology level, a second list was chosen with the additional requirement of deposited structure-factor data. All 8 chain lists are available on GitHub [31] (e.g. file `Top8000-SFbest_hom70_pdb_chain_list.csv`, at [http://github.com/rlabduke/reference\\_data](http://github.com/rlabduke/reference_data)). The "SF" lists are somewhat smaller, since some clusters may include no otherwise-acceptable structures with deposited data. At the 70% similarity level optimal for most data-mining purposes, the Top8000-SFbest\_hom70 list contains 7,419 chains and the Top8000-best\_hom70 list contains 7,957 chains. This entire set of datasets is therefore nicknamed the "Top8000" of structural biology's upper class.

All rotamer statistical analysis presented here used the Top8000-SFbest\_hom70 dataset

version. Deposited structure factors not only are required for electron-density based residue filtering, but they also enable examination of the model in the density map for individual examples of conformations deemed dubious or interesting. Of the 7,419 chains, 28 failed in the RSCC calculation described below, and 175 were dropped because  $\leq 20\%$  of their residues remained after residue-level filtering. The final rotamer dataset of 7,216 chains is listed in the supplementary material.

### 3.2 Residue-level Dataset Filters

Although the chain-level filters select for structures with high overall quality, local quality varies considerably within any model. In this context, *quality* refers to canonical macromolecular model validation criteria (sterics, geometry, and conformation) as well as the fit to electron density. An important aspect of this research is to include only amino-acid sidechains with sufficiently clear electron density to justify their given conformation. Our previous rotamer library was based on the Top500 dataset where only two residue-level filters were used: no clashes and no B factors  $\geq 40 \text{ \AA}^2$ . The B factor filter was then (in 2003) our best available proxy for electron-density quality; however, residues with dubious sidechain density still remained. To remedy this, the current dataset was also filtered by a local real-space correlation coefficient (RSCC) metric [32] and by local map value.

$$RSCC = \frac{\sum_{i=1}^n (o_i - \bar{o})(c_i - \bar{c})}{\sqrt{\sum_{i=1}^n (o_i - \bar{o})^2 * \sum_{i=1}^n (c_i - \bar{c})^2}} \quad (1)$$

The local RSCC looks at the correlation between the  $\sigma_A$ -weighted  $2mF_o$ -DF<sub>c</sub> and the F<sub>c</sub> (calculated from model) electron density maps for a given local region, in this case, around each individual atom. Equation 1 shows how the RSCC was calculated, as implemented in ***PHENIX***.  $o$  and  $c$  are  $\sigma$  values at grid points in the  $2mF_o$ -DF<sub>c</sub> and the F<sub>c</sub> maps, respectively, in a radius around the atom. The radius used is resolution dependent; a 1Å radius is used for resolutions below 1Å and a 1.5Å radius is used for resolutions between 1Å and 2Å.  $n$

is the number of grid points in the selected region and  $\bar{o}$  and  $\bar{c}$  are the local mean values for  $o$  and  $c$ , respectively. The RSCC alone is not an adequate density-fit metric for our purposes. However, local RSCC (for shape), local  $2mF_o$ - $DF_c$   $\sigma$  value at the atom coordinate (for height), and B factor (for spread) together create a satisfactory fit-to-density metric. We used *phenix.real\_space\_correlation* in the *PHENIX* software package [11] to calculate the RSCC and  $2mF_o$ - $DF_c$   $\sigma$  values at each atom for all structures in the Top8000. Per-residue filter values were then assigned by taking the worst atom B (greatest value), worst atom RSCC (least value), and worst atom  $2mF_o$ - $DF_c$   $\sigma$  value (least value) in each residue.

In selecting filter thresholds for the dataset, we had the goal of keeping a large number of residues while reliably eliminating those with dubious density. To do this, we analyzed the counts of residues remaining for several combinations of filter thresholds. The best balance between filter levels and a reasonably large number of residues yielded the following thresholds: worst B factor  $< 40$ , worst correlation coefficient  $> 0.7$ , and worst map value  $> 1.1$ . Upon visual inspection of residues and maps close to these filter thresholds, we determined that the selected combination of thresholds was indeed effective at keeping only residues with satisfactory electron density. Additionally, residues were required to have no all-atom clashes, an occupancy of 1.0, B factors  $> 1.0$  and all backbone atoms modeled, achieved indirectly by ensuring that  $\phi$ ,  $\psi$ ,  $\omega$  and  $\tau$  were defined for each residue. The latter test also drops sidechains of the first and last residue in each chain, which are somewhat less and differently affected by backbone interactions.

After all filters, the final reference dataset for this work contains more than a million residues, 983,574 of which are rotamer-relevant, non-Gly non-Ala residues. A csv file of all 983,574 residues in the rotamer-relevant reference dataset is available on GitHub (See `datasets/Top8000_rotamer_residues.csv` at [http://github.com/rlabduke/reference\\_data](http://github.com/rlabduke/reference_data)).

### 3.3 Determination of Distributions, Scores, and Contours

The described rotamer library’s primary use is for validation of sidechains in protein models, which flags an outlier if the sidechain has an extremely rare (and presumably high-energy) conformation. This assessment requires a scoring system based upon the multi-dimensional distribution of observed conformations in our high-quality, residue-filtered reference data. We have taken great care to ensure smooth, accurate, and robust contours dividing outlier from allowed. To achieve this we calculated smooth distributions in the multi-dimensional  $\chi$  space for each residue type, using an adaptive local-density-dependent kernel density estimation (KDE) [33]. Our method has two steps, both using a cosine kernel normalized to have an area or volume of 1.0. A cosine is used, rather than a gaussian, because it reaches zero at a well-defined edge. In the first step, the width of each cosine kernel is  $5^\circ$ . In the second step, the width of the kernel is varied, dependent on the density at that location, as calculated in the first step. Kernel widths are wider in sparse regions and narrower as density increases. The consequence of this is distributions that remain smooth in sparse regions but preserve the sharp transitions where occurrence frequency falls off quickly. A full explanation of this method can be found in [26]. The distributions are stored as a discrete grid in  $\chi$  space with coordinates and KDE values. Grid spacing is dependent on dimension number: in order of  $\chi$  dimensions, 1-4, the grid spacings are 1, 5, 8, and 10 degrees. Each datapoint residue can be assigned a value by interpolating its  $\chi$  values within the grid. One can then determine what grid value is just greater than for the lowest, say, 1% of the quality-filtered reference data. The grid values are then rescaled to represent those percentage values, which are known as the *rotamericity* of a sidechain conformation, or simply as the rotamer score. The rotamer-score grids are used in *PHENIX* validation (GUI, `phenix.rotalyze`, and `phenix.molprobity`) and the MolProbity web service, and they are available as plain-text numerical arrays on GitHub, under `datasets/Top8000_rotamer_pct_contour_grids` at [http://github.com/rlabduke/reference\\_data](http://github.com/rlabduke/reference_data).

To visualize the rotamer distributions, smooth contours are drawn at chosen levels using

our internal programs Silk, kin2Dcont, and kin3Dcont [34, 35]. For the 4-dimensional Lys and Arg cases, the 3-D plots for  $\chi_{1\mathbf{m}}$ ,  $\chi_{1\mathbf{t}}$ , and  $\chi_{1\mathbf{p}}$  are shown separately. For validation purposes, the interpolated grid value of a given multi- $\chi$  conformation is its rotamer score. In the new system developed here, a score of  $\leq 0.3\%$  qualifies as a rotamer outlier – its score is worse than 99.7% of the good data. Scores  $> 0.3\%$  and  $< 2.0\%$  are considered *allowed* while those  $\geq 2.0\%$  are considered *favored*, as is traditional for Ramachandran criteria [17, 26].

### 3.4 $\chi$ and Covalent Bond Angle Statistics

Maximum rotamer ranges were defined manually for each  $\chi$  dimension by inspecting the smoothed contours (See 3.3) and placing boundaries at saddle points between rotamers or to reasonably encompass the rotamer well. To avoid wrapping complications, circular statistics were used (See Equations 2 and 3); this is important for  $\chi$  distributions that cross zero, such as Asp **p0** and **m-30**, or for rotamers near  $90^\circ$  in symmetric aromatics. For each rotamer, we report a central value and a standard deviation ( $\sigma$ ), for both dihedral  $\chi$  values and covalent bond angles. The mean for the bond angles (Equation 4) and the  $\sigma$  for both  $\chi$  and bond angles (Equation 3) are calculated straightforwardly from the measures of each example in the filtered Top8000. Because of complex shapes, the central value for the  $\chi$  measures is the center-of-mass (COM) in the contoured data. This is found by collecting the stored contour grid points in each rotamer well and, for each dimension, calculating the COM (Equation 5).

$$\bar{x} = \frac{\sum_{i=1}^n (\sin \theta_i)}{n}, \bar{y} = \frac{\sum_{i=1}^n (\cos \theta_i)}{n} \quad (2)$$

$$\sigma = \sqrt{-2 \ln \left( \sqrt{\bar{x}^2 + \bar{y}^2} \right)} \quad (3)$$

$$\text{mean} = \arctan \left( \frac{\bar{x}}{\bar{y}} \right) \quad (4)$$

$$\text{COM} = \arctan \left( \frac{\frac{\sum_{i=1}^n (\sin \theta_i) * w_i}{W}}{\frac{\sum_{i=1}^n (\cos \theta_i) * w_i}{W}} \right) \quad (5)$$

where  $n$  is the number of grid points,  $\theta_i$  is the  $i^{th}$   $\chi$  coordinate in radians,  $w_i$  is the KDE value at the  $i^{th}$  coordinate and  $W$  is the sum of all KDE values in the rotamer well. To avoid wrapping issues when working with angular data, circular statistics breaks angle measures into its two unit circle components,  $x$  and  $y$ . This is part of what is being done in Equation 2, where  $\bar{x}$  and  $\bar{y}$  are the average  $x$  and  $y$  components, respectively, for all angular measures,  $\theta$ . In Equation 5, a weighted average of the  $x$  and  $y$  components is being calculated. The arctan function simply takes the  $\bar{x}$  and  $\bar{y}$  components and returns the corresponding angular value.

### 3.5 Rotamer Assignment and Score

For the purpose of setting the appropriate contour level for the outlier cutoff, a large test dataset was created from all PDBs in the Top8000 by including all chains and all residues in each PDB file (as a surrogate for entire new PDB files later being validated), hereafter called the unfiltered dataset. *phenix.rotalyze* was used to assign rotamer names and scores for each residue in the unfiltered dataset. The score is calculated from the individual residue's  $\chi$  values, by interpolating over the nearest contour grid points (see section 3.3). This was done twice, once using the new Top8000 contours and once using our previous Top500 contours. If the score is above the outlier threshold, then a rotamer name is assigned based on which rotamer well the given  $\chi$  angles fall within (the maximum well ranges were described in Section 3.4). Otherwise the side-chain conformation is classified as an *outlier*.

## 4 Results

In the Top8000 system there are a total of 235 named rotamers which reach above the 0.3% outlier cutoff and include  $\geq 4$  datapoints. Many of the central values have shifted somewhat from the Top500 system, both because of much more and higher-quality data and also because of the new center-of-mass definition (see section 3.4), which improves on both the modal and the common-atom values defined previously [1]. MolProbity’s “ultimate” rotamers and their parameters are given in two tables: Table S2 describes the frequency and count of each rotamer in the given residue type and Table S3 has central dihedral and covalent angle values for each rotamer. The latter numbers are available in CSV format on GitHub under `datasets/Top8000_rotamer_central_values` at [http://github.com/rabduke/BJH\\_Dissertation](http://github.com/rabduke/BJH_Dissertation). The more useful multi-dimensional distributions are also on GitHub, as linked in section 3.3.

The order-of-magnitude increase in size and quality of our reference dataset greatly improves signal-to-noise, allowing us to do a better job of the same functions as before. More importantly, it enables new features and new conclusions. The simplest view of a rotamer is of a favored sidechain conformation described by the list of dihedral angles, with ideal staggered  $\chi$  values between  $sp^3$  atoms and little preference for  $\chi$  between  $sp^3$  and  $sp^2$  atoms. The next level of detail adds a standard deviation for each  $\chi$ . However, rotamers are really much more complex, especially in a macromolecular interior where numerous interactions can compensate for the energy required to depart from ideal. We have high confidence in the reliability of our more complex and nuanced conformational distributions, as we took great care to include only physically plausible sidechains that fit clear electron density (see Section 3.2). As Figure 1 demonstrates, sidechain conformations do cluster in discrete regions of  $\chi$  space. Further, there is large variance in  $\chi$  values within a rotamer cluster, the shape is usually not axis-oriented, and its center is rarely at the nominally ideal staggered conformation. A rotamer is better thought of as a local energy well, with a potentially complex shape, describing the favorable extent around an allowed sidechain conformation.

## 4.1 Residue Filter Effects

Our strict residue-level filtering aims to include essentially only conformations with unassailable reliability, at some expense to the overall numbers. The fraction of residues kept after filtering differed between amino-acid types (Table S1). Residue-level filtering eliminated the most data, in order, for lysine, glutamate, and arginine. This makes sense since long sidechains are more susceptible to dynamics that can blur or eliminate density. Also, charged sidechains are usually found on the molecular surface interacting with mobile solvent. Hence, surface-exposed positions such as these, lacking strong interactions that can hold them in a disfavored conformation, have no need and no ability to adopt rare, unfavored rotamers. If such sidechains are modeled, they should be in one, or more often several, of the common rotamers.

## 4.2 Rotamer Evaluation

The primary purpose of this rotamer library is to analyze the rotamericity of protein sidechain conformations and robustly distinguish outliers. Rotamericity here refers to where a given conformation lies in  $\chi$  space relative to the calculated contours (see Section 3.3). This evaluation gives a rotamer score between 0 and 100, corresponding to the percentage of the high-quality reference data that lies outside that contour (i.e., how much of the reference data's scores are worse than the given residue). In addition, for scores above the outlier cutoff, the evaluation also assigns a rotamer name. Each local minimum, or cluster, in  $\chi$  space is given a name derived from the central  $\chi$  values of that cluster. Values for  $\chi$ s between  $sp^3$  hybridized atoms are named **m**, **p**, or **t** roughly corresponding to the staggered values  $-60^\circ$ ,  $+60^\circ$  and  $180^\circ$ .  $\chi$ s between  $sp^3$  and  $sp^2$  hybridized atoms are assigned a number which is the COM value (Section 3.4) of the given  $\chi$ , rounded off. For instance, a Gln rotamer with COM at  $\chi_1 = -174$ ,  $\chi_2 = -82$ , and  $\chi_3 = -22$  is named **tm-20**.

For validation purposes, users are often most interested in the binary issue of whether a sidechain conformation is rotameric or an outlier, i.e. falls within or outside the rotamer

distribution. With the Top500 data, contours were smooth and reproducible (between different datapoint selections) only out to the 1% level. Due to greater numbers and increased reliability in the filtered Top8000 distributions, a lower outlier cutoff is now feasible as well as desirable, since an even lower percentage of the reference datapoints are now dubious. We also preferred to keep outlier numbers for unfiltered, general data roughly equal between the previous and new systems. Therefore a count of the 1% outliers by residue type in the unfiltered dataset was performed using the Top500 contours, and those numbers were compared to outlier counts at several different Top8000 contour cutoffs. A new cutoff of 0.3% matched best, was found to behave smoothly, and happens also to match the  $3\sigma$  level for a normal distribution (3 out of 1000). Further, along with favored and outlier, a new category has now been added: *allowed*, for scores  $\geq 0.3\%$  and  $< 2.0\%$ . This new category lets users know if a given conformation is at the edge of the given rotamer’s distribution, close to the outlier region. This change matches the 3-part system, and the division at 2.0%, long judged useful for Ramachandran criteria [36, 18, 26].

An important question is how the new system will change rotamer analysis, specifically the number and identity of outliers. A rotamer outlier is simply a conformation that lies outside the outlier contours of the reference dataset, in this case either the Top500 or the Top8000. For residues that have low-energy sidechain conformations, nothing will change. However, a sidechain conformer in a strained position near the outlier edge could change status between the Top500 and Top8000 analyses. Figures 2 and S1 show Top500 and Top8000 allowed regions superimposed for all residues with two  $\chi$ s. For the most part, the contours overlap, but with some differences near the edges. Asp shows very little change, Trp has a significantly larger allowed region now than in the Top500 system, and the Ile Top8000 contours have tightened up relative to the Top500. Thus, differences are only in low-population regions of  $\chi$  space, and overall outlier counts will be nearly identical.

### 4.3 Dihedral and Bond Angle Deviations

As we have shown, rotamers are more complex than just a collection of mean  $\chi$  values and standard deviations. The allowed regions are often large and complex, and the mean or modal  $\chi$  angles can make fairly large deviations from "ideal". Such deviations usually occur in well-packed environments where the sidechain makes close atomic interactions, either van der Waals (vdW) or a hydrogen bond, with its own backbone (or within the sidechain). The  $\chi$  deviations help allow the contact to be close rather than a steric clash.

For sidechains that have  $sp^3$  atoms out to the  $\delta$  atom (Leu, Ile, Met, Glu, Gln, and Arg), there are four  $\chi_1\chi_2$  combinations that produce close sidechain-backbone interactions; **pm** and **mp** interact with the NH while **tm** and **pp** interact with the carbonyl C. Comparing statistics on all  $\chi_1\chi_2$  combinations for these amino acids in the filtered dataset (Table 1), we see that the backbone-interacting combinations are the least populated and that they have large  $\chi$  deviations from the ideal staggered positions, especially for  $\chi_2$ .

Breaking down these statistics by residue type reveals that Glu and Gln make up 54.3% of **mp**, 59.8% of **tm**, and 92.2% of **pm** examples. The likely reason for this is that the terminal oxygens on Glu and Gln can H-bond with their own NH for **pm** and **mp**. In **tm** a strong vdW interaction is made with the C-terminal peptide plane, and the sidechain amide or carboxyl group is generally further stabilized by one, or more often multiple, H-bonds. However, as for all sidechain-backbone interacting rotamers, the dihedral  $\chi$  deviations don't tell the whole story – covalent bond angles also must deviate to avoid steric clashes.

It has long been known that certain sidechain conformations can cause large deviations in the otherwise tight covalent bond angles [1]; this is seen in the scenarios above where sidechain-backbone interactions occur. One rotamer of this type is methionine **ppp**, which has just 48 examples (0.29% of Met) in the filtered Top8000. Figure 3 shows one specific example and demonstrates how deviations in *both*  $\chi$  dihedrals and covalent bond angles are needed to avoid a clash and form a favorable non-bonded interaction. All-atom contacts demonstrate the tight vdW packing. For this residue, the angles C-C $\alpha$ -C $\beta$ , C $\alpha$ -C $\beta$ -C $\gamma$ , and

$C\beta$ - $C\gamma$ - $S\delta$  each open up from Engh & Huber values, by  $+1.6^\circ$ ,  $+0.9^\circ$ , and  $+3.0^\circ$  respectively.

## 4.4 Interesting Rotamer Anecdotes

Consistently large deviations from ideal values in the  $\chi$  and covalent bond angles for some rotamers raise two questions, which are interesting and sometimes illuminating: (1) What is it about each of these rotamers that requires such deviations? (2) What are the structural features and roles for each of these rotamers? These questions were answered by inspecting individual examples from the filtered dataset. After an interesting rotamer was identified, at least five random examples of that rotamer were selected from the filtered dataset. To visualize interactions for the residue of interest, all-atom contacts were displayed in KiNG [37]. This process revealed commonalities of the rotamer environment, reasons for the deviations, and possible reasons for a rare rotamer frequency. What follows are select observations from this process.

### 4.4.1 Glutamate pm20 & mp0

Glutamate **pm20** and **mp0** make up 2.6% and 6.4% of all glutamates in our dataset and would not be considered rare in the context of this paper. These rotamers are of interest, however, for three reasons: (1) some  $\chi$  and bond angles have relatively large deviations from ideal (Table 2), (2) despite this, these rotamers are significantly more common than other rotamers with similar deviations, and (3) the  $\chi$  distributions for both show that all three  $\chi$  measures strongly depend on one another, forming correlated datapoint clusters (Figure 4a). In both **pm20** and **mp0**, the same carboxylate oxygen makes a strong H-bond with its own NH, but they take a different path to do so (Figure 4b). The large geometry deviations are required in order to avoid a steric clash of  $C\delta$  with the backbone. Glutamine can adopt equivalent arrangements, as previously noted even from much sparser data [38].

The strong sidechain-backbone H-bond, between atoms separated by only 5 covalent bonds, is an important characteristic of these conformations. It compensates for the ener-

getic cost of distortions better than would a vdW contact, and especially it gives rise to a characteristic interdependence of the three  $\chi$  angles: as one  $\chi$  changes, the others change in predictably compensating ways to preserve the favorable H-bond.

#### 4.4.2 Isoleucine pp

With just 249 examples in the filtered Top8000, **pp** is the rarest isoleucine rotamer. This rotamer epitomizes most rare ones, by having numerous surrounding vdW interactions to pack the sidechain into this specific rotamer state, which is rare because of strain from its need for large average bond and dihedral angle distortions. The modal  $\chi_2$  value is  $+24.3^\circ$  from staggered and bond angles  $C\alpha-C\beta-C\gamma 1$  and  $C\beta-C\gamma 1-C\delta 1$  open, relative to Engh & Huber, by  $2^\circ$  and  $1^\circ$ , respectively (Table 3). In **pp** the terminal methyl is in the "down" position approximately parallel to the course of the backbone (Figure 5). In all examples examined, local structure packed the methyl into this rare conformation by sterically prohibiting the more common **pt** rotamer (methyl "up"). The resulting bond-angle and dihedral openings are necessary to minimize a steric clash between  $C\delta 1$  and the backbone C.

#### 4.4.3 Methionine mpm

Methionine **mpm** is extremely rare, with only 13 examples in the filtered Top8000. The deviations from ideal are remarkable:  $C\alpha-C\beta-C\gamma$  and  $C\beta-C\gamma-S\delta$  open by more than  $2^\circ$  and  $3^\circ$ , respectively, and the modal  $\chi_3$  in **mpm** is a mere  $20^\circ$  from being eclipsed – probably permitted by the fact that  $\chi_3$  in Met has a lower rotational barrier than all-carbon tetrahedral torsions because of the longer C-S bond [39, 40]. Here, if  $\chi_3$  were any closer to staggered, the  $\epsilon$ -methyl would clash with its backbone NH.

Out of the 13 Met **mpm** examples in the filtered Top8000, 11 are structurally very similar, with much of the surrounding sequence identical. All 11 belong to the large superfamily of subtilisin-like serine proteases, which contains six families A-F. A large multi-sequence alignment ([41]) reveals that the Met is completely conserved in families A, B, and C, absent

in D and E, and mostly present in F. The Met occurs in the first turn of a helix that is disrupted  $n+3$  to the Met by a proline conserved in the same families (Figure 6a). Most significantly, immediately preceding the Met is the Ser of the canonical Asp-His-Ser catalytic triad (Figure 6b). Figure 6a shows the extensive contact between the catalytic Ser and the **mpm** Met. Also shown is the extensive contact between the Met and Ile 246, which excludes the possibility of Met  $\chi_3$  adopting either a **p** or **t** conformation. Although the identity of residue 246 changes across the 11 structures, the constraining hydrophobic contact is maintained.

This case illustrates that 70% "homology" filtering does not always produce independent examples when a rotamer is conserved for functional reasons. In the families where the Met is conserved, it seems structurally important as part of a motif packing tightly with the catalytic Ser. However, in families where the Met is missing, no other sidechain fills in to sterically position the active Ser. Therefore the functional reason for conservation must be more subtle, at the frontier of our understanding. Experimental research will be needed to untangle its effects on catalysis.

## 4.5 Conformations Between Tetrahedral and Planar Atoms

Due to differing physical constraints, dihedral angles between tetrahedral and planar atoms do not follow the well-clustered **ptm** conformations seen between two tetrahedral atoms: when one planar  $sp^2$  branch is staggered the other is eclipsed. This is reflected in the Top8000 filtered distributions where  $sp^3$ - $sp^2$  dihedrals (always in the final  $\chi$  of the sidechain) often show only weak preferences across their total range (e.g., see Figure 1 for  $\chi_2$  of histidine).

For the aspartate and asparagine contours shown in Figure 7, sometimes  $\chi_2$  has disallowed regions and sometimes the entire range is allowed. This fact makes it even more difficult, and less meaningful, to assign a central value and standard deviation that adequately describes the allowed conformations. Asp and Asn also show clearly that there is distinct, complex fine-structure of local clustering in the data distributions, especially evident for Asn and

for the diagonally streaked shoulder at the top of Asp  $\chi_1$  t. These datapoint clusters are primarily due to patterns of sidechain H-bonding to specific backbone donors or acceptors local in sequence, or occasionally due to especially favorable vdW packing against local backbone. For example, the Asn N $\delta$ 2 can either H-bond with the i-4 CO in a regular  $\alpha$ -helix in an **m-80** conformer or pack against that CO in the more common **m-20** conformer [38].

Both these issues confirm that a complex probability density function, such as our filtered and smoothed empirical contours, represents sidechain conformational preferences much better than a library of simple box or ellipsoid shapes. Its contour outline at low probability is a definitive way to flag rotamer outliers. For protein structure prediction or design, the details of favored local motifs at high contour levels should also be considered, preferably with the addition of information both about  $\phi,\psi$  values [14] and local secondary structure [1].

## 5 Discussion

### 5.1 In what sense is this ultimate?

Calling this set of rotamer-library distributions "ultimate" is a claim that requires both explanation and justification, beyond the mere fact that it follows after our "penultimate" rotamer library.

First, we claim it as ultimate for the rotamer aspect of model validation (where outlier status is pre-eminent) – such as done by the MolProbity website, in *PHENIX*, or by the wwPDB – but not necessarily for other purposes (such as protein design) that depend on accurate comparisons within the regions of high probability rather than on accurate delineation of the outer limits of those regions.

Second, changes from the penultimate distributions are fairly minor in quantitative terms; their importance lies in the new ability to support a 3-level evaluation, to identify very rare rotamers, and to specify rotamer-dependent bond-angle deviations. With a million residues

of high-reliability data, the sampling is now adequate for this task. The results are unlikely to change much in the future, except for some special cases such as disulfides or chemically modified amino acids. The multi-dimensional distributions of rotamer probability are freely available on GitHub for use in other software systems.

Third, what we do expect will improve in the future is a change in how sidechain modeling and validation is defined, driven both by expansion of data and compute power and by the needs of structure determination at lower resolution. Rather than separate Ramachandran and rotamer evaluation, we should move toward analyzing all backbone and sidechain torsional dimensions together [8], including allowance for the influence of secondary structure and local motifs. Thus our "ultimate" claim asserts the position that rotamer evaluation should no longer be updated but should evolve into something else.

## 5.2 Use of the Top8000 rotamer distributions in model building

Although the Top8000 rotamer distributions were created specifically for MolProbity's validation of sidechain conformations, they also can be used in model building for crystallography, design, homology, etc., as the Top500 library has been [42, 10, 43, 44, 11, 23]. To make model building computationally tractable, sidechains are initially fit as discrete conformations at rotamer values, then either minimized or tested at neighboring sample points using  $\sigma$  to inform the allowable range for each  $\chi$ . However, most rotamer distributions are far from normal and often have quite complex shapes, as can be seen in Figures 7 and 1. Thus the approach of using a central value and  $\sigma$  cannot adequately describe allowed sidechain conformations.

The contour values in our rotamer distributions correspond to the percentage of high-quality reference data that lies outside that contour. The consequence of this treatment is that the contour data represent a probability density function in  $\chi$  space describing sidechain conformations. We recommend using those contours to delimit sampling or minimization in  $\chi$  space, with the added benefit that the contour levels provide prior probabilities for each

position. If a contour level higher than 0.3% is chosen for this purpose, then rotamers which do not reach the chosen percentage level should be omitted. Depending on details of the algorithm used, it may or may not be beneficial to model significant rotamer-dependent bond-angle deviations.

### 5.3 I have a rotamer outlier; is it wrong?

An accurate macromolecular model gives unparalleled mechanistic knowledge at the single-molecule level. Many such models, of different cellular components, combined with structural perturbations such as point mutations and ligand binding, provide mechanistic detail on the cellular level. These methods not only provide interesting knowledge, they allow understanding crucial to the treatment of disease. The utility of a model in this process is correlated with how accurate it is. As such, structural biologists want to build models that come as close as possible to representing a valid state of the actual macromolecule. Unfortunately, the experimental data alone seldom provide all information needed to build an adequate model. Fortunately, empirical knowledge of chemical and macromolecular structure helps greatly, and in crystallography much of this knowledge (e.g. bond lengths and angles, chirality) is already part of the automated software. Structure validation such as MolProbity provides even more empirical knowledge, often highlighting errors that refinement was unable to fix.

Sometimes we get disturbing reports that people are trying to achieve a sort of "MolProbity Nirvana" by attempting to eliminate every single outlier. To approach eliminating all clashes is a worthy goal: in principle the clash target is zero, as sterics do not allow serious atomic overlaps. However, at high resolution a zero score would require the difficult reconciliation of occupancies for alternate conformations including waters, and in any case a few puzzles often remain. Another important point to keep in mind is that the primary purpose and usefulness of model validation is to help diagnose and correct places where the conformation has been fit in the wrong local energy well. Such corrections often matter to biological interpretation, and will be stable to further refinement. In contrast, a small shift

across the border into the allowed region will often shift right back again, and is not a very meaningful improvement.

However, the target is very definitely *not* zero for torsion validation. The idea that *every single* Ramachandran and rotamer conformation needs to be allowed is a misunderstanding. As outlined here, validation of torsion angles relies on contoured conformational distributions from quality-filtered datasets. The outer contour defines an outlier cutoff, meaning that even the filtered reference data has outliers but *not* that those conformations are wrong. There is room for valid torsion outliers, more so for sidechain rotamers than for backbone Ramachandran values. Even in 2005 the curve of rotamer-outlier percentage, fit as a function of resolution, was found to asymptote at 0.5% [9]. For quality-filtered data the rotamer outlier cutoff is now 0.3%, meaning that 3 out of 1,000 residues are expected to be valid rotamer outliers. For Ramachandran the outlier cutoff is 0.05%, so only a very low 5 out of 10,000 residues – but not zero – are expected to be valid Ramachandran outliers. In either case, to be accepted as valid, an outlier should have clear electron density to support its occurrence, and should have either H-bonds or tight packing to hold it in the presumably quite strained conformation. This is exemplified by the Asn rotamer outliers in Figure 8, which each have an eclipsed  $\chi_1$  but are validated by excellent density and three sidechain H-bonds. Valid outliers are more likely to have been selected and maintained by evolution if there is a functional need of some sort (folding, catalysis, binding) for that specific conformation; therefore such cases will probably be of great interest and will reward as well as require detailed examination.

## 6 References

### References

- [1] Simon C Lovell, J Michael Word, Jane S Richardson, and David C Richardson. The penultimate rotamer library. *Proteins: Structure, Function, and Genetics*, 40(3):389–408, 2000.
- [2] I. W. Davis, L. W. Murray, J. S. Richardson, and D. C. Richardson. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research*, 32(Web Server):W615–W619, jul 2004.
- [3] Henry Eyring. Steric hindrance and collision diameters. *J. Am. Chem. Soc.*, 54(8):3191–3203, Aug 1932.
- [4] Jay W. Ponder and Frederic M. Richards. Tertiary templates for proteins. *Journal of Molecular Biology*, 193(4):775–791, Feb 1987.
- [5] Roland L Dunbrack. Rotamer libraries in the 21st century. *Current Opinion in Structural Biology*, 12(4):431 – 440, 2002.
- [6] H. M. Berman. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, Jan 2000.
- [7] G N Ramachandran, C Ramakrishnan, and V Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–9, July 1963.
- [8] Jane S Richardson, Daniel A Keedy, and David C Richardson. “The Plot” thickens: more data, more dimensions, more uses. In M. Bansal and N. Srinivasan, editors, *Biomolecular Forms and Functions: A Celebration of 50 Years of the Ramachandran Map*, IISc Press-WSPC publication, pages 46–61. World Scientific Publishing Company Incorporated, 2013.
- [9] W. Bryan Arendall, Wolfram Tempel, Jane S. Richardson, Weihong Zhou, Shuren Wang, Ian W. Davis, Zhi-Jie Liu, John P. Rose, W. Michael Carson, Ming Luo, David C. Richardson, and Bi-Cheng Wang. A test of enhancing model accuracy in high-throughput crystallography. *J Struct Funct Genomics*, 6(1):1–11, mar 2005.
- [10] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and development of Coot. *Acta Crystallographica Section D*, 66(4):486–501, Apr 2010.
- [11] Paul D. Adams, Pavel V. Afonine, Gábor Bunkóczki, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, Li-Wei Hung, Gary J. Kapral, Ralf W. Grosse-Kunstleve, Airlie J. McCoy, Nigel W. Moriarty, Robert Oeffner, Randy J. Read, David C. Richardson, Jane S. Richardson, Thomas C. Terwilliger, and Peter H. Zwart. Phenix: a comprehensive python-based system for macromolecular structure solution. *Acta Cryst Sect D*, 66(2):213–221, Jan 2010.

- [12] Martyn D. Winn, Charles C. Ballard, Kevin D. Cowtan, Eleanor J. Dodson, Paul Emsley, Phil R. Evans, Ronan M. Keegan, Eugene B. Krissinel, Andrew G. W. Leslie, Airlie McCoy, Stuart J. McNicholas, Garib N. Murshudov, Navraj S. Pannu, Elizabeth A. Potterton, Harold R. Powell, Randy J. Read, Alexei Vagin, and Keith S. Wilson. Overview of the CCP 4 suite and current developments. *Acta Crystallogr D Biol Cryst*, 67(4):235–242, mar 2011.
- [13] R. P. Joosten, K. Joosten, S. X. Cohen, G. Vriend, and A. Perrakis. Automatic rebuilding and optimization of crystallographic structures in the protein data bank. *Bioinformatics*, 27(24):3392–3398, oct 2011.
- [14] Michael J. Bower, Fred E. Cohen, and Roland L. Dunbrack. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *Journal of Molecular Biology*, 267(5):1268–1282, apr 1997.
- [15] B. Kuhlman. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, Nov 2003.
- [16] Pablo Gainza, Kyle E. Roberts, and Bruce R. Donald. Protein design using continuous rotamers. *PLoS Computational Biology*, 8(1):e1002335, Jan 2012.
- [17] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2):283–291, Apr 1993.
- [18] R W Hooft, G Vriend, C Sander, and E E Abola. Errors in protein structures. *Nature*, 381(6580):272, May 1996.
- [19] Vincent B. Chen, W. Bryan Arendall, Jeffrey J. Headd, Daniel A. Keedy, Robert M. Immormino, Gary J. Kapral, Laura W. Murray, Jane S. Richardson, and David C. Richardson. MolProbity : all-atom structure validation for macromolecular crystallography. *Acta Cryst Sect D*, 66(1):12–21, Dec 2009.
- [20] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side chain conformations. *Journal of Biomolecular Structure and Dynamics*, 8(6):1267–1289, jun 1991.
- [21] Hannelore Schrauber, Frank Eisenhaber, and Patrick Argos. Rotamers: To be or not to be? *Journal of Molecular Biology*, 230(2):592–612, mar 1993.
- [22] M De Maeyer, J Desmet, and I Lasters. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Folding & design*, 2(1):53–66, January 1997.
- [23] Pablo Gainza, Kyle E. Roberts, Ivelin Georgiev, Ryan H. Lilien, Daniel A. Keedy, Cheng-Yu Chen, Faisal Reza, Amy C. Anderson, David C. Richardson, Jane S. Richardson, and Bruce R. Donald. *OSPREY: Protein Design with Ensembles, Flexibility, and Provable Algorithms*. Academic Press, 2013.

- [24] Roland L. Dunbrack and Fred E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, 6(8):1661–1681, aug 1997.
- [25] Randy J. Read, Paul D. Adams, W. Bryan Arendall, Axel T. Brunger, Paul Emsley, Robbie P. Joosten, Gerard J. Kleywegt, Eugene B. Krissinel, Thomas Lütteke, Zbyszek Otwinowski, Anastassis Perrakis, Jane S. Richardson, William H. Sheffler, Janet L. Smith, Ian J. Tickle, Gert Vriend, and Peter H. Zwart. A new generation of crystallographic validation tools for the protein data bank. *Structure*, 19(10):1395–1412, Oct 2011.
- [26] Simon C. Lovell, Ian W. Davis, W. Bryan Arendall, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, and David C. Richardson. Structure validation by  $C\alpha$  geometry:  $\phi$ ,  $\psi$  and  $C\beta$  deviation. *Proteins*, 50(3):437–450, Jan 2003.
- [27] J.Michael Word, Simon C. Lovell, Thomas H. LaBean, Hope C. Taylor, Michael E. Zalis, Brent K. Presley, Jane S. Richardson, and David C. Richardson. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *Journal of Molecular Biology*, 285(4):1711–1733, jan 1999.
- [28] R. A. Engh and R. Huber. Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Cryst Sect A*, 47(4):392–400, Jul 1991.
- [29] Maxim V. Shapovalov and Roland L. Dunbrack. Statistical and conformational analysis of the electron density of protein side chains. *Proteins: Structure, Function, and Bioinformatics*, 66(2):279–303, 2007.
- [30] J.Michael Word, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology*, 285(4):1735–1747, Jan 1999.
- [31] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. Social coding in github: Transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW ’12, pages 1277–1286, New York, NY, USA, 2012. ACM.
- [32] Gerard J. Kleywegt, Mark R. Harris, Jin-yu Zou, Thomas C. Taylor, Anders Wählby, and T. Alwyn Jones. The Uppsala Electron-Density Server. *Acta Crystallographica Section D*, 60(12 Part 1):2240–2249, Dec 2004.
- [33] Leo Breiman, William Meisel, and Edward Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, may 1977.
- [34] J. Michael Word. *All-Atom Small-Probe Contact Surface Analysis*. PhD thesis, Duke University, 2000.
- [35] Ian W. Davis. *Local Motion and Local Accuracy in Protein Backbone*. PhD thesis, Duke University, 2006.

- [36] Gerard J Kleywegt and T Alwyn Jones. Phi/psi-chology: Ramachandran revisited. *Structure*, 4(12):1395–1400, dec 1996.
- [37] Vincent B. Chen, Ian W. Davis, and David C. Richardson. KiNG (Kinemage, Next Generation): A versatile interactive molecular and scientific visualization program. *Protein Science*, 18(11):2403–2409, nov 2009.
- [38] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine rotamers: B-factor cutoff and correction of amide flips yield distinct clustering. *Proceedings of the National Academy of Sciences*, 96(2):400–405, jan 1999.
- [39] Glenn L. Butterfoss and Jan Hermans. Boltzmann-type distribution of side-chain conformation in proteins. *Protein Sci.*, 12(12):2719–2731, dec 2003.
- [40] Glenn L. Butterfoss, Jane S. Richardson, and Jan Hermans. Protein imperfections: separating intrinsic from extrinsic variation of torsion angles. *Acta Cryst Sect D*, 61(1):88–98, dec 2004.
- [41] Roland J. Siezen and Jack A. M. Leunissen. Subtilases: The superfamily of subtilisin-like serine proteases. *Protein Science*, 6(3):501–523, mar 1997.
- [42] Gerrit Langer, Serge X Cohen, Victor S Lamzin, and Anastassis Perrakis. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols*, 3(7):1171–9, January 2008.
- [43] Jeffrey J. Headd, Robert M. Immormino, Daniel A. Keedy, Paul Emsley, David C. Richardson, and Jane S. Richardson. Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place. *J Struct Funct Genomics*, 10(1):83–93, nov 2008.
- [44] Thomas C Terwilliger, Ralf W Grosse-Kunstleve, Pavel V Afonine, Nigel W Moriarty, Peter H Zwart, Li Wei Hung, Randy J Read, and Paul D Adams. Iterative model building, structure refinement and density modification with the phenix autobuild wizard. *Acta crystallographica. Section D, Biological crystallography*, 64(Pt 1):61–9, January 2008.

## 7 Figure Legends

**Figure 1 :**  $\chi_1$ - $\chi_2$  space for Leu, Ile, and His. The 0.3% contour is in red. Each point of the background data cloud represents a residue in the filtered Top8000 dataset. Crosshairs mark the nominally ideal staggered values between  $sp^3$  hybridized atoms, labeled as **m** ( $-60^\circ$ ), **t** ( $180^\circ$ ), and **p** ( $+60^\circ$ ).

**Figure 2 :** Areas in orange (from Top500 data) and in blue (from Top8000) fill the allowed regions for Asp, Trp, and Ile. The extensive areas in green are where the two systems both declare allowed conformations. See also Figure S1.

**Figure 3 :** Sidechain contacts for Met A 240 in 2bmo, which adopts the rare methionine rotamer **ppp**. (a) Sulfur contacts, modeled with ideal staggered dihedrals and Engh & Huber bond angles, (b) with Engh & Huber bond angles, but dihedrals as deposited, (c) for the deposited structure, and (d) all contacts for the deposited structure, showing its tight vdW packing.

**Figure 4 :** Panel (a) shows the 2.0% contour surfaces for the filtered Top8000 along with data points, as projected onto the  $\chi_2$ - $\chi_3$  plane for rotamers Glu **pm20** (orange) and **mp0** (blue). (b) shows how **pm20** and **mp0** both make a good H-bond (green dots) with the adjacent backbone NH, albeit through different conformations.

**Figure 5 :** Two different examples of Ile **pp**, both demonstrating the close C $\delta$ 1/C contact, and the extensive vdW interactions that prevent the more common **pt** conformation. (a) 1s99 Ile A 76, (b) 2wvx Ile C 235.

**Figure 6 :** 3D43 chain B, one of the 11 subtilisins containing a Met **mpm**. (a) The local arrangement of Ser250 (red) and Met251 (blue) on the helix, with the interacting Pro254 and Ile246; (b) rotated to show the active site, with its canonical catalytic triad in red and Met251 in blue.

**Figure 7 :** Filtered Top8000 Asp and Asn datapoints and outlier contours (red).  $\chi_2$  does not follow the **ptm** convention since it is an  $sp^3$   $sp^2$  dihedral.

**Figure 8 :** Shown is the  $\chi_1 = p$  slice of the Asn distribution as well as the allowed (green) and outlier (red) contours. Each point in the distribution represents one residue in the filtered dataset. There are several residues that lie outside the outlier contours – thus are outliers. Two examples that are far from any allowed contour (inside the circle) are shown with excellent  $2mF_o$ - $DF_c$ density. These valid outliers both exhibit extensive H-bonding which allows them to be held in an outlier conformation.

## 8 Tables

$\chi_1\chi_2$	n	%	$\chi_1$ mean	$\Delta$ ideal	$\chi_1$ sd	$\chi_2$ mean	$\Delta$ ideal	$\chi_2$ sd
<b>pp</b>	1301	0.4	60.7	0.7	8.3	84.0	<b>24.0</b>	10.4
pt	20733	5.6	64.1	4.1	7.7	177.0	3.0	11.4
<b>pm</b>	2086	0.6	69.5	<b>9.5</b>	9.3	-83.8	<b>23.8</b>	9.4
tp	50951	13.7	-178.6	1.4	8.9	63.3	3.3	7.8
tt	49913	13.4	-176.0	4.0	8.9	176.1	3.9	11.7
<b>tm</b>	2390	0.6	-172.5	<b>7.5</b>	9.5	-84.5	<b>24.5</b>	10.5
<b>mp</b>	8917	2.4	-71.3	<b>11.3</b>	12.0	78.2	<b>18.2</b>	14.5
mt	192806	51.9	-65.5	5.5	7.7	175.2	4.8	10.1
mm	42219	11.4	-62.8	2.8	8.9	-63.9	3.9	10.1

**Table 1:** All  $\chi_1\chi_2$  combinations in the filtered Top8000 for Leu, Ile, Met, Glu, Gln, and Arg. The backbone-interacting combinations are in **bold**, as are large deviations from ideal stagger.

Angle	E&H $\Delta$	$\chi$	stagger $\Delta$
<b>pm20</b> n=1442			
C $\alpha$ -C $\beta$ -C $\gamma$	+1.5°	1	+9.1°
C $\beta$ -C $\gamma$ -C $\delta$	+1.8°	2	-24.7°
C $\gamma$ -C $\delta$ -O $\epsilon 1$	+1.6°	3	-
<b>mp0</b> n=3568			
C $\alpha$ -C $\beta$ -C $\gamma$	+0.3°	1	-6.8°
C $\beta$ -C $\gamma$ -C $\delta$	+1.6°	2	+22.5°
C $\gamma$ -C $\delta$ -O $\epsilon 1$	+1.4°	3	-

**Table 2:** Mean angle differences, relative to Engh & Huber, and modal  $\chi$ -dihedral differences from stagger, for Glu **pm20**

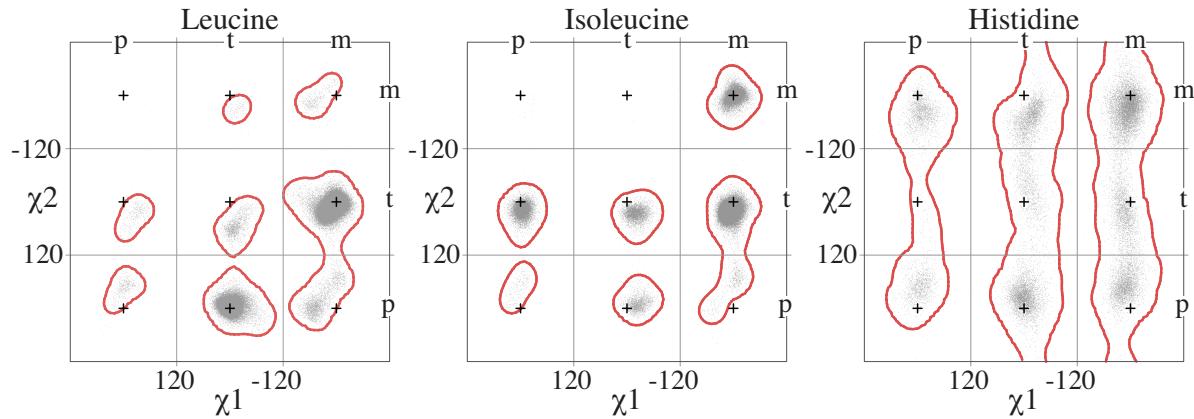
Angle	E&H $\Delta$	$\chi$	stagger $\Delta$
<b>pm20</b> n=249			
C $\alpha$ -C $\beta$ -C $\gamma 1$	+2.08°	1	-2.1°
C $\beta$ -C $\gamma 1$ -C $\delta 1$	+1.19°	2	+24.3°

**Table 3:** Mean bond-angle differences relative to Engh & Huber, and modal  $\chi$ -dihedral differences from stagger, for Ile pp

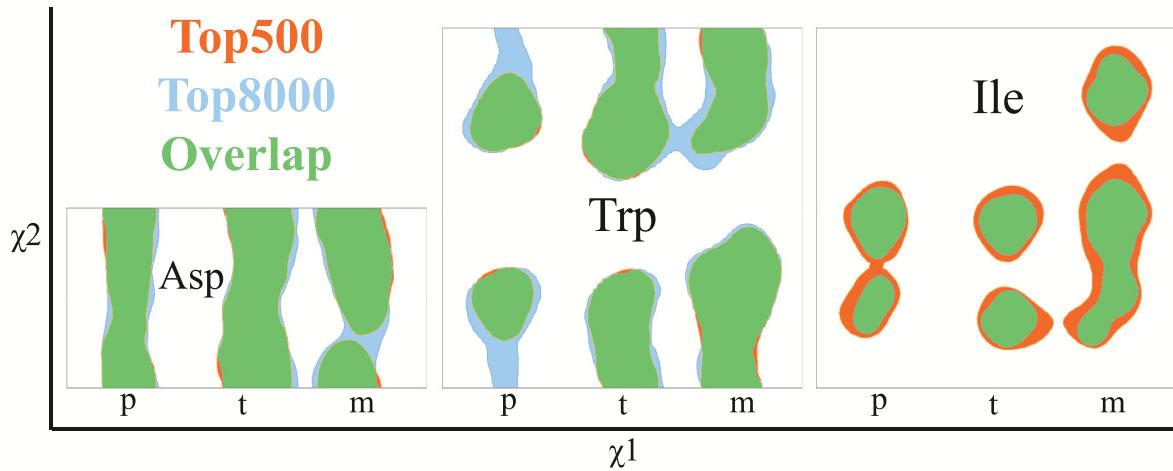
Angle	E&H $\Delta$	$\chi$	stagger $\Delta$
N-C $\alpha$ -C $\beta$	0.54°	1	-17.2°
C $\alpha$ -C $\beta$ -C $\gamma$	2.04°	2	+4.0°
C $\beta$ -C $\gamma$ -S $\delta$	3.15°	3	-41.8°

**Table 4:** Mean bond-angle differences, relative to Engh & Huber, and modal  $\chi$ -dihedral differences from stagger, for MET mpm.

## 9 Figures



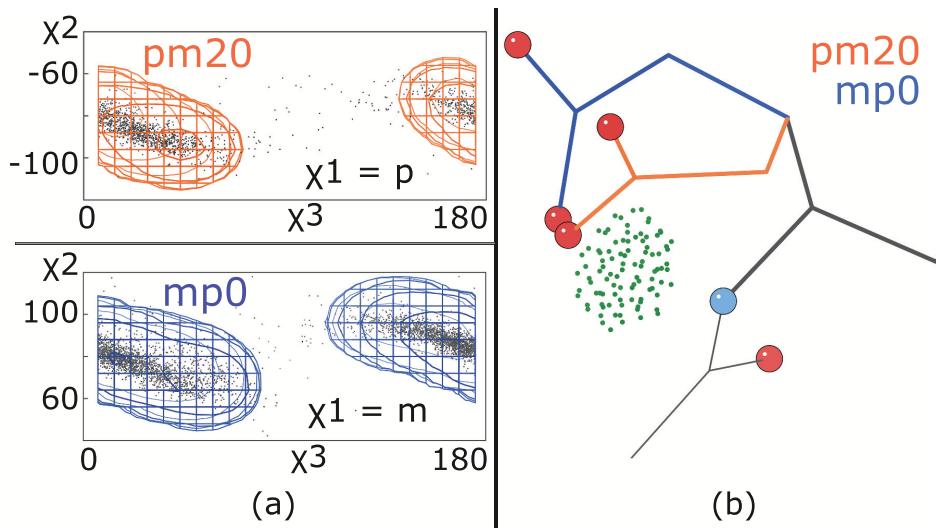
**Figure 1:**  $\chi_1$ - $\chi_2$  space for Leu, Ile, and His. The 0.3% contour is in red. Each point of the background data cloud represents a residue in the filtered Top8000 dataset. Crosshairs mark the nominally ideal staggered values between  $sp^3$  hybridized atoms, labeled as **m** ( $-60^\circ$ ), **t** ( $180^\circ$ ), and **p** ( $+60^\circ$ ).



**Figure 2:** Areas in orange (from Top500 data) and in blue (from Top8000) fill the allowed regions for Asp, Trp, and Ile. The extensive areas in green are where the two systems both declare allowed conformations. See also Figure S1.



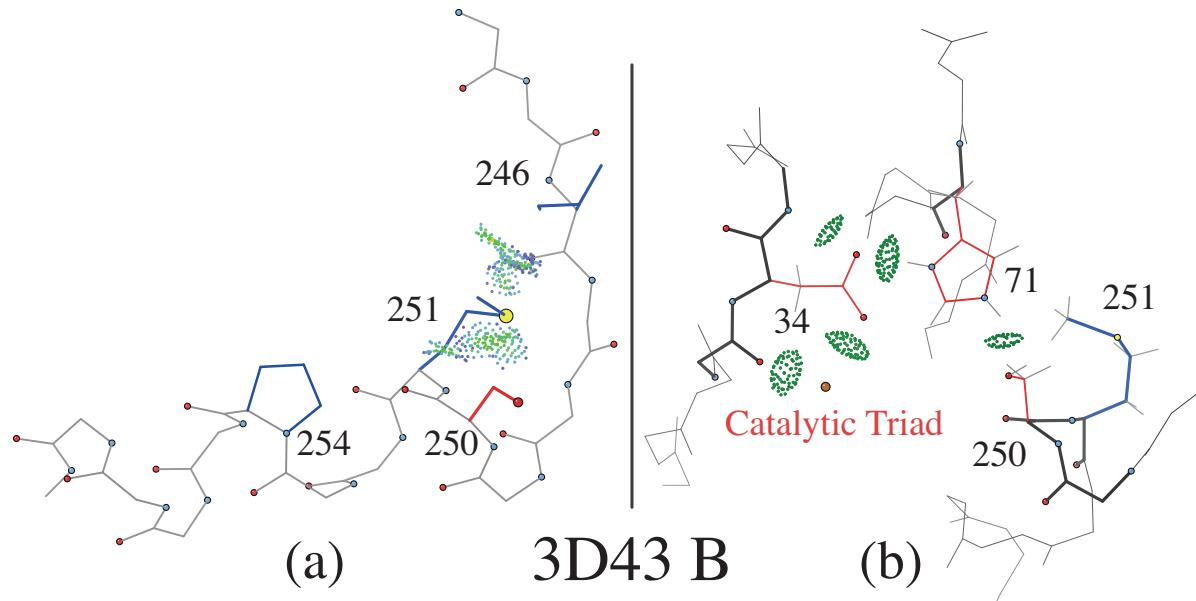
**Figure 3:** Sidechain contacts for Met A 240 in 2bmo, which adopts the rare methionine rotamer **ppp**. (a) Sulfur contacts, modeled with ideal staggered dihedrals and Engh & Huber bond angles, (b) with Engh & Huber bond angles, but dihedrals as deposited, (c) for the deposited structure, and (d) all contacts for the deposited structure, showing its tight vdW packing.



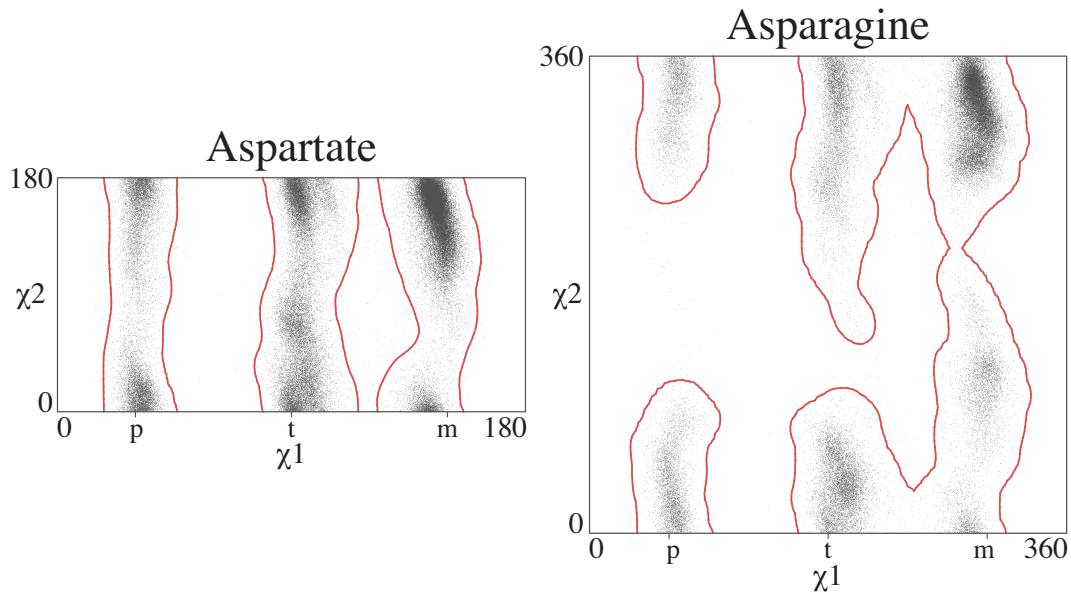
**Figure 4:** Panel (a) shows the 2.0% contour surfaces for the filtered Top8000 along with data points, as projected onto the  $\chi_2$ - $\chi_3$  plane for rotamers Glu **pm20** (orange) and **mp0** (blue). (b) shows how **pm20** and **mp0** both make a good H-bond (green dots) with the adjacent backbone NH, albeit through different conformations.



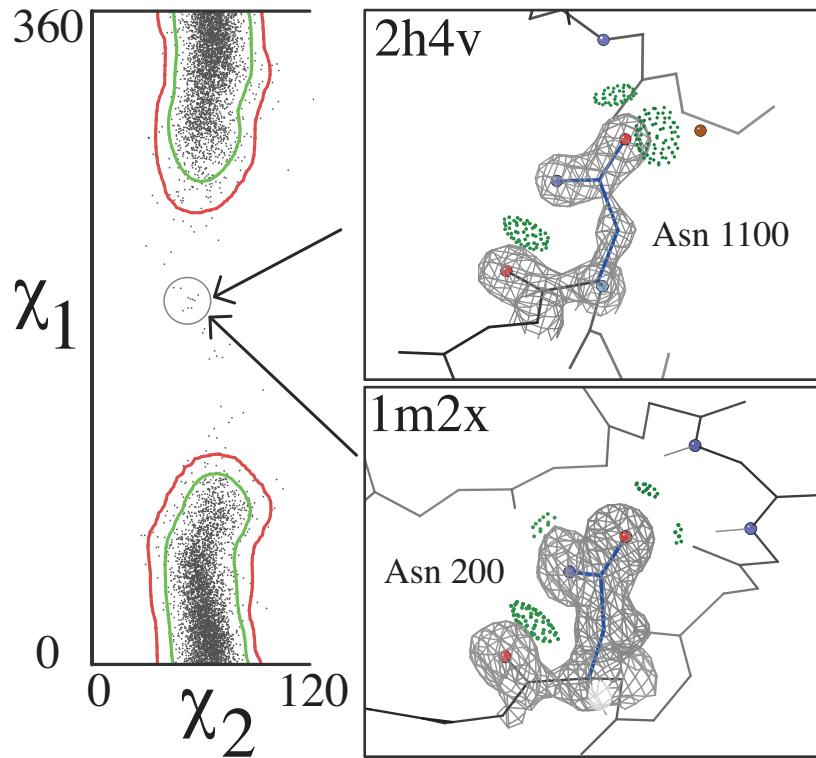
**Figure 5:** Two different examples of Ile **pp**, both demonstrating the close C $\delta$ 1/C contact, and the extensive vdW interactions that prevent the more common **pt** conformation. (a) 1s99 Ile A 76, (b) 2wvx Ile C 235.



**Figure 6:** 3D43 chain B, one of the 11 subtilisins containing a Met **mpm**. (a) The local arrangement of Ser250 (red) and Met251 (blue) on the helix, with the interacting Pro254 and Ile246; (b) rotated to show the active site, with its canonical catalytic triad in red and Met251 in blue.



**Figure 7:** Filtered Top8000 Asp and Asn datapoints and outlier contours (red).  $\chi_2$  does not follow the **ptm** convention since it is an  $\text{sp}^3 \text{ sp}^2$  dihedral.



**Figure 8:** Shown is the  $\chi_1 = \mathbf{p}$  slice of the Asn distribution as well as the allowed (green) and outlier (red) contours. Each point in the distribution represents one residue in the filtered dataset. There are several residues that lie outside the outlier contours – thus are outliers. Two examples that are far from any allowed contour (inside the circle) are shown with excellent  $2\text{mF}_o - \text{DF}_c$  density. These valid outliers both exhibit extensive H-bonding which allows them to be held in an outlier conformation.

## 10 Acknowledgments

We would like to thank Richardson lab members Dan Keedy and Bryan Arendall for setting up the chain-level Top8000, Jeff Headd for moving the MolProbity utilities to use the *PHENIX* cctbx toolbox, and Michael Prisant for help with GitHub. Also, thanks to *PHENIX* Project members Nigel Moriarty and Nat Echols for help with integration into *PHENIX* automation and GUI. This work was supported by NIH grants R01-GM073919 and ProjectIV of P01-GM063210, and by an NSF Graduate Research Fellowship to BJH.