PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

# PhiSiCal-Checkup: A Bayesian framework to validate amino acid conformations within experimental protein structures

Piyumi R. Amarasinghe[a], Lloyd Allison[a], Craig J. Morton[b], Peter J. Stuckey[a], Maria Garcia de la Banda[a], Arthur M. Lesk[c], and Arun S. Konagurthu[a,1]

Edited by Michael Levitt, Stanford University, Stanford, CA; received August 14, 2024; accepted October 4, 2024

**As structural biology and drug discovery depend on high-quality protein structures, assessment tools are essential. We describe a new method for validating amino-acid conformations: "PhiSiCal ($\phi\psi\chi$al) Checkup." Twenty new joint probability distributions in the form of statistical mixture models explain the empirical distributions of dihedral angles $\langle \omega, \phi, \psi, \chi_1, \chi_2, \ldots \rangle$ of canonical amino acids in experimental protein structures. Marginal and conditional probability distributions for subsets of dihedral angles are derived from these joint mixture models. Together, these distributions are employed to measure rapidly the information-theoretic "favorability" of any proposed experimental protein structure. The inferred statistical models and measures overcome several shortcomings and afford improvements over the current state of the art in amino-acid conformation verification. Experimental comparisons are made against current protein conformation verification software. In a number of examples, we pick up outliers that are invisible to current methods. We also calculate, as part of verification, the sensitivity of favorability to small changes in a proposed structure accounting for the precision of coordinates. In some cases a near neighbor of a proposed amino-acid conformation may be either less or more favorable. This raises the question, is the current reliance on fixed "thresholds" for validation a good thing? PhiSiCal-Checkup is freely available for online and offline (open-source) use from https://lcb.infotech.monash.edu.au/phisical/checkup.**

protein structure validation | amino acid conformation | conformation favorability | conformation outlier | Bayesian statistics

## Significance

Structural biology, biochemistry, evolutionary biology, medicine, and drug development all require high-quality protein structures. Reliable tools for assessing structures remain essential to ensure and maintain quality. We present a Bayesian method for analysis and validation of amino-acid conformations within protein structures: "PhiSiCal ($\phi\psi\chi$al) Checkup." This method overcomes major, long-standing shortcomings and provides significant improvements over the current state of the art. By introducing more-reliable information-theoretic measures of "favorability" of amino acid conformations, PhiSiCal-Checkup provides ways to analyze protein structures that were previously out of reach for experimentalists and structure biologists.

Experimental methods for protein structure determination produce raw measurements from which atomic coordinates are modeled as solutions to protein three-dimensional structures. The Worldwide Protein Data Bank (wwPDB) (1) administers the public archive (PDB) of protein coordinates (2) and scrutinizes them across a number of validation criteria, based on the guidelines of its validation task force (VTF) (3).

Key among the validation criteria is the knowledge-based assessment of amino-acid conformations. The current practice is to quantify the "favorability" of each amino acid's main chain conformation (described by $\langle \phi, \psi \rangle$ dihedral angles) and, separately, side chain conformation (described by $\langle \chi_1, \chi_2, \ldots \rangle$ dihedral angles) using their observed distributions within carefully curated datasets of protein structures (4). This quantification allows defining and detecting "outliers" from the distributions of conformational features. (Note that although outliers deserve further examination, they are not necessarily errors. Indeed, outliers that are correct often point to features of structural or functional interest. Conversely, non-outliers are not necessarily correct.) Popular programs to validate $\langle \phi, \psi \rangle$ and $\langle \chi_1, \chi_2, \ldots \rangle$ dihedral angles developed over the past three decades include PROCHECK (5), WHAT_CHECK (6), O (7), and MolProbity (8).

MolProbity represents the current state of the art, recommended by VTF (9). However, it has several limitations, some of which are acknowledged by its authors:

> "What we do expect will improve in the future is a redefinition of conformational validation...Rather than doing separate Ramachandran and rotamer evaluation, we should move toward analyzing all backbone and side-chain torsional dimensions together, including allowance for the influence of secondary structure and local motifs" (10).

(Others are discussed in *Results*, summarized in *SI Appendix*, Table ST1).

To overcome these limitations, we introduce "PhiSiCal ($\phi\psi\chi$al) Checkup," a comprehensive Bayesian and information-theoretic framework for validation of amino-acid conformations. Supporting PhiSiCal-Checkup are twenty new joint probability distributions, one for each of the canonical amino acids; each distribution treats all mainchain and sidechain dihedral angles together. These are inferred as statistical mixture models using our recently published inference methodology, PhiSiCal (11), based on

Author affiliations: [a]Department of Data Science and Artificial Intelligence, Monash University, Clayton, VIC 3800, Australia; [b]Biomedical Manufacturing Program, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Clayton, VIC 3168, Australia; and [c]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802

Author contributions: A.S.K. designed research; P.R.A. and A.S.K. performed research; P.R.A., L.A., A.M.L., and A.S.K. contributed new reagents/analytic tools; P.R.A., L.A., C.J.M., P.J.S., M.G.d.l.B., A.M.L., and A.S.K. analyzed data; C.J.M. introduced the validation problem; and P.R.A., A.M.L., and A.S.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.



[1]To whom correspondence may be addressed. Email: arun.konagurthu@monash.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2416301121/-/DCSupplemental.

Published January 2, 2025.

the unsupervised model selection criterion of Minimum Message Length (12, 13). [Although PhiSiCal provides the core methodology for inference of joint mixture models, specializing these models for the conformation-validation implemented by PhiSiCal-Checkup required several new methodological advances previously unexplored (*Materials and Methods*)].

In PhiSiCal-Checkup, each joint mixture model can be factorized into mutually consistent probability distributions to analyze any possible combination of dihedral angle terms. Specifically, for any amino-acid type containing $d$ dihedral angle terms ($\langle \underbrace{\omega, \phi, \psi, \chi_1, \chi_2, \ldots}_{d \text{ dihedral angle terms}} \rangle$), the corresponding joint mixture model (defining a distribution in a $d$-dimensional toroidal space, $\mathbb{T}^d$) can be transformed "on-the-fly" using Bayesian axioms of probability to derive $2^d - 2$ marginal and $3^d - 2^{d+1} + 1$ conditional probability distributions. The formal relationship between joint, marginal, and conditional distributions is governed by the Bayes theorem (14) making all amino acid–specific distributions mutually consistent.

Marginal probability distributions allow the unconstrained evaluation of *any* combination of the $d$ dihedral angle terms, of which the assessment of main chain $\langle \phi, \psi \rangle$ and side chain $\langle \chi_1, \chi_2, \ldots \rangle$ conformations are but two possibilities. Conditional distributions allow the evaluation of any subset of dihedral angles subject to observing specific values for others, providing meticulous new ways of analyzing amino-acid conformations. For example, the conditional probability distribution of the $\langle \chi_1, \chi_2, \ldots \rangle$ terms given some specific observed values for $\langle \phi, \psi \rangle$ (shorthand, $\langle \chi_1, \chi_2, \ldots \rangle \mid \langle \phi, \psi \rangle$) can be used to evaluate the favorability of the observed $\langle \chi_1, \chi_2, \ldots \rangle$ dihedral angles conditioned on the $\langle \phi, \psi \rangle$ observation. When the (continuous) $\langle \phi, \psi \rangle$ values correspond to, say, a (discrete) secondary structural state, the $\langle \chi_1, \chi_2, \ldots \rangle \mid \langle \phi, \psi \rangle$ probability distribution will automatically embody the secondary-structural constraint on the $\langle \chi_1, \chi_2, \ldots \rangle$ angles. In extension, the conditional distribution $\langle \chi_1, \chi_2, \ldots \rangle \mid \langle \omega, \phi, \psi \rangle$ constrains the evaluation of $\langle \chi_1, \chi_2, \ldots \rangle$ additionally on the cis/trans/twisted peptide state informed by the continuous value of $\omega$, beyond the state informed by $\langle \phi, \psi \rangle$. As another example, the conditional distribution of $\langle \phi, \psi \rangle \mid \omega$ allows evaluating $\langle \phi, \psi \rangle$ given any observed continuous value for $\omega$, thereby automatically accounting for $\omega$ being in cis/trans/twisted states. Data in ref. 15 shows the animations of amino acid–specific $\langle \phi, \psi \rangle \mid \omega$ distributions for varying $\omega$)

This work also introduces new statistical measures to validate amino-acid conformations. PhiSiCal-Checkup's validation of dihedral angles (in any combination of terms) is driven by the information-theoretic measure of lossless compression, quantified in bits. Favorability (or unfavorability) of any conformation can objectively be assessed based on the amount of compression gained (or lost) with respect to the raw bit-length of encoding the observed dihedral angles. Statistically, this measure of compression is equivalent to quantifying the log-likelihood-odds of any observation arising from its corresponding (amino acid–specific joint/marginal/conditional) mixture model compared to the raw (uniform) distribution. Further, since all statistical mixture models are continuous and differentiable at every point in the respective distributions' support, a gradient vector in the probability distribution can be computed for any observed dihedral angles. Using this gradient, sensitivity of the reported compression statistic to perturbation of dihedral angles can be analyzed accounting for the (im)precision of statement of atomic coordinates. This overcomes another limitation in the current state of art: the reliance on hard (non-overlapping) membership assignments of observations to one of three categories—"outlier," "allowed," and "favored"—that can yield misleading assignments, especially near the boundaries of those categories.

A configurable web server and standalone software implementing PhiSiCal-Checkup is available for immediate online and offline use: https://lcb.infotech.monash.edu.au/phisical/checkup. Validation of dihedral terms is fully customizable and supported by instructive visualizations.

## Results and Discussion

**Statistical Mixture Models for Validation of Dihedral Angles.** Twenty new joint statistical mixture models are inferred in this work, one for each of the canonical amino acids, using the reference protein structural data set of Williams et al. (4), curated specifically for the task of structure validation (*Materials and Methods* and *SI Appendix*, Table ST2). Each joint mixture model describes a continuous amino acid–specific probability distribution of its corresponding dihedral angle terms, $\langle \omega, \phi, \psi, \chi_1, \chi_2, \ldots \rangle$. Amino acid–specific marginal and conditional probability distributions are derived (on demand) from these inferred joint mixture models. Importantly, these distributions are also statistical mixture models, characterized in a reduced number of dimensions defined by the subset they explain (*Materials and Methods*).

The amino acid–specific joint mixture model, combined with the derived marginal and conditional mixture models provide the "basis set" of probability distributions for PhiSiCal-Checkup to interrogate dihedral angle conformations in any possible combination of terms. The computation of joint/marginal/conditional probabilities and other derived-statistics remains extremely efficient taking between 10 to 60 microseconds-per-computation on a standard (single-thread) computer.

**Compressibility of Observed Dihedral Angles.** PhiSiCal-Checkup uses the measures of Shannon information content (16) and compression to quantify the surprise of observing any amino-acid conformation. These measures are derived from the corresponding joint/marginal/conditional mixture model, depending on the combination of dihedral angle terms being observed. Compression (measured in bits) is statistically equivalent to computing the log-odds of an observation arising from its corresponding mixture model, compared to the uniform distribution. Compression (gain) of $+n$ bits yields $2^n$:1 odds in favor of the mixture model, whereas compression (loss) of $-n$ bits yields $1{:}2^n$ odds against it. Importantly, PhiSiCal-Checkup also assesses the sensitivity of the computed compression by locally perturbing the observation in the direction of its gradient vector in the probability distribution (*Materials and Methods*).

To understand compression, its sensitivity, and statistical odds, consider the case study of Chain A Arginine 368 (Arg-368) in the PDB coordinates of arabinoxylan arabinofuranohydrolase (AXAH) from *Bacillus subtilis* (3C7F). The conformation of Arg-368 in 3C7F (Chain A) yields the following dihedral angle observations: $\omega = -19.15°$, $\phi = -130.68°$, $\psi = 120.11°$, $\chi_1 = -67.61°$, $\chi_2 = 178.90°$, $\chi_3 = -174.16°$, $\chi_4 = -171.50°$, $\chi_5 = -179.13°$. Particularly, $\omega = -19.15°$ being discussed here is the rotation around the peptide bond connecting the preceding Glycine (Gly-367) and the current Arg-368. This suggests a strained, energetically unfavorable *cis*-conformation for Arg-368. However, this is not an error: checking the 1.55 Å

crystal structure coordinates of 3C7F, the carbonyl preceding Arg-368 is seen facilitating the binding of the Sodium ($Na^+$) metal ion (17).

Table 1 shows the compression and sensitivity statistics of observing Arg-368 in 16 combinations of dihedral angle terms (of the total 6,305 possible for Arginine). The individual observation of $\omega = -19.15°$ (ignoring all other dihedral angles) results in a loss of compression (−11.53 bits) using its corresponding Arginine-specific marginal mixture model. This gives $1{:}2^{11.53} \approx$ 1:3,000 odds against that mixture model. The compression's sensitivity is derived by perturbing $\omega = -19.15° \pm 2.5°$ in the direction of its gradient, causing compression to vary between $[-12.2, -10.9]$ bits (and odds between ≈[1:2,000, 1:4,700]) in the local neighborhood of that observation. Similarly, the observed values for $\langle \omega, \phi, \psi \rangle$ explained using its marginal mixture model result in a loss of compression of −9.22 bits (odds of $\approx$ 1 : 600) with a sensitivity of ±0.7 bits. The evaluation of all $\langle \omega, \phi, \psi, \chi_1, \chi_2, \ldots \rangle$ terms using the joint mixture model results in a loss of compression of −6.77 bits ($\approx$ 1 : 100 odds). In contrast, observing only the Arg-368's side chain ($\langle \chi_1, \chi_2, \ldots \rangle$) terms leads to a gain in compression of +11.24 bits (≈2,400:1 odds). All other dihedral angles (in combinations not involving the strained $\omega$) lead to gain in compression. Finally, the Arg-368's side chain conformation conditioned on the pair of Ramachandran angles taking the values $\langle \phi, \psi \rangle = \langle -130.68°, 120.11° \rangle$ (but ignoring the strained cis-$\omega = -19.15°$) leads to a gain in compression of +11.43 bits. However, when $\omega = -19.15°$ is also considered, the compression-gain drops sharply to 2.5 bits. This is equivalent to the statistical odds dropping from ≈2,800:1 to ≈6:1, thus quantifying the extent of surprise of observing that $\omega$ as part of Arg-368's overall conformation.

We note that, in the current state of the art (MolProbity), the surprise of the $\omega$ value at the peptide bond between Gly-367 and Arg-368 (that gives Arg-368 a *cis* conformation) is overlooked as it can only validate the $\langle \phi, \psi \rangle$ and $\langle \chi_1, \chi_2, \ldots \rangle$ terms (independently). In sum, such detailed analysis of observed amino-acid conformations, at varying granularity and constraints, is unique to PhiSiCal-Checkup.

**Quantifying Surprise of Observed Dihedral Angles.** From the relationship between statistical odds and compression, it follows that compression at zero bits yields the surprisal-odds of 1:1 (fifty–fifty), thus demarcating an objective boundary for the joint/marginal/conditional distributions: observations below zero become exponentially surprising and those above exponentially favorable.

Here, we explore how compression correlates with the empirical frequencies of dihedral angles. To achieve this we analyzed the percentile-rank distribution of compression across the 1,720,588 amino-acid conformations found in the reference dataset (4).

Fig. 1*A* tracks the bits of compression at varying percentile levels ({min, 0.05, 0.3, 2, 5, 10, 25, 50, 75, max}) for $\langle \phi, \psi \rangle$ and $\langle \chi_1, \chi_2, \ldots \rangle$ dihedral angle observations of each amino-acid type. (*SI Appendix*, Table ST3 provides a tabular view of the figure in raw numbers. The last row of this table summarizes the mean and SD of the compression values at each of those percentile levels.) Broadly, the compression values at each percentile mark for $\langle \phi, \psi \rangle$ show relatively lower dispersion about the mean compression values, compared to those of $\langle \chi_1, \chi_2, \ldots \rangle$. This is expected because the differences in the amino-acid types arise due to their side chain groups that show varying conformational-mobility and energetics (18). In the lower-half of the distribution below the median, the compression statistics of Proline $\langle \phi, \psi \rangle$ and $\langle \chi_1, \chi_2, \ldots \rangle$ observations diverge the most, followed by Glycine for $\langle \phi, \psi \rangle$ and Arginine for $\langle \chi_1, \chi_2, \ldots \rangle$. This likely arises due to Proline's cyclic-pyrrolidine side chain, Glycine's absence of $\beta$-carbon, and Arginine's side chain length.

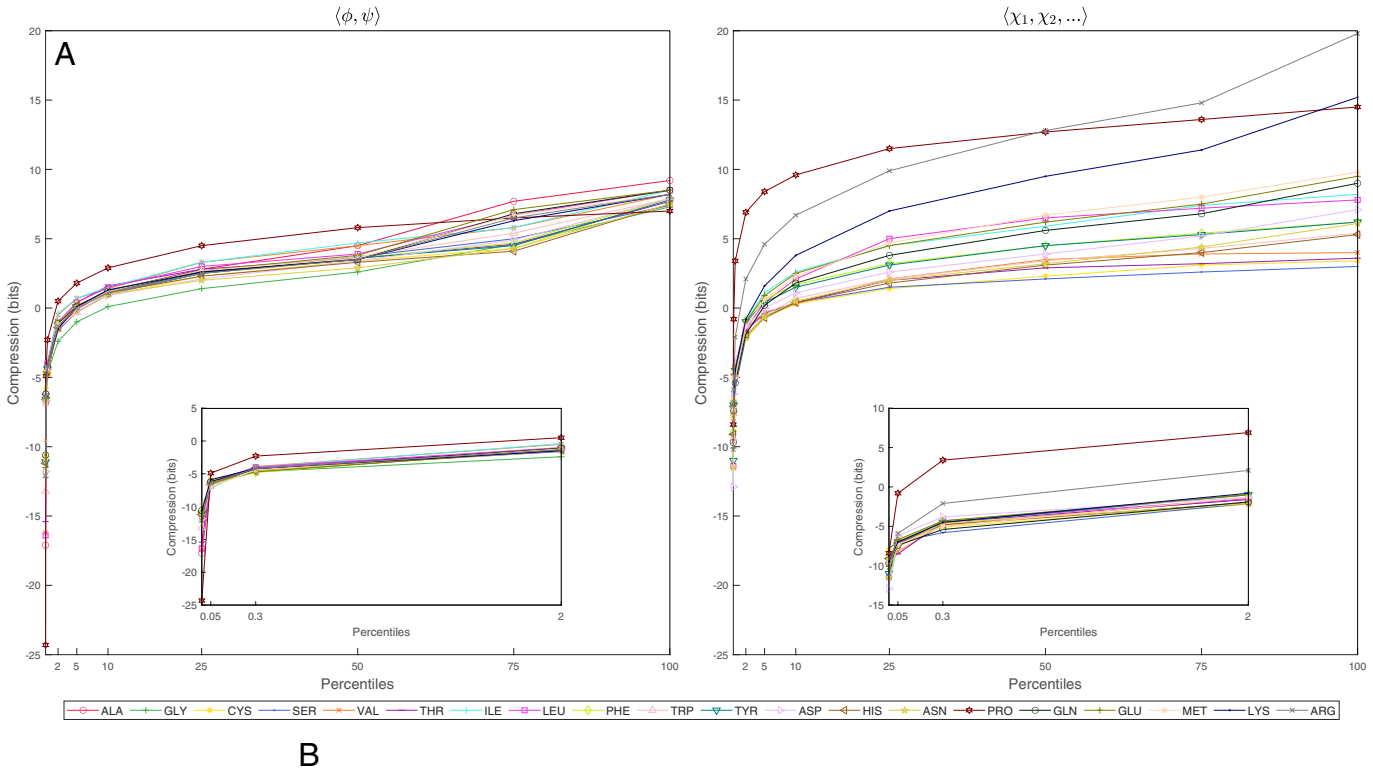
Focusing on the distributions' (infrequent) tails, Fig. 1*B* shows the proportions of $\langle \phi, \psi \rangle$ and $\langle \chi_1, \chi_2, \ldots \rangle$ observations that lose compression (<0 bits). For $\langle \phi, \psi \rangle$, this accounts for ∼5±1% of observations for most amino acids, except for Glycine with 9.1%, Valine and Isoleucine with 2.8% and Proline with 1.4% that deviate from this trend. For $\langle \chi_1, \chi_2, \ldots \rangle$, the proportions are more spread out, ranging from 0.1% for Proline to 8.2% for Cysteine. These results quantitatively highlight the differences in the distributions of dihedral angles across the 20 amino-acid types.

These differences were further explored by qualitatively analyzing the compression statistics for each amino acid using *all-pairs* compression contour maps derived from their corresponding (marginal) mixture models. Specifically, for each amino-acid type with $d$ dihedral terms $\langle \omega, \phi, \psi, \chi_1, \chi_2, \ldots \rangle$, $d$-choose-2 contour plots were generated for each possible pair of dihedral angle terms (19). Each 2D plot shows 1) the empirical distribution for that pair, 2) compression contour lines at 1-bit intervals between $[-5, +5]$ bits, and 3) gradient vectors showing the rate of change of probability (and hence compression) at 5° × 5° intervals. Visual examination of these plots again highlights the significant differences in the amino acid–specific distributions of dihedral angles.

These plots also illustrate the close-fit between the mixture models and the underlying empirical distributions. As an example, Fig. 2 shows two such plots for Proline. Specifically,

**Table 1. Compression and sensitivity of dihedral angles (in varying combinations) of Arg-368 in the protein coordinates of 3C7F Chain A**

| Observation (mixture model) | Compression (±sensitivity) | Observation (Mixture model) | Compression (±sensitivity) |
|---|---|---|---|
| $\omega_{(\text{marginal})}$ | $-11.5\ (^{+0.6}_{-0.7})$ bits | $\langle \omega, \phi, \psi \rangle_{(\text{marginal})}$ | $-9.2\ (^{+0.7}_{-0.7})$ bits |
| $\phi_{(\text{marginal})}$ | $1.1\ (^{+0.0}_{-0.0})$ bits | $\langle \phi, \psi \rangle_{(\text{marginal})}$ | $2.5\ (^{+0.3}_{-0.3})$ bits |
| $\psi_{(\text{marginal})}$ | $0.7\ (^{+0.3}_{-0.3})$ bits | $\langle \chi_1, \chi_2, \chi_3, \chi_4, \chi_5 \rangle_{(\text{marginal})}$ | $11.2\ (^{+0.0}_{-1.4})$ bits |
| $\chi_{1(\text{marginal})}$ | $3.4\ (^{+0.0}_{-0.1})$ bits | $\langle \phi, \psi \rangle \mid \omega_{(\text{conditional})}$ | $2.3\ (^{+0.1}_{-0.1})$ bits |
| $\chi_{2(\text{marginal})}$ | $3.2\ (^{+0.0}_{-0.0})$ bits | $\chi_1 \mid \langle \phi, \psi \rangle_{(\text{conditional})}$ | $3.0\ (^{+0.3}_{-0.5})$ bits |
| $\chi_{3(\text{marginal})}$ | $2.6\ (^{+0.2}_{-0.3})$ bits | $\langle \chi_1, \chi_2, \chi_3, \chi_4, \chi_5 \rangle \mid \langle \phi, \psi \rangle_{(\text{conditional})}$ | $11.4\ (^{+0.0}_{-1.4})$ bits |
| $\chi_{4(\text{marginal})}$ | $1.5\ (^{+0.1}_{-0.1})$ bits | $\langle \chi_1, \chi_2, \chi_3, \chi_4, \chi_5 \rangle \mid \langle \omega, \phi, \psi \rangle_{(\text{conditional})}$ | $2.5\ (^{+0.0}_{-0.0})$ bits |
| $\chi_{5(\text{marginal})}$ | $2.3\ (^{+0.0}_{-1.5})$ bits | $\langle \omega, \phi, \psi, \chi_1, \ldots, \chi_5 \rangle_{(\text{joint})}$ | $-6.8\ (^{+0.7}_{-0.7})$ bits |




B

| | ALA | GLY | CYS | SER | VAL | THR | ILE | LEU | PHE | TRP |
|---|---|---|---|---|---|---|---|---|---|---|
| $\langle\phi,\psi\rangle$: (% $<0$ bits) | 5.2% | 9.1% | 5.9% | 5.8% | 2.8% | 4.2% | 2.8% | 4.0% | 4.3% | 6.0% |
| $\langle\chi_1,\chi_2,\ldots\rangle$: (% $<0$ bits) | n/a | n/a | 8.2% | 7.5% | 6.6% | 7.0% | 3.1% | 4.2% | 3.4% | 6.6% |

| | TYR | ASP | HIS | ASN | PRO | GLN | GLU | MET | LYS | ARG |
|---|---|---|---|---|---|---|---|---|---|---|
| $\langle\phi,\psi\rangle$: (% $<0$ bits) | 4.5% | 4.5% | 4.4% | 4.5% | 1.4% | 4.5% | 4.8% | 5.1% | 4.7% | 5.1% |
| $\langle\chi_1,\chi_2,\ldots\rangle$: (% $<0$ bits) | 3.8% | 4.9% | 7.9% | 7.5% | 0.1% | 4.6% | 3.3% | 3.8% | 2.8% | 0.8% |

**Fig. 1.** (*A*) Bits of compression at varying percentile levels for amino acid–specific Ramachandran $\langle\phi,\psi\rangle$ (*Left*) and side chain $\langle\chi_1,\chi_2,\ldots\rangle$ (*Right*) dihedral angle terms, using 1,720,588 amino-acid conformations observed in the reference dataset (4). *Insets* show a zoomed-in view of the left-tail of the distributions. (*B*) Amino acid–specific percentile ranks with <0 bits of compression are shown. Note: Amino Acids Alanine (ALA) and Glycine (GLY) do not have side chain dihedral angle terms.

Fig. 2*A* displays the contour plot for Proline $\langle\phi,\psi\rangle$ terms. The contours correlate closely with the empirical frequencies of observed $\langle\phi,\psi\rangle$ angles (encoded in light-to-dark shades of blue). The +5-bit compression lines (innermost black contours) encompass two regions of high-probability (20) that peak at $\langle\phi,\psi\rangle \approx \langle -60°, +145°\rangle$ and $\approx \langle -60°, -30°\rangle$, where the norm of the gradient vectors in those regions approach 0. At −5 bits of compression (outermost, red line), the contour encompasses infrequent/low-probability regions in the valleys of Proline's $\langle\phi,\psi\rangle$ distribution. Importantly, points/observations with the same compression values can have significantly different gradients (and hence sensitivities to local perturbations), as illustrated by points labeled $P_1$ and $P_2$ in the figure. Locally perturbing $P_1$ in the direction of its gradient improves compression (and statistical odds) significantly more than it does at $P_2$.

Similarly, Fig. 2*B* shows the compression contours and gradient vectors for Proline's side chain dihedral angle pair $\langle\chi_2,\chi_3\rangle$. The contour lines again correlate closely with the empirical frequencies of that pair within the observed Proline conformations. This plot additionally overlays a set of 29 anomalous Proline observations, all from the same protein structure (3H8G), 1.5 Å Bestatin complex structure of leucine aminopeptidase from *Pseudomonas putida* (21). Surprisingly, in all 29 observations, $\chi_2$ is nearly 0° (*SI Appendix*, Table ST4). Examining the PDB validation report of 3H8G, none of these Prolines were earmarked as outliers by MolProbity. (Note, MolProbity ignores $\chi_2$ and $\chi_3$ dihedral angles in their evaluation for Proline and considers only $\chi_1$.) This illustrates how the contour-plots enabled by PhiSiCal-Checkup can be used to identify and examine surprising conformations that deviate from the observed distributions.

Altogether, the above analyses reveal quantitative and qualitative differences in the empirical distributions of dihedral angles across amino acids. These differences highlight the importance of using amino acid–specific distributions to validate dihedral angle conformations. Further, the analyses also reveal the variation of compression statistic at any fixed percentile threshold across amino-acid types. This throws into question the prevalent use of fixed percentile-thresholds to flag $\langle\phi,\psi\rangle$ and $\langle\chi_1,\chi_2,\ldots\rangle$ outliers in the current validation protocols (see discussion below).

**Comparison with MolProbity.** We compare PhiSiCal-Checkup with MolProbity. *SI Appendix*, Table ST1 summarizes the key differences between the two systems. Among the differences is the statistical test they rely on to validate amino-acid conformations. Central to MolProbity's method is the use of fixed percentile rank thresholds that do not change with amino-acid type. Specifically, MolProbity employs a 3-way classification of $\langle\phi,\psi\rangle$ and $\langle\chi_1,\chi_2,\ldots\rangle$ observations. Each observation is assigned a hard-membership to one of {outlier, allowed, favored}

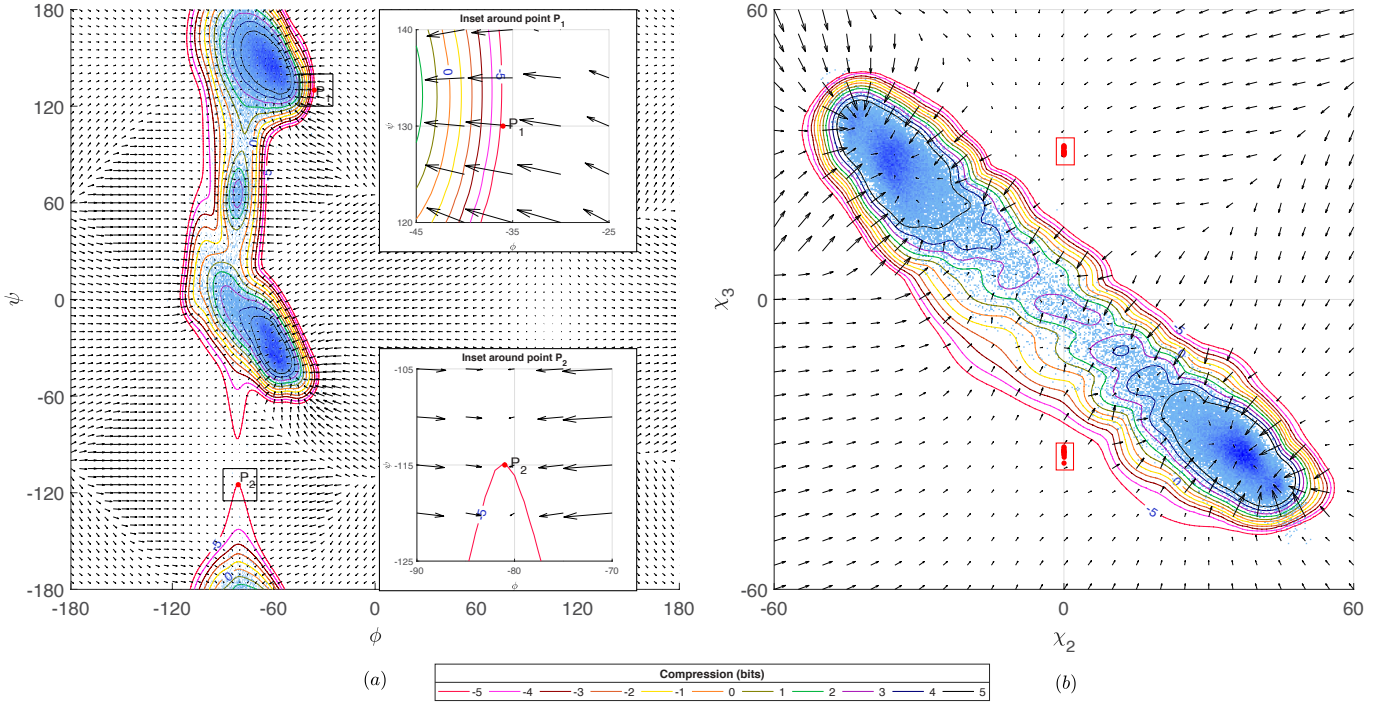



**Fig. 2.** Proline's empirical observations (shown as scatter points colored in light-to-dark shades of blue colored based on their 1° × 1° grid-frequencies), compression contours (multicolored lines) in 1-bit intervals between [−5, +5], and gradient vectors (black arrows) at 5° × 5° interval are shown above for (*A*) $\langle\phi, \psi\rangle$ and (*B*) $\langle\chi_2, \chi_3\rangle$ dihedral angle terms. The $\langle\phi, \psi\rangle$ plot also shows two points $P_1$ and $P_2$ on the −5 bit (outmost, red) compression line. The *Insets* around these two points highlight the differences in gradients (and hence sensitivities under local-perturbation). For $\langle\chi_2, \chi_3\rangle$ plot, the axes are truncated to [−60°, +60°], as there are no empirical observations beyond that in the reference dataset. Separately, a set of anomalous observations are highlighted (red points)—refer to the main text for discussion.

categories based on the percentile rank of the observation's score (derived from their normalized functions) within a set of scores precomputed for their reference data. MolProbity sets the outlier thresholds of 0.05% (significance-level 0.0005) for $\langle\phi, \psi\rangle$ and 0.3% (significance-level 0.003) for $\langle\chi_1, \chi_2, \ldots\rangle$ observations. For the allowed category, the percentile-threshold is set at 2% (significance-level 0.02) for both. Consequently, observations that fall above the 2nd percentile as per MolProbity's scoring earmarks the favored category.

In statistical parlance, MolProbity is relying on a $Z$-test for a hard 3-way clustering of $\langle\phi, \psi\rangle$ and $\langle\chi_1, \chi_2, \ldots\rangle$ observations. However, a $Z$-test is effective only if the test-statistic is normally distributed: only then can the significance-levels of 0.0005, 0.003, 0.02 correspond to Z-scores thresholds of $\sim \pm 3.5\sigma$, $\sim \pm 3\sigma$, and $\sim \pm 2.3\sigma$ from the respective mean values of scores in a two-tailed test. Importantly, our analysis finds no evidence of normality of the underlying probabilities (which the MolProbity's function-scores are approximating). *SI Appendix*, Figs. SF1 and SF2 show the distribution of probabilities for the $\langle\phi, \psi\rangle$ and $\langle\chi_1, \chi_2, \ldots\rangle$ conformations observed in the reference dataset, derived using PhiSiCal-Checkup's mixture models. This is quantitatively supported by the observed variance in the compression statistic across amino-acid types at 0.05, 0.3, and 2 percentile levels (*SI Appendix*, Table ST3). Therefore, PhiSiCal-Checkup avoids using percentile-rank thresholds in its method of validation.

Instead, PhiSiCal-Checkup relies on the information-theoretic measure of compression to quantify surprisal-odds of any observation. Compression at 0 bits defines an objective threshold below which observations grow exponentially surprising, and above which, exponentially favorable. Thus, PhiSiCal-Checkup uses compression of $\geq 0$ bits to earmark favored observations. For comparison with MolProbity, another threshold to flag outlier conformations requires to be defined, while noting that such a choice remains fully subjective and cannot be formally defended. A subjective outlier threshold at −4 bits of compression (or 1:16 odds) is thus chosen here as a default setting in PhiSiCal-Checkup. (Note, fractional odds (e.g., 1:16) should not be confused with frequentist percentile-based probabilities. In MolProbity, a score ranked at the 0.05th percentile gives a 1 in 2,000 *chance* of observing that score in their pre-computed set of scores. At the same percentile rank (refer *SI Appendix*, Table ST3), PhiSiCal-Checkup, on average, yields −6.4 bits of compression for $\langle\phi, \psi\rangle$ observations, or ∼1:85 surprisal-odds.)

This results in a 3-way categorization for PhiSiCal-Checkup based on compression: outlier $< -4$ bits, allowed $[-4, 0)$ bits and favored $\geq 0$ bits. A key difference compared with MolProbity's categorization is that PhiSiCal-Checkup permits overlapping membership-assignment based on the observation's gradient information (discussed below). The gradient accounts for the uncertainty (imprecision) of stated atomic coordinates and alerts users of observations that are close to the boundaries of PhiSiCal-Checkup's 3-way classification, and whose membership status is sensitive to minor perturbations of observed dihedral angles.

Table 2 summarizes the agreement/disagreement between the two systems. This is based on 3,624,568 $\langle\phi, \psi\rangle$ (Table 2*A*) and 3,027,146 $\langle\chi_1, \chi_2, \ldots\rangle$ (Table 2*B*) observations derived from 9,419 filtered-PDB structures (23). The corresponding details at the level of individual amino-acid types are available from refs. 24 and 25.

**Table 2. (*A* and *B*) 3 × 3 tables (confusion matrices) displaying the extent of agreement (main diagonal cells) and disagreement (off-diagonal cells) between PhiSiCal-Checkup (compression-based) and MolProbity (percentile-rank based) systems, performing 3-way ({outlier, allowed, and favored}) classification of 3,624,568 $\langle \phi, \psi \rangle$ observations and 3,027,146 $\langle \chi_1, \chi_2, \ldots \rangle$ observations**

**Confusion matrix for $\langle \phi, \psi \rangle$ observations**

| MolProbity \ PhiSiCal-Checkup (Hard membership thresholds →) | $<-4$ bits | $\geq -4 \& < 0$ bits | $\geq 0$ bits |
|---|---|---|---|
| $<0.05$ percentile | 3,887 (95.1%) | 191 (4.7%) | 9 (0.2%) |
| $\geq 0.05$ & $<2$ percentile | 13,185 (16.9%) | 56,752 (72.8%) | 8,003 (10.3%) |
| $\geq 2$ percentile | 329 (0.01%) | 91,844 (2.59%) | 3,450,368 (97.4%) |

A

**Confusion matrix for $\langle \chi_1, \chi_2, \ldots \rangle$ observations**

| MolProbity \ PhiSiCal-Checkup (Hard membership thresholds →) | $<-4$ bits | $\geq -4 \& < 0$ bits | $\geq 0$ bits |
|---|---|---|---|
| $<0.3$ percentile | 27,272 (75.3%) | 8,351 (23.0%) | 602 (1.7%) |
| $\geq 0.3$ & $<2$ percentile | 15,601 (15.6%) | 74,140 (74.3%) | 9,996 (10.0%) |
| $\geq 2$ percentile | 1,069 (0.03%) | 109,256 (3.77%) | 2,780,859 (96.2%) |

B

**Difference in membership numbers ($\langle \phi, \psi \rangle$)**
(after minor perturbation along the gradient)

| MolProbity \ PhiSiCal-Checkup (Hard membership thresholds →) | $<-4$ bits | $\geq -4 \& < 0$ bits | $\geq 0$ bits |
|---|---|---|---|
| $<0.05$ percentile | -368 | 365 | 3 |
| $\geq 0.05$ & $<2$ percentile | -5,963 | -8,892 | 14,855 |
| $\geq 2$ percentile | -262 | -51,354 | 51,616 |

C

**Difference in membership numbers ($\langle \chi_1, \chi_2, \ldots \rangle$)**
(after minor perturbation along the gradient)

| MolProbity \ PhiSiCal-Checkup (Hard membership thresholds →) | $<-4$ bits | $\geq -4 \& < 0$ bits | $\geq 0$ bits |
|---|---|---|---|
| $<0.3$ percentile | -6,025 | 5,697 | 328 |
| $\geq 0.3$ & $<2$ percentile | -11,401 | -3,392 | 14,793 |
| $\geq 2$ percentile | -928 | -87,132 | 88,060 |

D

Rows and columns represent MolProbity's and PhiSiCal-Checkup's respective 3-way assignments. The agreement/disagreement percentages with respect to MolProbity's membership assignments are shown in parentheses in each cell—row percentages add up to 100. (*C* and *D*) Difference matrices (corresponding to the confusion matrices shown above) quantifying the number of observations that change their membership in PhiSiCal-Checkup upon a minor perturbation of each observation in the direction of the observation's gradient in their probability distributions. Negative numbers indicate the reduction and positive indicate accretion in the corresponding column category for each row (sum of these differences in each row has to add up to 0).

Broadly, ∼97% of all $\langle \phi, \psi \rangle$ observations and ∼95% of all $\langle \chi_1, \chi_2, \ldots \rangle$ observations are in agreement between the two systems. However, this is dominated by the overrepresentation of observations assigned to the favored category by both systems. This arises because 1) the measure-space of this category (≥ 2% for MolProbity and ≥ 0 bits for PhiSiCal-Checkup) is disproportionately larger than that of the other two, and 2) the observations come from verified PDB structures which already embody this representational imbalance, with favored conformation forming the very basis for admission into PDB. With this drastic imbalance, any two systems are likely to display overwhelming agreement. We emphasize here that flagging outliers is a business fully in the tails of the distributions, so it becomes necessary to examine more carefully similarities and differences in the ∼3%/∼5% tails of the $\langle \phi, \psi \rangle$/$\langle \chi_1, \chi_2, \ldots \rangle$ observations.

Tables 2 *A* and *B* show the raw counts of agreement/disagreement between the two systems in the $3 \times 3 = 9$ possible combinations of assignments between the two systems: MolProbity (rows) and PhiSiCal-Checkup (columns). Each cell also shows the agreement/disagreement-percentage with respect to MolProbity's classification, with the sum of percentages in each row adding up to 100%. Analyzing the differences (off-diagonal cells), we observe that nearly all observations (99.7% for $\langle \phi, \psi \rangle$ and 98.8% for $\langle \chi_1, \chi_2, \ldots \rangle$) arise in cells that are ±1 distance from the main diagonal. This highlights the disagreement arising from observations being assigned to adjacent categories by the two systems: outlier⟷ allowed or allowed⟷favored.

Examining the compression statistics for observations that fall in ±1 off-diagonal cells, we observe that a significant proportion of them are close to the compression-based membership-boundaries defined by PhiSiCal-Checkup. This can be qualitatively visualized in the amino acid–specific contour plots for $\langle \phi, \psi \rangle$ shown in ref. 26. More quantitatively, Table 2 *C* and *D* demonstrate the effect of perturbation (in the direction of the gradient) on memberships (*Materials and Methods*). Each cell tracks the difference in the number of observations after perturbation compared to the raw counts shown in Tables 2 *A* and *B*. For $\langle \phi, \psi \rangle$ observations, ∼38% of the observations previously assigned to the outlier category by PhiSiCal-Checkup (first column) change membership and move into the allowed category (second column). Next, ∼40% of those previously classified as allowed by PhiSiCal-Checkup (second column), change membership and move into the favored category (third column). A similar trend is observed for $\langle \chi_1, \chi_2, \ldots \rangle$ observations, with ∼42% moving from outlier to allowed, and ∼44% moving from allowed to favored. This demonstrates the inherent limitation of using hard (non-overlapping) memberships, especially considering the uncertainty implicit in any statement of coordinates. To overcome this, PhiSiCal-Checkup alerts users of conformations with overlapping memberships using its gradient information.

Further, examining the sources of differences at the level of individual amino-acid types, other limitations in the state of the art come to the fore, enumerated below (also refer to refs. 24–26):

1. For $\langle \phi, \psi \rangle$ observations, MolProbity employs the same "general" model for 16 (non-{Glycine, Valine, Isoleucine, Proline}) amino-acid types. This ignores noticeable variations in the empirical distributions of $\langle \phi, \psi \rangle$ for these 16 amino acids, clearly observable from their contour plots (27). For this reason, across all evaluations, PhiSiCal-Checkup employs amino acid–specific probability distributions which model the empirical variations more accurately.
2. Specifically for Proline $\langle \phi, \psi \rangle$ observations, MolProbity infers two additional normalized-functions after grouping the Proline data in the reference dataset into two coarse bins, based on their observed $\omega$ value: $\omega \in [-30°, +30°]$ and $\omega \in ([-180°, -150°) \vee (+150°, +180°))$ (9). In contrast, PhiSiCal-Checkup employs formal and continuous $\langle \phi, \psi \rangle | \omega$ conditional probability distributions that are more accurate than the discretized (cis-only and trans-only) models of MolProbity.
3. For $\langle \chi_1, \chi_2, \ldots \rangle$ observations, no clear pattern emerges, with the differences spreading across all amino-acid types. As a proportion of each amino acid's number of observations, the differences between PhiSiCal-Checkup and MolProbity vary from ∼2 to 3% on the lower side (Tyrosine, Phenylalanine, Leucine, Isoleucine, and Proline) to ∼7 to 8% on the higher (Glutamic acid, Histidine, Asparagine, Lysine, Serine, Cysteine, Methionine). The remaining amino acids (Aspartic acid, Glutamine, Tryptophan, Valine, Threonine, Arginine) differ between $[4, 6]$%. We observe that the PhiSiCal-Checkup's mixture models fit the empirical distributions accurately (19). On the other hand, as the number of dihedral angles in the side chain increases, MolProbity uses increasingly coarse grid sizes and variable smoothing parameters to fit side chain-specific functions, contributing to the observed differences.

**Concluding Remarks and Future Direction.** The results presented above demonstrate the advances PhiSiCal-Checkup achieves to enable a comprehensive, consistent, and accurate evaluation of amino acid conformations of experimental protein coordinates. Where previously only $\langle \phi, \psi \rangle$ and (independently) $\langle \chi_1, \chi_2, \ldots \rangle$ observations could be validated, to varying consistency and accuracy, the current Bayesian method of PhiSiCal-Checkup supports amino-acid specific validation of dihedral angles in any combination of terms (conditional or otherwise) using strictly formal and mutually consistent statistical models, far-outstripping the scope of investigations currently possible. Further, the information-theoretic measure of 'favorability' along with the use of mathematical gradients to enable sensitivity analyses, allows experimentalists to quantify reliably the degree of surprise of any amino acid conformation while accounting for the uncertainty implicit in the protein coordinates.

Beyond these features, a significant effort has been directed toward engineering a configurable software that implements PhiSiCal-Checkup for immediate practical use by experimentalists. This is downloadable both as an open-source program written in C++ (for offline use), and as a web-server (for online use): https://lcb.infotech.monash.edu.au/phisical/checkup.

Several extensions to PhSiCal-Checkup are planned, earmarked as future work. Extending the current statistical models that analyze and assess individual amino-acid conformations, more generalized amino acid–specific statistical models are being constructed to permit assessment of short oligopeptide ($k \geq 1$-mer) conformations and their spatially interacting $k$-mer ensembles. The most basic in this line of planned extensions comes in the form of statistical models for pairs of interacting amino acids (i.e., pair of interacting 1-mers)—these extended models will permit ways to analyze and validate covalent (e.g., disulfide bridge) and non-covalent (e.g., hydrogen (−H) bonds) interactions that currently remain overlooked despite underpinning the protein 3D structure. Another study earmarked for immediate future work is to analyze the distribution of conformational angles of protein structures not determined by experimental methods but predicted using programs such as AlphaFold 3 (22). These predicted structures are increasingly being used in research and it would be interesting to compare the distributions with those derived from experimental coordinates.

## Materials and Methods

**Statistical Mixture Models.** A statistical mixture model describes a probability distribution expressed as a convex combination (i.e., a mixture) of component probability density functions. Formally, a parametric mixture model $\mathcal{M}$ composed of a mixture containing $|\mathcal{M}|$ component probability density functions is characterized as $\mathcal{M}(x) = \sum_{i=1}^{|\mathcal{M}|} w_i f_i(x|\Theta_i)$. Here, $x$ denotes any observation, $f_i(x|\Theta_i)$ denotes the $i$-th probability density function in the mixture with parameters $\Theta_i$, and $w_i$ denotes the component-weight such that $\sum_{i=1}^{|\mathcal{M}|} w_i = 1$. In unsupervised inference, all mixture parameters $\{|\mathcal{M}|, \{w_i\}_{\forall 1 \leq i \leq |\mathcal{M}|}, \{\Theta_i\}_{\forall 1 \leq i \leq |\mathcal{M}|}\}$ have to be inferred automatically from the observed data (i.e., from a set of observations of the form $X = \{x_1, x_2, \ldots, x_n\}$).

We recently described an unsupervised method to infer amino acid–specific joint mixture models from any given source collection of protein coordinates (11). The inference method relies on the Bayesian and information-theoretic criterion of Minimum Message Length (MML) (12, 13). Each inferred amino acid–specific mixture model describes a continuous joint probability distribution over its (vector of) dihedral angles terms ($\langle \omega, \phi, \psi, \chi_1, \chi_2, \ldots \rangle$). Each term in the vector is a continuous random variable in the range $(-180°, +180°]$.

For an amino acid type `aa` with $d$ dihedral angle terms, its corresponding joint mixture model $\mathcal{M}^{(\mathtt{aa})}_{(\text{joint})}(\omega, \phi, \psi, \chi_1, \chi_2, \ldots)$ defines a probability distribution on a wrapped multidimensional $d$-Torus ($\mathbb{T}^d$). In our work, each component of $\mathcal{M}^{(\mathtt{aa})}_{(\text{joint})}$ is a product of von Mises distributions, one for each dihedral angle term, thus defining a proper continuous probability density function in $\mathbb{T}^d$ space. For a set $X$ of observed vector of dihedral angles of an amino acid type `aa`, $\mathcal{M}^{(\mathtt{aa})}_{(\text{joint})}(\omega, \phi, \psi, \chi_1, \chi_2, \ldots)$ is inferred using the Bayesian criterion of MML as the mixture model that best explains $X$. (Refer to Amarasinghe et al. (11) for details of the inference method.)

**Deriving Marginal Probability Distributions.** Using the axioms of probability, amino acid–specific marginal probability distributions of any proper subset of dihedral angle terms can be derived from its corresponding joint mixture model.

Formally, let $A = \{\omega, \phi, \psi, \chi_1, \chi_2, \ldots\}$ denote the set of $d$ dihedral angle terms (random variables) for the amino acid type `aa`. Let $B$ define any non-empty, *proper* subset of terms in $A$. Then, the marginal probability distribution of the subset $B \subset A$ can be derived from the joint probability distribution of $A$ by (contour) integrating (out of the joint distribution) all dihedral angle terms $\{z_1, \ldots z_m\}$ in $B^{\complement} = A \setminus B$. (Note, if $|A| = d$ and $|B| = 0 < d' < d$, then $m = d - d'$.)

$$\mathcal{M}^{(\mathtt{aa})}_{(\text{marg.})}(B \subset A) = \underbrace{\oint \ldots \oint}_{\forall z_i \in B^{\complement}} \mathcal{M}^{(\mathtt{aa})}_{(\text{joint})}(A)\, dz_1 \ldots dz_m$$

From the computational side, we note that the marginal distribution for any subset of dihedral angle terms in $B$ is also a mixture model. It defines a continuous probability distribution in the subspace $\mathbb{T}^{d'} \subset \mathbb{T}^d$. Because each

component of the joint distribution is a product of von Mises distributions (over dihedral angle terms), a spatial projection of the joint mixture model $\mathcal{M}^{(aa)}_{(joint)}(A)$ from $\mathbb{T}^d$ space $\rightarrow$ $\mathbb{T}^{d'<d}$ space results in the corresponding $\mathcal{M}^{(aa)}_{(marg.)}(B)$. This allows us to efficiently compute any marginal mixture model on-the-fly (in real time) taking tens of microseconds on modern standalone computers (*Results*).

**Deriving Conditional Probability Distributions.** Amino acid–specific conditional probability distributions of *any* proper *subset* of the joint dihedral angle terms, upon observing specific values of another subset of dihedral angles terms can be derived using Bayes theorem.

As introduced above, let $B \subset A$ define any non-empty *proper* subset of $A$ containing $0 < d' < d$ dihedral angle terms. Further, let $C \subseteq B^{\complement} \subset A$ denote another subset containing $0 < d'' \leq d - d'$ terms. Assume the terms (i.e. random variables) in $C$ are observed to take specific dihedral angle values $c = \{c_1, c_2, \ldots c_{d''}\}$. Then, the conditional probability distribution for the subset $B$ after observing some specific values $c$ for the terms (random variables) in $C$ can be derived as:

$$\mathcal{M}^{(aa)}_{(cond.)}(B|C=c) = \frac{\mathcal{M}^{(aa)}_{(marg.)}(B \cup C = c)}{\Pr(C=c)}.$$

In the above equation, $\mathcal{M}^{(aa)}_{(marg.)}(B \cup C = c)$ denotes a mixture model over the dihedral angle terms (random variables) in $B$ *after* assigning the dihedral angle terms $C = c$ in $\mathcal{M}^{(aa)}_{(marg.)}(B \cup C)$. Further, $\Pr(C = c) \propto \mathcal{M}^{(aa)}_{(marg.)}(C = c)$ is the marginal probability of observing the dihedral angle terms (random variables) $C = c$.

As can be seen from the equation above, a conditional mixture model is deduced from the corresponding marginal mixture models, each of which is derived as a projection from the amino acid–specific joint mixture model. Hence, as before, the computation of any conditional mixture model in any combination of terms (i.e., subsets $B$ and $C$ of set $A$) can be performed highly efficiently on standalone computers in real time (*Results*).

**Measures of Shannon Information and Compression.** From the mathematical theory of communication (16), the Shannon information content of any observation $O$ is given by the relationship, $I(O) = -\log_2(\Pr(O))$ bits, where $\Pr(O)$ is the probability of that observation drawn from some source probability distribution. In this work, an observation $O$ involves observing specific values for the dihedral angles terms (in any combination of them). In PhiSiCal-Checkup, the information content in any observation $O$ (denoted here by $I_{\text{mixture}}(O)$) is evaluated by computing its probability $\Pr_{\text{mixture}}(O)$ under its corresponding amino-acid specific joint/marginal/conditional mixture model. The precise mixture model that has to be used to compute $I_{\text{mixture}}(O)$ is fully determined by the amino acid type of the observation and the combination of dihedral terms being evaluated.

Next, the measure of *compression* for the observation $O$ can be derived by comparing the Shannon information content using the mixture model against its uncompressed raw/null bit content: $\text{Compression}(O) = I_{\text{null}}(O) - I_{\text{mixture}}(O)$ bits. Note, the uncompressed bit content is the same as measuring the Shannon information content of $O$ with the uniform distribution as its source distribution.

Applying the negative-logarithm relationship between Shannon information content and source probabilities, compression is equivalent to computing the *log-likelihood* (or log-odds) of the observation $O$ arising from each of the two competing distributions as its source (mixture vs. uniform null): $\text{Compression}(O) = \log_2\left(\frac{\Pr_{\text{mixture}}(O)}{\Pr_{\text{null}}(O)}\right)$ bits. Thus, if an observation $O$ results in $+n$ bits of compression (i.e., gain of $n$ bits), the odds are $2^n$:1 in favor of $O$ arising from the mixture model. Conversely, if it results in $-n$ bits of compression (i.e., loss of $n$ bits), the odds are $1{:}2^n$ against the observation $O$ arising from the mixture model.

**Gradient.** All joint, marginal, and conditional amino acid–specific mixture models are continuous distributions that are differentiable at every point in their respective toroidal support. For example, the gradient vector for $\mathcal{M}^{(aa)}_{(joint)}(A)$ at any point $a = \{a_1, a_2, \ldots a_d\} \in \mathbb{T}^d$, is denoted by

$$\nabla\mathcal{M}^{(aa)}_{(joint)}(A=a) = \left[\frac{\partial\mathcal{M}^{(aa)}_{(joint)}(A=a)}{\partial a_1}, \frac{\partial\mathcal{M}^{(aa)}_{(joint)}(A=a)}{\partial a_2}, \ldots, \frac{\partial\mathcal{M}^{(aa)}_{(joint)}(A=a)}{\partial a_d}\right]^{\mathsf{T}}$$

For the mixture models involving the product of von Mises distributions, the result is in a closed mathematical form that can be computed on the fly. We note that the gradients for marginal and conditional mixture models yield similar expressions and characteristics since they also manifest as mixture models (except they are represented in the reduced dimensions $d'$ of the subset $B \subset A$).

Thus, for any (multidimensional) point $p$, its gradient vector ($\vec{v} = \nabla\mathcal{M}(p)$) gives the magnitude and direction of steepest ascent at $p$ in its corresponding amino acid–specific probability distribution (joint/marginal/conditional mixture model $\mathcal{M}$). Using this gradient, $p$ can be perturbed to a near-neighboring point $\tilde{p} = p \pm \lambda\hat{v}$, where $\hat{v}$ gives the direction cosines of the gradient vector $\vec{v}$. In practice, PhiSiCal-Checkup chooses $\lambda = 5$ for any $p$ (for dihedral terms stated in degrees). This constrains the norm of the projection of $\hat{p} - p$ in any (dihedral angle) dimension to $\leq 5°$.

**Difference between PhiSiCal-Checkup and PhiSiCal.** The inference of joint mixture models to support validation of amino acid conformations in PhiSiCal-Checkup is derived using the previously described methodology, PhiSiCal (11). Although PhiSiCal-Checkup builds on the inference-methodology of PhiSiCal, realizing a comprehensive framework specifically for dihedral angle validation required several novel additions and extensions.

The new set of joint mixture models inferred for PhiSiCal-Checkup includes the modeling of $\omega$ dihedral angle, along with all other (main chain and side chain) dihedral angles for each amino acid type–previous work, PhiSiCal, ignored $\omega$ from its joint probability distribution. Further, the mixture models of PhiSiCal were inferred on PDB50 and PDB50HighRes structural dataset that lacked residue-level filtering. Instead, PhiSiCal-Checkup uses the "Top2018" dataset curated by Williams et al. (4). This dataset comes with residue-level filtering to ensure only the "best parts" of high-quality protein residues are considered with "good electron density support for a physically acceptable model conformation" (4). Furthermore, the factorization of joint mixture models to derive marginal and conditional mixture models, along with the use of compression and gradient information (all described above) to support accurate validation of protein coordinates are unique to PhiSiCal-Checkup.

**Data, Materials, and Software Availability.** Supplementary Figures and Tables are included in *SI Appendix*. Supplementary Data has been deposited in FigShare (15, 19, 23–27). Software for online and offline use is available from https://lcb.infotech.monash.edu.au/phisical/checkup. All other data are included in the manuscript and/or *SI Appendix*.

**ACKNOWLEDGMENTS.** We thank Monash eResearch Centre and eServices for special job allocations on Monash high-performance computing clusters that facilitated this work.

1. H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **10**, 980 (2003).
2. V. P. Waman, C. Orengo, G. J. Kleywegt, A. M. Lesk, Three-dimensional structure databases of biological macromolecules. *Methods Mol. Biol.* **2449**, 43–91 (2022).
3. S. Gore *et al.*, Validation of structures in the Protein Data Bank. *Structure* **25**, 1916–1927 (2017).
4. C. J. Williams, D. C. Richardson, J. S. Richardson, The importance of residue-level filtering and the top2018 best-parts dataset of high-quality protein residues. *Protein Sci.* **31**, 290–300 (2022).
5. R. A. Laskowski, M. W. MacArthur, D. S. Moss, J. M. Thornton, PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
6. R. W. Hooft, G. Vriend, C. Sander, E. E. Abola, Errors in protein structures. *Nature* **381**, 272 (1996).

7. T. A. Jones, J. Y. Zou, S. W. Cowan, M. Kjeldgaard, Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. Sect. A: Found. Crystallogr.* **47**, 110–119 (1991).
8. S. C. Lovell *et al*., Structure validation by C$\alpha$ geometry: $\phi$, $\psi$ and C$\beta$ deviation. *Proteins: Struct., Funct., Bioinf.* **50**, 437–450 (2003).
9. R. J. Read *et al*., A new generation of crystallographic validation tools for the Protein Data Bank. *Structure* **19**, 1395–1412 (2011).
10. B. J. Hintze, S. M. Lewis, J. S. Richardson, D. C. Richardson, Molprobity's ultimate rotamer-library distributions for model validation. *Proteins: Struct., Funct., Bioinf.* **84**, 1177–1189 (2016).
11. P. R. Amarasinghe *et al*., Getting '$\phi\psi\chi$al' with proteins: Minimum message length inference of joint distributions of backbone and sidechain dihedral angles. *Bioinformatics* **39**, i357–i367 (2023).
12. C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length* (Springer, 2005).
13. L. Allison, *Coding Ockham's Razor* (Springer, 2018).
14. T. Bayes, An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR. *Philos. Trans. R. Soc. London* **53**, 370–418 (1763).
15. P. R. Amarasinghe *et al*., Supplementary Data 1. FigShare. https://figshare.com/s/0b349f998f795c45b109. Deposited 21 October 2024.
16. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
17. E. Vandermarliere *et al*., Structural analysis of a glycoside hydrolase family 43 arabinoxylan arabinofuranohydrolase in complex with xylotetraose reveals a different binding mechanism compared with other members of the same family. *Biochem. J.* **418**, 39–47 (2009).
18. O. Carugo, P. Argos, Correlation between side chain mobility and conformation in protein structures. *Protein Eng.* **10**, 777–787 (1997).
19. P. R. Amarasinghe *et al*., Supplementary Data 2. FigShare. https://figshare.com/s/1ee14426c6d49c7cbd2a. Deposited 21 October 2024.
20. H. K. Ganguly, G. Basu, Conformational landscape of substituted prolines. *Biophys. Rev.* **12**, 25–39 (2020).
21. A. Kale, T. Pijning, T. Sonke, B. W. Dijkstra, A. M. W. Thunnissen, Crystal structure of the leucine aminopeptidase from *Pseudomonas putida* reveals the molecular basis for its enantioselectivity and broad substrate specificity. *J. Mol. Biol.* **398**, 703–714 (2010).
22. J. Abramson *et al*., Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
23. P. R. Amarasinghe *et al*., Supplementary Data 3. FigS22hare. https://figshare.com/s/bda9c52606eb07780686. Deposited 21 October 2024.
24. P. R. Amarasinghe *et al*., Supplementary Data 4. FigShare. https://figshare.com/s/d65b9522eff486a5a5d4. Deposited 21 October 2024.
25. P. R. Amarasinghe *et al*., Supplementary Data 5. FigShare. https://figshare.com/s/b7bd8664f226dbb3d005. Deposited 21 October 2024.
26. P. R. Amarasinghe *et al*., Supplementary Data 6. FigShare. https://figshare.com/s/5e3ee64c43d69b21c88f. Deposited 21 October 2024.
27. P. R. Amarasinghe *et al*., Supplementary Data 7. FigShare. https://figshare.com/s/aebee230df911e48cb34. Deposited 21 October 2024.