Reviewed Preprint
v1 • May 14, 2026
Not revised

For correspondence:
zhang@nus.edu.sg

* These authors contributed equally

Competing interests: No competing interests declared

Funding: See page 10

Reviewing editor: Aaron Frank, Arrakis Therapeutics, United States

ProteinConformers: large-scale and energetically profiled descriptions of protein conformational landscapes

Yihang Zhou[1,2,*], Chen Wei[2,4,*], Minghao Sun[2], Lin Wang[6], Jin Song[7,8], Fanding Xu[2,9], Yang Li[1], Wei Zheng[5], Yang Zhang[1,2,3,6]

[1]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore • [2]School of Computing, National University of Singapore, Singapore, Singapore • [3]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore • [4]School of Economics and Management, Xi'an University of Posts & Telecommunications, Xi'an, China • [5]NITFID, School of Statistics and Data Science, AAIS, LPMC, and KLMDASR, Nankai University, Tianjin, China • [6]Center for AI and Computational Biology, Institute of Systems Medicine, Chinese Academy of Medical Sciences, Beijing, China • [7]School of Advanced Interdisciplinary Science, University of Chinese Academy of Sciences, Beijing, China • [8]State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China • [9]School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, China

eLife Assessment

This study presents a **useful** database resource containing protein conformations generated through molecular dynamics simulations, with extensive quality evaluation and benchmarking. While the database is well-constructed and professionally organized, the evidence supporting its claimed representation of protein conformational landscapes is **incomplete**, as the short simulation times and starting structure bias prevent true Boltzmann sampling of the conformational space.

https://doi.org/10.7554/eLife.110874.1.sa3

Abstract

Modeling protein conformational landscapes is essential for understanding dynamics, allostery, and drug discovery, yet existing resources lack diverse conformational coverage, energetic annotations, or benchmarking standards. ProteinConformers (*https://zhanggroup.org/ProteinConformers*) provides 2.7 million geometry-optimized conformations generated with a multi-seed molecular dynamics strategy, paired with 13.7 million energy evaluations and 5.5 million similarity annotations. It delivers continuous landscapes from non-native to near-native states, benchmarking framework for multi-conformation generators, and an interactive analysis platform.

Main

Understanding protein function requires capturing how structures dynamically interconvert across their conformational landscapes. Many proteins undergo allosteric and functionally critical transitions that occur at atomic resolution, and these motions are mostly governed by the underlying thermodynamic energy landscape[1-5]. Existing efforts to characterize protein conformational spaces fall into three broad strategies. Fragment-assembly methods construct decoy structures using Monte Carlo sampling[6-8]. Molecular dynamics (MD) simulations can generate conformations along trajectories initiated from experimentally resolved structures[9-13]. Ensemble extensions of deep learning based static structure predictors[14,15] can generate multiple conformers. Despite these advances, current datasets and samplers exhibit major limitations: MD-

based sampling is constrained by starting from native structures near the global energy minimum, existing conformer generators lack standardized benchmarks for assessing geometric plausibility and landscape-wide diversity, and available datasets provide limited energetic and structural similarity annotations, leaving the coupling between conformational variability and energetics insufficiently characterized.

To address these gaps, we developed ProteinConformers, a large-scale, energetically annotated resource of protein conformational landscapes. Using a multi-seed decoy sampling strategy that initiates MD simulations from hundreds of diverse starting conformations per protein, we generated over 2.7 million physically plausible structures spanning the full spectrum from non-native to near-native states, accompanied by 13.7 million energy evaluations and 5.5 million similarity annotations. ProteinConformers samples substantially broader conformational landscapes per protein than ATLAS database [11], while achieving atomic-level local plausibility comparable to the high-quality Top2018 dataset [16]. Moreover, it enables systematic benchmarking of multi-conformer generators through our proposed evaluation framework based on the curated ProteinConformers-lite dataset. A user-friendly web platform supports querying, visualization, analysis, and data access for the community.

ProteinConformers offers an extensive characterization of structural and energetic landscape diversity across 734 proteins. The conformational ensembles span diverse conformation status (Fig. 1A) with a wide range of sequence lengths from 33 to 949 residues and an average of 247 residues (Fig. 1B-C). For each protein, we computed TM-score and RMSD between all sampled conformations and their native structures to quantify the global geometric plausibility and measure the non-native to near-native conformational spectrum (Fig. S1A). Energetic profiles were generated using five widely used statistical and physics-based energy functions, RW [17], RWplus[17], EvoEF2[18], Rosetta[19], and FoldX[20] (Fig. S1B). Structural classification using ECOD[21] and CATH[22] annotations further confirms that the dataset spans diverse protein fold families (Fig. S2).

We constructed ProteinConformers-lite, a curated benchmark subset of the ProteinConformers dataset containing 381,546 MD refined conformers across 87 CASP14 and CASP15 proteins (Fig. S3). These targets were selected because their seed decoys are generally higher quality and the underlying structures more challenging. Unlike previous benchmarks centered on near-native fluctuations around PDB structures, ProteinConformers-lite spans a broader conformational spectrum than the ATLAS dataset, with proteins distributed across non-native to near-native regions (Fig. 1D) and exhibiting a wider per-protein TM-score range (Fig. S4). Near native states were defined as those with TM-score greater than 0.5, a threshold indicating preservation of global fold integrity [23].

We evaluated the physical plausibility of ProteinConformers-lite by comparing its geometric properties with the high-quality Top2018 reference dataset. The distributions of dihedral angles and bond lengths in ProteinConformers-lite closely matched those in the Top2018 dataset (Fig. 1E), demonstrating that the local geometry of our benchmark achieves consistent stereochemical quality. Analysis of Ramachandran outlier rates revealed a consistent decrease as TM-score increased from non-native to near-native values, eventually falling below the average rate observed in the Top2018 dataset (13%). This trend indicates that near-native conformations in ProteinConformers-lite exhibit comparable stereochemical quality with Top2018 dataset (Fig. 1F).

A dual-axis benchmark framework was established to assess the diversity of the protein conformational landscape by calculating the coverage rate of the generated conformers to the benchmark dataset (Materials and Methods, Fig. S5), and to quantify residue-pair geometric statistics across distance and orientation features via Conformation Geometry Map (CGM) and its similarity score $CGMS^{cos}$ and $CGMS^{mah}$ (Materials and Methods, Fig. S6).

Then we systematically benchmarked five representative models ($AlphaFlow^{Dis}_{MD}$ [15], $AlphaFlow^{Dis}_{PDB}$ [15], $ESMFlow^{Dis}_{MD}$ [15], $ESMFlow^{Dis}_{PDB}$ [15], and BioEmu [24]) using our framework and ProteinConformers-lite dataset. Diversity analysis shows that BioEmu achieves the highest coverage under strict 5 kJ/mol thresholds, indicating effective sampling of low-energy

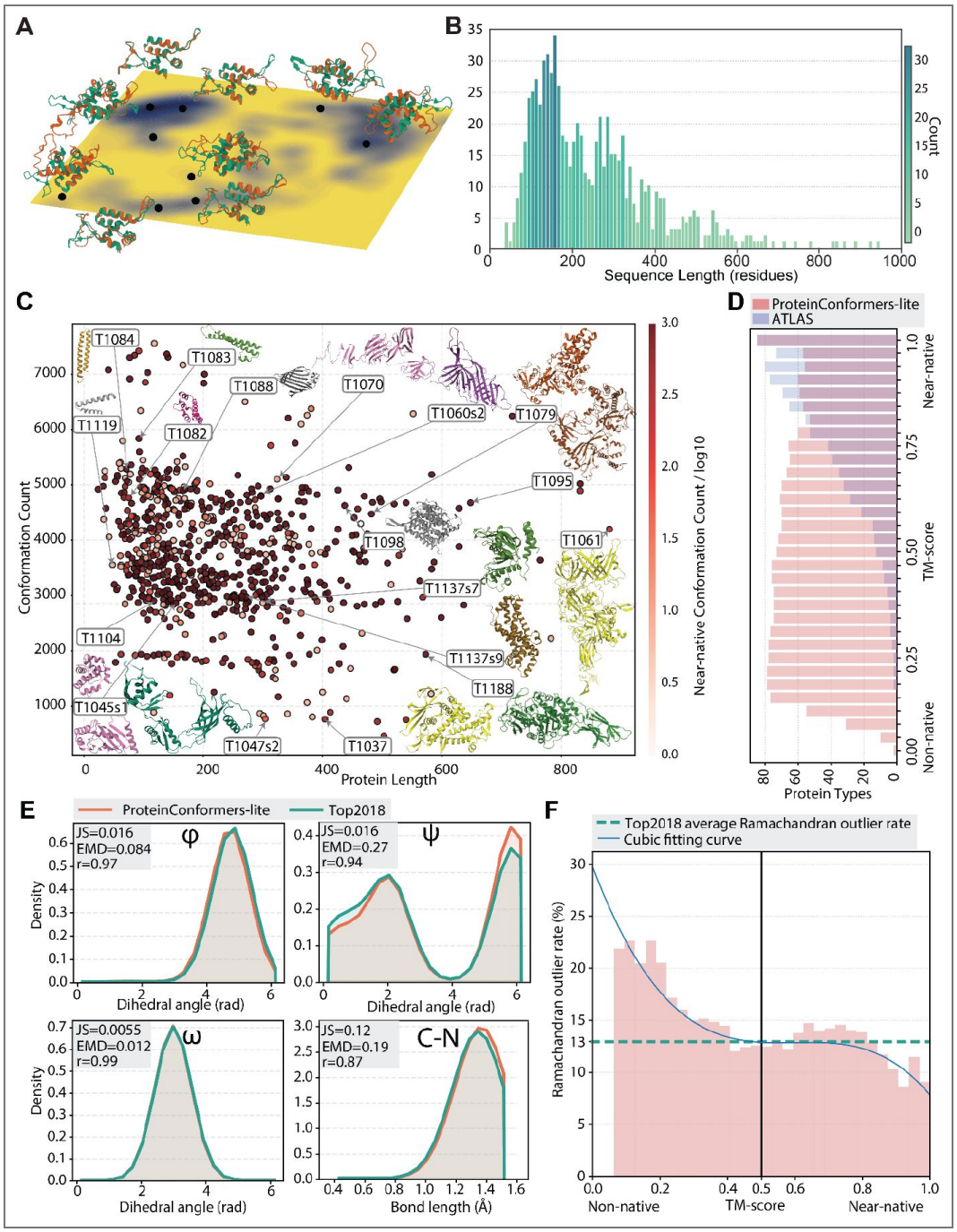

Figure 1. Overview and quality assessment of the ProteinConformers and ProteinConformers-lite datasets.

(A) Schematic conformational energy landscape with representative conformers positioned in distinct minima of T1035, illustrating the breadth of physically realistic states captured by ProteinConformers. (B) Sequence-length distribution of ProteinConformers. (C) Conformer counts per protein, sorted by sequence length; points are colored by the log10 of near-native (TM-score ⩾ 0.5) conformation counts, with representative structures shown as insets. (D) TM-score coverage across proteins from ProteinConformers-lite and sampled ATLAS dataset, binned into 32 equal-width intervals from non-native to near-native. (E) Comparison of φ, ψ, ω dihedral angle and C–N bond length distributions between ProteinConformers-lite and Top2018 dataset, quantified by Jensen–Shannon divergence (JS), Earth Mover's Distance (EMD), and Pearson correlation coefficient (r), demonstrating consistent local stereochemical quality. (F) Ramachandran outlier rates were averaged within 32 equal-width TM-score bins. The mean outlier rate of Top2018 dataset (13%, green dashed line) and the TM-score threshold of 0.5 separating non-native and near-native states are shown, with a cubic fit (blue curve) summarizing the trend.

regions, whereas distilled variants such as $\text{AlphaFlow}^{\text{Dis}}_{\text{MD}}$ and $\text{ESMFlow}^{\text{Dis}}_{\text{PDB}}$ display consistently reduced coverage, reflecting narrower exploration around basins (Table S1). Plausibility of CGMS^{cos} emphasizes directional agreement in residue-pair statistics. The results show high similarity on distance features but substantial degradation on orientation features, indicating that current models generally recover distance distributions better than orientation statistics. CGMS^{mah} instead evaluates global geometric consistency by measuring Mahalanobis distances in the joint covariance space and converting them into similarity scores using a Gaussian kernel, resulting in more balanced assessment across feature types. BioEmu attains the highest scores on distance components under both metrics, reflecting stronger geometric modeling of inter-residue separations. Under CGMS^{mah}, $\text{AlphaFlow}^{\text{Dis}}_{\text{MD}}$ achieves scores comparable to BioEmu, suggesting similar overall geometric plausibility. In general, models fine-tuned on MD data ($\text{AlphaFlow}^{\text{Dis}}_{\text{MD}}$ and $\text{ESMFlow}^{\text{Dis}}_{\text{MD}}$) show modest improvements in CGMS^{cos} and comparable performance in CGMS^{mah} relative to their non–MD-tuned counterparts, indicating that MD refinements provide limited gains in local geometric realism (Table S2).

To enable efficient exploration of ProteinConformers, we developed an interactive web portal integrating query, visualization, and data access functions. The homepage provides a searchable table of all 734 proteins with multi field filters, linking to detailed dashboards for each target (Fig. 2A). Dashboards offer interactive 3D visualization with real time alignment between native and decoy structures (Fig. 2B), as well as native distance and rotation maps for global comparison (Fig. 2C). Each protein includes a fully annotated decoy table with geometric, similarity, energetic, and secondary structure metrics (Fig. 2D). Users can perform dynamic, in-browser analysis of the conformational landscape by filtering decoys based on structural similarity or energetic thresholds, with all statistical summaries and visualizations updating in real time. Flexible data export is supported through customizable metric selection (Fig. 2E), and one-click bulk downloads provide complete per-protein datasets with associated metadata (Fig. 2F). The portal enables rapid, interactive analysis without local computation.

Taken together, ProteinConformers establishes a comprehensive, energetically annotated view of protein conformational landscapes, enabling systematic assessment of structural diversity and physical plausibility across modern multi-conformer generators. By integrating large-scale multi-seed sampling, detailed energetic profiling, and a unified evaluation framework, it provides a rigorous foundation for studying protein dynamics and allosteric mechanisms. Together with an interactive web platform for exploration and data access, ProteinConformers is positioned to support next-generation advances in protein conformation ensembles prediction, biomolecular modeling, and computational drug discovery.

Materials and methods

ProteinConformers Preparation

We manually downloaded all targets and corresponding predicted models (as decoys) submitted by participating groups worldwide from the Critical Assessment of protein Structure Prediction (CASP)[25] across seasons 5 to 15. At the time of this data preparation, CASP16 predictions were not available. We developed and applied a comprehensive pipeline to remove redundant and erroneous structures and to complete incomplete decoys according to the following steps.

First, we manually retrieved, decompressed, matched, and initially filtered all sequence and structure files from the CASP repository. We then matched the sequence with structure filenames by string-based file name comparison. For structure files with identical names, we appended suffixes derived from their parent folders to distinguish. We then applied SeqKit[26] to remove duplicate sequences. Finally, we manually inspected each pair. This process removed duplicate accessions, non-protein entries, and redundant sequences and associated predictions.

Second, we curated the structural files. Submissions from CASP teams frequently contained errors such as missing residues, incorrect residue ordering or numbering, missing atoms, unrecognized atom types, and erroneous bonding. Similar issues were also observed in the native PDB files

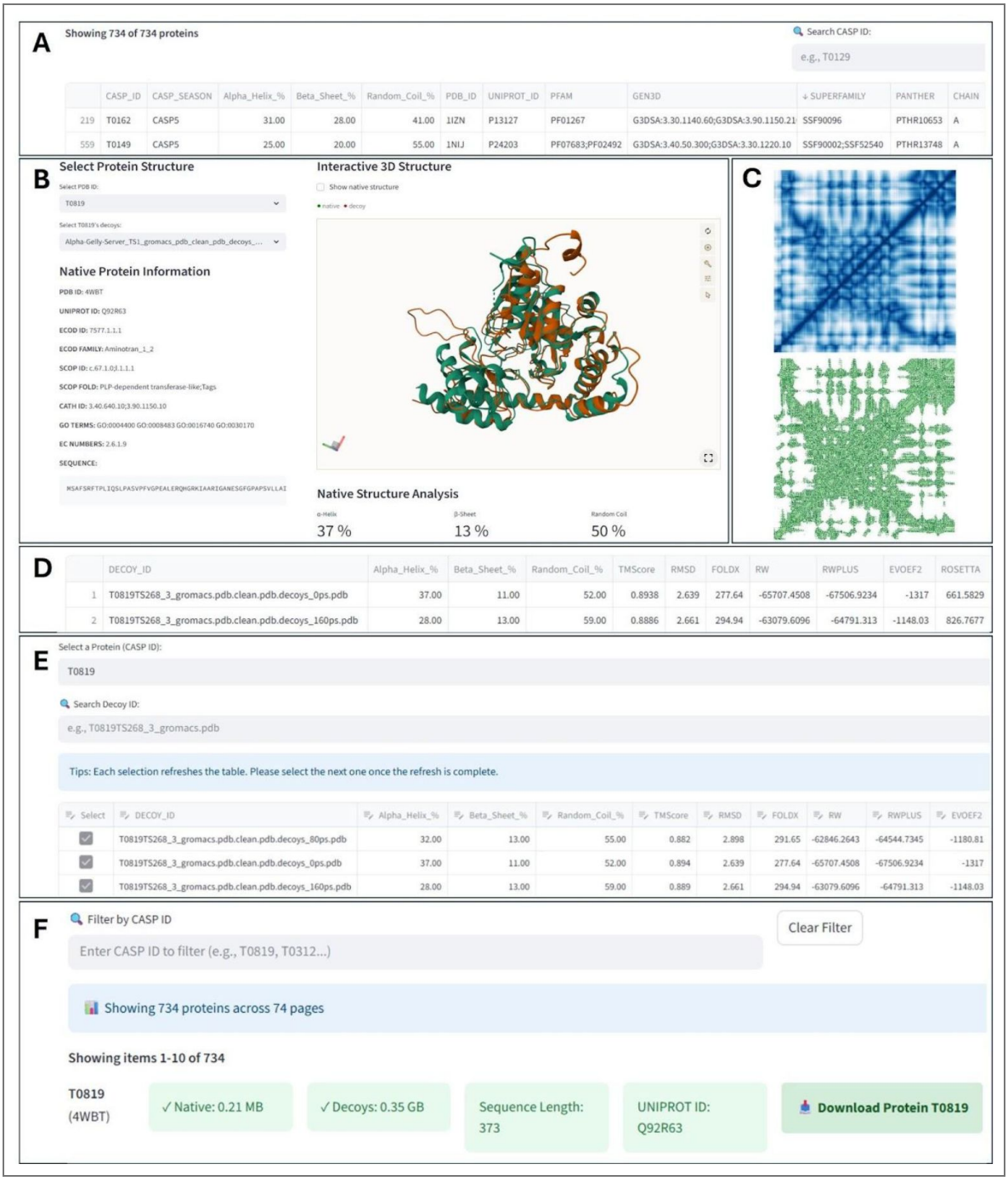

Figure 2. Interactive web interface of the ProteinConformers dataset.

(A) Main overview table displaying all 734 proteins in the dataset, with sortable columns for structural metadata and CASP-related annotations. (B) Dashboard view for a selected target (e.g. T0819), showing an interactive 3D alignment of the native structure (green) with a selected decoy (orange), along with basic protein metadata and secondary structure composition. (C) Visualization of the native structure's distance map (top) and orientation map (bottom), supporting global structural comparison. (D) Detailed table listing decoy models associated with the selected target, including secondary structure content, similarity scores, and energetic profiles. (E) Interface for selecting and filtering decoys by structural or energetic criteria. Selected decoys can be downloaded as a customized dataset. (F) Download panel summarizing metadata for the selected protein, including file sizes and sequence length, and providing options for downloading either the native structure or the full conformer set.

provided by CASP, together with inconsistencies between the reference sequence files and the sequences extracted from the native structures, which is often due to limited resolution in certain regions of the experimentally determined structures. To address these issues, we used the sequence files as the reference and performed a clean process. For each prediction, chains were extracted from structure files using Biopython [27] and aligned to the sequence reference. The alignments were classified into three mutually exclusive categories: *same* (full-length with identical residue order, for example, reference sequence ABCD vs. extracted sequence ABCD), *disorder* (a contiguous subsequence of the reference caused by order-preserved internal or terminal deletions, for example, reference ABCD vs. extracted ABD), and *mismatch* (any substitution, insertion, or reordered segment, for example, reference ABCD vs. BACD or ABED). Missing atoms were rebuilt with OpenBabel [28], while structures with unrecognized atom types or bonding errors were excluded. Oligomeric or hetero-complex models were excluded from the dataset. Based on the categories from the previous step, the *same* models were accepted without modification. The *disorder* models were retained only if their residue numbers matched the reference structure; otherwise, the coordinates would be automatically renumbered to restore one-to-one correspondence. All *mismatch* entries were discarded.

Third, when multiple experimental native structures were available for a target, we manually selected a single representative. Selection criteria included secondary validation in the PDB, sequence length, and structural resolution.

Fourth, additional filtering was performed based on data availability. A target was included in the benchmark only if at least one native structure and 100 decoys were available. For targets below this threshold, additional decoys would be generated using 3DRobot [8]. Targets that failed to obtain sufficient decoys were removed.

After these steps, all native structures and decoys underwent further physics-based optimization through MD simulations. Structures that crashed during MD process were abandoned.

Molecular Dynamics Protocol

We developed and applied a full-atom molecular dynamics simulation pipeline to optimize all conformations at atomic resolution, and all conformations that failed to converge during this process were excluded.

Atomic MD simulations are performed using GROMACS 2023 [29]. Each protein conformer follows the same workflow. Topology construction and solvation: The OPLS-AA force field is used for topology generation, together with the TIP3P water model. Each protein is centered in a dodecahedral box, and the box is filled with pre-equilibrated SPC216 water. Na^+ and Cl^- ions are added to neutralize the total charge. Steepest-descent minimization is applied until the largest force on any atom falls below 1000 kJ $mol^{-1}nm^{-1}$ (maximum 50,000 steps), thereby minimizing the energy, eliminating steric clashes and unrealistic geometries. The NVT phase (100 ps, 300 K) employed the V-rescale thermostat with positional restraints on heavy atoms to stabilize temperature. The subsequent NPT phase (100 ps, 1 bar) uses the Parrinello-Rahman barostat after partially releasing restraints to equilibrate density. Periodic-boundary artifacts are eliminated by recentering and re-imaging the trajectory. Restraints are removed, and a simulation between 125 ps to 375 ps is executed for simulation. Snapshots are extracted every 25 ps.

All the simulations are executed on high-performance computers equipped with AMD EPYC 7763 (64-core, 2.45 GHz) processors.

Energetic and Similarity Profiles

We annotated the energetic profile for each conformation using RW [17], RWplus [17], EvoEF2 [18], Rosetta (version: REF2015) [19], and FoldX (version: 20241231) [20].

We calculated pairwise conformational similarity metrics between each decoy to the native structure, including TM-score [23,30,31] and RMSD [32].

ProteinConformers Dataset statistics

ProteinConformers includes 734 native structures and 2,750,261 optimized atomistic conformations generated from 352,146 seed decoys. On average, each protein runs MD simulation from 480 different initial structures and has 3,747 conformations. Each conformation has 5 energetic profiles and 2 similarity metrics, which in total result in 13,751,305 energetic profiles and 5,501,990 similarity annotations. The 734 distinct proteins range in sequence length from 33 to 949 residues, and an average length of 247 residues. The aggregate cost is approximately 40 million CPU hours.

Benchmark Dataset

To highlight the capabilities of the ProteinConformer dataset and to facilitate systematic evaluation of existing protein conformation generative models, we developed a compact benchmarking suite, referred to as ProteinConformers-lite. The conformational landscapes in this dataset were collected from ProteinConformers, including conformers from CASP14 and CASP15. We selected the most recent two CASP seasons because their seed decoys are generally of higher quality and the corresponding targets are overall more challenging.

The ProteinConformers-lite benchmark dataset comprises 87 proteins with an average sequence length of 305 residues, a median of 255 residues, and a maximum length of 949 residues (Fig. S1). These 87 proteins include 40,387 seed decoys, 381,546 conformers, and 1,907,730 energetic annotations.

We depolyed $\text{AlphaFlow}^{\text{Dis}}_{\text{MD}}$ [15], $\text{AlphaFlow}^{\text{Dis}}_{\text{PDB}}$ [15], $\text{ESMFlow}^{\text{Dis}}_{\text{MD}}$ [15], $\text{ESMFlow}^{\text{Dis}}_{\text{PDB}}$ [15], and BioEmu [24] followed their official tutorials. For every software, we generated 3,000 conformations for each protein in ProteinConformers-lite, and assessed the diversity and plausibility of the generation using our metrics.

Top2018 dataset comparison

The Top2018 dataset is a high-quality curated reference collection of protein residues for plausibility reference [16]. We downloaded the complete Top2018 dataset, which contains 8,307 high-quality, low-redundancy protein chains. Ramachandran outlier rates were computed using PyRosetta [33] with the *Ramachandran* scoring function, where *phipsi_in_forbidden_rama* was applied to determine whether each residue dihedral angle falls into a disallowed region, excluding terminal residues. For bond length and bond angle statistics, we used Biopython [27] to calculate arithmetic means of the main chain bond lengths. Dihedral angles were first mapped onto the [0, 2π) interval and then averaged using the circular mean, which include ϕ ($C_{(i-1)}$-$N_{(i)}$-$C\alpha_{(i)}$-$C_{(i)}$), ψ ($N_{(i)}$-$C\alpha_{(i)}$-$C_{(i)}$-$N_{(i+1)}$), and ω ($C\alpha_{(i)}$-$C_{(i)}$-$N_{(i+1)}$-$C\alpha_{(i+1)}$).

ATLAS dataset comparison

The ATLAS dataset provides MD trajectories generated from a single initial conformation per protein. To enable a direct comparison of conformational diversity between ATLAS and ProteinConformers, we randomly selected 87 proteins from ATLAS (Table S3) as the same number of proteins in ProteinConformers-lite. For each target, trajectory files were downloaded and conformations were extracted following the official ATLAS processing protocol, yielding 30,000 conformations per protein. TM-scores were computed using the corresponding processed native structures as references.

Diversity Evaluation Metrics

To quantitatively benchmark generative models against this reference set, we adopt an energy-threshold-based overlap analysis over discretized 2D free energy landscapes. In this study, we adopt the dimensionality reduction technique Principal Component Analysis (PCA) to project the high-dimensional conformational data onto a low-dimensional space. Each conformer's energy is projected into a reduced-dimensional free energy landscape via PCA, followed by Boltzmann-weighting to estimate the free energy of each state on this landscape by weighting the contribution

of each conformation according to its probability as determined by the Boltzmann distribution. For a given energy threshold τ, which is a cutoff value used to determine which protein structures (conformations) are considered stable and realistic enough to be included in the analysis, we evaluate three complementary metrics between a reference ensemble A (ProteinConformers-lite) and a generated ensemble B, including *Interaction*, *Coverage*, and *Jaccard Index*.

Interaction measures the absolute number of shared low-energy regions:

$$\text{Interaction} = |A \cap B| = \sum_{i,j} \mathbf{1}\left[A_{i,j} < \tau \wedge B_{i,j} < \tau\right] \tag{1}$$

where *i*, *j* are the indices of the grid of the Boltzmann-weighted estimation of the energy landscape.

Coverage indicates the fraction of ProteinConformers-lite's low-energy regions recovered by the generative model:

$$\text{Coverage} = \frac{|A \cap B|}{|A|} \tag{2}$$

Jaccard Index provides a symmetric measure of shared low-energy regions:

$$\text{Jaccard Index} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3}$$

These metrics are evaluated under increasingly relaxed energy thresholds (5, 10, 20 kJ/mol), allowing assessment of both strict and broader sampling capabilities. The detailed comparison results are provided in Table S1.

Plausibility Evaluation Metrics

To evaluate the realism of generated ensembles, we propose the Conformation Geometry Map (CGM) and its derived Conformation Geometry Map Similarity (CGMS).

CGM Construction

Given a protein of length N and an ensemble of M conformations, $C = \left\{C^{(m)}\right\}_{m=1}^{\mathrm{M}}$, we compute four inter-residue geometric features $G \in [D, \Omega, \Theta, \Phi]$ for each conformation, following definitions from trRosetta [34,35] and DeepPotential [36] (Table 1).

Symbol	Type	Atoms Involved	Symmetry	Description
D	Euclidean Distance	$C_{\alpha,i}, C_{\alpha,j}$	Symmetric	Distance ($D_{ij} = D_{ji}$)
Ω	Dihedral Angle	$C_{\alpha,i}, C_{\beta,i}, C_{\beta,j}, C_{\alpha,j}$	Symmetric	Rotation about $C_{\beta,i} - C_{\beta,j}$ virtual axis
Θ	Dihedral Angle	$N_i, C_{\alpha,i}, C_{\beta,i}, C_{\beta,j}$	Asymmetric	Orientation of $C_{\beta,j}$ to residue i backbone
Φ	Planar Angle	$C_{\alpha,i}, C_{\beta,i}, C_{\beta,j}$	Asymmetric	Position of $C_{\beta,j}$ to $C_{\alpha,i} - C_{\beta,i}$ bond

Table 1. Geometric features used for CGM. For Ω, Θ, and Φ, a pseudo C_β is used for glycine.

For each residue pair (*i*, *j*), we calculate the mean (μ), standard deviation (σ), and skewness (γ) of each geometric feature across the ensemble. The calculation of angle features considers circularity. The resulting feature tensor is $P^G \in R^{3\times N\times N}$, and the full CGM is the concatenation P =

$[P^D, P^\Omega, P^\Theta, P^\Phi] \in R^{12\times N\times N}$. The Euclidean distance 20Å is used as the cutoff to filter the residue pair for CGMS.

CGMS Computation

Computation involves two variants: $\mathrm{CGMS}^{\cos}$ and $\mathrm{CGMS}^{\mathrm{mah}}$. For a given test dataset and the reference ProteinConformers-lite ensemble, the respective CGM matrices, denoted as CGM^{Test} and CGM^{Ref} are computed under the condition that the test dataset samples the same set of proteins as the reference. For each valid residue pair (i, j), a three-dimensional vector $u = (\mu, \sigma, \gamma)$ is extracted from both maps.

For $\mathrm{CGMS}^{\cos}$, the average cosine similarity is calculated to evaluate the overall directional correspondence between the descriptor vectors, with less sensitivity to local deviations. The metric is defined as the mean similarity over all aligned residue pairs:

$$\mathrm{CGMS}^{\cos} = \frac{1}{|C|}\sum_{(i,j)\in C}\frac{u^T{}_i u_j}{\|u_i\|\,\|u_j\|} \tag{4}$$

where C denotes the set of all valid residue pairs.

$\mathrm{CGMS}^{\mathrm{mah}}$ incorporates both directional and magnitude alignment through a composite Mahalanobis distance and a Gaussian radial basis function, imposing a stronger penalty on local discrepancies. The Mahalanobis distance is defined as:

$$\mathrm{d}_M(u_i, u_j) = \sqrt{(u_i - u_j)^T \Sigma^{-1}(u_i - u_j)} \tag{5}$$

where Σ in (5) is the covariance matrix between two vectors. This distance is then transformed via a Gaussian kernel to compute the similarity:

$$\mathrm{CGMS}^{\mathrm{mah}} = \frac{1}{|C|}\sum_{(i,j)\in C}\exp\left(-\frac{1}{2}\left(d_M(u_i,u_j)^2\right)\right) \tag{6}$$

Higher CGMS indicates better agreement with the native ensemble's inter-residue geometric statistics. The detailed comparison results are provided in Table S2.

Website implementation

The ProteinConformers portal is implemented as an interactive data-driven web application built with Streamlit (v1.32), an open-source Python framework for rapid development of scientific dashboards. The Streamlit service is deployed on a dedicated backend and is embedded within the main site through HTML, allowing seamless integration of dynamic content without coupling the application to the primary presentation layer. The enclosing HTML page is delivered by an Apache HTTP Server, thereby decoupling the interactive visualization engine from the static web infrastructure while preserving performance and security best practices.

The application architecture incorporates a multi-tab layout, implemented using the streamlit-option-menu library, which allows users to navigate between sections such as Home, Data Table, Structure Browser, Download Center, and About. Each section is backed by efficient data caching via @st.cache_data decorators, which reduce redundant I/O and enable fast stateful browsing across user sessions. The Home tab offers summary metrics and structural composition plots; the Table and Download tabs enable rich, multi-parameter filtering of over 2.7 million conformations using cross-linked annotations (CATH, SCOPe, ECOD, GO terms, EC numbers, etc.); while the Browser tab features an interactive 3D visualization module powered by the streamlit-molstar extension [37] for PDB file rendering, and it providing users with an interactive tool to inspect, rotate, and analyze protein conformers in detail. The interactive 2D charts, such as scatter plots and heatmaps, are rendered using the Plotly [38], allowing users to dynamically explore the data.

The portal supports real-time filtering through a combination of pandas [39] operations and Streamlit widgets, with built-in pagination to handle large tables and performance-optimized rendering for charts. Extensive custom CSS is injected to declutter the UI and suppress non-essential components such as default toolbars and deploy buttons, enhancing the user experience. Additionally, protein-specific decoy ensembles, native structures, and energy landscapes are integrated on demand through efficient file I/O and optional download modules, which estimate data size and dynamically build user-selected archives. The entire application is deployed on a web server to ensure stable and public accessibility for the research community.

Data availability

All processed data in ProteinConformers is freely accessible at *https://zhanggroup.org/ProteinConformers* without registration requirements. The benchmark dataset ProteinConformers-lite is also available at *https://huggingface.co/datasets/Jim990908/ProteinConformers/tree/main*. The benchmark codes and instructions are available at *https://github.com/auroua/ProteinConformers*.

Additional information

Funding

This work was supported in part by the Ministry of Education, Singapore (MOE-T1251RES2309 and MOE-T2EP20125-0014), the Agency for Science, Technology and Research (A*STAR), Singapore (IAF-PP H25J6a0034), and the National Research Foundation, Singapore (NRF-CRP33-2025-0048). Chen Wei also wants to acknowledge the financial support from the China Scholarship Council (Grant No. 202208615021) and the Humanities and Social Sciences Program of the Ministry of Education of China (24YJAZH169).

Authors' contribution

Ya.Z. contributed to the conception and design of the study; Yi.Z. generated the dataset; C.W., M.S., J.S. and F.X developed the website. Yi.Z., C.W., L.W., M.S., J.S. designed the methodology; C.W., M.S., L.W., J.S., F.X., Y.L., W.Z. contributed to the discussion and experiment design; Yi.Z., C.W., M.S., L.W., J.S., F.X., Y.L., Ya.Z. wrote the paper. All authors proofread and approved the final manuscript.

Funding

Funder	Grant reference number	Author
Ministry of Education, Singapore	MOE-T1251RES2309	Yang Zhang
Ministry of Education, Sinagpore	MOE-T2EP20125-0014	Yang Zhang
A*STAR, Singapore	IAF-PP H25J6a0034	Yang Zhang
National Research Foundation, Singapore	NRF-CRP33-2025-0048	Yang Zhang
China Scholarship Council (CSC)	202208615021	Chen Wei
Humanities and Social Sciences Program of the Ministry of Education of China	24YJAZH169	Chen Wei

Author ORCID iDs

Yihang Zhou: https://orcid.org/0000-0002-3125-7672
Yang Li: https://orcid.org/0000-0003-2480-1972
Yang Zhang: https://orcid.org/0000-0002-2739-1916

Additional files

Supplementary data. **Figure S1.** Example of similarity score and energetic annotation distributions in ProteinConformers. **Figure S2.** Distribution of fold classes in ProteinConformers. **Figure S3.** The 3D native structures of all 87 proteins in ProteinConformers-lite. **Figure S4.** Violin plot of TM-score range per protein. **Figure S5.** Example of free energy landscapes comparison from ProteinConformers-lite and generative models. **Figure S6.** Illustration of CGM and CGMS. **Table S1.** Diversity benchmark on ProteinConformers-lite under different energy thresholds. **Table S2.** Plausibility benchmark on ProteinConformers-lite. **Table S3.** IDs of sampled proteins from ATLAS

References

1. **Gunasekaran K.,** Ma B., Nussinov R. (2004) Is allostery an intrinsic property of all dynamic proteins?. *Proteins: Structure, Function, and Bioinformatics* **57**:433-443 https://doi.org/10.1002/prot.20232 | PubMed
2. **Guo J.,** Zhou H.-X. (2016) Protein allostery and conformational dynamics. *Chemical reviews* **116**:6503-6515 https://doi.org/10.1021/acs.chemrev.5b00590 | PubMed
3. **Wei G.,** Xi W., Nussinov R., Ma B. (2016) Protein ensembles: how does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell. *Chemical reviews* **116**:6516-6551 https://doi.org/10.1021/acs.chemrev.5b00562 | PubMed
4. **Noé F.,** Fischer S. (2008) Transition networks for modeling the kinetics of conformational change in macromolecules. *Current opinion in structural biology* **18**:154-162 https://doi.org/10.1016/j.sbi.2008.01.008 | PubMed
5. **Veitshans T.,** Klimov D., Thirumalai D. (1997) Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding and Design* **2**:1-22 https://doi.org/10.1016/s1359-0278(97)00002-3 | PubMed
6. **Zhang Y.** (2008) I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* **9**:1-8 https://doi.org/10.1186/1471-2105-9-40 | PubMed
7. **Zheng W.,** et al. (2025) Deep-learning-based single-domain and multidomain protein structure prediction with DI-TASSER. *Nature Biotechnology* **44**:641-653 https://doi.org/10.1038/s41587-025-02654-4 | PubMed
8. **Deng H.,** Jia Y., Zhang Y. (2016) 3DRobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics* **32**:378-387 https://doi.org/10.1093/bioinformatics/btv601 | PubMed
9. **Wang T.,** He X., Li M., Shao B., Liu T.-Y. (2023) AIMD-Chig: Exploring the conformational space of a 166-atom protein Chignolin with ab initio molecular dynamics. *Scientific Data* **10**:549 https://doi.org/10.1038/s41597-023-02465-9 | PubMed
10. **Shaw D.E.,** et al. (2010) Atomic-level characterization of the structural dynamics of proteins. *Science* **330**:341-346 https://doi.org/10.1126/science.1187409 | PubMed
11. **Vander Meersche Y.,** Cretin G., Gheeraert A., Gelly J.-C., Galochkina T. (2024) ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research* **52**:D384-D392 https://doi.org/10.1093/nar/gkad1084 | PubMed
12. **Mirarchi A.,** Giorgino T., De Fabritiis G. (2024) mdCATH: A large-scale MD dataset for data-driven computational biophysics. *Scientific Data* **11**:1299 https://doi.org/10.1038/s41597-024-04140-z | PubMed
13. **Liu C.,** et al. (2024) Dynamic PDB: A New Dataset and a SE (3) Model Extension by Integrating Dynamic Behaviors and Physical Properties in Protein Structures. *arXiv* https://doi.org/10.48550/arxiv.2408.12413
14. **Lewis S.,** et al. (2024) Scalable emulation of protein equilibrium ensembles with generative deep learning. *bioRxiv* https://doi.org/10.1101/2024.12.05.626885

15. **Jing B.,** Berger B., Jaakkola T. (2024) AlphaFold meets flow matching for generating protein ensembles. *arXiv* https://doi.org/10.48550/arxiv.2402.04845
16. **Williams C.J.,** Richardson D.C., Richardson J.S. (2022) The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues. *Protein Science* **31**:290-300 https://doi.org/10.1002/pro.4239 | PubMed
17. **Zhang J.,** Zhang Y. (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one* **5**:e15386 https://doi.org/10.1371/journal.pone.0015386 | PubMed
18. **Huang X.,** Pearce R., Zhang Y. (2020) EvoEF2: accurate and fast energy function for computational protein design. *Bioinformatics* **36**:1135-1142 https://doi.org/10.1093/bioinformatics/btz740 | PubMed
19. **Alford R.F.,** et al. (2017) The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation* **13**:3031-3048 https://doi.org/10.1021/acs.jctc.7b00125 | PubMed
20. **Delgado J.,** Radusky L.G., Cianferoni D., Serrano L. (2019) FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**:4168-4169 https://doi.org/10.1093/bioinformatics/btz184
21. **Cheng H.,** et al. (2014) ECOD: an evolutionary classification of protein domains. *PLoS computational biology* **10**:e1003926 https://doi.org/10.1371/journal.pcbi.1003926 | PubMed
22. **Sillitoe I.,** et al. (2021) CATH: increased structural coverage of functional space. *Nucleic acids research* **49**:D266-D273 https://doi.org/10.1093/nar/gkaa1079 | PubMed
23. **Xu J.,** Zhang Y. (2010) How significant is a protein structure similarity with TM-score= 0.5?. *Bioinformatics* **26**:889-895 https://doi.org/10.1093/bioinformatics/btq066 | PubMed
24. **Lewis S.,** et al. (2025) Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* **389**:eadv9817 https://doi.org/10.1126/science.adv9817 | PubMed
25. **Moult J.,** Pedersen J.T., Judson R., Fidelis K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**:ii-iv https://doi.org/10.1002/prot.340230303 | PubMed
26. **Shen W.,** Le S., Li Y., Hu F. (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one* **11**:e0163962 https://doi.org/10.1371/journal.pone.0163962 | PubMed
27. **Cock P.J.,** et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**:1422 https://doi.org/10.1093/bioinformatics/btp163 | PubMed
28. **O'Boyle N.M.,** Banck M., James C.A., Morley C., Vandermeersch T., Hutchison G.R. (2011) Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **3**:33 https://doi.org/10.1186/1758-2946-3-33
29. **Lindahl E.,** Hess B., Van Der Spoel D. (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual* **7**:306-317 https://doi.org/10.1007/s008940100045
30. **Zhang Y.,** Skolnick J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**:2302-2309 https://doi.org/10.1093/nar/gki524 | PubMed
31. **Zhang Y.,** Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**:702-710 https://doi.org/10.1002/prot.20264 | PubMed
32. **Reva B.A.,** Finkelstein A.V., Skolnick J. (1998) What is the probability of a chance prediction of a protein structure with an rmsd of 6 å?. *Folding and Design* **3**:141-147 https://doi.org/10.1016/s1359-0278(98)00019-4 | PubMed
33. **Chaudhury S.,** Lyskov S., Gray J.J. (2010) PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**:689-691 https://doi.org/10.1093/bioinformatics/btq007 | PubMed
34. **Du Z.,** et al. (2021) The trRosetta server for fast and accurate protein structure prediction. *Nature protocols* **16**:5634-5651 https://doi.org/10.1038/s41596-021-00628-9 | PubMed

35. **Yang J.,** Anishchenko I., Park H., Peng Z., Ovchinnikov S., Baker D. (2020) Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **117**:1496-1503 https://doi.org/10.1073/pnas.1914677117 | PubMed

36. **Li Y.,** Zhang C., Yu D.-J., Zhang Y. (2022) Deep learning geometrical potential for high-accuracy ab initio protein structure prediction. *Iscience* **25** https://doi.org/10.1016/j.isci.2022.104425 | PubMed

37. **Sehnal D.,** et al. (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic acids research* **49**:W431-W437 https://doi.org/10.1093/nar/gkab314 | PubMed

38. **Plotly Technologies Inc** (2015) Collaborative data science.

39. **McKinney W.** (2011) pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing* **14**:1-9

Peer reviews

Reviewer #1 (Public review):

Summary:

The authors describe a new database that rigorously explores protein conformations.

Strengths:

It is extremely well done, using state-of-the-art tools by a group at the top of the field of structural modeling. The evaluation of qualities and the benchmarking of the structures are outstanding, and it is expected that the new database will have a significant impact on the field.

Weaknesses:

The authors are using MD simulation to generate some of the structure, and therefore should have access to standard MD energies. I am surprised that no evaluation is provided based on these energies that can be extended to free energies.

https://doi.org/10.7554/eLife.110874.1.sa2

Reviewer #2 (Public review):

Summary:

The authors developed a dataset of protein conformations by running molecular dynamics simulations starting from both native and decoy conformations for a large number of proteins. These conformations were put together as a dataset for querying and downloading, along with their energies under different force fields. The authors suggest that such conformations represent the proteins' conformational landscape, so that they will be useful for evaluating methods generating multiple conformations of proteins.

Strengths:

The dataset is online and working. It has good documentation for others to use.

Weaknesses:

The biggest weakness is that the collected conformations very likely do not represent the true conformational landscape. To represent the conformational landscape, the structures need to be sampled based on the Boltzmann distribution. However, in this study, conformations are generated by running very short (125ps to 375ps) MD simulations starting from near-native

conformations and decoys. Such short simulations will produce small fluctuations around the starting conformations, so the distribution of conformations is largely dominated by the distribution of the initial conformations, which by one means are Boltzmann distributed. A conformation might be physically plausible, but it might have very small weight in the Boltzmann distribution. On the other hand, conformations with large weights might not be in the dataset.

https://doi.org/10.7554/eLife.110874.1.sa1

Reviewer #3 (Public review):

Summary:

This manuscript describes a web-based tool that allows researchers to compare large numbers of representative ("plausible") conformations of proteins. It also includes energetic analysis from multiple widely used structure-prediction methods.

Strengths:

This tool will likely be useful for students who want to learn more about the ensemble properties of proteins. The resource is well organized and it represents a large amount of computing resources.

Weaknesses:

It is not entirely clear how the database may be utilized by other groups to advance research. It could be helpful if the authors add a short section that provides example use cases that illustrate how this database can support new strategies for studying protein dynamics.

https://doi.org/10.7554/eLife.110874.1.sa0