ProteinConformers: Benchmark Dataset for Simulating Protein Conformational Landscape Diversity and Plausibility

Yihang Zhou[1*] **Chen Wei**[1,2*] **Matthew Minghao Sun**[1] **Jin Song**[3,4] **Yang Li**[1]
Lin Wang[5†] **Yang Zhang**[1,5†]

[1]National University of Singapore [2]Xi'an University of Posts & Telecommunications
[3]University of Chinese Academy of Sciences [4]Chinese Academy of Sciences
[5]Chinese Academy of Medical Sciences
[*]Equal contribution. [†]Correspondence author.

yihangjoe@foxmail.com, weichen@xupt.edu.cn
wenzhao30030@gmail.com, zhang@nus.edu.sg

Abstract

Understanding the conformational landscape of proteins is essential for elucidating protein function and facilitating drug design. However, existing protein conformation benchmarks fail to capture the full energy landscape, limiting their ability to evaluate the diversity and physical plausibility of AI-generated structures. We introduce **ProteinConformers**, a large-scale benchmark dataset comprising over **381,000** physically realistic conformations for **87** CASP targets. These were derived from more than **40,000** structural decoys via extensive all-atom molecular dynamics simulations totaling over **6 million CPU hours**. Using this dataset, we propose novel metrics to evaluate conformational diversity and plausibility, and systematically benchmark six protein conformation generative models. Our results highlight that leveraging large-scale protein sequence data can enhance a model's ability to explore conformational space, potentially reducing reliance on MD-derived data. Additionally, we find that PDB and MD datasets influence model performance differently, current models perform well on inter-atomic distance prediction but struggle with inter-residue orientation generation. Overall, our dataset, evaluation metrics, and benchmarking results provide the first comprehensive foundation for assessing generative models in protein conformational modeling. Dataset and instructions are available at `https://huggingface.co/datasets/Jim990908/ProteinConformers/tree/main`. Codes are stored at `https://github.com/auroua/ProteinConformers`. An interactive website locates at `https://zhanggroup.org/ProteinConformers`.

1 Introduction

In recent years, the field of protein structure prediction has rapidly shifted from a "single-conformation" paradigm toward "multi-conformation" modeling. While energy-based and deep learning–based methods such as I-TASSER[1], AlphaFold[2] and RoseTTAFold[3] deliver accurate static 3D structures, proteins under physiological conditions typically sample multiple functionally relevant states. The prediction of protein multi-conformations from conformational landscape is critical both for understanding molecular mechanisms and for demanding downstream applications. Yet, most studies to date have paid little systematic attention to balancing a model's ability to generate

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

multiple conformers with conformational diversity, or to examine the atom-level physical plausibility. There remains no benchmark dataset or metric suite that faithfully emulates a realistic folding funnel while simultaneously capturing polymorphism and ensuring physical plausibility.

To date, there is no analogous benchmark for multi-conformation generation. For single-conformation prediction, the Protein Data Bank (PDB)[4] serves as the golden standard, and its experimentally determined "native" structures naturally occupy global minima in the folding funnel. Some efforts have attempted to harvest a conformational ensemble by running extensive all-atom molecular dynamics (MD) simulations from the native state, yielding a wealth of near-native fluctuations around the crystallographic structure. However, the energy-based nature of MD inherently biases sampling toward the initial funnel basin[5, 6], and only extremely long simulations have any chance of escaping into more remote minima—at a computational cost that is both prohibitive and difficult to predict.

In terms of evaluation metrics, the community has established measures to assess single-conformation quality, such as TM-score[7], RMSD[8], and LDDT[9]. Metrics tailored to multi-conformation ensembles, however, remain in infancy. Recent proposals compute Jensen-Shannon (JS) divergence between ensembles based on pairwise distance distributions, or radius of gyration histograms, but still fall short of capturing atom-level distance and angle plausibility. Since a protein's conformation is uniquely defined by its internal geometry, a truly comprehensive metrics should directly evaluate the completeness and physical realism of all interatomic distance and torsional angles.

To address these gaps, we introduce **ProteinConformers** (Figure1). Starting from 87 Critical Assessment of protein Structure Prediction (CASP) [10, 11, 12] targets, we collected and filtered over 40,000 decoy conformers generated by hundreds of different traditional and AI based prediction algorithms. Each decoy was independently refined with all-atom MD simulations to resolve steric clashes and fine-tune geometry, yielding more than 381,000 physically realistic conformations that both span diverse regions of the folding landscape and satisfy stringent physicochemical constraints. To our knowledge, this is the first dataset of its kind, and none of the 87 proteins we release appear in prior multi-conformer benchmarks. Building this resource consumed over **6 million CPU hours**.

Built on this diverse and trustful dataset, we design a dual-axis evaluation framework. Axis one is **Conformational Diversity**, to quantify the breadth and representativeness of generated conformers across the overall energy funnel, with corresponding metrics. The axis is **Conformational Physics Plausibility**, which assesses the all-atom geometric and energetic validity of every generated conformer, with dedicated metrics.

Our preliminary benchmarking of six approaches[13, 14, 15] reveals that, while many methods accurately reproduce interatomic distance distributions, they generally under-explore the broader conformational landscape and exhibit significant shortcomings in torsion-angle sampling. Besides the basic task of ProteinConformers to understand how well current conformational generative methods explore the broader conformational landscape, this work seeks to provide some insight to the question: *Is it possible to develop generative models capable of superior energy landscape exploration without reliance on molecular dynamics (MD) simulation datasets?* By releasing **ProteinConformers** and its accompanying evaluation framework, we provide a large-scale, high-fidelity benchmark that will accelerate advances in protein structure prediction, multi-conformation modeling, computational biology, and rational drug discovery.

2 Related Work

2.1 Protein Conformational Landscape Exploration

Early efforts to map the protein folding funnel relied on coarse-grained sampling methods that rotate backbone or side-chain dihedrals under simplified force fields [16, 17], as well as fragment-assembly threading guided by Monte Carlo Simulation[18] algorithm, such as 3DRobot [19]. Although computationally efficient, these approaches suffer from low-accuracy energy functions and heavily reduced representations of protein geometry. To improve physical fidelity, recent work such as AIMD-Chig[20] has explored conformational space at the DFT level [21], but only for a single 166-atom peptide due to the prohibitive cost. Likewise, large-scale MD simulations initialized from experimental structures, such as BPTI [22], Atlas [23], mdCATH [24], Dynamic PDB [25], provide ensembles of near-native fluctuations, yet these trajectories remain confined to the primary funnel basin unless run for impractically long time, limiting exploration of more diverse conformations.

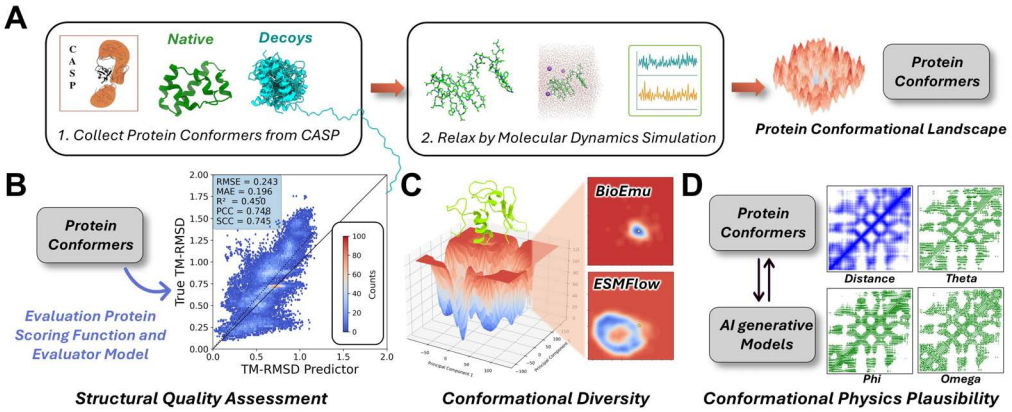


Figure 1: **Overview of the ProteinConformers benchmark and evaluation framework**. (A) We curate decoy ensembles for CASP targets and, for each decoy, perform all-atom MD refinement to sample across the folding landscape. (B) Beyond providing conformations, ProteinConformers includes a panel of per-conformation energy metrics, enabling auxiliary assessment of decoy quality. (C) We define diversity metrics that quantify how well generated conformations explore the overall conformational funnel. (D) We design physics-plausibility metrics to evaluate the atom-level realism based on inter-residual distances and torsion angles.

Unlike previous studies, we run MD simulations in batch with decoys at variant positions in energy landscapes generated by hundreds of different algorithms, which guarantees both diversity and plausibility of protein conformational landscape exploration.

2.2 Protein Conformation Generative Models

Traditional generative strategies commonly adopt coarse-grained movements to synthesize conformers [26, 27], but such models frequently produce steric clashes, and require post hoc side-chain reconstruction. With the advent of highly accurate single-structure predictors, a new class of diffusion [28, 29] and flow-based [30] generative models has emerged, fine-tuning backbones or post-processing predicted poses to yield all-atom ensembles. Examples include AlphaFlow [13], ESMFlow [13], BioEmu [14] and ESMDiff [15]. AlpahFlow and ESMFlow are two protein conformation generative models obtained by fine-tuning AlphaFold2 [2] and ESMFold [31] using flow matching framework, respectively. BioEmu [14] uses the trained AlphFold2 model to extract features and using a two-step training strategy to train a denoise diffusion model based on the extracted features to generate a collection of conformations that can reflect the equilibrium distribution of the structure of input sequences. ESMDiff [15] is fine-tuned from pre-trained ESM3 [32], and uses a conditional language model to capture sequence-specific conformational distributions. AFsample2 [33] generates ensembles by MSA (multiple sequence alignment) sampling based on AlphaFold2. AlphaFold3 [34] produces multiple conformations by initiating random noises during diffusion inference.

3 ProteinConformers Dataset

3.1 Source of ProteinConformers

CASP is a biennial competition and the CASP committee curates dozens of representative targets and invite global teams to predict the structures under blind conditions. The resulting predictions reflect the cutting-edge-tech in protein structure prediction and exhibit high conformational diversity due to the variety of methods employed by participating teams, which collectively enables extensive sampling of the protein folding landscape. However, these predicted structures are not guaranteed to be physically realistic at the atomic level. The MD simulation, by contrast, use physics-based force fields to refine protein conformations at full atomic resolution. Nevertheless, MD simulations are computationally expensive and conformational transitions are often limited to regions near the input structure or trapped in local minima[5, 6], resulting in high redundant conformations.

To address these complementary limitations, by integrating the strengths of both resources, we compiled the predictions from CASP and, after quality control, curated a dataset of 40,387 predicted structures for 87 proteins, and then performed MD simulations on each structure. Taking advantage of the large conformational shifts observed in the early phase of MD trajectories, we limited simulations to short timescales to reduce computational cost and redundancy. This resulted in a final dataset of 381,546 all-atom, physically refined protein conformations broadly sampling the folding landscape.

3.2 Preprocessing of Protein Conformers

To build ProteinConformers benchmark, we systematically collected, cleaned and corrected all prediction entries in CASP14 [11] and CASP15 [12]. In CASP14, nearly 100 teams participated in predicting structures for 90 modeling targets. CASP15 had a similar number of participants working on 127 targets. The predicted protein structures are processed according to the following steps.

First, we remove duplicate accessions, non-protein entries and redundant sequences to obtain 172 unique targets with corresponding predictions. **Second**, for each prediction, chain sequences are extracted with Biopython [35] and aligned to the corresponding reference. The alignment yields three mutually exclusive categories: *same* means full-length and residue-order identity; *disorder* means a contiguous subsequence of the reference (internal or terminal deletions, order preserved); *mismatch* means any substitution, insertion or reordered segment. Oligomeric or hetero-complex models are annotated and excluded from the benchmark. **Third**, based on the categories from the previous step, the *same* models are accepted without modification. The *disorder* models are retained only if their residue numbers already matched the reference structure; otherwise the coordinates are automatically renumbered to restore one-to-one correspondence. All *mismatch* entries are discarded. **Fourth**, when multiple experimental native structures exists, pairwise TM-scores are calculated to chose the reference native. Targets without a definitive match are resolved by manual inspection. **Fifth**, a target enters the benchmark only if it had at least one native structure and 100 decoys. For targets below this threshold, additional decoys are generated using 3DRobot [19].

3.3 MD Simulation Protocol

MD simulations are performed using GROMACS 2023 [36]. 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 total charge. **Energy minimization:** Steepest-descent minimization is applied until the largest force on any atom fell below $1000\,\mathrm{kJ\,mol^{-1}\,nm^{-1}}$ (maximum 50,000 steps), thereby eliminating steric clashes and unrealistic geometries. **Two-stage equilibration:** The NVT phase ($100\,\mathrm{ps},\ 300\,\mathrm{K}$) employed the V-rescale thermostat with positional restraints on heavy atoms to stabilize temperature. The subsequent NPT phase ($100\,\mathrm{ps},\ 1\,\mathrm{bar}$) uses the Parrinello–Rahman barostat after partially releasing restraints to equilibrate density. **Run and sampling:** Restraints are removed and a simulation between 125 ps to 375 ps is executed. Periodic-boundary artifacts are eliminated by recentering and re-imaging the trajectory. Snapshots are extracted every 25 ps. System energies (total, potential, and protein-only) are recorded throughout for subsequent energy-landscape modeling. **Computational resources:** Simulations are executed on high-performance computers equipped with AMD EPYC 7763 (64-core, 2.45 GHz) processors. Applying this protocol to 40,387 starting conformers from 87 proteins generated 381,546 all-atom refined conformations, at an aggregate cost over 6 million CPU hours. These trajectories and associated physical properties constitute the core of the ProteinConformers.

3.4 Analysis of ProteinConformers Dataset

ProteinConformers benchmark dataset is challenging. It comprises 87 CASP targets with an average sequence length of 305 residues and a median of 255 (Figure 2A). These targets, selected by the CASP committee, span a broad array of fold topologies—including predominantly α-helical proteins, β-sheet–rich folds, disordered regions, and multi-domains. Notably, 32 of the targets exceed 300 residues, and the largest reaches 949 residues. ProteinConformers provides both diverse and balanced protein conformational landscape ensembles. Each protein sampled 4,386 conformations on average. By classifying conformers as near-native ($\mathrm{TM\text{-}score} \geq 0.5$) or non-native ($\mathrm{TM\text{-}score} < 0.5$), we show that the dataset spans the full spectrum of conformational landscape, as visualized by

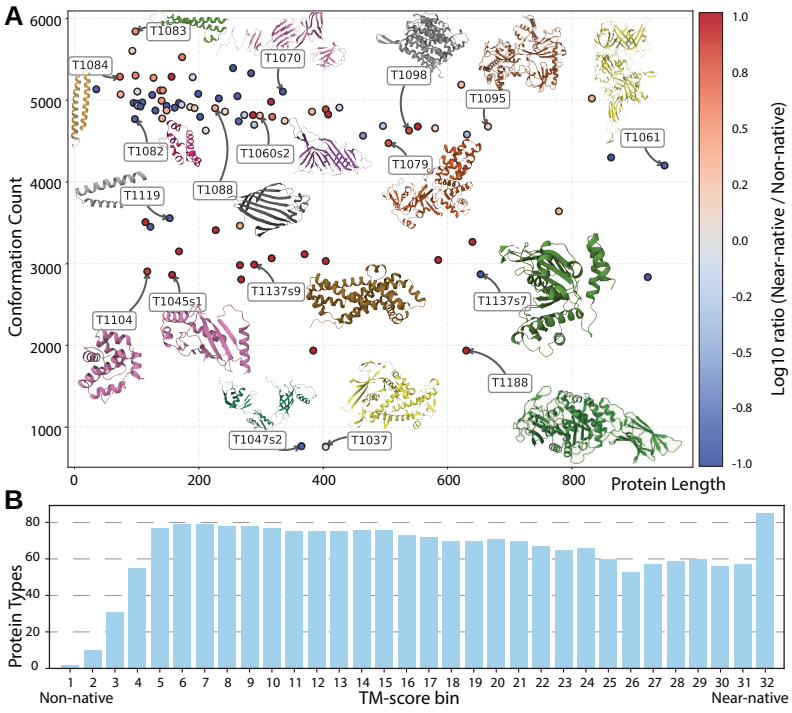


Figure 2: **Global distribution of the ProteinConformers dataset**. (A) Number of conformers per protein, sorted by sequence length (x-axis). The y-axis shows the total distinct conformational states captured for each target. Bar color encodes the log10-ratio of near-native to non-native conformations in our benchmark. Insets display 3D renderings of randomly selected proteins. (B) TM-score coverage histogram across the dataset. The x-axis divides TM-score into 32 equal-width bins from low (non-native) to high (near-native), and the y-axis indicates the number of proteins in each bin.

the log-ratio color scale in Figure 2A and TM-score distribution in Figure 2B. The TM-score distribution reveals a roughly uniform count of target proteins from non-native to near-native regions. In contrast to prior benchmarks that focus on sampling near-native conformers around PDB structures, ProteinConformers achieves uniform coverage across the entire TM-score range.

3.5 Conformation Similarity Predictor

Unlike other datasets, ProteinConformers also provides energy metrics for each conformation, including TM–score, RMSD, and atomic-level energy scores from EvoEF2 [38], RW [40], RWplus [40], FoldX [37], and Rosetta [39]. To illustrate the utility of these features, we explore a challenging task: predicting structural similarity metrics without access to native structures using only energy features. The similarity metrics are TM–score and TM–RMSD. Here, TM-RMSD is defined as

$$\text{TM-RMSD} = \text{TM-score} + \frac{1}{1 + \text{RMSD}} \tag{1}$$

which ranges in $(0, 2]$, with higher values indicating greater structural similarity. Identical conformations ($\text{TM-score} = 1$, $\text{RMSD} = 0$) result in a TM–RMSD of 2. Predicting such global and local similarity metrics without reference structures is a nontrivial task. While AlphaFold's pLDDT score is commonly used for model quality estimation, its pTM metric is derived from Evoformer internals and is not directly accessible. We train a simple conformation similarity predictor using the ProteinConformers' energy features (Figure 3). The 87 proteins are split into train, test and validation dataset with ratio of 0.8, 0.1 and 0.1. We use AutoGluon[41] to fit a simple linear regression model using the five energy terms and protein length as features. Despite its simplicity, the model achieved promising results: PCC of 0.743 for TM–score and 0.748 for TM–RMSD, and SCC of 0.776 and 0.745, respectively. These findings demonstrate that ProteinConformers' energy features are valuable resource for conformational analysis and quality prediction.

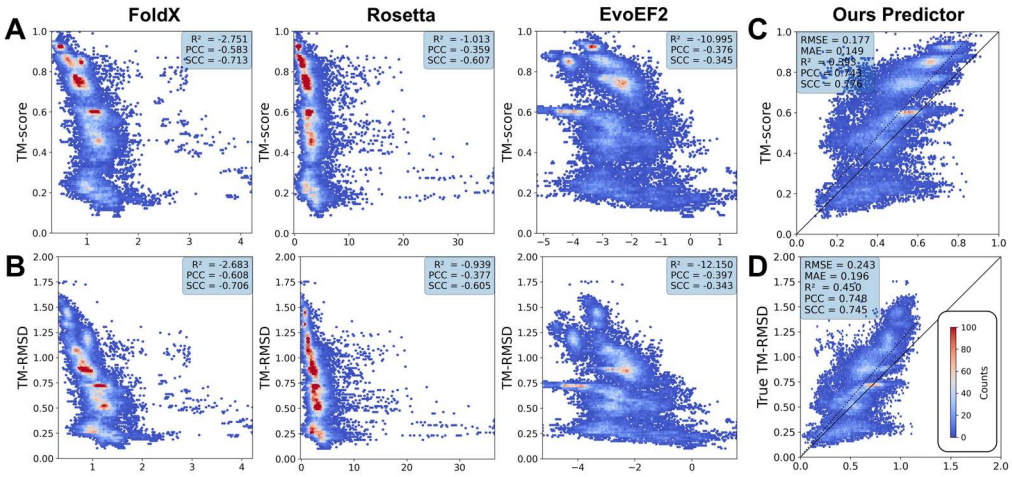


Figure 3: **Structural-quality assessment using classical energy functions and our learned Conformation Similarity Predictor.** (A) Density scatter plots of length-normalized potential energy (x-axis) versus TM-score (y-axis) for three classical scoring functions[37, 38, 39]. Color denotes point density; insets report R^2, Pearson correlation coefficient (PCC), and Spearman rank coefficient (SCC). (B) Analogous density plots of length-normalized potential energy (x-axis) versus TM-RMSD (y-axis). (C) True TM-score (y-axis) versus our predictor's estimated TM-score (x-axis). Inset statistics include RMSE, MAE, R^2, PCC, and SCC; the diagonal line denotes perfect agreement. (D) True TM-RMSD (y-axis) plotted against predicted TM-RMSD (x-axis).

4 Experiment

4.1 Evaluation Metrics and Compared Models

Based on the ProteinConformers dataset, we employ a coverage-based metric derived from free energy landscapes and introduce a new quantitative measure to assess two distinct aspects of the generated conformational ensembles: (i) **Coverage-based metrics**, which assesses the diversity of the protein conformational landscape by calculating the coverage rate of the generated conformers to the benchmark dataset, and (ii) **Protein Conformation Plausibility Map (PCPM)** and its derived **Protein Conformation Plausibility Score (PCPS)**, which evaluate the plausibility of the generated conformational ensembles in relation to known folding landscapes. To gain new insights into current conformational generative models, six different models were systematically compared. Two distilled variants of AlphaFlow were included: AlphaFlow$^{Dis}_{PDB}$, trained on experimental ensembles from PDB, and *AlphaFlow*$^{Dis}_{MD}$, trained on 300 K all-atom explicit-solvent MD trajectories. The same protocol is applied to ESMFlow$^{Dis}_{PDB}$ and ESMFlow$^{Dis}_{MD}$. For ESMDiff, the DDPM sampling paradigm with a step size of 1,000 is used. BioEmu is executed according to its official workflow. Each model generates 3,000 conformations for a curated subset of 18 proteins from the ProteinConformers.

4.1.1 Protein Conformational Landscape Diversity

Protein conformational landscape diversity is measured by the overlap of low-energy regions in free energy landscapes between ProteinConformers and generated ensembles (Figure 4). We project protein conformations onto a 2D space using Principal Component Analysis (PCA) via MDAnalysis [42, 43], using the first two principal components from ProteinConformers for each protein. Generated conformers are projected onto this same space. The 2D projections are discretized into a 64×64 grid. Conformer density in each bin constructs free energy landscapes ($pmf = -kT \log(hist)$, where $hist$ is normalized density and $kT = 2.494$ kJ/mol, more in supplementary materials). Area intersection, coverage, and Jaccard index (equations in Supplementary Material) quantify low-energy region overlap between generation and ProteinConformers.

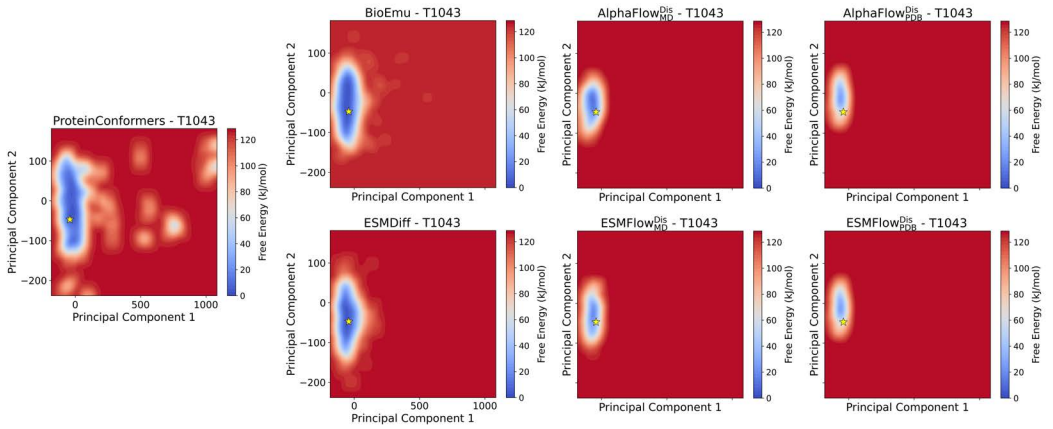


Figure 4: Case of free energy landscapes comparison from ProteinConformers and generative models.

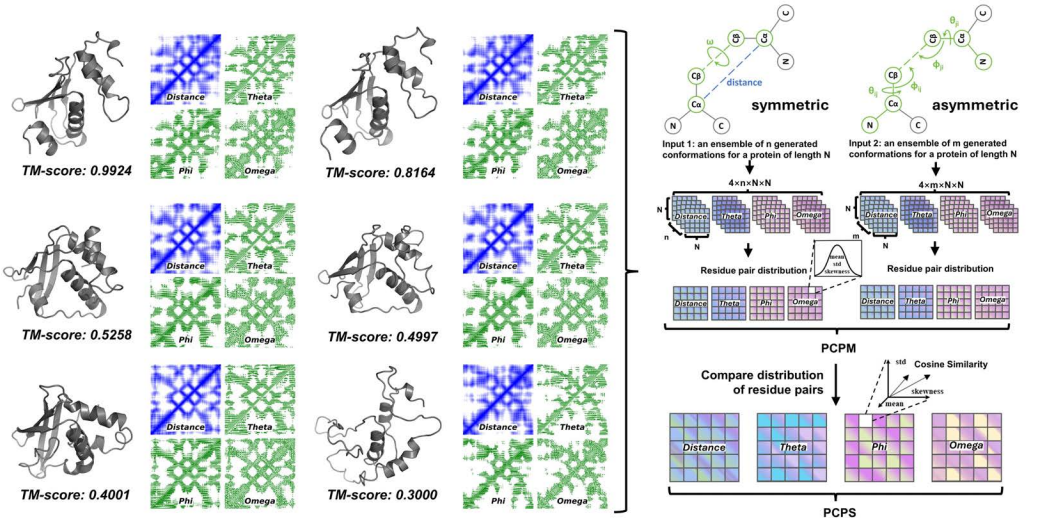


Figure 5: **Illustration of PCPM and PCPS.** Left: Six representative protein conformations with varying TM-scores and PCPMs. Upper Right: Geometric features construction: symmetric (distance D, dihedral Ω) and asymmetric (dihedral Θ, planar angle Φ). Lower Right: PCPM and PCPS construction workflow, comparing residue-pair distribution statistics (mean, std. dev., skewness) of an input ensemble's PCPM with a benchmark PCPM via cosine similarity.

4.1.2 Protein Conformation Plausibility Maps and Scores

Traditional metrics such as TM-score [7], compare a generated structure with a single native reference and therefore quantify *structural similarity* rather than *atomic-level plausibility of the conformational distribution*. To address this gap, we introduce the PCPM and its derived PCPS, which directly assess the local physical realism of an ensemble of conformations (Figure 5) .

PCPM Construction For a protein of length N and an ensemble of M conformations $\mathcal{C} = \{C^{(m)}\}_{m=1}^{M}$, we compute four inter-residue geometric feature maps $G \in \{D, \Omega, \Theta, \Phi\}$ for every conformation, analogous to trRosetta [44] and DeepPotential [45]. Feature definitions are in Table 1.

For every residue-pair $(i, j) \in \{1, \ldots, N\}$ of each conformation m in ensembles, we calculate the four geometric features. We then summarize the three central moments of the ensembles geometric features—mean (μ), standard deviation (σ), and skewness (γ), which results representation tensor for geometric feature map G is $\mathbf{P}^G \in \mathbb{R}^{3\times N\times N}$, where $\mathbf{P}^G_{ij} \in \{\mu, \sigma, \gamma\}$. The complete PCPM is the concatenation $\mathbf{P} = \left[\mathbf{P}^D, \mathbf{P}^\Omega, \mathbf{P}^\Theta, \mathbf{P}^\Phi\right] \in \mathbb{R}^{12\times N\times N}$.

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

Symbol	Type	Atoms Involved	Symmetry	Description
D	Euclidean Dist.	$C_{\alpha,i}, C_{\alpha,j}$	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

PCPS Computation Let $\mathbf{P}^{\mathrm{PC}}$ be ProteinConformer's PCPM and $\mathbf{P}^{\mathrm{B}}$ be the benchmark method's map. After aligning residue indices (discarding missing ones), for every entry (i, j) we build vectors $\mathbf{u}_{ij} = \big[\mu, \sigma, \gamma\big]^{\top}_{(\mathrm{PC})}$ and $\mathbf{v}_{ij} = \big[\mu, \sigma, \gamma, \big]^{\top}_{(\mathrm{B})}$, and compute cosine similarity $s_{ij} = \frac{\mathbf{u}_{ij}^{\top}\mathbf{v}_{ij}}{\|\mathbf{u}_{ij}\|\,\|\mathbf{v}_{ij}\|}$. Then $\mathrm{PCPS} = \frac{1}{|\mathcal{C}|}\sum_{(i,j)\in\mathcal{C}} s_{ij}$, where $\mathcal{C}$ is the set of aligned residue-pair entries. $\mathrm{PCPS} \in [-1, 1]$; higher values indicate better reproduction of atomic-level statistics of the authentic folding landscape.

We also used **Jensen-Shannon Divergence (JSD)** to evaluate the diversity, which measures the similarity between two probability distributions, P and Q.

$$\mathrm{JSD}(P||Q) = \frac{1}{2}D_{\mathrm{KL}}(P||M) + \frac{1}{2}D_{\mathrm{KL}}(Q||M), \text{ where } M = \frac{1}{2}(P + Q)$$

Specifically, we employ two JSD-based metrics to assess the structural diversity and compactness distributions.

JS Divergence of Pairwise Distances (JS-PwD): This metric measures the similarity of inter-residue distance distributions. For each Cα–Cα atom pair (i, j) separated by at least 3 residues, we treat it as a distinct "channel". For each channel, we collect the distance values across all conformations in generated ensembles and ProteinConformers to form two distributions. We then compute the JSD for each channel and report the final JS-PwD as the mean JSD across all channels.

JS Divergence of Radius of Gyration (JS-Rg): This metric measures the similarity of overall molecular compactness between generated ensembles and ProteinConformers. For each ensemble, we compute the Radius of Gyration (R_g) of all conformations, build normalized histograms, and calculate the Jensen–Shannon divergence between the two distributions.

Ramachandran outlier rate (Rama outlier %): This metric quantifies the fraction of residues whose backbone dihedral angles (ϕ, ψ) fall outside the favored Ramachandran regions. The final value is the mean outlier percentage across all conformations in the ensemble.

4.2 Evaluation Results

4.2.1 Protein Conformational Landscape Diversity

Figure 4 illustrates the two-dimensional free energy landscapes for protein T1043 as derived from the ProteinConformers dataset and six generative models: AlphaFlow$^{Dis}_{MD}$, AlphaFlow$^{Dis}_{PDB}$, ESMdiff, $^{Dis}_{MD}$, and ESMFlow$^{Dis}_{PDB}$, BioEmu, and ESMDiff. Low-energy basins (in blue) indicate regions of high conformer density, while high-energy regions (in red) are sparsely populated. The yellow star marks the native structure projected onto the same PCA space. ProteinConformers exhibits a more diverse and widely distributed free energy landscape. In contrast, generative models, especially the distilled variants, tend to produce narrower distributions.

Table 2 compares three free energy overlap metrics—interaction, coverage, and Jaccard index—across energy thresholds of 5, 10, and 20 kJ/mol. BioEmu performs best at 5 kJ/mol with strong interaction and Jaccard scores, indicating effective low-energy sampling. ESMDiff achieves the highest overlaps at 10 and 20 kJ/mol, reflecting broader conformational exploration. In contrast, distilled models like AlphaFlow$^{Dis}_{MD}$ and ESMFlow$^{Dis}_{MD}$ show consistently lower scores, suggesting narrower sampling near dominant energy basins.

ESMDiff, trained solely on PDB data without MD simulations, still shows strong conformational exploration ability (Table 2). This strength likely arises from its ESM3 foundation, pretrained on large-scale multimodal data, where massive sequence corpora provide implicit priors for sampling realistic conformations. Similar to PLAID [46], this highlights the potential of sequence-based pretraining for generating diverse, physically plausible protein structures.

Table 2: Protein Conformational Landscape with Different Energy Thresholds

Method	Energy Interaction(↑)			Energy Coverage(↑)			Energy Jaccard(↑)		
	5	**10**	**20**	**5**	**10**	**20**	**5**	**10**	**20**
$\text{AlphaFlow}^{\text{Dis}}_{\text{MD}}$	2.83	59.39	108.06	0.029	0.130	0.178	0.005	0.039	0.072
$\text{AlphaFlow}^{\text{Dis}}_{\text{PDB}}$	6.44	64.17	115.28	0.043	0.136	0.178	0.012	0.046	0.086
$\text{ESMFlow}^{\text{Dis}}_{\text{MD}}$	2.33	76.33	143.72	0.022	0.143	0.237	0.004	0.043	0.089
$\text{ESMFlow}^{\text{Dis}}_{\text{PDB}}$	6.44	64.17	115.28	0.043	0.136	0.179	0.012	0.046	0.086
AFsample2	1.14	13.86	90.43	0.004	0.047	0.116	0.003	0.020	0.070
AlphaFold3	7.47	60.00	104.59	0.039	0.081	0.126	0.021	0.045	0.076
BioEmu	**13.00**	91.11	146.78	0.079	0.178	0.250	**0.031**	0.089	0.138
ESMDiff	9.56	**139.78**	**213.17**	**0.140**	**0.299**	**0.317**	0.021	**0.102**	**0.158**

Note: The unit of the different energy thresholds is kJ/mol.

Our experimental results also show that the diffusion-based AlphaFold3 performs comparably to AlphaFlow, consistent with their similar generative paradigms. In contrast, AFsample2 exhibits the weakest performance, indicating that such perturbations only induce minor local structural variations and fail to capture global conformational diversity.

4.2.2 Protein Conformational Landscape Plausibility

The six generative models' PCPM is denoted as PCPM^{B}. We also computed three reference PCPM values on the ProteinConformers dataset itself: $\text{PCPM}^{\text{near-native}}$, using decoys with TM-score ≥ 0.5; and PCPM^{all}, using the full set of decoys regardless of TM-score. We then compare PCPM^{B} against each of the $\text{PCPM}^{\text{near-native}}$ and PCPM^{all} to obtain the PCPS, which reflects the physical realism of the generative models in different regions of the protein conformational landscape—non-native, near-native, and global. The results are summarized in Table 3.

Table 3: Performance of PCPS with benchmark models

	Near-native(↑)					All (↑)				
	D	O	P	T	A	D	O	P	T	A
$\text{AlphaFlow}^{\text{Dis}}_{\text{MD}}$	0.946	0.627	0.801	0.537	0.728	0.900	0.644	0.772	0.574	0.723
$\text{AlphaFlow}^{\text{Dis}}_{\text{PDB}}$	0.922	0.616	0.786	0.564	0.722	0.871	0.592	0.746	0.542	0.688
$\text{ESMFlow}^{\text{Dis}}_{\text{MD}}$	0.957	0.604	0.811	0.539	0.728	0.922	0.710	0.814	0.665	0.778
$\text{ESMFlow}^{\text{Dis}}_{\text{PDB}}$	0.859	0.516	0.710	0.468	0.638	0.819	0.603	0.700	0.578	0.675
BioEmu	0.967	**0.717**	**0.826**	**0.665**	**0.794**	0.935	0.764	**0.862**	**0.713**	**0.818**
ESMDiff	**0.971**	0.649	0.693	0.191	0.626	**0.967**	**0.807**	0.639	0.338	0.688

Note: Metrics abbreviations: D-Distance, O-Omega, P-Phi, T-Theta, A-Average.

Table 4: Comparison of models on JS metrics and Ramachandran outliers

Model	Avg. JS-PwD (↓)	Avg. JS-Rg (↓)	Rama Outlier % (↓)
$\text{AlphaFlow}^{\text{Dis}}_{\text{MD}}$	0.424	0.310	13.43%
$\text{AlphaFlow}^{\text{Dis}}_{\text{PDB}}$	0.445	0.412	9.93%
$\text{ESMFlow}^{\text{Dis}}_{\text{MD}}$	0.283	0.507	13.36%
$\text{ESMFlow}^{\text{Dis}}_{\text{PDB}}$	0.417	0.697	12.26%
AFsample2	0.490	0.311	6.61%
AlphaFold3	0.632	0.293	**2.66%**
BioEmu	0.235	0.206	10.69%
ESMDiff	**0.128**	**0.104**	20.41%

From Table 3, MD-trained AlphaFlow and ESMFlow outperform their PDB-trained counterparts across most metrics. While all models align well with native Cα distance distributions, they reproduce orientation angles poorly. This may stem from FAPE loss focusing on per-residue alignment without

enforcing inter-residue pose consistency, and from two-stage architectures that limit torsional realism. BioEmu offers the best overall physical plausibility, whereas ESMDiff, despite its superior distance modeling, still lacks angular accuracy.

We evaluated geometric and stereochemical quality across models Table4. ESMDiff achieved the best geometric similarity (lowest JS-PwD = 0.128, JS-Rg = 0.104) but had the highest Ramachandran outlier rate (20.41%). AlphaFold3 showed the most stereochemically plausible structures with only 2.66% outliers, while AFsample2 and BioEmu performed moderately. Diffusion- and flow-based variants exhibited intermediate to high outlier levels.

These results reveal a trade-off between geometric similarity and physical plausibility. Models optimized for distributional alignment with experimental ensembles (e.g., ESMDiff) tend to sacrifice stereochemical accuracy, while structure-prediction-based frameworks such as AlphaFold3 and AFsample2 maintain physically realistic torsion angles but explore a narrower conformational space. The consistency between Ramachandran statistics and PCPS trends reinforces the reliability of our plausibility assessment framework.

5 Discussion

This study introduces ProteinConformers, a large-scale, diverse and physically realistic benchmark dataset with corresponding evaluation metrics. We collect over 40,387 decoys from 87 CASP targets, refine each decoy with all-atom MD simulations and sample 381,546 conformations at a cost exceeding 6 million CPU hours. We further defined a dual-axis evaluation framework comprising conformational diversity and physical plausibility.

We applied ProteinConformers and corresponding evaluation metrics to benchmark generative models. Our results show that models trained on native structures excel at exploring near-native energy funnel, whereas those trained on MD trajectories yield superior atomic-level physical realism. Although all methods reproduce good inter-residue distance distributions, they underperform in sampling inter-residue torsion angles. We believe that integrating both PDB and MD data and focusing on torsion-angle learning will be crucial for improving conformational plausibility.

We observe that current multi-conformer generators perform poorly on torsion-angle prediction, likely due to weak coupling between local geometry and long-range constraints. To address this, we suggest three complementary directions: introducing long-range orientation modules to propagate angular dependencies across residues; incorporating joint supervision of backbone and side-chain torsions to enhance stereochemical realism; and applying energy-based angular regularization to provide physics-guided priors that improve plausibility without compromising diversity.

ProteinConformers has several limitations. Short MD simulations may miss high-energy conformations, and the decoy set lacks fragment assembly or physics-based diversity. The similarity predictor is preliminary and unvalidated at scale. Due to computational limits, only 3,000 conformations per protein were generated, preventing full model benchmarking. Dataset non-overlap hinders direct comparison, and CASP-derived seeds may bias landscape coverage. Short MD windows miss rare transitions, limiting Boltzmann sampling. The absence of enhanced-sampling methods reduces exploration of high-energy states, and generalizability to experimental ensembles remains untested. Plausibility metrics, such as Ramachandran outliers, may also be sensitive to parameter choices.

Future work will diversify decoy generation pipelines by combining fragment assembly, coarse-grained physics simulations and generative modeling for more systematic coverage. We will also extend the MD sampling protocol by integrating enhanced sampling techniques and longer trajectory. Finally, we will develop an online platform that provides data browsing, metric computation and model evaluation in a unified environment, facilitating use by the computational biology and drug design communities.

6 Acknowledgements

This work is supported in part by Singapore Ministry of Education (T1 251RES2309) and National University of Singapore Startup Grants (A-8001129-00-00, A-0009651-30-00).

References

[1] Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The I-TASSER suite: protein structure and function prediction. *Nature Methods*, 12(1):7–8, 2015.

[2] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[3] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[4] Joel L Sussman, Dawei Lin, Jiansheng Jiang, Nancy O Manning, Jaime Prilusky, Otto Ritter, and Enrique E Abola. Protein data bank (pdb): database of three-dimensional structural information of biological macromolecules. *Biological Crystallography*, 54(6):1078–1084, 1998.

[5] Michael Levitt. Protein folding by restrained energy minimization and molecular dynamics. *Journal of molecular biology*, 170(3):723–764, 1983.

[6] Donald Hamelberg, John Mongan, and J Andrew McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of chemical physics*, 120(24):11919–11929, 2004.

[7] Yang Zhang and Jeffrey Skolnick. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005.

[8] Vladimir N Maiorov and Gordon M Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of molecular biology*, 235(2):625–634, 1994.

[9] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. LDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

[10] John Moult, Jan T. Pedersen, Richard Judson, and Krzysztof Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–iv, 1995.

[11] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021.

[12] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1539–1549, 2023.

[13] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22277–22303. PMLR, 21–27 Jul 2024.

[14] Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, Victor García Satorras, Osama Abdin, Bastiaan S. Veeling, Iryna Zaporozhets, Yaoyi Chen, Soojung Yang, Arne Schneuing, Jigyasa Nigam, Federico Barbero, Vincent Stimper, Andrew Campbell, Jason Yim, Marten Lienen, Yu Shi, Shuxin Zheng, Hannes Schulz, Usman Munir, Cecilia Clementi, and Frank Noé. Scalable emulation of protein equilibrium ensembles with generative deep learning. *bioRxiv*, 2024. preprint.

[15] Jiarui Lu, Xiaoyin Chen, Stephen Zhewen Lu, Chence Shi, Hongyu Guo, Yoshua Bengio, and Jian Tang. Structure language models for protein conformation generation. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024.

[16] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, 1975.

[17] Shoji Takada. Coarse-grained molecular simulations of large biomolecules. *Current Opinion in Structural Biology*, 22(2):130–137, 2012. Theory and simulation/Macromolecular assemblages.

[18] Christopher Z Mooney. *Monte Carlo Simulation*. Number 116. Sage, 1997.

[19] Haiyou Deng, Ya Jia, and Yang Zhang. 3DRobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics*, 32(3):378–387, 2016.

[20] Tong Wang, Xinheng He, Mingyu Li, Bin Shao, and Tie-Yan Liu. AIMD-Chig: Exploring the conformational space of a 166-atom protein Chignolin with ab initio molecular dynamics. *Scientific Data*, 10(1):549, 2023.

[21] C Fonseca Guerra, JG Snijders, G t Te Velde, and E Jan Baerends. Towards an order-N DFT method. *Theoretical Chemistry Accounts*, 99:391–403, 1998.

[22] David E. Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, Michael P. Eastwood, Joseph A. Bank, John M. Jumper, John K. Salmon, Yibing Shan, and Willy Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[23] Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1):D384–D392, 2024.

[24] Antonio Mirarchi, Toni Giorgino, and Gianni De Fabritiis. mdcath: A large-scale md dataset for data-driven computational biophysics. *Scientific Data*, 11(1):1299, 2024.

[25] Ce Liu, Jun Wang, Zhiqiang Cai, Yingxu Wang, Huizhen Kuang, Kaihui Cheng, Liwei Zhang, Qingkun Su, Yining Tang, Fenglei Cao, Limei Han, Siyu Zhu, and Yuan Qi. Dynamic PDB: A new dataset and a SE(3) model extension by integrating dynamic behaviors and physical properties in protein structures, 2024.

[26] Shina CL Kamerlin, Spyridon Vicatos, Anatoly Dryga, and Arieh Warshel. Coarse-grained (multiscale) simulations in studies of biophysical and chemical systems. *Annual Review of Physical Chemistry*, 62(1):41–64, 2011.

[27] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, 2016.

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020.

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[30] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[31] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[32] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

[33] Yogesh Kalakoti and Björn Wallner. Afsample2 predicts multiple conformations and ensembles with alphafold2. *Communications Biology*, 8(1):373, 2025.

[34] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[35] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[36] Erik Lindahl, Berk Hess, and David Van Der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual*, 7:306–317, 2001.

[37] Javier Delgado, Leandro G. Radusky, Damiano Cianferoni, and Luis Serrano. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics*, 35(20):4168–4169, 2019.

[38] Xiaoqiang Huang, Robin Pearce, and Yang Zhang. EvoEF2: accurate and fast energy function for computational protein design. *Bioinformatics*, 36(4):1135–1142, 2020.

[39] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J. Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691, 2010.

[40] Jian Zhang and Yang Zhang. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLOS ONE*, 5(10):1–13, 2010.

[41] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander J. Smola. Autogluon-tabular: Robust and accurate automl for structured data. *CoRR*, abs/2003.06505, 2020.

[42] Naveen Michaud-Agrawal, Elizabeth J. Denning, Thomas B. Woolf, and Oliver Beckstein. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, 32(10):2319–2327, July 2011.

[43] Richard J. Gowers, Max Linke, Jonathan Barnoud, Tyler J. E. Reddy, Manuel N. Melo, Sean L. Seyler, Jan Domański, David L. Dotson, Sébastien Buchoux, Ian M. Kenney, and Oliver Beckstein. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In *Proceedings of the 15th Python in Science Conference*, pages 98–105, 2016.

[44] Zongyang Du, Hong Su, Wenkai Wang, Lisha Ye, Hong Wei, Zhenling Peng, Ivan Anishchenko, David Baker, and Jianyi Yang. The trRosetta server for fast and accurate protein structure prediction. *Nature protocols*, 16(12):5634–5651, 2021.

[45] Yang Li, Chengxin Zhang, Dong-Jun Yu, and Yang Zhang. Deep learning geometrical potential for high-accuracy ab initio protein structure prediction. *Iscience*, 25(6), 2022.

[46] Amy X. Lu, Wilson Yan, Sarah A. Robinson, Kevin K. Yang, Vladimir Gligorijevic, Kyunghyun Cho, Richard Bonneau, Pieter Abbeel, and Nathan Frey. Generating all-atom protein structure from sequence-only training data. *bioRxiv*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly and accurately state the paper's main contributions and scope. Diverse and physically plausible conformational spaces, named ProteinConformers, are obtained for each protein by applying structural decoys and molecular dynamics simulations to CASP-provided monomer structures. It paves the way for future algorithms to jointly optimize conformational diversity and physical realism.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have disscussed the limitation in Sec. 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All code, data, and detailed experiment settings are (or will be) made available to ensure full reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have provide open access to the data and code at `https://huggingface.co/datasets/Jim990908/ProteinConformers/tree/main`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and test details are shown in the code, see Section 3.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: To evaluate the quality of the ProteinConformers dataset and gain new insights into current conformational generative models, six different models are systematically compared in Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information on the computer resources is provided in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Equations for Conformational Energy Landscape Overlap Analysis

To quantify the similarity between the protein conformations generated by AI-based models and those in the ProteinConformers dataset, the following three commonly used overlap metrics are employed: Interaction overlap, coverage, and the Jaccard index. These metrics evaluate the extent of agreement in low-energy regions between the protein conformers from different models of the same protein, based on a specified energy threshold.

Let $A = \{A_{i,j}\}$ and $B = \{B_{i,j}\}$ where $i, j \in [0, N]$, denote the two-dimensional free energy landscapes corresponding of two conformational ensembles. Each element $A_{i,j}$ and $B_{i,j}$ represents the free energy value at a specific grid point in the conformational energy landscape. For a given energy threshold τ (e.g., 40 kJ/mol), the number of shared low-energy conformations is defined as:

$$|A \cap B| = \sum_{i,j=1}^{N} \mathbf{1}[A_{i,j} < \tau \wedge B_{i,j} < \tau]$$

where $N = 63$, and $\mathbf{1}[\cdot]$ is the indicator function, which returns 1 if the condition inside is true and 0 otherwise.

The low energy area of different conformational free energy landscape under different threshold are given by:

$$|A| = \sum_{i,j=1}^{N} \mathbf{1}[A_{i,j} < \tau], \qquad |B| = \sum_{i,j=1}^{N} \mathbf{1}[B_{i,j} < \tau]$$

Using the above definitions, the overlap metrics are computed as follows:

- **Interaction**:

$$\text{Interaction} = |A \cap B|$$

- **Coverage** (proportion of low-energy conformations in A also found in B):

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

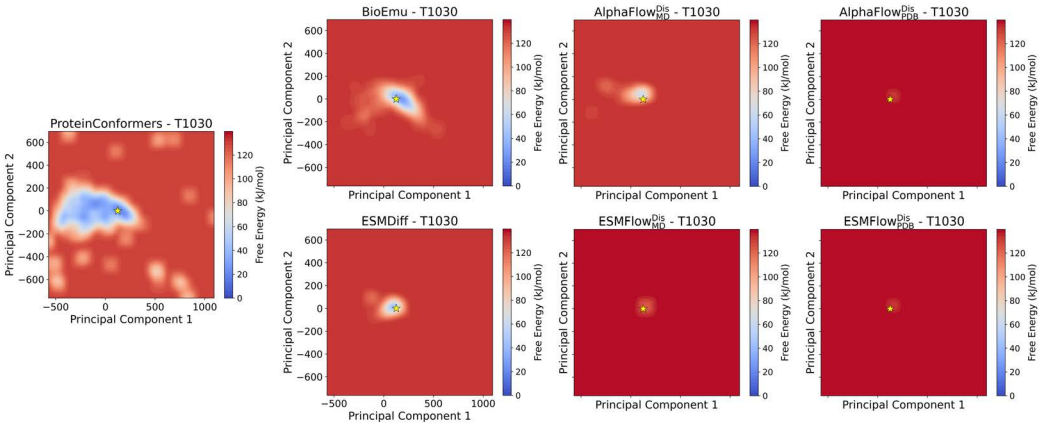
- **Jaccard Index** (symmetric overlap metric between both sets):

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

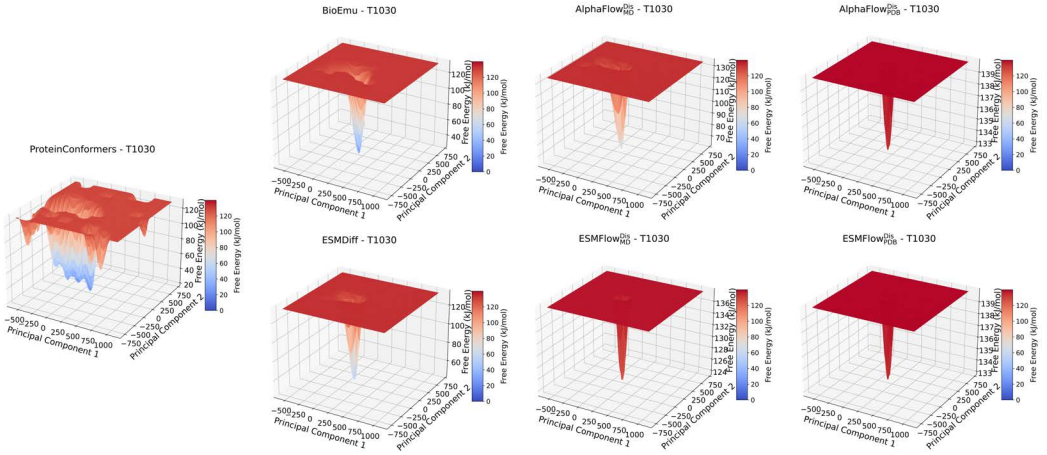
A.2 Free Energy Landscapes Comparison

This section provides additional figures comparing the conformational landscapes of protein conformers from ProteinConformers with those generated by AI models.

A.3 Overview the ProteinConformers proteins

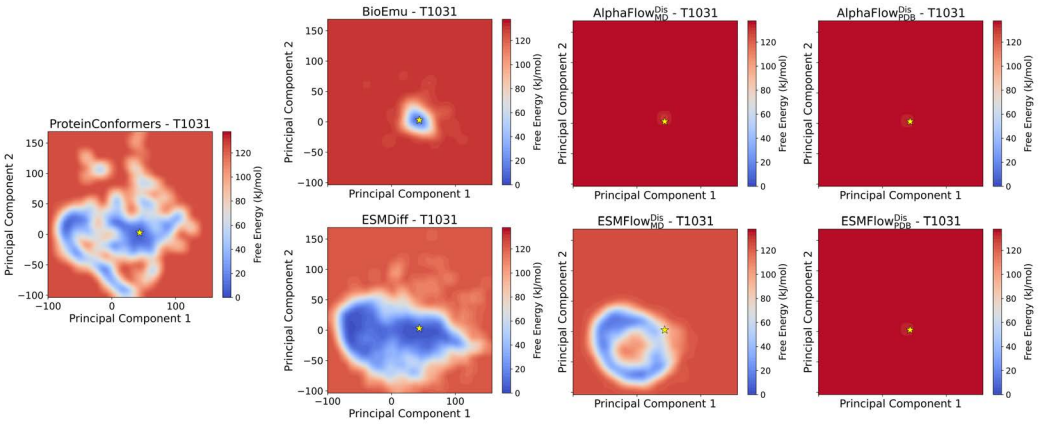


(a) Comparison of 2D conformational landscapes.

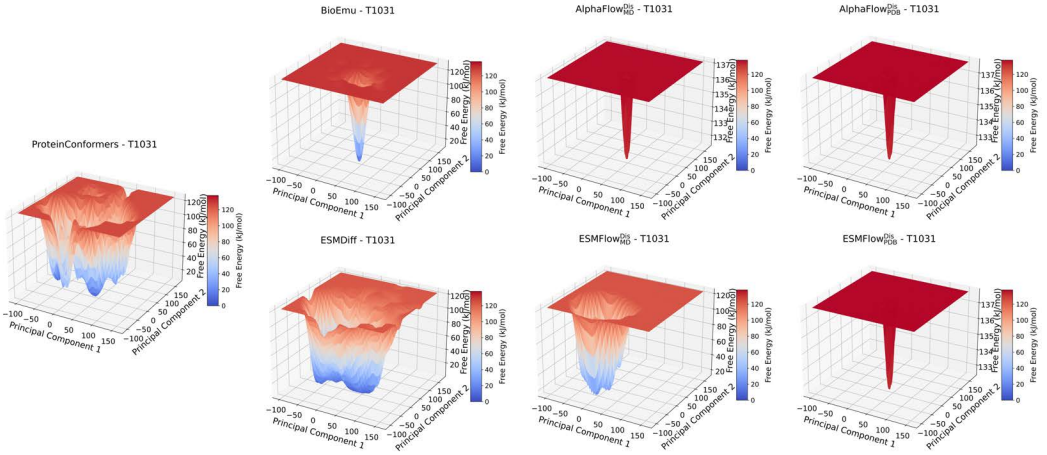


(b) Comparison of 3D conformational landscapes.

Figure 6: Comparison of conformational landscapes for protein T1030, generated by ProteinConformers and protein conformation generative models.

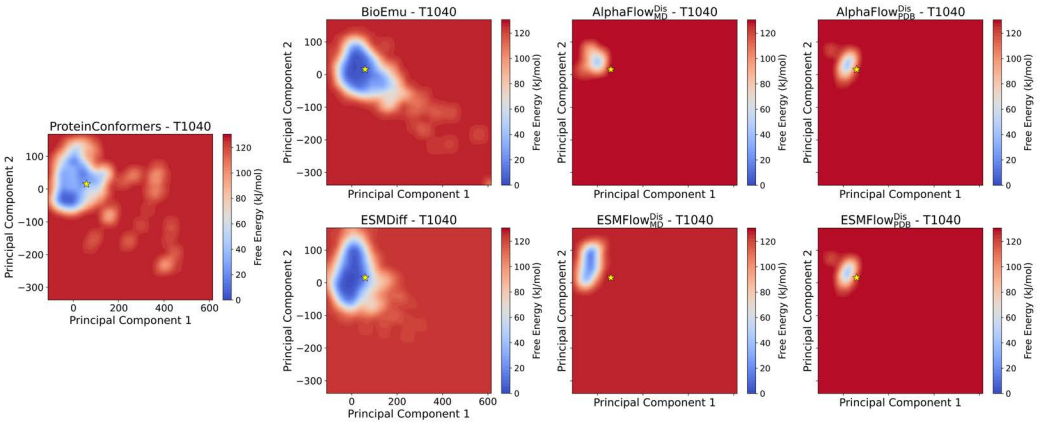


(a) Comparison of 2D conformational landscapes.

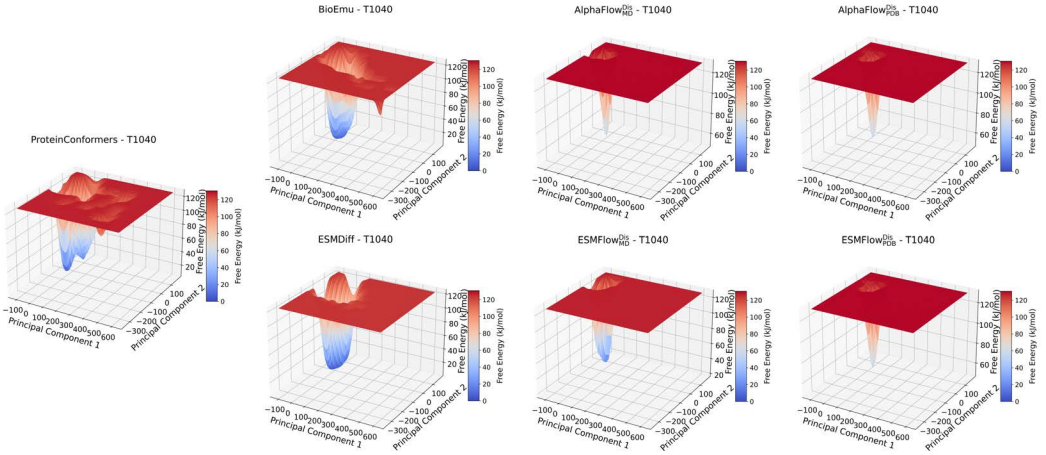


(b) Comparison of 3D conformational landscapes.

Figure 7: Comparison of conformational landscapes for protein T1031, generated by ProteinConformers and protein conformation generative models.

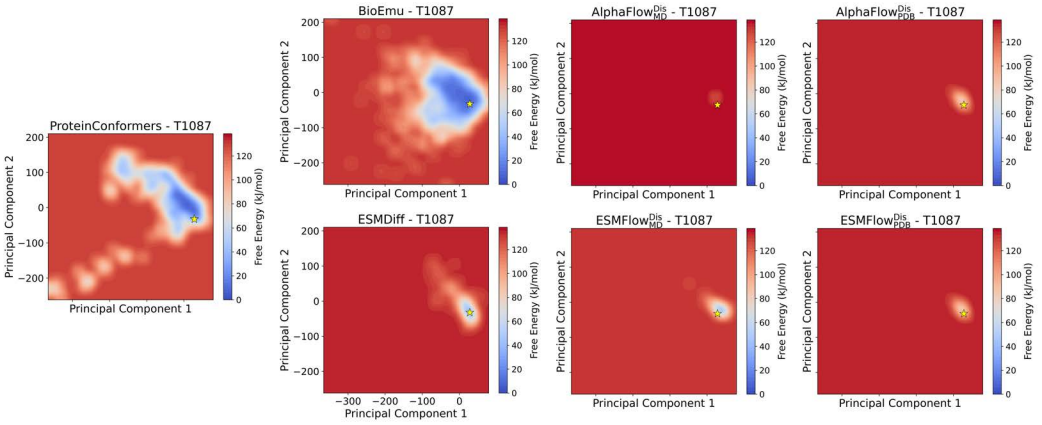


(a) Comparison of 2D conformational landscapes.

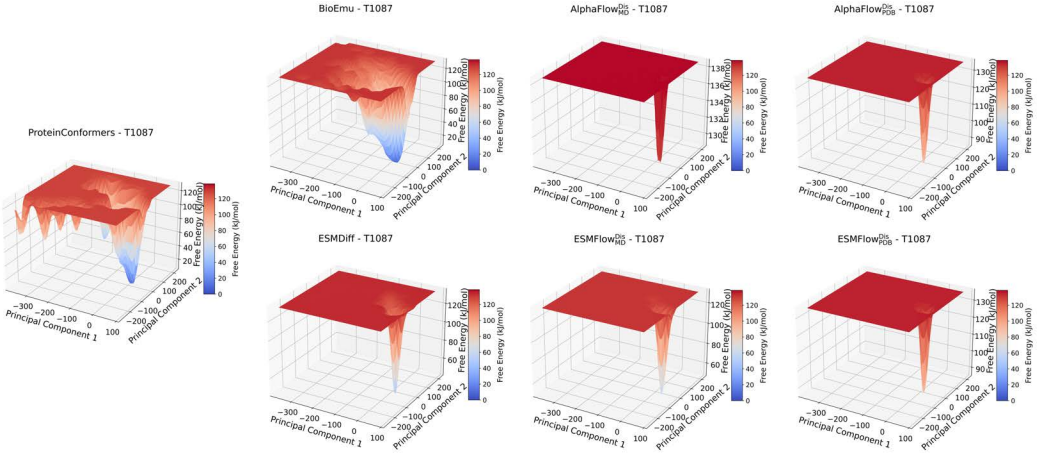


(b) Comparison of 3D conformational landscapes.

Figure 8: Comparison of conformational landscapes for protein T1040, generated by ProteinConformers and protein conformation generative models.



(a) Comparison of 2D conformational landscapes.



(b) Comparison of 3D conformational landscapes.

Figure 9: Comparison of conformational landscapes for protein T1087, generated by ProteinConformers and protein conformation generative models.

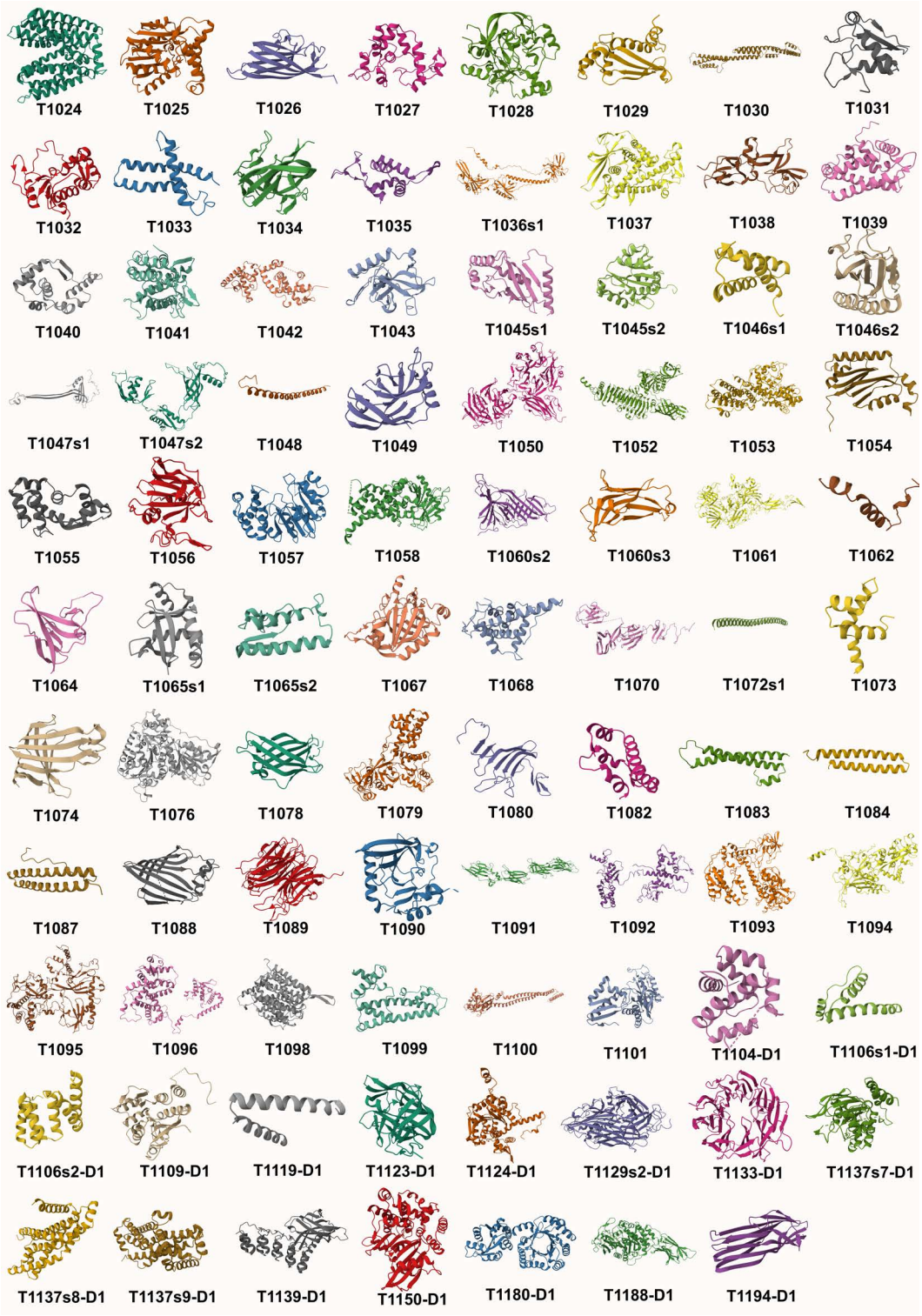


Figure 10: The 3D native structures of all 87 proteins in ProteinConformers.