Bioinformatics, 2026, **42**(4), btag156
https://doi.org/10.1093/bioinformatics/btag156
Published: 2 April 2026
Original Paper | Structural bioinformatics

OXFORD

Protein FID: improved evaluation of protein structure generative models

Felix Faltings[1,*], Hannes Stark[1], Tommi Jaakkola[1], Regina Barzilay[1]

[1]CSAIL, MIT, Cambridge, MA 02139, United States
*Corresponding Author. Felix Faltings, CSAIL, MIT, Cambridge, MA 02139, United States (faltings@mit.edu)
Associate Editor: Lenore Cowen

Abstract

Motivation: **Protein structure generative models have seen a recent surge of interest, but meaningfully evaluating them computationally is an active area of research. While current metrics have driven useful progress, they do not capture how well models sample the design space represented by the training data. We argue for a protein Frechet Inception Distance (FID) metric to supplement current evaluations with a measure of distributional similarity in a semantically meaningful latent space.**

Results: **Our FID behaves desirably under protein structure perturbations and correctly recapitulates similarities between protein samples: it correlates with optimal transport distances and recovers FoldSeek clusters and the CATH hierarchy. Evaluating current protein structure generative models with FID shows that they fall short of modeling the distribution of PDB proteins.**

Availability: **Code is available at: https://github.com/ffaltings/protfid.**

Introduction

Progress in generative protein design requires access to accurate and reliable evaluation measures. Although experimental validation remains the gold standard, in silico measures are essential to quickly develop and compare machine learning models. However, even though state-of-the-art models perform very well on current metrics, their success in practical design applications has remained relatively limited (Glasscock *et al*. 2023, Ingraham *et al*. 2023, Watson *et al*. 2023). For example, Ingraham *et al*. (2023) report an experimental success rate of generated structures of 3%, while achieving much higher in silico scores. Moreover, as models continue to improve, they are beginning to outgrow current metrics, with some recent models reporting near-perfect performance (Campbell *et al*. 2024), making it difficult to compare models. Continuing advancement in this area thus requires additional, more powerful evaluation metrics to track progress.

Currently, the most commonly used in silico metrics for protein design are designability, novelty and diversity (Ingraham *et al*. 2023, Yim *et al*. 2023a, 2023b, Lin and AlQuraishi, 2023, Watson *et al*. 2023, Campbell *et al*. 2024). A structure is considered designable if there exists a sequence that folds into it. In practice, designability is evaluated by generating sequences conditioned on the generated structure and checking whether any of the sampled sequences fold back into the given structure using a folding model. On the other hand, diversity looks at how different the model generated outputs are from each other, usually assessed by looking at the number of distinct clusters over the output space. Finally, novelty checks the number of memorized samples produced by the model.

Problematically, none of these metrics capture how well the model samples the design space represented in the training data. For example, a model could generate highly diverse, novel, and designable proteins without ever generating any beta sheets, yet beta sheets may be necessary to solve some design problems. In fact, many generative models have been observed to over-sample alpha-helices at the expense of other secondary structures (Lin *et al*. 2024, Campbell *et al*. 2024). Some authors address this evaluation issue by reporting the proportion of secondary structures in model-generated proteins, but the same problem persists for CATH domains (Orengo *et al*. 1997) or any other as-of-yet undiscovered classifications.

To address this gap, we consider using the Frechet Inception Distance (FID) introduced in Heusel *et al*. (2017) and whose application to proteins was previously explored in Geffner *et al*. (2025) and Lu *et al*. (2025). Building on this, we demonstrate the FID's utility as a new metric with extensive experiments across different design choices. The FID approximates the Wasserstein distance between the distribution generated by a model and a reference distribution. Hence, in contrast to designability, which evaluates *individual* structures, the FID considers the entire *distribution* of

Received: 9 August 2025. **Revised:** 11 March 2026. **Accepted:** 18 March 2026

structures generated by the model, where a lower FID means the model captures the reference distribution better. Note that having a low FID is different from *memorizing* the training distribution. A model can achieve a low FID while producing entirely novel structures. Our experiments will show that a PDB sample that has no structures in common with the reference distribution still achieves a better FID than current state-of-the-art models.

To overcome issues with high-dimensional data, both distributions are first embedded into a latent space that captures meaningful features of the data. The FID is then computed as the Wasserstein distance between Gaussian approximations of the two distributions. To adapt the FID metric to proteins we curate a reference set of structures from the Protein Data Bank (PDB) and explore several different methods for computing embeddings based on pretrained models, including GearNet (Zhang *et al*. 2022) and ESM3 (Hayes *et al*. 2024). Because it approximates the Wasserstein distance, the FID penalizes models if they miss any part of the reference distribution, thus addressing the issue raised above. We validate this in our experiments, which show that the FID can detect if models undersample FoldSeek (Van Kempen *et al*. 2024) or CATH clusters. Our experiments also show that the FID cannot be improved by generating invalid structures. For example, it can discriminate between PDB structures and structures folded by AlphaFold3 (AF3) (Abramson *et al*. 2024) without any multiple sequence alignments (MSAs).

We then show that state-of-the-art models still have significantly higher FIDs than natural proteins. Moreover, we find that computing FIDs using different embedding dimensions and different sampling sizes leads to consistent conclusions, including the ranking between different models. This highlights the robustness of the FID as a metric for model comparison. Finally, we connect the observed gap in FIDs between natural and generated proteins to additional statistical differences. We find that the average order of inter-residue contacts in PDB structures is much higher than in generated proteins and that generated proteins display a higher redundancy of tertiary motifs (TERMs) (Mackenzie *et al*. 2016). We additionally find that changing the generative model's sampling parameters to improve FIDs also leads to better contact orders and less TERM redundancy.

Results

We first empirically show that generative models are saturating current metrics. Following this, we present and validate the FID as a new evaluation metric. We use it to evaluate a slate of state-of-the-art generative models, and find a gap between generated structures and natural structures. We further illustrate the observed gap in FID by looking at other statistical differences. Finally, for some models with a tunable sampling hyperparameter, we show that tuning for FID also reduces these other statistical differences.

Current metrics are starting to be saturated

We show here here that recent models achieve designabilities much higher than the designability of natural proteins, making it challenging to further improve models based on designability alone. To demonstrate this, we consider single chain proteins that were released between September 2021 and March 2024—outside of the training data of the folding model we employ, but still in distribution. Each protein is a FoldSeek cluster representative of a 100–500 length protein set. FoldSeek is a fast structural clustering algorithm designed to approximate clustering based on TMScore (Zhang and Skolnick 2004), and taking cluster representatives ensures the structures are not redundant. The resulting set contains 1,024 structures. On this set we evaluate designability by sampling 8 sequences from ProteinMPNN (Dauparas *et al*. 2022), folding them with ESMFold (Lin *et al*. 2023), and reporting the original structure as designable if one of the 8 refolds is within 2 Å of the original. Fig. 1 shows the resulting designabilities broken down by length. The left panel shows the designability when using sequences generated by ProteinMPNN, and the right panel uses the native protein sequences. Across all lengths, we see that a quarter of PDB structures are not considered designable. Even for shorter proteins, the designability is much lower than that achieved by generative models, and this number drops even further when considering the native protein sequences instead of those generated by ProteinMPNN. This displays how state-of-the-art protein structure generative models such as Multiflow (Campbell *et al*. 2024) that reach 99% designability for generations of similar length scales could be considered over-optimized for this metric.

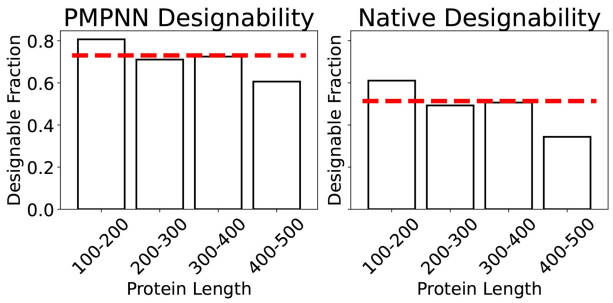


Figure 1. Designability of PDB Proteins. Fraction of designable structures in a curated set of PDB entries, broken down by length for sequences designed by ProteinMPNN and native sequences. The red horizontal line indicates the mean over the whole set.

FID metric as a useful tool for evaluation

A natural way to evaluate generative models is to measure the distributional discrepancy between the target distribution we want to sample from and the distribution of samples actually produced by the model. There are various ways to measure differences between probability distributions, such as the Kullback-Leibler divergence, or Wasserstein distance. However, estimating these for high dimensional distributions like protein structures is exceedingly difficult. The key idea of FID is to instead measure the difference in a lower dimensional embedding space using Gaussian approximations. Concretely, given two sets of samples, we first compute their embeddings. The distributions of the embedded samples are then approximated by Gaussians, and the 2-Wasserstein distance between the Gaussians, which can be computed with a closed-form formula,

is the FID. For images, the embedding model is a pretrained image classifier (Heusel *et al*. 2017), and generative models are evaluated by computing the FID to a reference set of images. For proteins, we explored several different embedding models and found GearNet to work well. We use this in all of the following results, but investigate other embedding methods in the Appendix, available as supplementary data at *Bioinformatics* online. As detailed in the Methods, we also found it useful to first project the GearNet embeddings to a lower dimension. In our results we use a dimension of 32, but we sweep other dimensions in the Appendix, available as supplementary data at *Bioinformatics* online. We find that the FID is highly robust, exhibiting consistent behavior down to as few as 4 dimensions, as demonstrated in the next section when we evaluate current generative models. Finally, we construct a reference set of structures by taking 4,991 filtered structures from the PDB, each one of which is a FoldSeek cluster representative. In addition to this, we construct a separate test set of 467 cluster representatives with no structural overlap to the reference. See Methods for details.

FID detects perturbations

We first check that the FID can discriminate between physically plausible and implausible structures. As our plausible structures, we take our test set of structures. To obtain implausible structures, we apply the following perturbations to the test set,

1) *Jitter*: We apply Gaussian noise to the positions of the protein's atoms.
2) *Refolding*: We fold the proteins using AlphFold3, but without using any MSAs. These refolded structures constitute a form of perturbation, since their RMSDs to the original PDB structures are significantly worse than when folding using MSAs.

Because the resulting perturbed structures will be out of distribution for the pretrained embedding model, their embeddings will differ from the embeddings of natural proteins and we should be able to detect this in the FIDs. The results in Fig. 2 show that the FID not only detects these perturbations, but also increases monotonically with the severity of the perturbation. It is therefore not possible to improve the FID by sacrificing the quality of the generated structures.

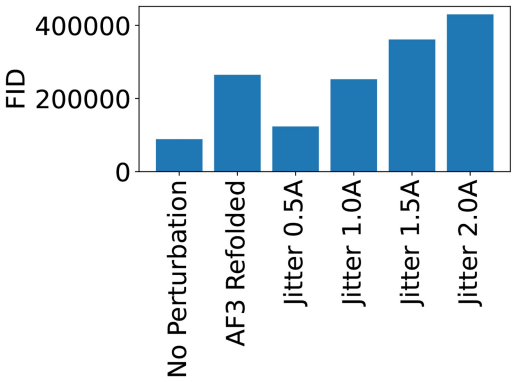


Figure 2 FID detects perturbations. FIDs of perturbed PDB samples. A random sample of 1,000 proteins was taken from the PDB. FIDs were computed between the reference and perturbed and unperturbed versions of the structures.

FID recapitulates FoldSeek clustering

Because the FID is computed in a latent space, it is possible that the embeddings lose some structural information. We therefore verify that the FID can discriminate between more or less similar sets of structures by checking that it recapitulates the FoldSeek clustering.

Each FoldSeek cluster represents a distribution of similar structures, where each cluster is more similar to itself than to other clusters. We check whether the FID recapitulates this by looking at the FIDs between random samples from FoldSeek clusters. For each cluster, we draw two random non-overlapping samples. We then compute FIDs between all pairs of clusters, where we always compute the FID between the first and second samples so that on the same cluster we compute the distance between two *different* samples.

We also check whether the FID can capture the *degree* to which different clusters differ from each other based on the TMScore, a trusted measure of structural similarity. Because the TMScore is a metric between structures and not distributions, we compute the optimal transport distance between clusters using the TMScore as the base metric, which we call the OT-TMScore. See Methods for details. Intuitively, this metric looks at how well two samples can be soft matched based on TMScores.

As can be seen in Fig. 3, the FID between two samples from the same cluster is much lower than between two different clusters, thereby recapitulating the FoldSeek clustering. The FID also shows agreement with the OT-TMScore, with a Pearson correlation coefficient of 0.61 ($p = 1.94 \times 10^{-22}$) and 0.2 ($p = 0.28$) when only considering different clusters (i.e. off diagonal entries in the figure). In Fig. 13 in the appendix, available as supplementary data at *Bioinformatics* online we plot the FID against the OT-TMScore and see that the FID picks up on clusters deemed very different based on the OT-TMScore.

FID captures fold diversity at multiple levels

As a distributional metric, the FID should detect whether a generative model is undersampling the variety of folds represented in the data distribution. For example, it should penalize the

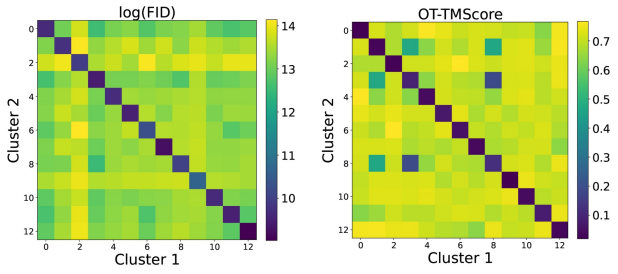


Figure 3 FID recapitulates FoldSeek clusters and correlates with OT-TMScore. *Left*: Pairwise FIDs between FoldSeek clusters. Note that FID is displayed in natural log-scale for better visualization. *Right*: Pairwise OT-TMScores between FoldSeek clusters.

model for undersampling CATH clusters. To confirm this, we propose a *diversity race* experiment.

Consider a set of clustered structures. We can compute the FID of the set to a reference set after successively removing clusters in a random order. We can then compare this against removing structures at random, without regard for the clustering. Ideally, the FID should increase faster when removing entire clusters, since removing structures at random would leave most clusters represented in the remaining set.

We can think of this as a race taking place over multiple rounds. Each racer is a set of structures. At the first round, the racers are identical. At each subsequent round, each racer drops a subset of structures. One racer will drop a cluster of structures, while the other will drop a random subset. In both cases, the FID to the reference set will increase, but they may do so at different rates.

With this analogy, we can generalize to a hierarchical clustering like CATH, where we now have more than two racers–one for each level of the hierarchy, and one baseline racer that drops structures at random. Here, we would expect the ranking of the racers to reflect the CATH hierarchy, where the higher level racer should win out, followed in order by the lower levels, and, finally, the baseline. Indeed, removing all structures from a broader category should give a higher FID than removing many narrower categories.

Because the sizes of the clusters are not uniform, dropping different clusters will result in sets of different sizes, which may affect the FIDs. A racer should not win simply because it drops more structures in each round. To account for this, each racer always drops the same fixed number of structures in each round, but prioritizes which structures to drop based on the clustering. See Methods for a detailed description.

In Fig. 4, we carry out two such diversity races. One for a FoldSeek clustering, which results in 2 racers–the clustered order and the random order–and the hierarchical CATH clustering, which results in 5 racers–one for each of the four levels in CATH and the random order. To estimate standard deviations, we rerun the races 50 times, which allows us to ascertain that our FIDs recover both of the clusterings and meaningfully capture structural diversity. The fact that the racers in the CATH experiment are ordered according to the hierarchy of classes also suggests that the FID captures diversity at *multiple* levels.

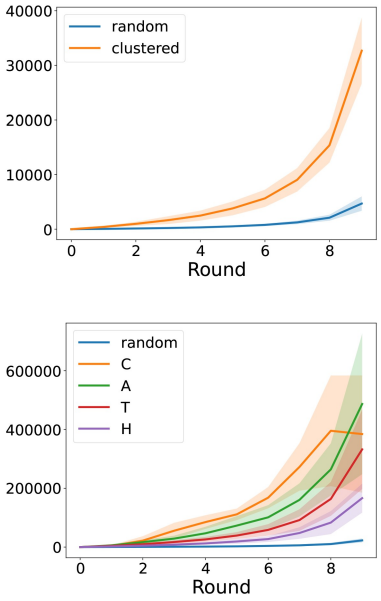


Figure 4 Diversity races. In our "races," we increase the sizes of several "racer" sets (starting from empty sets) by adding samples from a fixed reference set. We observe how their FIDs to the reference set drop at different speeds since some racers receive less diverse samples from clustered versions of the reference set, and since our FIDs capture meaningful structural diversity. *Right*: Race conducted with CATH clusters, with one racer for each level of the hierarchy.

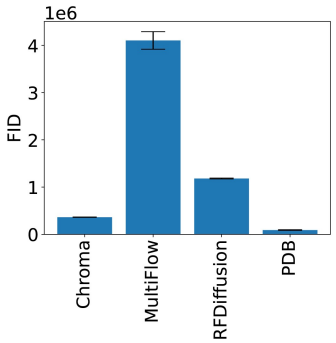


Figure 5 FIDs of Generative Models. FIDs achieved by state-of-the-art protein structure generative models and a set of PDB structures distinct from the reference set. The lengths of the generated proteins were chosen to match the length distribution of the PDB sample.

Evaluating state-of-the-art generative models

Having validated the FID as a metric, we use it to evaluate several state-of-the-art generative models. Our results show a large remaining FID gap between generated and natural structures, which is corroborated by our observation that generated structures have lower complexity and higher substructure redundancy compared to PDB structures.

Current generative models are still far from natural proteins

We use the FID to evaluate samples from the following generative models: MultiFlow (Campbell *et al.* 2024), RFDiffusion (Watson *et al.* 2023), and Chroma (Ingraham *et al.* 2023), which we compare to our PDB test set. For each model, we generate 467 structures, to match our test set. We control for sequence length by generating one sample for each length observed in the test set, so that the resulting sets all have the same distribution of lengths. The results in Fig. 5 show that state-of-the-art models still achieve FIDs substantially worse than natural proteins. For reference, the gap between Chroma, the best performing model, and natural structures is around 1.5 times that of structures refolded by AF3 without MSAs. In other words, the distribution generated by chroma is about 1.5 times worse than the distribution of structures obtained by folding natural

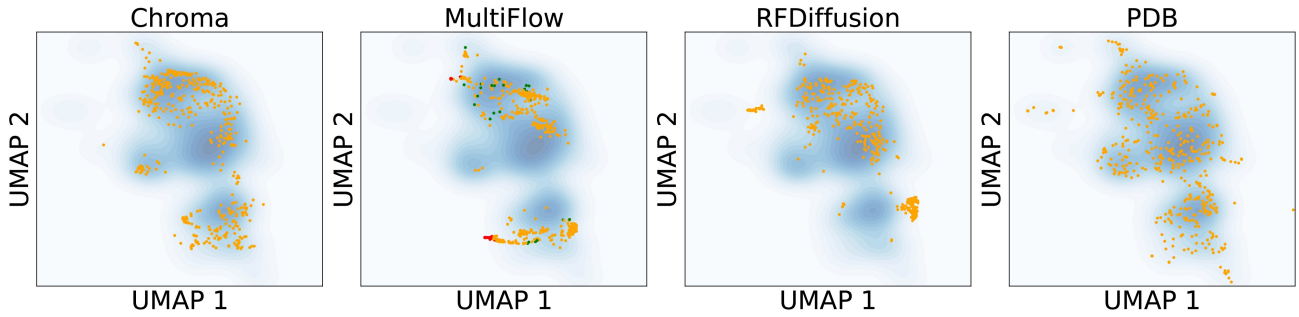
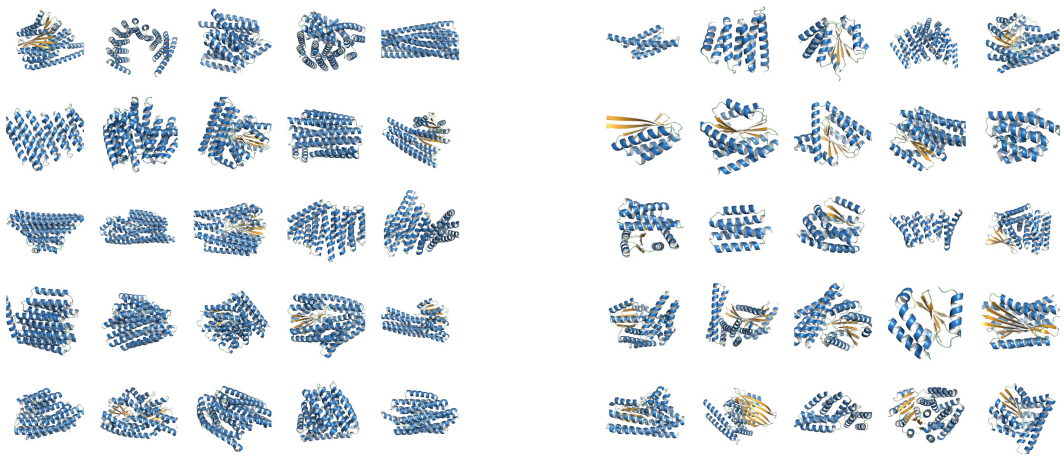


Figure 6 Embedding UMAPs. We visualized the UMAP embeddings of the representations for different samples of generated structures. The blue coloring shows a kernel density estimate of the reference PDB distribution. The orange points show the samples. For MultiFlow, the red points show the points with a negative influence on the FID (left side of Fig. 7) and green points show samples with a positive influence (right side of Fig. 7).



(a) MultiFlow samples with negative influence on FID. (b) Multiflow samples with positive influence on FID.

Figure 7 FID prefers structures with diverse motifs. Samples whose weights had a positive gradient (left) and negative gradient (right) for the FID. A positive gradient indicates samples that have a negative effect on the FID and vice versa for negative gradients.

proteins with AF3. The distribution of MultiFlow, the worst model, is about 23 times worse.

We also visualize the embeddings of the generated structures and the PDB structures in Fig. 6, where it is clear that the PDB structures better overlap with the reference set. We also see that Chroma has the best agreement with the reference, followed by RFDiffusion and Multiflow, which is consistent with the FID ranking.

Finally, in order to better illustrate the gap between generated structures and natural structures, we estimate which samples improve or worsen the FID to the reference set the most. We first assign weights to each generated sample, indicating how much each sample contributes to the FID. For example, assigning a weight of 0 is the same as removing a sample entirely. We then consider the derivative of the FID with respect to these weights and visualize which samples have a positive or negative influence, corresponding respectively to negative and positive derivatives. See Methods for more details. We see in Fig. 7 that for MultiFlow, the samples with a negative influence all consist of large alpha-helical bundles, whereas the positive influence samples are more diverse, also confirming our previous observation that the generative models tend to undersample beta sheets. However, we note that the FID is not only capturing a difference in secondary structure, since some of the negative samples also contain beta sheets.

These samples are also shown in the UMAP plots of Fig. 6. We see there that the negative samples all lie on the edge of the reference distribution, while the positive samples tend to fall in under sampled regions of the reference.

FID is robust to sample sizes and embedding dimensions

As an evaluation measure, it is important that the FID gives consistent results. We thus recompute the FIDs of the generative models above using varying sample sizes and varying embedding dimensions as shown in Fig. 12 in the Appendix, available as supplementary data at *Bioinformatics* online. We see that even with a sample size as small as 50 structures, the ordering of the models is preserved and the variances are relatively small. Similarly, when projecting to as low as 4 PCA dimensions, we

still observe a consistent ranking between models. This demonstrates the robustness of the FID as an evaluation metric.

Generative models generate proteins with lower complexity than natural proteins

Another possible explanation for the gap in FIDs between generated and natural proteins is that natural proteins may be more complex. As a proxy to address the complexity of protein structures, we count the number of higher order contacts in the structure. More specifically, we consider potential contacts (PCs) as defined in Mackenzie *et al*. (2016), where two backbone positions form a PC if their side chains are close to each other when considering all combinations of rotamers and side chains at both positions. We then define the order of a contact as the number of distinct secondary structures that appear between the contact positions. For example, a contact that occurs within the same secondary structure, such as in an alpha-helix, would have an order of 0. The average and median contact orders for samples from the PDB and generative models are reported in Table 1. We see that on average, the PDB contains structures with significantly higher order contacts than generated proteins.

Table 1 Natural structures have more high order contacts than generated structures.

	Contact Order	
Set	**Mean**	**Median**
multiflow	3.3	2
chroma	4.0	2
rfdiffusion	1.9	0
pdb	7.6	3

Average and median order of contacts across all proteins in a random sample from the PDB and different generative models. A contact is a pair of backbone positions that could potentially form a contact. The order of the contact is the number of secondary structures between the pairs of positions. Contacts in the same secondary structure have order 0.

Generative models generate proteins with low substructure diversity

To explain the gap in FIDs between natural and generated proteins, we compare their diversity of substructures. Specifically, we consider the number of tertiary structural motifs (TERMs) from Mackenzie *et al*. (2016) needed to explain each set of structures. A TERM is a small neighborhood around an amino acid, including potentially contacting amino acids. A structure can then be seen as a set of backbone atom positions and potential contacts (PCs) as defined above. A TERM is said to cover a set of positions and PCs if it can be aligned to those positions with a low enough RMSD. A structure is said to be explained by a set of TERMs if they cover all its backbone positions and PCs (Mackenzie *et al*. 2016). To assess the substructure diversity of a set of structures, we take the top 1000 TERMs from the original paper (Mackenzie *et al*. 2016). We then look at how much of the structures can be explained using those TERMs. As can be seen in Fig. 10, the generated proteins can be explained with much fewer TERMs than the PDB sample, indicating they contain a lower diversity of substructural motifs.

Improving FID can improve substrcture diversity and complexity

Frame-based flow models (Bose *et al*. 2023, Campbell *et al*. 2024) have previously found that using different hyperparameters at sampling time than at training time leads to better designability. In particular, the noise schedule of the frame rotations is annealed much faster than the translations. Here we show that, conversely, using the original noise schedule improves FIDs. We sample from MultiFlow either with the hyperparameters from the paper (Exponential), or the hyperparameters used during training (Linear), and find that the training hyperparameters lead to better FIDs (Fig. 8a), higher average contact orders (Fig. 8b), and more substructural diversity (Fig. 8c), but lower designability. We also see improved coverage in the UMAP projections (Fig. 9). Thus, we see that changing the sampling schedule to improve FIDs also improves the substructure diversity and complexity of generated proteins.

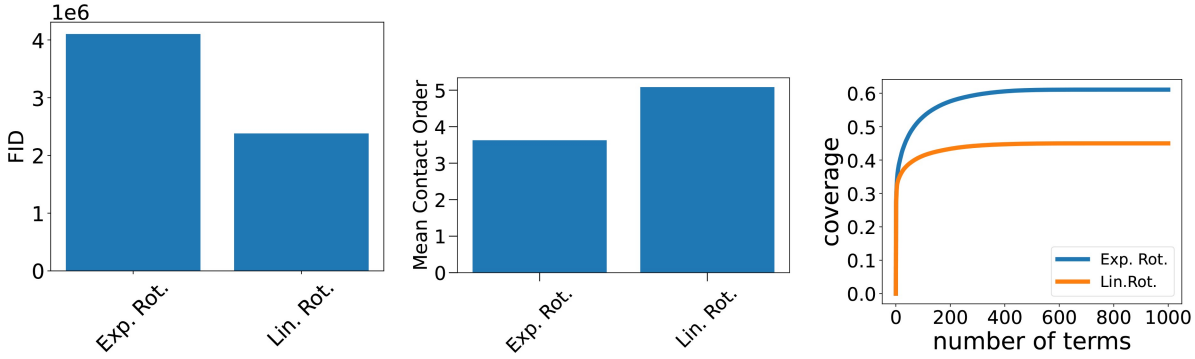


Figure 8 Effect of MultiFlow Sampling Parameters. We compare samples from MultiFlow using the exponential schedule for rotations used in the original paper, versus the linear schedule the model was trained with. We report the FID (a), average contact order (b) and TERM coverage (c). The linear schedule leads to better results across all three metrics.

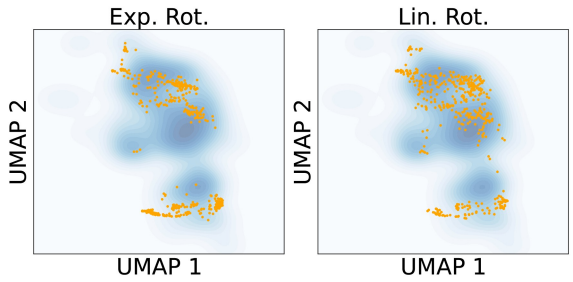


Figure 9 Changing sampling hyperparameters increases overlap of samples with reference. UMAP embeddings of proteins generated by MultiFlow when using the exponential rotation schedule from the paper versus the linear rotation schedule used during training. We see better overlap between model generated samples and reference clusters when using the linear schedule.

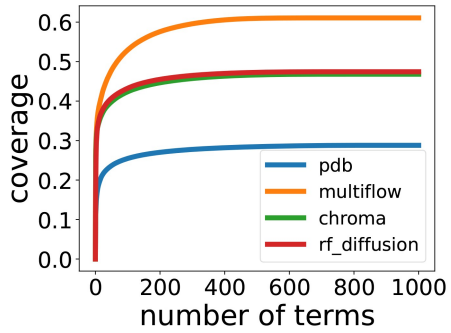


Figure 10 TERM coverage. Number of TERMs from Mackenzie et al. [2016] needed to cover samples from the PDB and generative models.

Discussion

In this work, we argue that current evaluation metrics of generative protein structure models insufficiently evaluate how well generated structures represent the training data. We thus propose to use the FID, which has proven successful in computer vision. We show that the FID can penalize models for lack of diversity, or for missing parts of the training distribution, and discriminates between physically plausible and implausible structures. Novel structures also still achieve good FIDs, showing that the FID does not simply reward memorization of training samples. We then re-evaluate state-of-the-art generative models in terms of FIDs and find that the generated structures are still far from natural structures, even though they achieve near-perfect designability. We also find that these results are robust, giving consistent results for different sample sizes or embedding dimensions used to compute the FIDs. Moreover, we find that we can improve the FID of some of these models by changing their sampling hyperparameters, leading to a corresponding improvement in substructure diversity and structure complexity.

One limitation that is not fully explored in this work is the embeddings used to compute the FID. While we use pretrained embedding models, there may be better ways to learn these embeddings specifically for the purpose of computing FIDs. These embeddings could then also be tuned towards specific types of errors that protein design practitioners care about.

In conclusion, the FID appears to be a viable evaluation measure, and a strong complement to existing measures like designability. This additional metric is all the more pertinent as current metrics begin to saturate. Indeed, we found that recent models achieve designabilities close to 99%, whereas natural proteins in the PDB of a similar length are only around 80% designable. This makes it unclear whether further optimizing designability will lead to meaningful progress. Reducing the FID gap between natural and generated proteins should thus serve as a new target for improving generative models.

Methods

Here we present all the computational methods used to produce our results.

FID formula

Given two samples of structures $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_M\}$, we first compute embeddings $E(x_i) \in \mathbb{R}^p$, and $E(y_j) \in \mathbb{R}^p$, of dimension p. Let E_X denote the $p \times N$ matrix of embeddings $[E(x_1), \ldots, E(x_N)] \in \mathbb{R}^{p \times N}$, and similarly for E_Y. We approximate the distributions of the embeddings using Gaussians. This amounts to estimating parameters,

$$\mu_X = \frac{1}{N}\sum_{i=1}^{N} E(x_i), \quad \Sigma_X = \frac{1}{N}\sum_{i=1}^{N}(E(x_i) - \mu_X)(E(x_i) - \mu_X)^T, \qquad (1)$$

with μ_Y and Σ_Y defined similarly. The FID is then the Wasserstein distance between these Gaussian approximations, which is given by the formula,

$$\mathcal{W}_2(\mu_X, \Sigma_X, \mu_Y, \Sigma_Y) = \|\mu_X - \mu_Y\|_2^2 + \mathrm{Tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_Y\Sigma_X)^{1/2}). \qquad (2)$$

Embeddings

In our experiments, we use the embeddings from GearNet. Since GearNet only takes the backbone as input, we can evaluate any generative model that generates protein backbones. This includes models like Chroma, RFDiffusion and Multiflow. We found that the covariance matrices of the embeddings were nearly degenerate, with only a few dominant principal components (see Fig. 14 in the appendix, available as supplementary data at *Bioinformatics* online). This lead to numerical issues when computing the matrix square roots in the FID formula. We thus found it beneficial to project onto the first 32 principal components, giving embeddings of dimension 32. We compute the PCA using the embeddings of both the reference set and the evaluation set in order to emphasize differences between the two.

Reference set

In order to construct our reference set, we start from the filtered training set of OpenFold (Ahdritz *et al*. 2022), filtered for structures of 100–500 residues obtained either via X-ray diffraction or electron microscopy with a resolution under 3 Å. We cluster the

resulting set using FoldSeek, and only keep structures with a single identified protein chain, since all the generative models we consider only generate single-chain structures. We cluster the resulting set using FoldSeek and split into a reference and test set. Finally, we sample a representative from each cluster to give a representative, diverse, and non-redundant set of structures. This results in 4,991 reference structures and 467 test structures. We verify that the test set does not have significant structural overlap with the reference by computing the highest TMScore of each structure in the test set to any structure in the reference using the Foldseek easy-search software (Van Kempen *et al*. 2024). The resulting mean TMScore of 0.24 indicates there is no significant overlap.

Optimal transport TM score

The TMScore is a similarity metric between two proteins structures. Here we describe how to turn it into a metric between sets of structures. Given two sets of proteins $P_1, \ldots, P_N$ and $Q_1, \ldots, Q_M$, we can compute a matrix of pairwise TMScores,

$$D_{ij} = \text{TMScore}(P_i, Q_j). \tag{3}$$

We turn this into a symmetric[1] cost matrix,

$$C_{ij} = 1 - (D_{ij} + D_{ji})/2 \tag{4}$$

The OT-TMScore between the two samples is the solution to the optimal transport problem,

$$\begin{aligned} \min_{P_{ij}} \quad & \sum_{ij} P_{ij} C_{ij} \\ \text{subject to} \quad & \\ & \sum_{j} P_{ij} = 1 \\ & \sum_{i} P_{ij} = 1 \\ & P_{ij} \geq 0 \end{aligned}$$

Informally, this metric looks at how well the structures in each sample can be (soft) matched to each other based on the TMScore.

Diversity race

Here we describe the details of the diversity race in Section FID Metric as a Useful Tool for Evaluation. The race proceeds over L rounds. Each racer, $r_1, \ldots, r_M$ is a set of structures. Let r_i^j denote racer i at round j. The sets are decreasing in size, $r_i^1 \supset r_i^2 \supset \ldots \supset r_i^L$, and each racer starts off identical $r_i^1 = r$, where r contains N structures. At each round, we compute the FID to a reference set R, $s_i^l = \text{FID}(r_i^l, R)$.

Each racer differs in which structures are dropped at each round. Consider racer i, which prioritizes dropping structures according to clusters, $c_1, \ldots, c_K$, where $\cup_{i=1}^K c_i = r$. We order the N structures in r according to the clustering. If $n_1, \ldots, n_K$ are the sizes of the clusters, then the first n_1 structures are from cluster c_1, the next n_2 from c_2, and so on. The racer then drops structures according to this ordering, thus prioritizing dropping a whole cluster before moving on to the next one.

[1] The TMScore is not symmetric.

FID optimization

In order to probe the influence of each sample on the FID, we compute the FID using a weighted sample and consider the derivative of the FID with respect to the weights. Samples with a negative impact on the FID should receive positive gradients (increasing the FID is worse). We largely follow the methodology in Kynkäänniemi *et al*. (2022). Given a sample of structures x_i, we compute embeddings $E(x_i)$. We then assign weights w_i, and compute the weighted mean,

$$\mu_X(w) = \sum_{i=1}^{N} w_i E(x_i), \tag{5}$$

And covariance,

$$\Sigma_X(w) = \sum_{i=1}^{N} w_i (E(x_i) - \mu_X)(E(x_i) - \mu_X)^T, \tag{6}$$

Where we parameterize the weights as $w_i = \frac{1}{\sum_{i=1}^{N} \exp(u_i)} \exp(u_i)$ so that they are nonnegative and sum to 1. We then compute the FID as before using a differentiable implementation and compute the derivatives when $w_i = 1, \forall i$.

TERM analyses

TERM analyses are performed using the MASTER (Zhou and Grigoryan, 2020) and ConFind (Holland *et al*. 2018) software packages to respectively search for TERM matches and identify potential contacts. To compute the coverage of a target set of structures, we first identify all potential contacts and residues, forming a set of elements U. We then take the 1,000 first TERMs from Mackenzie *et al*. (2016) to use as query TERMs. Using MASTER, each query is matched to TERMs in the target set. Potential contacts and residues in the matched target TERMs are considerd covered by the query. To estimate how to cover U with the least number of query TERMs, we greedily select TERMs that cover the most remaining elements in U, as done in Mackenzie *et al*. (2016).

Acknowledgements

We thank Sergey Ovchinnikov for helpful discussions, Gevorg Grigoryan for help with the TERM analyses, and the reviewers and editor for their constructive feedback.

Author contributions

Felix Faltings(Conceptualization [Equal], Data curation [Equal], Formal analysis [Lead], Methodology [Equal], Validation [Lead], Visualization [Lead], Writing—original draft [Lead], Writing—review & editing [Equal]), Hannes Stärk(Conceptualization [Equal], Data curation [Supporting], Formal analysis [Supporting], Investigation [Supporting], Methodology [Supporting], Validation [Supporting], Visualization [Supporting], Writing—

original draft [Supporting]), and Tommi S Jaakkola(Funding acquisition [Equal], Resources [Equal], Supervision [Equal], Writing—review & editing [Equal]), Regina Barzilay(Funding acquisition [Equal], Resources [Equal], Supervision [Equal], Writing—review & editing [Equal])

Supplementary data

Supplementary data are available at Bioinformatics online.

Conflict of interest

None declared.

Funding

This work was supported by the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) consortium; the NSF Expeditions grant (award 1918839) Understanding the World Through Code; DSO Singapore grant on next generation techniques for protein ligand binding; a grant from Siemens Corporation for inverse design; the Abdul Latif Jameel Clinic; and a grant from Quanta.

Data and software availability

Code is available at https://github.com/ffaltings/protfid. Code for reproducing experiments is available on zenodo (10.5281/zenodo.18911655).

References

Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* 2024;**630**:493–500.

Ahdritz G, Bouatta N, Floristean C *et al.* OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat Methods* 2024;**21**:1514–24.

Bose AJ, Akhound-Sadegh T, Huguet G *et al.* Se (3)-stochastic flow matching for protein backbone generation *arXiv preprint arXiv : 2310.02391*, 2023. preprint: not peer reviewed.

Campbell A, Yim J, Barzilay R *et al.* Generative flows on discrete state-spaces: enabling multimodal flows with applications to protein co-design. In: *Proceedings of the 41st International Conference on Machine Learning*, pp. 5453–512, 2024.

Dauparas J, Anishchenko I, Bennett N *et al.* Robust deep learning–based protein sequence design using proteinmpnn. *Sci* 2022;**378**:49–56.

Geffner T, Didi K, Zhang Z *et al.* Proteina: Scaling flow-based protein structure generative models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Glasscock CJ, Pecoraro R, McHugh R *et al.* Computational design of sequence-specific DNA-binding proteins. *Nat Struct Mol Biol* 2023;**32**:2252–61.

Hayes T, Rao R, Akin H *et al.* Simulating 500 million years of evolution with a language model. *Sci* 2025;**387**:850–8.

Heusel M, Ramsauer H, Unterthiner T *et al.* GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Adv Neural Inf Process Syst* 2017;**30**.

Holland J, Pan Q, Grigoryan G. Contact prediction is hardest for the most informative contacts, but improves with the incorporation of contact potentials. *PLoS One* 2018; **13**:e0199585.

Ingraham J, Baranov M, Costello Z *et al.* Illuminating protein space with a programmable generative model. *Nature* 2023; **623**:1070–8.

Kynkäänniemi T, Karras T, Aittala M *et al.* The role of imagenet classes in Fréchet inception distance. In: *The Eleventh International Conference on Learning Representations*, 2022

Lin Y, AlQuraishi M. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. In: *International Conference on Machine Learning.* PMLR, 2023.

Lin Y, Lee M, Zhang Z *et al.* Out of many, one: designing and scaffolding proteins at the scale of the structural universe with genie 2. 2024. *arXiv preprint arXiv: 2405.15489*. preprint: not peer reviewed.

Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Sci* 2023;**379**:1123–30.

Lu T, Liu M, Chen Y *et al.* Assessing generative model coverage of protein structures with shapes. *Cell Systems* 2025;**16**.

Mackenzie CO, Zhou J, Grigoryan G. Tertiary alphabet for the observable protein structural universe. *Proc Natl Acad Sci U S A* 2016;**113**:E7438–47.

Orengo CA, Michie AD, Jones S *et al.* Cath–a hierarchic classification of protein domain structures. *Struct* 1997;**5**:1093–108.

Van Kempen M, Kim SS, Tumescheit C *et al.* Fast and accurate protein structure search with foldseek. *Nat Biotechnol* 2024; **42**:243–6.

Watson JL, Juergens D, Bennett NR *et al.* De novo design of protein structure and function with rfdiffusion. *Nature* 2023; **620**:1089–100.

Yim J, Campbell A, Foong AYK *et al.* Fast protein backbone generation with se(3) flow matching. arXiv preprint arXiv:2310.05297. 2023a.

Yim J, Trippe BL, De Bortoli V *et al.* Se (3) diffusion model with application to protein backbone generation. In: *International Conference on Machine Learning*. PMLR, 2023b.

Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct Funct Bioinforma* 2004;**57**:702–10.

Zhang Z, Xu M, Jamasb A *et al.* Protein representation learning by geometric structure pretraining. In: *The Eleventh International Conference on Learning Representations*, 2022.

Zhou J, Grigoryan G. A c++ library for protein sub-structure search. *bioRxiv*, 2020;2020–04.

Bioinformatics, 2026, 42, 1–9
https://doi.org/10.1093/bioinformatics/btag156
Original Paper