Evaluation Metrics for Protein Structure Generation

Anonymous Authors[1]

Abstract

Generative models have become increasingly popular for sampling novel proteins. To compare and evaluate these models, we need metrics that can assess the quality of the generated structures. We propose a set of standardized metrics for benchmarking protein generation. We experimentally show that these metrics can measure differences between proteins on a distributional level, as well as quantify the novelty, diversity and designability of the generated proteins.

1. Introduction

Proteins are biological macromolecules, made up of a sequence of amino acids with 20 different naturally occurring types, and serve an incredibly diverse set of functionalities, playing a role in almost all biological processes. Deep learning has recently made tremendous progress in protein science from determining the three-dimensional structure of a protein given its sequence (Jumper et al., 2021) to the task of protein-ligand docking (Corso et al., 2023). Furthermore, an increasing number of works are using generative models to computationally generate new but physically realistic protein structures that could lead to new treatments and speed up drug discovery. Recent protein generation methods have utilised generative adversarial networks (Anand & Huang, 2018), variational autoencoders (Harteveld et al., 2022) and now most commonly diffusion models (Anand & Achim, 2022; Ingraham et al., 2022; Trippe et al., 2023; Watson et al., 2022; Wu et al., 2022a). In protein generation, we are given a set of known proteins from an unknown distribution $p_{data}(\mathbf{x})$ found in nature. The goal of protein generation is then to sample new proteins from the same distribution. Although many works are tackling this problem, we currently have limited methods for assessing the outputs of these models in-silico. This is critical for comparing models and furthering the field, as well as to decrease experimental time when evaluating the vast number of sampled proteins. Standardised benchmarks enable quantitative comparisons between independently proposed models and have facilitated methodological progress in computer vision (Deng et al., 2009) and protein structure prediction (Moult et al., 2020) in the past. Some standard benchmarks have been developed for sequence generation (Castorina et al., 2023). We focus on assessing the ability of protein structure generative models to sample new proteins from the distribution of proteins found in nature. For this, we want metrics which can measure if sampled proteins are physically realistic, novel, diverse and designable.

Our *contributions* are as follows:

- We extend and propose novel metrics for the evaluation of protein generative models.
- We show that our distributional measure is more sensitive to changes in protein topology than previous approaches.
- We experimentally verify that our metrics measure the novelty, diversity, and designability of sampled proteins.
- We implement our metrics as a benchmark suite for the comparison of protein generative models.

2. Background

In other areas such as molecule, image, and graph generation, models are assessed on their ability to mimic the training distribution (distributional learning) (Preuer et al., 2018; Southern et al., 2023). For protein structure generation, we also care about domain specific metrics such as the novelty, designability and diversity of the generated proteins. We provide a background on current metrics used and the limitations for each.

2.1. Metrics for Protein Generation

2.1.1. Distribution learning

For comparing distributions of proteins, we want a metric which is zero when the distributions are the same and which increases as one distribution gets 'further' from the other. In previous works, a descriptor function was combined with a divergence measure, such as KL-divergence, to compare two sets of proteins. Different descriptor functions have been utilised such as residue angles, secondary structure counts,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ramachandran angles (Wu et al., 2022a) and residue-residue distances (Anand & Achim, 2022) and divergences between these descriptors on the true and sampled distribution are given. However, these approaches are not expressive enough to be sensitive to fine-grained topological differences (see Section 3.1) and therefore cannot pick up on important structural dissimilarities between distributions. We want a more expressive descriptor which can differentiate more distributions of proteins than using angles and distances, and which is stable with respect to perturbations of the input.

2.1.2. Novelty

Novelty measures the amount to which the generated samples differ significantly from the reference set. This is important to ensure that the model is not merely memorising the input. To measure the novelty of the sampled proteins, previous work measured the maximum TMscore (Zhang & Skolnick, 2005) between the generated and the training set of proteins or between the generated set and any chain in the Protein Data Bank (PDB) using FoldSeek (van Kempen et al., 2022). Although (Xu & Zhang, 2010), looked at the significance of the TMscore in terms of topological similarity at a specific value, we still lack understanding on what extent these scores imply that the sampled proteins are novel. Additionally, the current metric is dependent on the training set of proteins used and so it is difficult to compare and benchmark models.

2.1.3. Coverage and Diversity

It is important for generative models to cover all relevant modes of the target distribution and to avoid phenomena such as 'mode collapse'. This allows us to generate novel proteins across the full protein space. Previously, internal diversity of generated proteins has been measured using Max-Cluster (Herbert & Sternberg, 2008) to hierarchically cluster proteins with a 0.5 TM-score threshold (Yim et al., 2023). The diversity is then given by (number of clusters) / (number of samples). Although these methods show that the generated proteins are different from each other and no mode collapse is occurring, they do not give an indication of how much the sampled proteins cover the protein space.

2.1.4. Designability

Designability is defined as the total number of amino acid sequences that can fold to a target protein structure. This has previously been approximated with a self-consistency evaluation (Trippe et al., 2023): sequences for the protein are generated with a sequence design model (e.g. ProteinMPNN (Dauparas et al., 2022)), these sequences are then input to a structure prediction model (eg. AlphaFold (Jumper et al., 2021)) and the agreement between the predicted structure and the original protein is measured. In (Wu et al., 2022a), they show that 87% of natural structures have an scTM ≥ 0.5 which should give an upper bound on the metric. Although this measure may be useful in practice, it effectively measures the invertibility of the sequence and structure prediction models and has not been shown to be highly correlated with known protein designability. There are also parameters such as scTM threshold, sequence and structure design models, and number of generated sequences which benefit different models, making it difficult to compare approaches.

3. Method and Evaluation

As outlined, current metrics are not sensitive to topological differences, do not fully capture protein designability or coverage, and are often dependent on the training set used. Therefore, we propose and extend new metrics to form a benchmark suite for comparing protein generative models.

3.1. Distribution learning

Generative modelling aims to capture a given data distribution as accurately as possible. To measure success, we therefore seek a metric that is zero when the training distribution and the distribution of generated samples are the same and increases as the distributions get more dissimilar.

Definition Similarly to the image (Fréchet Inception distance) and small molecule community (Fréchet ChemNet distance) (Preuer et al., 2018), we propose to measure the distance between the final layer embeddings of a neural network between the generated and the training set. We choose ProteinMPNN (Dauparas et al., 2022) as our model, since it works directly on structures without sequence information, and we use the maximum mean discrepancy (MMD) (Gretton et al., 2012) to get the distances between distributions. The biased empirical estimate of MMD between two samples $\mathcal{X} = \{x_1, ..., x_n\}$ and $\mathcal{Y} = \{y_1, ..., y_m\}$ can be computed as

$$\text{MMD}_k^2(\mathcal{X}, \mathcal{Y}) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(y_i, y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j) \quad (1)$$

and depends on the choice of the kernel function k (O'Bray et al., 2021). In our experiments, we use a Gaussian kernel applied to the neural network embeddings $\phi(x) \in \mathbb{R}^d$.

Evaluation To assess whether our metric is expressive enough to distinguish topologies and behaves well as two distributions get topologically more equivalent. We measure the correlation between topological overlap and different measures including secondary structure counts, C_α

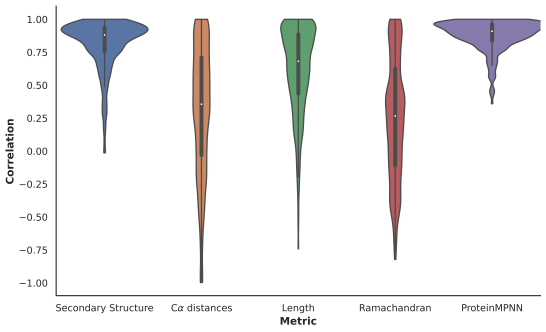

Figure 1: Correlation between topological overlap and different measures on CATH dataset.

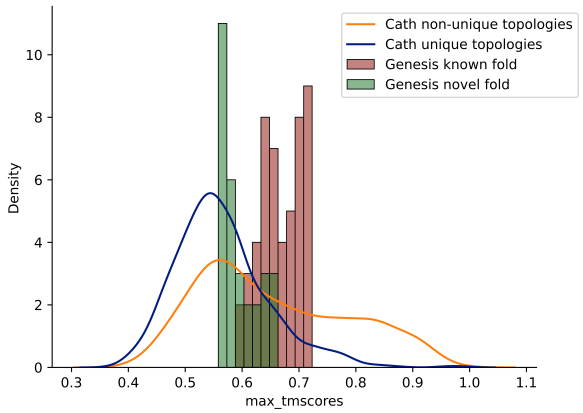

Figure 2: Distributions of TMscores to CATH reference set using protein structures generated from the Genesis model when the generation is conditioned on a novel fold and when the generation is conditioned on a known topology.

distances, protein length, ϕ, ψ dihedral angles and our proposed measure using ProteinMPNN embeddings over 100 experiments. To define topological overlap, we randomly select two sets of $n = 100$ proteins from a particular CATH class of which a predefined fraction $d_{\text{topo}} \in [0, 1]$ have the same label on the CATH architecture level.

In Figure 1 we see that our measure is the most correlated with topological overlap, whilst secondary structure also performs well. This is not surprising given that a lot of the architectures in CATH are defined through the presence of certain secondary structures.

3.2. Novelty

Instead of assessing novelty in the context of a specific training set, we propose using a standardised set based on the CATH database for comparisons of generated structures. This allows for easy comparison of models and ensures that the standardised set covers the full topological space of natural proteins.

Definition We first create a reference distribution of TM-scores for artificial proteins with a novel topology. To this end, we remove a single topology from the CATH dataset and compute the maximum TM-score between any protein from our query set with this topology to all proteins from other topology classes. We repeat this leave-one-out procedure to create a reference distribution of maximum TM-scores for *novel topologies*. In a similar vein, we calculate maximum TM-scores for all query proteins to the CATH set without leaving out a topology, thereby creating a reference distribution for *non-novel topologies*. To evaluate a set of generated proteins, we perform the same steps to obtain a third distribution of maximum TM-scores which can be compared to the reference distributions of novel and non-novel proteins. Our topological novelty metric is defined as

$$\textbf{Novelty}(\mathcal{X}) = 1 - \frac{d(\mathcal{X}, \mathcal{X}_{\text{novel}})}{d(\mathcal{X}, \mathcal{X}_{\text{novel}}) + d(\mathcal{X}, \mathcal{X}_{\text{non-novel}})}, \quad (2)$$

with d being the KL-divergence. It will be high whenever the distribution is close to the novel and far from the non-novel reference distribution.

Evaluation To showcase our approach, we evaluate two sets of proteins generated from the Genesis model described in (Harteveld et al., 2022). The first set is sampled whilst being conditioned on a native topology and the second set is generated with a fold that is not observed in nature. From 2, we find that we get a high novelty for protein structures with novel topologies (novelty score = 0.779) whilst being lower for structures with known natural topologies that are close to the training distribution (novelty score = 0.050).

3.3. Coverage

We propose a metric to measure the ability of the model to sample the full topological space of proteins. Unlike previous approaches that focus on internal diversity measures, here we explicitly measure the coverage of the generative model.

Definition To gauge the coverage of the (known) structural space, we define a diversity metric based on the 'spread' of similarity values to all CATH topologies. We represent each topology in CATH with a single protein and measure the TMscore from the sampled proteins to each topology. We can then represent the TMscores as an $N \times T$ matrix, $\boldsymbol{M}$, where N is the number of generated samples and $T = 1470$

Table 1: Coverage metric when sampling 50 proteins from the CATH database with different strategies.

Sampling	Coverage Metric
Topology	0.00088
Architecture	0.00095
Class	0.00231
Any	0.00276

is the number of CATH topologies. A high score at position (i, j) in the matrix corresponds to a topology j which protein i is close to. As we want sampled proteins to cover topological space, we wish columns of the matrix to have high variance - we want different proteins to be close to different CATH topologies. We can thus define our coverage measure as

$$\textbf{Coverage}(\mathcal{X}) = \frac{1}{T}\sum_{j=1}^{T} \text{Var}(m_{1,j}, ..., m_{N,j}), \quad (3)$$

Additionally, we can use similar principles to create an internal diversity measure. Here, instead of calculating TMscores to CATH, we can calculate pairwise TMscores of the generated proteins.

Evaluation In order to validate the approach, we first sample N proteins with a given topology, then N proteins with a given architecture, then N proteins in a certain class and finally N proteins in the full CATH database to see that our metric is increasing as the size of the CATH space that we are covering increases. We set $N = 50$ and see that Table 1 is in line with our expectations of the measure.

3.4. Designability

We want a metric that will be able to approximate the designability of a given structure. Additionally, we would like to not use structure prediction models such as Omegafold (Wu et al., 2022b) or AlphaFold (Jumper et al., 2021) given that these require a sequence which another model is required to generate.

Definition The total number of amino acid sequences that can fold into a target protein structure is known as designability. We focus on the diversity of protein sequences, where a large diversity means that dissimilar sequences and a large coverage of the sequence space can fold into a given structure. To measure this we use the ProteinMPNN model (Dauparas et al., 2022) which predicts sequences that fold into a target structure. We generate n sequences and calculate the mean edit distance of the generated proteins. This gives us an approximation for the diversity of sequences that can fold into the structure and thus the protein's designability.

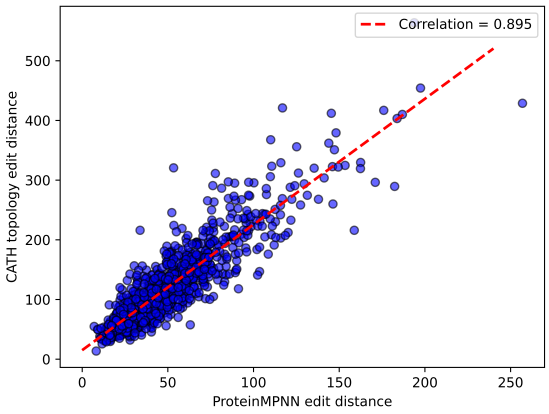

Figure 3: Mean edit distance of sequences generated from CATH protein structures using ProteinMPNN vs the mean edit distance of the CATH sequences in the protein topology which the protein belongs to.

Evaluation It is difficult to have a ground truth designability score for a single protein. However, we propose to approximate the designability of a protein with the designability of its topology. This allows us to utilize datasets in which many sequences map to the same topology. Using the CATH database, we score the designability of a topology using the mean edit distance of sequences in that topology. A high mean edit distance implies that there are a lot of dissimilar sequences that fold into a similar structure. We want a metric, which given a single protein structure, can estimate the designability of the topology in which it is in. We generate 10 sequences using ProteinMPNN and calculate the mean edit distance of the generated proteins. From Figure 3, we see that this metric is well correlated (correlation = 0.895) with the mean edit distance of CATH sequences with that topology and thus can be used to approximate to the designability of the protein topology.

4. Conclusion

We have described metrics for evaluating protein generation in terms of the novelty, designability and coverage of the sampled structures. Additionally, we have outlined a distributional metric for evaluating the differences between two distributions of protein structures which, unlike previous approaches, is sensitive to topological differences. We believe these metrics are important for the community in multiple ways, from benchmarking current and new approaches to speeding up drug discovery by aligning the evaluation with important metrics in the discovery pipeline.

References

Anand, N. and Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models, 2022.

Anand, N. and Huang, P. Generative modeling for protein structures. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Castorina, L. V., Petrenas, R., Subr, K., and Wood, C. W. PDBench: evaluating computational methods for protein-sequence design. *Bioinformatics*, 39(1), 01 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad027. btad027.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking, 2023.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Harteveld, Z., Southern, J., Defferrard, M., Loukas, A., Vandergheynst, P., Bronstein, M., and Correia, B. Deep sharpening of topological features for de novo protein design. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

Herbert, A. and Sternberg, M. Maxcluster: a tool for protein structure comparison and clustering, 2008.

Ingraham, J., Baranov, M., Costello, Z., Frappier, V., Ismail, A., Tie, S., Wang, W., Xue, V., Obermeyer, F., Beam, A., and Grigoryan, G. Illuminating protein space with a programmable generative model. *bioRxiv*, 2022. doi: 10.1101/2022.12.01.518682.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Topf, M. Critical assessment of techniques for protein structure prediction, fourteenth round. *CASP 14 Abstract Book*, 2020.

O'Bray, L., Horn, M., Rieck, B., and Borgwardt, K. Evaluation metrics for graph generative models: Problems, pitfalls, and practical solutions. *arXiv preprint arXiv:2106.01098*, 2021.

Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., and Klambauer, G. Fréchet chemnet distance: A metric for generative models for molecules in drug discovery, 2018.

Southern, J., Wayland, J., Bronstein, M., and Rieck, B. Curvature filtrations for graph generative model evaluation, 2023.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. S. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*, 2023.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Söding, J., and Steinegger, M. Foldseek: fast and accurate protein structure search. *bioRxiv*, 2022. doi: 10.1101/2022.02.07.479398.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., Bortoli, V. D., Mathieu, E., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, 2022. doi: 10.1101/2022.12.09.519842.

Wu, K. E., Yang, K. K., van den Berg, R., Zou, J. Y., Lu, A. X., and Amini, A. P. Protein structure generation via folding diffusion, 2022a.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., and Peng, J. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022b. doi: 10.1101/2022.07.21.500999.

Xu, J. and Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26(7): 889–895, April 2010.

Yim, J., Trippe, B. L., Bortoli, V. D., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. Se(3) diffusion model with application to protein backbone generation, 2023.

Zhang, Y. and Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, 33(7):2302–2309, April 2005.