Joint Protein Sequence-Structure Co-Design via Equivariant Diffusion

Ria Vinod
Brown University
ria_vinod@brown.edu

Kevin K. Yang
Microsoft Research
yang.kevin@microsoft.com

Lorin Crawford
Microsoft Research, Brown University
lorin_crawford@brown.edu

Abstract

Protein macromolecules are known to play key roles in cellular processes. Solving inverse design problems can allow us to control targeted cellular processes by designing proteins optimized for downstream tasks. However, current fixed-backbone protein design methods are limited to generating one type of secondary structure for a set of design candidates, that are learned from distributions of a single modality (either sequence or structure). To this end, we propose a diffusion-based generative modelling method that co-designs sequence and structure properties for an arbitrary distribution of proteins structures by optimizing over a function of a downstream protein task. We demonstrate preliminary results of an equivariant joint diffusion process for 2 modalities, with the goal of scaling to more modalities.

1 Introduction and Related Work

Protein structure and sequence generation has been approached from several angles: through multiple sequence alignment [1], standard generative models [2], adaptive sampling methods [3], message passing neural networks [4], and more recently, diffusion-based generative methods [5], [6]. However, these works generate candidates from one type of protein representation, relying on either sequence or 3D structure information to capture the variance that influences downstream task performance. Generated structures or sequence identities are then inferred by using auxiliary oracles to predict sequence recovery [4] or fold patterns [7]. As a result, nontrivial errors in approximations are propagated in the inference process and generated candidates perform poorly when evaluated on a set downstream tasks. Consequently, many deep-learning based inverse-design problems, where models are trained to implicitly 'learn the protein language' and apply a learned set of rules to a latent distribution to sample novel protein structures from. However, success with this method is limited by the quality and quantity of the reference data used to train these models, and so generated candidates perform poorly in downstream tasks as they result in unstable or invalid structures – which only become known during insilico testing. New representation learning methods that learn a rules-based design protocol conditioned on physical and biomechanical priors are critical to be able to generate design candidates for specified target responses.

To combat this, we propose a new diffusion-based generative modeling approach that seeks to learn a representation of biological and physical prior information about a set of protein measurement distributions that measure the same protein set. The complete view of a biological system can be thought of to be recursively defined by protein conformations, molecular motions, and atomic configurations. Proteins can be measured and represented through sequence identities, 3D structures, homologous or evolutionary information, microscopy, energetics, molecular dynamics simulation

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

data, enzyme activity (canonical reaction information), and text descriptions from data banks (in addition to other measurements). From each of these measurement modes, we propose learning over a mixture of distributions to implicitly learn which measurement mode captures the most variance at each reconstructive time step in the reverse diffusion process. Without knowing the relative contributions of different measurements (or modes) to protein fitness (as defined by the downstream task) a priori, this method can scale to work with multiple distributions to define a more granular set of design rules in approaching the protein-design problem.

2 Method

2.1 Diffusion Models

We build on the approach of Hoogeboom et al. [8] which uses a E(3) equivariant diffusion model to learn to diffuse on categorical and continuous data types. We simultaneously diffuse on a concatenated representation of 3 different distributions: 3D atom information, associated attributes [5] and a latent representation of sequence identities associated with a residue. The set of samples is thus modified to $(x_i, h_i, s_i)_{i=1,\ldots,M}$, where $x_i \in \mathbf{R^3}$, $h_i \in \mathbf{R^{nf}}$, $s_i \in \mathbf{R^{128}}$ (number of features from the latent embedding representation. We use a concatenated representation of $[\mathbf{x}, \mathbf{h}, \mathbf{s}]$ during the diffusion process (see Appendix A for details).

The h_i features are sinusoidal positional edge embeddings, developed in [5], which facilitate learning interactions between the linear structure of sequential residues. They are therefore invariant to E(n) transformations as proved in [9].

2.2 Data

$\mathbf{X_{structure}}$: 3D coordinates for C-α backbone atoms of a set of monomeric protein backbones. The training set was constructed by filtering through the Protein Data Bank for monomeric proteins with residues lengths between 48 and 128, and a resolution under 5 . The resulting set consisted of 5156 samples with 14 different corresponding CATH secondary structures (see appendix).

$\mathbf{X_{sequence}}$: Sequence identities for the same 5156 PDB entries. Using variable length sequences (with padding), we obtain a latent space representation from the learned embeddings in [10] for the "discrete" diffusion case. The forward diffusion process is then applied to the latent variable distribution. see appendix A for details.

3 Preliminary Results

We evaluate generated structures using Alpha Fold secondary structure predictions for generated sequences.

We perform 3 individual diffusion model training processes:

- Structure-only diffusion: We train a diffusion model on only 3D $C - \alpha$ backbone atom coordinates and the associated node attributes, as demonstrated in [cite satoras].
- Sequence-only diffusion: We train a diffusion model on the latent representation of discrete protein sequences. We extract the learned embeddings from the CARP model [10] for the given protein sequence data and perform the full forward diffusion process on the latent representation. We then use the CARP decoder to reconstruct generated sequences from the latent space to feature space.
- Sequence and structure joint diffusion: We perform the full joint diffusion on a concatenated representation of 3D atom information, node attributes and a latent representation of sequence identities, $[\mathbf{x}, \mathbf{h}, \mathbf{s}]$.

To prove the validity of the assumption that the diffusion model that learns correlations between sequence and structure feature distributions generate more realistic structures, we compare generated sequences from the sequence and structure diffusion model to the generated sequences from the sequence diffusion model. We evaluate the validity of the sequences by obtaining the Alpha Fold [7] folding pattern and comparing it to the original label that the sequence is most similar to from

the training set. We show that for a random set of 10 structures that we trained on, the joint sequence-structure diffusion model has a higher accuracy for structure validation.

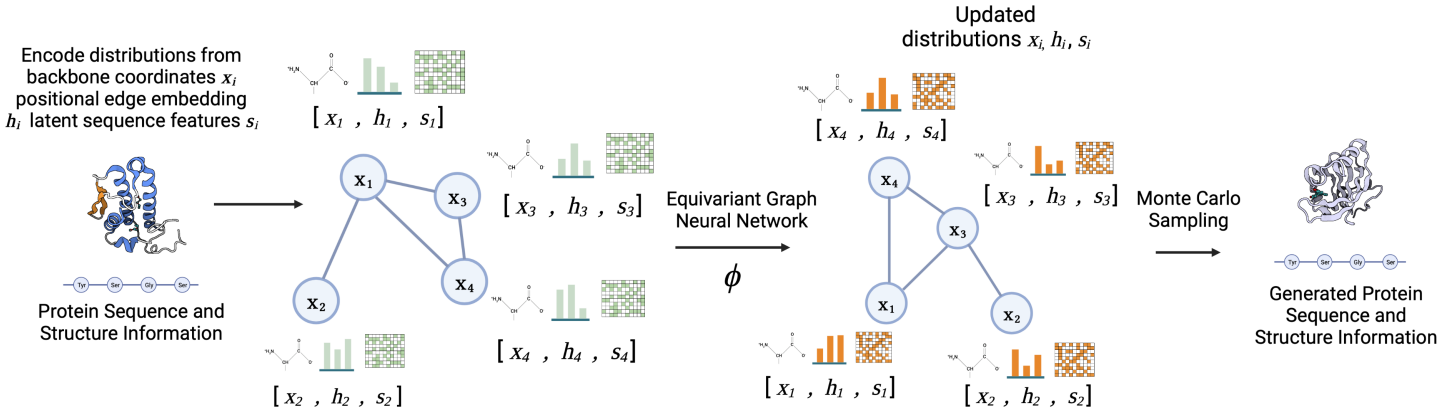

Figure 1: Schematic of graph convolutional network with data representation of 3D atom data, node attribute data, sequence data, modified from [9].

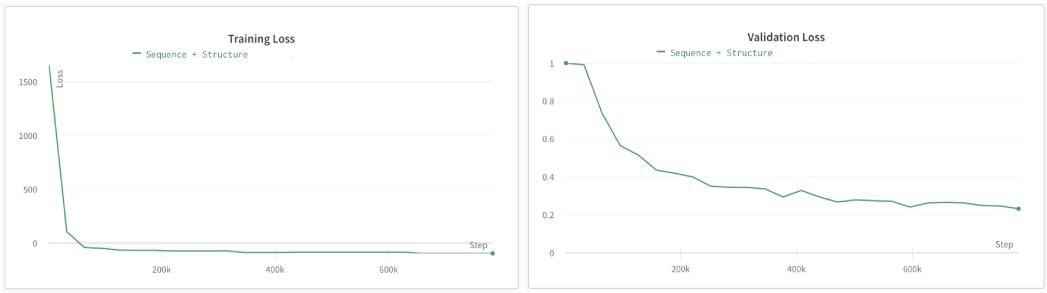

Figure 2: Training and validation loss curves for joint sequence-structure diffusion.

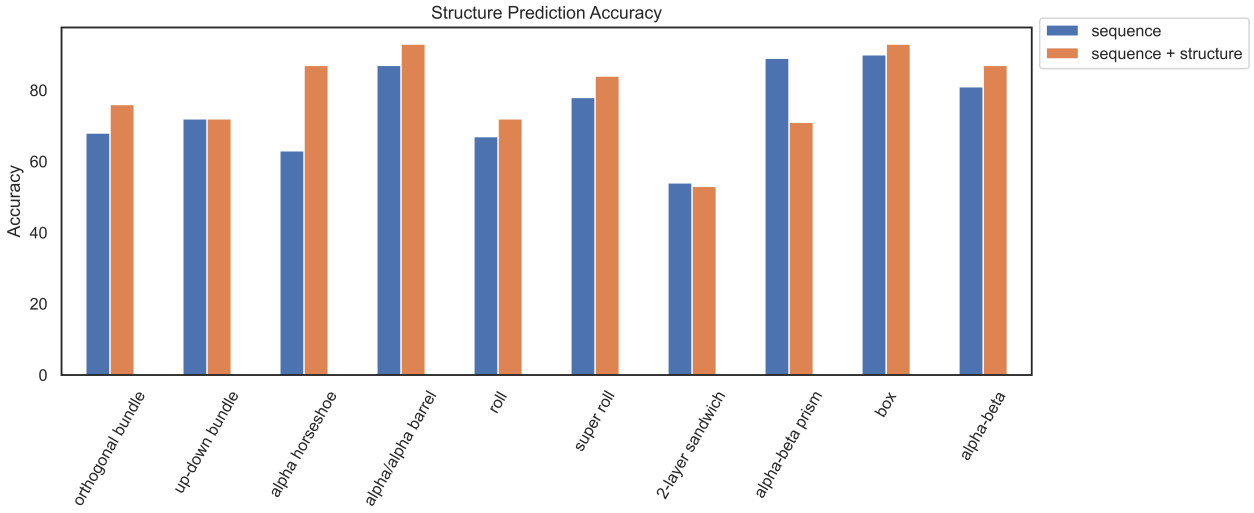

Figure 3: Structure prediction validation from generated sequences using Alpha Fold [7].

4 Future Directions

Our future directions are multi-fold:

- Validate the generated structures through protein backbone geometrical constraints.
- Condition the model on energetics and other empirical measurement data to be able to optimize for downstream protein tasks (such as binding and localization).
- Design a weighting scheme to learn which modes of protein data measurements capture the most variance in feature space and latent space to guide the reconstruct process for targeted protein binding tasks.

5 Conclusion

This very preliminary work of joint diffusion to learn how measurement data of the same biological system can be understood as hierarchical distributions through a diffusion modeling process. Furthermore, we anticipate that we will be able to exploit this hierarchy for more granular and targeted design tasks. We are particularly excited about this initial work as it shows potential for success in design tasks across multiomic scales.

A Appendix

A.1 Protein Notation

A: the set of 20 naturally occurring amino acids
N-residue protein: $s \in A^N$
C-α backbone coordinates (3D): $\mathbf{x} = [[\mathbf{x}_{\alpha_1}, \mathbf{x}_{\beta_2}], [\mathbf{x}_{\alpha_2}, \mathbf{x}_{\beta_2}], ...[\mathbf{x}_{\alpha_N}, \mathbf{x}_{\beta_N}]]^{\mathbf{T}} \in \mathbb{R}^{\mathbf{N},\mathbf{3}}$
Protein structure label: $\mathbf{y}$

Structures:

- mainly alpha: orthogonal bundle, up-down bundle, alpha horseshoe, alpha solenoid, alpha/alpha barrel
- mainly beta: ribbon, single sheet, roll, beta barrel, clam, sandwich, trefoil, prism
- alpha beta: roll, super roll, alpha-beta barrel, 2-layer sandwich, alpha-beta prism, box, alpha-beta complex

A.2 Forward Diffusion

- Diffusion on the joint sequence/structure distribution simultaneously, both approach the isotropic Gaussian distribution
- 2 simultaneous diffusion processes for structure and sequence distributions with respective noise steps $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$
  - $q_{struct}(\mathbf{x}_{\mathbf{struct},1:T}|\mathbf{x}_{\mathbf{struct},0}) = \prod_{t=1}^{T} q(\mathbf{x}_{\mathbf{struct},t}|\mathbf{x}_{\mathbf{struct},t-1})$
  - $q_{seq}(\mathbf{x}_{\mathbf{seq},1:T}|\mathbf{x}_{\mathbf{seq},0}) = \prod_{t=1}^{T} q(\mathbf{x}_{\mathbf{seq},t}|\mathbf{x}_{\mathbf{seq},t-1})$
- Forward process starts from sample $\mathbf{x}^{(0)}$ from data distribution q with density $q(\mathbf{x}^{(0)})$
- Cosine variance schedule, $\beta^{(1)}, \beta^{(2)}, ...\beta^{(N)}$:
  $\beta_t = \text{clip}(1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}, 0.999) \quad \bar{\alpha}_t = \frac{f(t)}{f(0)}$ where $f(t) = \cos\left(\frac{t/T+s}{1+s} \cdot \frac{\pi}{2}\right)$ (can be tuned in the process).
- Iteratively add noise at each timestep t, when $t = T$, $\mathbf{x}^{(T)} \sim \mathcal{N}(\mathbf{x}^{(T)}; 0, I)$

A.3 Reverse Process

- We choose ϵ_θ to be the $\mu_{\theta t}, \beta_t$ predictor; a standard roto-translational invariant neural network, such that
  - $\mu_\theta(\mathbf{x}^{(t)}, t) = \frac{1}{\sqrt{\alpha^{(t)}}}(\mathbf{x}^{(t)} - \frac{\beta^{(t)}}{\sqrt{1-\overline{\alpha}^{(t)}}}\epsilon_\theta(\mathbf{x}^{(t)}, t))$
  - $\alpha^{(t)} := 1 - \beta^{(t)}$
  - $\overline{\alpha}^{(t)} := \prod_{s=1}^{t} \alpha^{(t)}$
- During training:
  - Marginally sample $\mathbf{x}^{(t)} \sim q(\mathbf{x}^{(t)} \mid \mathbf{x}^{(0)})$ from the forward process as $\mathbf{x}^{(t)} = \sqrt{\overline{\alpha}^{(t)}\mathbf{x}^{(0)}} + \sqrt{1-\alpha^{(t)}\epsilon}$
  - Minimize the objective: $T^{-1}\sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}^{(0)}|\mathbf{x}^{(t)})}[\|\epsilon - \epsilon_\theta(\mathbf{x}^{(t)}, t)\|^2]$ by some stochastic optimization algorithm where $Loss_{total} = Loss_{structure} + Loss_{sequence}$. Weighting the sequence/structure losses relative to each other will be explored in training.
- Generate samples: By simulating the reverse process, sample noise at $t = T : \mathbf{x}^{(T)} \sim \mathcal{N}(0, I)$
- From $p_\theta(\mathbf{x}^{(0)})$ for $t = T-1, ...0$, simulate $\mathbf{x}^{(t)} \sim p_\theta(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t+1)})$

References

[1] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[2] Anvita Gupta and James Zou. Feedback gan (fbgan) for dna: A novel feedback-loop architecture for optimizing protein functions. *arXiv preprint arXiv:1804.01694*, 2018.

[3] David H Brookes and Jennifer Listgarten. Design by adaptive sampling. *arXiv preprint arXiv:1810.03714*, 2018.

[4] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, page eadd2187, 2022.

[5] Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

[6] Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

[7] 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.

[8] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pages 8867–8887. PMLR, 2022.

[9] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

[10] Kevin K Yang, Alex X Lu, and Nicolo K Fusi. Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv*, 2022.

[11] Athanasios Baltzis, Leila Mansouri, Suzanne Jin, Björn E Langer, Ionas Erb, and Cedric Notredame. Highly significant improvement of protein sequence alignments with alphafold2. *Bioinformatics*, 2022.

[12] Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models. *bioRxiv*, 2022.

[13] Boqiao Lai, Jinbo Xu, et al. End-to-end deep structure generative model for protein design. *bioRxiv*, 2022.

[14] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[15] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):1–10, 2022.

[16] Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of Medicinal Chemistry*, 2022.

[17] Hirofumi Kobayashi, Keith C Cheveralls, Manuel D Leonetti, and Loic A Royer. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nature methods*, 19(8):995–1003, 2022.

[18] Amir Motmaen, Justas Dauparas, Minkyung Baek, Mohamad Abedi, David Baker, and Philip Harlan Bradley. Peptide binding specificity prediction using fine-tuned protein structure prediction networks. *bioRxiv*, 2022.

[19] Xiaomin Fang, Fan Wang, Lihang Liu, Jingzhou He, Dayong Lin, Yingfei Xiang, Xiaonan Zhang, Hua Wu, Hui Li, and Le Song. Helixfold-single: Msa-free protein structure prediction by using protein language model as an alternative. *arXiv preprint arXiv:2207.13921*, 2022.