From Static Structures to Ensembles: Studying and Harnessing Protein Structure Tokenization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Protein structure tokenization converts 3D structures into discrete or vectorized representations, enabling the integration of structural and sequence data. Despite many recent works on structure tokenization, the properties of the underlying discrete representations are not well understood. In this work, we first demonstrate that the successful utilization of structural tokens in a language model for structure prediction depends on using rich, pre-trained sequence embeddings to bridge the semantic gap between the sequence and structural "language". The analysis of the structural vocabulary itself then reveals significant semantic redundancy, where multiple distinct tokens correspond to nearly identical local geometries, acting as "structural synonyms". This redundancy, rather than being a flaw, can be exploited with a simple "synonym swap" strategy to generate diverse conformational ensembles by perturbing a predicted structure with its structural synonyms. This computationally lightweight method accurately recapitulates protein flexibility, performing competitively with state-of-the-art models. Our study provides fundamental insights into the nature of discrete protein structure representations and introduces a powerful, near-instantaneous method for modeling protein dynamics. Source code is available here.

1 Introduction

The convergence of deep learning and vast protein databases has given rise to powerful protein models that can decipher the intricate rules governing protein sequence, structure, and function [29, 25, 30]. Trained on billions of protein sequences, protein language models (PLMs) such as ESM demonstrate remarkable transfer learning capabilities across downstream tasks [16]. The rapid development of protein structure prediction models, such as AlphaFold, solves the long-standing challenge of predicting static 3D protein structures with remarkable accuracy [12].

While these breakthroughs are powerful, they largely treat sequence and structure as separate domains. In many applications, especially in protein design tasks like binder design [28] and functional site scaffolding [26], it requires joint understanding and generation of both modalities. This highlights the need for multi-modal models that jointly process protein one-dimensional sequences and three-dimensional structures [27]. A fundamental obstacle in developing such models is how to combine complex, continuous structural data with discrete amino acid tokens in a unified representation suitable for deep learning. To overcome this issue, recent approaches have converged on the concept of **protein structure tokenization**, discretizing the continuous 3D space into a finite vocabulary using techniques like the Vector Quantized Variational Autoencoder (VQ-VAE) [8, 23, 6]. This approach enables modeling the sequence of amino acids and protein structure in a unified language model [27].

Despite the promise of this paradigm, several fundamental questions remain unanswered. First, what is the most effective way to integrate the distinct modalities of protein sequence and discrete structure

Submitted to the AI for Science workshop (NeurIPS 2025). Do not distribute.

within a single generative framework? While a simple multilayer perceptron (MLP) adaptor is an intuitive starting point, it may not adequately bridge the gap between these different informational streams. Second, the intrinsic properties of the learned structural vocabularies are largely unexplored. Are these tokens distinct and orthogonal, or have the models learned a robust and potentially redundant set of representations? Understanding the "grammar" and "synonymy" of this structural language is crucial for interpreting and improving these models.

In this work, we investigate these questions by analyzing the properties of the VQ-VAE structural tokens and their application in structure prediction with a GPT-based generative model. We first demonstrate that the method of integrating sequence and structure information is critical, with pre-trained ESM3 sequence embeddings outperforming original ProGen2 sequence embeddings for accurate structure prediction. We then provide direct evidence of semantic redundancy within the structural codebook, showing that distinct tokens often decode to nearly identical structures. The semantic redundancy of the codebook, which is a "flaw" for next-token prediction, actually can be employed to explore the flexibility of protein structures. This naturally leads us to study a compelling question: *can the discrete representations learned by the VQ-VAE be leveraged for tasks beyond static prediction, offering a new avenue to model protein dynamics?* By creating a "synonym dictionary" based on this redundancy, we introduce a novel "synonym swap" strategy. Our results show that this method can generate conformational ensembles whose statistical properties, measured by Root Mean Square Fluctuation (RMSF), are highly correlated with those from traditional MD simulations. This study, therefore, not only sheds light on the nature of discrete structural representations but also establishes a computationally efficient method for generating realistic protein conformational ensembles, opening new possibilities for the study of protein dynamics.

2 Preliminaries

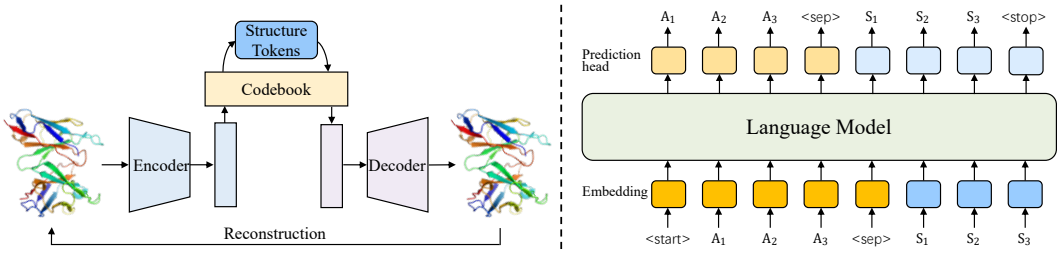

Figure 1: The VQ-VAE (left) discretizes continuous protein structures into a finite set of "Structure Tokens". These tokens are then used in an autoregressive language model (right) that predicts a sequence of structural tokens conditioned on the amino acid sequence.

2.1 Discrete representation of protein structures

The discretization of continuous 3D protein structures into a finite set of tokens has emerged as a powerful strategy for applying natural language processing techniques to protein science. Traditional methods usually rely on the domain knowledge about protein structures and discretize protein structures with hard-coded rules [2, 5]. Recently, a particularly effective approach for this task has been the Vector Quantized Variational Autoencoder (VQ-VAE), which learns a "vocabulary" of discrete tokens representing local structural motifs [22]. Recent works use the machine-learned vocabulary to show that complex protein folds could be successfully represented as sequences of these learned tokens [32, 8].

As shown in, the process of VQ-VAE for protein structure consists of three key components.

- Encoder: A neural network that takes as input the atomic coordinates of the protein and compresses this geometric information into a continuous latent vector.
- Codebook: The continuous vector from the encoder is mapped to the nearest vector in a learned codebook via a nearest-neighbor lookup. The index of this codebook vector becomes the discrete structural token.
- Decoder: A second neural network that takes a sequence of discrete tokens, retrieves their corresponding vectors from the codebook, and uses this sequence of embeddings to reconstruct the protein structure.

These structural vocabularies can serve as a fundamental component of a unified generative model, including multimodal diffusion language models, masked language models, and autoregressive GPT models [27, 8, 6], enabling sophisticated protein design and analysis. There are also efforts focused on systematically benchmarking different tokenization schemes and developing improved recipes to guide future research [31].

2.2 Autoregressive sequence-structure language modeling

With a protein structure tokenizer, we can represent both the protein sequence and its corresponding structure as discrete tokens. A unified generative model can then be developed. This paradigm enables the joint modeling of these two disparate modalities, facilitating tasks such as protein folding and de novo protein design.

In the context of protein structure prediction, the goal is to model the conditional probability of the structural token sequence (S_{struct}) given the amino acid sequence (S_{seq}), denoted as $P(S_{struct}|S_{seq})$. An autoregressive model, such as GPT [21], can be used to model $P(S_{struct}|S_{seq})$ by factorizing the probability sequentially:

$$P(S_{struct}|S_{seq}) = \prod_{t=1}^{L} P(s_t|S_{seq}, s_{<t}), \tag{1}$$

where s_t is the structural token at position t, and $s_{<t}$ represents all the preceding structural tokens. At each step of the generation process, the model takes the full amino acid sequence and the sequence of structural tokens predicted so far as input. It then outputs a probability distribution over the entire structural token vocabulary for the next position, t. A token is sampled from this distribution, appended to the sequence of predictions, and the process is repeated until the full structure is generated. This generative framework enables the direct prediction of a protein's 3D structure from its primary sequence, forming the basis of the predictive model used in this study.

3 The gap between the structure and sequence semantics

3.1 A study of structure tokenization with a GPT-like model for protein structure prediction

Model architecture. Given the discrete tokenization of the protein structures, it is a natural choice for jointly modeling the protein structure and sequence with a causal language model. Following Liu et al. [17], the network architecture is shown in Figure 1. We choose ProGen2-medium (764M) [20] as the protein language model.

For the input structure tokens, we use a simple linear layer to align the structure token embeddings with the PLM embedding space. Specifically, the input structure tokens are first passed through an embedding layer to get a continuous presentation Z. A trainable projection matrix $\mathbf{W}_{struct}$ is then applied to convert Z to the same dimensionality as the PLM embedding space. As the model needs to predict the structure tokens, a simple linear layer with weight $\mathbf{W}_{head}$ is added to the PLM as a new prediction head.

We use two settings for the sequence token embedding. The first is to use the original `nn.embedding` layer of ProGen2. Considering the gap between the sequence and structure modalities, the second setting is using the pre-trained ESM3 sequence embedding followed by a simple linear aligner with a weight matrix $\mathbf{W}_{seq}$.

Two-stage training. To align the structure token embeddings with the PLM embedding, we first keep the PLM weights frozen and train the projection matrix $\mathbf{W}_{struct}$ (and $\mathbf{W}_{seq}$ and in the second setting) and the structure head matrix $\mathbf{W}_{head}$ by maximizing the likelihood of Eq. 1. After embedding alignment, we perform a full fine-tuning to update both the weights of ProGen2 and the linear layers trained in the first stage.

Training data. For the first-stage training, we utilize the AFDB SwissProt data [25]. For the second stage fine-tuning, we use both AFDB structures and the single-chain structures from PDB. Since the ProGen2 model has a maximal sequence length of 1024, we crop the sequences to a maximal length of 512 to model the sequence and structure tokens together.

Table 1: The performance of structure prediction. StructGPT: ProGen2 model with the original ProGen2 sequence embedding; StructGPT w. ESM3_emb: ProGen2 model with the ESM3 sequence embedding. The results of ESM3 are from Zhang et al. [32]

	TMScore			RMSD (Å)		
Model	CAMEO	CASP14	CASP15	CAMEO	CASP14	CASP15
StructGPT	0.523	0.329	0.383	11.87	17.19	18.99
StructGPT w. ESM3_emb	0.784	0.580	0.639	5.43	10.24	11.74
ESM3	0.781	0.575	0.625	5.74	10.29	14.69

Structure prediction performance. After training, we generate the structure tokens conditioned on the sequence tokens using top-p sampling [9] with $p = 0.9$ and temperature $T = 0.7$. The generated structure tokens are then decoded to obtain the protein structure prediction. The performance is tested on three datasets, including CAMEO, CASP14, and CASP15 [7, 14, 15]. Our experiments reveal a stark difference in performance between the two sequence embeddings (Table 1). The model trained using the original ProGen2 sequence embedding fails to produce globally coherent structures. The model trained using the ESM3 embeddings demonstrates strong predictive performance, comparable to ESM3 [8], as measured by the TM-score [33] and the root-mean-square deviation (RMSD) [13].

By looking into the training curve (Figure 2), we find that with the original ProGen2 sequence embedding, while the cross-entropy of the CAMEO test set (i.e., the testing loss) continues to decrease, the structural accuracy of the generated proteins, as measured by TM-score, although still increasing very slowing, stays at a poor performance level with near 90,000 training steps. In contrast, the training curve with the ESM3 sequence embedding shows much faster convergence and a more stable performance plateau. Moreover, the two cases have similar cross-entropy losses. This observation leads us to hypothesize that the training objective itself might be complicated by the nature of the structural vocabulary. Specifically, the size of the ESM3 VQ-VAE codebook is relatively large (4096). If distinct tokens can represent structurally similar geometries, the model would be penalized during training for predicting a valid "structural synonym" that deviates from the one specific token in the reference structure. This suggests the existence of semantic redundancy within the VQ-VAE's learned vocabulary.

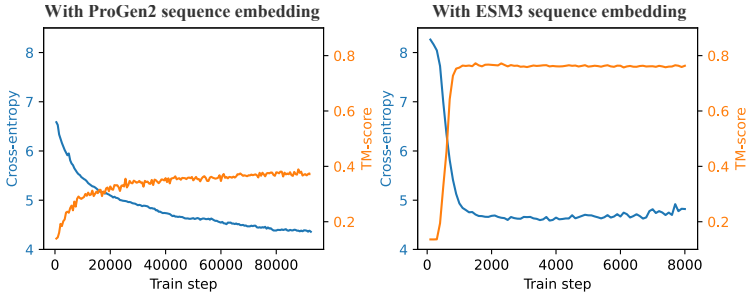

Figure 2: The training curve of the GPT model for protein structure prediction with different sequence embeddings.

3.2 Semantic redundancy in the ESM3 structural codebook

From the above result, we hypothesize that the ESM3 VQ-VAE codebook contains multiple tokens for similar geometric motifs. We thus turn to directly analyzing the properties of the structural vocabulary to test this hypothesis. To quantify the similarity between different structural tokens, we perform a direct analysis of the learned codebook. The latent vectors, corresponding to the structure tokens, are extracted from the ESM3 structure VQ-VAE model.

First, to visually inspect the relationships within the tokens, we use the t-SNE (t-Distributed Stochastic Neighbor Embedding) visualization [18], which projects the high-dimensional token embeddings into a 2D space. As shown in Figure 3, the t-SNE projection illustrates the existence of numerous dense clusters, indicating there are grouped tokens representing similar structural motifs in the embedding space.

Second, to establish a quantitative measure of similarity, we calculate the pairwise Euclidean distance between all token latent vectors. From the clustered heatmap, it is easy to see that the ESM3 structure

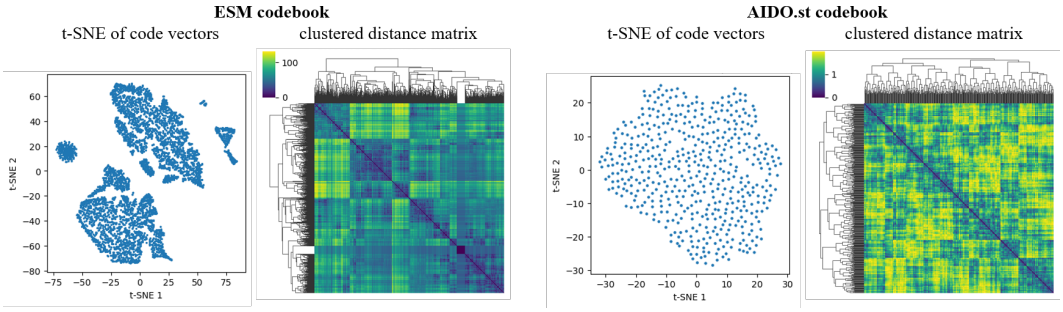

Figure 3: t-SNE visualization and the distance matrix of the code vectors. The ESM3 codebook shows distinct, well-defined clusters, while the t-SNE of the AIDO.st codebook [32] vectors are uniformly distributed in the 2D space.

tokens have strong, well-separated clusters. In contrast, the AIDO.st VQ-VAE codebook [32], which has a smaller number of codes ($m = 512$), shows a more diffuse and less defined clustering pattern.

Based on this, we construct a "synonym dictionary" by defining any two tokens as synonyms if the Euclidean distance between their latent vectors is below a threshold. Specifically, given a codebook $\mathcal{V} = \{\mathbf{v}_k\}_{k=1}^m$ of m different codes, the "synonym dictionary" of code $\mathbf{v}_k$ is defined by $\mathcal{S}_k = \{i\}_{\|\mathbf{v}_i - \mathbf{v}_k\|_2 < \tau}$, where τ is a threshold hyperparameter that can balance structural preservation and diversity. Based on visual inspection of the distance distribution, we currently set $\tau = 10$. The dictionary $\mathcal{S}_k$ serves as the basis for our subsequent perturbation studies. Given a structure, we first encode it into a sequence of structure tokens. Each token is then replaced by a random token in its "synonym dictionary" and the perturbed structural sequence is decoded into 3D structures. The resulting 3D structures are very similar to the original structure. The RMSD between the perturbed and original structures is consistently low, often less than 2.0 Å, confirming their structural equivalence. This indicates the semantic redundancy of the ESM3 codebook, which explains why a GPT model is statistically uncertain about which specific token to predict next, yet still can point towards the correct local geometry.

Table 2: The average TM-score and RMSD of the structures decoded from the perturbed structure tokens.

	TM-score	RMSD (Å)
CAMEO	0.933	1.744
CASP14	0.938	1.969
CASP15	0.849	4.136

4 Exploiting redundancy to generate dynamic conformational ensembles

Our finding in the previous section demonstrates that the structural vocabulary of ESM3 is not a minimal set of building blocks but a robust, flexible, and highly redundant language. The semantic redundancy of the codebook, which is a flaw for next-token prediction, is actually a feature that reflects the inherent flexibility of protein structures. Together with the structure decoder, the subtle structural variations in the "synonymous" tokens may reflect the natural, low-energy fluctuations a protein experiences in its native state. If this hypothesis is true, perturbing a ground truth structure by swapping its tokens with synonyms could provide a computationally inexpensive method to generate a realistic conformational ensemble, offering a rapid alternative to Molecular Dynamics (MD) simulations for studying protein flexibility [1].

4.1 Method

To test our hypothesis, we employ a simple "synonym swap" strategy. As shown in Figure 4, we first encode a given experimental structure of a target protein into a sequence of structure tokens. A new sequence of structure tokens is then generated by randomly replacing each token k from the original sequence with one of its synonyms, as defined in the synonym dictionary $\mathcal{S}_k$. The perturbed token sequence is then decoded back into a 3D protein structure.

To validate our approach, we use the 82 test proteins [10] from the ATLAS database [24]. Following the work in Jing et al. [10], we generate 250 perturbed structures for each target as the conformational ensembles.

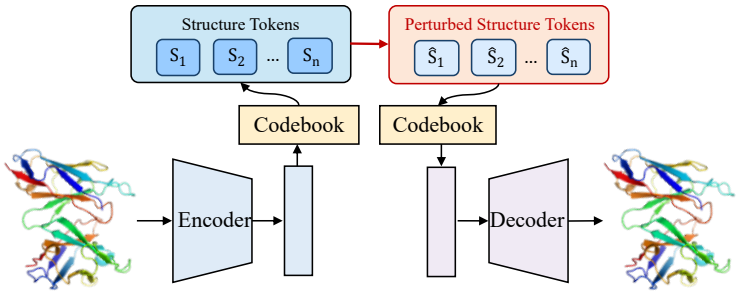

Figure 4: Perturbation of the structure tokens for exploring the conformational ensemble space.

4.2 Experiment

To rigorously evaluate the quality of our generated ensembles, we compare our "synonym swap" method to two state-of-the-art models, MDGen and AlphaFlow [10, 11], using the ground-truth trajectories from Molecular Dynamics (MD) simulations. To compare two ensembles, we use a suite of metrics designed to assess both protein flexibility and the conformational distribution. The results, summarized in Table 3, demonstrate that our computationally efficient method is highly competitive, particularly in capturing protein-specific flexibility.

A key measure of an ensemble's utility is its ability to reproduce the flexibility profile of individual proteins. The root mean square fluctuation (RMSF) quantifies the fluctuation of individual residues. The median Pearson correlation (r) between the generated and MD RMSF for each protein is 0.84. This result is highly competitive with the top-performing method, AlphaFlow (0.85), and outperforms MDGen (0.71), confirming that token perturbation effectively captures the unique dynamic fingerprint of individual proteins. Note that MDGen and AlphaFlow are both trained on the ATALS MD data, while our "synonym swap" method is totally **training-free**.

Table 3: The median of results on the 82 test targets in ATLAS; the results of MDGen and AlphaFlow from Jing et al. [10, 11]

	Ours	MDGen	AlphaFlow
Pairwise RMSD r ↑	0.38	0.48	0.48
Per target RMSF r ↑	0.84	0.71	0.85
Global RMSF r ↑	0.41	0.50	0.50
MD PCA $\mathcal{W}_2$ dist. ↓	1.83	1.89	1.52
Joint PCA $\mathcal{W}_2$ dist. ↓	2.54	-	2.25

To assess how well the ensembles capture the dominant modes of motion, we use the 2-Wasserstein distance ($\mathcal{W}_2$) between the generated and MD distributions after projecting them onto the principal components (PCs) of the positional distribution. A lower distance indicates a better match. Our method achieved an MD PCA $\mathcal{W}_2$ distance of 1.83, slightly better than MDGen (1.89) and competitive with AlphaFlow (1.52), indicating that the generated conformations occupy the same principal dynamic spaces as those explored by the MD simulation.

While our method can accurately capture per-protein flexibility and conformational distribution, other metrics that assess the global properties of the conformational space show the limitation of our current simple approach. The correlation of the Pairwise RMSD matrices and the Global RMSF correlation is lower for our method compared to the other two benchmarked methods. This suggests that while individual flexibility profiles are accurate, capturing the absolute scale of motion across a diverse dataset is more challenging (Figure 5). "Synonym swapping" creates local perturbations and may not be sufficient to capture the large-scale, cooperative motions that dictate global structural changes. This provides a direction for future work, such as exploring perturbations of token sequences rather than individual tokens to model more complex dynamics or combining with other techniques like MSA subsampling [3].

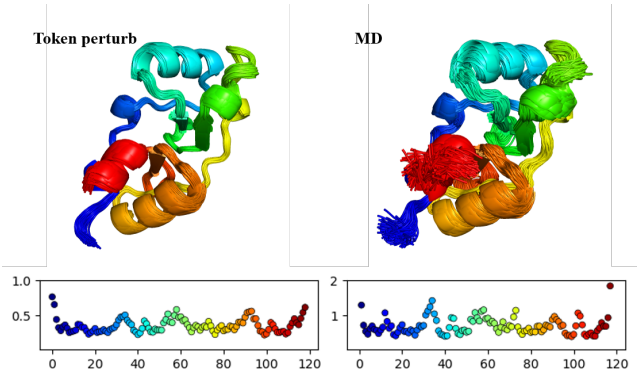

Figure 5: Protein ensembles for `6uof_A` generated by token perturbation and MD, and the Cα RMSFs indexed by the residue id (Pearson $r = 0.81$).

5 Conclusion

In this paper, we investigate the intrinsic properties of these discrete structural representations and explore how they can be leveraged for tasks beyond static prediction. We first establish that effective integration of modalities is critical, showing that a GPT model using pre-trained sequence embeddings significantly outperforms one using simple token concatenation. We then demonstrate that the structural codebook contains considerable semantic redundancy, where distinct tokens decode to nearly identical local structures. Finally, we harness this redundancy by developing a "synonym swap" strategy, showing it can generate conformational ensembles whose dynamic properties are highly correlated with those from computationally expensive Molecular Dynamics (MD) simulations [4, 19]. Our findings provide a deeper understanding of discrete structural representations and offer a novel, efficient method for modeling protein dynamics.

References

[1] 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.

[2] Alexandre G de Brevern. New assessment of a structural alphabet. *In silico biology*, 5(3): 283–289, 2005.

[3] Diego Del Alamo, Davide Sala, Hassane S Mchaourab, and Jens Meiler. Sampling alternative conformational states of transporters and receptors with alphafold2. *Elife*, 11:e75751, 2022.

[4] Ron O Dror, Robert M Dirks, JP Grossman, Huafeng Xu, and David E Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41 (1):429–452, 2012.

[5] Janani Durairaj, Mehmet Akdel, Dick de Ridder, and Aalt DJ van Dijk. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics*, 36 (Supplement_2):i718–i725, 2020.

[6] Zhangyang Gao, Cheng Tan, Jue Wang, Yufei Huang, Lirong Wu, and Stan Z Li. Foldtoken: Learning protein language via vector quantization and beyond. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. doi: 10.1609/aaai.v39i1.31998.

[7] Jürgen Haas, Alessandro Barbato, Dario Behringer, Gabriel Studer, Steven Roth, Martino Bertoni, Khaled Mostaguir, Rafal Gumienny, and Torsten Schwede. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins: Structure, Function, and Bioinformatics*, 86:387–398, 2018.

[8] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

[9] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[10] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *International Conference on Machine Learning*, pp. 22277–22303. PMLR, 2024.

[11] Bowen Jing, Hannes Stärk, Tommi Jaakkola, and Bonnie Berger. Generative modeling of molecular dynamics trajectories. *arXiv preprint arXiv:2409.17808*, 2024.

[12] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[13] Wolfgang Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 34(5):827–828, 1978.

[14] 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.

[15] 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.

[16] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[18] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[19] Mitchell D Miller and George N Phillips. Moving beyond static snapshots: Protein dynamics and the protein data bank. *Journal of Biological Chemistry*, 296, 2021.

[20] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[22] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[23] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

[24] 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.

[25] Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

[26] Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L Watson, Karla M Castro, Robert Ragotte, Amijai Saragovi, Lukas F Milles, Minkyung Baek, et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.

[27] Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Dplm-2: A multimodal diffusion protein language model. *arXiv preprint arXiv:2410.13782*, 2024.

[28] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[29] Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pp. 38749–38767. PMLR, 2023.

[30] Yang Xue, Zijing Liu, Xiaomin Fang, and Fan Wang. Multimodal pre-training model for sequence-based prediction of protein-protein interaction. In *Machine Learning in Computational Biology*, pp. 34–46. PMLR, 2022.

[31] Xinyu Yuan, Zichen Wang, Marcus Collins, and Huzefa Rangwala. Protein structure tokenization: Benchmarking and new recipe. *arXiv preprint arXiv:2503.00089*, 2025.

[32] Jiayou Zhang, Barthelemy Meynard-Piganeau, James Gong, Xingyi Cheng, Yingtao Luo, Hugo Ly, Le Song, and Eric Xing. Balancing locality and reconstruction in protein structure tokenizer. *bioRxiv*, pp. 2024–12, 2024.

[33] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.