LM2Protein: A Structure-to-Token Protein Large Language Model

Chang Zhou[1*] **Yuheng Shan**[2*] **Pengan Chen**[1] **Xiangyu Shi**[3]
Zikang Wang[4] **Yanting Li**[5] **Jiyue Jiang**[1]
[1] The Chinese University of Hong Kong [2] National University of Singapore
[3] Beijing Jiaotong University [4] The Hong Kong Polytechnic University
[5] Hong Kong University of Science and Technology (Guangzhou)
{changzhoupro10, sxysxygm}@gmail.com, shan.yuheng@u.nus.edu,
{chenpengan, jiangjy}@link.cuhk.edu.hk,
zikang.wang@connect.polyu.hk, yli106@connect.hkust-gz.edu.cn

Abstract

Proteins are critical for various molecular functions, relying on their precise tertiary structures. This structure-sequence relationship is complex and degenerate, meaning multiple sequences can fold into a similar structure. The challenges in protein prediction, design, and modification increase with sequence complexity, while research on RNA-protein interactions, especially RNA-binding proteins (RBPs), is gaining importance. Large-scale pre-trained language models (LLMs) have shown promising results in handling biological sequences by treating them as natural language, though integrating spatial structures remains complex due to the need for specialized visual and 3D modeling approaches. We introduce a method to integrate protein 3D structural data within a sequence processing framework, converting 3D coordinates into discrete structure tokens using a VQ-VAE-like network. This simplifies the handling of 3D data, avoiding complex pipelines and facilitating a unified sequence-to-sequence processing model. Our approach demonstrates strong performance across a range of tasks, achieving high sequence recovery in inverse folding and protein-conditioned RNA design. These outstanding results demonstrate significant potential for application in complex biological systems research.

1 Introduction

Proteins play crucial roles in molecular biology and biochemistry, from molecular catalysis to signal transduction, with diverse functions relying on their precisely folded three-dimensional structures (Ferruz et al., 2022; Madani et al., 2023; Yue and Dill, 1992). However, there is no strict one-to-one mapping between protein sequences and their 3D structures (Yue and Dill, 1992; Zhang et al., 2024), and increasing sequence length and complexity poses significant challenges for prediction, design, and modification. Meanwhile, growing research on nucleic acid-protein interactions, especially RNA-binding proteins, calls for rational RNA design based on RBP sequences or structures to enhance binding or catalytic functions.

LLMs demonstrate powerful capabilities across tasks (Achiam et al., 2023; Jiang et al., 2025a; Wang et al., 2025; Jiang et al., 2025b), leveraging self-supervised learning on large corpora to capture long-range dependencies and complex patterns. This approach extends to biological sequences such as proteins and nucleic acids. However, spatial structures often require additional computer vision or 3D pipelines, increasing complexity, resource usage, and potential mismatches in multimodal integration.

To efficiently incorporate 3D structural information into a unified sequence framework, we propose discretizing protein 3D coordinates into "structural tokens" via a VQ-VAE-like network. This mapping from continuous 3D coordinates to discrete indices removes the need for elaborate visual or 3D pipelines, instead treating protein structures as sequential tokens. Consequently, various tasks—sequence-to-structure, structure-to-sequence, and protein-nucleic acid co-design—can be unified under a sequence-to-sequence (Seq2Seq) paradigm, maintaining autoregressive modeling strengths and leveraging LLM capabilities in capturing long-range dependencies and intricate pattern recognition.

The main contributions of this paper can be summarized as follows: (1) We propose a unified structure representation method: converting protein 3D coordinates into structure tokens through the discretization of VQ-VAE, enabling the autoregressive sequence modeling paradigm of traditional LLMs to be directly applied to protein structure learning. (2) Within this unified framework, we have successfully implemented four distinct tasks: Protein sequence-to-structure prediction, achieving a TM-

[*] Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 7023–7029
November 4-9, 2025 ©2025 Association for Computational Linguistics

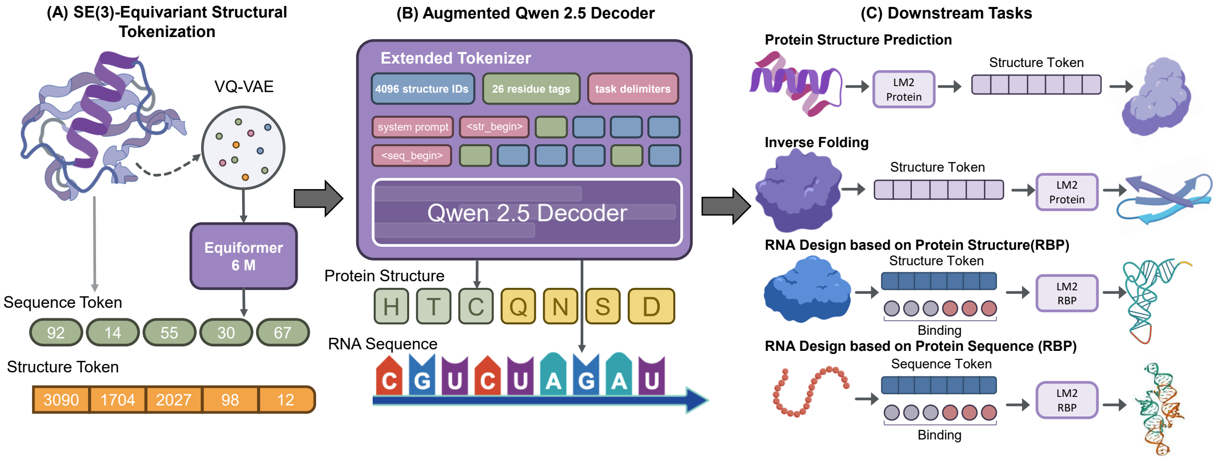

Figure 1: The **LM2Protein** Framework. The model consists of three main stages. **(A)** SE(3)-Equivariant Structural Tokenization: 3D atomic coordinates of a protein are processed by a lightweight Equiformer encoder and a VQ-VAE to be discretized into a sequence of structural tokens. **(B)** Augmented Decoder-Only Language Model: A Qwen 2.5 decoder, whose tokenizer is extended to include the 4096 structural tokens, processes both amino acid sequences and structural token sequences under a unified format. **(C)** Unified Downstream Generative Tasks: The single Seq2Seq model is capable of performing four distinct tasks: protein structure prediction, inverse folding, and RNA design conditioned on either an RBP's sequence or its tokenized structure.

score of 54.80%. Protein structure-to-sequence design (inverse folding), achieving a sequence recovery of 51.00%. RNA design based on protein structure, achieving a high sequence recovery rate. (3) By encoding structures discretely, we avoid the cumbersome and unstable multi-modal integration in 3D geometric or visual pipelines, significantly simplifying the training and inference processes. (4) The effectiveness and versatility demonstrated in the protein-nucleic acid domain offer a viable path for further expansion into larger-scale and more complex biological systems.

2 Method

2.1 Problem Definition

In this work, we aim to develop a unified framework capable of handling biomolecular sequence to sequence tasks, leveraging both sequences and structure representations of proteins. Here we formalize these four tasks:

Task1 - Protein Structure Prediction The objective is to predict the corresponding 3D structure given protein's amino acids $A = (a_1, a_2, \ldots, a_n)$. We form the structure in a sequence of discrete tokens $T = (t_1, t_2, \ldots, t_m)$. We adopt a VQ-VAE (Van Den Oord et al., 2017) based structural tokenizer to convert point cloud into that sequence of discrete tokens, allowing structural data to be processed similarly to sequence data. The details will be introduced in Section 2.2

Task2 - Protein Sequence Design The objective is to predict the corresponding amino-acid sequence of the input protein's 3D structure. It's the inverse task of protein folding, given tokenized structure $T = (t_1, t_2, \ldots, t_m)$, and predicting $A = (a_1, a_2, \ldots, a_n)$.

Task3 - RNA Design Based on RBP Sequence This task aims to generate an RNA sequence that satisfies the binding requirements of RBPs for a given protein sequence. Formally, given a protein sequence $A = (a_1, a_2, \ldots, a_n)$, the goal is to generate a corresponding RNA sequence $R = (r_1, r_2, \ldots, r_k)$. To provide fine-grained control over the binding sites, we additionally introduce a binding-type annotation $B = (b_1, b_2, \ldots, b_n)$, where $b_j \in \{0, 1\}$ indicates whether the j-th residue is involved in RNA binding.

Task4 - RNA Design Based on RBP Structure This task generates an RNA sequence that satisfies RBP binding requirements based on a given protein structure. Formally, given a sequence of structural token $T = (t_1, t_2, \ldots, t_m)$, the goal is to generate an RNA sequence $R = (r_1, r_2, \ldots, r_k)$.

2.2 Protein Tokenization

To enable sequence-based processing of protein structures, we adopt a VQ-VAE-based tokenization approach inspired by (Zhang et al., 2024) and implemented in the ESM-3 repository (Hayes et al., 2025). This approach transforms coordinates into a sequence of discrete tokens, making it compatible with autoregressive language models.

SE(3)-Equivariant Encoding. The 3D coordinates of protein atoms are first processed by a lightweight SE(3)-equivariant structure encoder $\mathcal{E}$, implemented as a 6M-parameter multi-layer Equiformer. This encoder computes residue-level continuous latent descriptors $\mathbf{z}_i = \mathcal{E}(\mathbf{r}_i) \in \mathbb{R}^d$, where $\mathbf{r}_i$ represents the local 30-atom point cloud surrounding the i-th residue. The equivariant transform $\mathcal{E}$ ensures that both local and global structural patterns are captured consistently across different global poses of the same structure.

Vector Quantization. The continuous descriptors $\mathbf{z}_i$ are then mapped to their nearest entries in a pre-trained 4096-vector codebook $\mathcal{C} = \{c_0, c_1, \ldots, c_{4095}\}$ using nearest-neighbor search: $y_i = \mathrm{VQ}(\mathbf{z}i) = \arg\min_k \|\mathbf{z}_i - c_k\|_2$. This quantization step converts the continuous latent representations into discrete structural tokens $T = (y_1, y_2, \ldots, y_m)$.

Sequence-to-Sequence Modeling. The tokenized tokens represent the structure of a protein, reformulating the four tasks in uniform sequence-to-sequence problems. The tokenized representations bridge the gap between geometric point cloud and sequence-based language models, supporting unified modeling across multiple biomolecular tasks.

2.3 Language Interface

To expose structural information to a text decoder, we augment the original Qwen 2.5 (Team, 2024) tokenizer with three groups of symbols: (i) task delimiters, (ii) 26 residue tags, (iii) integer tokens for the 4096 codebook indices. This minimal extension (4 k+ tokens) allows the model to jointly process amino-acid text and structure tokens without architectural changes.

2.4 Causal Decoder

We fine-tune a Qwen 2.5 (Team, 2024) with a fixed `SYSTEM_PROMPT` that instructs the model to output structure tokens. Only positions corresponding to `<str_begin>...<str_end>` contribute to the standard cross-entropy loss; earlier tokens are masked. Flash-Attention is enabled but no other architectural modifications are made.

2.5 Training Protocol

The model is trained for a handful of epochs on a split of protein structures ($\sim$200 k), holding out 10 % for validation. We use AdamW with a peak learning rate of $2{\times}10^{-5}$ and a short linear warm-up. Training runs on multiple GPUs under DeepSpeed ZeRO with tensor parallelism; checkpointing retains the most recent snapshots to ensure recoverability. All hyper-parameters follow common practice for billion-scale language models and are provided in the supplementary material.

3 Experiments

3.1 Experiment Setting

All experiments are conducted on a distributed computing cluster equipped with eight NVIDIA RTX 4090 GPUs. We implement our models using PyTorch 2.6.0 and Hugging Face Transformers. For distributed training, we utilize DeepSpeed with tensor parallelism and ZeRO-1 optimization strategy across multiple GPUs to efficiently train our large-scale model.

The fine-tuning of Qwen-2.5-1.5B is performed with a mini-batch size of 2 sequences per GPU, accumulating gradients for 8 iterations before performing parameter updates, resulting in an effective batch size of 16 sequences per GPU. We employ Flash-Attention to optimize memory usage and computational efficiency. To manage the varying lengths of protein sequences and structural tokens, we implement a dynamic batching strategy with a maximum sequence length of 4096 tokens.

3.2 Datasets

Regarding the protein data, we collect protein datasets from the AlphaFold DB (Varadi et al., 2022) to serve as our training set, and utilized protein data from the BFVD database (Kim et al., 2025) as our test set.

Concerning the RBP data, we gather complex structures from EMDB (wwp, 2024) and PDB (Burley et al., 2017).

3.3 Baselines

In the field of protein structure prediction, we are selected Alphafold3 (Abramson et al., 2024) as our baseline model. For protein sequence generation, we have chosen ESM-IF[1], PiFold (Gao et al., 2022), and ProteinMPNN (Dauparas et al., 2022) as our

[1] https://github.com/facebookresearch/esm

baselines. In RNA design, RNAFlow (Nori and Jin, 2024) is selected as the baseline.

3.4 Evaluation Metrics

We use the **TM-score** and **RMSD** to compare the similarity between structures, and employ **recovery rate** and **sequence similarity** to compare the similarity between sequences.

4 Results and Analysis

4.1 Protein Structure Prediction

In the task of protein structure prediction (Table 1), compared to AlphaFold3, which is based on deep neural networks and has an enormous number of parameters, our LM2Protein (with only about 1.5 billion parameters) slightly underperforms in the TM-score (54.80% vs. 62.52%). However, it performs comparably in terms of RMSD (6.9428 Å vs. 6.4000 Å), indicating from another perspective that the predicted structures have a reasonably good level of convergence. Overall, although LM2Protein's performance still lags behind AlphaFold3, it can achieve usable and reasonable three-dimensional structure predictions under conditions of fewer parameters.

Models	TM-score	RMSD
AlphaFold3	62.52	6.4000 Å
LM2Protein	54.80	6.9428 Å

Table 1: Comparison of structure prediction performance between AlphaFold3 and LM2Protein models.

4.2 Protein Sequence Design

In the task of protein sequence design, LM2Protein significantly outperforms ProteinMPNN, PiFold, and ESM-IF in both Recovery Rate and Sequence Similarity metrics. This demonstrates that our discrete approach can more effectively capture and reconstruct the features of protein sequences while maintaining high sequence similarity, thereby generating amino acid sequences that better match the target structure.

Models	Recovery	Sequence Similarity
ProteinMPNN	34.26	29.61
PiFold	30.35	27.02
ESM-IF	30.43	26.32
LM2Protein	51.00	39.30

Table 2: Protein sequence design results comparing ProteinMPNN, PiFold, ESM-IF, and LM2Protein.

4.3 RNA design based on RBP Sequence

In the task of RNA design based on RBP sequence, LM2Protein demonstrates exceptional sequence generation capabilities. It achieves a recovery rate of 73.7%, a more than threefold improvement compared to the 21.0% of the baseline model, RNAFlow. This significant advantage validates that LM2Protein's discrete representation method can effectively capture and leverage protein sequence information to generate RNA sequences that better meet specific binding requirements.

However, despite the outstanding performance in sequence generation, the model faces challenges at the tertiary structure level. When the generated RNA sequences are evaluated for their structural accuracy, the metrics are an RMSD of 10.0282 Å, and a TM-score of 42.32%.

Models	Recovery	Sequence Similarity
RNAFlow	21.0	28.0
LM2Protein	73.7	72.5

Table 3: Comparison of RNA design performance between RNAFlow and LM2Protein models.

4.4 RNA design based on RBP Structure

In the task of RNA design based on protein structures, LM2Protein achieve high Recovery, significantly outperforming RNAFlow. This demonstrates that the model can effectively use a discrete representation of protein structures to generate target RNA sequences with greater precision. However, the 3D structures of the generated RNA sequences show room for improvement. The designed sequences achieve an RMSD of 17.5493 Å, and a TM-score of 34.53%.

Models	Recovery	Sequence Similarity
RNAFlow	26.0	32.0
LM2Protein	72.1	70.7

Table 4: Comparison of RNA design performance between RNAFlow and LM2Protein models.

5 Related Work

5.1 Protein Representation and Language Modeling

Deep learning models can accurately predict 3D conformations from sequences (Jumper et al., 2021; Lin et al., 2023; Li et al., 2025). Existing language models (e.g., ESM (Rives et al., 2021)) or generative models (e.g., ProtGPT2 (Ferruz et al., 2022), ProGen (Madani et al., 2023)) often omit explicit structural context or rely on specialized geometric encoders. We adopt structural tokenization, mapping 3D coordinates to discrete indices, which allows a standard Seq2Seq model to uniformly process both structures and sequences.

5.2 Inverse Protein Folding

Inverse folding aims to design an amino acid sequence for a given protein backbone (Yue and Dill, 1992). The field has shifted from computationally expensive physics-based methods (e.g., ROSETTA; Das and Baker, 2008) to being dominated by deep learning models. Current leading models, such as PROTEINMPNN (Dauparas et al., 2022) which uses message-passing, and ESM-IF (Hsu et al., 2022) which employs Geometric Vector Perceptrons, rely on custom geometric networks to process continuous 3D coordinates. This structure-conditioned generative paradigm has also been extended to other molecular design tasks; for instance, SURFPRO (Song et al., 2024) designs sequences from molecular surfaces, and DS-PROGEN (Li et al., 2025) uses surfaces and structural designs to generate RNA sequences.

5.3 Generative Design of RNA for RNA-Binding Proteins

Designing RNA sequences to bind specific RNA-binding proteins is a key challenge in synthetic biology. However, current computational methods are either purely predictive, identifying binding sites without generating new sequences (Shen et al., 2025; Xu et al., 2023), or employ decoupled generative pipelines. Flow matching-based models like RNAFlow rely on the interplay between an inverse folding module and a separate structure prediction network (RF2NA) to iteratively generate structures and sequences (Nori and Jin, 2024).

6 Conclusion

We introduced LM2Protein, a unified framework that processes discrete structural tokens within a LLM to bridge protein sequence and 3D geometry. This simplified approach achieves high sequence recovery in protein-conditioned RNA design. However, it also reveals a critical challenge: The model captures local sequence motifs but struggles with global folding constraints. Future work will focus on integrating biophysical priors to enhance structural accuracy and extending the framework to other molecular systems, paving the way for a universal model for in silico molecular engineering.

Limitations

The reliance on VQ-VAE for structural tokenization may underrepresent rare or anomalous protein topologies. Limited structural resolution may constrain the model's ability to capture high-fidelity 3D features, thus falling short of physics-based simulations. Scaling up to larger or more intricate molecular systems demands further optimization to balance model complexity and computational costs. For RBP-based RNA design, the generated sequences still leave room for improvement with respect to accurately capturing the target tertiary structure.

In addition, we only use the AI tool to polish the language of the paper.

Ethnics Statement

This paper does not involve issues related to ethics.

Acknowledgment

We want to thank our anonymous area chair and reviewers for their feedback.

References

2024. Emdb—the electron microscopy data bank. *Nucleic acids research*, 52(D1):D456–D465.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, and 1 others. 2024. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. 2017. Protein data bank (pdb): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*, pages 627–641.

Rhiju Das and David Baker. 2008. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, 77(1):363–382.

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, and 1 others. 2022. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348.

Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. 2022. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, and 1 others. 2025. Simulating 500 million years of evolution with a language model. *Science*, page eads0018.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. 2022. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pages 8946–8970. PMLR.

Jiyue Jiang, Pengan Chen, Jiuming Wang, Dongchen He, Ziqin Wei, Liang Hong, Licheng Zong, Sheng Wang, Qinze Yu, Zixian Ma, and 1 others. 2025a. Benchmarking large language models on multiple tasks in bioinformatics nlp with prompting. *arXiv preprint arXiv:2503.04013*.

Jiyue Jiang, Zikang Wang, Yuheng Shan, Heyan Chai, Jiayi Li, Zixian Ma, Xinrui Zhang, and Yu Li. 2025b. Biological sequence with language model prompting: A survey. *arXiv preprint arXiv:2503.04135*.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, and 1 others. 2021. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

Rachel Seongeun Kim, Eli Levy Karin, Milot Mirdita, Rayan Chikhi, and Martin Steinegger. 2025. Bfvd—a large repository of predicted viral protein structures. *Nucleic Acids Research*, 53(D1):D340–D347.

Yanting Li, Jiyue Jiang, Zikang Wang, Ziqian Lin, Dongchen He, Yuheng Shan, Yanruisheng Shao, Jiayi Li, Xiangyu Shi, Jiuming Wang, and 1 others. 2025. Ds-progen: A dual-structure deep language model for functional protein design. *arXiv preprint arXiv:2505.12511*.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, and 1 others. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, and 1 others. 2023. Large language models generate functional protein sequences across diverse families. *Nature biotechnology*, 41(8):1099–1106.

Divya Nori and Wengong Jin. 2024. Rnaflow: Rna structure & sequence design via inverse folding-based flow matching. *arXiv preprint arXiv:2405.18768*.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and 1 others. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.

Xilin Shen, Yayan Hou, Xueer Wang, Chunyong Zhang, Jilei Liu, Hongru Shen, Wei Wang, Yichen Yang, Meng Yang, Yang Li, and 1 others. 2025. A deep learning model for characterizing protein-rna interactions from sequences at single-base resolution. *Patterns*, 6(1).

Zhenqiao Song, Tinglin Huang, Lei Li, and Wengong Jin. 2024. Surfpro: Functional protein design based on continuous surface. *arXiv preprint arXiv:2405.06693*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, and 1 others. 2022. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444.

Zhenyu Wang, Zikang Wang, Jiyue Jiang, Pengan Chen, Xiangyu Shi, and Yu Li. 2025. Large language models in bioinformatics: A survey. *arXiv preprint arXiv:2503.04490*.

Yiran Xu, Jianghui Zhu, Wenze Huang, Kui Xu, Rui Yang, Qiangfeng Cliff Zhang, and Lei Sun. 2023. Prismnet: predicting protein–rna interaction using in vivo rna structural information. *Nucleic Acids Research*, 51(W1):W468–W477.

Kaizhi Yue and Ken A Dill. 1992. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences*, 89(9):4163–4167.

Jiayou Zhang, Barthelemey Meynard-Piganeau, James Gong, Xingyi Cheng, Yingtao Luo, Hugo Ly, Le Song, and Eric Xing. 2024. Balancing locality and reconstruction in protein structure tokenizer. *bioRxiv*.