DeProt: Protein language modeling with quantizied structure and disentangled attention

Mingchen Li[2,3]† **Yang Tan**[2,3]† **Bozitao Zhong**[1] **Ziyi Zhou**[1] **Huiqun Yu**[3]
Xinzhu Ma[2] **Wanli Ouyang**[2] **Liang Hong**[1,2*] **Bingxin Zhou**[1*] **Pan Tan**[1,2*]
[1] Shanghai Jiao Tong University, China
[2] Shanghai Artificial Intelligence Laboratory, China
[3] East China University of Science and Technology, China
{hongl3liang,bingxin.zhou,tpan1039}@sjtu.edu.cn

Abstract

Protein language models have exhibited remarkable representational capabilities in various downstream tasks, notably in the prediction of protein functions. Despite their success, these models traditionally grapple with a critical shortcoming: the absence of explicit protein structure information, which is pivotal for elucidating the relationship between protein sequences and their functionality. Addressing this gap, we introduce DeProt, a Transformer-based protein language model designed to incorporate protein sequences and structures. It was pre-trained on millions of protein structures from diverse natural protein clusters. DeProt first serializes protein structures into residue-level local-structure sequences and use a graph neural network based auto-encoder to vectorized the local structures. Then, these vectors are quantized and formed a discrete structure tokens by a pre-trained codebook. Meanwhile, DeProt utilize disentangled attention mechanisms to effectively integrate residue sequences with structure token sequences. Despite having fewer parameters and less training data, DeProt significantly outperforms other state-of-the-art (SOTA) protein language models, including those that are structure-aware and evolution-based, particularly in the task of zero-shot mutant effect prediction across 217 deep mutational scanning assays. Furthermore, DeProt exhibits robust representational capabilities across a spectrum of supervised-learning downstream tasks. Our comprehensive benchmarks underscore the innovative nature of DeProt's framework and its superior performance, suggesting its wide applicability in the realm of protein deep learning. For those interested in exploring DeProt further, the code, model weights, and all associated datasets are accessible at: https://github.com/ginnm/DeProt.

1 Introduction

Proteins have a myriad of diverse functions that underlie the complexity of life [1]. Protein Language Models (PLMs), inspired from Natural Language Processing methods, have heralded a new era in bioinformatics and structural biology. These models have become indispensable tools for protein representation. They effectively capture the semantic and grammatical features inherent within protein sequences through unsupervised training on extensive sequence datasets. As the databases of protein sequences have grown exponentially over the past decades, data-driven approaches for modeling proteins at scale made substantial progress. These PLMs have been trained on diverse protein sequence databases and different unsupervised objectives. They have shown remarkable proficiency in various protein related tasks.

*Corresponding authors. †Equal contribution.

Preprint. Under review.

While the primary sequences of protein provide valuable insights, it is the three-dimensional structures of protein that offer a more profound understanding of their function. Despite the critical role of structural information, the development of large-scale pre-trained models that utilize this data remains nascent, primarily hindered by the scarcity of protein structure data compared to sequence data. Nevertheless, the groundbreaking AlphaFold2 [2] has precipitated a paradigm shift, accurately predicting protein structures and thereby inflating the repository of structural data to an unprecedented scale, with hundreds of millions of structures now available [3, 4]. This expansion allows for the development of large pre-trained models based on structure data. Initiatives like ESM-IF1, leverage AlphaFold's predicted structures to enrich the datasets, enhancing the performance of the inverse folding task through augmented data volumes. Yet, challenges persist in directly incorporating structural data into protein language models, notably due to potential discrepancies between predicted structures and their empirical counterparts, which may lead to overfitting [5]. Recent works have tried to tackle this by converting 3D structures into a sequence of discrete tokens using quantization method like foldseek [6]. This approach not only decreases the risk of overfitting but also facilitates the incorporation of structure data into Transformer architectures. Such models include ProtsT5 [7] and SaProt [5], which build upon existing sequence models. For instance, SaProt extends the ESM model [8], while ProtsT5 evolves from the ProtT5 model [9]. Collectively, these models demonstrate that incorporating structural information significantly enhances their representational capacity. However, methods like foldseek, which focuses on encoding geometric information of individual amino acids, may overlook the nuances of protein local structures, such as catalytic sites or binding pockets, which are critical to protein functionality [10].

In this work, we propose an innovative framework, DeProt(**D**is**e**ntangled **Prot**ein sequence-structure model), tailored for protein structure-sequence pre-training. DeProt innovatively quantizes protein structures to mitigate overfitting and is adeptly engineered to amalgamate sequence and structure tokens. The framework features a module for protein structure serialization and quantization, integrated with a Transformer architecture that utilizes disentangled attention to harmonize protein sequences with their corresponding structures. Contrary to approaches that encode geometry information of individual residue, DeProt utilizes graph neural network (GNN) to represent the local structure of a residue, enabling it to exchange information with other residues in the local-structure through the edges of the graph. These sequences are then vectorized via a pre-trained encoder and subsequently quantized into structure token sequences using a pre-trained codebook. A novel disentangled attention mechanism is employed to fuse the structured token sequence with the residue sequence, explicitly modeling the ternary relationship among primary sequence, three-dimensional structure, and the relative positions of residues. We posit that this disentangled attention mechanism will capture more intricate features of protein sequences and structures compared to the prevalent self-attention mechanisms utilized by models such as SaProt or ProtsT5. DeProt underwent pre-training on an extensive database of 18 million protein structures from a subset of AlphaFoldDB, utilizing masked language modeling [4]. Our experimental results indicate that DeProt surpasses previous structure-aware Protein Language Models (PLMs) in both zero-shot mutant effect prediction and downstream supervised fine-tuning tasks, underscoring the substantial promise of integrating structural data with protein sequences.

2 Related work

2.1 Protein representation models

Based on their input modality, protein representation models can be divided into three categories: residue sequence-based models, structure-based models and hybrid structure-sequence models.

2.1.1 Residue-sequence based models

Residue sequence based models treat proteins as a sequence of residue tokens, using the Transformer model [11] for unsupervised pre-training on extensive datasets of sequence. These models are mostly inspired by the pre-training work of natural language. Based on the pre-training objective, current residue sequence based models can be further divided into BERT-based models [12], GPT-based models [13] and Span-Mask based models. BERT-style models, including ESM-series models [14, 15, 8], ProteinBert[16], TAPE [17]. Models that use a GPT-style auto-regressive training goal include Tranception [18], ProGen2 [19], ProtGPT2 [20], and models that use Span-mask as the

training objective include Ankh [21], ProtT5 [9], xTrimo [22]. In addition to the basic sequence models, there are also models used for representing multiple sequence alignments, such as MSA-Transformer [23].

2.1.2 Structure-based models

Protein structures play a dominant role in protein functionality. Thus, models that incorporate structure information often yield more accurate predictions. Recently, various techniques have been applied in learning protein structure representation, including CNN-based models [24], Graph Neural Networks such as GVP [25], IEConv [26], GearNet [27] and ProtLGN [28] have demonstrated significant versatility in integrating protein-specific features through node or edge attributes. Moreover, the recent advancements in protein folding models [2, 8, 29] have provided structure-based models with access to extensive datasets of protein structures. This has led to a growing interest in developing pre-training models that leverage protein structure information [7, 5, 27].

2.1.3 Hybird structure-sequence models

Hybrid models, which incorporate both sequence and structure information of proteins, offer more effective representations of proteins. For example, the LM-GVP[30] model employs ProtBERT-BFD [16] embeddings as input features for the GVP [25] model, while ESM-GearNet [31] investigates various methods of integrating ESM-1b [14] representations with GearNet [27]. Similarly, the recent ProtSSN [32] model leverages ESM-2 [8] embeddings as input for the EGNN [33] model, resulting in notable advancements. Both ESM-IF1 [34] and MIF-ST [35] target inverse folding, utilizing the structure to predict corresponding protein residues, whereas ProstT5 [7] focuses on the transformation between residue sequences and their structure token sequences [6] as a pre-training objective. SaProt [5] constructs a structure-aware vocabulary using structure tokens generated by foldseek [6]. Both SaProt and ProstT5 extensively utilize large structure databases [3] for their pre-training datasets. DeProt is also a hybrid structure-sequence model. Compared to previous work, DeProt develops an advanced model architecture and a method for structure quantization.

2.2 Protein structure quantization

Most protein structure models used to characterize proteins utilize continuous representations of protein structures, such as coordinates and dihedral angles. However, directly using these continuous features in the pre-training process can lead to overfitting [5]. This issue arises from the mismatch between the training set (derived from AlphaFold predictions) and the test set (from the wet-lab experiments). Quantization serves as a crucial means to eliminate these differences. There have been a few methods for the quantification of protein structures, which are generally designed for protein structure classification. These include the traditional secondary structure DSSP [36] defined by the physical location of the tertiary structure, and foldseek [6] developed based on VQ-VAE [37]. Among them, SaProt and ProstT5 utilize the serialized model of foldseek. We also tried foldseek, but the results were not satisfactory, and foldseek can only be applied to entire proteins. To address this issue, we have developed a protein serialization and quantization method focused on local-structure.

3 Method

DeProt is a new Transformer-based protein language model, which enhances the performance of protein language models by considering the amino acid information and 3D structure information of proteins using two novel techniques: structure quantization (Section 3.1) and sequence-structure disentangled attention (Section 3.2).

3.1 Structure quantization module

The structure quantization module aims to transform a residue's local structure in a protein into a discrete token. Initially, the local structure is encoded into a dense vector using a pre-trained structure encoder. Subsequently, a pre-trained k-means clustering model assigns a category label to the local structure based on the embedded vector. Finally, the category label is the structure token for the residue. The process of structure serialization is shown in Figure 1.

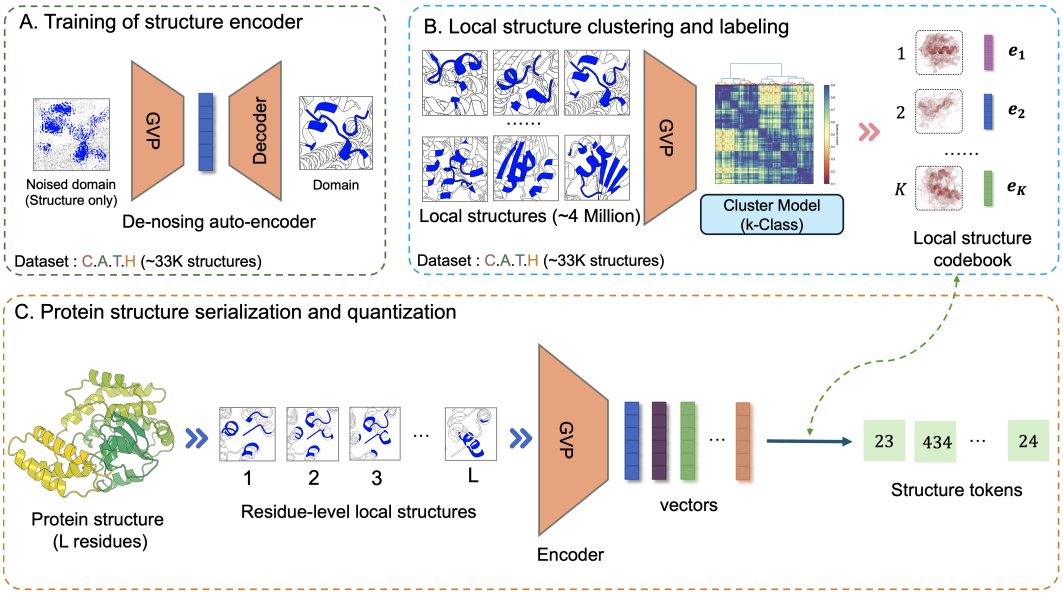


Figure 1: Structure serialization and quantization. (A) Training the structure encoder. (B) Local structure clustering and labeling. (C) Converting a protein structure to structure token sequence.

Structure construction. We represent a structure using graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ and $\mathbf{E}$ represent the residue-level nodes and edges, respectively. For any given node $\boldsymbol{v} \in \boldsymbol{V}$, it contains only the coordinate information of the residue, but does not include any type information of the residue itself. This ensures that the structure encoder is solely focused on the structure. The edge set $\boldsymbol{E} = \{\boldsymbol{e}_{ij}\}$ includes all i, j for which $\boldsymbol{v}_j$ is one of the 40 nearest neighbors of v_i, determined by the distance between their $\boldsymbol{C}_\alpha$ atoms. The features of the edges include the angle and distance between two residues, as well as the positional encoding.

Structure encoder. Based on the definition of structure mentioned above, we use geometric vector perceptrons (GVP) as the structure encoder to embed local-structures [25]. The GVP can be represented as a structure feature extraction function $\pi_\theta(\boldsymbol{G}) \in \mathbb{R}^{l \times d}$, where l is the number of nodes, d is the embedding dimension and θ is trainable parameters. We integrate GVP with a decoder that includes a position-wise multi-layer perceptron (MLP) to form an auto-encoder model. The entire model is trained using a de-noising pre-training objective. In this process, Gaussian noise is randomly added to the structure coordinates (Figure 1A). The model is then tasked with recovering the structure to its original, noise-free state. After being trained on the CATH 4.3 dataset [38], consisting of approximately 33,185 protein domains, we exclude the decoder and utilize solely the mean pooled output of the encoder as the ultimate representation of structures. Despite the structure encoder being trained on domains, it can effectively be applied to structures of varying sizes, including local structures. Therefore, for a specified graph $\boldsymbol{G}$, the encoding process can be described by: $\boldsymbol{r} = \frac{1}{l}\sum_{i=1}^{l} \pi_\theta(\boldsymbol{g}_i)$, where $\boldsymbol{g}_i$ represents the features or the sub-graph associated with the i-th node in the graph $\boldsymbol{G}$, and $\pi_\theta(\boldsymbol{g_i}) \in \mathbb{R}^d$ is the output of the encoder for the i-th node. Here, $\boldsymbol{r} \in \mathbb{R}^d$ is the mean pooled output of the encoder and the vectorized representation of the structure.

Local structure clustering and labeling. We utilize clustering to quantize structure vectors. To achieve this, we first build a large local-structure dataset derived from the C.A.T.H dataset [38], which involves creating a local-structure for each residue in the protein. This process yielded about 4 million local structures. Then we employ the structure encoder to encoding these local-structures into vectors. Based on the Euclidean distance, we utilize the k-means algorithm to cluster these vectors into K distinct groups and we get a code-book of structure vectors, as shown in Figure 1B. This codebook contains K vectors, denoted as $\{\boldsymbol{e}\}_{i=1}^{K}$. For any given structure, we first encode it into a vector using the structure encoder. Then the nearest vector $\boldsymbol{e}_j$ within the codebook is identified, with j serving as the structure token for the local-structure. And the symbol K is also referred to as the structure vocab size.

In general, for a residue at position i in a protein sequence, we first construct a graph $\boldsymbol{G}_i$ only based on the 3D structure information, and then encode it into a vector $\boldsymbol{r}_i$ using a structure encoder. Then we use the codebook to assign a class $s_i \in 1, 2, ..., K$ to this vector as the structure token for the residue's local-structure. Then the entire protein sequence can be serialized then quantized into a sequence of structure tokens (Figure 1C).

3.2 Sequence-structure disentangled attention

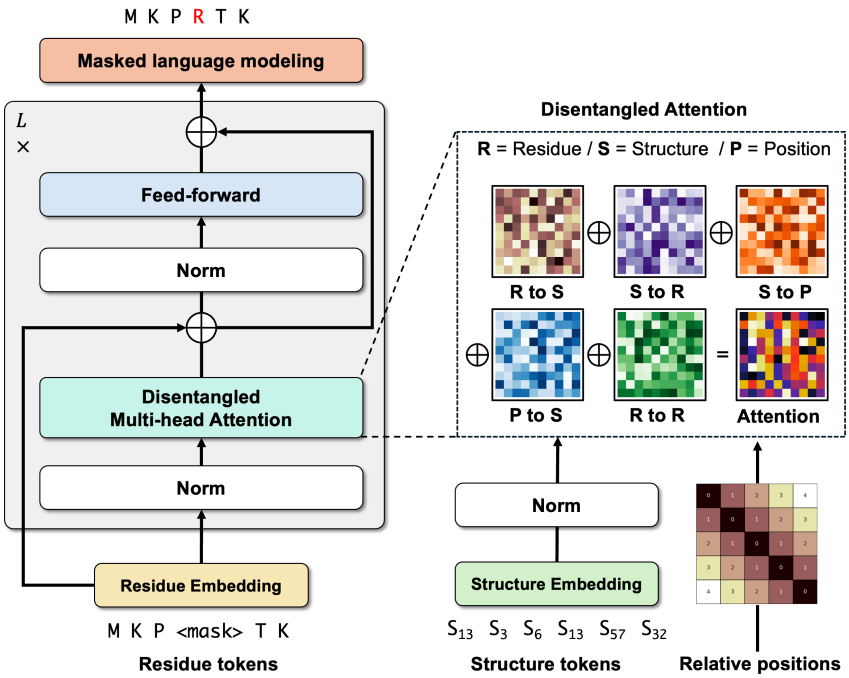


Figure 2: Model architecture of DeProt. DeProt is a Transformer-style model and the difference is that DeProt uses disentangled attention instead of self-attention [11].

Inspired by DeBerta [39], we use an expanded form of disentangled attention to combine the attention of residual sequences and structure sequences as well as relative positions. Specifically, for a residue at position i in a protein sequence, it is represented by three vectors: $\boldsymbol{R}_i$ denotes its residue token hidden state, $\boldsymbol{S}_i$ represents the embedding of residue-level local structure, and $\boldsymbol{P}_{i|j}$ is the embedding of relative position with the token at position j. The calculation of the cross attention score between residue i and residue j can be decomposed into nine components as:

$$\begin{aligned}\boldsymbol{A}_{i,j} &= \{\boldsymbol{R}_i, \boldsymbol{S}_i, \boldsymbol{P}_{i|j}\} \times \{\boldsymbol{R}_j, \boldsymbol{S}_j, \boldsymbol{P}_{j|i}\}^\top \\ &= \boldsymbol{R}_i\boldsymbol{R}_j^\top + \boldsymbol{R}_i\boldsymbol{S}_j^\top + \boldsymbol{R}_i\boldsymbol{P}_{j|i}^\top \\ &+ \boldsymbol{S}_i\boldsymbol{R}_j^\top + \boldsymbol{S}_i\boldsymbol{S}_j^\top + \boldsymbol{S}_i\boldsymbol{P}_{j|i}^\top \\ &+ \boldsymbol{P}_{i|j}\boldsymbol{R}_j^\top + \boldsymbol{P}_{j|i}\boldsymbol{S}_j^\top + \boldsymbol{P}_{j|i}\boldsymbol{P}_{j|i}^\top\end{aligned} \tag{1}$$

The attention weight of a residue pair can be calculated by adding nine attention scores. This is achieved by using separate matrices based on the residue tokens, structure tokens, and relative positions. These matrices are utilized for various interactions such as *residue-to-residue, residue-to-structure, residue-to-position, structure-to-residue, structure-to-structure, structure-to-position, position-to-residue, position-to-structure, and position-to-position*. Since our approach focuses on the masked language modeling task for residue tokens, we found that the terms structure-to-structure, structure-to-position, position-to-structure, and position-to-position do not significantly contribute additional information. Consequently, these terms are removed from our implementation of the attention weight calculation. As shown in Figure 2, our sequence-structure disentangled attention mechanism includes 5 types of attention.

In the following, we use single-head attention as an example to demonstrate the operation of sequence-structure disentangled attention. To begin, we define the relative position of the i-th to the j-th residue, denoted as $\delta(i, j)$:

$$\delta(i,j) = \begin{cases} 0 & \text{if } i-j \leq -l \\ 2k-1 & \text{if } i-j \geq l \\ i-j+l & \text{otherwise} \end{cases} \tag{2}$$

where, l represents the maximum relative distance we consider, which is set to $l = 1024$ in our implementation.

Similar to standard self-attention operation [11], the computation of query, key for structure, residue and relative position, and the value for sequence is as follows:

$$\begin{aligned} \boldsymbol{Q}^r &= \boldsymbol{R}\boldsymbol{W}_r^q, \boldsymbol{K}^r = \boldsymbol{R}\boldsymbol{W}_r^k, \boldsymbol{V}^r = \boldsymbol{R}\boldsymbol{W}_r^v \\ \boldsymbol{Q}^s &= \boldsymbol{S}\boldsymbol{W}_s^q, \boldsymbol{K}^s = \boldsymbol{S}\boldsymbol{W}_s^k \\ \boldsymbol{Q}^p &= \boldsymbol{P}\boldsymbol{W}_p^q, \boldsymbol{K}^p = \boldsymbol{S}\boldsymbol{W}_p^k \end{aligned} \tag{3}$$

And the the attention score $\hat{\boldsymbol{A}}_{i,j}$ from residue i to residue j is calculated as follows:

$$\hat{\boldsymbol{A}}_{i,j} = \underbrace{\boldsymbol{Q}_i^r(\boldsymbol{K}_j^r)^\top}_{\text{(a) residue-to-residue}} + \underbrace{\boldsymbol{Q}_i^r(\boldsymbol{K}_j^s)^\top}_{\text{(a) residue-to-structure}} + \underbrace{\boldsymbol{Q}_i^r\boldsymbol{K}_{\delta(i,j)}^p{}^\top}_{\text{(b) residue-to-position}} + \underbrace{\boldsymbol{K}_j^s(\boldsymbol{Q}_i^r)^\top}_{\text{(a) structure-to-residue}} + \underbrace{\boldsymbol{K}_j^p\boldsymbol{Q}_{\delta(j,i)}^r{}^\top}_{\text{(b) position-to-residue}} \tag{4}$$

where $\boldsymbol{Q}_i^r$ represents the i-th row of the matrix $\boldsymbol{Q}^r$, and $\boldsymbol{K}_j^r$ denotes the j-th row of $\boldsymbol{K}^r$. Similarly, $\boldsymbol{Q}_i^s$ and $\boldsymbol{K}_j^s$ are the i-th and j-th rows of $\boldsymbol{Q}^s$ and $\boldsymbol{K}^s$, respectively. The term $\boldsymbol{K}_{\delta(i,j)}^p$ refers to the row in $\boldsymbol{K}^p$ indexed by the relative distance $\delta(i, j)$, and $\boldsymbol{Q}_{\delta(j,i)}^p$ refers to the row in $\boldsymbol{Q}^p$ indexed by the relative distance $\delta(j, i)$. To normalize the attention scores, a scaling factor of $\frac{1}{\sqrt{5d}}$ is applied to $\boldsymbol{A}$. This scaling is crucial for ensuring the stability of model training [39], particularly when dealing with large-scale language models. All the $\hat{\boldsymbol{A}}_{ij}$ form the attention matrix, and the final output hidden vector is $\boldsymbol{H}_o$:

$$\boldsymbol{H}_o = \text{softmax}(\hat{\boldsymbol{A}})V^r \tag{5}$$

which is used as the input for the hidden state of the next layer of residues.

3.3 Pre-training objective

DeProt is pre-trained using structure conditioned masked language modeling. In this approach, each input sequence $\boldsymbol{x}$ is noised by substituting a fraction of the residues with a special mask token. The objective of DeProt is to predict the original tokens that have been masked in the input sequence, utilizing both the corrupted sequence and its structure token sequence $\boldsymbol{s}$ as context:

$$\boldsymbol{\mathcal{L}}_{MLM} = \boldsymbol{E}_{\boldsymbol{x}\sim\boldsymbol{X}}\boldsymbol{E}_{\boldsymbol{M}} \sum_{i\in\boldsymbol{M}} -\log p(\boldsymbol{x}_i|\boldsymbol{x}_{/\boldsymbol{m}}, \boldsymbol{s}) \tag{6}$$

Based on state-of-the-art protein language models, we randomly select 15% of the indices $\boldsymbol{M}$ for masking, substituting the residue at each selected index i with a mask token. The training objective is to minimize the negative log likelihood for each masked residue $\boldsymbol{x}_i$, based on the partially masked sequence $\boldsymbol{x}/\boldsymbol{M}$ and the unmasked structure token, serving as contextual information. Therefore, in order to accurately predict the masked tokens, the model must not only learn the dependencies between residues but also consider the relationship between residues and structures.

3.4 Pre-training configuration

Dataset. The pre-training dataset is curated from AlphaFoldDB [3], which contains more than 214 million structures predicted by AlphaFold [2]. We select protein structures that exhibit high prediction confidence as indicated by their pLDDT scores and employ Foldseek clustering [4] to reduce structure redundancy by 90%. Consequently, we compile a database comprising 18.8 million structures. From

this collection, we randomly select 100,000 structures for validation and another 100,000 for testing, enabling us to monitor the perplexity throughout the training process.

Hyper-parameters DeProt is trained on a cluster of DGX-A800 GPU (8×80G) server in BF16 precision for three days. The model has 12 transformer layers, 12 attention heads, and 768 embedding dims with 3172 feed-forward embedding dimensions using GELU activation function. We train with 4096 tokens per micro batch for 500,000 steps. We use AdamW [40] as our optimizer with β_1 and β_2 set to 0.9 and 0.98, and a weight decay value of 0.01. We warm up the learning rate from 0 to 0.0002 over the first 2000 steps, then decay it by a cosine schedule to the 0. We use a dropout rate of 0.1 and clip gradients using a clipping value of 1.0. For the tokenization of the protein data, we use the residue-level tokenizer which is adopted in several PLMs [14, 8, 15].

4 Experiments

In this section, we comprehensively evaluated the representation ability of DeProt, including evaluating zero-shot mutant effect prediction tasks on the ProteinGYM [41] and assessing its supervised fine-tuning abilities on various function prediction tasks [42, 43, 44, 24, 45, 46, 17]. Additionally, we conducted ablation experiments on the the clustering number K of the structure quantization model and the model architecture as well as and the degree of structure information required, identifying the factors that contribute most importantly. We summarize the experimental results as follows:

- DeProt achieves state-of-the-art performance in zero-shot mutant effect prediction tasks, especially in stability prediction.
- DeProt outperforms other structure-aware protein language models in all protein-level tasks. This outcome once again demonstrates that integrating structure information into PLMs leads to superior protein representation.
- The selection of K, i.e., the clustering model number of structure quantization model, impacts the performance. We found that setting $K = 2048$ yields the best results. Moreover, the structure quantization model outperforms other structure serialization methods in DeProt architecture, including foldSeek [6] and secondary structure [36].
- Various types of attentions contribute uniquely, with the attention from structure-to-sequence being particularly crucial. Utilizing structure information is essential for enhancing the performance of the DeProt. Without leveraging structure information, the effectiveness of the DeProt model cannot surpass that of sequence-based protein language models.

4.1 Zero-shot mutant effect prediction

Zero-shot scoring method. Previous studies have demonstrated that protein language models, when trained on extensive and varied protein sequence databases, are capable of predicting experimental measurements of protein mutants function without further supervision. For those protein language models that are trained with masked language modeling objective, the calculation of mutation scores can be formalized as follows:

$$\sum_{i \in \boldsymbol{M}} \log P(\boldsymbol{x}_i = \boldsymbol{m}_i | \boldsymbol{x}) - \log P(\boldsymbol{x}_i = \boldsymbol{w}_i | \boldsymbol{x}) \tag{7}$$

Here $\boldsymbol{M}$ is the mutations set, and $\boldsymbol{M} = \{(i, \boldsymbol{m}_i, \boldsymbol{w}_i) | i = 1, 2, ..., |\boldsymbol{M}|\}$ is a set of triplets, where $i \in \mathbb{N}$ represents the position of a point mutation, $\boldsymbol{m}_i$ is the residue after the point mutation, and $\boldsymbol{w}_i$ is the original residue of the point mutation. $\boldsymbol{x}$ is the sequence of amino acids of the wild type.

We slightly modify the formula above to adapt to DeProt, where the structure sequence is an additional condition to score mutants:

$$\sum_{i \in \boldsymbol{M}} \log P(\boldsymbol{x}_i = \boldsymbol{m}_i | \boldsymbol{x}, \boldsymbol{s}) - \log P(\boldsymbol{x}_i = \boldsymbol{w}_i | \boldsymbol{x}, \boldsymbol{s}) \tag{8}$$

Here, $\boldsymbol{s}$ is the structure token sequence of the wild type.

Model Type	Model	ProteinGYM			ProteinGYM-Stability		
		ρ_s	NDCG	Top-recall	ρ_s	NDCG	Top-recall
Evolution-based	EVE	0.439	0.781	0.230	0.491	0.838	0.283
	GEMME	0.457	0.777	0.211	0.519	0.828	0.240
	MSA-Transformer	0.434	0.779	0.217	0.421	0.835	0.279
Sequence-based	Tranception	0.434	0.779	0.220	0.476	0.824	0.279
	ESM-1v	0.374	0.732	0.211	0.437	0.791	0.265
	ESM-2	0.414	0.747	0.217	0.523	0.834	0.300
	DeProt (-structure)	0.392	0.741	0.184	0.430	0.750	0.191
Inverse-folding	ESM-IF	0.422	0.748	0.223	0.624	**0.864**	0.344
	MIF-ST	0.401	0.765	0.226	0.485	0.839	0.300
Ensemble Models	Trancepiton-EVE	0.457	0.786	0.230	0.501	0.839	0.279
	ESM-1v*	0.407	0.749	0.211	0.477	0.811	0.245
Sequence-Structure models	SaProt	0.426	0.758	0.221	0.552	0.855	0.315
	DeProt(K=4096)	0.491	0.784	0.239	0.650	0.841	0.330
	DeProt(K=2048)	**0.502**	**0.793**	**0.252**	**0.663**	0.851	**0.348**
	DeProt(K=1024)	0.476	0.783	0.236	0.626	0.825	0.306
	DeProt(K=512)	0.474	0.783	0.234	0.618	0.819	0.296
	DeProt(K=128)	0.460	0.769	0.225	0.589	0.801	0.275
	DeProt(K=20)	0.422	0.750	0.195	0.539	0.781	0.224
	DeProt(foldseek)	0.474	0.775	0.217	0.571	0.790	0.231
	DeProt(dssp)	0.439	0.760	0.204	0.527	0.774	0.233

Table 1: Comparison of zero-shot mutation prediction performance between DeProt and other models. ρ_s is the Spearman rank correlation evaluation metric. DeProt(foldseek) uses tokens serialized by foldseek [6] as structural tokens, while DeProt(dssp) uses secondary structure [36] tokens as structural tokens. The K in the parentheses following DeProt represents the number of clustering model.

Datasets. To evaluate the effectiveness of DeProt in zero-shot mutant effect prediction, we conducted experiments using ProteinGym [41]. It contains over 3.5 millions mutated sequences across 217 protein assays, spanning different functions, including stability, activity, expression, organism fitness and binding affinity. However, most proteins in the benchmark lack a human-measured structure. Consequently, we utilize AlphaFold2 [2] to generate the structures of wild-type sequences.

Baselines. We compare DeProt with the current state-of-the-art models, including sequence-based models, structure-sequence models, inverse folding models, and evolutionary models. For sequence models, we compared ESM1v [15], ESM2 [8], and Tranception [18]. For sequence-structure models, we compared SaProt [5]. For inverse folding models, we compared ESM-IF [34] and MIF-ST [35] models. For evolutionary models, we compared GEMME [47], MSA-Transformer [23], and EVE [48]. In addition, we also compared ensemble models, including ESM1v-ensemble [15] and Tranception-EVE [49].

Metrics. We use the Spearman's rank correlation coefficient (ρ_s) to evaluate model performance, which is the most commonly used metric for the evaluation of protein mutant effect predictions.Additionally, we compute the Normalized Discounted Cumulative Gains (NDCG) to prioritize models that assign the highest scores to sequences exhibiting the most significant DMS values. We further assess model effectiveness using Top-recall, selecting the top 10% of DMS values as our benchmark. The details of these metrics are detailed in Appendix D.

Results. Table 4.1 shows the zero-shot performance results on the ProteinGYM. Based on the results, we draw several noteworthy conclusions:

- DeProt outperforms all baselines on zero-shot mutant effect predictions of ProteinGYM. Especially, it achieves the best performance in predicting stability, aligning with the previous findings that models incorporating structure information typically perform better in stability predictions [41].

- DeProt is compatible with other sequence structure quantization methods, while our structure quantization module outperforms other structure quantization methods, including foldseek [6] and DSSP [36]. And replacing the structure quantization module with foldseek, which is used by SaProt, can still achieve competitive performance, despite SaProt being based on the ESM-2 architecture, having more than five times parameters as DeProt. This demonstrates the effectiveness of DeProt architecture.

Model	# Params	DeepLoc	Metal Ion Binding	Thermostability	GO-MF	GO-BP	GO-CC
		Acc%	Acc%	ρ_s	F1-Max	F1-Max	F1-Max
ESM-2	650M	91.96	71.56	0.680	0.668	0.345	0.411
ESM-1b	650M	92.83	73.57	0.708	0.661	0.320	0.392
MIF-ST	643M	91.76	75.08	0.694	0.627	0.239	0.248
GearNet	42M	89.18	71.26	0.571	0.650	0.354	0.404
SaProt-35M	35M	91.96	71.56	0.680	0.668	0.345	0.411
SaProt-650M	650M	93.55	75.75	0.724	**0.678**	0.356	0.414
DeProt	110M	**94.68**	**76.67**	**0.724**	0.672	**0.387**	**0.483**

Table 2: Comparison of supervised fine-tuning on downstream tasks. The parameters refer to all parameters of the models, including the sequence and structure components.

- The structure quantization achieved the best performance when the clustering number is $K = 2048$. As the K value gradually increases from 20 to 2048, we observe an improvement in performance. However, when the K value increases to 4096, we notice a decline in its Zero-shot performance, so we did not continue testing higher K values. When the structure module is completely removed, that is, when K is set to 0, we find a significant decrease in the performance of the DeProt model, which demonstrates the importance of the structure module to the DeProt model.

4.2 Supervised fine-tuning for downstream tasks

To comprehensively evaluate DeProt, we conducted benchmarks across 4 downstream protein-related tasks, including thermostability prediction, Metal Ion Binding prediction, protein localization prediction and GO annotations prediction. Since DeProt achieve its best zero-shot performance when K is set to 2048, we select $K = 2048$ as the default configuration for DeProt. Unless specifically noted, DeProt in this section represents DeProt$(K = 2048)$. These downstream task datasets are detailed as follows:

- **Thermostability**. The task is to predict the thermostability values of proteins using the "Human-cell" divisions from the Thermostability task in FLIP [42]. For this regression task, the Spearman correlation coefficient is utilized as the evaluation metric to evaluate the prediction results.
- **DeepLoc** (Protein Sub-cellular Localization). The task is to output a probability distribution across two sub-cellular localization categories for a protein. This is a binary classification task, and we utilize accuracy as the metric to evaluate the predictions. This dataset was introduced by DeepLoc [44] and we use the original data split.
- **Metal Ion Binding**. The task is to predict whether there exist metal ion-binding sites within a protein. This is also a binary classification task, and we utilize accuracy as the metric to evaluate the predictions. This dataset was introduced by TAPE [17], and we use the original data split.
- **GO annotations prediction**. This task is to predict Gene Ontology terms to evaluate the model's ability to predict protein functions. This task was introduced by DeepFRI [24], and we use three types of GO labels: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). This is a multi-label classification task, and we evaluate the model using the Max F1-Score.

Baselines We compared DeProt with protein language models including ESM-2[8], ESM-1b [14], and the sequence-structure model SaProt [5] (two parameter versions, 35M, 650M), MIF-ST [35], as well as the protein structure representation model GearNet [27].

Fine-tuning settings

To ensure fair comparisons, we fine-tuned DeProt using a fixed set of hyper-parameters. We use for the Adam optimizer with β_1 set to 0.9, β_2 to 0.98, and applied an $L2$ weight decay of 0.001. The batch size was maintained at 64 (If 64 causes the GPU memory to explode, we will reduce the batch size and then use gradient accumulation to achieve the same batch size.) and the learning rate was set at 0.00003, except for Go annotation prediction, where it was adjusted to 0.00001 due to that there are hundreds of labels. We fine-tuned all model parameters for 200 epochs, and we choose

	DeepLoc	ProteinGYM	Perplexity
DeProt(K=4096)	94.16	0.491	**8.880**
DeProt(K=2048)	**94.68**	**0.502**	9.033
DeProt(K=1024)	93.41	0.476	9.333
DeProt(K=512)	93.71	0.474	9.577
DeProt(K=128)	93.18	0.460	10.021
DeProt(K=20)	93.18	0.422	10.719
DeProt(foldseek)	93.24	0.474	10.049
DeProt(dssp)	93.21	0.439	10.009

Table 3: Ablation experiment results of different structure vocab sizes.

	DeepLoc	ProteinGYM	Perplexity
DeProt	**94.68**	**0.502**	**9.033**
DeProt (- P2R)	91.46	0.478	9.173
DeProt (- R2P)	92.46	0.466	9.410
DeProt (- R2S)	92.19	0.438	12.142
DeProt (- S2R)	91.38	0.475	9.355
DeProt (- position)	72.34	0.052	13.885
DeProt (- structure)	89.65	0.392	12.190

Table 4: The ablation experiment results with different attentions.

the best checkpoints based on validation set performance. Following SaProt [5], we downloaded all protein structures identified by Uniprot IDs from AFDB [3], and any proteins not found in AFDB were excluded.

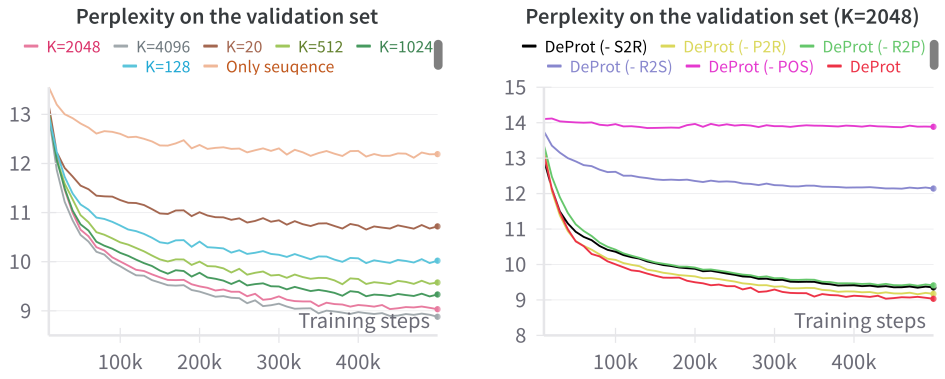
Results The results on the supervised fine-tuning downstream tasks are shown in Table 4.2. It can be observed that on most datasets, the performance of DeProt can surpass both protein language models and sequence-structure models. This result shows the importance of incorporating structure information into protein representation models. Moreover, compared to other structure-sequence models, DeProt model is more capable of effectively integrating protein sequence and structure information. Therefore, our model not only achieves state-of-the-art (SOTA) performance on Zero-shot mutation prediction tasks but also have competitive performance in supervised learning tasks.

4.3 Ablation Study

Since DeProt has multiple attentions and different structure vocab size, we need to justify the contribution of different components. We use the zero-shot mutant effect prediction capability on ProteinGYM, accuracy on the DeepLoc supervised downstream task, and perplexity on the pre-trained test set to conduct the ablation experiments.

Ablation on structure vocab size K. The results are shown in Table 4.3. It can be noted that there is a negative relation between perplexity on the test set and structure vocab size. However, the zero-shot mutant effect prediction performance and transferability achieve the optimal performance at a structure vocab size of $K = 2048$. So we take $K = 2048$ as the default configuration for the DeProt model. To more clearly demonstrate the relationship between the size of the structure vocab and perplexity, we have marked in Figure 3(a) the change in the perplexity on the validation set with training steps. It can be observed that as the size of the structured vocabulary increases, the convergence of DeProt improves progressively. We can also note that when k is set to 0, i.e., removing the structure module, the perplexity of the model significantly increases. This observation underscores the effectiveness of incorporating additional structure information in enhancing the representation capabilities and zero-shot performance of protein language model. Additionally, we compared the performance of other structure serialization and quantization models on the DeProt model, including DSSP [36] and foldseek [6]. We found that their performance metrics are similar to our protein structure quantization module with $K = 128$.

Ablation on disentangled attention. As shown in Figure 2, the sequence-structure disentangled attention incorporates five types of attentions, making it crucial to verify the necessity of each. We individually removed each type of attention from DeProt (selecting $K = 2048$) and conducted



(a) Ablation of structure vocab size.

(b) Ablation of attentions.

Figure 3: Ablation experiment. The change in Perplexity on the validation set for different models during the training process. The horizontal axis represents training steps, and the vertical axis represents perplexity.

experiments. Additionally, we evaluated the performance of DeProt when completely omitting positional or structure information. The results are shown in Table 4.3. We also show the changes in perplexity on the validation set for models with different attentions in Figure 3(b). The term "-S2R" denotes the removal of structure-to-residue attention, "-P2R" denotes the removal of position-to-residue attention, "-R2P" denotes the removal of residue-to-position attention, and "-R2S" denotes the removal of residue-to-structure attention. The term "-POS" denotes the removal of position encodings. The findings indicate that removing any type of attention leads to a performance decline. Moreover, it was observed that for perplexity and zero-shot mutant effect prediction, discarding position-to-residue or structure-to-residue attentions has a relatively minor impact on the model, whereas removing residue-to-structure attention has the most significant effect. Overall, every type of disentangled attention in DeProt is crucial, illustrating the rationality of the DeProt model.

5 Conclusion

In this paper, we propose a Transformer-based protein model, DeProt, that integrates structure information and residue information. DeProt was pre-trained on 18 million protein structures using a masked language modeling objective. The experiments demonstrate that DeProt achieves state-of-the-art performance on zero-shot mutation prediction and excels in supervised learning tasks, even when operating with fewer parameters and less training data . We also introduce a novel method for protein structure serialization and quantization. First, we trained a de-noising auto-encoder based on graph neural networks (GNNs) on protein structures from PDB data bank. Then, we constructed a code-book using the k-means algorithm, which allows us to find the nearest label in the code-book to quantitatively represent the representation vector of a protein structure. For a protein structure including L residues, we can represent it as L local-structures, and then quantitatively represent these local-structures to get structure tokens. This discrete representation method can reduce the storage space needed to represent protein structures and avoid over-fitting. We also proposed a new type of disentangled attention to improve the self-attention mechanism. The disentangled attention mechanism includes not only the self-attention of residues but also the cross attention between residues and sequence positions, and residues and structure tokens. The experimental results show that our attention module can effectively combine sequence and structure information. Although DeProt achieved impressive performance, there are still some potential directions for improvement, including expanding the use of larger structure datasets to train larger parameter versions of DeProt, using distillation methods to improve the extraction and encoding process of local-structures, and reducing computational costs, etc.

6 Acknowledgements

This work was supported by the grants from the National Science Foundation of China (grant number 12104295), the Computational Biology Program of Shanghai Science and Technology Commission (23JS1400600), Shanghai Jiao Tong University Scientific and Technological Innovation Funds (21X010200843), the Student Innovation Center at Shanghai Jiao Tong University, and Shanghai Artificial Intelligence Laboratory

References

[1] William P Jencks. *Catalysis in chemistry and enzymology*. Courier Corporation, 1987.

[2] 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.

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

[4] Inigo Barrio-Hernandez, Jingi Yeo, Jürgen Jänes, Milot Mirdita, Cameron LM Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao, and Martin Steinegger. Clustering predicted structures at the scale of the known protein universe. *Nature*, 622(7983):637–645, 2023.

[5] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024.

[6] 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*, pages 1–4, 2023.

[7] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Martin Steinegger, and Burkhard Rost. Prostt5: Bilingual language model for protein sequence and structure. *bioRxiv*, pages 2023–07, 2023.

[8] 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.

[9] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

[10] Hongyuan Lu, Daniel J Diaz, Natalie J Czarnecki, Congzhi Zhu, Wantae Kim, Raghav Shroff, Daniel J Acosta, Bradley R Alexander, Hannah O Cole, Yan Zhang, et al. Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature*, 604(7907):662–667, 2022.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] 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.

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

[15] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc., 2021.

[16] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[17] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[18] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022.

[19] 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.

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

[21] Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568*, 2023.

[22] Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*, 2024.

[23] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

[24] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

[25] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021.

[26] Pedro Hermosilla, Marco Schäfer, Matej Lang, Gloria Fackelmann, Pere-Pau Vázquez, Barbora Kozlikova, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. In *International Conference on Learning Representations*, 2021.

[27] Zuobai Zhang, Minghao Xu, Arian Rokkum Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *The Eleventh International Conference on Learning Representations*, 2023.

[28] Bingxin Zhou, Lirong Zheng, Banghao Wu, Yang Tan, Outongyi Lv, Kai Yi, Guisheng Fan, and Liang Hong. Protein engineering with lightweight graph denoising neural networks. *bioRxiv*, pages 2023–11, 2023.

[29] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pages 2022–07, 2022.

[30] Zichen Wang, Steven A Combs, Ryan Brand, Miguel Romero Calvo, Panpan Xu, George Price, Nataliya Golovach, Emmanuel O Salawu, Colby J Wise, Sri Priya Ponnapalli, et al. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific reports*, 12(1):6832, 2022.

[31] Zuobai Zhang, Minghao Xu, Aurelie Lozano, Vijil Chenthamarakshan, Payel Das, and Jian Tang. Enhancing protein language model with structure-based encoder and pre-training. In *ICLR 2023 - Machine Learning for Drug Discovery workshop*, 2023.

[32] Yang Tan, Bingxin Zhou, Lirong Zheng, Guisheng Fan, and Liang Hong. Semantical and topological protein encoding toward enhanced bioactivity and thermostability. *bioRxiv*, pages 2023–12, 2023.

[33] 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.

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

[35] Kevin K Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection*, 36:gzad015, 2023.

[36] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[37] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[38] Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.

[39] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[41] Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 64331–64379. Curran Associates, Inc., 2023.

[42] Christian Dallago, Jody Mou, Kadina E Johnston, Bruce Wittmann, Nick Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[43] Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. Exploring evolution-aware & -free protein language models as protein function predictors. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 38873–38884. Curran Associates, Inc., 2022.

[44] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

[45] Xiao-Yong Pan, Ya-Nan Zhang, and Hong-Bin Shen. Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *Journal of proteome research*, 9(10):4992–5001, 2010.

[46] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. Peer: A comprehensive and multi-task benchmark for protein sequence understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35156–35173. Curran Associates, Inc., 2022.

[47] Elodie Laine, Yasaman Karami, and Alessandra Carbone. Gemme: a simple and fast global epistatic model predicting mutational effects. *Molecular biology and evolution*, 36(11):2604–2619, 2019.

[48] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

[49] Pascal Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora Susan Marks. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022.