# ProtCLIP: Function-Informed Protein Multi-Modal Learning

**Hanjing Zhou**[1,2,3*] **Mingze Yin**[1,2*], **Wei Wu**[4], **Mingyang Li**[3], **Kun Fu**[3], **Jintai Chen**[5†], **Jian Wu**[2,6†], **Zheng Wang**[3†]

[1]College of Computer Science and Technology, Zhejiang University,
[2]State Key Laboratory of Transvascular Implantation Devices of The Second Affiliated Hospital, Zhejiang University,
[3]Alibaba Cloud Computing,
[4]School of Artificial Intelligence and Data Science, University of Science and Technology of China,
[5]AI Thrust, Information Hub, HKUST(Guangzhou),
[6]Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence
{zhj85393, mzyin256, jtchen721}@gmail.com, wujian2000@zju.edu.cn,
{sangheng.lmy, fukun.fu, wz388779}@alibaba-inc.com, urara@mail.ustc.edu.cn

## Abstract

Multi-modality pre-training paradigm that aligns protein sequences and biological descriptions has learned general protein representations and achieved promising performance in various downstream applications. However, these works were still unable to replicate the extraordinary success of language-supervised visual foundation models due to the ineffective usage of aligned protein-text paired data and the lack of an effective function-informed pre-training paradigm. To address these issues, this paper curates a large-scale protein-text paired dataset called *ProtAnno* with a property-driven sampling strategy, and introduces a novel function-informed protein pre-training paradigm. Specifically, the sampling strategy determines selecting probability based on the sample confidence and property coverage, balancing the data quality and data quantity in face of large-scale noisy data. Furthermore, motivated by significance of the protein specific functional mechanism, the proposed paradigm explicitly model protein static and dynamic functional segments by two segment-wise pre-training objectives, injecting fine-grained information in a function-informed manner. Leveraging all these innovations, we develop ProtCLIP, a multi-modality foundation model that comprehensively represents function-aware protein embeddings. On 22 different protein benchmarks within 5 types, including protein functionality classification, mutation effect prediction, cross-modal transformation, semantic similarity inference and protein-protein interaction prediction, our ProtCLIP consistently achieves SOTA performance, with remarkable improvements of 75% on average in five cross-modal transformation benchmarks, 59.9% in GO-CC and 39.7% in GO-BP protein function prediction. The experimental results verify the extraordinary potential of ProtCLIP serving as the protein multi-modality foundation model.

## 1 Introduction

Proteins are essential functional units of cells, responsible for performing a wide range of vital and versatile functions crucial to life. Mirroring the language-supervised pre-training paradigm towards powerful and unified vision representations (Radford et al. 2021; Ramesh et al. 2022; Girdhar et al. 2023; Junnan et al. 2023), previous work has explored in the pre-training of multi-modality Protein Language Models (PLMs) by aligning protein sequences with textual function descriptions to achieve function-centric protein representations (Zhang et al. 2022; Xu et al. 2023; Wu, Chang, and Zou 2024; Yin et al. 2024). However, these works were still unable to replicate the extraordinary success of image-text foundation models, and have shown to discard fine-grained protein functional information (Wu, Chang, and Zou 2024), which results in the suboptimal performance on cross-modal transformation (Wang et al. 2024) and localization prediction (Xu et al. 2023). Literature has summarized that the success of visual foundation models primarily stems from **the efficient utilization of large-scale data** (Radford et al. 2021; Chen et al. 2024) and **a holistic multi-modal pre-training framework** (Zhang et al. 2023; Pujin et al. 2023), which points to two inherent obstacles that hinder further progress in multi-modal protein-biotext pre-training:

**(i) Absence of large-scale datasets and ineffective data usage.** Large-scale aligned dataset is an indispensable part of obtaining powerful multi-modality foundation models. However, biotexts describing protein functions are much harder to construct than image captions, as often requiring detailed annotated process including manual review by experts or computational analysis by machines. This highlights the pressing need of large-scale multi-modal datasets containing protein sequences with high-quality functional annotations across multiple attribute domains. Even with large-scale protein-biotext pairs, it is non-trivial to effectively inject biological property information into PLMs during multi-modal pre-training. This is primarily because the machine-analyzed process leads to numerous noisy labels (*i.e.*, less accurate annotations) (Bairoch and Apweiler 2000). Currently, there is still a lack of efficient learning techniques to effectively utilize large-scale proteins with noisy annotations for protein-biotext pre-training.

**(ii) Lack of a function-informed pre-training paradigm.** Unlike the alignment of natural image-text pairs, the under-

*These authors contributed equally.
†Corresponding authors

standing of proteins is strongly influenced by their specific functional mechanism, which has been largely neglected by previous research yet. Proteins perform specific biological functions depending on their corresponding functional domains in 3D structural spaces. The amino acids at these active site are contiguous or discrete in 1D protein sequences. In this paper, we introduce the static and dynamic functional segment, new concepts which directly determine the specific protein functions and should be primarily focused during the alignment with biological function descriptions. However, we find existing protein-biotext pre-training works directly take after the original CLIP methodology for coarse-grained alignment, discarding the fine-grained information of protein unique functional mechanism (*i.e.*, static or dynamic functional segments primarily determine protein specific functions and properties), which significantly prevents the better performance of protein-biotext pre-training.

Our work proposes a step towards constructing a universally applicable protein multi-modality foundation model aligning biological and natural language. We present ProtCLIP, consistently alleviates the aforementioned two intrinsic problems and introduces remarkable innovations in multiple dimensions including the pre-training data, sampling strategy, and multi-modality objectives.

We first construct a high-quality protein-biotext paired dataset *ProtAnno* with sparse version (*ProtAnno-S*) and dense version (*ProtAnno-D*), derived from the existing protein function database (Consortium 2019). ProtCLIP employs *ProtAnno-D* comprising 251.5 million aligned pairs for large-scale protein-biotext pre-training, which is the same order of magnitude as large-scale image-text pre-training. Since there exist some inevitable noisy annotations in *ProtAnno-D* (caused by machine-annotated bias), we propose a novel property-driven sampling strategy motivated by (Berthelot et al. 2019; Li, Socher, and Hoi 2020). Compared to the vanilla uniformly sampling, the proposed sampling strategy decides the selecting probability based on the sample confidence and property coverage, simultaneously balancing the data quality and data quantity in face of large-scale noisy labels. Furthermore, a function-informed pre-training paradigm is constructed motivated by significance of the protein functional mechanism. Within such paradigm, we utilize CLIP loss (Radford et al. 2021) to inject coarse-grained information, and two segment-wise objectives are designed to capture fine-grained information of the static and dynamic functional segments. Concretely, on the one hand, we design a cross-modality reconstruction module to recover the masked static segments based on knowledge from both modalities. On the other hand, the property prototype is exploited to aggregate dynamic segments in an unsupervised way. The resulting property-grouped dynamic segments are contrasted with property prototypes within the same protein-biotext pair, mitigating the mutual interference across multiple attribute domains.

Evaluated by extensive experiments, ProtCLIP sets new state-of-the-art on 22 important yet challenging protein benchmarks within five types. For protein classification engineering and mutation effect prediction, the superiority of ProtCLIP in representation learning attributes to incorporation of multi-modal information (*e.g.*, 59.9%/39.7% improvements in Go-CC/GO-BP benchmarks). For cross-modal transformation, ProtCLIP surpasses baselines by a significant margin (75% improvement). For semantic similarity inference and protein-protein interaction prediction, ProtCLIP ranks the best, which verifies effectiveness of the proposed data-efficient and function-informed multi-modal learning.

## 2 Methods

In this section, we first describe the curated multi-modal dataset, *ProtAnno*, and the property-driven sampling strategy to enhance data usage effectiveness. Next, we introduce the model architectures and our novel function-informed pre-training paradigm, which incorporates holistic multi-modal pre-training objectives to capture both coarse-grained and fine-grained information. Finally, we depict the overall loss function used for protein-biotext pre-training.

| Dataset | Conf-L1 | Conf-L2 | Conf-L3 | Conf-L4 | Conf-L5 |
|---|---|---|---|---|---|
| *ProtAnno-S* | 0.1982 | 0.0980 | **0.6777** | 0.0229 | 0.0032 |
| *ProtAnno-D* | 0.0013 | 0.0057 | 0.3269 | **0.6661** | 0.0000 |

Table 1: Data distribution of *ProtAnno-S* and *ProtAnno-D* with different sample confidence. We highlight the confidence where protein entries are mostly concentrated in **bold**.

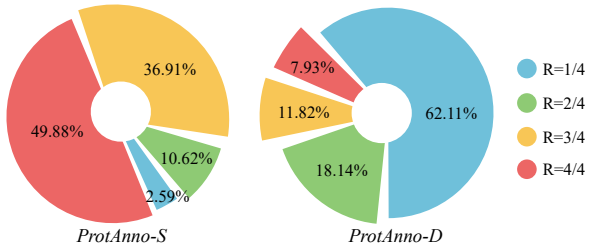


Figure 1: Data distribution of *ProtAnno-S* and *ProtAnno-D* with different property coverage.

### 2.1 Pre-training data

**Dataset Curation** To enable pre-training of the protein multi-modality foundation model aligning biological and natural language, it is essential to build dataset containing large-scale pairs of protein sequences and textual property descriptions. Our pre-training data is sourced from SwissProt and trEMBL (Bairoch and Apweiler 2000), containing proteins with textual descriptions. We align protein sequences with meticulously selected properties to curate *ProtAnno*, which is available in sparse version (*ProtAnno-S*) and dense version (*ProtAnno-D*). *ProtAnno-S* includes 0.5 million manually reviewed protein-biotext pairs with higher annotation quality, whereas *ProtAnno-D* comprises 251.5 million mostly computationally analyzed protein-biotext pairs which are less accurate due to the machine-annotated bias. To gain more insights into the dataset, we conduct extensive quantitative analyses, and display the compositional structure of *ProtAnno* with varying confidence $C$ and property coverage $R$ in Table 1 and Figure 1.

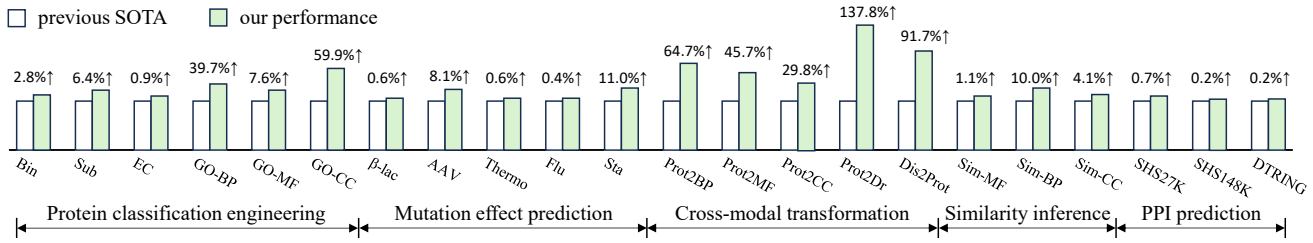


Figure 2: Comparison results on 22 downstream benchmarks within five types. ProtCLIP consistently achieves the state-of-the-art performance on all these tasks. PPI: protein-protein interaction.

**Property-driven Sampling Strategy** For protein-biotext pre-training, most prior works only used scarce proteins with manually reviewed annotations (equivalent to *ProtAnno-S*), and the attempt to incorporate plentiful computationally analyzed proteins (equivalent to *ProtAnno-D*) has been unsuccessful, declaring "*data quality could be more important than data quantity.*" (Xu et al. 2023). However, we question and rethink this issue, and propose the property-driven sampling strategy which integrate the merits of the multi-modality data quality and data quantity. Specifically, the main considerations for sampling probability are sample confidence $C$ and property coverage $R$ and data size $N$. Note that the smaller the confidence, the more reliable the entry is, and $C \in \{1, 2, 3, 4, 5\}$, $R \in \{1/4, 2/4, 3/4, 4/4\}$. Initially, we discard machine-annotated entries with $C = 4, 5$ (less accurate) and $R = 1/4, 2/4$ (low coverage) for comprehensive property understanding. Next, rather than uniform sampling, we explicitly build the sampling distribution according to the aforementioned three factors. The likelihood of selecting protein entries from cluster $u$ with $\{C_u, R_u, N_u\}$ during multi-modality pre-training is defined as:

$$P = \frac{C_u^{-3} \cdot \sqrt{R_u} \cdot N_u}{\sum_{i,j,k} C_i^{-3} \cdot \sqrt{R_j} \cdot N_k}. \tag{1}$$

In this paper, we perform large-scale protein-biotext pre-training exploiting *ProtAnno-D*, in conjunction with the proposed property-driven sampling strategy.

## 2.2 Model Architecture

The overview of our framework is displayed in Figure 3, which contains a protein encoder and a biotext encoder. The protein encoder is a protein language model for learning biological features from protein sequences and we use pre-trained ESM-2-650M (Lin et al. 2023) here. The biotext encoder is a text language model for learning linguistic features from biotext descriptions and we use PubMedBERT (Gu et al. 2021) here. Initialization with these two pre-trained large models significantly facilitates pre-training process by providing decent representations in the early stage of training.

## 2.3 Function-informed Pre-training Paradigm

To accomplish the holistic function-informed multi-modal pre-training, we jointly optimize four protein-biotext pre-training objectives, with two classic ones and two newly proposed segment-wise ones, customized for learning locality-aware and fine-grained information of protein specific functional mechanism.

**Global Contrastive Loss** Global Contrastive loss (GC) learning aligns representations of two modalities by encouraging positive pairs to have higher similarity in contrast to the negative pairs. Considering the effectiveness of $\boldsymbol{L}_{\text{GC}}$ for multi-modal understanding in many previous works (Radford et al. 2021; Junnan et al. 2023; Su et al. 2022) from different domains, we perform it to realize global alignment of protein-biotext. Given a batch of sequence-text pairs $\{(S_i, T_i)\}_{i=1}^K$, $\boldsymbol{L}_{\text{GC}}$ is composed of two symmetric standard InfoNCE loss:

$$\begin{aligned}\boldsymbol{L}_{\text{GC}} = -\frac{1}{2}\Bigg[&\mathbb{E}_{p(S,T)}\Big(log\frac{\exp(\texttt{sim}(S_i, T_i)/\tau_1)}{\sum_{j=1}^K \exp(\texttt{sim}(S_i, T_j)/\tau_1)}\Big) \\ &+ \mathbb{E}_{p(S,T)}\Big(log\frac{\exp(\texttt{sim}(T_i, S_i)/\tau_1)}{\sum_{j=1}^K \exp(\texttt{sim}(T_i, S_j)/\tau_1)}\Big)\Bigg],\end{aligned} \tag{2}$$

where $\texttt{sim}(;)$ is the consine similarity and $\tau_1$ denotes the temperature parameter that controls the softmax distribution.

**Biotext-guided Static Segment Reconstruction (BSR)** Given the global contrastive objective modeling coarse-grained information, the fine-grained information of static and dynamic segments are ubiquitous, which primarily determines protein specific functions and properties. To capture such locality-aware information of static segments, we propose Biotext-guided Static segment Reconstruction (BSR) to reconstruct corrupted static segments using information from both modalities. Specifically, given a sequence of protein residues $S = \{x_1, x_2, \ldots, x_n\}$, we sample $l$ consecutive tokens as a static segment at a time, until the total sampling length reaches 15% of $S$. In other words, we execute sampling iterations to prepare a random set of static segments $\{e_1, e_2, \ldots, e_m\}$ with $e_i \in S$ for subsequent masking and reconstruction. At each iteration, we randomly select the starting point of each segment and its length $l$ follows a discrete uniform distribution between 5 and 10. Note that all static segments are non-overlapping and their total length accounts for 15% of $S$.

Given the selected diverse static segments, we introduce a novel cross-modality reconstruction module to reconstruct masked segments according to the biotext functional descriptions, as displayed in Figure 3. Specifically, the protein sequence with masked segments $e^m$ and biotext $T$ are

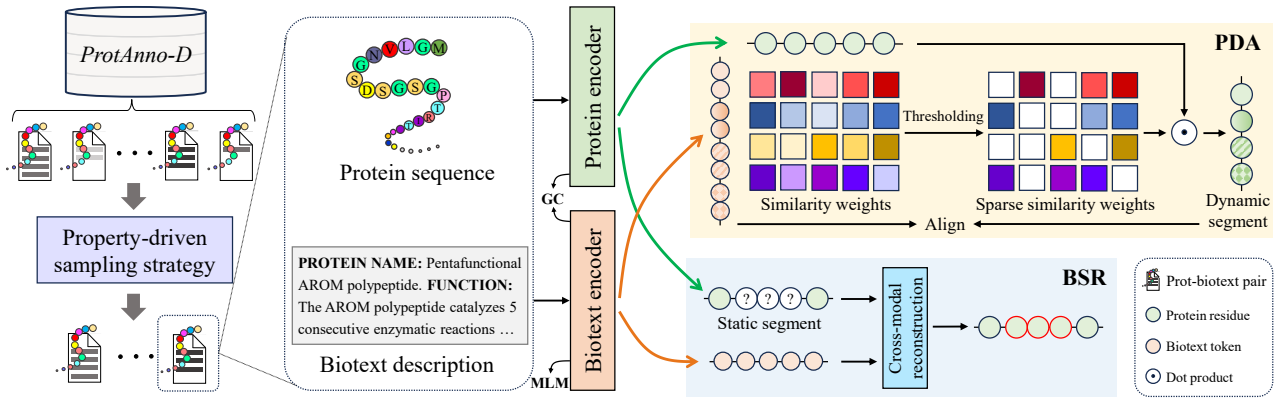


Figure 3: Overview of ProtCLIP. We curate a large-scale protein-biotext dataset *ProtAnno* with a property-driven sampling strategy, and proposes a function-informed pre-training paradigm containing two segment-wise objectives BSR and PDA.

fed into a cross-attention module to obtain the fused representation by attending to all tokens along the biological property description. Then a MLP with the GELU activation (Hendrycks and Gimpel 2016) and layer normalization (Ba, Kiros, and Hinton 2016) serves as the reconstruction head. Formally, the loss function for BSR is:

$$\boldsymbol{L}_{\text{BSR}} = \mathbb{E}_{p(T,e^m)} H(\Phi(T, e^m), y_e), \tag{3}$$

where $\Phi(T, e^m)$ is the predicted probability of protein sequence with masked static segments $e^m$, and $y_e$ is the corresponding ground truth. $H(;)$ is the cross-entropy function.

**Property-grouped Dynamic Segment Alignment (PDA)** To capture the fine-grained information of dynamic segments, we propose Property-grouped Dynamic Segment Alignment (PDA), optimizing the alignment between property-grouped dynamic segments and corresponding property descriptions.

Specifically, a prototype memory bank is constructed to approximate property descriptive sentences, without any need to accurately retain redundant information such as syntax. Then the property prototype is exploited to aggregate dynamic segments in an unsupervised way, which are more flexible than static segments in BSR. Provided property description prototypes of biotext $T = \{a_1, a_2, a_3, a_4\}$ and the corresponding sequence of residues $S = \{x_1, x_2, \ldots, x_n\}$, we first compute similarity weights as:

$$w_{ij} = a_i \cdot x_j, \; i = 1, 2, 3, 4, \; j = 1, 2, \ldots, n, \tag{4}$$

where $w_{ij} \in \mathbb{R}$ and $\cdot$ is the inner product. Then min-max normalization is applied along the residue dimension to normalize $w_{ij}$ to $[0, 1]$. After that, some non-functional protein residues are discarded by sparsifying the similarity weights with a threshold $\theta$:

$$\hat{w}_{ij} = \begin{cases} w_{ij}, & \text{if } w_{ij} \geq \theta \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Eventually, we obtain the property-grouped dynamic segments by multiplying similarity weights and protein residues:

$$e_i = \{\hat{w}_{ij} x_j \mid j = 1, 2, \ldots, n\}, \; i = 1, 2, 3, 4. \tag{6}$$

Property-grouped dynamic segment alignment is conducted to align these dynamic segments with property descriptions within the same protein- biotext pair, mitigating the mutual interference across multiple attribute domains:

$$\begin{aligned} \boldsymbol{L}_{\text{PDA}} = -\frac{1}{2} \Big[ & \mathbb{E}_{p(e,a)} \Big( log \frac{\exp(\texttt{sim}(e_i, a_i)/\tau_2)}{\sum_k \exp(\texttt{sim}(e_i, a_k)/\tau_2)} \Big) \\ & + \mathbb{E}_{p(e,a)} \Big( log \frac{\exp(\texttt{sim}(e_i, a_i)/\tau_2)}{\sum_k \exp(\texttt{sim}(e_i, a_k)/\tau_2)} \Big) \Big], \end{aligned} \tag{7}$$

where $\texttt{sim}(;)$ represents the consine similarity and $\tau_2$ denotes the temperature parameter that controls the softmax distribution.

Aiming to extract the essential knowledge of protein sequences, we select the most relevant residues based on their similarities to each property description, resulting in segments of variable lengths. Owing to such variable length, dynamic segments are flexible to capture information of consecutive or non-consecutive functional residues, excluding redundant and non-functional ones. Additionally, the threshold $\theta$ directly influences the segment length by determining different number of zero values in each row of the similarity weights, which decouples similarities of individual residues to different property descriptions. In essence, the thresholding operation allows for different properties to match different residues that are the most relevant, thereby forming dynamic segments.

## 2.4 Overall Loss Function

The overall loss function of ProtCLIP comprises four terms. Global contrastive loss $\boldsymbol{L}_{\text{GC}}$ learns coarse-grained information, while biotext-guided static segment reconstruction $\boldsymbol{L}_{\text{BSR}}$ and property-grouped dynamic segment alignment $\boldsymbol{L}_{\text{PDA}}$ focuses on fine-grained information. And we keep the protein masked language modeling $\boldsymbol{L}_{\text{MLM}}$ to preserve unimodal knowledge when injecting multi-modality information from biological texts. We optimize these terms jointly via a weighted sum with hyper-parameters $\lambda_1$ and $\lambda_2$:

$$\boldsymbol{L} = \boldsymbol{L}_{\text{GC}} + \lambda_1 \boldsymbol{L}_{\text{BSR}} + \lambda_2 \boldsymbol{L}_{\text{MLM}} + \boldsymbol{L}_{\text{PDA}}. \tag{8}$$

During the training process, we observe a significant mutual interference between segment-level reconstruction $\boldsymbol{L}_{\text{BSR}}$ and token-level reconstruction $\boldsymbol{L}_{\text{MLM}}$, and set $\lambda_1 + \lambda_2 = 1$. The investigation of their equilibrium is in Section 3.7.

# 3 Experiments

In this section, we first introduce some training setups, and then provide configurations and result discussions about five types of downstream applications (Figure 4) on totally 22 benchmarks. Eventually, the analysis of ablation experiments are presented to further validate the effectiveness of our pre-training objectives.

## 3.1 Training Setups

We build our codes upon the PyTorch framework and conduct experiments on 64 Tesla V100 GPUs with 10,000 GPU hours. An Adam optimizer is used (learning rate: $1.0 \times 10^{-5}$, weight decay: 0) to train the model. The batch size is 2048 and 512 for pre-training and downstream experiments. Within the function-informed pre-training paradigm, we set hyper-parameters $\theta = 0.3, \lambda_1 = 0.7, \lambda_2 = 0.3$.

## 3.2 Protein Classification Engineering

**Configurations** Protein classification engineering aims to classify protein locations and functions. For location classification, we consider two such problems from DeepLoc (Almagro Armenteros et al. 2017), subcellular localization prediction (Sub) with 10 categories and binary localization prediction (Bin) with 2 categories. For function classification, we employ two benchmarks (Gligorijević et al. 2021) namely Enzyme Commission (EC) number prediction and Gene Ontology (GO) term prediction. On GO benchmark, there are three branches that predict molecular function (GO-MF), biological process (GO-BP) and cellular component (GO-CC). The compared baselines include three parts: (a) two traditional protein encoders CNN (Shanehsazzadeh, Belanger, and Dohan 2020), LSTM (Rao et al. 2019); (b) four single-modal PLMs ProtBERT (Elnaggar et al. 2022), OntoProtein (Zhang et al. 2022), ESM-1b (Rives et al. 2021), ESM2 (Lin et al. 2023)); (c) one multi-modal PLM ProtST-ESM2 (Xu et al. 2023). The evaluation metrics are accuracy for location prediction, and AUPR and $F_{max}$ for function prediction.

**Results** Table 2 (left) and Table 3 show that ProtCLIP establishes state-of-the-art results on all six classification benchmarks under both linear probing and full tuning settings. Moreover, ProtCLIP performs best on protein classification engineering among all five type of downstream tasks.

## 3.3 Mutation Effect Prediction

**Configurations** Mutation effect prediction is a regression task that predicts the effect of residue mutations on protein fitness. We utilize $\beta$-lactamase ($\beta$-lac) landscape from PEER (Xu et al. 2022), Fluorescence (Flu) and Stability (Sta) landscapes from TAPE (Rao et al. 2019), and AAV and Thermostability (Thermo) landscapes from FLIP (Dallago et al. 2021). The baselines remain the same as mentioned in Section 3.2. The performance is measured by Spearman's $\rho$. Moreover, we evaluate ProtCLIP and PLMs under both linear probing and full tuning settings on location prediction and mutation effect prediction tasks.

| Modality | Method | **Loc class** (Acc %) | | **Effect pred** (Spearman's $\rho$) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Bin | Sub | $\beta$-lac | AAV | Thermo | Flu | Sta |
| Traditional models trained from scratch | | | | | | | | |
| Single | CNN | 82.67 | 58.73 | 0.781 | 0.746 | 0.494 | 0.682 | 0.637 |
| | LSTM | 88.11 | 62.98 | 0.139 | 0.125 | 0.564 | 0.494 | 0.533 |
| PLMs under linear probing | | | | | | | | |
| Single | ProtBERT | 81.54 | 59.44 | 0.616 | 0.209 | 0.562 | 0.339 | 0.697 |
| | OntoProtein | 84.87 | 68.34 | 0.471 | 0.217 | 0.605 | 0.432 | 0.688 |
| | ESM-1b | 91.61 | 79.82 | 0.528 | 0.454 | 0.674 | 0.430 | 0.750 |
| | ESM2 | 91.32 | 80.84 | 0.559 | 0.374 | 0.677 | 0.456 | 0.746 |
| Multiple | ProtST-ESM2 | 92.52 | 83.39 | 0.565 | 0.398 | 0.681 | 0.499 | 0.776 |
| | **ProtCLIP** | 94.39 | 83.65 | 0.565 | 0.532 | 0.682 | 0.503 | 0.795 |
| PLMs under full tuning | | | | | | | | |
| Single | ProtBERT | 91.32 | 76.53 | 0.731 | 0.794 | 0.660 | 0.679 | 0.771 |
| | OntoProtein | 92.47 | 77.59 | 0.757 | 0.791 | 0.662 | 0.630 | 0.731 |
| | ESM-1b | 92.40 | 78.13 | 0.839 | 0.821 | 0.669 | 0.679 | 0.694 |
| | ESM2 | 91.72 | 78.67 | 0.867 | 0.817 | 0.672 | 0.677 | 0.718 |
| Multiple | ProtST-ESM2 | 92.52 | 80.22 | 0.879 | 0.825 | 0.682 | 0.682 | 0.738 |
| | **ProtCLIP** | **95.08** | **85.34** | **0.884** | **0.892** | **0.686** | **0.685** | **0.819** |

Table 2: Results on location classification (Loc class) and mutation effect prediction (Effect pred) tasks. We highlight the best results in **bold**.

| Modality | Method | **EC** | | **GO-BP** | | **GO-MF** | | **GO-CC** | |
|---|---|---|---|---|---|---|---|---|---|
| | | AUPR | $F_{max}$ | AUPR | $F_{max}$ | AUPR | $F_{max}$ | AUPR | $F_{max}$ |
| Traditional model trained from scratch | | | | | | | | | |
| Single | CNN | 0.540 | 0.545 | 0.165 | 0.244 | 0.380 | 0.354 | 0.261 | 0.387 |
| | LSTM | 0.032 | 0.082 | 0.130 | 0.248 | 0.100 | 0.166 | 0.150 | 0.320 |
| PLMs under full tuning | | | | | | | | | |
| Single | ProtBERT | 0.859 | 0.838 | 0.188 | 0.279 | 0.464 | 0.456 | 0.234 | 0.408 |
| | OntoProtein | 0.854 | 0.841 | 0.284 | 0.436 | 0.603 | 0.631 | 0.300 | 0.441 |
| | ESM-1b | 0.884 | 0.869 | 0.332 | 0.452 | 0.630 | 0.659 | 0.324 | 0.477 |
| | ESM2 | 0.888 | 0.874 | 0.340 | 0.472 | 0.643 | 0.662 | 0.350 | 0.472 |
| Multiple | ProtST-ESM2 | 0.898 | 0.878 | 0.342 | 0.482 | 0.647 | 0.668 | 0.364 | 0.487 |
| | **ProtCLIP** | **0.906** | **0.908** | **0.567** | **0.574** | **0.696** | **0.691** | **0.582** | **0.541** |

Table 3: Results on function classification task. We highlight the best results in **bold**.

**Results** Table 2 illustrates that ProtCLIP consistently ranks the best among other baselines. We can observe that although traditional models (*e.g.*, CNN) pose strong competition in mutation effect prediction, ProtCLIP still retains the lead, especially on Stability benchmark in full tuning setting.

## 3.4 Cross-modal Transformation

**Configurations** Cross-modal transformation matches the transformed embedding with candidates from the target modality, where embeddings from ProtCLIP are transformed by an extra transformation module. Following (Wang et al. 2024), we leverage the raw knowledge graph (KG) data and undertake some preprocessing steps, with the training/validation/test split of 80%/10%/10%. The baselines are BioBridge (Wang et al. 2024) and three knowledge graph embedding methods (ComplEx (Trouillon et al. 2016), DistMult (Yang et al. 2015), RotatE (Sun et al. 2019)). We use mean reciprocal rank (MRR) as the metric.

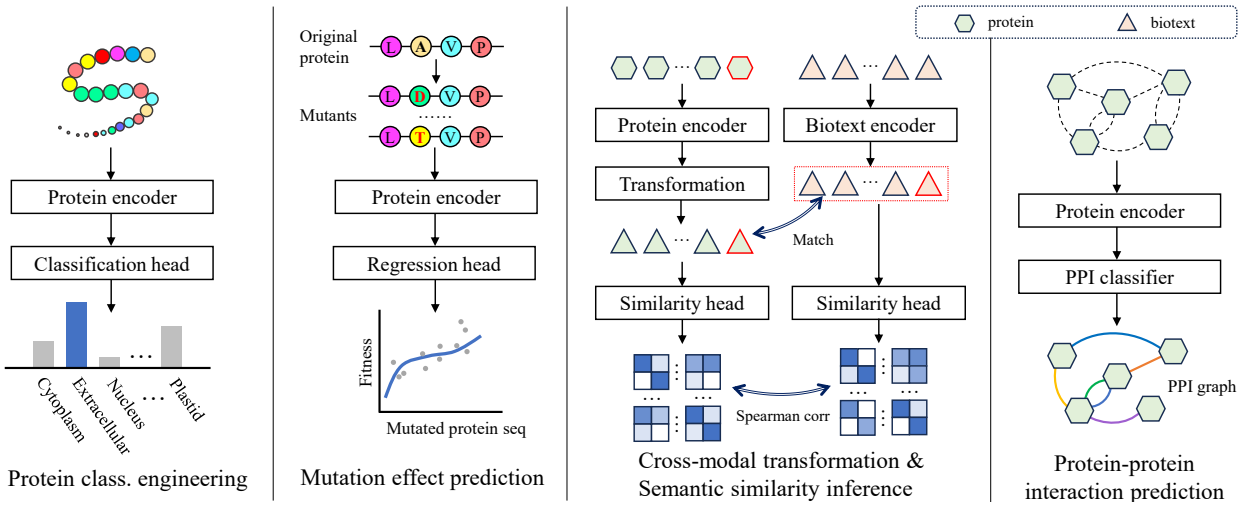


Figure 4: An overview of downstream tasks within five types.

| Method | Prot2BP | Prot2MF | Prot2CC | Prot2Drug | Disease2Prot |
|---|---|---|---|---|---|
| ComplEx | 0.084 | 0.100 | 0.099 | 0.079 | 0.059 |
| DistMult | 0.054 | 0.089 | 0.095 | 0.044 | 0.033 |
| RotatE | 0.079 | 0.119 | 0.107 | 0.125 | 0.070 |
| BioBridge | 0.136 | 0.326 | 0.319 | 0.172 | 0.084 |
| **ProtCLIP** | **0.224** | **0.475** | **0.414** | **0.409** | **0.161** |

Table 4: Mean reciprocal rank (MRR) results on cross-modal transformation task. Prot: protein.

| Method | Sim (Spearman's $\rho$) | | | PPI (F1 score) | | |
|---|---|---|---|---|---|---|
| | MF | BP | CC | SHS27K | SHS148K | STRING |
| ESM2-3B | 0.33 | 0.42 | 0.23 | 0.732 | 0.733 | 0.834 |
| KeAP | 0.41 | 0.41 | 0.40 | 0.733 | 0.726 | 0.834 |
| BioBridge | 0.91 | 0.80 | 0.73 | 0.739 | 0.739 | 0.836 |
| **ProtCLIP** | **0.92** | **0.88** | **0.76** | **0.744** | **0.740** | **0.838** |

Table 5: Results on semantic similarity inference (Sim) and protein-protein interaction prediction (PPI) tasks.

**Results** Table 4 reports our remarkable enhancement over all baselines. The first three baselines are traditional KG encoders trained from scratch, which lack flexibility, while BioBridge cannot fully unleash the potential of PLMs. Instead, ProtCLIP compensates for their shortcomings and incorporates flexibility, data-efficiency and high performance. Particularly, ProtCLIP is 2.4 $\times$ better than the best baseline for "Prot2Drug" and 2 $\times$ better for "Prot2BP" and "Disease2Prot", which signals the superiority of ProtCLIP in multimodal understanding.

### 3.5 Semantic Similarity Inference

**Configurations** Semantic similarity inference computes the relevance between predicted and groundtruth similarity matrices (Unsal et al. 2022). our goal is to evaluate the extent to which the encoded protein embeddings can capture biomolecular functional similarity (*i.e.*, BP, CC, MF). The predicted matrix contains pairwise Manhattan Similarities of the encoded protein embeddings, while the groundtruth stores pairwise Lin Similarities of the protein associated BP, MF, and CC. We compare ProtCLIP with three baselines (*i.e.*, ESM2-3B (Lin et al. 2023), KeAP (Zhou et al. 2023), BioBridge (Wang et al. 2024)). The metric is Spearman's $\rho$.

**Results** In Table 5 (left), ProtCLIP achieves the best performance over other baselines. In particular, ProtCLIP surpasses the vanilla ESM2-3B by a large margin, demonstrating the proposed data-efficient and function-informed multimodal learning is generally beneficial to the unimodal PLM.

### 3.6 Protein-Protein Interaction Prediction

**Configurations** Protein-protein interaction (PPI) prediction seeks to classify 7 interaction types of a pair of proteins. Following (Zhang et al. 2022), we extract the protein embeddings with ProtCLIP and baselines, which serve as the input for a graph neural network model to be trained on the PPI network. The baselines remain the same as mentioned in Section 3.5. Additionally, F1 score is reported on SHS27K (Chen et al. 2019), SHS148K (Chen et al. 2019) and STRING (Lv et al. 2021) datasets for evaluation.

**Results** Table 5 (right) presents average results on three benchmarks. ProtCLIP performs the best and exceeds the prior state-of-the-art BioBridge owing to its pre-training on the enormous dataset *ProtAnno-D* with the property-driven sampling strategy.

### 3.7 Ablation Study

We conduct extensive ablation experiments from multiple aspects. Unless otherwise specified, ESM-2-150M serves as the protein encoder and we evaluate on three downstream benchmarks from different types in ablation experiments.

**Ablation study on Pre-training Data** As seen in Section 2.1, we curate a new dataset *ProtAnno* with a property-driven sampling strategy. Table 6 displays comparison of different pre-training data organization. Obviously, single dataset pre-training and pretrain+finetune (first pretrained on machine-annotated data, then fine-tuned on manually-reviewed data) are inferior to the model pre-trained on

| Pre-training data | Sub | EC | | Prot2MF |
|---|---|---|---|---|
| | Acc % | AUPR | $F_{max}$ | MRR |
| *ProtAnno-S* | 72.41 | 0.216 | 0.282 | 0.246 |
| *ProtAnno-D* | 73.72 | 0.282 | 0.309 | 0.256 |
| Pretrain+finetune | 74.98 | 0.312 | 0.404 | 0.283 |
| Our sampling strategy | **75.77** | **0.384** | **0.441** | **0.299** |

Table 6: Analysis on pre-training data. Pretrain+fintune: first pretrained on low accurate data, then fine-tuned on high accurate data. Property-driven sampling strategy: pretrained on *ProtAnno-D* with the proposed sampling strategy.

| Config | Sub | EC | | Prot2MF |
|---|---|---|---|---|
| | Acc % | AUPR | $F_{max}$ | MRR |
| w/o $\boldsymbol{L}_{\text{BSR}}$ | 76.09 | 0.189 | 0.254 | 0.282 |
| w/o $\boldsymbol{L}_{\text{PDA}}$ | 73.64 | 0.136 | 0.227 | 0.210 |
| Full loss | **76.52** | **0.204** | **0.320** | **0.312** |

Table 7: Ablation study on pre-training objectives.

*ProtAnno-D* with the proposed sampling strategy. Such phenomenon demonstrates that low-quality data still holds potential value if subjected to elaborate processing and sampling, and *ProtAnno* strikes a good balance between data quality and data quantity.

**Ablation Study on Pre-training Objectives** Table 7 reports results with full or partial pre-training objectives. We can observe that both PDA and BSR are essential for injecting fine-grained information, and the absence of PDA leads to a more significant drop compared to the lack of BSR. Such results signal the competence of our function-informed paradigm for protein-biotext multi-modal learning.

**Ablation Study on Loss Weights** In Figure 5, different values of loss weights $\lambda_1$ yield different ablation results on two location classification benchmarks. Due to evident advantages, the ultimate weights are $\lambda_1 = 0.7$ and thus $\lambda_2 = 1 - \lambda_1 = 0.3$.

# 4 Related Work

## 4.1 Multi-modal Image-Text Pre-training

In an effort to overcome the limitations of single-modality learning (Zhou et al. 2024), multi-modal image-text pre-training has been introduced to learn and align visual and textual representations by pre-training the model on large-scale image-text pairs. There are numerous representative methods, such as CLIP (Radford et al. 2021), LLaVA (Liu et al. 2023a), LLaVA-Med (Li et al. 2023), BLIP families (Junnan et al. 2022, 2023), etc. Despite their impressive performance, previous methods have only learned coarse-grained representations. Motivated by this, many recent works (Ioana et al. 2024; Fuying et al. 2022; Pujin et al. 2023; Chaoyi et al. 2023; Yao et al. 2021) propose fine-grained losses or techniques to focus on localized details. However, most of them are specifically tailored for image-text alignment, and cannot seamlessly be applied to multi-modal protein-biotext pre-training.

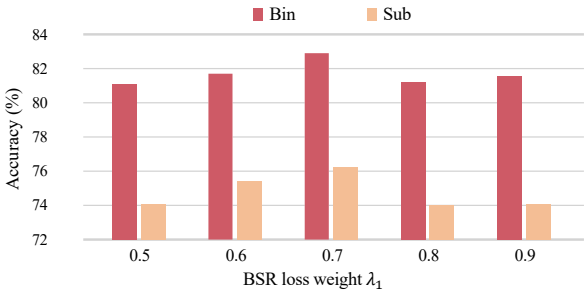


Figure 5: Ablation study on loss weights.

## 4.2 Multi-modal Protein-Biotext Pre-training

Recently, models that jointly pre-train protein sequences and biotext descriptions have gradually drawing the attention of researchers. OntoProtein (Zhang et al. 2022) first incorporates knowledge graphs to enhance protein representation with external biological descriptions. Chroma (Ingraham et al. 2023) conducts text-guided protein backbone editing towards desired properties and functions. Meanwhile, ProtDT (Liu et al. 2023b) is a newly proposed multi-modal framework that aligns the representations of proteins and biotexts for protein design. ProtST (Xu et al. 2023) has shown a tremendous performance on exploiting biomedical function annotations to enhance protein sequence understanding. Additionally, a novel multi-modal framework for the accurate prediction of protein functional descriptions in free text format is proposed by (Abdine et al. 2024). BioBridge (Wang et al. 2024) introduces a bridge module to learn transformations between protein, molecule and biotext foundation models. Nevertheless, existing works of protein-biotext alignment primarily exploit the global alignment objective proposed by CLIP (Radford et al. 2021), without utilizing protein specific functional mechanism to fully facilitate fine-grained understanding of protein and biotext.

# 5 Conclusion

This paper has accomplished data-efficient and function-informed multi-modal learning of proteins and biotexts. We build the *ProtAnno* dataset with large-scale aligned protein sequences and functional descriptions. The property-driven sampling strategy is introduced to strike a balance between data quality and data quantity for pre-training, thereby facilitating the effective harnessing of large-scale noisy data. Inspired by the intricate mechanisms of protein functionality, we novelly adopt a function-informed pre-training paradigm with newly proposed segment-wise objectives to explicitly model protein static and dynamic segments. Such paradigm seamlessly integrates multi-modality information from coarse-grained to fine-grained levels, culminating in the holistic function-centric protein representation. We also identified that ProtCLIP achieves the new state-of-the-art results on 22 protein downstream benchmarks. In the future, we envision that ProtCLIP has the potential to serve as the protein multi-modality foundation model to promote controllable protein discovery and optimization in real-world scenarios.

## Acknowledgments

This research was partially supported by National Natural Science Foundation of China under grants No. 12326612, Zhejiang Key R&D Program of China under grant No. 2023C03053, the Opening Foundation of the State Key Laboratory of Transvascular Implantation Devices, grant No. SKLTID2024003, and Alibaba Research Intern Program.

## References

Abdine, H.; Chatzianastasis, M.; Bouyioukos, C.; and Vazirgiannis, M. 2024. Prot2Text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10757–10765.

Almagro Armenteros, J. J.; Sønderby, C. K.; Sønderby, S. K.; Nielsen, H.; and Winther, O. 2017. Deeploc: prediction of protein subcellular localization using deep learning. *Nature Communications*, 33(21): 3387–3395.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bairoch, A.; and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1): 45–48.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. MixMatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*.

Chaoyi, W.; Xiaoman, Z.; Ya, Z.; Yanfeng, W.; and Weidi, X. 2023. MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis. In *International Conference on Computer Vision*, 21315–21326.

Chen, M.; Ju, C. J.-T.; Zhou, G.; Chen, X.; Zhang, T.; Chang, K.-W.; Zaniolo, C.; and Wang, W. 2019. Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics*, 35(14): i305–i314.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Computer Vision and Pattern Recognition*, 24185–24198.

Consortium, U. 2019. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1): D506–D515.

Dallago, C.; Mou, J.; Johnston, K. E.; Wittmann, B. J.; Bhattacharya, N.; Goldman, S.; Madani, A.; and Yang, K. K. 2021. FLIP: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; and Rost, B. 2022. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7112–7127.

Fuying, W.; Yuyin, Z.; Shujun, W.; Varut, V.; and Lequan, Y. 2022. Multi-Granularity Cross-modal Alignment for Generalized Medical Visual Representation Learning. In *Advances in Neural Information Processing Systems*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Computer Vision and Pattern Recognition*, 15180–15190.

Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; et al. 2021. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1): 1–14.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1): 1–23.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Ingraham, J. B.; Baranov, M.; Costello, Z.; Barber, K. W.; Wang, W.; Ismail, A.; Frappier, V.; Lord, D. M.; Ng-Thow-Hing, C.; Van Vlack, E. R.; et al. 2023. Illuminating protein space with a programmable generative model. *Nature*, 623(7989): 1070–1078.

Ioana, B.; Anastasija, I.; Matthias, B.; Goker, E.; Matko, B.; Christos, K.; Alexey A., G.; Matthias, M.; Charles, B.; Razvan, P.; and Jovana, M. 2024. Improving fine-grained understanding in image-text pre-training. In *International Conference on Machine Learning*.

Junnan, L.; Dongxu, L.; Caiming, X.; and Steven, H. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*.

Junnan, L.; Dongxu, L.; Silvio, S.; and Steven, H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Advances in Neural Information Processing Systems*.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*.

Liu, S.; Li, Y.; Li, Z.; Gitter, A.; Zhu, Y.; Lu, J.; Xu, Z.; Nie, W.; Ramanathan, A.; Xiao, C.; Tang, J.; Guo, H.; and Anandkumar, A. 2023b. A Text-guided Protein Design Framework. *arXiv preprint arXiv:2302.04611*.

Lv, G.; Hu, Z.; Bi, Y.; and Zhang, S. 2021. Learning unknown from correlations: graph neural network for

inter-novel-protein interaction prediction. *arXiv preprint arXiv:2105.06709*.

Pujin, C.; Li, L.; Junyan, L.; and Yijin, H. 2023. PRIOR: Prototype Representation Joint Learning from Medical Images and Reports. In *International Conference on Computer Vision*, 21304–21314.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, P.; Canny, J.; Abbeel, P.; and Song, Y. 2019. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; and Fergus, R. 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.

Shanehsazzadeh, A.; Belanger, D.; and Dohan, D. 2020. Is Transfer Learning Necessary for Protein Landscape Prediction? *arXiv preprint arXiv:2011.03443*.

Su, B.; Du, D.; Yang, Z.; Zhou, Y.; Li, J.; Rao, A.; Sun, H.; Lu, Z.; and Wen, J. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, E.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*.

Unsal, S.; Atas, H.; Albayrak, M.; Turhan, K.; Acar, A. C.; and Dogan, T. 2022. Learning functional properties of proteins with language models. *Nature Communications*, 4(3): 227–245.

Wang, Z.; Wang, Z.; Srinivasan, B.; Ioannidis, V. N.; Rangwala, H.; and Anubhai, R. 2024. BioBridge: Bridging Biomedical Foundation Models via Knowledge Graphs. In *International Conference on Learning Representations*.

Wu, K. E.; Chang, H.; and Zou, J. 2024. ProteinCLIP: enhancing protein language models with natural language. *bioRxiv*.

Xu, M.; Yuan, X.; Miret, S.; and Tang, J. 2023. ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts. In *International Conference on Machine Learning*.

Xu, M.; Zhang, Z.; Lu, J.; Zhu, Z.; Zhang, Y.; Chang, M.; Liu, R.; and Tang, J. 2022. PEER: A comprehensive and multi-task benchmark for protein sequence understanding. In *Advances in Neural Information Processing Systems*.

Yang, B.; tau Yih, S. W.; He, X.; Gao, J.; and Deng, L. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations*.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training. *arXiv preprint arXiv:2111.07783*.

Yin, M.; Zhou, H.; Zhu, Y.; Lin, M.; Wu, Y.; Wu, J.; Xu, H.; Hsieh, C.-Y.; Hou, T.; Chen, J.; and Wu, J. 2024. Multi-Modal CLIP-Informed Protein Editing. *Health Data Science*, 4: 0211.

Zhang, N.; Bi, Z.; Liang, X.; Cheng, S.; Hong, H.; Deng, S.; Zhang, Q.; Lian, J.; and Chen, H. 2022. OntoProtein: Protein Pretraining With Gene Ontology Embedding. In *International Conference on Learning Representations*.

Zhang, X.; Wu, C.; Zhang, Y.; WeidiXie; and Wang, Y. 2023. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1): 4542.

Zhou, H.; Yin, M.; Chen, J.; Chen, D.; and Wu, J. 2024. Group-On: Boosting One-Shot Segmentation with Supportive Query. *arXiv preprint arXiv:2404.11871*.

Zhou, H.-Y.; Fu, Y.; Zhang, Z.; Cheng, B.; and Yu, Y. 2023. Protein representation learning via knowledge enhanced primary structure reasoning. In *International Conference on Learning Representations*.