ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts

Minghao Xu [*] [†] [1 2] Xinyu Yuan [*] [1 2] Santiago Miret [3] Jian Tang [1 4 5]

Abstract

Current protein language models (PLMs) learn protein representations mainly based on their sequences, thereby well capturing co-evolutionary information, but they are unable to explicitly acquire protein functions, which is the end goal of protein representation learning. Fortunately, for many proteins, their textual property descriptions are available, where their various functions are also described. Motivated by this fact, we first build the **ProtDescribe** dataset to augment protein sequences with text descriptions of their functions and other important properties. Based on this dataset, we propose the **ProtST** framework to enhance **Prot**ein **S**equence pre-training and understanding by biomedical **T**exts. During pre-training, we design three types of tasks, *i.e.*, unimodal mask prediction, multimodal representation alignment and multimodal mask prediction, to enhance a PLM with protein property information with different granularities and, at the same time, preserve the PLM's original representation power. On downstream tasks, ProtST enables both supervised learning and zero-shot prediction. We verify the superiority of ProtST-induced PLMs over previous ones on diverse representation learning benchmarks. Under the zero-shot setting, we show the effectiveness of ProtST on zero-shot protein classification, and ProtST also enables functional protein retrieval from a large-scale database without any function annotation. Source code and model weights are available at `https://github.com/DeepGraphLearning/ProtST`.

1. Introduction

Proteins serve as the mainstay governing diverse biological processes and life itself, inducing important applications in drug discovery (Teague, 2003) and healthcare (Organization & University, 2007). Recent studies have proven the great promise of machine learning methods in predicting protein structures (Jumper et al., 2021; Baek et al., 2021) and functionality (Meier et al., 2021; Gligorijević et al., 2021). Among these methods, protein language models (PLMs) (Elnaggar et al., 2020; Rives et al., 2021; Lin et al., 2022) pre-trained on large-scale protein sequence corpus succeed in acquiring powerful protein representations, which boost protein structure and function prediction (Xu et al., 2022b).

Most existing PLMs (Elnaggar et al., 2020; Lu et al., 2020; Rives et al., 2021; Lin et al., 2022) learn protein representations based only on their sequences, which can well capture co-evolutionary information but cannot explicitly acquire protein functions and other important properties like their subcellular locations. Acquiring such function and property information is actually the end goal of protein representation learning. Fortunately, for many proteins, we can get access to their textual property descriptions in which their diverse functions are also described. This fact motivates us to study protein sequence representation learning enriched with diverse protein properties described by biomedical texts.

To our best knowledge, OntoProtein (Zhang et al., 2022a) is the only existing PLM that explicitly captures protein properties. However, it learns a closed set of properties over a *fixed biological knowledge graph* and thus can hardly generalize to unknown properties of new proteins. In comparison, by modeling *textual protein property descriptions*, we can flexibly model the generalization from known properties to unknown ones based on the semantic correlation of their text descriptions, as shown by our zero-shot experiments (Secs. 4.3 and 4.4).

To attain biomedical-text-enhanced protein sequence representation learning, we first build the **ProtDescribe** dataset, a paired dataset of protein sequences and textual property descriptions. We resort to the Swiss-Prot database (Bairoch & Apweiler, 2000) for high-quality protein annotations and construct each protein's property description with the se-

[*]Equal technical contribution. [†]Project lead. [1]Mila - Québec AI Institute [2]Université de Montréal [3]Intel Labs [4]HEC Montréal [5]CIFAR AI Research Chair. Correspondence to: Minghao Xu <minghao.xu@mila.quebec>, Santiago Miret <santiago.miret@intel.com>, Jian Tang <jian.tang@hec.ca>.

Proceedings of the 40^th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

lected annotations of it. ProtDescribe incorporates the information of protein names, protein functions, subcellular locations and protein families, and these properties are described by biomedical texts with rich expressions.

Based on this dataset, we propose the **ProtST** framework to enhance protein sequence pre-training and understanding by biomedical texts. During ProtST pre-training, to preserve the beneficial representation power of a conventional PLM on capturing co-evolutionary information, we adopt the **Unimodal Mask Prediction** task for masked protein modeling. On such basis, two multimodal pre-training tasks are designed to inject different granularities of pertinent protein property information into a PLM: **Multimodal Representation Alignment** injects integrated and general property information into the PLM, in which a biomedical language model is used to extract structured text representations of different property descriptions, and protein sequence representations are aligned to the corresponding text representations; **Multimodal Mask Prediction** models the fine-grained dependencies between residues in a protein sequence and property-descriptive words in its property description, in which a fusion module is employed to derive multimodal representations of residues and words, and, based on these fused multimodal representations, masked residues and words are predicted. For downstream applications, ProtST can conduct supervised learning with only the PLM and can also perform zero-shot prediction based on the aligned representation space of protein sequences and text descriptions.

We investigate the PLMs trained under ProtST by representation learning and zero-shot prediction. For representation learning, we verify their superior performance over previous masked language modeling and knowledge-enhanced PLMs on 11 standard benchmarks for protein localization prediction, fitness landscape prediction and protein function annotation (Sec. 4.2). For zero-shot protein classification, ProtST-induced zero-shot classifiers show better data efficiency against various few-shot classifiers (Sec. 4.3.2), and are proven to be able to enhance the performance of supervised learning models via ensemble (Sec. 4.3.3). For zero-shot text-to-protein retrieval, we verify the effectiveness of ProtST on retrieving functional proteins from a large-scale database without any function annotation (Sec. 4.4).

2. Preliminaries

2.1. Problem Definition

In the pre-training phase, we study the problem of learning informative protein sequence representations guided by the proteins' associated biomedical text descriptions. In this problem, a protein $P = (S, T)$ is represented by an amino acid sequence $S = [s_1, s_2, \cdots, s_n]$ with n amino acids (*a.k.a.*, residues) and a text description $T = [t_1, t_2, \cdots, t_m]$ with m word tokens. Given a pre-training dataset with N proteins $\mathcal{P} = \{P_1, P_2, \cdots, P_N\}$, our goal is to extract effective protein representations by fully utilizing the information from their sequences and descriptions. The extracted protein representations are expected to boost various downstream tasks by supervised learning or zero-shot prediction.

2.2. Protein Language Models

Protein language models (PLMs) (Elnaggar et al., 2020; Rives et al., 2021; Meier et al., 2021; Lin et al., 2022) pre-trained on large-scale protein sequence corpus have shown impressive results on protein function (Meier et al., 2021) and structure (Lin et al., 2022) prediction. PLMs are commonly trained by masked protein modeling, in which partial residues are masked at input and predicted based on the context. In this work, we select three state-of-the-art PLMs, ProtBert (Elnaggar et al., 2020), ESM-1b (Rives et al., 2021) and ESM-2 (Lin et al., 2022), as baselines and seek to enhance their representation power by modeling biomedical texts at the same time as protein sequence modeling.

2.3. Biomedical Language Models

Compared to the texts from general domains like newswire and Web, biomedical texts differ a lot in terms of vocabulary and expressions. To tackle such differences, language models specific to the biomedical domain (Beltagy et al., 2019; Lee et al., 2020; Gu et al., 2021) are actively studied. In this work, we employ a performant biomedical language model, PubMedBERT (Gu et al., 2021), to represent the biomedical text descriptions of proteins.

3. Method

In this section, we first motivate the proposed ProtST framework and present its general picture in Sec. 3.1, and then elucidate the design of pre-training tasks in Sec. 3.2, followed by discussing the connections with and advantages over previous works in Sec. 3.3.

3.1. Motivation and Overview

Motivation: Existing PLMs (Elnaggar et al., 2020; Lu et al., 2020; Rives et al., 2021; Lin et al., 2022) learn protein representations primarily based on their sequences, which can well capture co-evolutionary information but cannot explicitly acquire various protein properties like protein functions and subcellular locations. By acquiring such property information, the effectiveness of a PLM can be further improved, considering that the protein properties studied in pre-training and downstream tasks can correlate with each other (Bhardwaj & Lu, 2005).

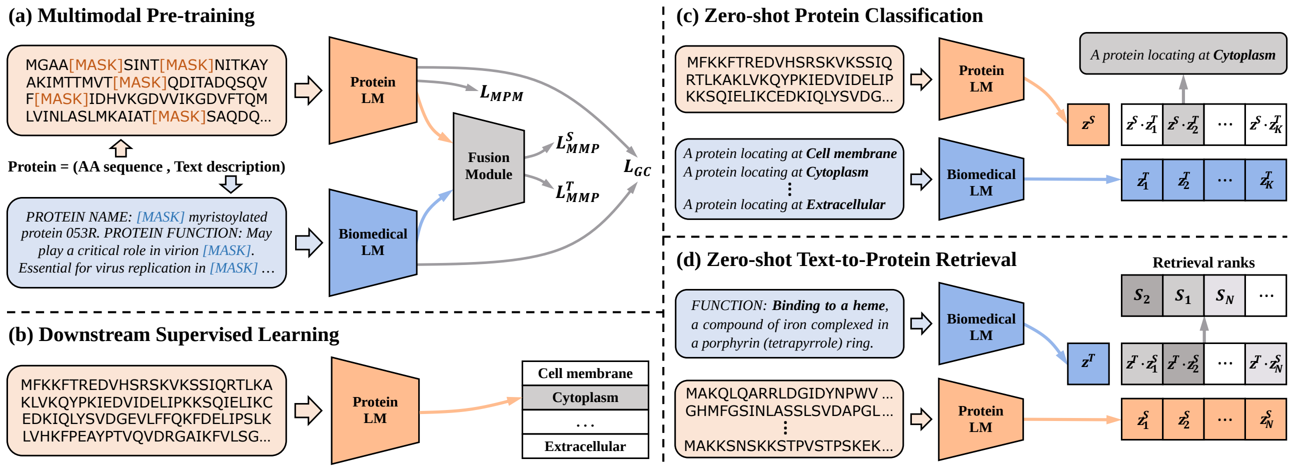

(a) Multimodal Pre-training
MGAA[MASK]SINT[MASK]NITKAY AKIMTTMVT[MASK]QDITADQSQV F[MASK]IDHVKGDVVIKGDVFTQM LVINLASLMKAIAT[MASK]SAQDQ…
Protein = (AA sequence , Text description)
PROTEIN NAME: [MASK] myristoylated protein 053R. PROTEIN FUNCTION: May play a critical role in virion [MASK]. Essential for virus replication in [MASK] …
Protein LM
Biomedical LM
Fusion Module
L_{MPM}
L^S_{MMP}
L^T_{MMP}
L_{GC}
(b) Downstream Supervised Learning
MFKKFTREDVHSRSKVKSSIQRTLKA KLVKQYPKIEDVIDELIPKKSQIELIKC EDKIQLYSVDGEVLFFQKFDELIPSLK LVHKFPEAYPTVQVDRGAIKFVLSG…
Protein LM
Cell membrane
Cytoplasm
…
Extracellular
(c) Zero-shot Protein Classification
MFKKFTREDVHSRSKVKSSIQ RTLKAKLVKQYPKIEDVIDELIP KKSQIELIKCEDKIQLYSVDG…
Protein LM
A protein locating at Cytoplasm
z^S
$z^S \cdot z^T_1$
$z^S \cdot z^T_2$
…
$z^S \cdot z^T_K$
A protein locating at Cell membrane
A protein locating at Cytoplasm
A protein locating at Extracellular
Biomedical LM
z^T_1
z^T_2
…
z^T_K
(d) Zero-shot Text-to-Protein Retrieval
FUNCTION: Binding to a heme, a compound of iron complexed in a porphyrin (tetrapyrrole) ring.
Biomedical LM
Retrieval ranks
S_2
S_1
S_N
…
z^T
$z^T \cdot z^S_1$
$z^T \cdot z^S_2$
…
$z^T \cdot z^S_N$
MAKQLQARRLDGIDYNPWV … GHMFGSINLASSLSVDAPGL… MAKKSNSKKSTPVSTPSKEK…
Protein LM
z^S_1
z^S_2
…
z^S_N

Figure 1: **Graphical illustration of ProtST framework**. (a) A protein language model (PLM) is first pre-trained along with a biomedical language model (BLM) and a fusion module to jointly model protein sequences and biomedical texts. (b) After this multi-modal pre-training, the PLM can be used individually for supervised learning on downstream tasks. (c) The couple of pre-trained PLM and BLM can perform zero-shot protein classification using only label descriptions. (d) The paired PLM and BLM can also retrieve functional proteins from a large-scale database without any function annotation.

To gain such improvement, we curate the ProtDescribe dataset that augments protein sequences with text descriptions of their diverse properties (see Sec. 4.1 for details). By injecting such property information into protein sequence representations, we aim at (1) a PLM that is more effective than previous ones on various downstream tasks under supervised learning, and (2) it can further enable zero-shot prediction through the generalization of text descriptions between known protein properties and unknown ones.

ProtST Framework: To attain these goals, we first perform multi-modal pre-training of sequences and texts and then apply the pre-trained model to three types of downstream applications (framework overview is shown in Fig. 1):

- **Multimodal Pre-training:** Given the ProtDescribe dataset, we train a PLM together with a biomedical language model (BLM) and a fusion module to model the paired protein sequences and text descriptions. We consider three kinds of pre-training tasks, *i.e.*, unimodal mask prediction, multimodal representation alignment and multimodal mask prediction, to capture the protein property information with different granularities and also preserve the PLM's original representation power.
- **Downstream Supervised Learning:** After such pre-training, the PLM is enriched by the useful property information within biomedical texts. For downstream tasks with labeled proteins, we can employ the PLM individually to solve the tasks by supervised learning.
- **Zero-shot Protein Classification:** When a protein classification task occurs without any labeled data, ProtST enables zero-shot classification. Specifically, the classification result can be determined by the representation similarity comparison between the query protein and all labels, thanks to the aligned representation space of protein sequences and label descriptions.
- **Zero-shot Text-to-Protein Retrieval:** Based on the aligned representation space, ProtST also allows us to retrieve functional proteins from a large-scale database by using only the text descriptions of protein functions, in which no function annotation is required.

3.2. Pre-training Tasks: Joint Modeling of Protein Sequences and Biomedical Texts

During ProtST pre-training, we aim to learn informative protein sequence representations guided by biomedical texts. To start this process with decent representations of protein sequences and biomedical texts, we use pre-trained PLM (*i.e.*, ProtBert (Elnaggar et al., 2020), ESM-1b (Rives et al., 2021) or ESM-2 (Lin et al., 2022)) and pre-trained BLM (*i.e.*, PubMedBERT (Gu et al., 2021)) for initialization. During training, we tune the parameters of PLM and freeze those of BLM, since the pre-trained BLM is sufficient for extracting semantically meaningful representations from biomedical texts, and it is computationally expensive to tune both PLM and BLM simultaneously. ProtST involves the following pre-training tasks for representation learning.

Unimodal Mask Prediction: The PLM for initialization is pre-trained by masked protein modeling (MPM), *i.e.*, predicting masked residues based on the protein sequence context. This task can capture co-evolutionary information by modeling residue type dependency. To preserve such unimodal information when injecting the cross-modality information from biomedical texts, we keep an MPM loss function $\mathcal{L}_{\mathrm{MPM}}$ for ProtST pre-training. Specifically, for each protein sequence, we randomly mask 15% residue tokens and predict each masked token based on its contextualized representation extracted by the PLM, where $\mathcal{L}_{\mathrm{MPM}}$ is formulated as a cross-entropy loss to measure the cost.

Multimodal Representation Alignment: The biomedical text representations learned by a pre-trained BLM can well reflect the semantics of the texts (Jin et al., 2019; Gu et al., 2021). Therefore, when given protein property descriptions, the BLM can extract semantically meaningful text representations of proteins. Thanks to this capability, by aligning protein sequence representations to their associated text representations, we can naturally inject protein property information into sequence representations.

To realize such alignment, we perform contrastive learning between protein sequences and their text descriptions. Given a batch of M proteins $\{P_i = (S_i, T_i)\}_{i=1}^M$, we use the PLM to extract protein sequence representations $\{z_i^S\}_{i=1}^M$ and the BLM to derive text description representations $\{z_i^T\}_{i=1}^M$. A standard InfoNCE loss (Oord et al., 2018) $\mathcal{L}_{\mathrm{GC}}$ is defined to maximize the representation similarity between corresponding sequences and texts and minimize the similarity between negative pairs:

$$\begin{aligned}\mathcal{L}_{\mathrm{GC}} = -\frac{1}{2M}\sum_{i=1}^{M}\Bigg(&\log\frac{\exp(z_i^S \cdot z_i^T/\tau)}{\sum_{j=1}^{M}\exp(z_i^S \cdot z_j^T/\tau)}\\ &+\log\frac{\exp(z_i^S \cdot z_i^T/\tau)}{\sum_{j=1}^{M}\exp(z_j^S \cdot z_i^T/\tau)}\Bigg),\end{aligned} \tag{1}$$

where, under multi-GPU data parallelism, we gather whole-batch samples separated on different GPUs to form negative pairs and thus term the loss $\mathcal{L}_{\mathrm{GC}}$ as a *global contrastive (GC) loss* following the convention (Singh et al., 2022), and τ denotes a learnable temperature parameter.

Multimodal Mask Prediction: Although the general dependency between the whole protein sequences and full text descriptions can be well modeled by $\mathcal{L}_{\mathrm{GC}}$, $\mathcal{L}_{\mathrm{GC}}$ alone does not capture the dependency between the residues in a protein sequence and the words in its text description. Such fine-grained cross-modality interdependency is actually ubiquitous. For example, *a soluble protein* (descriptive words) always co-occurs with charged and polar surface residues (Capaldi & Vanderkooi, 1972); *high thermostability* (descriptive words) and high amounts of hydrophobic residues are correlated with each other (Kumar et al., 2000), *etc.* To capture such interdependency, we propose a novel pre-training task that encourages the model to recover the corrupted protein sequence (or text description) based on the information from both modalities.

Specifically, given a protein sequence $S = [s_1, s_2, \cdots, s_n]$ and its corresponding text description $T = [t_1, t_2, \cdots, t_m]$, we first randomly mask 15% residues in the protein sequence and 15% words in the text description. Upon the corrupted inputs, we employ the PLM to extract residue representations $Z^S = [z_1^s, z_2^s, \cdots, z_n^s]$ and utilize the BLM to extract word representations $Z^T = [z_1^t, z_2^t, \cdots, z_m^t]$. A **fusion module** with both self- and cross-attention is then used to model the interdependency between residues and words, in which each residue and word updates its representation by attending to all the tokens along both protein sequence and text description (we state the detailed architecture in Appendix A). The fusion module produces the fused residue representations $\tilde{Z}^S = [\tilde{z}_1^s, \tilde{z}_2^s, \cdots, \tilde{z}_n^s]$ and the fused word representations $\tilde{Z}^T = [\tilde{z}_1^t, \tilde{z}_2^t, \cdots, \tilde{z}_m^t]$, in which each residue/word representation combines the information from both modalities. Based on $\tilde{Z}^S$ and $\tilde{Z}^T$, we perform *multimodal mask prediction (MMP)* to recover masked residues and words, where a cross-entropy loss $\mathcal{L}_{\mathrm{MMP}}^S$ measures the cost on protein sequence, and another cross-entropy loss $\mathcal{L}_{\mathrm{MMP}}^T$ measures the cost on text description, inducing the overall MMP loss $\mathcal{L}_{\mathrm{MMP}} = \mathcal{L}_{\mathrm{MMP}}^S + \mathcal{L}_{\mathrm{MMP}}^T$.

Overall Pre-training Objective: During the pre-training process, we seek to minimize the loss functions of all pre-training tasks simultaneously:

$$\min_{\theta}\ \mathcal{L}_{\mathrm{MPM}} + \mathcal{L}_{\mathrm{GC}} + \mathcal{L}_{\mathrm{MMP}}, \tag{2}$$

where θ denotes all learnable parameters including those of the PLM, the fusion module and all projection/prediction heads. We state the detailed architectures of these modules in Appendix A.

3.3. Discussion

Now we discuss the connections of our method with previous works and emphasize its advantages.

Advantages over Self-Supervised PLMs: Previous self-supervised PLMs (Elnaggar et al., 2020; Rives et al., 2021; Lin et al., 2022) and the proposed ProtST-induced ones can both capture co-evolutionary information hidden in protein sequences by masked protein modeling. On this basis, ProtST-induced PLMs further utilize the supervision from textual protein property descriptions, and they are guided to acquire whole-protein properties by multimodal representation alignment and acquire residue-level properties by multimodal mask prediction.

Advantages over OntoProtein (Zhang et al., 2022a): Similar to our approach, OntoProtein also seeks to enhance a self-supervised PLM by involving protein property information. In comparison, ProtST could be more effective mainly in two aspects. (1) **Diversity of considered properties:** OntoProtein retrieves Gene Ontology terms (Zhang et al., 2022a) to cover protein functions and locations; besides these two kinds of properties, ProtST additionally includes

Table 1: Statistics of the ProtDescribe dataset.

Field	Name	Function	Location	Family
#Covered samples	553,052	460,936	350,929	512,276
Coverage	100%	83.3%	63.5%	92.6%

protein names and families which are useful to indicate protein structural and functional similarity (Murzin et al., 1995). (2) **Property modeling manner:** OntoProtein learns a closed set of protein properties under the context of a *fixed biological knowledge graph*, which limits its ability to generalize to unknown properties of new proteins, while ProtST can flexibly model such generalization based on the semantic correlation of text descriptions between known and unknown properties, leading to decent zero-shot prediction capability (studied in Secs. 4.3 and 4.4).

4. Experiments

4.1. Pre-training Setups

Pre-training Dataset: To inject protein property information into PLMs, we build the ProtDescribe dataset with 553,052 aligned pairs of protein sequence and property description. Specifically, we employ the Swiss-Prot (Bairoch & Apweiler, 2000) database to provide annotations of various protein properties, in which we select four property fields: (1) "*Protein Name*" gives the full protein name recommended by the UniProt consortium (Consortium, 2019); (2) "*Function*" depicts diverse functions owned by a protein; (3) "*Subcellular Location*" describes the location and topology of a mature protein in the cell; (4) "*Similarity*" provides information about the protein families that a protein belongs to. A complete property description is formed by concatenating these four fields in order, where missing fields are skipped (see Appendix B.1 for the detailed concatenation scheme and examples). Tab. 1 presents the statistics of how each field covers the whole dataset.

Protein Language Models: We seek to enhance three performant PLMs, *i.e.*, ProtBert (Elnaggar et al., 2020), ESM-1b (Rives et al., 2021) and ESM-2 (Lin et al., 2022), by tuning their weights through the proposed ProtST pre-training. We name the PLMs after this pre-training phase as **ProtST-ProtBert**, **ProtST-ESM-1b** and **ProtST-ESM-2**. For ProtBert, we employ the ProtBert-BFD version which is trained on the BFD database (Steinegger & Söding, 2018). For ESM-2, we adopt the ESM-2-650M model so as to fairly compare with ESM-1b under the same model size.

Biomedical Language Models: By default, we utilize the PubMedBERT-abs (Gu et al., 2021) trained on PubMed abstracts to extract representations of protein property descriptions. We study another model version, PubMedBERT-full trained with additional full-text articles, in Appendix E.2.

Training Configurations: An Adam optimizer (Kingma & Ba, 2014) (learning rate: 1.0×10^{-5}, weight decay: 0) is used to train the whole model for 20 epochs on 4 Tesla V100 GPUs. More settings are introduced in Appendix B.1.

4.2. Representation Learning

4.2.1. Experimental Setups

Downstream Benchmark Tasks. We adopt 11 benchmark tasks within three task types (the "*Abbr.*" below denotes the abbreviated task name in Tab. 2 and 3):

- **Protein Localization Prediction** seeks to predict the subcellular locations of proteins. We consider two such problems from DeepLoc (Almagro Armenteros et al., 2017), the subcellular localization prediction (*Abbr.*, Sub) with 10 location categories and the binary localization prediction (*Abbr.*, Bin) with 2 location categories. We follow the official dataset splits.
- **Fitness Landscape Prediction** aims to predict the effect of residue mutations on protein fitness. We employ the β-lactamase (*Abbr.*, β-lac) landscape from PEER (Xu et al., 2022b), the AAV and Thermostability (*Abbr.*, Thermo) landscapes from FLIP (Dallago et al., 2021), and the Fluorescence (*Abbr.*, Flu) and Stability (*Abbr.*, Sta) landscapes from TAPE (Rao et al., 2019). For AAV, we use the "two_vs_many" dataset splits; for Thermostability, we adopt the "human_cell" splits; we follow the only default splits on all other tasks. In Appendix C, we further show the results on ProteinGym (Notin et al., 2022).
- **Protein Function Annotation** seeks to annotate a protein with multiple functional labels. We employ two standard benchmarks proposed by DeepFRI (Gligorijević et al., 2021), *i.e.*, Enzyme Commission (EC) number prediction and Gene Ontology (GO) term prediction. The GO benchmark is split into three branches to predict molecular function (*Abbr.*, GO-MF), biological process (*Abbr.*, GO-BP) and cellular component (*Abbr.*, GO-CC). Following Zhang et al. (2022b), we use the dataset splits under 95% sequence identity cutoff for both EC and GO.

Baselines: We adopt four protein sequence encoders trained from scratch, *i.e.*, CNN (Shanehsazzadeh et al., 2020), ResNet (Rao et al., 2019), LSTM (Rao et al., 2019) and Transformer (Rao et al., 2019), as naive baselines. We focus on comparing with four performant PLMs, *i.e.*, ProtBert (Elnaggar et al., 2020), OntoProtein (Zhang et al., 2022a), ESM-1b (Rives et al., 2021) and ESM-2 (Lin et al., 2022).

Training and Evaluation: We train with an Adam optimizer for 100 epochs on localization and fitness prediction tasks and for 50 epochs on function annotation tasks. For localization and fitness prediction, all PLMs are evaluated under both fix-encoder learning and full-model tuning settings, and only full-model tuning is used for PLMs on

Table 2: Benchmark results on protein localization and fitness landscape prediction. We use three color scales of blue to denote the first, second and third best performance. *Abbr.*, Loc.: Localization; pred.: prediction; Acc: accuracy.

Model	**Loc. pred. (*Acc%*)**		**Fitness pred. (*Spearman's* ρ)**					
	Bin	**Sub**	**β-lac**	**AAV**	**Thermo**	**Flu**	**Sta**	**Mean ρ**
Protein sequence encoders trained from scratch								
CNN	82.67	58.73	0.781	0.746	0.494	**0.682**	0.637	0.668
ResNet	78.99	52.30	0.152	0.739	0.528	0.636	0.126	0.436
LSTM	88.11	62.98	0.139	0.125	0.564	0.494	0.533	0.371
Transformer	75.74	56.02	0.261	0.681	0.545	0.643	0.649	0.556
PLMs *w/* fix-encoder learning								
ProtBert	81.54	59.44	0.616	0.209	0.562	0.339	0.697	0.485
OntoProtein	84.87	68.34	0.471	0.217	0.605	0.432	0.688	0.483
ESM-1b	91.61	79.82	0.528	0.454	0.674	0.430	0.750	0.567
ESM-2	91.32	80.84	0.559	0.374	0.677	0.456	0.746	0.562
ProtST-ProtBert	92.29	78.49	0.569	0.219	0.621	0.376	0.719	0.501
ProtST-ESM-1b	**92.87**	82.00	0.578	0.460	0.680	0.523	0.766	0.601
ProtST-ESM-2	92.52	**83.39**	0.565	0.398	0.681	0.499	**0.776**	0.584
PLMs *w/* full-model tuning								
ProtBert	91.32	76.53	0.731	0.794	0.660	0.679	0.771	0.727
OntoProtein	92.47	77.59	0.757	0.791	0.662	0.630	0.731	0.714
ESM-1b	92.40	78.13	0.839	0.821	0.669	0.679	0.694	0.740
ESM-2	91.72	78.67	0.867	0.817	0.672	0.677	0.718	0.750
ProtST-ProtBert	91.78	78.71	0.863	0.804	0.673	0.679	0.745	0.753
ProtST-ESM-1b	92.35	78.73	**0.895**	**0.850**	0.681	**0.682**	0.751	**0.772**
ProtST-ESM-2	92.52	80.22	0.879	0.825	**0.682**	**0.682**	0.738	0.761

Table 3: Benchmark results on protein function annotation. We use three color scales of blue to denote the first, second and third best performance.

Model	**EC**		**GO-BP**		**GO-MF**		**GO-CC**	
	AUPR	$\mathrm{F}_{\max}$	AUPR	$\mathrm{F}_{\max}$	AUPR	$\mathrm{F}_{\max}$	AUPR	$\mathrm{F}_{\max}$
Protein sequence encoders trained from scratch								
CNN	0.540	0.545	0.165	0.244	0.380	0.354	0.261	0.387
ResNet	0.137	0.187	0.166	0.280	0.281	0.267	0.266	0.403
LSTM	0.032	0.082	0.130	0.248	0.100	0.166	0.150	0.320
Transformer	0.187	0.219	0.135	0.257	0.172	0.240	0.170	0.380
PLMs *w/* full-model tuning								
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
ESM-2	0.888	0.874	0.340	0.472	0.643	0.662	0.350	0.472
ProtST-ProtBert	0.876	0.856	0.286	0.440	0.615	0.648	0.314	0.449
ProtST-ESM-1b	0.894	**0.878**	0.328	0.480	0.644	0.661	**0.364**	**0.488**
ProtST-ESM-2	**0.898**	**0.878**	**0.342**	**0.482**	**0.647**	**0.668**	**0.364**	0.487

function annotation, since it is hard to solve the multiple binary classification problems on EC and GO with fixed protein representations. More training details are stated in Appendix B.2.

For all models on all tasks, we select the checkpoint for evaluation based on the validation set performance, and all results are reported on the seed 0. We measure the classification accuracy for localization prediction and the Spearman's ρ for fitness prediction. Following Gligorijević et al. (2021), function annotation tasks are measured by AUPR and $\mathrm{F}_{\max}$ whose detailed definitions are in Appendix B.2.

4.2.2. Experimental Results

We report the benchmark results on localization and fitness prediction in Tab. 2 and report function annotation results in Tab. 3. Based on the benchmark results, we have the following observations:

ProtST-induced PLMs clearly outperform the vanilla PLMs. It is observed that: (1) ProtST-ProtBert outperforms the vanilla ProtBert on 21 out of 24 benchmark metrics (including both fix-encoder learning and full-model tuning ones); (2) ProtST-ESM-1b surpasses the vanilla ESM-1b on 22 out of 24 benchmark metrics; (3) ProtST-ESM-2 outperforms the vanilla ESM-2 on all 24 benchmark metrics. These results demonstrate that ProtST pre-training is generally beneficial to different PLMs, which boosts their performance on diverse downstream tasks.

ProtST-ProtBert performs consistently better than OntoProtein under fair comparison. ProtST-ProtBert and OntoProtein can be fairly compared with each other, since they both adopt ProtBert as the initial PLM. ProtST-ProtBert surpasses OntoProtein on 22 out of 24 benchmark metrics, which verifies the superiority of the proposed pre-training dataset and pre-training tasks.

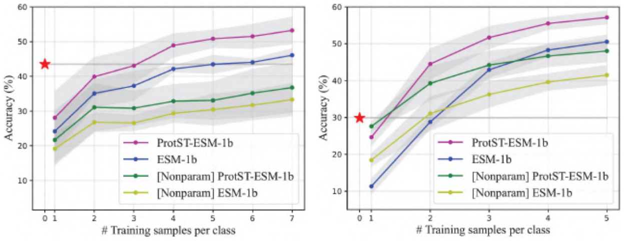
ProtST-ESM-1b performs best on fitness prediction, and ProtST-ESM-2 performs best on localization prediction and function annotation. We can observe that: (1) ProtST-ESM-1b achieves the best performance on 4 out of 6 benchmark metrics for fitness prediction; (2) ProtST-ESM-2 obtains the highest localization prediction accuracy on average, and it performs best on 7 out of 8 benchmark metrics for function annotation. We therefore recommend these two PLMs as new state-of-the-arts.

4.3. Zero-shot Protein Classification

4.3.1. Experimental Setups

Zero-shot Protein Classification based on Aligned Representation Space: A ProtST-induced PLM naturally allows zero-shot protein classification, thanks to its aligned representation space of protein sequences and text descriptions. In specific, given the sequence S of a query protein and the label descriptions $\{T_i\}_{i=1}^{K}$ of all K classes, we employ the PLM to extract protein representation z^S and use the jointly learned BLM to extract label representations $\{z_i^T\}_{i=1}^{K}$. We then derive classification logits $\{y_i\}_{i=1}^{K}$ by comparing the dot product similarity between protein and label representations: $y_i = z^S \cdot z_i^T / \tau$ $(i = 1, \cdots, K)$, which follows the formula of InfoNCE loss in Eq. (1). Softmax is performed upon these logits to derive classification probabilities.

Benchmark Tasks: In this part of experiments, we adopt two protein classification tasks as benchmarks: (1) the *subcellular localization prediction* task which is same as the one introduced in Sec. 4.2.1; (2) the *reaction classification* task proposed by Hermosilla et al. (2020) which reformulates the EC number prediction task introduced in Sec. 4.2.1 as a classification task with 384 reaction classes. We follow the official dataset splits for both tasks.



(a) Subcellular localization prediction

(b) Reaction classification

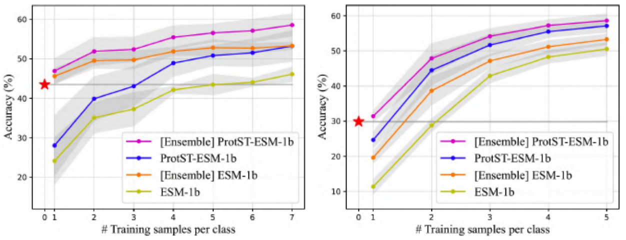
Figure 2: **Zero-shot ProtST-ESM-1b outperforms few-shot classifiers.** The horizontal line with a red star denotes the zero-shot performance of ProtST-ESM-1b. All few-shot results are averaged over seeds 0, 1, 2, 3 and 4, and gray intervals denote standard deviations.

Prompt Engineering: To extract discriminative label representations, we have tried three types of prompt templates to describe protein function/location labels. (1) *Name only*: a label is described only by the name of a function or location (*e.g.*, "Cytoplasm"); (2) *Natural language*: the name is embedded into a natural language template (*e.g.*, "A protein locating at Cytoplasm"); (3) *Pre-training template*: the name is embedded into the template used during ProtST pre-training (*e.g.*, "SUBCELLULAR LOCATION: Cytoplasm"). The pre-training template is empirically verified to be more effective than other two templates, and thus it is used across all experiments of this section. The comparisons among these templates are provided in Appendix B.3.

4.3.2. Data Efficiency of Zero-shot Classifier

Baselines: We study the data efficiency of zero-shot ProtST-ESM-1b by comparing it with n-shot classifiers ($n \geqslant 1$) which employ n training samples per class for prediction. We adopt four baselines: (1) the ProtST-ESM-1b with supervised fine-tuning, (2) the ESM-1b with supervised fine-tuning, (3) the nonparametric ProtST-ESM-1b classifier, and (4) the nonparametric ESM-1b classifier. We follow Khandelwal et al. (2019) to design the nonparametric classifiers which predict based on the relations between test sample and training samples, and they well fit the few-shot prediction setting. We elucidate such classifiers in Appendix B.3.

Results: For subcellular localization prediction (Fig. 2(a)), the zero-shot ProtST-ESM-1b matches the performance of 3-shot supervised ProtST-ESM-1b and the performance of 5-shot supervised ESM-1b, and the zero-shot classifier outperforms two 7-shot nonparametric classifiers. For reaction classification (Fig. 2(b)), the zero-shot ProtST-ESM-1b surpasses the 1-shot performance of supervised and nonparametric ProtST-ESM-1b, and it aligns the 2-shot performance of supervised and nonparametric ESM-1b. These results demonstrate the data efficiency of ProtST-induced zero-shot classifiers. In particular, they can be helpful in the downstream tasks with limited or even no labeled proteins by making educated predictions using only label descriptions.



(a) Subcellular localization prediction

(b) Reaction classification

Figure 3: **Zero-shot ProtST-ESM-1b enhances few-shot classifiers' performance via ensemble.** The horizontal line with a red star denotes the zero-shot performance of ProtST-ESM-1b. All few-shot results are averaged over seeds 0, 1, 2, 3 and 4, and gray intervals denote standard deviations.

Table 4: **Zero-shot ProtST-ESM-1b enhances full-shot classifiers' performance via ensemble.** *Abbr.*, loc.: localization; Acc: accuracy.

Model	Subcellular loc. (*Acc%*)	Reaction (*Acc%*)
ProtST-ESM-1b	82.00	86.73
[Ensemble] ProtST-ESM-1b	**82.37**	**87.14**
ESM-1b	79.82	80.54
[Ensemble] ESM-1b	**80.20**	**83.03**

4.3.3. Enhancing Supervised Learning with Zero-shot Classifier

Ensemble of Supervised Learning Model and Zero-shot Classifier: We study how zero-shot ProtST-ESM-1b can boost supervised learning models via ensemble. Specifically, we combine the classification logits produced by a supervised learning model and the zero-shot classification logits as below: $\{y_k = y_k^{\text{sup}} + \alpha\, y_k^{\text{zero}}\}_{k=1}^{K}$ (K is the number of classes), where α controls the contribution of the zero-shot classifier. Empirically, we set α as the ratio of the zero-shot classifier's validation set performance over the validation performance of the supervised learning model.

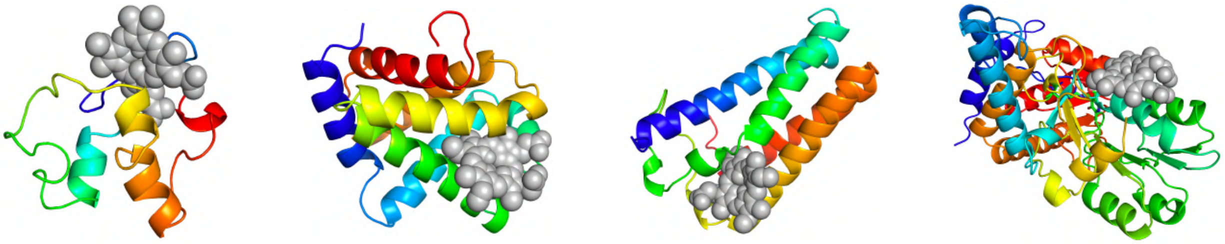
Baselines: We employ ProtST-ESM-1b and ESM-1b with supervised fine-tuning on downstream tasks as baselines. We consider fine-tuning under both the few-shot setting and the full-shot setting (*i.e.*, trained with all training samples). Based on these supervised models, we seek to utilize zero-shot ProtST-ESM-1b to enhance their performance.

Results: According to Fig. 3 and Tab. 4, we can observe that zero-shot ProtST-ESM-1b succeeds in enhancing the performance of all few-shot and full-shot baselines on both benchmarks. These results verify that ProtST-induced zero-shot classifiers are useful tools to enhance supervised learning models, which is realized by refining decision boundaries.

4.4. Zero-shot Text-to-Protein Retrieval

Zero-shot Text-to-Protein Retriever: Based on the protein-text aligned representation space, ProtST enables us to retrieve functional proteins from a large-scale database with-

Prompt **- FUNCTION: Binding to a heme, a compound composed of iron complexed in a porphyrin (tetrapyrrole) ring.**



(*1st*) 2N91-A:
- Affinity: **-7.3 (kcal/mol)**
- GO-MF label: **Bind**

(*2nd*) 1YHU-A:
- Affinity: **-7.9 (kcal/mol)**
- GO-MF label: **Bind**

(*3rd*) 5B3I-A:
- Affinity: **-8.1 (kcal/mol)**
- GO-MF label: **Bind**

(*4th*) 5VPR-A:
- Affinity: **-7.4 (kcal/mol)**
- GO-MF label: **Non-bind**

Figure 4: Zero-shot text-to-protein retrieval of heme binders based on ProtST-ESM-1b.

Table 5: Swiss-Prot *v.s.* TrEMBL on protein property coverage.

Dataset	Name	Function	Location	Family
Swiss-Prot	100%	83.3%	63.5%	92.6%
TrEMBL	100%	24.0%	51.5%	78.0%

Table 6: Swiss-Prot *v.s.* TrEMBL as pre-training data source, compared on downstream representation learning tasks. *Abbr.*, Loc.: localization prediction; Fit.: fitness prediction; Fix-enc.: fix-encoder learning; Full-m.: full-model tuning.

Dataset	Loc. (mean Acc%)		Fit. (mean ρ)	
	Fix-enc.	Full-m.	Fix-enc.	Full-m.
Swiss-Prot	**87.44**	**85.54**	**0.601**	**0.772**
TrEMBL	86.68	85.13	0.597	0.762

out any function annotation. To be specific, the PLM is first employed to extract the representations $\{z_i^S\}_{i=1}^N$ of all proteins in the database. During the retrieval process, given the text description (*i.e.*, prompt) T of a protein function, the BLM is used to extract its representation z^T, and all proteins are then ranked based on their representation similarity $\{\epsilon_i = z_i^S \cdot z^T\}_{i=1}^N$ with the prompt.

Experimental Setups: We use ProtST-ESM-1b to retrieve the Gene Ontology (GO) dataset introduced in Sec. 4.2.1. We build each prompt by adding the "FUNCTION:" prefix before the molecular function definition from GO.

Results: In Fig. 4, we visualize the top-4 retrieved candidates of heme binders. We present the text prompt, the docking result of each candidate binding with heme (AutoDock Vina (Trott & Olson, 2010) is used for docking), the binding affinity predicted by AutoDock Vina (the lower the better), and the GO molecular function labels of heme binding. We can observe that the top-3 candidates are annotated as heme binders by GO, and the 4th candidate owns decent binding affinity though annotated as non-binding (only 0.54% proteins are annotated as heme binders in the GO dataset). These results verify the effectiveness of ProtST-ESM-1b on retrieving heme binders. We provide more case studies in Appendix D. Other visualization results are in Appendix F.

Table 7: Ablation study of pre-training losses on ProtST-ESM-1b. *Abbr.*, Loc.: localization prediction; Fit.: fitness prediction; Func.: function annotation; Fix-enc.: fix-encoder learning; Full-m.: full-model tuning. Blue denotes the largest decay.

Config	Loc. (mean Acc%)		Fit. (mean ρ)		Func. (mean $\mathbf{F}_{\max}$)
	Fix-enc.	Full-m.	Fix-enc.	Full-m.	
Full loss	87.44	85.54	0.601	0.772	0.627
w/o $\mathcal{L}_{\text{MPM}}$	87.40$_{(\downarrow 0.05\%)}$	85.12$_{(\downarrow 0.49\%)}$	0.593$_{(\downarrow 1.33\%)}$	0.766$_{(\downarrow 0.78\%)}$	0.625$_{(\downarrow 0.32\%)}$
w/o $\mathcal{L}_{\text{GC}}$	86.34$_{(\downarrow \mathbf{1.26\%})}$	85.21$_{(\downarrow 0.39\%)}$	0.579$_{(\downarrow \mathbf{3.66\%})}$	0.758$_{(\downarrow 1.81\%)}$	0.613$_{(\downarrow \mathbf{2.23\%})}$
w/o $\mathcal{L}_{\text{MMP}}$	87.41$_{(\downarrow 0.03\%)}$	84.97$_{(\downarrow \mathbf{0.67\%})}$	0.588$_{(\downarrow 2.16\%)}$	0.751$_{(\downarrow \mathbf{2.72\%})}$	0.615$_{(\downarrow 1.91\%)}$

4.5. Ablation Study

Effect of Pre-training Data Source: In this project, besides Swiss-Prot, we also tried to use TrEMBL (Bairoch & Apweiler, 2000) as the data source to construct ProtDescribe. Compared to Swiss-Prot with high-quality human annotations for around 500K proteins, TrEMBL contains a larger number of over 200M annotated proteins, while the TrEMBL annotations are given by computational tools and are thus less accurate and have lower protein property coverage (as shown in Tab. 5).

The results in Tab. 6 show that the ProtST-ESM-1b pre-trained on the smaller while higher-quality Swiss-Prot-based dataset performs better. Therefore, for the multimodal pre-training of protein sequences and biomedical texts, data quality could be more important than data quantity.

Effect of Pre-training Losses: Tab. 7 reports the averaged performance of ProtST-ESM-1b by using full or partial pre-training losses (per-task results are in Appendix E.1). By removing any of three pre-training losses, performance decay occurs on all three types of tasks. Such phenomenon verifies the necessity of each ProtST pre-training loss, where $\mathcal{L}_{\text{GC}}$ and $\mathcal{L}_{\text{MMP}}$ inject different granularities of protein property information into a PLM, and $\mathcal{L}_{\text{MPM}}$ preserves the PLM's original representation power.

Effect of PLM: According to the results in Tabs. 2 and 3, we can observe that the strength of a ProtST-induced PLM correlates with the strength of its initial PLM. To be specific, the better performance of ESM-1b and ESM-2 over ProtBert is inherited by their ProtST-induced variants.

5. Related Work

Protein Representation Learning: Learning effective protein representations is of great importance for machine learning guided protein understanding. Existing works learn protein representations in two ways: (1) Sequence-based methods model protein sequences on evolutionary scale (Elnaggar et al., 2020; Rives et al., 2021; Lin et al., 2022) or on individual protein families (Bileschi et al., 2019; Meier et al., 2021; Biswas et al., 2021); (2) Structure-based methods seek to represent different levels of protein structures including residue-level structures (Gligorijević et al., 2021; Zhang et al., 2022b; Xu et al., 2022a), all-atom structures (Jing et al., 2020; Zhang et al., 2023) and protein surfaces (Gainza et al., 2020; Sverrisson et al., 2021). Our work aims to enhance protein sequence representation learning by using textual protein property descriptions.

Multimodal Representation Learning: It has been broadly studied how to learn better image (Radford et al., 2021; Singh et al., 2022), video (Luo et al., 2020; Xu et al., 2021), speech (Chung et al., 2020; Qian et al., 2021) and molecule (Edwards et al., 2021; Liu et al., 2022) representations by incorporating text supervision, while such study is lacked for proteins. OntoProtein (Zhang et al., 2022a) learns protein representations under the context of a knowledge graph; ProGen (Madani et al., 2020) incorporates protein function labels to generate functional proteins. However, these two works investigate less the effect of biomedical texts. Our work takes the initiative of enhancing protein sequence representation learning by biomedical texts.

6. Conclusions and Future Work

In this work, we propose the ProtST framework to study how textual protein property descriptions can boost protein sequence pre-training and understanding. We build the ProtDescribe dataset that aligns protein sequences with their diverse property descriptions. ProtST pre-training injects the property information with different granularities into a protein language model (PLM). The ProtST-induced PLMs are verified to be generally effective on various downstream applications including supervised learning, zero-shot protein classification and zero-shot text-to-protein retrieval.

The current ProtDescribe dataset is limited in the coverage of protein sequences and textual property descriptions, which motivates us to resort to massive biomedical articles in PubMed (Canese & Weis, 2013) for information extraction. In addition, we plan to extend the ProtDescribe dataset by incorporating protein structures and study biomedical text enhanced protein structure representation learning. Also, we will go beyond text-to-protein retrieval towards text-guided controllable protein design.

Acknowledgments

The authors would like to thank Meng Qu, Zhaocheng Zhu, Zuobai Zhang and Hesham Mostafa for their helpful discussions and comments.

This project is supported by Intel-MILA partnership program, the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant, the Canada CIFAR AI Chair Program, collaboration grants between Microsoft Research and Mila, Samsung Electronics Co., Ltd., Amazon Faculty Research Award, Tencent AI Lab Rhino-Bird Gift Fund, a NRC Collaborative R&D Project (AI4D-CORE-06) as well as the IVADO Fundamental Research Project grant PRF-2019-3583139727.

References

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

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Bairoch, A. and Apweiler, R. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 28(1):45–48, 2000.

Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Bhardwaj, N. and Lu, H. Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics*, 21(11):2730–2738, 2005.

Bileschi, M. L., Belanger, D., Bryant, D., Sanderson, T., Carter, B., Sculley, D., DePristo, M. A., and Colwell, L. J. Using deep learning to annotate the protein universe. *BioRxiv*, pp. 626507, 2019.

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. Low-n protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

Canese, K. and Weis, S. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1), 2013.

Capaldi, R. A. and Vanderkooi, G. The low polarity of many membrane proteins. *Proceedings of the National Academy of Sciences*, 69(4):930–932, 1972.

Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., and Schomburg, D. Brenda, the elixir core data resource in 2021: new developments and updates. *Nucleic Acids Research*, 49(D1):D498–D508, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chung, Y.-A., Zhu, C., and Zeng, M. Splat: Speech-language joint pre-training for spoken language understanding. *arXiv preprint arXiv:2010.02295*, 2020.

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

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

Edwards, C., Zhai, C., and Ji, H. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607, 2021.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M., and Correia, B. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

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. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):1–14, 2021.

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

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hermosilla, P., Schäfer, M., Lang, M., Fackelmann, G., Vázquez, P. P., Kozlíková, B., Krone, M., Ritschel, T., and Ropinski, T. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *arXiv preprint arXiv:2007.06252*, 2020.

Jin, Q., Dhingra, B., Cohen, W. W., and Lu, X. Probing biomedical embeddings from language models. *arXiv preprint arXiv:1904.02181*, 2019.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J., and Dror, R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kumar, S., Tsai, C.-J., and Nussinov, R. Factors enhancing protein thermostability. *Protein engineering*, 13(3):179–191, 2000.

Laine, E., Karami, Y., and Carbone, A. Gemme: a simple and fast global epistatic model predicting mutational effects. *Molecular biology and evolution*, 36(11):2604–2619, 2019.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., and Anandkumar, A. Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789*, 2022.

Lu, A. X., Zhang, H., Ghassemi, M., and Moses, A. M. Self-supervised contrastive learning of protein representations by mutual information maximization. *BioRxiv*, 2020.

Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., and Zhou, M. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

Marquet, C., Heinzinger, M., Olenyi, T., Dallago, C., Erckert, K., Bernhofer, M., Nechaev, D., and Rost, B. Embeddings from protein language models predict conservation and variant effects. *Human genetics*, 141(10):1629–1647, 2022.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995.

Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N., and Madani, A. Progen2: exploring the boundaries of protein language models. *arXiv preprint arXiv:2206.13517*, 2022.

Notin, P., Dias, M., Frazer, J., Hurtado, J. M., Gomez, A. N., Marks, D., and Gal, Y. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Organization, W. H. and University, U. N. *Protein and amino acid requirements in human nutrition*, volume 935. World Health Organization, 2007.

Qian, Y., Bianv, X., Shi, Y., Kanda, N., Shen, L., Xiao, Z., and Zeng, M. Speech-language pre-training for end-to-end spoken language understanding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7458–7462. IEEE, 2021.

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

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Shanehsazzadeh, A., Belanger, D., and Dohan, D. Is transfer learning necessary for protein landscape prediction? *arXiv preprint arXiv:2011.03443*, 2020.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Steinegger, M. and Söding, J. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1): 1–8, 2018.

Sverrisson, F., Feydy, J., Correia, B. E., and Bronstein, M. M. Fast end-to-end learning on protein surfaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15272–15281, 2021.

Teague, S. J. Implications of protein flexibility for drug discovery. *Nature reviews Drug discovery*, 2(7):527–541, 2003.

Trott, O. and Olson, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., and Feichtenhofer, C. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

Xu, M., Guo, Y., Xu, Y., Tang, J., Chen, X., and Tian, Y. Eurnet: Efficient multi-range relational modeling of spatial multi-relational data. *arXiv preprint arXiv:2211.12941*, 2022a.

Xu, M., Zhang, Z., Lu, J., Zhu, Z., Zhang, Y., Ma, C., Liu, R., and Tang, J. Peer: A comprehensive and multi-task benchmark for protein sequence understanding. *arXiv preprint arXiv:2206.02096*, 2022b.

Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., and Chen, H. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022a.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022b.

Zhang, Z., Xu, M., Lozano, A., Chenthamarakshan, V., Das, P., and Tang, J. Physics-inspired protein encoder pretraining via siamese sequence-structure diffusion trajectory prediction. *arXiv preprint arXiv:2301.12068*, 2023.

Zhu, Z., Shi, C., Zhang, Z., Liu, S., Xu, M., Yuan, X., Zhang, Y., Chen, J., Cai, H., Lu, J., et al. Torchdrug: A powerful and flexible machine learning platform for drug discovery. *arXiv preprint arXiv:2202.08320*, 2022.

A. Model Architecture for Pre-training

Fusion Module: The fusion module extracts multimodal representations from the unimodal representations of protein sequence and text description. As shown in Fig. 5, each *fusion layer* of this module receives a sequence of residue representations $Z^S = [z_1^s, z_2^s, \cdots, z_n^s] \in \mathbb{R}^{n\times d}$ and a sequence of word representations $Z^T = [z_1^t, z_2^t, \cdots, z_m^t] \in \mathbb{R}^{m\times d}$ (d denotes the hidden dimension), and the layer updates each residue/word representation by attending to all residues and all words. Specifically, two sets of projection matrices (W_q^S, W_k^S, W_v^S) and (W_q^T, W_k^T, W_v^T) are respectively used to derive the queries, keys and values for protein sequence and text description as below (each projection matrix is in $\mathbb{R}^{d\times d}$):

$$Q^S = Z^S W_q^S, \quad K^S = Z^S W_k^S, \quad V^S = Z^S W_v^S, \quad (3)$$

$$Q^T = Z^T W_q^T, \quad K^T = Z^T W_k^T, \quad V^T = Z^T W_v^T, \quad (4)$$

where $Q^S, K^S, V^S \in \mathbb{R}^{n\times d}$ are the queries, keys and values for protein sequence, and $Q^T, K^T, V^T \in \mathbb{R}^{m\times d}$ are the queries, keys and values for text description. Multi-head self- and cross-attention are then applied to update each residue and word representation as below:

$$\tilde{Z}^S = \frac{1}{2}\big(\mathrm{MHA}(Q^S, K^S, V^S) + \mathrm{MHA}(Q^S, K^T, V^T)\big), \quad (5)$$

$$\tilde{Z}^T = \frac{1}{2}\big(\mathrm{MHA}(Q^T, K^T, V^T) + \mathrm{MHA}(Q^T, K^S, V^S)\big), \quad (6)$$

where $\tilde{Z}^S \in \mathbb{R}^{n\times d}$ and $\tilde{Z}^T \in \mathbb{R}^{m\times d}$ are the updated residue and word representations, and $\mathrm{MHA}(\cdot,\cdot,\cdot)$ denotes the multi-head attention operation (Vaswani et al., 2017).

In our implementation, each fusion layer contains 8 attention heads, and we equip the fusion module with a single fusion layer so as to restrict the capacity of fusion module and facilitate the representation power of PLM. Upon the fused residue and word representations produced by the fusion module, multimodal mask prediction is performed.

Projection Head for Multimodal Representation Alignment: Following SimCLR (Chen et al., 2020), we use a two-layer MLP (with ReLU nonlinearity in between) to project the protein sequence representation extracted by the PLM, and another two-layer nonlinear MLP is employed to project the text description representation extracted by the BLM. The projected sequence and text representations are then used to compute the global contrastive loss defined in Eq. (1).

Prediction Head for Masked Protein Modeling (MPM): Based on the residue representations extracted by the PLM, we utilize a two-layer MLP (with ReLU nonlinearity in between) to predict the type of each residue token masked at input.

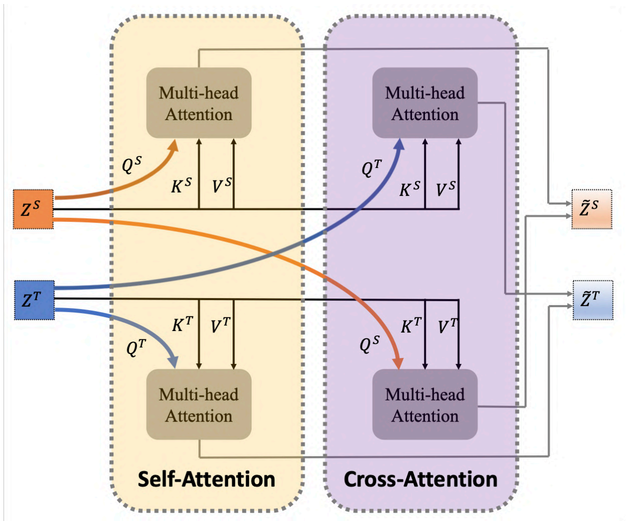


Figure 5: **Architecture of the fusion layer.** This layer fuses the protein representation and the text representation by querying over them with self-attention and cross-attention.

Prediction Head for Multimodal Mask Prediction (MMP): Upon the fused residue representations output from the fusion module, a two-layer MLP (with ReLU nonlinearity in between) is used to predict the type of each residue token masked at input protein sequence. Upon the fused word representations produced by the fusion module, another two-layer nonlinear MLP is employed to predict each word token masked at input text description.

B. More Experimental Setups

B.1. More Pre-training Setups

Pre-training Data Curation: We add prefixes to denote annotations from different fields, *i.e.*, "PROTEIN NAME" for the protein name field, "FUNCTION" for the protein function field, "SUBCELLULAR LOCATION" for the subcellular location field, and "SIMILARITY" for the protein family field. The complete protein property description is formed by concatenating all annotations of the protein in the order of (1) protein name, (2) protein function, (3) subcellular location, and (4) protein family. In Tab. 8, we present several property descriptions coupled with the Swiss-Prot entry names of their corresponding proteins.

Training Configurations: We list the training configurations of three ProtST-induced PLMs in Tab. 9. In general, an Adam optimizer with the constant learning rate of 1.0×10^{-5} is used to train the model for 20 epochs on 4 Tesla V100 GPUs, where ProtST-ProtBert adopts the batch size of 16 (4 proteins per GPU), and ProtST-ESM-1b and ProtST-ESM-2 adopt the batch size of 12 (3 proteins per GPU). Since the PLM is pre-trained, we set its learning rate as 1.0×10^{-6}, *i.e.*, one tenth of other modules. The weights of PubMed-

Table 8: Examples of property descriptions in the ProtDescribe dataset. We index each description with the Swiss-Prot entry name of its corresponding protein.

Entry name	Description
14336_ORYSJ	PROTEIN NAME: 14-3-3-like protein GF14-F. FUNCTION: Is associated with a DNA binding complex that binds to the G box, a well-characterized cis-acting DNA regulatory element found in plant genes. SUBCELLULAR LOCATION: Cytoplasm. Nucleus. SIMILARITY: Belongs to the 14-3-3 family.
053R_FRG3G	PROTEIN NAME: Putative myristoylated protein 053R. FUNCTION: May play a critical role in virion formation. Essential for virus replication in vitro. SUBCELLULAR LOCATION: Host membrane; Multi-pass membrane protein.
1A16_ORYSJ	PROTEIN NAME: 1-aminocyclopropane-1-carboxylate synthase 6. FUNCTION: Catalyzes the formation of 1-aminocyclopropane-1-carboxylate, a direct precursor of ethylene in higher plants (By similarity). Required for the regulation of starch grain size in endosperm. SUBCELLULAR LOCATION: Plastid, amyloplast membrane. Note=Localizes to the amyloplast membrane surrounding starch grains in endosperm, pollen, and pericarp. SIMILARITY: Belongs to the class-I pyridoxal-phosphate-dependent aminotransferase family.
17KD_RICPR	PROTEIN NAME: 17 kDa surface antigen. SUBCELLULAR LOCATION: Cell outer membrane; Lipid-anchor. SIMILARITY: Belongs to the rickettsiale 17 kDa surface antigen family.
1A1D_CYBSA	PROTEIN NAME: 1-aminocyclopropane-1-carboxylate deaminase. FUNCTION: Catalyzes a cyclopropane ring-opening reaction, the irreversible conversion of 1-aminocyclopropane-1-carboxylate (ACC) to ammonia and alpha-ketobutyrate. SIMILARITY: Belongs to the ACC deaminase/D-cysteine desulfhydrase family.
1AP1_BRAOT	PROTEIN NAME: Floral homeotic protein APETALA 1-1. FUNCTION: Transcription factor that promotes early floral meristem identity in synergy with LEAFY. Displays a redundant function with CAULIFLOWER in the up-regulation of LEAFY. Required subsequently for the transition of an inflorescence meristem into a floral meristem, and for the normal development of sepals and petals in flowers. Regulates positively B class homeotic proteins (By similarity). SUBCELLULAR LOCATION: Nucleus.

Table 9: ProtST pre-training configurations. *Abbr.*, lr.: learning rate; bs.: batch size.

Model	optimizer	lr.	bs.	#epochs	train time
ProtST-ProtBert	Adam	1.0×10^{-5}	16	20	117h 10min
ProtST-ESM-1b	Adam	1.0×10^{-5}	12	20	205h 36min
ProtST-ESM-2	Adam	1.0×10^{-5}	12	20	206h 12min

BERT are frozen along the whole process. To reduce the memory cost, we truncate the protein sequences that have more than 450 residues to the length of 450, where the truncation starts from a random residue before the last 450 ones. Following MoCo (He et al., 2020), we initialize the temperature parameter τ in Eq. (1) as 0.07 and optimize it along the training process.

B.2. More Representation Learning Setups

Architecture of Prediction Heads: Following the default settings in TorchDrug (Zhu et al., 2022), the prediction of each task is performed by a two-layer MLP with ReLU nonlinearity in between. To be specific, given the protein representation, the MLP head is used to predict classification logits for localization prediction, regression score for fitness prediction and per-function classification logits for function annotation.

Table 10: Configurations of fix-encoder learning and full-model tuning on three task types. *Abbr.*, lr.: learning rate; bs.: batch size; MSE: mean squared error; CE: cross entropy; BCE: binary cross entropy.

Task	optimizer	lr.	bs.	#epochs	loss
fix-encoder learning					
Localization	Adam	5.0×10^{-5}	128	100	CE
Fitness	Adam	5.0×10^{-5}	128	100	MSE
full-model tuning					
Localization	Adam	2.0×10^{-4}	12	100	CE
Fitness	Adam	2.0×10^{-4}	24	100	MSE
Annotation	Adam	1.0×10^{-4}	8	50	BCE

Training Configurations: In Tab. 10, we present the detailed configurations of fix-encoder learning and full-model tuning on three task types, which mainly follows the configurations used in PEER benchmark (Xu et al., 2022b). For full-model tuning, the learning rate of the PLM is set as one tenth of the value in Tab. 10. The protein sequence encoders trained from scratch do not use smaller learning rates. All experiments are conducted on 4 Tesla V100 GPUs.

Evaluation Metrics: The protein function annotation tasks are measured by AUPR and $\mathrm{F}_{\max}$. We clarify their defini-

Table 11: Zero-shot protein classification performance under different prompt templates. *Abbr.*, Acc: accuracy; loc.: localization.

Prompt template	Label	Subcellular loc. (*Acc%*)	Reaction (*Acc%*)
Name only	Name	25.68	25.27
Natural language	Name	36.24	26.93
Pre-training template	**Name**	**43.49**	**29.85**
Pre-training template	Description	29.90	21.91

tions as below:

(1) **AUPR** denotes the pair-centric area under precision-recall curve. It computes the average precision scores for all protein-function pairs, which is exactly the micro-average precision score for the multiple binary classification problem.

(2) $\mathbf{F_{max}}$ denotes the protein-centric maximum F-score. Given a decision threshold $t \in [0, 1]$, it first calculates the precision and recall for each protein:

$$\text{precision}_i(t) = \frac{\sum_f \mathbb{1}[f \in P_i(t) \cap T_i]}{\sum_f \mathbb{1}[f \in P_i(t)]}, \tag{7}$$

$$\text{recall}_i(t) = \frac{\sum_f \mathbb{1}[f \in P_i(t) \cap T_i]}{\sum_f \mathbb{1}[f \in T_i]}, \tag{8}$$

where f denotes a functional term of EC or GO, T_i is the set collecting all experimentally determined functions for protein i, $P_i(t)$ denotes the predicted functions for protein i whose scores are at least t, and $\mathbb{1}[\cdot]$ represents the indicator function. The precision and recall are then averaged over all proteins:

$$\text{precision}(t) = \frac{1}{M(t)} \sum_i \text{precision}_i(t), \tag{9}$$

$$\text{recall}(t) = \frac{1}{N} \sum_i \text{recall}_i(t), \tag{10}$$

where N is the total number of proteins, and $M(t)$ denotes the number of proteins that contain at least one prediction larger than t, *i.e.*, $|P_i(t)| > 0$.

Finally, the F_{max} score is computed as the maximum value of F-measure over all thresholds:

$$\text{F}_{\text{max}} = \max_t \left\{ \frac{2 \cdot \text{precision}(t) \cdot \text{recall}(t)}{\text{precision}(t) + \text{recall}(t)} \right\}. \tag{11}$$

B.3. More Zero-shot Protein Classification Setups

Prompt Engineering for Subcellular Localization Prediction: Based on the information provided by DeepLoc (Almagro Armenteros et al., 2017), we consider two label formats, the *name* of each subcellular location (*i.e.*, the "Location" field in the Tab. 1 of DeepLoc paper) and the *description* of each location (*i.e.*, the "Sublocations" field in the Tab. 1 of DeepLoc paper). We further embed the labels into three prompt templates: (1) *Name only*: only the label itself is used; (2) *Natural language*: the label is embedded into the template "A protein locating at {`label`}."; (3) *Pre-training template*: the label is embedded into the template "SUBCELLULAR LOCATION: {`label`}".

According to the results in Tab. 11, we can observe that the pre-training template clearly outperforms other two templates on the subcellular localization prediction task, which mainly owes to the alignment of text format across pre-training and zero-shot prediction. It is shown that representing the labels with location names leads to better performance than using location descriptions, since the location names better fit the biomedical text distribution that the BLM is trained on. Based on these results, we represent the labels with the location names coupled with the pre-training prompt template on this task.

Prompt Engineering for Reaction Classification: Same as subcellular localization prediction, we also use two sets of label notations for reaction classification, *i.e.*, the *name* and the *description*. (1) The *name* refers to the composition of the enzyme class name and its alternative names, allowing unambiguous identification of each enzyme class. (2) The *description* further adds the scientific comments that discuss each class of enzymes in depth, which are extracted from scientific articles published by the International Union of Biochemistry and Molecular Biology (IUBMB). We retrieve all the information from Chang et al. (2021).

We embed such label information into three prompt templates: (1) *Name only*: the concatenation of the name and alternative names of an enzyme class, *i.e.*, "{`Name`} {`AlterNames`}"; (2) *Natural Language*: the label is incorporated into a natural-language-like template "A {`Name`} enzyme. This enzyme is also known as {`AlterNames`}."; (3) *Pre-training template*: the label is merged into the template used for pre-training, *i.e.*, "FUNCTION: {`Name`} {`AlterNames`}" (scientific comments "{`Comments`}" are appended after the names if the *description* is used).

According to Tab. 11, the pre-training template performs the best on the reaction classification task, mainly thanks to the consistent format of text descriptions between pre-training and zero-shot prediction. Injecting detailed scientific comments does not bring further benefits to the zero-shot performance. Therefore, we represent each enzyme class with its name and alternative names along with the pre-training prompt template for this task.

Nonparametric Few-shot Classifier: We adopt the nonparametric classifier proposed by Khandelwal et al. (2019) as baseline. Specifically, given n-shot K-class training samples $\{\{(S_i^k, y_i^k = k)\}_{i=1}^n\}_{k=1}^K$ composed of pairs of protein sequence and label, we employ the PLM to extract the rep-

Table 12: Performance comparison of PLMs on ProteinGym Substitution benchmark. *Abbr.*, retr.: retrieval.

Model	ProtST-ESM-1b	ESM-1b	ESM-1v	Tranception L (*w/o* retr.)	Progen2 XL
Model Type	PLM	PLM	PLM	PLM	PLM
UniProt-level Mean ρ	**0.412**	0.358	0.372	0.401	0.402

Table 13: ProtST-ESM-1b *v.s.* alignment-based methods on ProteinGym Substitution benchmark.

Model	ProtST-ESM-1b	EVE	GEMME	ProtST-ESM-1b + GEMME
Model Type	PLM	Align	Align	Hybrid
UniProt-level Mean ρ	0.412	0.443	0.459	**0.464**

resentations $\{\{z_i^k\}_{i=1}^n\}_{k=1}^K$ of all protein sequences. When a test protein S' comes, the nonparametric classifier first extracts its representation z' via the PLM and then derives its classification logits $\{y'_k\}_{k=1}^K$ by computing its representation similarity with each training protein:

$$y'_k = \sum_{i=1}^{n} \exp\left(-||z' - z_i^k||_2^2\right), \quad k = 1, \cdots, K. \qquad (12)$$

Softmax is performed upon these logits to derive classification probabilities. Such a classifier predicts based on the relations between test sample and training samples, which well fits the few-shot setting. In our experiments, the nonparametric classifier based on ESM-1b and the one based on ProtST-ESM-1b serve as two baselines for zero-shot classifiers.

C. Experimental Results on ProteinGym

C.1. Comparisons of Protein Language Models (PLMs)

Baselines. We compare the proposed ProtST-ESM-1b with four performant PLMs, *i.e.*, ESM-1b (Rives et al., 2021), ESM-1v (Meier et al., 2021), Tranception L (*w/o* retrieval) (Notin et al., 2022) and Progen2 XL (Nijkamp et al., 2022). Note that, for fair comparison, we do not include the PLMs with model ensemble (*e.g.*, VESPA (Marquet et al., 2022)) and the PLMs with inference-time retrieval (*e.g.*, Tranception L w/ retrieval (Notin et al., 2022)). We report the UniProt-level Mean Spearman's ρ.

Results. Under such a fair comparison, in Tab. 12, ProtST-ESM-1b achieves the best performance. In particular, compared with ESM-1b (*i.e.*, the initial PLM that ProtST-ESM-1b is based on), ProtST-ESM-1b obtains a significant performance gain with 15.1% relative improvement. This result demonstrates the effectiveness of the proposed multimodal training, which injects protein property knowledge into the ESM-1b and enhances its downstream fitness prediction performance.

Table 14: Ablation study of pre-training losses on localization and fitness prediction. *Abbr.*, Loc.: Localization; pred.: prediction; Acc: accuracy. Gray denotes the performance decay.

Model	**Loc. pred. (*Acc%*)**		**Fitness pred. (*Spearman's ρ*)**					
	Bin	**Sub**	**β-lac**	**AAV**	**Thermo**	**Flu**	**Sta**	**Mean ρ**
			Fix-encoder learning					
ProtST-ESM-1b	92.87	82.00	0.578	0.460	0.680	0.523	0.766	0.601
ProtST-ESM-1b (w/o $\mathcal{L}_{\mathrm{MPM}}$)	92.52	82.28	0.558	0.475	0.680	0.522	0.730	0.593
ProtST-ESM-1b (w/o $\mathcal{L}_{\mathrm{GC}}$)	92.12	80.55	0.560	0.448	0.684	0.467	0.738	0.579
ProtST-ESM-1b (w/o $\mathcal{L}_{\mathrm{MMP}}$)	92.81	82.00	0.544	0.479	0.681	0.504	0.731	0.588
			Full-model tuning					
ProtST-ESM-1b	92.35	78.73	0.895	0.850	0.681	0.682	0.751	0.772
ProtST-ESM-1b (w/o $\mathcal{L}_{\mathrm{MPM}}$)	92.64	77.59	0.894	0.842	0.681	0.685	0.726	0.766
ProtST-ESM-1b (w/o $\mathcal{L}_{\mathrm{GC}}$)	91.67	78.75	0.891	0.798	0.674	0.686	0.741	0.758
ProtST-ESM-1b (w/o $\mathcal{L}_{\mathrm{MMP}}$)	91.90	78.03	0.902	0.804	0.677	0.678	0.696	0.751

Table 15: Ablation study of pre-training losses on function annotation. Gray denotes the performance decay.

Model	**EC**		**GO-BP**		**GO-MF**		**GO-CC**	
	AUPR	$\mathrm{F}_{\max}$	AUPR	$\mathrm{F}_{\max}$	AUPR	$\mathrm{F}_{\max}$	AUPR	$\mathrm{F}_{\max}$
				Full-model tuning				
ProtST-ESM-1b	0.894	0.878	0.328	0.480	0.644	0.661	0.364	0.488
ProtST-ESM-1b (w/o $\mathcal{L}_{\mathrm{MPM}}$)	0.898	0.873	0.324	0.483	0.642	0.660	0.350	0.482
ProtST-ESM-1b (w/o $\mathcal{L}_{\mathrm{GC}}$)	0.894	0.870	0.322	0.463	0.638	0.656	0.327	0.462
ProtST-ESM-1b (w/o $\mathcal{L}_{\mathrm{MMP}}$)	0.890	0.871	0.328	0.456	0.635	0.659	0.340	0.473

C.2. Comparisons with Alignment-based Methods

Baselines. In this experiment, we involve two alignment-based methods, *i.e.*, EVE (Frazer et al., 2021) and GEMME (Laine et al., 2019), for comparison. We further investigate the ensemble of ProtST-ESM-1b and GEMME. We report the UniProt-level Mean Spearman's ρ.

Results. In Tab. 13, it is observed that the alignment-based methods are superior over ProtST-ESM-1b, since they additionally utilize the homologous information within sequence alignments, which is not utilized by ProtST-ESM-1b. However, by combining the normalized predictions of ProtST-ESM-1b and GEMME, the ensemble model "ProtST-ESM-1b + GEMME" outperforms these two SOTA alignment-based methods. This result verifies the complementary knowledge hidden in ProtST-ESM-1b and an alignment-based model in terms of fitness prediction. Therefore, it will be a promising direction to study the combination of these two lines of methods. We leave this as our future work.

D. More Zero-shot Text-to-Protein Retrieval Results

In Fig. 10, we study four more sets of text-to-protein retrieval of ligand binders based on ProtST-ESM-1b. For each study, we visualize the text prompt and the top-4 retrieved candidates. For each candidate, we present the docking result of it binding with the ligand, the binding affinity and its GO molecular function label of binding with the ligand, where AutoDock Vina (Trott & Olson, 2010) is used to estimate docking pose and binding affinity. It is observed that, among the top-4 candidates, ProtST-ESM-1b succeeds in retrieving 3 GO-annotated ATP binders (only

Table 16: Ablation study of BLM on localization and fitness prediction. ProtST-ESM-1b serves as the base model. *Abbr.*, Loc.: Localization; pred.: prediction; Acc: accuracy.

BLM	**Loc. pred. (*Acc%*)**		**Fitness pred. (*Spearman's ρ*)**					
	Bin	**Sub**	**β-lac**	**AAV**	**Thermo**	**Flu**	**Sta**	**Mean ρ**
			Fix-encoder learning					
PubMedBERT-abs	92.87	82.00	**0.578**	**0.460**	0.680	**0.523**	**0.766**	**0.601**
PubMedBERT-full	**93.04**	**82.28**	0.548	0.458	**0.682**	0.507	0.744	0.588
			Full-model tuning					
PubMedBERT-abs	92.35	78.73	0.895	**0.850**	**0.681**	**0.682**	**0.751**	**0.772**
PubMedBERT-full	**92.87**	**78.77**	**0.899**	0.785	0.672	0.680	0.722	0.752

Table 17: Ablation study of BLM on function annotation. ProtST-ESM-1b serves as the base model.

BLM	**EC**		**GO-BP**		**GO-MF**		**GO-CC**	
	AUPR	$\mathrm{F_{max}}$	AUPR	$\mathrm{F_{max}}$	AUPR	$\mathrm{F_{max}}$	AUPR	$\mathrm{F_{max}}$
				Full-model tuning				
PubMedBERT-abs	0.894	**0.878**	**0.328**	**0.480**	**0.644**	**0.661**	0.364	**0.488**
PubMedBERT-full	**0.905**	**0.878**	0.323	0.475	0.630	0.652	**0.374**	0.485

3.99% proteins are annotated as ATP binders in GO), 3 GO-annotated GTP binders (only 1.18% proteins are annotated as GTP binders in GO), 2 GO-annotated P5P binders (only 0.17% proteins are annotated as P5P binders in GO), and 2 GO-annotated NAD+ binders (only 0.05% proteins are annotated as NAD+ binders in GO). The rest candidates annotated as non-binding also own decent binding affinity, *e.g.*, the better binding affinity of protein 2AKA-B (*without* ATP binder annotation) against protein 6EAC-A (*with* ATP binder annotation), the better binding affinity of protein 5DHG-A (*without* NAD+ binder annotation) against protein 3GFB-A (*with* NAD+ binder annotation), *etc.* These results demonstrate the general effectiveness of ProtST-ESM-1b on retrieving the binders of diverse ligands. In the future work, we will study how ProtST enables zero-shot text-to-protein retrieval of other types of functional proteins, *e.g.*, antigen binders, toxic substance binders, transcription factors, *etc.*

E. More Ablation Study

E.1. Ablation Study of Pre-training Losses

In Tabs. 14 and 15, we report the performance of ProtST-ESM-1b on all benchmark tasks by using full or partial pre-training losses. It can be observed that: (1) removing the loss $\mathcal{L}_{\mathrm{MPM}}$ leads to performance decay on 16 out of 24 benchmark metrics; (2) removing the loss $\mathcal{L}_{\mathrm{GC}}$ leads to decay on 20 out of 24 benchmark metrics; (3) removing the loss $\mathcal{L}_{\mathrm{MMP}}$ diminishes model performance on 19 out of 24 benchmark metrics. Therefore, all pre-training losses are necessary to maximize the effectiveness of a ProtST-induced PLM, where $\mathcal{L}_{\mathrm{GC}}$ and $\mathcal{L}_{\mathrm{MMP}}$ inject different granularities of protein property information into a PLM, and $\mathcal{L}_{\mathrm{MPM}}$ preserves the PLM's original representation power.

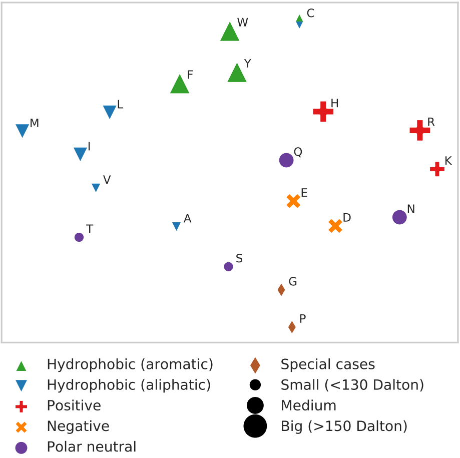


Figure 6: Amino acid representations learned by the linear layer for unimodal mask prediction (ProtST-ESM-1b is used).

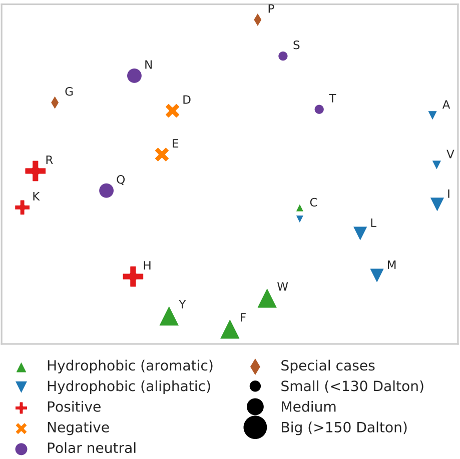


Figure 7: Amino acid representations learned by the linear layer for multimodal mask prediction (ProtST-ESM-1b is used).

E.2. Ablation Study of Biomedical Language Model

PubMedBERT owns two versions: (1) the PubMedBERT-abs trained by using only PubMed abstracts, and (2) the PubMedBERT-full trained by using additional PubMed Central full-text articles. In this experiment, we compare the effectiveness of these two models by respectively using them as the BLM of ProtST-ESM-1b.

Tabs. 16 and 17 report the performance comparison of these two models on all benchmark tasks. We can observe that: (1) PubMedBERT-full outperforms PubMedBERT-abs on all four benchmark metrics of localization prediction; (2) PubMedBERT-abs performs better than PubMedBERT-full on 10 out of 12 benchmark metrics of fitness prediction; (3) PubMedBERT-abs outperforms PubMedBERT-full on 5 out of 8 benchmark metrics of function annotation. Therefore, PubMedBERT-full does not show superiority over PubMedBERT-abs in ProtST pre-training, which owes to the fact that the protein property descriptions in the ProtDe-

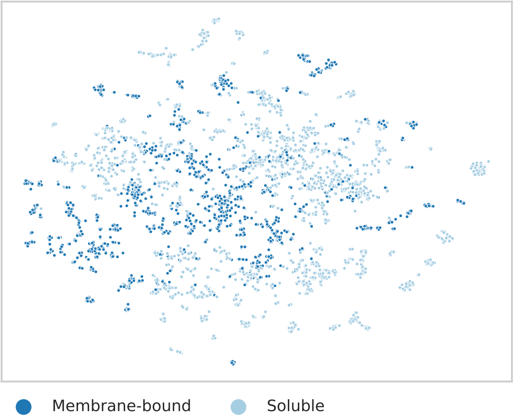


Figure 8: Visualization of protein representations on the binary localization prediction dataset (ProtST-ESM-1b is used).

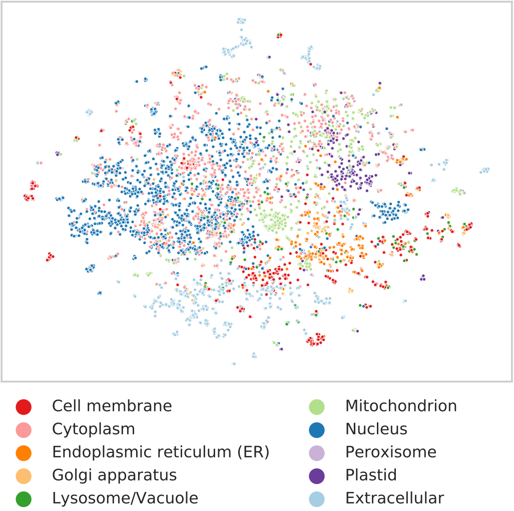


Figure 9: Visualization of protein representations on the subcellular localization prediction dataset (ProtST-ESM-1b is used).

scribe dataset are more like abstracts than full-text articles.

F. More Visualization

Well-trained PLMs should have the capacity to extract structural, functional, and even evolutionary features of proteins. As a result, the learned representations in PLMs are expected to have certain intrinsic organization patterns in the embedding space to capture these protein characteristics. To demonstrate the effectiveness of ProtST-ESM-1b, we use t-SNE (Van der Maaten & Hinton, 2008) to visualize such information at different scales from amino acid decompositions to protein functional properties.

Biophysical Properties of Amino Acids: It is known that the biophysical properties of amino acids, such as hydrophobicity, aromaticity and charge, highly influence the biological structures of proteins and therefore their biological functions as well. To investigate if ProtST-ESM-1b captures such intrinsic features, we apply t-SNE to the two linear layers used for unimodal mask prediction and multimodal mask prediction. As shown in Figs. 6 and 7, hydrophobic and polar residues exhibit clear distinct clusterings, even to the level of aliphatic *v.s.* aromatic. The clustering is also coherent in terms of the charge and size of the amino acids.

Biological and Biochemical Properties of Proteins: As introduced in Sec. 4.1, our proposed ProtDescribe dataset provides ProtST-ESM-1b with direct access to knowledge like protein subcellular localizations, which refers to a specific region within a cell where the proteins can be found. For a protein, such locations can influence its activity and interaction with other molecules, thus helping the PLMs to better capture the biological and biomedical protein functions. To validate this assumption, we adopt the datasets used in two protein localization prediction tasks, *i.e.*, the subcellular localization prediction and the binary localization prediction. With t-SNE, we project protein representations to the 2-dimensional space for these two benchmark datasets. In Figs. 8 and 9, certain clustering patterns of different cellular locations are observed.

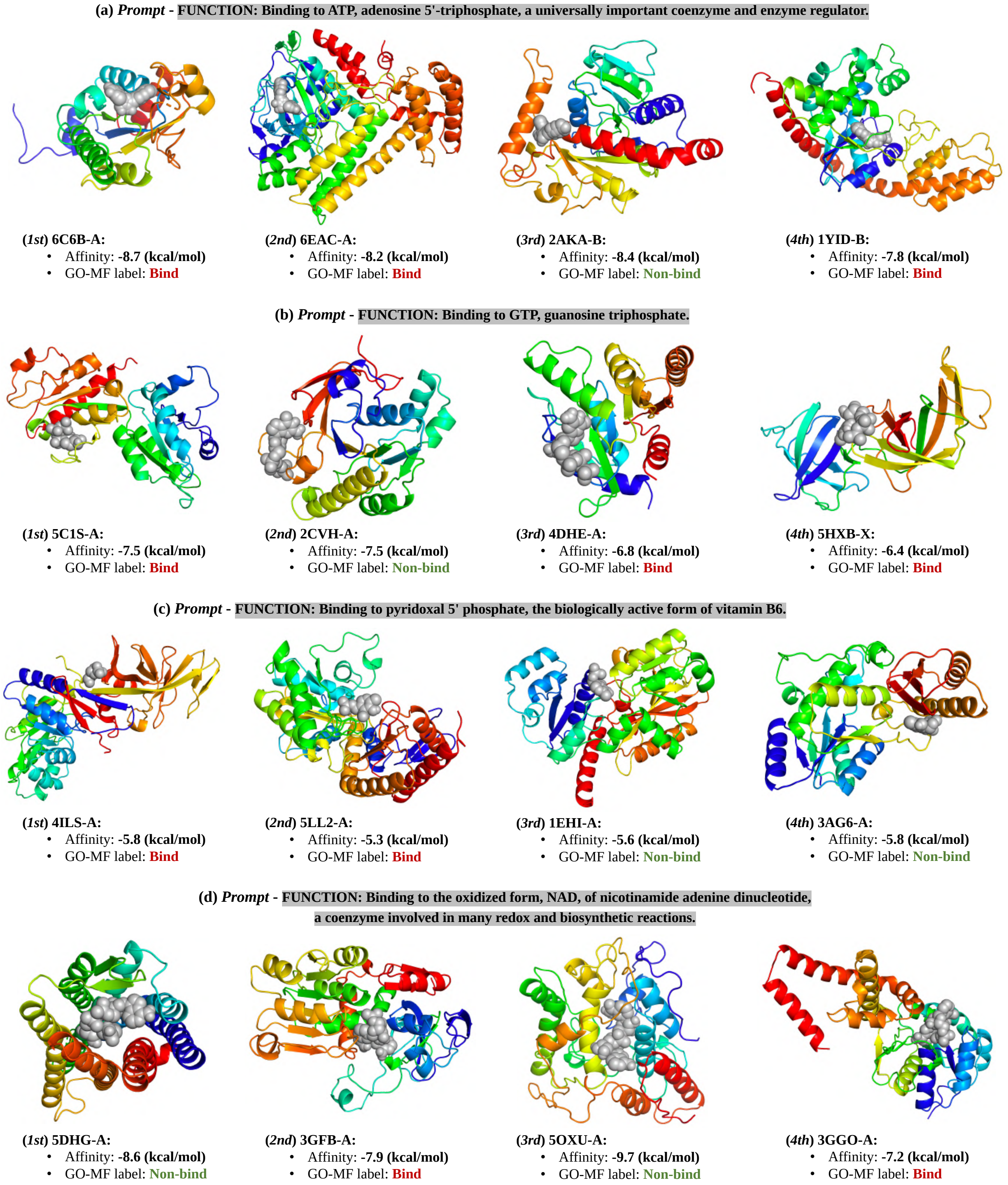

(a) Prompt - FUNCTION: Binding to ATP, adenosine 5'-triphosphate, a universally important coenzyme and enzyme regulator.
(1st) 6C6B-A:
• Affinity: -8.7 (kcal/mol)
• GO-MF label: Bind
(2nd) 6EAC-A:
• Affinity: -8.2 (kcal/mol)
• GO-MF label: Bind
(3rd) 2AKA-B:
• Affinity: -8.4 (kcal/mol)
• GO-MF label: Non-bind
(4th) 1YID-B:
• Affinity: -7.8 (kcal/mol)
• GO-MF label: Bind
(b) Prompt - FUNCTION: Binding to GTP, guanosine triphosphate.
(1st) 5C1S-A:
• Affinity: -7.5 (kcal/mol)
• GO-MF label: Bind
(2nd) 2CVH-A:
• Affinity: -7.5 (kcal/mol)
• GO-MF label: Non-bind
(3rd) 4DHE-A:
• Affinity: -6.8 (kcal/mol)
• GO-MF label: Bind
(4th) 5HXB-X:
• Affinity: -6.4 (kcal/mol)
• GO-MF label: Bind
(c) Prompt - FUNCTION: Binding to pyridoxal 5' phosphate, the biologically active form of vitamin B6.
(1st) 4ILS-A:
• Affinity: -5.8 (kcal/mol)
• GO-MF label: Bind
(2nd) 5LL2-A:
• Affinity: -5.3 (kcal/mol)
• GO-MF label: Bind
(3rd) 1EHI-A:
• Affinity: -5.6 (kcal/mol)
• GO-MF label: Non-bind
(4th) 3AG6-A:
• Affinity: -5.8 (kcal/mol)
• GO-MF label: Non-bind
(d) Prompt - FUNCTION: Binding to the oxidized form, NAD, of nicotinamide adenine dinucleotide, a coenzyme involved in many redox and biosynthetic reactions.
(1st) 5DHG-A:
• Affinity: -8.6 (kcal/mol)
• GO-MF label: Non-bind
(2nd) 3GFB-A:
• Affinity: -7.9 (kcal/mol)
• GO-MF label: Bind
(3rd) 5OXU-A:
• Affinity: -9.7 (kcal/mol)
• GO-MF label: Non-bind
(4th) 3GGO-A:
• Affinity: -7.2 (kcal/mol)
• GO-MF label: Bind

Figure 10: Zero-shot text-to-protein retrieval of (a) ATP binders, (b) GTP binders, (c) P5P binders, and (d) NAD+ binders based on ProtST-ESM-1b.