

Functional Protein Design with Local Domain Alignment

Chaohao Yuan^{2*}, Songyou Li³, Geyan Ye¹, Yikun Zhang⁴, Long-Kai Huang¹, Wenbing Huang³,
Wei Liu¹, Jianhua Yao¹, Yu Rong^{1†}

¹Tencent AI Lab

²Tsinghua University

³Renmin University of China

⁴Peking University

yuanch22@mails.tsinghua.edu.cn, songyou_li@126.com, blazerye@tencent.com,
yikun.zh@hotmail.com, hlongkai@gmail.com, hwenbing@126.com, topliu@tencent.com,
jianhua.yao@gmail.com, yu.rong@hotmail.com

Abstract

The core challenge of *de novo* protein design lies in creating proteins with specific functions or properties, guided by certain conditions. Current models explore to generate protein using structural and evolutionary guidance, which only provide indirect conditions concerning functions and properties. However, textual annotations of proteins, especially the annotations for protein domains, which directly describe the protein’s high-level functionalities, properties, and their correlation with target amino acid sequences, remain unexplored in the context of protein design tasks. In this paper, we propose **Protein-Annotation Alignment Generation (PAAG)**, a multi-modality protein design framework that integrates the textual annotations extracted from protein database for controllable generation in sequence space. Specifically, within a multi-level alignment module, PAAG can explicitly generate proteins containing specific domains conditioned on the corresponding domain annotations, and can even design novel proteins with flexible combinations of different kinds of annotations. Our experimental results underscore the superiority of the aligned protein representations from PAAG over 7 prediction tasks. Furthermore, PAAG demonstrates a nearly sixfold increase in generation success rate (24.7% vs 4.7% in zinc finger, and 54.3% vs 8.7% in the immunoglobulin domain) in comparison to the existing model. We anticipate that PAAG will broaden the horizons of protein design by leveraging the knowledge from between textual annotation and proteins.

1 Introduction

Protein design Marshall et al. (2019) is a crucial task for its immense potential on drug discovery Sliwoski et al. (2014), enzyme engineering Sterner et al. (2008), immunengineering Swartz et al. (2012) and so on. The generation of proteins with specific properties, behaviors, or functions, such as optimizing the binding affinity to given molecules Houk et al. (2003) or incorporating a particular ion-binding site Regan (1995), is known as *de novo* protein design. This process presents a significant challenge due to the vast space of protein sequences and the complexity of protein functions. Recently, machine learning models have shown profound potential for protein design. The existing studies mostly rely on the structural Watson et al. (2023) or evolutionary information Alamdari et al. (2023a) as the guidance to design proteins. However, in many cases, these conditions can only offer indirect

*Work is done when Chaohao Yuan worked as an intern in Tencent AI Lab.

†Yu Rong is the corresponding author.

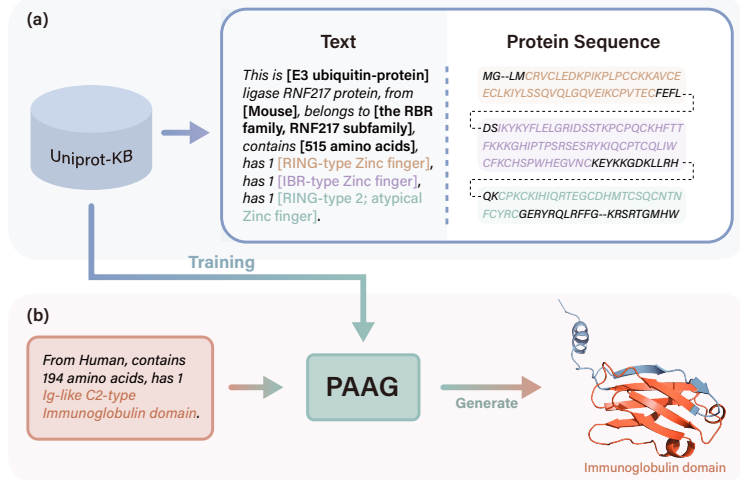


Figure 1: (a) The example of property annotations (in bold) and domain annotations (in colors). (b). The illustration of annotation-guided protein design with PAAG. Given the input of textual description within immunoglobulin domain annotation, PAAG can generate the proteins containing immunoglobulin domain.

guidance towards the desired protein design targets to their inherent ambiguity. For example, the same protein sequence segment can be either act as receptors to regulate synaptic function Stathakis et al. (1997) or helpers to locate target proteins to specific subcellular locations Mohandas et al. (1995).

For instance, as one of crucial functional domains of DNA/RNA binding proteins Cassandri et al. (2017), the zinc-finger domain naturally has many variants, such as C2H2 type, CCHC type and Zinc ribbon type. These variants exhibit significant differences in both structural and evolution features among them which is hard covered by structural and evolutionary conditions. On the contrary, the “zinc-finger” annotation from the protein database inherently provides a more effective means of describing the high-level knowledge span across both sequences and structures. Hence, we aim to investigate the following question: *Is it possible to leverage such textual annotations to guide the delicate controllable protein design?*

Recently, several primary attempts have been made to leverage such textual annotations to guide the protein generation. Examples include extracting tags from labels to create unique tokens Madani et al. (2023), training an individual protein caption model to guide the diffusion generation process Ingraham et al. (2023), and incorporating the an overall text description through a global language-to-protein alignment model Liu et al. (2023b). However, current models cannot flexibly combine the different conditions and lack of the capability for fine-grained control, such as specifying the generation of particular domains.

To fill these gaps, in this paper, we introduce a novel framework, **Protein-Annotation Alignment Generation (PAAG)**, that enables annotation-guided protein design by aligning protein sequences with their textual annotations. Specifically, we first consider both property and domain annotations in proteins and design a multi-level alignment module to align the representations of sequences and annotations extracted from the existing encoders in both global and local level. For the generation task, PAAG utilizes an autoregressive decoder to generate protein sequences guided by the aligned representation of textual annotations. Additionally, PAAG employs an end-to-end training pipeline that joint training of alignment and generation tasks without freezing the parameters in sequence and text encoders. This joint training enhances the understanding of the complex and flexible annotation condition, resulting in improved guided generation. Figure 1 demonstrates an example of annotation-guided generation.

In experiments, we first investigate the quality of protein representations from PAAG by seven predictive downstream tasks. PAAG surpasses state-of-the-art benchmarks with an average relative improvement of 1.5%. Subsequently, three protein generation tasks are conducted to assess the capabilities of PAAG. In the case of unconditional generation, PAAG produces sequences exhibiting the highest degree of novelty while maintaining the distribution of natural proteins. For the two

conditional protein design tasks, PAAG demonstrates a nearly threefold increase in generation success rate (24.7% vs 4.7% in zinc finger, and 54.3% vs 8.7% in the immunoglobulin domain)³ in comparison to the existing model.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to propose a new annotation-guided paradigm for protein design, which can simultaneously integrate local and global annotation information. Our proposed framework PAAG is the first approach that can generate proteins containing specific domains, guided by their corresponding annotations with high success rate.
- PAAG features a multi-level alignment module for handling annotation and protein data alignment at various granularities. Joint training of alignment and generation tasks allows the model to produce improved protein representations, consequently boosting performance in predictive and generative tasks.
- Comprehensive experiments on 7 predictive and 3 generative tasks showcase PAAG’s superiority compared to the existing methods. Notably, PAAG is not only capable of generating proteins that include a single domain, but it also successfully generates proteins adhering to the flexible conditions of multiple annotation combinations.

2 Preliminaries

2.1 Protein and Its Textual Annotations

The primary structure of a protein can be represented as an amino acid sequence $S = (s_1, s_2, \dots, s_n)$, where s_i is the i -th amino acids chosen from 20 different amino acids which are represented as 20 characters.

Given a protein S , the annotation-sequence pair set $\mathcal{D}_S = \{(T_k, S_{i:j}^k)\}_{k=1}^{K_S}$ is constructed by extracting the correspondence between textual annotations and protein domains & properties from protein database, such as UniProtKB Consortium (2022), where T_k is the textual annotations of the k -th domain $S_{i:j}^k$ of the protein and K_S is the number of annotation-domain pairs of the protein S . For the property annotation, the corresponding domain is the entire sequence, i.e., $S_{1:|S|}$. This annotation-domain pair set \mathcal{D}_S provides comprehensive knowledge about the protein S and, therefore, plays an important role in training the protein generation model.

2.2 Encoders for Proteins and Annotations

We adopt a protein encoder (PE) to generate both the protein and its domain representations based on the amino acid sequence as $z_S = f_{PE}(S)$, where f_{PE} is the protein encoder and z_S is the embedding of S .

For textual annotations, we generate their representations using the pre-trained language model f_{LM} as $z_{\mathcal{A}_S} = f_{LM}(G(\mathcal{A}_S))$, where $\mathcal{A}_S \subseteq \{T_k | (T_k, S_{i:j}^k) \in \mathcal{D}_S\}$ is a subset of the annotations in the annotation-domain pair set for a protein S , $G()$ is a template function which converts a set of annotations to a textual description, and $z_{\mathcal{A}}$ is the embedding of the annotation set. Note that the annotation subset can cover all annotations in \mathcal{D}_S or just one. If \mathcal{A}_S covers only one annotation T_0 , we can skip the template function and directly generate the embedding of T_0 using f_{LM} .

We employ transformer-based protein encoder and text encoder as our base model and initialize them with pre-trained models. Specifically, we employ SciBERTBeltagy et al. (2019), which is pre-trained on computer science and biological datasets, to initialize the text encoder f_{LM} . The pre-trained protein encoder, ProtBERTEInaggar et al. (2021), is used to initialize the protein encoder f_{PE} .

3 Methodology

In this section, we propose a multi-modal framework, **Protein-Annotation Alignment Generation(PAAG)**, which enables the flexible annotations-guided protein design. Figure 2 il-

³This is the success rate with quality threshold $e = 1$, i.e., SR_1 .

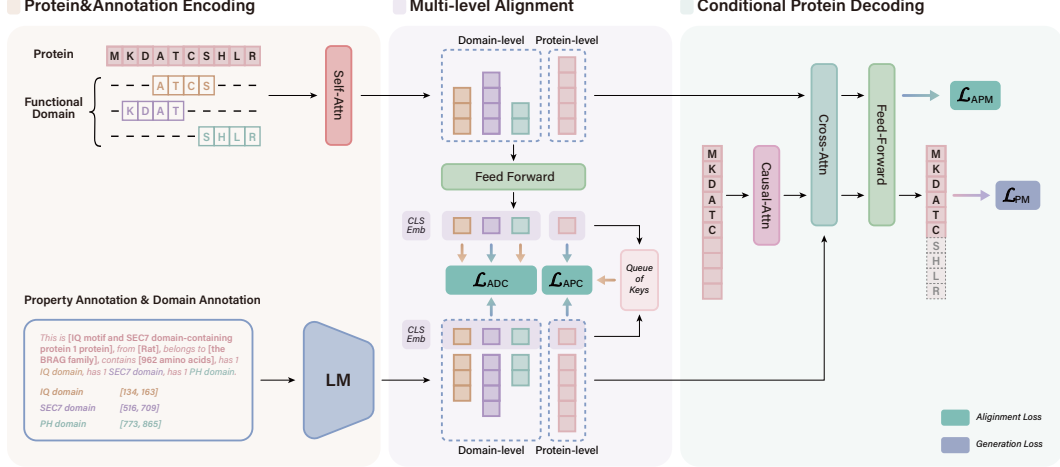


Figure 2: The overall framework of PAAG. The same parameters share the same color. PAAG contains three modules. (1) Protein & Annotation Encoding module encode the input protein sequence & domains and corresponding annotations to the embeddings. (2) Multi-level alignment module projects the protein and annotation embeddings into and employs Annotation-Protein Contrastive (APC) loss, Annotation-Domain Contrastive (ADC) loss and Annotation-Protein Matching (APM) loss to align them in a same latent space. (3) Conditional Protein Decoding accepts the annotation embedding as input and generate the protein sequence.

lustrates the overall framework of PAAG. In the following, we present the details of each component of PAAG.

3.1 Conditional Protein Decoding

Our ultimate goal is generating functional proteins conditioned on a set of annotations $\mathcal{A} = \{T_k\}_{k=1}^K$. We will first employ template function G to translate these annotations as textual descriptions and then utilize a language model to extract the representation of the annotation set $z_{\mathcal{A}} = f_{\text{LM}}(G(\mathcal{A}))$. Then the protein sequence is generated using a decoder f_D based on this representation as $S = f_D(z_{\mathcal{A}})$.

Protein Decoder. We adopt an auto-regressive protein decoder f_D that can receive the condition from the language model. Specifically, the decoder will first process the protein sequence through causal attention layers to enable auto-regressive generation, then integrate the information from annotations via cross-attention layers followed by feed-forward layers. Note that the causal attention layers are initialized using the same weights as the self-attention layers in the protein encoder and the feed-forward layers share the same parameters as the protein encoder. Here, the parameter-sharing mechanism enables higher training efficiency Li et al. (2023). The cross-attention layers are randomly initialized and trained from scratch.

Protein Modeling (PM) Loss. To guide the learning of the model, we estimate the PM loss as

$$\mathcal{L}_{PM} = - \sum_{i=1}^l \log p(S_i | S_{<i}; \mathcal{A}), \quad (1)$$

where l is the length of the protein S and $p(S_i | S_{<i}; \mathcal{A})$ is the predicted probability of the i -th amino acid given all previous amino acids and the annotation set.

3.2 Multi-level Protein and Annotation Alignment

To better integrate the information between proteins and annotations, we aim to align the multi-level representations of proteins and annotations. Specifically, we conduct local alignment and global alignment by performing contrastive learning at domain level and protein level, respectively.

We measure the alignment score using the cosine similarity between the embeddings of protein and the annotation set, which is defined as

$$s(\mathcal{A}, S) = \frac{\langle \text{Proj}_{\mathcal{A}}(z_{\mathcal{A}}), \text{Proj}_S(z_S) \rangle}{\|\text{Proj}_{\mathcal{A}}(z_{\mathcal{A}})\| \|\text{Proj}_S(z_S)\|}, \quad (2)$$

where $\text{Proj}_a(z) = \mathbf{W}_a z + \mathbf{b}$ projects the input z into a latent space with dimension h . $\mathbf{W}_a \in \mathbb{R}^{h \times |z|}$ and $\mathbf{b} \in \mathbb{R}^h$ are trainable parameters. During alignment, we encourage the matched (positive) pairs to have representations with higher similarity s than unmatched (negative) pairs. We next will explain in detail how to construct the positive and negative pairs in our multi-level framework.

3.2.1 Local Alignment

Given a protein S , and its annotation-domain pairs set \mathcal{D}_S , in domain-level, we aim to align the representation of the domain $S_{i:j}^k$ and its corresponding annotation T_k and use all annotation-domain pairs in \mathcal{D}_S as positive pair. To construct the negative pairs, we randomly sample the sub-regions outside the domain $S_{i:j}^k$, i.e. $S_{1:i-1}^k \cup S_{j+1:L}^k$, as the negative samples $S_{i:j}^{k-}$ of the domain.

Annotation-Domain Contrastive(ADC) Loss: We adopt InfoNCE loss as the ADC loss to align the representations of local annotation z_T and functional domain $z_{S_{i:j}}$ as

$$\mathcal{L}_{\text{ADC}} = -\frac{1}{K_S} \sum_{k=1}^{K_S} \log \frac{\exp(s(S_{i:j}^k, T_k)/\tau)}{\sum_{n=1}^N \exp(s(S_{i:j}^{n-}, T_n)/\tau)}, \quad (3)$$

where K_S is the number of functional domains for the protein S , N is the number of negative samples, and τ is a learnable temperature parameter.

The local alignment enables PAAG to explicitly learn the relation between the functional domain and its annotation. Therefore, we can use the model to controllably generate the specific functional domain given annotations.

3.2.2 Global Alignment

To enable global alignment, we utilize template function $G(\cdot)$ to construct protein-level textual description and form the positive protein-description pairs as $\{(G(\mathcal{A}_S), S)\}$. Since multiple annotations and the entire protein will be more complex, to enlarge the number of negative samples for global alignment, we follow the setting of MOCO He et al. (2020) to construct momentum encoders f_m as

$$f_m \leftarrow m f_m + (1 - m) f. \quad (4)$$

where the encoder f can be either protein encoder f_{PE} or text encoder f_{LM} , and m is the momentum hyperparameter. We follow implementation details in Li et al. (2021) and Li et al. (2022) to construct the momentum encoders for both encoders.

The momentum encoders extract consistent features to increase the number of negative samples, and the dynamic dictionaries will store these features.

Annotation-Protein Contrastive (APC) Loss is designed to align the representations of global properties $z_{\mathcal{A}}$ with protein sequence z_S . Specifically, for each protein sequence and annotation set, we calculate the softmax-normalized sequence-to-annotation and annotation-to-sequence similarity as:

$$p_m^{s2a}(S) = \frac{\exp(s(S, A_m)/\tau)}{\sum_{m=1}^M \exp(s(S, A_m)/\tau)}, \quad (5)$$

$$p_m^{a2s}(A) = \frac{\exp(s(A, S_m)/\tau)}{\sum_{m=1}^M \exp(s(A, S_m)/\tau)}, \quad (6)$$

where τ is a learnable temperature parameter, A_m and S_m indicate their representations will be extracted by respective momentum encoders.

Denote $\vec{y}^{a2s}(A)$ and $\vec{y}^{s2a}(S)$ as the ground-truth one-hot similarity, where negative pairs have a probability of 0 and the positive pair has a probability of 1. The annotation-protein contrastive loss is defined as the cross-entropy H between \vec{p} and \vec{y} :

$$\mathcal{L}_{\text{APC}} = \frac{1}{2} \mathbb{E}_{(S,A) \sim D} [\text{H}(\vec{y}^{\text{s2a}}(S), \vec{p}^{\text{s2a}}(S)) + \text{H}(\vec{y}^{\text{a2s}}(A), \vec{p}^{\text{a2s}}(A))]. \quad (7)$$

Annotation-Protein Matching (APM) Loss. Inspired by Li et al. (2022), we introduce a multimodal encoder f_{ME} , which integrates the representation of annotations and protein as $z_{S,A}^M = f_{\text{ME}}(S, G(\mathcal{A}))$, to identify whether the given pairs of protein description $G(\mathcal{A})$ and protein S are matched or not. Specially, the multimodal encoder f_{ME} is transformer-based and shares the parameters of self-attention and feed-forward layers with f_{PE} and has additional cross-attention layers between self-attention layers and feed-forward layers to integrate the text information. The cross-attention layers share the same cross-attention parameters as the decoder.

Annotation-Protein Matching (APM) Loss aims to facilitate the learning of multimodal representations. Additionally, we use the hard negative strategy, where we first select the most similar negative pairs and then use these most challenging negative pairs to optimize the model through in the training of the model with the APM loss, enabling the encoder to learn informative representations. To construct the APM loss, we extract the multi-modal representation as $z_{S,A}^M = f_{\text{ME}}(S, G(\mathcal{A}))$ and use a classifier to classify its probability of being positive or negative, which is denoted by $\vec{p}^{\text{APM}}(\mathcal{A}, S)$. And we compute the cross-entropy between the ground-truth label $\vec{y}^{\text{APM}}(A, S)$ indicating the protein-description pair being positive or negative to obtain the APM loss as

$$\mathcal{L}_{\text{APM}} = \mathbb{E}_{(A,S)} \text{H}(\vec{y}^{\text{APM}}(A, S), \vec{p}^{\text{APM}}(A, S)). \quad (8)$$

3.3 Training Objectives

In the training of PAAG, we optimize four objective functions. These functions are designed to align the representations between annotations and protein sequences, integrate the two modalities, and reconstruct the protein sequence. In contrast to ProteinDT Liu et al. (2023a), which splits the training process into three separate stages, PAAG jointly optimizes these four objective functions in an end-to-end manner.

The overall pretraining objective of PAAG is :

$$\mathcal{L} = \mathcal{L}_{\text{ADC}} + \mathcal{L}_{\text{APC}} + \mathcal{L}_{\text{APM}} + \mathcal{L}_{\text{PM}}. \quad (9)$$

3.4 Annotation-guided Protein Design

After obtaining the model F of PAAG, we can use F to design the proteins with given textual annotations. In the following, we given the definition of annotation-guided protein design.

Definition 3.1 (Annotation-guided Protein Design) *Given an annotation set $\mathcal{A} = \{T_k\}_{k=1}^K$ with K annotations, a generative model F can leverage the information from $\{T_k\}_{k=1}^K$ to generate the protein S which satisfies the condition described in $\{T_k\}_{k=1}^K$. Quantitatively, this task aims to maximize the following objective function:*

$$\max \sum_{T \in \mathcal{A}} M_T(S), \text{ s.t. } S = F(\mathcal{A}). \quad (10)$$

where $M_T()$ is corresponding metric function for the annotation T .

In this paper, we mainly focus on two type of measures, *functional-domain measure* and *global-property measure*.

- **Functional-domain metric:** for the annotation describing a certain functional domain, $M_T()$ will invoke a profile hidden Markov models from Pfam Mistry et al. (2021) to search for the optimal match within protein S regarding to the domain described by annotation T and assign an e-value s to this match. Given an e-value threshold e , we have:

$$M_{T,e}(S) = \mathbb{1}_S(s < e), \quad (11)$$

where $\mathbb{1}_S(\text{cond})$ outputs 1 when $\text{cond} = \text{True}$, otherwise, outputs 0.

- **Global-property metric:** for the annotations describing protein’s global properties, $M_T()$ employ a pre-defined oracle to determine if the given S has the property described by the annotation T and output a score s in the range of $[0, 1]$. Here, 0 denotes the absence of the property, while 1 signifies its presence.

4 Experiment

In this section, we extensively evaluate PAAG from two aspects: (1). quality of aligned protein representation for the predictive tasks; (2) evaluation on the unconditional sequence generation and protein design with textual annotations.

4.1 Construction of ProtAnnotation Dataset

To enable multi-level alignment, we build the ProtAnnotation dataset with annotation-sequence pair set for protein design task. Specifically, we select all proteins with “Domain” entry in UniProtKB Consortium (2022) to build ProtAnnotation dataset, resulting 129,727 proteins. The domain annotations are extracted by these “Domain” entries including the domain description and start & end index of this domain. Additionally, we select four properties as the property annotations, that is, “protein_name”, “organism_name”, “length” and “SIMILARITY”. The textual description is assembled by the template function G , which can accept any subset of annotation as input. More details, including the data examples are deferred in Appendix A.2.

4.2 Quality of Aligned Representation

We first conduct multiple experiments on predictive tasks. to evaluate the quality of protein representation produced by PAAG.

Settings To make a fair comparison with ProtST, we use the same proteins in dataset ProtDescribe Xu et al. (2023) to first pretrain PAAG, followed by full-model fine-tuning on various downstream tasks. For the full-model fine-tuning, we add a task head for each task and fine-tune the model for 100 epochs. We use the validation set to select the model and report the results on random seed 0, adhering to the same settings as in ProtST Xu et al. (2023). More details are deferred in Appendix A.6.

Benchmark Tasks: We adopt 7 downstream tasks within two task types as the benchmark task.

- **Protein Localization Prediction** aims to forecast the subcellular locations of proteins. Derived from DeepLoc Almagro Armenteros et al. (2017), we focus on two similar tasks. The subcellular localization prediction (*Abbr. as*, Sub) encompasses 10 location categories and binary localization prediction (*Abbr. as*, Bin) that includes 2 location categories, soluble and membrane-bound. The splits of data follow the original split in DeepLoc.
- **Fitness Landscape Prediction** is primarily focused on the prediction of the effects of residue mutations on the fitness of proteins. We evaluate our models on β -lactamase (*Abbr.*, β -lac) landscape from PEER Xu et al. (2022), 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). The splits of data follow the splitting setting of ProtST Xu et al. (2023)

Baselines: We adopt two types of baseline. The first type is the models train from scratch, including CNN Shانهsazzadeh et al. (2020), ResNet Rao et al. (2019), LSTM Rao et al. (2019), Transformer Rao et al. (2019). The second type is the pretrained models with full model tuning, including OntoProtein Zhang et al. (2022), ProtBert Elnaggar et al. (2021) and ESM2 Lin et al. (2023). For the ProtST and PAAG, we train two variants with different initialization weights from ProtBert and ESM2 to verify the enhancement of text knowledge on protein representations from different protein encoders. We utilize distinct subscripts to denote the initialized parameters, such as $\text{PAAG}_{\text{ProtBert}}$ and $\text{PAAG}_{\text{ESM2}}$.

Results: Table 1 reports the results of all models on seven baselines. As illustrated in Table 1, PAAG achieves superior performance in comparison to the baselines on all 7 tasks. These results highlight the robust generalization capabilities of PAAG for downstream tasks. Specifically, PAAG outperforms the vanilla pretrained models, ProtBert and ESM2, in all cases, indicating that PAAG can further enhance the quality of protein representation by incorporating the knowledge from textual annotations. Furthermore, to assess the impact of multimodal training on enhancing protein representations, we visualize the relative improvements of ProtST and PAAG in comparison to the vanilla pre-trained models in Figure 3. As observed in Figure 3, multi-modal training has a positive influence on the performance of downstream tasks in both ProtST and PAAG. Additionally, the

Table 1: Results on protein localization prediction and protein landscape prediction benchmarks. Bold in green and underlined numbers indicate the best and the second best result, respectively.

Model	Loc. pred. (Acc%)		Fitness pred. (Spearman’s ρ)				
	Bin \uparrow	Sub \uparrow	ρ -lac \uparrow	AAV \uparrow	Thermo \uparrow	Flu \uparrow	Sta \uparrow
Models trained from scratch							
CNN	82.67	58.73	0.781	0.746	0.494	0.682	0.637
ResNet	78.99	52.30	0.152	0.739	0.528	0.636	0.126
LSTM	88.11	62.98	0.139	0.125	0.564	0.494	0.533
Transformer	75.74	56.02	0.261	0.681	0.545	0.643	0.649
Models with full model tuning							
OntoProtein	92.47	77.59	0.757	0.791	0.662	0.630	0.731
ProtBert	91.32	76.53	0.731	0.794	0.660	0.679	0.771
ESM2	91.72	78.67	0.867	0.817	0.672	0.677	0.718
ProtST _{ProtBert}	91.78	78.71	0.863	0.804	0.673	0.679	<u>0.745</u>
ProtST _{ESM2}	<u>92.52</u>	<u>80.22</u>	<u>0.879</u>	<u>0.825</u>	<u>0.682</u>	0.682	0.738
PAAG _{ProtBert}	92.63	78.96	0.820	0.825	0.668	<u>0.680</u>	0.788
PAAG _{ESM2}	92.46	81.30	0.888	0.839	0.684	0.682	0.737

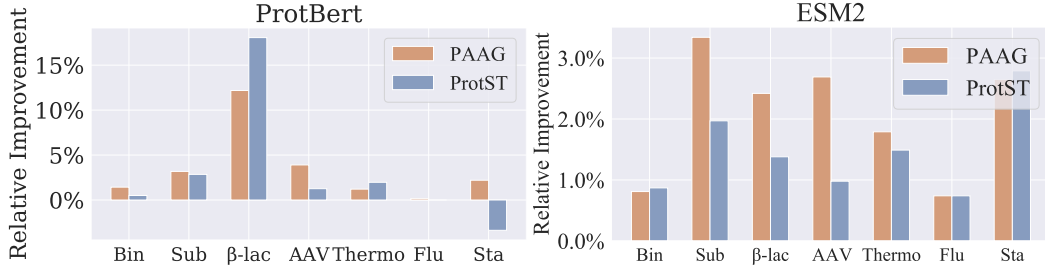


Figure 3: Visualization of jointly generation with domain and property annotations. PAAG is capable of integrating domain and property annotations.

performance improvement of PAAG surpasses that of ProtST in 10 out of 14 cases, further validating the effectiveness of PAAG in augmenting the quality of protein representation.

4.3 Unconditional Protein Generation

To verify the learning effect of the decoder, we compare the ability of different models in unconditional generation task. A good decoder is able to generate protein sequences that conform to the distribution of the training set while simultaneously exhibiting adequate novelty.

Setting: In unconditional generation task, we only specify the length of generated proteins and conditions. We sample the same length from natural proteins, to ensue a fair comparisons across different models.

Baselines: We compare PAAG with two representative protein design models, *i.e.*, ProGen Madani et al. (2023) and Chroma Ingraham et al. (2023). Furthermore, we introduce two naive baselines: Random_{Uniform} and Random_{Empirical}. Random_{Uniform} generates protein sequence by randomly selecting amino acids based on a uniform distribution, while Random_{Empirical} adheres to the empirical amino acid distribution observed in the training dataset. Additionally, we also report the results of the sequence set sampled from natural proteins, denoted as Natural, to serve as a reference. The details of baselines are deferred to Appendix A.5.

Evaluation metrics: To evaluate the quality of generated protein sequences, we employ three metrics: Distinct-n, Diversity and Novelty. Suppose \mathcal{S} is the protein sequence set.

- **Distinct-n** Li et al. (2016) is a classical metric in natural language processing that measures textual diversity of generated text by counting distinct n-grams. We use this metric to assesses the fraction of repetitive sequence motifs in sequences from \mathcal{S} , which exhibits the biological importance Andrade et al. (2000). A higher Distinct-n suggests fewer repetitive amino acid segments. We set $n = 2$ here.

Table 2: Results of the unconditional generation task. The values in parentheses $\Delta\%$ represent the absolute relative difference from the values of Natural.

Model	Distinct-n ($\Delta\%$)	Diversity($\Delta\%$)	Novelty \uparrow
Natural	235.76(0)	0.829(0)	-
Random _{Uniform}	273.89(16.17%)	0.847(2.17%)	0.713
Random _{Empirical}	243.04(3.09%)	0.834(6.03%)	0.721
ProGEN	165.54(29.78%)	0.845(1.93%)	0.374
Chroma	175.72(25.47%)	0.855(3.14%)	0.638
PAAG	231.88(1.66%)	0.815(1.69%)	0.766

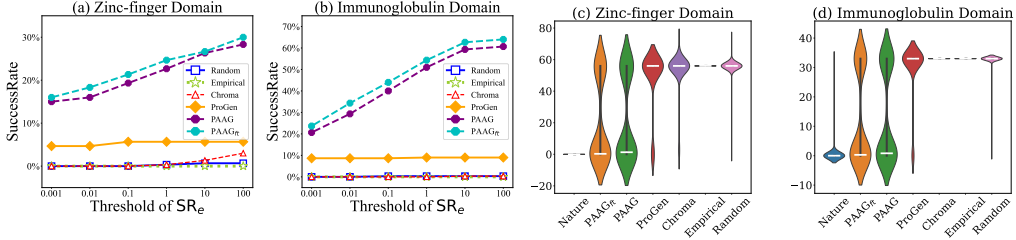


Figure 4: Figure (a) and (b) show the SR_e on zinc-finger domain and immunoglobulin domain over all models. Figure (c) and (d) show their distributions of e-value. White bar indicates the mean e-value of each set. PAAG consistently exhibits better performance on all metrics compared with other models. Fine-tuning also introduces additional improvement for PAAG.

- **Diversity** measures the dissimilarity of sequences in S . We employ Mmseq2 Steinegger and Söding (2017) to compute the dissimilarity between each pair of sequences, and utilize the mean of these dissimilarities as Diversity. A higher Diversity signifies a greater diversity in S .
- **Novelty** measures the novelty of sequences in S compared to a reference set. We take UniprotKB Consortium (2022) as the reference set. For each sequences in S , we employ Mmseq2 Steinegger and Söding (2017) to return the dissimilarity score of the most similar sequence in UniprotKB. Novelty is defined as the mean of dissimilarity score for all sequences in S . A higher Novelty indicates the generated sequences exhibit substantial novelty comparing with the reference set.

Results: Table 2 shows the results of unconditional generation under three metrics. From Table 2, we can observe that:

- PAAG achieves the highest novelty, indicating PAAG captures the intrinsic relationship between amino acids, rather than merely memorizing protein sequences in the training set.
- PAAG exhibits the closet Distince-n score compared with natural proteins, which proves that proteins generated by PAAG possess similar amino acid distribution as natural proteins.
- PAAG also maintains the closest diversity to natural proteins, proving PAAG has the similar distribution at protein level.
- ProGen and Chroma have considerably lower Distince-n score in comparison to PAAG and natural proteins, implying an abundance of repetitions within their generated protein sequence.

In a summary, PAAG is capable of generating high-quality protein sequences, aided by the alignment process.

4.4 Protein Design with Domain Annotations

In this section, we evaluate the performance of PAAG in generating proteins under the given domain annotations.

Settings: We utilize two biologically significant domains, zinc-finger domain Klug and Rhodes (1987) and immunoglobulin domain Brummendorf (1995), as the target domain annotation to generate the proteins respectively. For each case, we generate $N = 300$ protein sequences given the

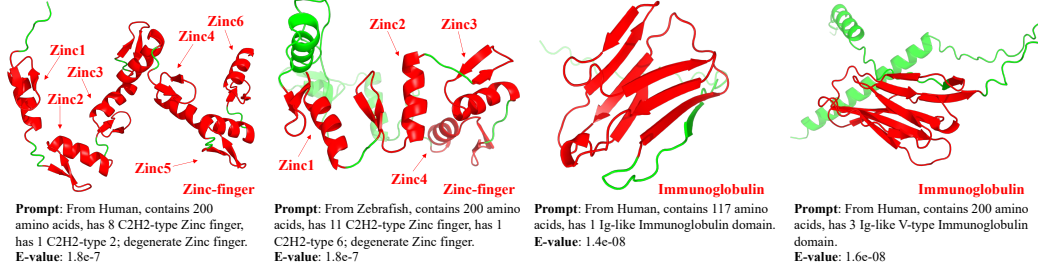


Figure 5: Visualization of the generated results on zinc-finger and immunoglobulin domain. The corresponding prompt and generation qualify (e-value) is listed below.

length of the proteins and the annotation set containing the domain annotations of “zinc-finger” or “immunoglobulin domain”. For all models, the protein sequences are generated with the same length sampled from natural proteins that have corresponding domain in UniprotKB.

Based on (11), we further define a metric: success rate $SR_e = \frac{M_{T,e}(S)}{N}$ to measure the generation quality with the proportion of proteins that successfully identify the specific domain with the quality threshold e .

Model training: We train PAAG on ProtAnnotation for total 100 epochs, employing a learning rate of 3e-5 and incorporating a warm-up phase with a batch size of 32. Additionally, we extract a subset of ProtAnnotation which only contains proteins with “zinc-finger” and “immunoglobulin domain”. This subset is subsequently utilized to fine-tune the model over 5 epochs, adopting a learning rate of 1e-5 and maintaining the batch size at 32. We denote this fine-tuned version as $PAAG_{ft}$. More details are deferred to Appendix A.6.

Baselines: We adopt Chroma Ingraham et al. (2023) and ProGen Madani et al. (2023) as the baselines due to their public availability of model weights and their capacity to accept text and keywords as conditional inputs for protein generation. ProGen utilizes keywords as its condition tags. Here, we take the keywords correspond to zinc-finger (“KW-0863”) and immunoglobulin domain (“KW-0393”) in Uniprot to generate functional proteins. Chroma adopts ProCap to understand the textual conditions to guide its diffusion process. We give the same text prompts to Chroma to enable its controllable generation. We also include two trivial baselines $Random_{Uniform}$ and $Random_{Empirical}$ as the blank references.

Results: In Figure 4 (a) and (b), we plot the curves depicting the variation in success rate SR_e for different models in two protein design tasks as the quality threshold e varies from 0.001 to 100. Additionally, in Figure 4 (c) and (d), we employ violin plots to illustrate the distribution of e-value e for generated sequences matched in Pfam. Due to Pfam’s limitations, e-values exceeding 100 are adjusted to the maximum e-value score among all method results for the current task. We also show the distribution of e-values calculated on natural proteins with domains as a reference. From Figure 4, we can observe that:

- PAAG significantly outperforms the other baselines in success rates SR_1 across all tasks by a large margin, e.g. , 51% versus 9% in the immunoglobulin domain task with quality threshold $e = 1$. Furthermore, the fine-tuned version $PAAG_{ft}$ can further improve the success rates (54.3% on SR_1) consistently, indicating that the finetuning process can help to generate proteins with high quality domains.
- ProGen outperforms other baseline methods, maintaining consistent results as the threshold changes. This may be due to ProGen remembering the zinc and Ig domain sequences present in the training data and output them when using keywords and the first four amino acid sequences as prompts for generation.
- Chroma’s performance is not good, resembles the unconditioned results. This may be because Chroma’s training text primarily focuses on structural descriptions, making it less sensitive to the domain annotations.

Visualizations: We provide the visualization of generated proteins in Figure 5. In Figure 5, all generated domains are marked in red. The textual descriptions generated by the annotation set and

Table 3: The result of protein design with “membrane-bound” and “soluble” annotations.

Annotation	Total	Matched	Success Rate
membrane-bound	660	351	53.18%
soluble	894	811	90.72%
overall	1554	1162	74.78%

Table 4: Ablation study of the effectiveness of each component. Performance of SR_1 for zinc-finger and immunoglobulin domains.

\mathcal{L}_{ADC}	\mathcal{L}_{APC}	\mathcal{L}_{APM}	Zinc-finger	Immunoglobulin
	✓	✓	0.00%	0.33%
✓		✓	39.00%	18.33%
✓	✓		41.33%	19.33%
✓	✓	✓	51.00%	22.67%

the e-value e of this sequence are listed below. The protein structure are folded by OmegaFold Wu et al. (2022). As observed in Figure 5, the generated sequences can accurately produce the target domain as per the condition given in the annotation set. Interestingly, in scenarios such as two zinc finger case, when the prompt specifies the presence of multiple zinc finger domains, PAAG generates multiple functional domains in response. However, PAAG falls to capture the precise numbers of these domains, which can be a direction for future improvement.

4.5 Protein Design with Property Annotations

In this section, we explore the potential of PAAG to generate proteins with certain properties guided by property annotations.

Settings: We employ the subcellular location of proteins as an example property. An additional dataset, termed ProtLocation, is generated by extracting subcellular location labels, encompassing both Bin (2-class) and Sub (10-class) tasks as delineated in Section 4.2, from Deeploc Almagro Armenteros et al. (2017). These labels are then incorporated into annotation-sequence pairs derived from Uniprot. ProtLocation includes 10100 proteins for training, while a separate set of 2434 test proteins is reserved for constructing the annotation set for generation. More details can be found in Appendix

Evaluation protocol: For the generated sequences, we employ the official server provided by Deeploc Almagro Armenteros et al. (2017) and Deeploc2 Thumhuri et al. (2022) as the pre-defined oracle to construct the Global-property metric $M_T()$ for predicting binary location label and 10-class label respectively. A generation is deemed successful if the predicted label aligns with the input annotation label.

Results: Table 3 depicts the generation result of PAAG on binary localization annotation. PAAG achieves overall 74.78% success rate, indicating PAAG captures the difference between soluble and membrane-bound, and can generate properties given corresponding property annotations. As for a more challenging setting, PAAG achieves overall 49.71% success rate for the 10-class subcellular localization annotation, still indicating a strong correlation between property annotations and the generated proteins by PAAG. The detailed confusion matrix about 10-class subcellular localization annotation are depicted in Figure 8 in Appendix.

4.6 Ablation Study

To ascertain the contribution of each component towards the generation of the functional proteins, the ablation study reports the SR_1 of zinc-finger and immunoglobulin domains in the absence of each alignment loss, as presented in Table 4. We observe that \mathcal{L}_{ADC} is the key to the high SuccessRate of PAAG. The SuccessRate decreases to 0% without \mathcal{L}_{ADC} , underscoring the importance of incorporating domain range into the learning framework. Furthermore, \mathcal{L}_{APC} and \mathcal{L}_{APM} also enhance the SuccessRate of generating high-quality immunoglobulin domains by 4.33% and 3.33%, respectively. Moreover, \mathcal{L}_{APC} and \mathcal{L}_{APM} are more important to zinc-finger domain, by improving the SR_1 by 12% and 9.67%.

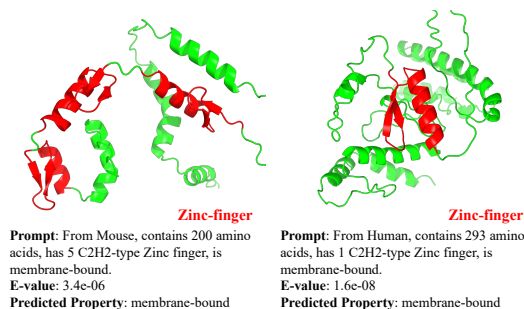


Figure 6: Visualization of jointly generation with domain and property annotations. PAAG is capable of integrating domain and property annotations.

4.7 Case study: Joint Generation with Domain and Property Annotations

Due to the flexibility of combining annotations, PAAG can jointly generate the proteins with both specific domain and property. Specifically, we generate the zinc-finger proteins in membrane-bound by the model in Section 4.5. We showcase two examples of the generated results in Figure 6. As shown in Figure 6, PAAG can successfully generate proteins under the conditions derived from both domain and property annotations, demonstrating its potential in complex protein design tasks.

5 Related Work

Multimodal Representation Learning. By harnessing the potential of extensive image-text pair data, the Contrastive Language-Image Pretraining (CLIP) model, as proposed by Radford et al. (2021), employs contrastive learning to align the representations between image and text modalities. Following by CLIP, many image-text pertaining model are proposed, such as BILP Li et al. (2022), BLIP-2 Li et al. (2023) and ClipCap Mokady et al. (2021). Beyond the image-text pretraining, several studies introduce the more modalities, such as videos Xu et al. (2021), audios Tang et al. (2023) and even molecules Liu et al. (2022) into a unified representation. Specifically, for multimodal learning on protein sequences, OntoProtein Zhang et al. (2022) first learns protein representations by combining them with textual descriptions in a knowledge graph. ProtST Xu et al. (2023) constructs a large-scale dataset containing aligned pairs of protein sequences and property descriptions, and pretrain a protein-biotext model to improve performance on downstream predictive task and enables zero-shot retrieval.

Protein Generation Model. With huge success of language models Devlin et al. (2019), several studies Elnaggar et al. (2021); Shin et al. (2021); Ferruz et al. (2022); Rives et al. (2019) treat protein sequences consisting chains of amino acids as a type of “languages” and pretrain models on millions of protein sequences. Upon the pretrained model, they generate the protein sequences in an autoregressive manner. In addition to the autoregressive model, Evodiff Alamdari et al. (2023b) extracts the evolutionary information from protein sequences and proposes an evolution-guided diffusion model to generate protein sequences. Given the significance of structural information for protein function, a group of methods Hsu et al. (2022); Ingraham et al. (2019); Tan et al. (2022); Jing et al. (2020), known as inverse folding, utilize the structure as a conditional input, allowing for the generation of amino acid sequences. Some studies Shi et al. (2022); Watson et al. (2023) integrate both sequential and structural information and propose a co-design model that accepts sequences and structures as conditions. Recently, ProteinDT Liu et al. (2023b) first proposes a text-sequence alignment framework, enabling its capabilities for text-guided protein generation and editing. Chroma Ingraham et al. (2023) trains a protein caption model ProCap and utilize it as a classifier guidance to generate proteins via a diffusion model.

6 Conclusion

In this paper, we present PAAG, a multi-modality framework that first incorporates the rich annotation information derived from protein database, achieving the superior performance in various applications, such as representation learning and annotation-guided protein design. Crucially, we demonstrate that it is possible to use the flexible combinations of various kinds of textual annotations to guide the protein design process. We hope that PAAG will expand the possibilities of protein design

and establish a robust foundation for future advancements in protein-related applications. Future research directions include functional protein editing, co-design of protein sequence-structure within alignment frameworks, and exploring the potential of larger protein dataset with lower annotation quality.

References

- Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. 2023a. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv* (2023), 2023–09.
- Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex X. Lu, Nicolo Fusi, Ava P. Amini, and Kevin K. Yang. 2023b. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv* (2023). <https://doi.org/10.1101/2023.09.11.556673> arXiv:<https://www.biorxiv.org/content/early/2023/09/12/2023.09.11.556673.full.pdf>
- José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 21 (2017), 3387–3395.
- Miguel A Andrade, Chris P Ponting, Toby J Gibson, and Peer Bork. 2000. Homology-based method for identification of protein repeats using statistical significance estimates. *Journal of molecular biology* 298, 3 (2000), 521–537.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv:1903.10676 [cs.CL]
- T Brummendorf. 1995. Cell adhesion molecules. 1. Immunoglobulin superfamily. *Protein profile* 2 (1995), 963–1108.
- Matteo Cassandri, Artem Smirnov, Flavia Novelli, Consuelo Pitolli, Massimiliano Agostini, Michal Malewicz, Gerry Melino, and Giuseppe Raschella. 2017. Zinc-finger proteins in health and disease. *Cell death discovery* 3, 1 (2017), 1–12.
- The UniProt Consortium. 2022. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* 51, D1 (11 2022), D523–D531. <https://doi.org/10.1093/nar/gkac1052> arXiv:<https://academic.oup.com/nar/article-pdf/51/D1/D523/48441158/gkac1052.pdf>
- Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. 2021. FLIP: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv* (2021), 2021–11.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]
- Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. 2021. ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. <https://doi.org/10.1109/TPAMI.2021.3095381>
- Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications* 13, 1 (2022), 4348.
- Yaron Geffen, Yanay Ofran, and Ron Unger. 2022. DistilProtBert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics* 38, Supplement_2 (2022), ii95–ii98.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

- KN Houk, Andrew G Leach, Susanna P Kim, and Xiyun Zhang. 2003. Binding affinities of host-guest, protein-ligand, and protein-transition-state complexes. *Angewandte Chemie International Edition* 42, 40 (2003), 4872–4897.
- Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. 2022. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*. PMLR, 8946–8970.
- John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. 2019. Generative models for graph-based protein design. *Advances in neural information processing systems* 32 (2019).
- John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. 2023. Illuminating protein space with a programmable generative model. *Nature* 623, 7989 (2023), 1070–1078.
- Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. 2020. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411* (2020).
- A Klug and D Rhodes. 1987. Zinc fingers: a novel protein fold for nucleic acid recognition. In *Cold Spring Harbor symposia on quantitative biology*, Vol. 52. Cold Spring Harbor Laboratory Press, 473–482.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). Association for Computational Linguistics, San Diego, California, 110–119. <https://doi.org/10.18653/v1/N16-1014>
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.
- Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 6637 (2023), 1123–1130.
- Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. 2023a. A Text-guided Protein Design Framework. *arXiv:2302.04611 [cs.LG]*
- Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2022. Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789* (2022).
- Shengchao Liu, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Anthony Gitter, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. 2023b. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611* (2023).
- Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* 41, 8 (2023), 1099–1106.
- Liam R Marshall, Oleksii Zozulia, Zsofia Lengyel-Zhand, and Ivan V Korendovych. 2019. Minimalist de novo design of protein catalysts. *ACS catalysis* 9, 10 (2019), 9265–9275.

- Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. 2021. Pfam: The protein families database in 2021. *Nucleic acids research* 49, D1 (2021), D412–D419.
- T.K. Mohandas, X.-N. Chen, L.B. Rowe, E.H. Birkenmeier, A.S. Fanning, J.M. Anderson, and J.R. Korenberg. 1995. Localization of the Tight Junction Protein Gene TJP1 to Human Chromosome 15q13, Distal to the Prader-Willi/Angelman Region, and to Mouse Chromosome 7. *Genomics* 30, 3 (1995), 594–597. <https://doi.org/10.1006/geno.1995.1281>
- Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734* (2021).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. 2019. Evaluating protein transfer learning with TAPE. *Advances in neural information processing systems* 32 (2019).
- Lynne Regan. 1995. Protein design: novel metal-binding sites. *Trends in biochemical sciences* 20, 7 (1995), 280–285.
- Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. 2019. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *PNAS* (2019). <https://doi.org/10.1101/622803>
- Amir Shanehsazzadeh, David Belanger, and David Dohan. 2020. Is transfer learning necessary for protein landscape prediction? *arXiv preprint arXiv:2011.03443* (2020).
- Chence Shi, Chuanrui Wang, Jiarui Lu, Bozitao Zhong, and Jian Tang. 2022. Protein sequence and structure co-design with equivariant translation. *arXiv preprint arXiv:2210.08761* (2022).
- Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. 2021. Protein design and variant prediction using autoregressive generative models. *Nature communications* 12, 1 (2021), 2403.
- Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W Lowe. 2014. Computational methods in drug discovery. *Pharmacological reviews* 66, 1 (2014), 334–395.
- Dean G Stathakis, Kevin B Hoover, Zhiyong You, and Peter J Bryant. 1997. Human postsynaptic density-95 (PSD95): location of the gene (DLG4) and possible function in nonneural as well as in neural tissues. *Genomics* 44, 1 (1997), 71–82.
- Martin Steinegger and Johannes Söding. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* 35, 11 (2017), 1026–1028.
- Reinhard Sterner, Rainer Merkl, and Frank M. Raushel. 2008. Computational Design of Enzymes. *Chemistry & Biology* 15, 5 (2008), 421–423. <https://doi.org/10.1016/j.chembio.2008.04.007>
- Melody A Swartz, Sachiko Hirose, and Jeffrey A Hubbell. 2012. Engineering approaches to immunotherapy. *Science translational medicine* 4, 148 (2012), 148rv9–148rv9.
- Cheng Tan, Zhangyang Gao, Jun Xia, Bozhen Hu, and Stan Z Li. 2022. Generative de novo protein design with global context. *arXiv preprint arXiv:2204.10673* (2022).
- Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-Any Generation via Composable Diffusion. *arXiv preprint arXiv:2305.11846* (2023).

- Vineet Thummuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. 2022. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research* 50, W1 (2022), W228–W234.
- Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. 2023. De novo design of protein structure and function with RFdiffusion. *Nature* 620, 7976 (2023), 1089–1100.
- Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. 2022. High-resolution de novo structure prediction from primary sequence. *BioRxiv* (2022), 2022–07.
- Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. *CoRR* abs/2109.14084 (2021). arXiv:2109.14084 <https://arxiv.org/abs/2109.14084>
- Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. 2023. ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts. *arXiv preprint arXiv:2301.12040* (2023).
- Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. 2022. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems* 35 (2022), 35156–35173.
- Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. 2022. OntoProtein: Protein Pretraining With Gene Ontology Embedding. arXiv:2201.11147 [q-bio.BM]

A Experimental settings

A.1 General settings

Backbone Models of PAAG: Unless otherwise specified, we utilize the weights from ProtBert-BFD Elnaggar et al. (2021) as the initialization of our protein encoder. For the text encoder, we choose SciBert Beltagy et al. (2019) as the initial backbone. The scarcity of training data does not allow us to train a large decoder, we utilize a lighter decoder, initialized by DistilProtBert Geffen et al. (2022). A decoder that contains more parameters is also our future direction.

Training Configurations: We train PAAG on ProtAnnotation, using the AdamW optimizer (with a learning rate of $3e-5$ and zero weight decay) for 100 epochs. Our generation experiments are conducted on 16 NVIDIA Tesla A100-SX4-40GB GPUs.

A.2 ProtAnnotation and the template function $G(\cdot)$

Figure 5 depicts additional statistics of ProtAnnotation. We demonstrate several example data samples in ProtAnnotation as well as the corresponding textual description generated by the template function $G(\cdot)$.

Table 5: The statistics of ProtAnnotation

# of proteins	# of distinct domains	average # of domains per protein	average length
129,727	1,416	1.60	419.40

The explanation of four property annotations:

- **protein_name:** The protein name is a naming method used to describe the function, characteristics, or origin of a protein. These names usually contain information about the protein’s structure, function, substrate specificity, and biological process.

- **organism_name:** Organism names are used to identify and classify different species of living organisms, including bacteria, fungi, plants, and animals. These names usually consist of the genus and species of the organism, and sometimes include additional information such as strain or cultivar.
- **length:** The number of amino acids in the protein sequence.
- **SIMILARITY:** In the context of biology and protein classification, “similarity” refers to the shared characteristics or features among different proteins. This can include similar structures, functions, or evolutionary origins. Proteins with high similarity are often grouped into the same family or subfamily.

Table 6: Examples of annotations and description of protein in ProtAnnotation

Entry name	Domain Annotation	Property Annotation	Textual Description G(D)
Q8W4R8	[38, 95], Ubiquitin-like; degenerate domain [158, 459], PI3K/PI4K catalytic domain	organism_name: Mouse-ear cress, protein_name: Phosphatidylinositol 4-kinase gamma 6, length: 622, SIMILARITY: Belongs to the PI3/PI4-kinase family, Type II PI4K subfamily	This is Phosphatidylinositol 4-kinase gamma 6 protein , from Mouse-ear cress , belongs to the PI3/PI4-kinase family , Type II PI4K subfamily , contains 622 amino acids , has 1 Ubiquitin-like; degenerate domain , has 1 PI3K/PI4K catalytic domain .
Q8XAW7	[5, 241], ABC transporter domain [252, 495], ABC transporter domain	organism_name: Escherichia coli O157:H7, protein_name: Ribose import ATP-binding protein RbsA 1, length: 501, SIMILARITY: Belongs to the ABC transporter superfamily, Ribose importer (TC 3.A.1.2.1) family	This is Ribose import ATP-binding protein RbsA 1 protein , from Escherichia coli O157:H7 , belongs to the ABC transporter superfamily , Ribose importer (TC 3.A.1.2.1) family , contains 501 amino acids , has 2 ABC transporter domain .
Q8Y2D7	[254, 288], FPG-type Zinc finger	organism_name: Pseudomonas solanacearum, protein_name: Formamidopyrimidine-DNA glycosylase, length: 288, SIMILARITY: Belongs to the FPG family	This is Formamidopyrimidine-DNA glycosylase protein , from Pseudomonas solanacearum , belongs to the FPG family , contains 288 amino acids , has 1 FPG-type Zinc finger .

A.3 Biological background of zinc-finger and immunoglobulin domain

We provide some introductions of zinc-finger and immunoglobulin domain from biological perspective, elucidating their significance and functions within the context of molecular biology.

- **Zinc-finger:** Zinc finger Klug and Rhodes (1987) domains are compact protein structural motif with multiple protrusions that interact with target molecules. Initially identified in the transcription factor TFIIIA from *Xenopus laevis* as a DNA-binding motif, these domains are now recognized for their ability to bind DNA, RNA, proteins, and lipids. Their binding characteristics depend on the amino acid sequence, the linker structure, the higher-order structures, and the number of fingers. Often found in clusters, each finger within the Zinc finger domain can have unique binding specificities. Despite being present in several unrelated protein superfamilies, these domains maintain stable structures and play a crucial role in diverse cellular processes such as gene transcription, translation, mRNA trafficking, and more. Zinc-binding motifs, part of the Zinc finger domains, rarely undergo significant conformational changes upon target binding.
- **Immunoglobulin domain:** Immunoglobulin (Ig) domains Brummendorf (1995), also referred to as immunoglobulin-like domains or Ig-like domains, are protein domains that of importance in the immune system and are found in a wide range of proteins, both within and beyond the immune system. These domains are characterized by their highly conserved structure, which typically consists of 70 to 110 amino acid residues that form a compact, globular fold. The Ig domains are ubiquitous in nature and can be found in a diverse array of proteins, including antibodies, T-cell receptors, major histocompatibility complex (MHC) molecules, cell adhesion molecules, and various receptor proteins. The primary function of Ig domains is to facilitate protein-protein interactions, particularly in the context of the immune system, where they mediate the recognition, binding, and neutralization of foreign antigens.

An illustration of Zinc-finger and Immunoglobulin domains is shown in Figure 7.

A.4 The evaluation metric for the unconditional generation

In the computation of the Distinct-n metric, we assign a value of n to 2. For the Novelty metric, the default parameters of Mmseq2 are employed with an exception for the e-value threshold, which

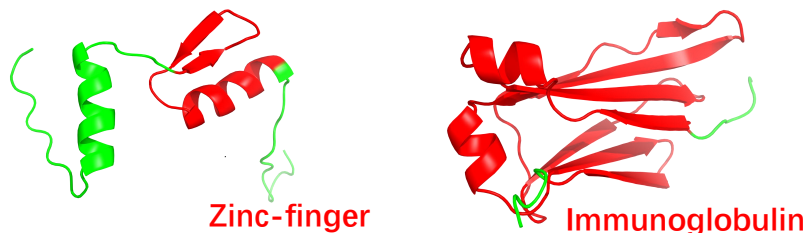


Figure 7: Visualization of typical proteins containing Zinc-finger (P26634, left figure) protein and Immunoglobulin domain (A0M8Q6, right figure).

is designated as 100. In the case of the Diversity metric, the default parameters of Mmseq2 are again utilized, but with alterations to the e-value threshold and sensitivity, set to 1000000 and 15 respectively. During the application of Mmseq2 for the calculation of Novelty and Diversity metrics, the dissimilarity for unmatched sequences is assumed to be 1.0.

A.5 More details of baselines

ProGen Madani et al. (2023) first forms each property of protein as keyword tags. In the training stage, ProGen prepends the keyword tags to the corresponding amino acid sequence, and try to reconstruct the amino acid sequence in the auto-regressive manner. ProGen consequently has the ability to generate the protein sequence given the keyword tags. During controllable generation, ProGen first translates desired properties into keyword tags, then can controllably generate protein sequence.

Chroma Ingraham et al. (2023) is also a protein design model that but focus more on generating the protein backbone structure. Specifically, Chroma is a diffusion model that can generate protein backbone with desired properties and functions by accepting the classifier guidance. Therefore, Chroma introduces ProCap, a protein-to-text model, as the classifier guidance to enable text-guided protein design. With an additional sequence sampling model, Chroma is capable of sampling both structure and sequence based on the conditions.

In unconditional generation, PAAG only accept the length l of the protein sequence. We also define this is the k -th protein we generated Consequently, we set the template function as “protein number k , contains l amino acids.”

While generate proteins with functional domains, we randomly sample the organism name, length and similarity from natural proteins that have corresponding functional domain. We specify the generative hyper-parameters in A.8. When given property annotations. Deeploc Almagro Armenteros et al. (2017) contains training data and test data in its original split. We utilize template function $G()$ to form the textual descriptions of test data. Then PAAG can generate protein with properties given these descriptions.

A.6 Training configurations

A.6.1 Predictive Task

In downstream prediction tasks, PAAG utilizes the embedding in the aligned space. Specifically, we not only use the protein encoder to extract the protein representations. We also keep the projector head to extract the more informative aligned features. Furthermore, we also incorporate dropout rate and weight decay to improve the generalization ability.

A.6.2 Protein Design with Domain Annotations

While training PAAG with ProtAnnotation, we introduce momentum contrast He et al. (2020), and we set the queue size as 16384, and momentum as 0.995. In the contrastive learning, we initialize the learnable temperature as 0.07. Moreover, we align the our latent space in 256 dimension with one linear layer. Also, following the setting of Li et al. (2022), to enable soft labeling, we initialize the learnable alpha as 0.4. When fine-tuning on the dataset only contains zinc-finger and immunoglobulin

domains, we keep these hyper-parameters but decrease queue size to 4096 since the number training data also decreases.

Besides, we also set the weight decay as 0.05 for AdamW during the learning.

A.6.3 Protein Design with Property Annotations

Due to the scarcity of training data, we conduct fine-tuning process based on the models trained in Section 4.4 for 100 epochs. We mostly keep hyper-parameters as training on ProtAnnotation but decrease the learning rate to 1e-5 without warming up.

For binary localization properties, soluble and membrane-bound, we set template function $G()$ as “is soluble” or “is membrane-bound”. For 10-class subcellular localization, $G()$ is “locates in {location}”.

Table 7: Configurations of full-model tuning of PAAG-ProtBert on three task types. *Abbr.*, lr.: learning rate; lrr.: learning rate ratio; dr.: dropout rate; wd.: weight decay; bs.: batch size; MSE: mean squared error; CE: cross entropy; BCE: binary cross entropy.

Task	optimizer	lr.	lrr.	dr.	wd.	bs.	#epochs	loss
Localization								
Bin	Adam	2.0×10^{-5}	0.15	0	0	4	100	BCE
Sub	Adam	2.0×10^{-5}	0.15	0.1	0	4	100	CE
Fitness								
ρ-lac	Adam	2.0×10^{-4}	0.02	0	3.0×10^{-4}	36	100	MSE
AAV	Adam	9.0×10^{-4}	0.02	0	0	36	100	MSE
Thermo	Adam	2.0×10^{-4}	0.02	0	0	6	100	MSE
Flu	Adam	2.0×10^{-4}	0.02	0.3	4.0×10^{-4}	36	100	MSE
Sta	Adam	5.0×10^{-4}	0.02	0.7	0	36	100	MSE

Table 8: Configurations of full-model tuning of PAAG-ESM-2 on three task types. *Abbr.*, lr.: learning rate; lrr.: learning rate ratio; dr.: dropout rate; wd.: weight decay; bs.: batch size; MSE: mean squared error; CE: cross entropy; BCE: binary cross entropy.

Task	optimizer	lr.	lrr.	dr.	wd.	bs.	#epochs	loss
Localization								
Bin	Adam	1.0×10^{-5}	0.15	0	0	4	100	BCE
Sub	Adam	2.0×10^{-5}	0.15	0	0	4	100	CE
Fitness								
ρ-lac	Adam	1.0×10^{-4}	0.02	0	0	16	100	MSE
AAV	Adam	8.0×10^{-4}	0.02	0	0	16	100	MSE
Thermo	Adam	3.0×10^{-4}	0.02	0.6	0	2	100	MSE
Flu	Adam	3.0×10^{-4}	0.02	0	0	16	100	MSE
Sta	Adam	8.0×10^{-5}	0.02	0	0	16	100	MSE

A.7 Additional experimental results

A.7.1 Results of protein generation with 10-class subcellular location annotation

We represent the confusion matrix of 10-class subcellular location annotation result in Figure 8 for clearer illustration.

Cell membrane	25	40	78	40	2	25	5	5	0	2
Cytoplasm	1	317	12	26	1	1	16	43	0	3
Endoplasmic reticulum	4	56	59	22	3	10	3	2	0	1
Extracellular	3	62	25	264	3	7	7	5	0	0
Golgi apparatus	1	25	20	6	5	3	2	1	0	0
Lysosome/Vacuole	2	13	16	10	1	12	0	3	0	0
Mitochondrion	2	142	19	22	2	3	73	19	0	16
Nucleus	1	253	8	13	0	0	7	379	0	2
Peroxisome	0	20	4	2	0	0	1	1	0	0
Plastid	1	38	8	12	0	1	8	3	0	76
Cell membrane		Cytoplasm	Endoplasmic reticulum	Extracellular	Golgi apparatus	Lysosome/Vacuole	Mitochondrion	Nucleus	Peroxisome	Plastid

Figure 8: Confusion matrix of protein design within 10 subcellular localization annotations. Numbers indicate the count of generated instances given the condition annotation in the column while predicted as row annotation.

A.8 Hyper-parameters

A.8.1 Hyper-parameters of generative task

In our generative task, we utilize nucleus sampling in our decoder. Specifically, we will sample the amino acids based on their probability instead of always choosing the amino acids with the highest probability. We set the Top-p hyperparameter as 0.9 in nucleus sampling, which means the decoder will only consider the most likely amino acids that make up 90% of the probability mass. Besides, we set the repetition penalty of decoder as 1.2.

A.8.2 Hyper-parameters of the predictive task

The hyperparameters for the predictive task concerning PAAG-ProtBert and PAAG-ESM-2 can be found in Table 7 and Table 8, respectively.