DEMYSTIFY THE SECRET FUNCTION IN PROTEIN SEQUENCE VIA CONDITIONAL DIFFUSION MODELS

[1]Yaoyao Xu, [3]Xuxi Chen, [3]Tong Wang, [4]Huan He, [2,5]Tianlong Chen,* [5]Manolis Kellis
[1]The Chinese University of Hong Kong, Shenzhen
[2]University of North Carolina at Chapel Hill [3]University of Texas at Austin
[4]University of Pennsylvania [5]Massachusetts Institute of Technology
xuyaoyao@cuhk.edu.cn, {xxchen,tong.wang}@utexas.edu
Huan.He@pennmedicine.upenn.edu
tianlong@cs.unc.edu, manoli@mit.edu

ABSTRACT

Generating accurate functional annotations for protein sequences presents a significant challenge, especially when dealing with lengthy captions that contain concise descriptions. Recent advancements in diffusion models have shown impressive empirical performance in sequence-to-sequence generation tasks. In this paper, we propose **ProCDM**, a conditional diffusion generative model that utilizes protein sequence representations to generate functional descriptions for proteins. **ProCDM** employs a contrastive learning framework to extract and align protein embeddings with their functionality and then generates functional descriptions by denoising within the continuous embedding space. Our approach, **ProCDM**, demonstrates the capability to generate a wide range of functional descriptions for proteins that align with their actual functionality. Comprehensive experiments are conducted on the EC-Caption datasets to evaluate the effectiveness of our proposal.

1 INTRODUCTION

Accurately predicting protein functions from protein sequences is a crucial task. Traditionally, it relied on extensive biochemical experiments (Thompson et al., 2020). Despite the effort, the experimental annotations available for these sequences remain scarce, comprising less than 1% of the total (Dohan et al., 2021). In recent years, deep learning-based methods have emerged as powerful tools for protein functional annotation (Sanderson et al., 2023; Yu et al., 2023).

In this work, we aim to provide a textual description of unseen protein's functions. Current deep learning methods only predict categorical labels, which only capture a small fraction of the information about protein function, and hinder a comprehensive understanding of protein functionalities. As a result, non-experts face challenges when attempting to utilize these sparse annotations to characterize proteins comprehensively and accurately predict their properties. However, by harnessing the power of natural language, we can offer more detailed descriptions that are easier to understand compared to task-specific predictions and concise annotations.

We present a novel framework, referred to as **ProCDM**, that focuses on encoding protein information and generating comprehensive textual descriptions. To achieve this, we utilize a contrastive learning approach on a pretrained ESM model (Lin et al., 2023a) to encode amino acid sequences of proteins and align them with their functional annotations, specifically the EC number, which is the numerical code used to classify and categorize enzymes based on the reactions they catalyze (Bairoch, 2000). We employ a classifier-free conditional diffusion model inspired by the DDPM framework to generate textual descriptions for protein functions, which iteratively eliminates noise from a Gaussian distribution and results in concise and coherent descriptions. To facilitate the training and evaluation of protein captioning models, we introduce a protein-caption dataset called EC-Caption. This dataset consists of approximately 144k protein-text pairs, encompassing a wide range of proteins

*Correspondence

with diverse functions. We showcase the effectiveness of our proposed approach through extensive experiments conducted on the EC-Caption dataset. Experimental results showcase the effectiveness of **ProCDM** in generating protein descriptions.

2 Problem Formulation and Preliminaries

Extensive research has been carried out on the utilization of diffusion models in sequence-to-sequence generation tasks. In this study, we delve into the application of diffusion models specifically to the sequences of proteins. Our objective is to develop a model capable of generating **textual description** that captures the functional characteristics of proteins. Unlike traditional diffusion models, our approach focuses primarily on tackling the encoding of biological information pertaining to proteins and ensuring its effectiveness.

Preliminaries. Diffusion models have achieved remarkable success in generating high-quality samples across various modalities, including images, videos, and texts. As a score-based generative model falling under the category of latent variable generative models, the diffusion model consists of three fundamental components: the forward process, the reverse process, and the sampling procedure (Chang et al., 2023). The key objective of diffusion models is to gain insights into the probability distribution that underlies a given dataset. This is accomplished by capturing the latent structure of the data through modeling the diffusion process experienced by the data points within their latent space (Song et al., 2020; Ho & Salimans, 2022).

Forward Process. Given $\boldsymbol{x}_0$ sampled from a data distribution $q(\cdot)$, the forward diffusion process aims to add a small amount of Gaussian noise to the sample in T steps, producing a sequence of samples with noises, namely $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$. The step sizes, which are controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^T$. The forward process process could be modeled as Equation 1:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{1}$$

The data sample $\mathbf{x}_0$ gradually loses its distinguishable features as the step t becomes larger. Eventually when $T \to \infty$, $\mathbf{x}_T$ is equivalent to an isotropic Gaussian distribution.

Backward process. We will be able to recreate the true sample from a Gaussian noise input $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, by reversing the above forward process and sample from $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$. we train a model p_θ to approximate these conditional probabilities to run the reverse diffusion process, and the sequential reverse process could be modeled as Equation 2:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{2}$$

DiffuSeq (Gong et al., 2023) is a diffusion model designed for text generation tasks such as open-ended sentence generation in a seq2seq framework. Unlike traditional approaches such as, (Dhariwal & Nichol, 2021; Kim et al., 2022; Shi et al., 2023) that rely on separate classifiers, it is a classifier-free framework that directly predicts the conditional probability of generating a target sentence based on the context. This unique paradigm allows for comprehensively capturing the data distribution and effectively utilizing conditional guidance.

Protein Functional Annotations. Given a protein sequence (denoted as $\mathbf{w}_{\text{seq}}$) and a textual description of the protein's function (denoted as $\mathbf{w}_{\text{func}}$), our objective is to train a model that takes $\mathbf{w}_{\text{seq}}$ as input and generates a sentence that accurately describes its functionality, closely aligning with $\mathbf{w}_{\text{func}}$. f_θ is denoted as our transformer model which takes the noised input and predicts the denoised output in their embedding space.

3 ProCDM: A Diffusion Framework for Protein Caption

We present **ProCDM**, a framework that focuses on generating protein functional description conditioning on their sequences. The overall framework is shown in Figure 1.

Forward Process. For the forward process, **ProCDM** follows the protocol of Diffusion LM (Li et al., 2022) to map the protein sequences (as "source" sequences) and corresponding text descriptions of

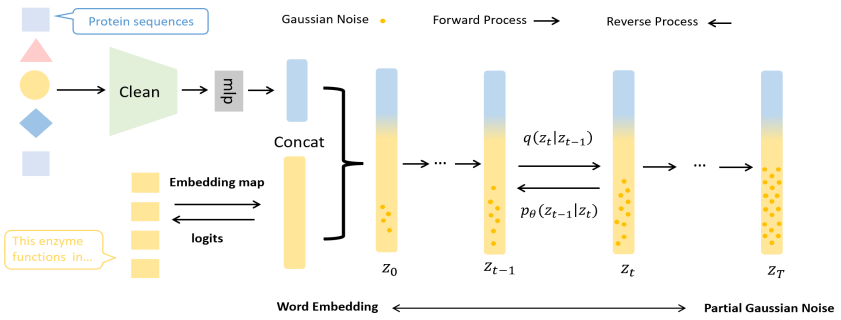


Figure 1: The conditional diffusion process of our model. The source protein sequences are converted into the target continuous space using a functional embedding extractor. Simultaneously, the textual description of protein functions undergoes transformation using a separate embedding layer. The partial Gaussian noise is iteratively added on the continuous space $\mathbf{z_t}$.

their functions (as "target" sequences) from discrete to continuous embedding space. **ProCDM** employs different embedding mapping functions for protein sequences and their functional descriptions due to their differing modalities. For protein sequences, we leverage CLEAN (Yu et al., 2023), a contrastive learning framework for protein functional annotation, to train a feature extractor and obtain the embeddings (denoted as $\mathbf{E}_{\text{seq}}$) of protein sequences $\mathbf{w}_{\text{seq}}$. This allows for better encoding of functionally related information (refer to Section 4.2) thereby improving the quality of generated textual descriptions. As for the functional description, we use a learnable word embedding function to extract their embeddings (denoted as $\mathbf{E}_{\text{func}}$) from textual description $\mathbf{w}_{\text{func}}$.

After obtaining the respective embeddings of proteins' sequences and functional description, $\mathbf{E}_{\text{seq}}$ and $\mathbf{E}_{\text{func}}$ are concatenated as a unified embedding, referred to as $\mathbf{E}$, which is subsequently fed as the input to the forward process of the diffusion model. This extension of the original forward chain to a new Markov transition, $q_\phi\left(\mathbf{z_0}|\mathbf{w}^{\text{seq}\oplus\text{fun}}\right) = \mathcal{N}\left(\mathbf{E}, \beta_0\mathbf{I}\right)$, enables the adaptation of discrete inputs. For brevity, we simplify the notation using $\mathbf{z_t} = \mathbf{E}^{\mathbf{t}}_{\text{seq}} \oplus \mathbf{E}^{\mathbf{t}}_{\text{func}}$, where $\mathbf{E}^{\mathbf{t}}_{\text{seq}}$ and $\mathbf{E}^{\mathbf{t}}_{\text{func}}$ represent the hidden state of the protein sequences and the functional description at time step t, respectively. To achieve the final distribution, **ProCDM** introduce Gaussian noise through the conditional diffusion forward process $q_\phi\left(\mathbf{z_t}|\mathbf{z_{t-1}}\right)$, which **selectively injects noises only into $\mathbf{E}^{\mathbf{t}}_{\text{func}}$ at each time step**.

Conditional Denoising Process. In the denoising process, **ProCDM** iteratively removes the noise at each time step, aiming to recover $\mathbf{E}^{\mathbf{t}}_{\text{func}}$ to $\mathbf{E}^{\mathbf{0}}_{\text{func}}$. Specifically, we train f_θ to model the conditional denoising process in a classifier-free manner following the settings of DiffuSeq (Gong et al., 2023). The training objective $\mathcal{L}_{\mathbf{VLB}}$ is defined as follows:

$$\begin{aligned}\min_\theta \mathcal{L}_{\mathbf{VLB}} &= \min_\theta \left[\sum_{t=2}^{T} \|\mathbf{z_0} - f_\theta\left(\mathbf{z_t}, t\right)\|^2 + \|\mathbf{E} - f_\theta\left(\mathbf{z_1}, 1\right)\|^2 - \log p_\theta\left(\mathbf{w}^{\text{seq}\oplus\text{fun}}|\mathbf{z_0}\right)\right] \\ &\rightarrow \min_\theta \left[\sum_{t=2}^{T} \|\mathbf{E}^{\mathbf{0}}_{\text{func}} - \tilde{f}_\theta\left(\mathbf{z_t}, t\right)\|^2 + \|\mathbf{E}_{\text{func}} - \tilde{f}_\theta\left(\mathbf{z_1}, 1\right)\|^2 + \mathcal{R}\left(\|\mathbf{z_0}\|^2\right)\right],\end{aligned} \tag{3}$$

where $p_\theta(\mathbf{z_t}|\mathbf{z_{t-1}}) = \mathcal{N}(\mathbf{z_{t-1}}; \mu_\theta(\mathbf{z_t}, t), \sigma_\theta(\mathbf{z_t}, t))$. $\mu_\theta\left(\cdot\right)$ and $\sigma_\theta\left(\cdot\right)$ is the parameterization of the predicted mean and standard deviation of $q\left(z_{t-1}|z_t\right)$ in forward process. In Equation 3, $\tilde{f}_\theta\left(\mathbf{z_t}, t\right)$ represents our recovered $\mathbf{E}^{\mathbf{0}}_{\text{func}}$, and $\mathcal{R}\left(\|\mathbf{z_0}\|^2\right)$ represents the regularization on embedding learning. We use the different embedding functions for the source and target and apply an MLP layer to map the source to the target's embedding space to match their dimension.

Training and Inference Sampling Method. To mitigate inadequate training caused by prolonged diffusion steps, **ProCDM** adopts the importance sampling technique (Nichol & Dhariwal, 2021) as the alternative to uniform time-step sampling. The formulation of its loss objective is subsequently expressed as $\mathcal{L}_{\text{vlb}} = \mathbb{E}_{t\sim p_t}\left[\frac{L_t}{p_t}\right]$, where $p_t \propto \sqrt{\mathbb{E}\left[L_t^2\right]}$ and $\sum p_t = 1$. Samples are generated by conditioning on the random variable $\mathbf{z}_T$, drawn from a standard normal distribution $\mathcal{N}(0, I)$. Our experiments indicate that integrating the rounding technique (Gong et al., 2023) into the target embedding during each denoising step has led to the emergence of duplicated sequences. Consequently,

in **ProCDM**, we opt against employing the rounding process. Instead, we extend the denoising process to maintain the diversity within the generated descriptions of protein functions.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Tasks. We collect proteins' sequences and their corresponding functional description from the Kyoto Encyclopedia of Genes and Genomes (KEGG) dataset [1]. We randomly select seven thousand protein-description pairs for inclusion in our study, with a train/valid/test split of 60%/20%/20%. We set two tasks for evaluating the performance of **ProCDM**: *(1) Long Description Generation (Long for short)*, where the model is instructed to generate a functional description that has maximally 50 words; and *(2) Short Description Generation)*, where the model is required to generate description that has no more than 12 words. To generate the ground-truth label, we summarize the crawled original functional description into short sequences with GPT-4. More details of the summarization process can be found in Appendix B.2.

Baselines. We compare the performance of **ProCDM** with both autoregressive (AR) and non-autoregressive (NAR) models that are commonly employed for sequence-to-sequence generation tasks. For the AR models, we use the encoder-decoder architecture of the Transformer (Vaswani et al., 2017) and LSTM (Bahdanau et al., 2016), which are extensively employed for machine translation tasks. As for the NAR models, we employ a fine-tuned large pre-trained language model (PLM), specifically GPT-2 (Radford et al., 2019), known for its remarkable success across various sequence-to-sequence tasks. Additionally, we perform experiments on baseline models using various pretrained feature extractors to obtain protein embeddings (*i.e.*, CLEAN, and ESM-2) and compare their performances with randomly initialized feature extractors. Detailed implementations can be found in Appendix B.3.

Evaluation Metrics. We assess the generated sequences focus on two aspects, namely the quality and diversity of the sequences. To evaluate the quality, we employ established metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores. However, it is important to note that metrics solely based on string similarity may not be optimal for open-ended generation scenarios. Therefore, we incorporate BERTScore (Zhang et al., 2019) as an additional metric for assessment. Additionally, we measure the diversity aspect using the dist-1 metric, which quantifies the intra-diversity within each generated sentence by quantifying the occurrence of repeated words.

Implementation. Our model is based on `BERT-base-uncased` (Devlin et al., 2018), where the time-step embedding is plugged akin to the position embedding. The maximum protein sequence length is 300, with an embedding dimension d of 128. We set the diffusion steps T to be 2000, with a square-root noise schedule. Due to the presence of biological words, we use the vocabulary list and tokenizer of `PubmedBERT` (Gu et al., 2020) in our model for better handling. We generate embeddings for protein sequences using ESM-2 with 150M number of parameters.

4.2 EXPERIMENTS RESULTS

Based on the results presented in Table 1, our model demonstrates superior performance in generating diverse and high-quality text captions for protein sequences across all four tasks compared to baseline models. Table 3 in the appendix presents generated captions and reference captions for diverse protein sequences produced by our model, along with the corresponding UniProt ID indicating the protein category. The findings indicate that utilizing fixed CLEAN embedding for protein sequences leads to better outcomes than fine-tuning ESM for protein sequence embedding. These results suggest the necessity of functional alignment between protein and text modalities to obtain a distribution of text captions conditioned on protein sequences.

Figure 3 displays the t-SNE visualization of 800 pairs of protein sequences' text captions from the testing set, along with the corresponding captions generated by our model in the embedding space. The clusters correspond to distinct protein sequences that align with identical or similar functional text descriptions, often observed among proteins sharing the same EC number. The figure highlights

[1] https://www.kegg.jp/kegg/annotation/enzyme.html

Task	Method	BLEU	R-L	Score	dist-1	Avg. Len
Long Caption	**ProCDM** (CLEAN)	**0.149**	**0.262**	**0.614**	0.838	36.14
	ProCDM (ESM-2)	0.127	0.199	0.540	0.897	48.26
	LSTM	0.018	0.150	0.430	0.287	25.4
	Transformer	0.079	0.208	0.570	0.900	26.29
	GPT-2	0.034	0.130	0.546	**0.930**	21.2
Summary	**ProCDM** (CLEAN)	0.182	0.267	**0.681**	**0.940**	14.55
	ProCDM (ESM-2)	**0.227**	**0.302**	0.674	0.833	14.76
	LSTM	0.068	0.201	0.591	0.319	14.61
	Transformer	0.000	0.000	0.339	0.077	15
	GPT-2	0.013	0.061	0.494	0.919	7.455
Short Caption	**ProCDM** (CLEAN)	**0.169**	**0.336**	**0.620**	0.960	11.99
	LSTM	0.080	0.289	0.589	0.930	12
	GPT-2	0.013	0.081	0.546	**0.970**	10.20
EC Number	**ProCDM** (CLEAN)	**0.134**	**0.452**	**0.786**	0.870	6
	LSTM	0.001	0.004	0.319	0.750	6
	GPT-2	0.096	0.326	0.752	**0.883**	3.98

Table 1: The overall results of different methods on different captioning tasks. **ProCDM** (CLEAN) represents using fixed protein representation extracted from CLEAN in **ProCDM**, **ProCDM** (ESM-2) represents fine-tuned representation extracted from ESM-2 in **ProCDM**. The best results are bold.

the difficulty faced by the fine-tuned ESM in capturing the data distribution of captions. To address this challenge, we experimented with incorporating an additional MLP layer after the ESM encoder and applying L2 norm loss to align the protein and text distributions in the embedding space during training. However, the results indicate that directly learning the alignment between modalities under the diffusion process remains challenging, so we choose to use CLEAN which gives the aligned representation by pretraining. The generated caption samples can be found in Table 3. To enhance the accuracy assessment of our model, we randomly choose two protein sequences that share similar functions but have low alignment scores. The alignment between these proteins, as depicted in Figure 2, indicates a sequence identity of merely 34% based on MMseqs2 (Mirdita et al., 2021). **ProCDM** generates identical descriptions for these two proteins, aligning with their recorded functions in the UniProt database. Notably, our model with diffusion outperforms the baseline AR and NAR models in terms of generating more diverse sentences with fewer repeated words.

Open sequence generation tasks often encounter challenges related to the limited diversity of generated text and an increased occurrence of repetitive words. Sequence-based diffusion models commonly rely on two mechanisms: denoising the discrete space (tokens) and the continuous space (word embedding). Inadequate training in either of these mechanisms can result in degenerate solutions, such as generating a substantial number of repeated words or sentences, and lead to a low diversity score as shown in Figure 5. This issue becomes particularly pronounced in sequence diffusion within the word embedding space. Besides, conditional diffusion also faces the problem that the diffusion process could neglect source conditions in training (Ye et al., 2023). Several approaches have been proposed to address this issue. These solutions include techniques such as noise manipulation, integration of anchor loss within the denoising process, and word embedding normalization (Gao et al., 2022). In our experiments, we observe that the condition's distribution plays a crucial role in guiding the diffusion model to generate data that aligns with the desired distribution. It influences the effectiveness of generated captions based on the given conditions with the same sampling algorithm. Specifically, when the condition distribution is highly uniform, it poses challenges in effectively guiding the denoising process toward generating captions that satisfy the conditional probability. Figure 4 illustrates that the protein sequences embedding obtained from CLEAN exhibit more prominent clusters compared to those extracted from the ESM encoder. Consequently, utilizing embeddings from CLEAN is expected to yield superior performance in caption generation, even when employing the same implementation.

5 Conclusion

This paper addresses the challenging task of generating accurate functional descriptions of proteins' functions. We propose **ProCDM**, a conditional diffusion generative model specifically designed for

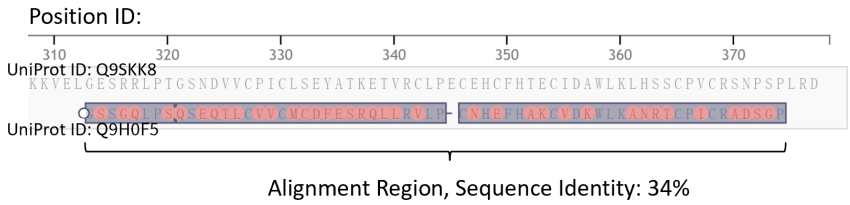


Figure 2: Two different Protein sequences that have the same functions but with a low alignment score. The highlight region indicates their sequence alignment positions which are marked by UniProt tools with the cut-off to be 0.1. The generated captions from **ProCDM** of these two sequences are the same, indicating they have similar functions even though their alignment score is only 34%.

sequence-to-sequence generation tasks. **ProCDM** introduces a novel approach to leveraging protein sequence representations for generating functional descriptions of proteins. By employing a contrastive learning framework, **ProCDM** effectively extracts and aligns protein embeddings with their corresponding functionality. This alignment process forms the foundation for generating functional descriptions by denoising within the continuous embedding space. The findings presented in this paper contribute valuable insights and pave the way for further advancements in the field of protein sequence captions and bioinformatics research.

REFERENCES

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR 2015*, 2016.

Amos Bairoch. The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305, 2000.

Ziyi Chang, George A Koulieris, and Hubert PH Shum. On the design fundamentals of diffusion models: A survey. *arXiv preprint arXiv:2306.04542*, 2023.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

David Dohan, Andreea Gane, Maxwell L Bileschi, David Belanger, and Lucy Colwell. Improving protein function annotation via unsupervised pre-training: Robustness, efficiency, and insights. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2782–2791, 2021.

Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. Difformer: Empowering diffusion model on embedding space for text generation. *arXiv preprint arXiv:2212.09412*, 2022.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations, ICLR*, 2023.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Heeseung Kim, Sungwon Kim, and Sungroh Yoon. Guided-tts: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning*, pp. 11119–11133. PMLR, 2022.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *Science*, 2023a.

Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In *International Conference on Machine Learning*, pp. 21051–21064. PMLR, 2023b.

Milot Mirdita, Martin Steinegger, F Breitwieser, Johannes Söding, and E Levy Karin. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics*, 37(18):3029–3031, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2002.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Theo Sanderson, Maxwell L Bileschi, David Belanger, and Lucy J Colwell. Proteinfer, deep neural networks for protein functional inference. *Elife*, 12:e80942, 2023.

Changhao Shi, Haomiao Ni, Kai Li, Shaobo Han, Mingfu Liang, and Martin Renqiang Min. Exploring compositional visual generation with latent classifier guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 853–862, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Michael C Thompson, Todd O Yeates, and Jose A Rodriguez. Advances in methods for atomic resolution macromolecular structure determination. *F1000Research*, 9, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. Dinoiser: Diffused conditional sequence learning by manipulating noises. *arXiv preprint arXiv:2302.10025*, 2023.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

A Related Works

A.1 Conditional Diffusion Model

Considerable studies have been conducted on the application of diffusion models for text generation. GENIE (Lin et al., 2023b), for instance, employs a progressive transformation of a random noise sequence into a coherent text sequence. It encourages the diffusion-decoder to reconstruct a text paragraph from a corrupted version. Austin et al. (2021) extended the scope of discrete text diffusion models by introducing the concept of an absorbing state ([MASK]). However, discrete diffusion models face challenges in scaling one-hot row vectors and are limited to generating text samples solely in discrete space without incorporating conditional constraints. In contrast, the Diffusion-LM (Li et al., 2022) and Analog Bits (Chen et al., 2022) models introduce innovative language models that utilize continuous latent representations. These models employ distinct mapping functions to establish a connection between the discrete and continuous realms of text. Additionally, Diffseq (Gong et al., 2023) specializes in SEQ2SEQ tasks for text generation in the continuous space with conditional constraints.

B More Experiment Settings

B.1 Protein Public dataset

We list the datasets we have used in our experiments below:

1. Uniprot, which is available at `https://www.uniprot.org/`.
2. PDB, which is available at `https://www.rcsb.org/`.
3. KEGG (based on EC number), available at `https://www.kegg.jp/kegg/annotation/enzyme.html`.

B.2 Prompts to generate a summary for protein sequences' functional descriptions

We use the following prompt to query GPT-4 to generate the description of functions for our EC-Caption dataset:

> "This is a functional description of a protein. Based on your biological knowledge, please pick the five most important words in this sentence to best summarize the function of this protein. Please separate these five words with a space, just answer the five words. Functional description: xxx xx xxx xx."

B.3 Hyperparameter Setting

We list the hyperparameter we used in Table 2.

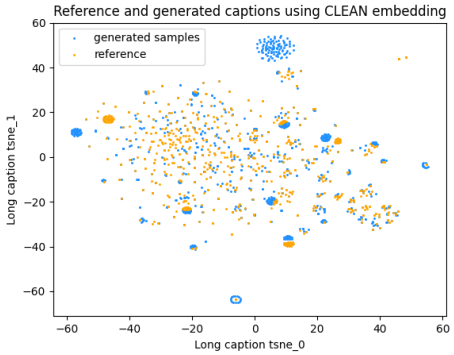
C More Experimental Results

We present the generation performance after different numbers of training steps in Table 5. In Figure 3 and Figure 4 we provide the t-SNE visualization of the embeddings of the description of functions and the protein sequences, respectively.

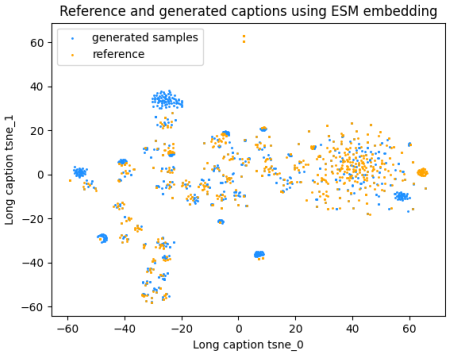
Sampled Generation Results. In Table 3, we present generation results generated by **`ProCDM`** and compare with the references.

Hyperparameter	**ProCDM**	GPT-2	LSTM	Transformer
Learning Rate	1×10^{-4}	5×10^{-4}	1×10^{-4}	5×10^{-4}
Seed	102	42	1234	1234
Hidden Dimension	128	-	512	256
Batch Size	10	10	10	10
Protein Sequence Length	300	300	300	300
Vocab	PubMed	GPT-2	PubMed	PubMed
Init PLM	CLEAN/ESM-2 (150M)	-	CLEAN/ESM-2 (150M)	CLEAN/ESM-2 (150M)
Noise Schedule	Sqrt	-	-	-
Schedule Sampler	Loss-Aware	-	-	-
Diffusion Steps	2000	-	-	-
Learning Steps	50000	10000	10000	20000
Warm-up Steps	-	100	-	-
ϵ	-	1×10^{-8}	-	-
Dropout	-	-	0.5	0.1

Table 2: The hyperparameters adopted by **ProCDM** and baseline models.

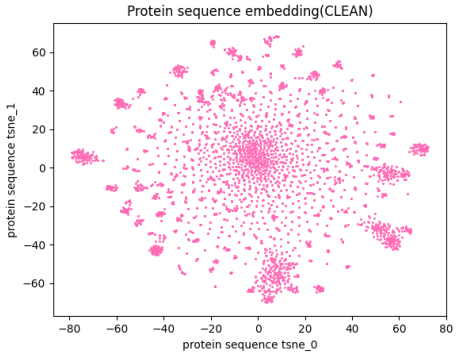


(a) Samples generated by our model using CLEAN's embedding.

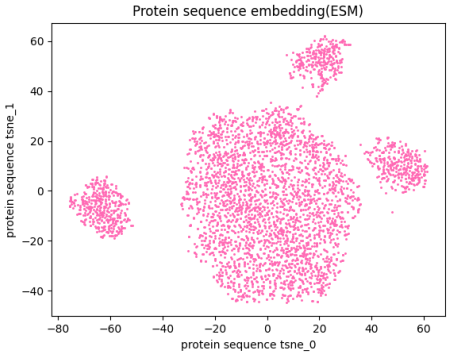


(b) Samples generated by our model fine-tuning ESM's embedding.

Figure 3: t-SNE of textual description's embeddings for samples in the testing dataset. The samples in blue color represent the generated captions' embedding and the samples in yellow represent the reference captions' embedding.



(a) Protein sequences' embedding from CLEAN.



(b) Protein sequences' embedding from ESM-2.

Figure 4: Protein sequences' embedding of training set. Embedding extracted from CLEAN has more clusters which give better conditions in the diffusion process than ESM-2.

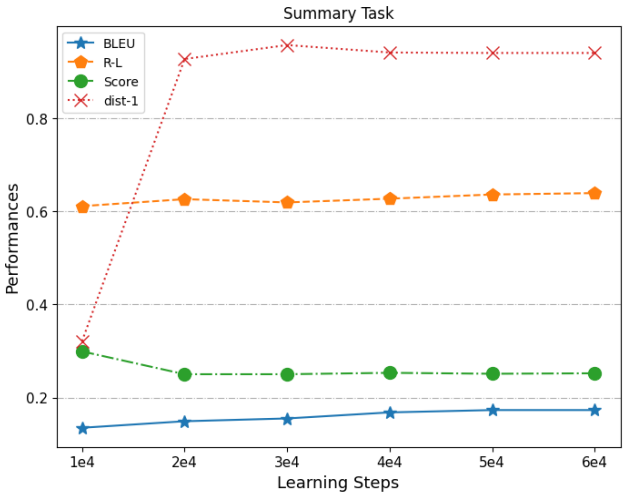


Figure 5: The horizontal axis represents the number of learning steps, and the vertical axis represents the metrics for generating captions. As can be seen from the curve, the overall score gradually increases and converges as the learning steps increase. Based on dist-1, from 10000 to 20000 steps, the diversity of captions generated increases greatly, which indicates that insufficient training will result in a large number of word repetitions.

Samples	**ProCDM**	Reference	UniProt Id
1	”[CLS] ec 3 . 1 . 21 . 4 is a type 2 site - specific deoxyrib onuclease enzyme involved in dna cleavage and is part of a large group of enzymes . [CLS]”	”EC 3.1.21.4 is a type 2 site-specific deoxyribonuclease enzyme involved in DNA cleavage and is part of a large group of enzymes.”	Q07605
2	”[CLS] EC two . 1 . 4 . 46 enzyme a type methylation of sarc which r 1 2 coli converting beta from part coli and phosphate group at fructose form 2 pathway 6 catalyzes membrane by glucose metabolism dimethyl intermediate porphobilinogen regulating the adenosyl”	”EC 2.5.1.61 (hydroxymethylbilane synthase) catalyzes the formation of uroporphyrinogen III from porphobilinogen.”	A8FF13
3	”[CLS] rna helicases utilize atp hydrolysis to unw ind rna with either 3 ’ to 5 ’ or 5 ’ to 3 ’ polarity. some also unw ind dna . they may be identical to dna helicases.[CLS]”	”RNA helicases utilize ATP hydrolysis to unwind RNA with either 3’ to 5’ or 5’ to 3’ polarity. Some also unwind DNA. They may be identical to DNA helicases.”	O45244
4	”[CLS] the enzyme catalyzes biosynthesis use nadh transfer activated ov residues specificity binding c can processes membrane chain pathway eukaryotes position bacteria active group by protons chain activity reaction. [CLS]”	”The enzyme catalyzes the breakdown of (L-cysteinylglycine)-S-conjugate into L-cysteine and glycine, by transferring a L-glutamate molecule to the (L-cysteinylglycine) molecule.	Q6IE08
5	”[CLS] cer ul oplasm in is a multic op per oxidase enzyme that oxid izes fe (ii) to fe (iii) in the blood plasma , allowing for incorporation into proteins . an iron oxidizing bacterium enzyme contains heme.[CLS]”	Ceruloplasmin is a multicopper oxidase enzyme that oxidizes Fe(II) to Fe(III) in the blood plasma, allowing for incorporation into proteins. An iron-oxidizing bacterium enzyme contains heme a.”	P41822

Table 3: Samples generated from the testing set with **ProCDM** and their corresponding protein sequences’ UniProt ID and reference captions.