A Multi-Modal Contrastive Diffusion Model for Therapeutic Peptide Generation

Yongkang Wang[1*], **Xuan Liu**[1*], **Feng Huang**[1], **Zhankun Xiong**[1], **Wen Zhang**[1,2 3†]

[1]College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2]Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan 430070, China
[3]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China
{wyky481, lx666, fhuang233, xiongzk}@webmail.hzau.edu.cn, zhangwen@mail.hzau.edu.cn

Abstract

Therapeutic peptides represent a unique class of pharmaceutical agents crucial for the treatment of human diseases. Recently, deep generative models have exhibited remarkable potential for generating therapeutic peptides, but they only utilize sequence or structure information alone, which hinders the performance in generation. In this study, we propose a Multi-Modal Contrastive Diffusion model (MMCD), fusing both sequence and structure modalities in a diffusion framework to co-generate novel peptide sequences and structures. Specifically, MMCD constructs the sequence-modal and structure-modal diffusion models, respectively, and devises a multi-modal contrastive learning strategy with inter-contrastive and intra-contrastive in each diffusion timestep, aiming to capture the consistency between two modalities and boost model performance. The inter-contrastive aligns sequences and structures of peptides by maximizing the agreement of their embeddings, while the intra-contrastive differentiates therapeutic and non-therapeutic peptides by maximizing the disagreement of their sequence/structure embeddings simultaneously. The extensive experiments demonstrate that MMCD performs better than other state-of-the-art deep generative methods in generating therapeutic peptides across various metrics, including antimicrobial/anticancer score, diversity, and peptide-docking.

Introduction

Therapeutic peptides, such as antimicrobial and anticancer peptides, are a unique class of pharmaceutical agents that comprise short chains of amino acids, exhibiting significant potential in treating complex human diseases (Jakubczyk et al. 2020). Traditionally, therapeutic peptides are discovered through a comprehensive screening of sequence spaces using phage/yeast display technologies (Muttenthaler et al. 2021) or computational tools trained for scoring desired properties (Lee et al. 2017; Lee, Wong, and Ferguson 2018). However, the combinatorial space of possible peptides is vast and only a small solution satisfies therapeutic requirements; thus, such screening methods based on brute force can be time-consuming and costly.

In recent years, deep generative models (DGMs) have demonstrated success in generating images (Liu and Chilton 2022), texts (Iqbal and Qureshi 2022), proteins (Wu et al. 2021), and also gained popularity in peptides. DGMs explored a more expansive chemical space that affords the creation of structurally novel peptides, by training neural networks to approximate the underlying distribution of observed or known ones (Wan, Kontogiorgos, and Fuente 2022). For example, autoregression-based methods depicted peptide sequences as sentences composed of residue tokens, so that the problem can be solved by predicting residue arrangement via recurrent neural networks (RNN) (Müller, Hiss, and Schneider 2018; Capecchi et al. 2021). Variational autoencoder (VAE)-based methods generated new peptide sequences by sampling from the latent space learned through an encoder-decoder architecture, with or without therapeutic properties as conditional constraints (Ghorbani et al. 2022; Szymczak et al. 2023b). Generative adversarial network (GAN)-based methods trained the generator and discriminator using known data, which compete against each other to generate new peptides (Tucs et al. 2020; Oort et al. 2021; Lin, Lin, and Lane 2022). Nowadays, diffusion models (Yang et al. 2023) are prevalent in the generation of protein sequences and structures, owing to their superior capability in fitting distributions compared to prior techniques (Shi et al. 2023; Wu et al. 2022). Likewise, these advanced diffusion models can be extended to peptide generation and are expected to deliver favorable outcomes.

Despite the commendable progress of efforts above, they focused on generating either sequences (i.e., residue arrangements) or structures (i.e., spatial coordinates of backbone atoms), ignoring that models fusing information from both modalities may outperform their uni-modal counterparts (Huang et al. 2021). However, how to effectively integrate the multi-modal information and capture their consistency in peptide generation is a major challenge. Additionally, compared with generation tasks for images, texts, and proteins that involve millions of labeled samples, public datasets for therapeutic peptides typically contain only thousands of sequence or structure profiles, induced by the high cost of *in vitro* screening. This limited amount of available data may result in overfitting (Webster et al. 2019), which confines generated outcomes within a restricted distribution, consequently compromising the model's generalization abil-

[*]These authors contributed equally.
[†]Corresponding authors.

ity. How to fully leverage existing peptide data, such as therapeutic and non-therapeutic peptides, to enhance the generation performance could be regarded as another challenge.

To address these challenges, we propose a **M**ulti-**M**odal **C**ontrastive **D**iffusion model for therapeutic peptide generation, named **MMCD**. Specifically, we build a multi-modal framework that integrates sequence-modal and structure-modal diffusion models for co-generating residue arrangements and backbone coordinates of peptides. To ensure consistency between the two modalities during the generation process, we bring in an inter-modal contrastive learning (Inter-CL) strategy. Inter-CL aligns sequences and structures, by maximizing the agreement between their embeddings derived from the same peptides at each diffusion timestep. Meanwhile, to avoid the issue of inferior performance caused by limited therapeutic peptide data, we incorporate substantial known non-therapeutic peptides as data augmentations to devise an intra-modal CL (Intra-CL). Intra-CL differentiates therapeutic and non-therapeutic peptides by maximizing the disagreement of their sequence/structure embeddings at each diffusion timestep, driving the model to precisely fit the distribution of therapeutic peptides. Overall, the main contributions of this work are described as follows:

- We propose a multi-modal diffusion model that integrates both sequence and structure information to co-generate residue arrangements and backbone coordinates of therapeutic peptides, whereas previous works focused only on a single modality.
- We design the inter-intra CL strategy at each diffusion timestep, which aims to maximize the agreement between sequence and structure embeddings for aligning multi-modal information, and maximize the disagreement between therapeutic and non-therapeutic peptides for boosting model generalization.
- Extensive experiments conducted on peptide datasets demonstrate that MMCD surpasses the current state-of-the-art baselines in generating therapeutic peptides, particularly in terms of antimicrobial/anticancer score, diversity, and pathogen-docking.

Related Works

Diffusion Model for Protein Generation

Diffusion models (Song and Ermon 2019; Trippe et al. 2023) are devoted to learning the noise that adequately destroys the source data and iteratively remove noise from the prior distribution to generate new samples, which have emerged as cutting-edge methods for numerous generation tasks, especially in proteins (Wu et al. 2022; Cao et al. 2023). For example, Liu et al. (2023) proposed a textual conditionally guided diffusion model for sequence generation. Hoogeboom et al. (2022) introduced ProtDiff with an E(3) equivariant graph neural network to learn a diverse distribution over backbone coordinates of structures. Luo et al. (2022) considered both the position and orientation of antibody residues, achieving an equivariant diffusion model for sequence-structure co-generation. Despite their success, the fusion of both sequence and structure modalities in diffusion models has not been comprehensively investigated, and their potential for peptide generation remains unexplored. To fill this gap, we implement a peptide-oriented diffusion model capable of sequence-structure co-generation and multi-modal data fusion.

Contrastive Learning

Being popular in self-supervised learning, contrastive learning (CL) allows models to learn the knowledge behind data without explicit labels (Xia et al. 2022; Zhu et al. 2023). It aims to bring an anchor (i.e., data sample) closer to a positive/similar instance and away from many negative/dissimilar instances, by optimizing their mutual information in the embedding space. Strategies to yield the positive and negative pairs often dominate the model performance (Zhang et al. 2022). For example, Yuan et al. (2021) proposed a multi-modal CL to align text and image data, which encourages the agreement of corresponding text-image pairs (positive) to be greater than those of all non-corresponding pairs (negative). Wu, Luu, and Dong (2022) designed a CL framework that makes full use of semantic relations among text samples via efficient positive and negative sampling strategies, to mitigate data sparsity for short text modeling. Zhang et al. (2023b) augmented the protein structures using different conformers, and maximized the agreement/disagreement between the learned embeddings of same/different proteins, aiming to learn more discriminative representations. However, these CL strategies have yet to be extended to peptide-related studies. Therefore, we devise the novel CL strategy in peptide generation, which serves as an auxiliary objective to enforce sequence-structure alignment and boost model performance.

Methodology

In this section, we formulate the peptide co-generation problem for sequence and structure. Subsequently, we elaborately enumerate the components of our method MMCD, including the diffusion model for peptide generation and the multi-modal contrastive learning strategy. The overview of MMCD is illustrated in Figure 1.

Problem Formulation

A peptide with N residues (amino acids) can be represented as a sequence-structure tuple, denoted as $X = (S, C)$. $S = [s_i]_{i=1}^{N}$ stands for the sequence with $s_i \in \{ACDEFGHIKLMNPQRSTVWY\}$ as the type of the i-th residue, and $C = [c_i]_{i=1}^{N}$ stands for the structure with $c_i \in \mathbb{R}^{3*4}$ as Cartesian coordinates of the i-th residue (involving four backbone atoms N-C$_\alpha$-C-O). Our goal is to model the joint distribution of X based on the known peptide data, so that sequences (i.e., residue types) and structures (i.e., residue coordinates) of new peptides can be co-generated by sampling the distribution.

Diffusion Model for Peptide Generation

The diffusion model defines the Markov chains of processes, in which latent variables are encoded by a *forward diffusion process* and decoded by a *reverse generative process*

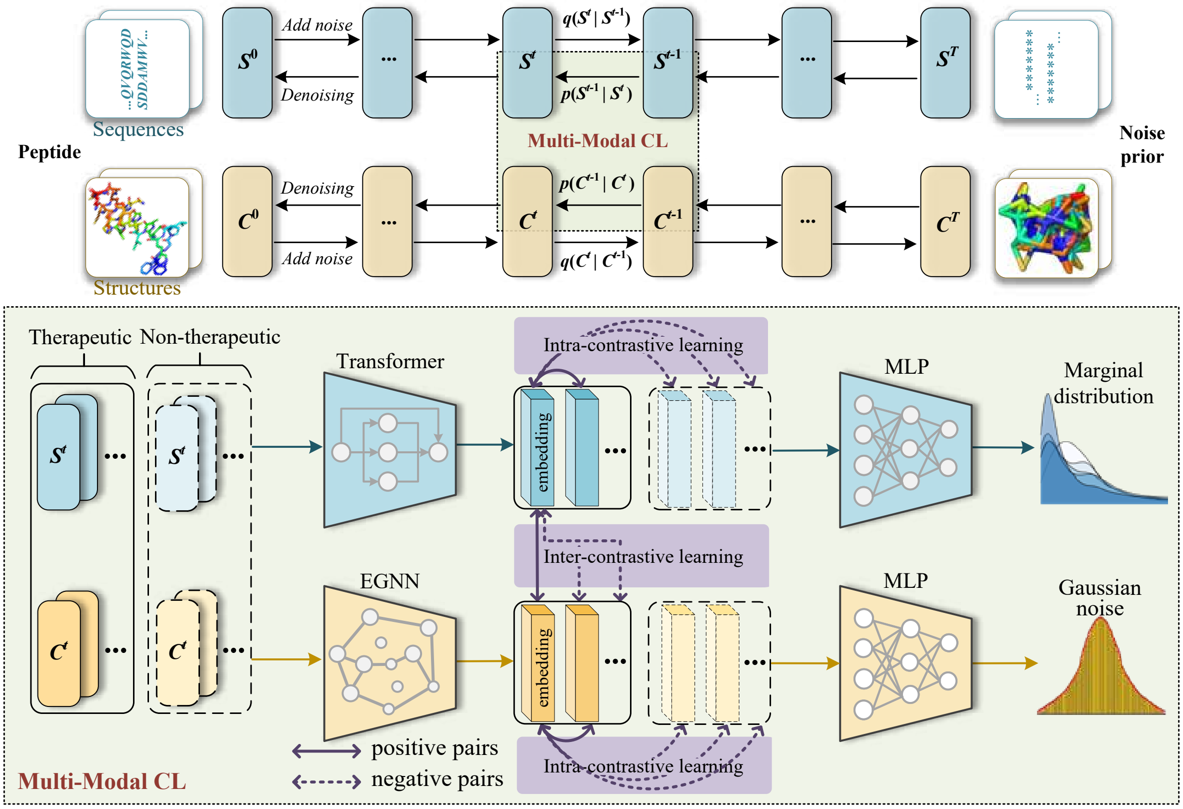

Figure 1: Overview of the MMCD. MMCD consists of a diffusion model for the peptide sequence-structure co-generation and multi-modal contrastive learning (CL). The diffusion model involves a forward process ($q(\cdot|\cdot)$) for adding noise and a reverse process ($p(\cdot|\cdot)$) for denoising at each timestep t. The reverse process utilizes a transformer encoder (or EGNN) to extract embeddings from sequences S (or structures C), and a sequence (or structure)-based MLP to map embeddings to the marginal distribution (or Gaussian) noise. The multi-modal CL includes an Inter-CL and an Intra-CL, which aims to align sequence and structure embeddings, and differentiate therapeutic and non-therapeutic peptide embeddings.

(Sohl-Dickstein et al. 2015). Let $X^0 = (S^0, C^0)$ denotes the ground-truth peptide and $X^t = (S^t, C^t)$ for $t = 1, ..., T$ to be the latent variable at timestep t. The peptide generation can be modeled as an evolving thermodynamic system, where the forward process $q(X^t|X^{t-1})$ gradually injects small noise to the data X^0 until reaching a random noise distribution at timestep T, and the reverse process $p_\theta(X^{t-1}|X^t)$ with learnable parameters θ learns to denoise the latent variable X^t towards the data distribution (Luo et al. 2022).

Diffusion for Peptide Sequence. Following Anand and Achim (2022), we treat residue types as categorical data and apply discrete diffusion to sequences, where each residue type is characterized using one-hot encoding with 20 types. For the forward process, we add noise to residue types using the transition matrices with the marginal distribution (Austin et al. 2021; Vignac et al. 2023) (see details in Appendix A). For the reverse process, the diffusion trajectory is parameterized by the probability $q(S^{t-1} \mid S^t, \hat{S}^0)$ and a network $\hat{p}_\theta$ is defined to predict the probability of S^0 (Austin et al. 2021), that is:

$$p_\theta\left(S^{t-1} \mid S^t\right) = \prod_{1 \le i \le N} q(s_i^{t-1} \mid S^t, \hat{S}^0) \cdot \hat{p}_\theta(\hat{S}^0 \mid S^t) \quad (1)$$

where s_i^t denotes the one-hot feature for the i-th residue in the sequence S at timestep t, and $\hat{S}^0$ is the predicted probability of S^0. In this work, we design the $\hat{p}_\theta$ as follows:

$$\hat{p}_\theta\left(\hat{S}^0 \mid S^t\right) = \prod_{1 \le i \le N} \text{Softmax}\left(\hat{s}_i^0 \mid \mathcal{F}_s\left(h_i^t\right)\right) \quad (2)$$

where h_i^t is the input feature of residue i with the diffusion noise at time t (the initialization of h_i^t is provided in Appendix A). $\mathcal{F}_s$ is a hybrid neural network to predict the noise of residue types from the marginal distribution, and then the noise would be removed to compute the probability of $\hat{s}_i^0$. Softmax is applied over all residue types. Here, we implement $\mathcal{F}_s$ with a transformer encoder and an MLP. The former learns contextual embeddings of residues from the sequence, while the latter maps these embeddings to the noises of residue types. The learned sequence embedding (defined as $\mathcal{S}$) involves downstream contrastive learning strategies.

Diffusion for Peptide Structure. As the coordinates of atoms are continuous variables in the 3D space, the forward process can be defined by adding Gaussian noise to atom coordinates (Ho, Jain, and Abbeel 2020) (see details in Appendix A). Following Trippe et al. (2023), the reverse process can be defined as:

$$p_\theta(c_i^{t-1} \mid C^t) = \mathcal{N}(c_i^{t-1} \mid \mu_\theta(C^t, t), \beta^t I) \tag{3}$$

$$\mu_\theta\left(C^t, t\right) = \frac{1}{\sqrt{\alpha^t}}\left(c_i^t - \frac{\beta^t}{\sqrt{1-\overline{\alpha}^t}}\epsilon_\theta\left(C^t, t\right)\right) \tag{4}$$

where c_i refers to coordinates of the i-th residue in the structure C; β is the noise rate, formally $\alpha^t = 1 - \beta^t$, $\overline{\alpha}^t = \prod_{\tau=1}^t (1-\beta^\tau)$; the network ϵ_θ is used to gradually recover the structural data by predicting the Gaussian noise. In this work, we design the ϵ_θ as follows:

$$\epsilon_\theta(C^t, t) = \mathcal{F}_c\left(r_i^t, h_i^t\right) \tag{5}$$

where r_i represents the coordinates of residue i, h_i is the residue feature, and $\mathcal{F}_c$ is a hybrid neural network for predicting Gaussian noises at timestep t. Similar to sequence diffusion, we implement $\mathcal{F}_c$ with an equivariant graph neural network (EGNN) (Satorras, Hoogeboom, and Welling 2021) and an MLP. The former learns spatial embeddings of residues from the structure (formalized as a 3D graph), while the latter maps these embeddings to Gaussian noises. The learned structure embedding (defined as $\mathcal{C}$) also involves downstream contrastive learning strategies.

Diffusion Objective. Following previous work (Anand and Achim 2022), we decompose the objective of the peptide diffusion process into sequence loss and structure loss. For the sequence loss $\mathcal{L}_S^t$, we aim to minimize the cross-entropy (CE) loss between the actual and predicted residue types at timestep t:

$$\mathcal{L}_S^t = \frac{1}{N}\sum_{1\le i\le N} \text{CE}\left(s_i^0, \hat{p}_\theta(\hat{s}_i^0 | S^t)\right) \tag{6}$$

For the structure loss $\mathcal{L}_C^t$, the objective is to calculate the mean squared error (MSE) between the predicted noise ϵ_θ and standard Gaussian noise ϵ at timestep t:

$$\mathcal{L}_C^t = \frac{1}{N}\sum_{1\le i\le N} \left\|\epsilon_i - \epsilon_\theta(C^t, t)\right\|^2 \tag{7}$$

Multi-Modal Contrastive Learning Strategy

When multiple modal data (e.g., sequence and structure) coexist, it becomes imperative to capture their consistency to reduce the heterogeneous differences between modalities, allowing them to be better fused in generation tasks. Mutual information (MI) is a straightforward solution to measure the non-linear dependency (consistency) between variables (Liu et al. 2023); thus, maximizing MI between modalities can force them to align and share more crucial information. Along this line, we bring in contrastive learning (CL) to align sequences and structures by maximizing their MI in the embedding space. Specifically, we devise CL strategies for each diffusion timestep t, as follows:

Inter-CL. For a peptide, we define its sequence as the anchor, its structure as the positive instance, and the structures of other peptides in a mini-batch as the negative instances. Then, we maximize the MI of positive pair (anchor and positive instance) while minimizing the MI of negative pairs (anchor and negative instances), based on embeddings learned from the networks $\hat{p}_\theta$ and ϵ_θ. Further, we establish a 'dual' contrast where the structure acts as an anchor and sequences are instances. The objective is to minimize the following InfoNCE-based (Chen et al. 2020) loss function:

$$\mathcal{L}_{\text{inter}}^t = -\frac{1}{2}\left[\log\frac{E\left(\mathcal{S}_i^t, \mathcal{C}_i^t\right)}{\sum_{j=1}^M E\left(\mathcal{S}_i^t, \mathcal{C}_j^t\right)} + \log\frac{E\left(\mathcal{C}_i^t, \mathcal{S}_i^t\right)}{\sum_{j=1}^M E\left(\mathcal{C}_i^t, \mathcal{S}_j^t\right)}\right] \tag{8}$$

where $\mathcal{S}_i/\mathcal{C}_i$ is the sequence/structure embeddings of i-th peptide in the mini-batch, $E(\cdot,\cdot)$ is the cosine similarity function with the temperature coefficient to measure the MI score between two variables, M is the size of a mini-batch.

In addition, the used diffusion model can only remember confined generation patterns if therapeutic peptide data for training is limited, which may lead to inferior generalization towards novel peptides. To alleviate this issue, we introduce contrastive learning to boost the generative capacity of networks $\hat{p}_\theta$ and ϵ_θ by enriching the supervised signals. However, it is unwise to construct positive instances by performing data augmentations on therapeutic peptides, as even minor perturbations may lead to significant functional changes (Yadav, Kumar, and Singh 2022). Hence, our focus lies on employing effective strategies for selecting negative instances. In this regard, we collect non-therapeutic peptides from public databases to treat them as negative instances, and maximize the disagreement between embeddings of therapeutic and non-therapeutic peptides. In detail, we devise an Intra-CL strategy for each diffusion timestep t, as follows:

Intra-CL. In a mini-batch, we define the sequence of a therapeutic peptide i as the anchor, and the sequence of another therapeutic peptide j as the positive instance, while the sequences of non-therapeutic peptides k are regarded as negative instances. Similar to Inter-CL, we then maximize/minimize the MI of positive/negative pairs. And we also establish a structure-oriented contrast by using structures of therapeutic and non-therapeutic peptides to construct the anchor, positive, and negative instances. The objective is to minimize the following loss function (Zheng et al. 2021):

$$\begin{aligned}\mathcal{L}_{\text{intra}}^t = -\frac{1}{M}\sum_{j=1, j\neq i}^{M} 1_{y_i=y_j}\Bigg(\log\frac{E\left(\mathcal{S}_i^t, \mathcal{S}_j^t\right)}{\sum_{k=1}^M 1_{y_i\neq y_k} E\left(\mathcal{S}_i^t, \mathcal{S}_k^t\right)} \\ + \log\frac{E\left(\mathcal{C}_i^t, \mathcal{C}_j^t\right)}{\sum_{k=1}^M 1_{y_i\neq y_k} E\left(\mathcal{C}_i^t, \mathcal{C}_k^t\right)}\Bigg)\end{aligned} \tag{9}$$

where y_i represents the class of peptide i (i.e., therapeutic or non-therapeutic). $1_{y_i=y_j}$ and $1_{y_i\neq y_k}$ stand for the indicator functions, where the output is 1 if $y_i = y_j$ (peptides i and j belong to the same class) or $y_i \neq y_k$ (the types of peptides i and k are different); otherwise the output is 0. The indicator function filters therapeutic and non-therapeutic peptides from the data for creating positive and negative pairs.

Methods	AMP			ACP		
	Similarity↓	Instability↓	Antimicrobial↑	Similarity↓	Instability↓	Anticancer↑
LSTM-RNN	39.6164	45.0862	0.8550	36.9302	47.0669	0.7336
AMPGAN*	38.3080	51.5236	0.8617	-	-	-
HydrAMP*	31.0662	59.6340	0.8145	-	-	-
WAE-PSO*	-	-	-	41.2524	42.5061	0.7443
DiffAB	28.9849	43.3607	0.8024	31.4220	36.0610	0.6669
SimDiff	25.5385	41.1629	0.8560	28.8245	33.0405	0.7222
MMCD	**24.4107**	**39.9649**	**0.8810**	**27.4685**	**31.7381**	**0.7604**

’*’ represents that the method relies on domain-specific biological knowledge. ’-’ represents that the method is unsuitable for the current task. For example, AMPGAN and HydrAMP are only designed for the AMP generation.

Table 1: Results for the sequence generation

Methods	AMP			ACP	
	Ramachandran↑	RMSD↓	Docking↑	Ramachandran↑	RMSD↓
APPTEST	69.6576	2.7918	1362	67.9826	2.8055
FoldingDiff	72.4681	2.5118	1574	72.0531	2.6033
ProtDiff	71.3078	2.5544	1533	69.7589	2.4960
DiffAB	72.9647	2.3844	1608	71.3225	2.5513
SimDiff	76.1378	2.1004	1682	76.6164	2.4118
MMCD	**80.4661**	**1.8278**	**1728**	**78.2157**	**2.0847**

Table 2: Results for the structure generation.

The reason behind the design of Intra-CL is intuitive. First, the non-therapeutic class naturally implies opposite information against the therapeutic class, and hence it makes the model more discriminative. Second, the fashion to maximize the disagreement between classes (1) can induce biases in the embedding distribution of therapeutic peptides, identifying more potential generation space, and (2) can explicitly reinforce embedding-class correspondences during diffusion, maintaining high generation fidelity (Zhu et al. 2022). Further analysis is detailed in the ablation study.

Model Training

The ultimate objective function is the sum of the diffusion process for sequence and structure generation, along with the CL tasks for Intra-CL and Inter-CL:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{t\sim\text{Uniform(1...T)}}\left[\alpha\left(\mathcal{L}_S^t + \mathcal{L}_C^t\right) + (1-\alpha)\left(\mathcal{L}_{\text{intra}}^t + \mathcal{L}_{\text{inter}}^t\right)\right] \tag{10}$$

where α represents a hyperparameter to balance the contributions of different tasks. The Uniform(1...T) shows the uniform distribution for the diffusion timesteps. The implementation details of MMCD and the sampling process of peptide generation can be found in Appendix A.

Experiments

Experimental Setups

Datasets. Following previous studies (Thi Phan et al. 2022; Zhang et al. 2023a), we collected therapeutic peptide data from public databases, containing two biological types, i.e., antimicrobial peptides (AMP) and anticancer peptides (ACP). Among these collected peptides, a portion of them only have 1D sequence information, without 3D structure information. Then, we applied Rosetta-based computational tools (Chaudhury, Lyskov, and Gray 2010) to predict the missing structures based on their sequences. Finally, we compiled two datasets, one containing 20,129 antimicrobial peptides and the other containing 4,381 anticancer peptides. In addition, we paired an equal number of labeled non-therapeutic peptides (collected from public databases) with each of the two datasets, exclusively for the contrastive learning task.

Baselines. We compared our method with the following advanced methods for peptide generation at sequence and structure levels. For the sequence generation, the autoregression-based method LSTM-RNN (Müller, Hiss, and Schneider 2018), the GAN-based method AMPGAN (Oort et al. 2021), and the VAE-based methods including WAE-PSO (Yang et al. 2022) and HydrAMP (Szymczak et al. 2023a) are listed as baselines. For the structure generation, we took APPTEST (Timmons and Hewage 2021) as a baseline, which combines the neural network and simulated annealing algorithm for structure prediction. Moreover, we extended diffusion-based methods for protein generation to peptides. The diffusion-based methods for structure generation (e.g., FoldingDiff (Wu et al. 2022) and ProtDiff (Trippe et al. 2023)) and the sequence-structure co-design (e.g., DiffAB(Luo et al. 2022) and SimDiff(Zhang et al. 2023b)), are considered for the comparison separately in the sequence and structure generation.

Evaluation Protocol. Here, we required each model (ours and baselines) to generate 1,000 new peptides, and then evaluated the quality of generated peptides with the following metrics. For the sequence, **similarity** score is used to

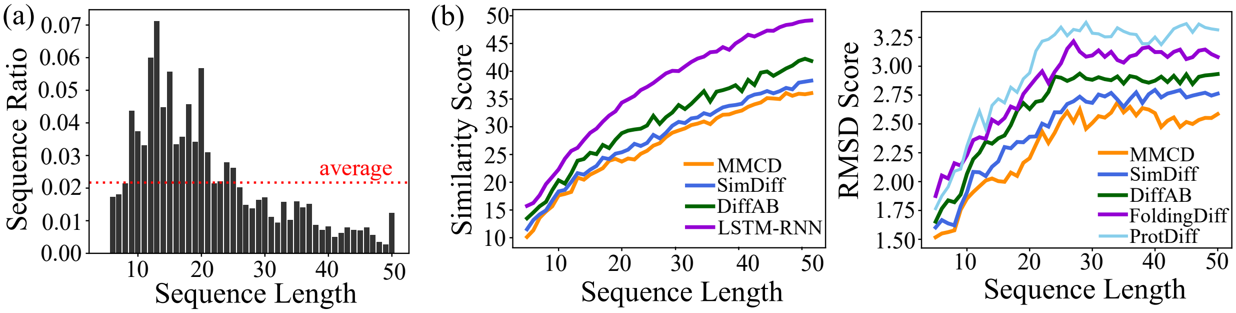

Figure 2: (a) The sample ratio under different sequence lengths in the AMP dataset, where the red line is the average ratio. (b) The similarity and RMSD scores of MMCD and baselines across different sequence lengths.

quantify how closely the generated sequences match existing ones, with a lower score indicating higher novelty; **instability** score (Müller et al. 2017) indicates the degree of peptide instability; **antimicrobial/anticancer** score evaluates the probability of peptides having therapeutic properties. For the structure, **Ramachandran** score (Hollingsworth and Karplus 2010) accesses the reliability of peptide structures; **RMSD** score measures the structural similarity between generated and existing peptides, with a lower score indicating higher authenticity; **docking** score (Flórez-Castillo et al. 2020) evaluates the binding degree of antimicrobial peptides to bacterial membrane proteins (PDB ID: 6MI7). We only reported the average metrics over all generated peptides for each method in the experimental results. Detailed information about the datasets, baselines, metrics, and implementations can be found in Appendix B. Our code, data, and appendix are available on GitHub (https://github.com/wyky481l/MMCD)

Experimental Results

Performance Comparison. In the results of sequence generation under two datasets (as shown in Table 1), MMCD exhibited lower similarity and instability scores than all baselines, suggesting its good generalization ability in generating diverse and stable peptides. Meanwhile, MMCD surpassed all baselines with higher antimicrobial and anticancer scores across AMP and ACP datasets, highlighting its strong potential for generating therapeutic peptides. Beyond that, we noticed that diffusion-based baselines (e.g., SimDiff, DiffAB) exhibit higher stability and diversity but lower therapeutic scores compared to baselines that incorporate biological knowledge (e.g., AMPGAN, HydrAMP, WAE-PSO, details in Appendix B). By contrast, MMCD introduced biological knowledge into the diffusion model by designing the contrastive learning of therapeutic and non-therapeutic peptides, thereby delivering optimality across various metrics.

For the results of structure generation (as shown in Table 2), MMCD also outperformed all the baselines and exceeded the best baselines (DiffAB and SimDiff) by 23.3% and 12.9% in RMSD scores, 10.2% and 5.6% in Ramachandran scores, and 7.4% and 2.7% in docking scores for AMP dataset. The higher Ramachandran score and lower RMSD score of MMCD underlined the reliability of our generated peptide structures. Especially in peptide docking, we found that MMCD shows the best docking score compared with baselines, which indicates great binding interactions with the target protein. Overall, MMCD is superior to all baselines in both sequence and structure generation of peptides, and its impressive generative ability holds great promise to yield high-quality therapeutic peptides.

Performance on Different Sequence Lengths. In our dataset, sequence lengths of different peptides exhibited substantial variation, with the number of residues ranging from 5 to 50 (Figure 2-a). We required models to generate 20 new peptides (sequences or structures) at each sequence length. Note that two methods, AMPGAN and HydrAMP, were excluded from the comparison because they cannot generate peptides with fixed lengths. From the generated results on the AMP dataset (Figure 2-b), MMCD exceeded the baselines in terms of similarity and RMSD scores at each sequence length. With the increasing sequence lengths, there is a general trend of increased similarity and RMSD scores across all methods. One possible reason for this trend is that designing longer peptides becomes more complex, given the more prominent search space involved. Additionally, the scarcity of long-length peptides poses challenges in accurately estimating the similarity between generated and known peptides. In summary, these observations supported that MMCD excels at generating diverse peptides across different lengths, especially shorter ones.

Ablation Study

To investigate the necessity of each module in MMCD, we conducted several comparisons between MMCD with its variants: (1) MMCD (w/o Inter-CL) that removes the Inter-CL task, (2) MMCD (w/o Intra-CL) that removes the Intra-CL task, and (3) MMCD (w/o Inter-CL & Intra-CL) that removes both Inter-CL and Intra-CL tasks. The comparisons were operated on both AMP and ACP datasets, and the results are shown in Table 3 and Appendix Table 1. When the Inter-CL was removed (w/o Inter-CL), we observed a decline in all metrics for peptide sequence and structure generation, implying the importance of aligning two modalities via CL. The variant (w/o Intra-CL) results signified that using the CL to differentiate therapeutic and non-therapeutic peptides contributes to the generation. As expected, the per-

Methods	AMP			ACP		
	Similarity↓	Instability↓	Antimicrobial↑	Similarity↓	Instability↓	Anticancer↑
MMCD (w/o InterCL & IntraCL)	27.4794	42.5359	0.8013	31.2820	34.6888	0.6996
MMCD (w/o IntraCL)	26.6889	41.2631	0.8584	28.9782	33.0268	0.7513
MMCD (w/o InterCL)	24.9079	41.7646	0.8494	28.0143	33.9816	0.7352
MMCD	24.4107	39.9649	0.8810	27.4685	31.7381	0.7604

Table 3: Ablation study on the sequence-level generation task.

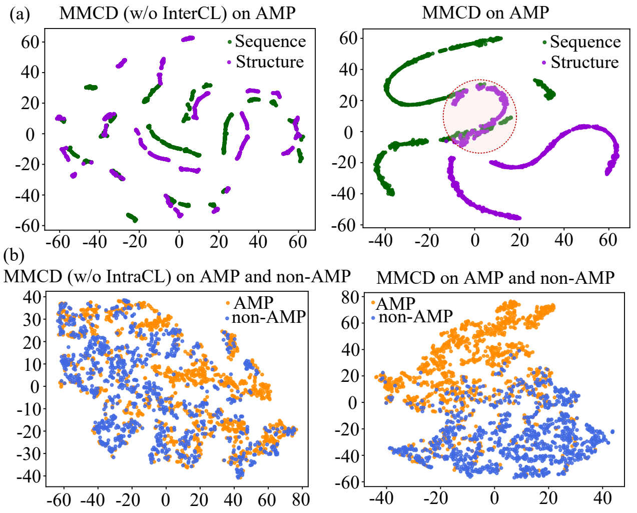

Figure 3: (a) The t-SNE for structure and sequence embeddings of therapeutic peptides (AMP data) obtained from MMCD (w/o Inter-CL) and MMCD. (b) The t-SNE for embeddings (including structures and sequences) of therapeutic (AMP) and non-therapeutic (non-AMP) peptides obtained from MMCD (w/o Intra-CL) and MMCD.

formance of MMCD dropped significantly after removing both Inter-CL and Intra-CL (w/o Inter-CL & Intra-CL).

To better understand the strengths of Inter-CL and Intra-CL, we performed the t-SNE (Van der Maaten and Hinton 2008) visualization using the learned embeddings of peptides on the AMP dataset. As illustrated in Figure 3-a, Inter-CL effectively promoted the alignment of sequence and structure embeddings, facilitating the shared crucial information (dashed circle) to be captured during diffusion. The t-SNE of Intra-CL (Figure 3-b) also revealed that it better distinguished therapeutic peptides from non-therapeutic ones in the embedding distribution. And the resulting distribution bias may identify more potential generation space, thus leading to higher quality and diversity of therapeutic peptides generated by MMCD. Overall, MMCD with all the modules fulfilled superior performance, and removing any modules will diminish its generation power.

Peptide-docking Analysis

To test the validity of generated peptide structures, we conducted a molecular-docking simulation. Here, a peptide was randomly selected from the AMP dataset as the reference, and the methods (Figure 4) were employed to generate corresponding structures based on the sequence of the reference peptide (see details in Appendix C). The lipopolysaccharide on the outer membrane of bacteria (Li, Orlando, and Liao 2019) was selected as the target protein for molecular docking. Then, we extracted the residues within a 5Å proximity between peptides (i.e., the reference and generated structures) and the active pocket of target protein in docking complexes, to visualize their binding interactions (Miller et al. 2021). Of these docking results, all methods yielded a new structure capable of binding to the target protein, and our method exhibited the highest docking scores and displayed binding residues most similar to the reference structure. This prominent result underscored the reliability and therapeutic potential of our method for peptide generation.

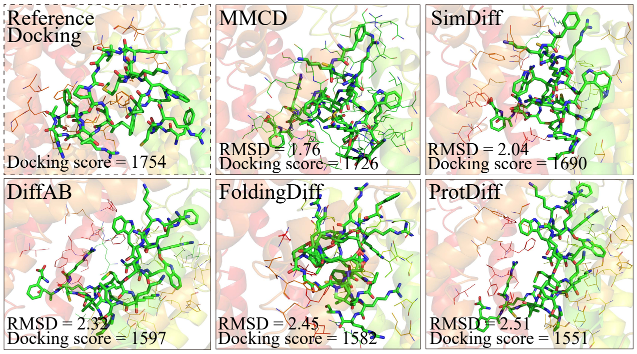

Figure 4: Docking analysis (interactive visualization between target protein and peptides) of the reference and generated structures by MMCD and baselines. Thick lines represent the residues of peptides, and the thin lines show the binding residues for protein-peptide complexes.

Conclusion

In this work, we propose a multi-modal contrastive diffusion model for the co-generation of peptide sequences and structures, named MMCD. MMCD is dedicated to leveraging a multi-modal contrastive learning strategy to capture consensus-related and difference-related information behind the sequences/structures and therapeutic/non-therapeutic peptides, enhancing the diffusion model to generate high-quality therapeutic peptides. The experimental results unequivocally demonstrate the capability of our method in co-generating peptide sequence and structure, surpassing state-of-the-art baseline methods with advantageous performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62372204, 62072206, 61772381, 62102158); Huazhong Agricultural University Scientific & Technological Self-innovation Foundation; Fundamental Research Funds for the Central Universities (2662021JC008, 2662022JC004). The funders have no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

References

Anand, N.; and Achim, T. 2022. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. arxiv:2205.15019.

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and van den Berg, R. 2021. Structured Denoising Diffusion Models in Discrete State-Spaces. In *Advances in Neural Information Processing Systems*, volume 34, 17981–17993. Curran Associates, Inc.

Cao, H.; Tan, C.; Gao, Z.; Xu, Y.; Chen, G.; Heng, P.-A.; and Li, S. Z. 2023. A Survey on Generative Diffusion Model. arxiv:2209.02646.

Capecchi, A.; Cai, X.; Personne, H.; Köhler, T.; van Delden, C.; and Reymond, J.-L. 2021. Machine Learning Designs Non-Hemolytic Antimicrobial Peptides. *Chem Sci*, 12(26): 9221–9232.

Chaudhury, S.; Lyskov, S.; and Gray, J. J. 2010. PyRosetta: A Script-Based Interface for Implementing Molecular Modeling Algorithms Using Rosetta. *Bioinformatics*, 26(5): 689–691.

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

Flórez-Castillo, J. M.; Rondón-Villareal, P.; Ropero-Vega, J. L.; Mendoza-Espinel, S. Y.; Moreno-Amézquita, J. A.; Méndez-Jaimes, K. D.; Farfán-García, A. E.; Gómez-Rangel, S. Y.; and Gómez-Duarte, O. G. 2020. Ib-M6 Antimicrobial Peptide: Antibacterial Activity against Clinical Isolates of Escherichia Coli and Molecular Docking. *Antibiotics*, 9(2): 79.

Ghorbani, M.; Prasad, S.; Brooks, B. R.; and Klauda, J. B. 2022. Deep Attention Based Variational Autoencoder for Antimicrobial Peptide Discovery.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates, Inc.

Hollingsworth, S. A.; and Karplus, P. A. 2010. A Fresh Look at the Ramachandran Plot and the Occurrence of Standard Structures in Proteins. 1(3-4): 271–283.

Hoogeboom, E.; Satorras, V. G.; Vignac, C.; and Welling, M. 2022. Equivariant Diffusion for Molecule Generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, 8867–8887. PMLR.

Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; and Huang, L. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34: 10944–10956.

Iqbal, T.; and Qureshi, S. 2022. The Survey: Text Generation Models in Deep Learning. *J King Saud Univ-com*, 34(6, Part A): 2515–2528.

Jakubczyk, A.; Karaś, M.; Rybczyńska-Tkaczyk, K.; Zielińska, E.; and Zieliński, D. 2020. Current Trends of Bioactive Peptides—New Sources and Therapeutic Effect. *Foods*, 9(7): 846.

Lee, E. Y.; Lee, M. W.; Fulan, B. M.; Ferguson, A. L.; and Wong, G. C. L. 2017. What Can Machine Learning Do for Antimicrobial Peptides, and What Can Antimicrobial Peptides Do for Machine Learning? *Interface Focus*, 7(6): 20160153.

Lee, E. Y.; Wong, G. C. L.; and Ferguson, A. L. 2018. Machine Learning-Enabled Discovery and Design of Membrane-Active Peptides. *Bioorgan Med Chem*, 26(10): 2708–2718.

Li, Y.; Orlando, B. J.; and Liao, M. 2019. Structural Basis of Lipopolysaccharide Extraction by the LptB2FGC Complex. *Nature*, 567(7749): 486–490.

Lin, E.; Lin, C.-H.; and Lane, H.-Y. 2022. De novo peptide and protein design using generative adversarial networks: an update. *Journal of Chemical Information and Modeling*, 62(4): 761–774.

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

Liu, V.; and Chilton, L. B. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, 1–23. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9157-3.

Luo, S.; Su, Y.; Peng, X.; Wang, S.; Peng, J.; and Ma, J. 2022. Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models for Protein Structures.

Miller, E. B.; Murphy, R. B.; Sindhikara, D.; Borrelli, K. W.; Grisewood, M. J.; Ranalli, F.; Dixon, S. L.; Jerome, S.; Boyles, N. A.; Day, T.; Ghanakota, P.; Mondal, S.; Rafi, S. B.; Troast, D. M.; Abel, R.; and Friesner, R. A. 2021. Reliable and Accurate Solution to the Induced Fit Docking Problem for Protein–Ligand Binding. *J Chem Theory Comput*, 17(4): 2630–2639.

Müller, A. T.; Gabernet, G.; Hiss, J. A.; and Schneider, G. 2017. modlAMP: Python for Antimicrobial Peptides. *Bioinformatics*, 33(17): 2753–2755.

Müller, A. T.; Hiss, J. A.; and Schneider, G. 2018. Recurrent Neural Network Model for Constructive Peptide Design. *J Chem Inf Model*, 58(2): 472–479.

Muttenthaler, M.; King, G. F.; Adams, D. J.; and Alewood, P. F. 2021. Trends in peptide drug discovery. *Nature reviews Drug discovery*, 20(4): 309–325.

Oort, C. M. V.; Ferrell, J. B.; Remington, J. M.; Wshah, S.; and Li, J. 2021. AMPGAN v2: Machine Learning Guided Design of Antimicrobial Peptides.

Satorras, V. G.; Hoogeboom, E.; and Welling, M. 2021. E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, 9323–9332. PMLR.

Shi, C.; Wang, C.; Lu, J.; Zhong, B.; and Tang, J. 2023. Protein Sequence and Structure Co-Design with Equivariant Translation. arxiv:2210.08761.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2256–2265. PMLR.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Szymczak, P.; Możejko, M.; Grzegorzek, T.; Jurczak, R.; Bauer, M.; Neubauer, D.; Sikora, K.; Michalski, M.; Sroka, J.; Setny, P.; Kamysz, W.; and Szczurek, E. 2023a. Discovering Highly Potent Antimicrobial Peptides with Deep Generative Model HydrAMP. *Nat Commun*, 14(1): 1453.

Szymczak, P.; Możejko, M.; Grzegorzek, T.; Jurczak, R.; Bauer, M.; Neubauer, D.; Sikora, K.; Michalski, M.; Sroka, J.; Setny, P.; et al. 2023b. Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nature Communications*, 14(1): 1453.

Thi Phan, L.; Woo Park, H.; Pitti, T.; Madhavan, T.; Jeon, Y.-J.; and Manavalan, B. 2022. MLACP 2.0: An Updated Machine Learning Tool for Anticancer Peptide Prediction. *Comput Struct Biotec*, 20: 4473–4480.

Timmons, P. B.; and Hewage, C. M. 2021. APPTEST Is a Novel Protocol for the Automatic Prediction of Peptide Tertiary Structures. *Brief Bioinform*, 22(6): bbab308.

Trippe, B. L.; Yim, J.; Tischer, D.; Baker, D.; Broderick, T.; Barzilay, R.; and Jaakkola, T. 2023. Diffusion Probabilistic Modeling of Protein Backbones in 3D for the Motif-Scaffolding Problem. arxiv:2206.04119.

Tucs, A.; Tran, D. P.; Yumoto, A.; Ito, Y.; Uzawa, T.; and Tsuda, K. 2020. Generating Ampicillin-Level Antimicrobial Peptides with Activity-Aware Generative Adversarial Networks. *ACS Omega*, 5(36): 22847–22851.

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

Vignac, C.; Krawczuk, I.; Siraudin, A.; Wang, B.; Cevher, V.; and Frossard, P. 2023. DiGress: Discrete Denoising Diffusion for Graph Generation. arxiv:2209.14734.

Wan, F.; Kontogiorgos, H. D.; and Fuente, d. l. N. C. 2022. Deep Generative Models for Peptide Design. *Digital Discovery*, 1(3): 195–208.

Webster, R.; Rabin, J.; Simon, L.; and Jurie, F. 2019. Detecting overfitting of deep generative networks via latent recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11273–11282.

Wu, K. E.; Yang, K. K.; van den Berg, R.; Zou, J. Y.; Lu, A. X.; and Amini, A. P. 2022. Protein Structure Generation via Folding Diffusion. arxiv:2209.15611.

Wu, X.; Luu, A. T.; and Dong, X. 2022. Mitigating Data Sparsity for Short Text Topic Modeling by Topic-Semantic Contrastive Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2748–2760.

Wu, Z.; Johnston, K. E.; Arnold, F. H.; and Yang, K. K. 2021. Protein sequence design with deep generative models. *Current Opinion in Chemical Biology*, 65: 18–27.

Xia, C.; Feng, S.-H.; Xia, Y.; Pan, X.; and Shen, H.-B. 2022. Fast protein structure comparison through effective representation learning with contrastive graph neural networks. *PLoS computational biology*, 18(3): e1009986.

Yadav, N. S.; Kumar, P.; and Singh, I. 2022. Structural and functional analysis of protein. In *Bioinformatics*, 189–206. Elsevier.

Yang, L.; Yang, G.; Bing, Z.; Tian, Y.; Huang, L.; Niu, Y.; and Yang, L. 2022. Accelerating the Discovery of Anticancer Peptides Targeting Lung and Breast Cancers with the Wasserstein Autoencoder Model and PSO Algorithm. *Brief Bioinform*, 23(5): bbac320.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion Models: A Comprehensive Survey of Methods and Applications. arxiv:2209.00796.

Yuan, X.; Lin, Z.; Kuen, J.; Zhang, J.; Wang, Y.; Maire, M.; Kale, A.; and Faieta, B. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6995–7004.

Zhang, H.; Saravanan, K. M.; Wei, Y.; Jiao, Y.; Yang, Y.; Pan, Y.; Wu, X.; and Zhang, J. Z. H. 2023a. Deep Learning-Based Bioactive Therapeutic Peptide Generation and Screening. *J Chem Inf Model*, 63(3): 835–845.

Zhang, Z.; Xu, M.; Lozano, A.; Chenthamarakshan, V.; Das, P.; and Tang, J. 2023b. Pre-Training Protein Encoder via Siamese Sequence-Structure Diffusion Trajectory Prediction. arxiv:2301.12068.

Zhang, Z.; Zhao, Y.; Chen, M.; and He, X. 2022. Label Anchored Contrastive Learning for Language Understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1437–1449.

Zheng, M.; Wang, F.; You, S.; Qian, C.; Zhang, C.; Wang, X.; and Xu, C. 2021. Weakly Supervised Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10042–10051.

Zhu, Y.; Wu, Y.; Olszewski, K.; Ren, J.; Tulyakov, S.; and Yan, Y. 2022. Discrete contrastive diffusion for cross-modal and conditional generation. *arXiv preprint arXiv:2206.07771*.

Zhu, Y.; Wu, Y.; Olszewski, K.; Ren, J.; Tulyakov, S.; and Yan, Y. 2023. Discrete Contrastive Diffusion for Cross-Modal Music and Image Generation. arxiv:2206.07771.