
Latent Retrieval Augmented Generation of Cross-Domain Protein Binders

Zishen Zhang^{1,2} Xiangzhe Kong^{1,2} Wenbing Huang^{3*} Yang Liu^{1,2*}

¹Dept. of Comp. Sci. & Tech., Tsinghua University

²Institute for AIR, Tsinghua University

³Gaoling School of Artificial Intelligence, Renmin University of China

Abstract

Designing protein binders targeting specific sites, which requires to generate realistic and functional interaction patterns, is a fundamental challenge in drug discovery. Current structure-based generative models are limited in generating interfaces with sufficient rationality and interpretability. In this paper, we propose **Retrieval-Augmented Diffusion for Aligned interface(RADiAance)**, a new framework that leverages known interfaces to guide the design of novel binders. By unifying retrieval and generation in a shared contrastive latent space, our model efficiently identifies relevant interfaces for a given binding site and seamlessly integrates them through a conditional latent diffusion generator, enabling cross-domain interface transfer. Extensive experiments show that **RADiAance** significantly outperforms baseline models across multiple metrics, including binding affinity and recovery of geometries and interactions. Additional experimental results validate cross-domain generalization, demonstrating that retrieving interfaces from diverse domains, such as peptides, antibodies, and protein fragments, enhances the generation performance of binders for other domains. Our work establishes a new paradigm for protein binder design that successfully bridges retrieval-based knowledge and generative AI, opening new possibilities for drug discovery.

1 Introduction

Designing binders that target specific sites on proteins is a fundamental problem with broad implications in drug discovery, structural biology, and beyond [54, 45]. The core challenge involves engineering molecular interfaces that exhibit desired interaction patterns with a target surface, such as hydrogen bonding, hydrophobic packing, and π - π stacking [45]. Traditional approaches rely on sampling and optimizing within physical or statistical energy landscapes [3, 8, 22], while recent advances have shifted towards direct structure-based generation of binders with deep generative models, such as diffusion and flow matching [53, 40].

Despite these advances, current methods remain limited in rationality and interpretability due to their inability to leverage prior knowledge from existing structurally similar interfaces. Most approaches conduct generation solely conditional on the target binding site [30, 40], ignoring the wealth of reusable interaction patterns present in existing protein complexes. In contrast, it is routine for human experts (e.g. medicinal chemists or structural biologists) to rely on similar known interfaces to guide rational design [48, 17]. We thus argue that a better generative paradigm should identify and incorporate analogous interfaces from existing databases to improve both rationality and interpretability. However, incorporating interface retrieval into the generation process poses several challenges. First, the retriever should identify relevant interfaces from a large database using only the binding site as input. Existing retrieval methods typically assume a known binder structure (e.g., for inverse

*Correspondence to Wenbing Huang <hwenbing@126.com>, Yang Liu <liuyang2011@tsinghua.edu.cn>

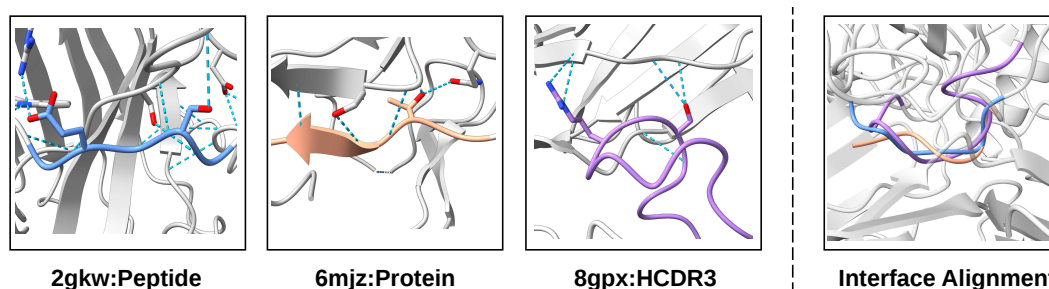


Figure 1: Visualization of interface similarity across antibodies, proteins, and peptides, highlighting similar interaction patterns among diverse binder types.

folding [51]), which is inapplicable in our setting. Second, while binders may vary in form (e.g., cyclic peptides, linear peptides, antibodies) due to different biochemical requirements, the underlying interaction motifs are often generalizable across binders of different types and domains, as illustrated in Fig. 1. Therefore, the similarity metric used in the retriever should capture shared interactions between cross-domain interfaces and binding sites. Moreover, a mechanism is needed to integrate the retrieved information into the generative process, effectively focusing on relevant parts for different generated regions.

To address the challenges, this work introduces **Retrieval-Augmented Diffusion for Aligned interface (RADiAance)**, a novel framework that unifies retrieval and generation in a shared contrastive latent space. We begin by constructing a comprehensive interface database spanning peptides [30], antibodies [14], and protein fragments [30], where each entry comprising a binding site and a corresponding ligand interface. We then train an all-atom variational autoencoder (VAE) [31] augmented with contrastive learning to derive an interaction-aligned and retrieval-friendly latent space. The encoder maps each binding site and interface independently to separate latent vectors, which are supervised by a contrastive loss to align for positive pairs and repel for negative ones. At inference time, the encoder is used to map database entries of interface pairs into key-value vectors, enabling fast and accurate retrieval using simple dot-product similarity. A diffusion model is then trained in the same latent space, conditioned on the retrieved interface embeddings. These embeddings act as compressed yet informative prompts, integrated through a combination of cross-attention and residual MLPs [50]. This enables the model to leverage relevant priors (such as residue preferences, contact geometries, and binding orientations) from the retrieved examples, resulting in rational, interpretable binder generations with physically plausible interfaces.

We summarize our contributions as follows:

- **Contrastive Latent Space for Cross-Domain Interface Retrieval.** We propose an all-atom VAE that is trained contrastively to align latent representations between binding sites and their interfaces. This latent space enables accurate and a unified similarity metric across diverse binder types such as antibodies, peptides, and protein fragments.
- **Retrieval-Augmented Latent Diffusion.** To the best of our knowledge, we are the first to apply retrieval-augmented generation for the sequence-structure codesign of protein binders. We introduce a latent diffusion model that operates in the same latent space as the retriever, seamlessly incorporating retrieved interface embeddings via cross-attention and residual conditioning to guide generation.
- **Extensive Benchmarking and Analysis.** We benchmark **RADiAance** on peptide and antibody design tasks, demonstrating significant improvements over strong baselines in recovering sequence, structure, and interaction patterns. Further analysis confirms that the gains stem from the ability of our model in mimicking retrieved interaction motifs. We also show that leveraging interfaces from other binder types (e.g., antibodies to aid peptide generation) improves performance, verifying the rationality of cross-domain transfer.

2 Related Work

Peptide Design Peptide design has transitioned from classical energy-based sampling [8, 22] to deep generative modeling [35]. Recent approaches, such as PepFlow [36] and PPFlow [39], employ multi-

modal flow matching to jointly model residue types, backbone orientations, C_α positions, and side-chain dihedrals, setting new benchmarks in peptide generation. PepGLAD [30] advances this direction by applying geometric latent diffusion models for sequence–structure co-design. Parallel efforts using language models, including PeptideGPT [47], generate functional peptides from sequence space and filter candidates through structure prediction tools like AlphaFold [27] to ensure 3D viability.

Antibody Design Antibody design has similarly evolved from traditional physical sampling [3] to deep learning frameworks [46, 4, 41]. Early methods focused on CDR loop generation without antigen context, while later works such as MEAN [28] and DiffAb [40] introduced antigen-conditioned generation. Diffusion-based models like AbDiffuser [41] and GeoAB [38] further improve structure-aware antibody design by incorporating physical constraints, while HERN [26] captures long-range dependencies via hierarchical graph decoding. Additionally, language-model-based approaches have gained prominence, PALM-H3 [19] and MAGE [52] adopt encoder-decoder architectures to generate CDR sequences or full variable regions conditioned on antigen features.

Retrieval-Augmented Generation(RAG) RAG integrates external knowledge into generative models to enhance fidelity, interpretability, and controllability. Initially developed for NLP tasks like question answering [34, 18], RAG frameworks pair a retriever with a generator to access dynamic knowledge during inference. In vision, retrieval-conditioned diffusion models [10, 11] improve image synthesis by incorporating structure-consistent references. Recently, RAG has emerged as a powerful paradigm for targeted molecular design. For instance, f-RAG introduces a dynamic fragment vocabulary to improve small-molecule generation via fragment-level retrieval [55]. Structure-based approaches such as IRDiff leverage high-affinity ligands as retrieval references to guide diffusion-based 3D molecular generation [32], while READ incorporates pocket-matched scaffold embeddings into SE(3)-equivariant diffusion to ensure both structural validity and affinity optimization [23]. RADAb further explored retrieval for antibody design but was limited to sequence inverse folding with fixed structures [51].

Despite these advances, existing methods do not address the challenge of protein binder co-design, which requires simultaneous modeling of sequence and structure, nor do they provide a unified retrieval framework across modalities such as peptides and antibodies. Our method, **RADiAnce**, bridges this gap by learning a contrastive latent space that aligns interaction interfaces across diverse binder modalities and by introducing a retrieval-augmented latent diffusion model that enables cross-domain motif retrieval to guide joint sequence–structure design.

3 Method

We begin the introduction of our method by establishing the notations and definitions (§3.1). Our framework consists of a contrastive atomic variational autoencoder (§3.2) to define a compressed latent space for target-based retrieval, and a conditional latent diffusion model defined on the same latent space, which learns from the retrieved interfaces (§3.3).

3.1 Notations and Definitions

We represent a molecular complex as a pair of graphs $(\mathcal{X}, \mathcal{Y})$, where $\mathcal{X}$ denotes the binder molecule and $\mathcal{Y}$ represents the binding site on the target protein. Each graph, either $\mathcal{X}$ or $\mathcal{Y}$, comprises a node set $\mathcal{V} = \{(\mathbf{A}_i, \vec{\mathbf{X}}_i)\}$ and an edge set $\mathcal{E} = \{(\mathbf{B}_i, \mathbf{B}_{ij}) \mid i \neq j\}$, where $\mathbf{A}_i \in \mathbb{Z}^{n_i}$ and $\vec{\mathbf{X}}_i \in \mathbb{R}^{n_i \times 3}$ are the element types and 3D coordinates of the n_i atoms in block i , with $\mathbf{B}_i$ and $\mathbf{B}_{ij}$ denoting intra- and inter-block chemical bonds. We treat each standard amino acid as one block, and use principal subgraph decomposition algorithm to decompose other entities (e.g. non-standard amino acids) into blocks of frequently occurring chemical motifs. The resulting vocabulary $\mathbb{V}$ thus extends beyond standard amino acids and also includes principal subgraphs derived from non-standard components and we denote the type of block i as s_i [31]. The binding site $\mathcal{Y}$ is defined as residues within a 10Å distance cutoff from the binder, based on C_β atoms. We construct a key-value database $\mathbb{D} = \{(\mathcal{Y}_k, \mathcal{X}_k) \mid k = 1, 2, 3, \dots\}$, using the binding site $\mathcal{Y}$ as the key and the corresponding binder interface $\mathcal{X}$ as the value. Given a query binding site $\mathcal{Y}$, our framework first retrieves potential reference interfaces $\mathbb{T}(\mathcal{Y} \mid \mathbb{D}) = \{\mathcal{X}_k \mid (\mathcal{Y}_k, \mathcal{X}_k) \in \mathbb{D}, \mathcal{Y}_k \text{ similar to } \mathcal{Y}\}$ from the database and then generates a new binder by modeling the conditional distribution $p_\theta(\mathcal{X} \mid \mathcal{Y}, \mathbb{T}(\mathcal{Y} \mid \mathbb{D}))$.

3.2 Contrastive Latent Space

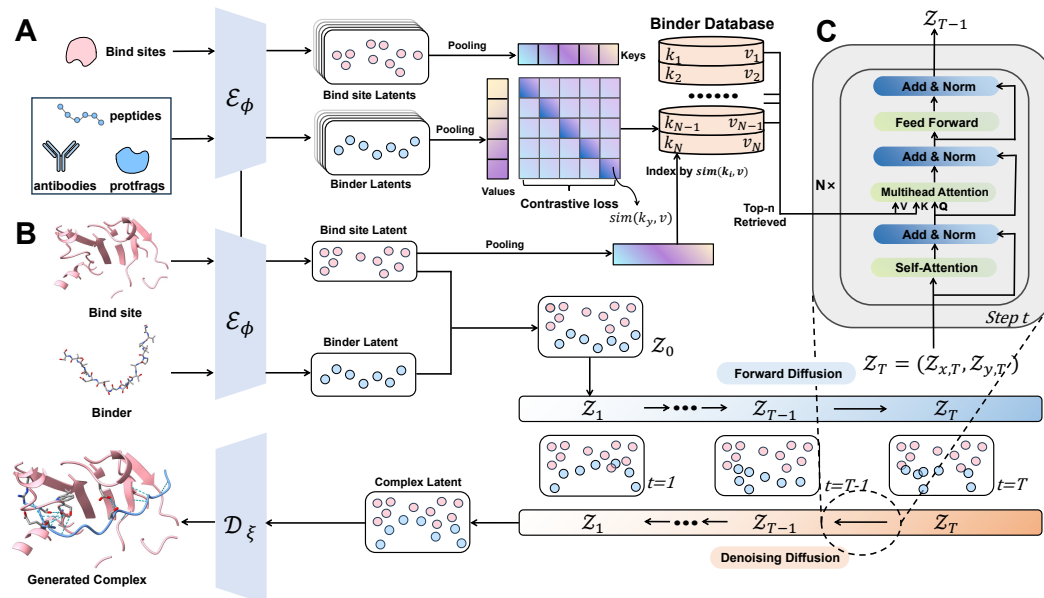


Figure 2: **Overview of RADiAance.** (A) Cross-domain binding sites and binders are encoded into key/value latents and trained with a contrastive loss to derive a retrievable binder database. (B) Contrastive VAE aligns binding site and binder latents for accurate retrieval and conditional diffusion. (C) Conditional diffusion generator leverages the retrieved latents through cross-attention at every reverse step, progressively refining noisy features into sterically and chemically consistent complexes.

To enable reliable retrieval and informative conditioning during generation, we construct a compressed contrastive latent space using an atomic variational autoencoder [31], which is trained to support both interface retrieval and atom-level reconstruction. The encoder projects the binder $\mathcal{X}$ and binding site $\mathcal{Y}$ independently into two latent point clouds, $\mathcal{Z}_x$ and $\mathcal{Z}_y$, while the decoder reconstructs the binder $\mathcal{X}$ conditioned on $\mathcal{Y}$ and $\mathcal{Z}_y$:

$$\mathcal{Z}_x = \mathcal{E}_\phi(\mathcal{X}), \quad \mathcal{Z}_y = \mathcal{E}_\phi(\mathcal{Y}), \quad \hat{\mathcal{X}} = \mathcal{D}_\xi(\mathcal{Z}_x, \mathcal{Z}_y, \mathcal{Y}), \quad (1)$$

where latent variable $\mathcal{Z}_{x/y} = \{z_i, \vec{z}_i\}$ consists of a scalar embedding $z_i = \mathbf{W}_z \mathbf{h}_i \in \mathbb{R}^d$ derived from a trainable linear transformation of the hidden states $\mathbf{h}_i$, and a 3D coordinate $\vec{z}_i \in \mathbb{R}^3$ for each block i . The decoder first predicts the block (i.e. residue) type from z_i and then reconstructs atomic coordinates $\vec{\mathbf{X}}_i$ via a flow matching process using a Gaussian prior centered at $\vec{z}_i$. The reconstruction loss includes cross-entropy (CE) over block types and mean squared error (MSE) over time-dependent vector fields:

$$\mathcal{L}_{\text{rec}}(i) = \text{CE}(p(\hat{s}_i), p(s_i)) + \mathbb{E}_{t \sim U(0,1)} \left[\text{MSE}(\hat{\mathbf{V}}_i^t, \mathbf{V}_i^t) \right], \quad (2)$$

where $\hat{s}_i$ and $\hat{\mathbf{V}}_i^t \in \mathbb{R}^{n_i \times 3}$ denote the reconstructed block type and the vector field at time step t . Note that z_i and $\vec{z}_i$ are sampled from the encoded distributions $\mathcal{N}(\mu_i, \sigma_i)$ and $\mathcal{N}(\vec{\mu}_i, \vec{\sigma}_i)$, respectively. To regularize the latent space, we apply Kullback–Leibler (KL) divergence to both scalar and coordinate embeddings [31] with prior distributions $\mathbb{N}(z_i; \mathbf{0}, \mathbf{I})$ and $\mathbb{N}(\vec{z}_i; \vec{\mathbf{c}}_i, \mathbf{I})$, where $\vec{\mathbf{c}}_i$ denotes the center of mass for block i :

$$\mathcal{L}_{\text{KL}}(i) = \lambda_1 \cdot D_{\text{KL}}(\mathcal{N}(\mathbf{0}, \mathbf{I}) \| \mathcal{N}(\mu_i, \text{diag}(\sigma_i))) + \lambda_2 \cdot D_{\text{KL}}(\mathcal{N}(\vec{\mathbf{c}}_i, \mathbf{I}) \| \mathcal{N}(\vec{\mu}_i, \text{diag}(\vec{\sigma}_i))), \quad (3)$$

where λ_1 and λ_2 define the strengths for constraining on the latent variables. To further endow the latent space with the ability to retrieve relevant interfaces given the binding site $\mathcal{Y}$, we implement contrastive training between the encoded embeddings of the binding site and the binder interfaces. Specifically, the graph-level embeddings of the binding site and the binder are derived by averaging the hidden states as $\mathbf{k} = \frac{1}{\|\mathcal{Y}\|} \sum_{i \in \mathcal{Y}} \mathbf{h}_i$, and $\mathbf{v} = \frac{1}{\|\mathcal{X}\|} \sum_{i \in \mathcal{X}} \mathbf{h}_i$, respectively. Subsequently, the

training objective maximize the similarity for authentic pairs $\{(\mathbf{k}_p, \mathbf{v}_q) \mid p = q\}$ and minimize it for mismatched pairs $\{(\mathbf{k}_p, \mathbf{v}_q \mid p \neq q\}$ in a double-sided manner [16]:

$$\mathcal{L}_{\text{retrieval}}(p) = -\log \frac{\exp(\langle \mathbf{k}_p, \mathbf{v}_p \rangle / \tau)}{\sum_{q=1}^N \exp(\langle \mathbf{k}_p, \mathbf{v}_q \rangle / \tau)} - \log \frac{\exp(\langle \mathbf{v}_p, \mathbf{k}_p \rangle / \tau)}{\sum_{q=1}^N \exp(\langle \mathbf{v}_p, \mathbf{k}_q \rangle / \tau)}, \quad (4)$$

where τ , N , and $\langle \cdot, \cdot \rangle$ are the temperature, the batch size, and the standard Euclidean inner product. The temperature $\tau > 0$ controls the sharpness of the similarity distribution. The total training objective over the dataset $\mathcal{D}$ combines reconstruction, KL regularization, and contrastive retrieval:

$$\mathcal{L}(\mathcal{D}) = \sum_{p \in \mathcal{D}} \left(\sum_{i \in \mathcal{X}_p} \mathcal{L}_{\text{rec}}(i) + \sum_{i \in \mathcal{X}_p} \mathcal{L}_{\text{KL}}(i) + \mathcal{L}_{\text{retrieval}}(p) \right). \quad (5)$$

After training, the database $\mathbb{D}$ containing complexes of interest is processed by the contrastive encoder to derive key-value pairs $\mathbb{D}^{kv} = \{(\mathbf{k}_k, \mathbf{v}_k) \mid k = 1, 2, 3, \dots\}$. During training and inference of the diffusion model, given an arbitrary binding site $\mathcal{Y}$, we use the encoder to obtain its key embedding $\mathbf{k}$, which is then used to query $\mathbb{D}^{kv}$ via inner product similarity. Additionally, to prevent trivial memorization and ensure generalization, we exclude the ground-truth ligand of the query complex from the retrieval database during this querying step. The top- n nearest neighbors are retrieved based on this process and serve as template candidates for subsequent diffusion-based generation on $\mathcal{X}$, which are denoted as $\mathbb{T}^v = \{\mathbf{v}_k \mid k = 1, 2, \dots, n\}$.

3.3 Retrieval-Conditioned Latent Diffusion

Latent Diffusion Given that the variational autoencoder (VAE) encodes the binding site and binder into continuous latent representations, we directly define a diffusion process over these latent variables to model the conditional distribution $p(\mathcal{Z}_x | \mathcal{Z}_y, \mathbb{T}^v)$, where $\mathbb{T}^v$ contains the value vectors of the retrieved template interfaces. The forward process gradually corrupts the initial latent variables with Gaussian noise across a predefined number of steps T , transforming them into the standard Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Conversely, the reverse process denoises these samples step by step to recover the original latent vectors. Let $\bar{\mathbf{u}}_i^t = [\mathbf{z}_i^t, \mathbf{z}_i^t]$ denote the state of block i at timestep t , the forward process can be defined as follows [31]:

$$q(\bar{\mathbf{u}}_i^t | \bar{\mathbf{u}}_i^{t-1}) = \mathcal{N}(\bar{\mathbf{u}}_i^t; \sqrt{1 - \beta^t} \cdot \bar{\mathbf{u}}_i^{t-1}, \beta^t \mathbf{I}), \quad (6)$$

$$q(\bar{\mathbf{u}}_i^t | \bar{\mathbf{u}}_i^0) = \mathcal{N}(\bar{\mathbf{u}}_i^t; \sqrt{\bar{\alpha}^t} \cdot \bar{\mathbf{u}}_i^0, (1 - \bar{\alpha}^t) \mathbf{I}), \quad (7)$$

where $\bar{\alpha}^t = \prod_{s=1}^{s=t} (1 - \beta^s)$, and β^t is the noise scale following the cosine schedule [43]. Therefore, state at time t can be sampled given $\bar{\mathbf{u}}_i^0$ as $\bar{\mathbf{u}}_i^t = \sqrt{\bar{\alpha}^t} \bar{\mathbf{u}}_i^0 + \sqrt{1 - \bar{\alpha}^t} \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The reverse process learns to denoise the latent samples [21] with retrieved templates as contexts:

$$p_{\theta}(\bar{\mathbf{u}}_i^{t-1} | \mathcal{Z}_x^t, \mathcal{Z}_y, \mathbb{T}^v) = \mathcal{N}(\bar{\mathbf{u}}_i^{t-1}; \bar{\boldsymbol{\mu}}_{\theta}(\mathcal{Z}_x^t, \mathcal{Z}_y, \mathbb{T}^v), \beta^t \mathbf{I}), \quad (8)$$

$$\bar{\boldsymbol{\mu}}_{\theta}(\mathcal{Z}_x^t, \mathcal{Z}_y, \mathbb{T}^v) = (\bar{\mathbf{u}}_i^t - \frac{\beta^t}{\sqrt{1 - \bar{\alpha}^t}} \epsilon_{\theta}(\mathcal{Z}_x^t, \mathcal{Z}_y, \mathbb{T}^v, t)[i]) / \sqrt{\alpha^t}, \quad (9)$$

where $\alpha^t = 1 - \beta^t$. The parameterization of the denoising kernel ϵ_{θ} will be illustrated in the following paragraph. The training process minimizes the discrepancy between the predicted and true noise added during the forward process:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t \sim \mathcal{U}(1 \dots T)} [\sum_i \|\epsilon_i - \epsilon_{\theta}(\mathcal{Z}_x^t, \mathcal{Z}_y, \mathbb{T}^v, t)[i]\|^2 / |\mathcal{Z}_x^t|], \quad (10)$$

where $|\mathcal{Z}_x^t|$ denotes the size of the latent point cloud.

Template Integration The denoising kernel $\epsilon_{\theta}(\mathcal{Z}_x^t, \mathcal{Z}_y, \mathbb{T}^v, t)$ is instantiated by an E(3)-equivariant transformer (EPT) [25] augmented with a cross-attention mechanism to incorporate information from the retrieved templates $\mathbb{T}^v$. This cross-attention module is placed between the self-attention and feed-forward layers in the original EPT architecture (Figure 2). Let $\mathbf{H} \in \mathbb{R}^{m \times h}$ denote the invariant hidden states of the m latent nodes (blocks) with hidden dimension h , and let $\mathbf{T} \in \mathbb{R}^{n \times h}$ represent the embeddings of the retrieved top- n templates. The queries, keys, and values are computed as:

$$\mathbf{Q} = \mathbf{H} \mathbf{W}^Q, \quad \mathbf{K} = \mathbf{T} \mathbf{W}^K, \quad \mathbf{V} = \mathbf{T} \mathbf{W}^V, \quad (11)$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{h \times h}$ are learnable matrices for linear transformations. The hidden states are subsequently updated using a residual connection with the cross-attention output:

$$\mathbf{H}' = \text{LN}(\mathbf{H} + \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{h}})\mathbf{V}), \quad (12)$$

where LN denotes the layernorm [6]. This design enables the model to effectively integrate contextual information from the retrieved templates $\mathbb{T}$, guiding the refinement of latent variables toward mimicking key interaction patterns. We have also explored alternative conditioning mechanisms for incorporating template information, but these variants exhibit downgraded performance compared to the current cross-attention strategy, as demonstrated in Appendix D.

3.4 Inference Algorithm

During inference, the framework encodes the target binding-site graph $\mathcal{Y}$ into a key $\mathbf{k}_y$, retrieves the top- K latent binders from the database via inner-product similarity to form the prompt set $\mathbb{T}^v$, and performs prompt-conditioned reverse diffusion from Gaussian noise to a clean latent $\mathcal{Z}_0$, which is then decoded into the binder $\hat{\mathcal{X}}$. Owing to the modularity of retrieval and generation, the latent database $\mathbb{D}$ can be freely replaced at inference. The overall workflow is as follows:

Algorithm 1 Inference Workflow

```

1: Input: Binding site structure  $\mathcal{Y}$ 
2:  $\mathbf{k}_y \leftarrow \mathcal{E}_\phi(\mathcal{Y})$                                 {Encode binding site into query key}
3: for each  $\mathbf{v}^{(j)} \in \mathbb{D}$  do
4:    $s_j \leftarrow \langle \mathbf{k}_y, \mathbf{v}^{(j)} \rangle$                     {Compute similarity by inner product}
5: end for
6:  $\mathcal{I} \leftarrow \text{TopK}(\{s_j\})$                             {Indices of top- $K$  highest similarity scores}
7:  $\mathbb{T}^v \leftarrow \{\mathbf{v}^{(k)} : k \in \mathcal{I}\}$                 {Retrieve binder latents to form prompt features}
8:  $\mathcal{Z}_T \sim \mathcal{N}(0, \mathbf{I})$                                 {Initialize latent variable from Gaussian}
9: for  $t = T$  down to 1 do
10:   $\epsilon \leftarrow \epsilon_\theta(\mathcal{Z}_t^t, \mathcal{Z}_y, \mathbb{T}^v, t)$                 {Predict noise}
11:   $\mathcal{Z}_{t-1} \leftarrow \text{Denoise}(\mathcal{Z}_t, \epsilon)$                 {Refine latent variable}
12: end for
13:  $\hat{\mathcal{X}} \leftarrow \mathcal{D}_\xi(\mathcal{Z}_0)$                             {Decode final latent to molecular graph}
14: Output: Generated binder structure  $\hat{\mathcal{X}}$ 

```

4 Experiments

We evaluate **RADiA**nc from two key aspects: **Retrieval Reliability** (Section 4.1) and **Generation Performance** (Section 4.2), using peptide, antibody, and protein fragments from existing literature. For peptide design, we adopt PepBench [30], which includes 4,157 training and 114 validation complexes, with 93 test cases from the LNR benchmark [49]. For antibodies, we follow the literature [31] to use 9,473 training and 400 validation entries from SABDab [14], and 60 test cases from the RAbD benchmark [3]. We further include ProtFrag [30], a dataset of 70,498 monomer-derived protein fragments, to assess the cross-domain referencing ability of our framework. To prevent data leakage, all retrieval operations are strictly limited to training data. Lastly, we demonstrate the potential of **RADiA**nc in real-world scenarios with a pipeline for de novo antibody design without relying on predefined frameworks (Section 4.3).

4.1 Retrieval Reliability

Setup We first assess whether the latent embeddings produced by the variational encoder can reliably retrieve relevant interfaces, as accurate retrieval is essential for successful downstream generation. Although the model is trained on cross-domain data, retrieval evaluation is conducted using only the in-domain database (i.e., antibody or peptide) to create a more stringent testing scenario. Ground-truth binders are also included in the database for this setting, to evaluate whether the models can accurately retrieve the ground truths. We further perform ablation studies to assess

the effect of cross-domain training and the importance of jointly optimizing the VAE and contrastive objectives. As a baseline, we include random retrieval to illustrate the degree of structural and biological relevance achieved by our Contrastive VAE (CVAE).

Metrics We evaluate retrieval reliability from two complementary perspectives: recall and precision. To assess recall, we compute the proportion of cases where the ground-truth binder ranks within the top $N\%$ of retrieved candidates (RC), denoted as **RC- $N\%$** . For precision, we assess the similarity between the top-10 retrieved interfaces and the ground-truth binder in terms of shared interaction patterns (e.g., hydrogen bonds, hydrophobic contacts, etc) [2]. This is quantified by Interaction Type Overlap (**ITO**), which measures agreement between two distributions consisting of interaction types, with details provided in Appendix B.

Table 1: Results of recovery metrics for retrieval models. Columns “Pep”, “Ab”, and “ProtFrag” indicate whether peptide, antibody, and protein fragment data are used during training (✓ for included, ✗ for excluded). “All loss” = ✗ indicates using contrastive loss only.

Model	Pep	Ab	ProtFrag	All loss	ITO(%)	RC-0.1%	RC-0.5%	RC-5%
Antibody								
Random	✗	✗	✗	✓	17.71	0.1	0.5	5.0
CVAE	✗	✓	✗	✓	28.21	25.00	68.33	100.0
	✓	✓	✗	✓	34.98	60.00	85.00	100.0
	✓	✓	✓	✗	38.85	5.00	30.00	100.0
	✓	✓	✓	✓	43.93	66.67	96.67	100.0
Peptide								
Random	✗	✗	✗	✓	36.55	0.1	0.5	5.0
CVAE	✓	✗	✗	✓	53.39	3.23	5.38	25.81
	✓	✓	✗	✓	53.64	4.30	7.53	33.33
	✓	✓	✓	✗	54.02	2.15	8.60	31.18
	✓	✓	✓	✓	61.41	11.58	22.58	67.74

Results Table 1 supports the following three key findings: 1) Our contrastive variational autoencoder (CVAE) reliably retrieves relevant interfaces, with RC-5% achieving 100% for antibodies and 67.74% for peptides; 2) Exposure to diverse interface types during training significantly enhances generalization, as incorporating cross-domain training, particularly with intra-protein interfaces from the ProtFrag dataset, improves performance to a large extent; 3) Joint training with both contrastive and VAE losses is essential, as removing the VAE loss (i.e., using an encoder-only model) significantly degrades performance, likely because the reconstruction objective reinforces memorization of meaningful interaction patterns between binding sites and binders. These results demonstrate that our CVAE enables effective retrieval with high recall and precision, forming a solid foundation for retrieval-augmented generation.

4.2 Generation Performance

Metrics To evaluate our model’s effectiveness in designing peptide–protein and antibody–antigen complexes, we adopt several established structural and biochemical metrics. Amino Acid Recovery (**AAR**) measures sequence fidelity as the proportion of residues in the generated peptide that match the reference, based on canonical sequence alignment [42, 20]. Complex **RMSD** quantifies structural accuracy via the root mean square deviation of C_α coordinates, after aligning binders on the target protein. Interaction Site Match (**ISM**) assesses how well key physicochemical interactions (e.g., hydrogen bonds, hydrophobic contacts) are reproduced, reflecting interface realism [2], computation details are in Appendix A. The change in binding free energy ($\Delta\Delta G$) captures affinity shifts, where negative values indicate stronger binding. **IMP** reports the fraction of targets for which generated binders outperform the native ligand in predicted affinity. For each test complex, we sample 100 candidates per model and evaluate all metrics to ensure statistical robustness.

4.2.1 Peptide Sequence-Structure Codesign

Baselines We compare our method against several representative generative frameworks. RFDiffusion [53] generates peptide backbones, followed by one round of inverse folding using ProteinMPNN [13] and Rosetta-based atomistic relaxation [5]. To avoid overfitting physical metrics,

we restrict this process to a single iteration. We also include peptide-specific models such as PepFlow [36] and PepGLAD [30], which leverage multi-modal flow matching and latent diffusion to jointly model sequence and structure. In addition, we evaluate UniMoMo [31], a unified framework for small molecules, peptides, and antibodies, using its best peptide-generation configuration without incorporating the small molecule dataset.

Results As shown in Table 2, our method consistently outperforms all baselines across multiple evaluation metrics. Compared to existing frameworks, our model achieves higher recovery of the ground-truth sequences and structures, effectively leveraging knowledge from existing interfaces. Notably, our **RADiAnce** surpasses baselines by a large margin in terms of ISM, indicating it excels at copying relevant interaction patterns from retrieved interfaces. Further ablation shows improved generation when retrieving cross-domain data over single-domain H.

Table 2: Results for peptide sequence-structure codesign.

Model	AAR (%)	RMSD (Å)	$\Delta\Delta G$ (kJ/mol)	IMP (%)	ISM (%)
RFDiffusion	34.68	4.69	24.78	5.38	28.38
PepFlow	35.47	2.87	15.71	14.13	27.83
PepGLAD	38.62	2.74	15.26	16.13	32.63
UniMoMo	38.69	2.31	2.409	40.86	49.13
RADiAnce	39.42	2.29	1.963	41.94	52.15

4.2.2 Antibody Sequence-Structure Codesign

Setup We define Complementarity-Determining Regions (CDRs) according to the Chothia numbering system [12], as adopted in previous literature Luo et al. [40], Kong et al. [31]. For each antigen-antibody complex in the test set, we randomly mask one CDR loop and treat the remaining framework and unmasked CDRs as conditioning context. The model is then tasked with generating the missing CDR regions.

Baselines We benchmark our approach against representative state-of-the-art generative models in molecular and antibody design: MEAN [28] uses a multi-channel SE(3)-equivariant GNN to iteratively generate CDR sequences and structures. DyMEAN [29] extends MEAN to full-atom generation and handles uncertain antibody docking poses. DiffAb [40] applies diffusion modeling to jointly generate amino acid types, C_α positions, and orientations. GeoAB [37] combines heterogeneous residue encoders with geometric energy priors for realistic structure generation. UniMoMo [31] unifies the generation of small molecules, peptides, and antibodies within a shared representation framework.

Results As shown in Table 3, our method consistently outperforms all baselines on CDR sequence-structure co-design. It achieves notable improvements in sequence recovery, structural accuracy, and binding affinity across all CDR regions, demonstrating the benefits of learning from retrieved interfaces. Moreover, consistent with earlier peptide results, cross-domain retrieval further boosts performance over single-domain retrieval (Appendix H), highlighting the ability of our model to generalize shared interaction patterns across different domains. We also provide a qualitative example in Appendix G, illustrating how the model leverages retrieved interfaces from other domains to correct and enhance generated interactions.

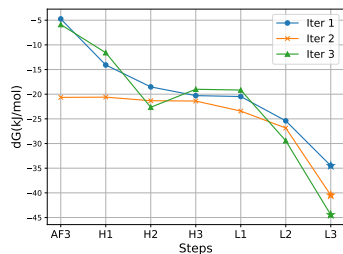


Figure 3: Free energy changes during iterative CDR design.

4.3 Antibody Design Without Bound Structural scaffold

To better align with real-world applications, we propose a pipeline for designing all six CDRs of an antibody without access to its bound structure. Beginning with a predefined framework, we iteratively

Table 3: Results for antibody CDRs sequence-structure codesign.

Model	CDR	AAR	RMSD	$\Delta\Delta G$	IMP	ISM	CDR	AAR	RMSD	$\Delta\Delta G$	IMP	ISM
MEAN		76.78	0.9660	3.194	20.00	66.11		68.84	1.075	4.828	28.30	53.42
dyMEAN		84.25	0.8061	18.42	1.695	56.10		81.28	0.8731	17.88	1.695	52.23
GeoAB-R		80.00	0.5472	254.34	25.00	69.30		79.00	0.4670	2.757	28.33	55.13
DiffAb	H1	83.25	0.6278	-3.625	87.93	75.02	L1	83.25	0.6278	-3.424	93.10	75.02
GeoAB-D		85.48	0.4290	-6.925	83.33	74.76		80.65	0.3352	-6.692	78.33	56.32
Unimomo		86.72	0.3728	-6.247	88.33	98.01		80.28	0.3177	2.374	93.33	96.37
RADiAnce		90.83	0.2977	-8.221	96.67	98.36		86.20	0.2907	-7.170	95.00	98.31
MEAN		54.16	0.9928	49.69	6.67	26.98		64.24	0.7657	13.08	10.00	72.37
dyMEAN		73.62	0.9453	18.45	3.390	48.35		84.35	0.7514	17.58	0.00	59.41
GeoAB-R		54.21	0.7125	30.23	16.67	27.67		75.67	0.4742	5.691	35.00	72.58
DiffAb	H2	67.51	0.4378	-1.895	77.58	29.11	L2	72.87	0.3910	-8.397	86.21	72.12
GeoAB-D		75.04	0.4828	3.982	56.67	28.95		83.06	0.3462	-5.067	76.67	73.16
Unimomo		75.74	0.2207	-2.710	88.33	86.31		84.36	0.1592	-4.779	95.00	96.55
RADiAnce		79.20	0.2135	-2.370	83.33	87.83		86.68	0.1558	-3.857	95.00	98.52
MEAN		26.19	1.076	104.88	8.30	12.81		53.72	1.100	9.835	18.33	49.52
dyMEAN		31.65	1.123	24.23	1.695	41.67		72.86	0.8722	18.39	3.390	57.16
GeoAB-R		32.04	1.682	42.71	0.33	11.67		64.16	0.7213	5.161	35.00	51.27
DiffAb	H3	51.84	1.584	-1.580	56.90	27.40	L3	64.16	0.8005	-6.685	86.21	57.70
GeoAB-D		44.82	1.565	34.73	41.67	22.50		64.16	0.8005	-3.445	80.00	54.76
Unimomo		52.34	1.040	-7.438	65.00	71.52		72.36	0.5003	-8.797	95.00	90.94
RADiAnce		54.66	0.9443	-6.236	71.67	71.64		76.75	0.4478	-9.207	90.00	94.82

alternate between CDR redesign using our **RADiAnce** and structure prediction with AlphaFold3 [1], progressively refining both the sequence and structure toward an optimal binder.

To design a novel antibody targeting the HIV-1 primary receptor CD4 (PDBID: 3O2D) [15], we use Ramucirumab (DrugBank ID: DB05578) as the structural scaffold, which originally targets VEGFR2 [24] but with good drug developability. We identify the CD4 epitope via AlphaFold3-guided docking [1], and iteratively redesign the complementarity-determining regions (CDRs) by knocking out and regenerating each loop using our model. After each iteration, the antibody is relaxed with Rosetta [5] and re-docked against CD4. This redesign-dock cycle is repeated three times to progressively refine binding. We use PyRosetta to estimate binding free energy (ΔG), considering

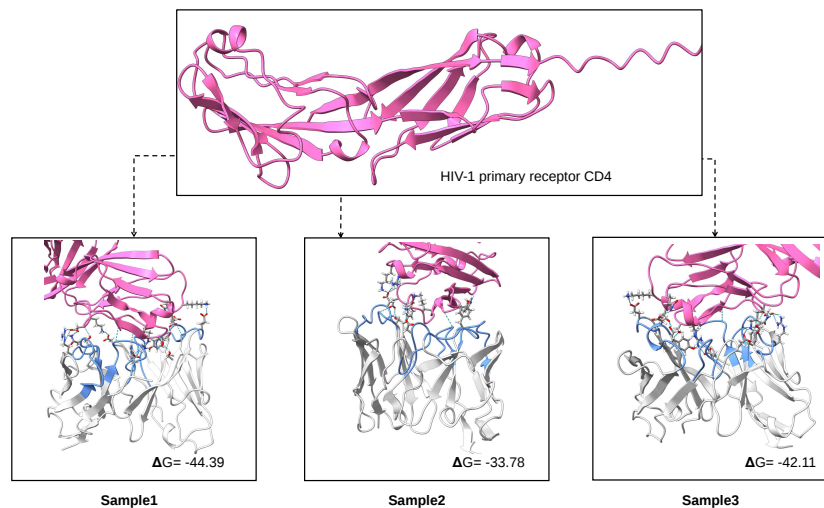


Figure 4: Examples of de novo antibody designs targeting the HIV-1 receptor CD4. Each case shows the final docked complex with the redesigned antibody interacting with the target epitope.

designs with $\Delta G < 0$ as indicative of favorable antigen engagement, even without a fixed antibody framework. We visualize the final antibody-antigen complex and highlight hydrogen bonds at the interface (Figure 4). Figure 3 illustrates the ΔG trajectory over design iterations, showing consistent affinity improvements, with the final design reaching $\Delta G < 0$. These results highlight the potential

Table 4: Comparison of retrieval strategies for HCDR3 and peptide codesign tasks.

Strategy	Retrieved	HCDR3				Peptide			
		Similarity	AAR	RMSD	Diversity	Similarity	AAR	RMSD	Diversity
Top-N	0	N/A	51.55	0.9525	0.0524	N/A	37.94	2.38	0.5042
Top-N	1	57.62±5.3	53.26	0.9431	0.0404	55.56±4.5	39.03	2.38	0.4846
Top-N	10	54.65±4.9	54.66	0.9443	0.0484	51.67±3.6	39.12	2.31	0.5129
Top-N	20	50.43±4.8	54.63	0.9682	0.0485	50.24±3.3	39.42	2.29	0.5227
Top-N	40	46.04±4.7	54.06	0.9653	0.0494	47.68±3.1	38.16	2.28	0.5273
Reverse-N	10	-49.86±16.7	52.31	0.9718	0.0468	-26.09±12.5	36.33	3.09	0.5267
Random	10	17.34±6.9	51.72	0.9665	0.0649	11.78±8.16	38.04	2.56	0.5580

Table 5: Performance with adaptive retrieval for HCDR3 and Peptide codesign tasks.

Settings	AAR (%)	RMSD (Å)	Diversity	$\Delta\Delta G$ (kJ/mol)	FoldX ΔG (kJ/mol)	ISM (%)	DockQ
HCDR3							
Fixed Top-N	54.63	0.9682	0.0485	-6.236	-8.4250	71.64	0.9582
Adaptive Retrieval	54.71	0.9147	0.0523	-8.938	-8.6180	71.61	0.9604
Peptide							
Fixed Top-N	39.42	2.29	0.5227	1.963	-7.99	52.15	0.7698
Adaptive Retrieval	39.22	2.27	0.5400	1.427	-7.97	52.89	0.7762

of our framework to generate *de novo* antibodies capable of strong and specific epitope binding without requiring a pre-solved complex.

5 Analysis

5.1 Effect of Retrieval Quantity and Quality

To better understand how retrieval influences our model performance, we evaluate **RADiA**nce under different retrieval modes and varying numbers of retrieved samples on codesign tasks.

Results Table 4 summarizes the relationship between retrieval strategy and model performance. Increasing the number of retrieved samples generally improves performance, but excessively large pools introduce low-similarity samples that add noise and degrade quality. Random and reverse retrieval perform poorly for the same reason, while the no-retrieval baseline is markedly worse than any retrieval-based setting, highlighting the importance of retrieval. We also find that larger pools and the no-retrieval case yield higher diversity, whereas retrieving only one sample produces the lowest. This aligns with our expectation, as limited retrieval can overly constrain the generative space.

5.2 Adaptive Retrieval Approach

Building on the results above, we replace fixed top-*N* retrieval with an adaptive similarity cutoff: only neighbors above a task-specific threshold are retained, allowing the model to automatically use more samples when many are relevant and fewer when they are not. As shown in Table 5, this adaptive strategy consistently improves multiple metrics by preserving the benefits of larger retrieval pools while avoiding noise from less relevant neighbors.

6 Conclusions and Limitations

We propose **RADiA**nce, a retrieval-augmented framework that unifies retrieval and generation in a contrastive latent space to improve rational and interpretable binder design. A contrastive VAE enables efficient cross-domain motif retrieval, while cross-attention-guided latent diffusion enhances sequence–structure codesign. Despite its effectiveness, performance depends on retrieval quality, motivating future work on improving structural descriptors and exploring more robust and structure-aware conditional integration strategies. Beyond methodological advances, our framework enables a transparent and interpretable binder design process, potentially accelerating biologic development through more controllable generation of therapeutically meaningful molecules.

Acknowledgements

This work is jointly supported by the National Key R&D Program of China (No.2022ZD0160502), the National Natural Science Foundation of China (No. 61925601, No. 62376276, No. 62276152), and Beijing Nova Program (20230484278).

References

- [1] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.
- [2] M. F. Adasme, K. L. Linnemann, S. N. Bolz, F. Kaiser, S. Salentin, V. J. Haupt, and M. Schroeder. PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic Acids Research*, 49(W1):W530–W534, May 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab294. URL <https://doi.org/10.1093/nar/gkab294>. _eprint: <https://academic.oup.com/nar/article-pdf/49/W1/W530/38841758/gkab294.pdf>.
- [3] J. Adolf-Bryfogle, O. Kalyuzhnyi, M. Kubitz, B. D. Weitzner, X. Hu, Y. Adachi, W. R. Schief, and R. L. Dunbrack Jr. Rosettaantibodydesign (rabd): A general framework for computational antibody design. *PLoS computational biology*, 14(4):e1006112, 2018.
- [4] R. Akbar, P. A. Robert, C. R. Weber, M. Widrich, R. Frank, M. Pavlović, L. Scheffer, M. Chernigovskaya, I. Snapkov, A. Slabodkin, et al. In silico proof of principle of machine learning-based antibody design at unconstrained scale. In *MAbs*, volume 14, page 2031482. Taylor & Francis, 2022.
- [5] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O’Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.
- [6] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [7] S. Basu and B. Wallner. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.
- [8] G. Bhardwaj, V. K. Mulligan, C. D. Bahl, J. M. Gilmore, P. J. Harvey, O. Cheneval, G. W. Buchko, S. V. Pulavarti, Q. Kaas, A. Eletsky, et al. Accurate de novo design of hyperstable constrained peptides. *Nature*, 538(7625):329–335, 2016.
- [9] S. Bheemireddy and N. Srinivasan. Computational study on the dynamics of mycobacterium tuberculosis rna polymerase assembly. In *Prokaryotic Gene Regulation: Methods and Protocols*, pages 61–79. Springer, 2022.
- [10] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer. Retrieval-Augmented Diffusion Models. *Advances in Neural Information Processing Systems*, 35:15309–15324, Dec. 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/62868cc2fc1eb5cdf321d05b4b88510c-Abstract-Conference.html.
- [11] T.-H. Chiang, Y.-C. Tseng, and Y.-C. Tseng. A multi-embedding neural model for incident video retrieval. *Pattern Recognition*, 130:108807, Oct. 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2022.108807. URL <https://www.sciencedirect.com/science/article/pii/S0031320322002886>.
- [12] C. Chothia and A. M. Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of molecular biology*, 196(4):901–917, 1987.
- [13] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas, N. Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

- [14] J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C. M. Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014.
- [15] M. M. Freeman, M. S. Seaman, S. Rits-Volloch, X. Hong, C.-Y. Kao, D. D. Ho, and B. Chen. Crystal structure of hiv-1 primary receptor cd4 in complex with a potent antiviral antibody. *Structure*, 18(12):1632–1641, 2010.
- [16] B. Gao, B. Qiang, H. Tan, Y. Jia, M. Ren, M. Lu, J. Liu, W.-Y. Ma, and Y. Lan. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36:44595–44614, 2023.
- [17] Z. Geng, S. Xie, Y. Xia, L. Wu, T. Qin, J. Wang, Y. Zhang, F. Wu, and T.-Y. Liu. De novo molecular generation via connection-aware motif mining. *arXiv preprint arXiv:2302.01129*, 2023.
- [18] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang. Retrieval Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938. PMLR, Nov. 2020. URL <https://proceedings.mlr.press/v119/guu20a.html>. ISSN: 2640-3498.
- [19] H. He, B. He, L. Guan, Y. Zhao, F. Jiang, G. Chen, Q. Zhu, C. Y.-C. Chen, T. Li, and J. Yao. De novo generation of SARS-CoV-2 antibody CDRH3 with a pre-trained generative large language model. *Nature Communications*, 15(1):1–19, Aug. 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-50903-y. URL <https://www.nature.com/articles/s41467-024-50903-y>. Number: 1 Publisher: Nature Publishing Group.
- [20] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.
- [21] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [22] P. Hosseinzadeh, P. R. Watson, T. W. Craven, X. Li, S. Rettie, F. Pardo-Avila, A. K. Bera, V. K. Mulligan, P. Lu, A. S. Ford, et al. Anchor extension: a structure-guided approach to design cyclic peptides targeting enzyme active sites. *Nature Communications*, 12(1):3384, 2021.
- [23] Z. Huang, L. Yang, X. Zhou, C. Qin, Y. Yu, X. Zheng, Z. Zhou, W. Zhang, Y. Wang, and W. Yang. Interaction-based retrieval-augmented diffusion models for protein-specific 3d molecule generation. In *International Conference on Machine Learning*, 2024.
- [24] M. Javle, E. C. Smyth, and I. Chau. Ramucirumab: successfully targeting angiogenesis in gastric cancer. *Clinical cancer research*, 20(23):5875–5881, 2014.
- [25] R. Jiao, X. Kong, Z. Yu, W. Huang, and Y. Liu. Equivariant pretrained transformer for unified geometric learning on multi-domain 3d molecules. *arXiv preprint arXiv:2402.12714*, 2024.
- [26] W. Jin, R. Barzilay, and T. Jaakkola. Antibody-antigen docking and design via hierarchical structure refinement. In *International Conference on Machine Learning*, pages 10217–10227. PMLR, 2022.
- [27] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.
- [28] X. Kong, W. Huang, and Y. Liu. Conditional antibody design as 3d equivariant graph translation. In *The Eleventh International Conference on Learning Representations*, 2023.
- [29] X. Kong, W. Huang, and Y. Liu. End-to-end full-atom antibody design. In *International Conference on Machine Learning*, pages 17409–17429. PMLR, 2023.
- [30] X. Kong, Y. Jia, W. Huang, and Y. Liu. Full-atom peptide design with geometric latent diffusion. *Advances in Neural Information Processing Systems*, 37:74808–74839, 2025.

- [31] X. Kong, Z. Zhang, Z. Zhang, R. Jiao, J. Ma, K. Liu, W. Huang, and Y. Liu. Unimomo: Unified generative modeling of 3d molecules for de novo binder design, 2025. URL <https://arxiv.org/abs/2503.19300>.
- [32] S. Lee, K. Kreis, S. P. Veccham, M. Liu, D. Reidenbach, S. Paliwal, A. Vahdat, and W. Nie. Molecule generation with fragment retrieval augmentation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 132463–132490. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ef107bb91dae96368c856b1064370bd0-Paper-Conference.pdf.
- [33] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, et al. Imgt®, the international immunogenetics information system®. *Nucleic acids research*, 37(suppl_1):D1006–D1012, 2009.
- [34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- [35] J. Li, T. Chen, S. Luo, C. Cheng, J. Guan, R. Guo, S. Wang, G. Liu, J. Peng, and J. Ma. Hotspot-driven peptide design via multi-fragment autoregressive extension. *arXiv preprint arXiv:2411.18463*, 2024.
- [36] J. Li, C. Cheng, Z. Wu, R. Guo, S. Luo, Z. Ren, J. Peng, and J. Ma. Full-atom peptide design based on multi-modal flow matching. In *Forty-first International Conference on Machine Learning*, 2024.
- [37] H. Lin, Y. Huang, O. Zhang, Y. Liu, L. Wu, S. Li, Z. Chen, and S. Z. Li. Functional-group-based diffusion for pocket-specific molecule generation and elaboration. *Advances in Neural Information Processing Systems*, 36, 2024.
- [38] H. Lin, L. Wu, Y. Huang, Y. Liu, O. Zhang, Y. Zhou, R. Sun, and S. Z. Li. Geoab: Towards realistic antibody design and reliable affinity maturation. In *Forty-first International Conference on Machine Learning*, 2024.
- [39] H. Lin, O. Zhang, H. Zhao, D. Jiang, L. Wu, Z. Liu, Y. Huang, and S. Z. Li. Ppflow: Target-aware peptide design with torsional flow matching. *bioRxiv*, pages 2024–03, 2024.
- [40] S. Luo, Y. Su, X. Peng, S. Wang, J. Peng, and J. Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.
- [41] K. Martinkus, J. Ludwiczak, W.-C. Liang, J. Lafrance-Vanasse, I. Hotzel, A. Rajpal, Y. Wu, K. Cho, R. Bonneau, V. Gligorić, et al. Abdiffuser: full-atom generation of in-vitro functioning antibodies. *Advances in Neural Information Processing Systems*, 36, 2024.
- [42] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.
- [43] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.
- [44] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- [45] S. Sachdeva, H. Joo, J. Tsai, B. Jasti, and X. Li. A rational approach for creating peptides mimicking antibody binding. *Scientific reports*, 9(1):997, 2019.
- [46] K. Saka, T. Kakuzaki, S. Metsugi, D. Kashiwagi, K. Yoshida, M. Wada, H. Tsunoda, and R. Teramoto. Antibody design using lstm based deep generative model from phage display library for affinity maturation. *Scientific reports*, 11(1):5852, 2021.

- [47] A. Shah, C. Guntuboina, and A. B. Farimani. Peptide-GPT: Generative Design of Peptides using Generative Pre-trained Transformers and Bio-informatic Supervision, Oct. 2024. URL <http://arxiv.org/abs/2410.19222>. arXiv:2410.19222 [cs].
- [48] H. Su, G. Chen, H. Xie, W. Li, M. Xiong, J. He, H. Hu, W. Zhao, Q. Shao, M. Li, et al. Structure-based design of potent and selective inhibitors targeting ripk3 for eliminating on-target toxicity in vitro. *Nature Communications*, 16(1):1–15, 2025.
- [49] T. Tsaban, J. K. Varga, O. Avraham, Z. Ben-Aharon, A. Khramushin, and O. Schueler-Furman. Harnessing protein folding neural networks for peptide–protein docking. *Nature communications*, 13(1):176, 2022.
- [50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [51] Z. Wang, Y. Ji, J. Tian, and S. Zheng. Retrieval Augmented Diffusion Model for Structure-informed Antibody Design and Optimization, Oct. 2024. URL <http://arxiv.org/abs/2410.15040>. arXiv:2410.15040 [cs].
- [52] P. T. Wasdin, N. V. Johnson, A. K. Janke, S. Held, T. M. Marinov, G. Jordaan, L. Vandenabeele, F. Pantouli, R. A. Gillespie, M. J. Vukovich, C. M. Holt, J. Kim, G. Hansman, J. Logue, H. Y. Chu, S. F. Andrews, M. Kanekiyo, G. A. Sautto, T. M. Ross, D. J. Sheward, J. S. McLellan, A. A. Abu-Shmais, and I. S. Georgiev. Generation of antigen-specific paired chain antibody sequences using large language models, Mar. 2025. URL <https://www.biorxiv.org/content/10.1101/2024.12.20.629482v3>. Pages: 2024.12.20.629482 Section: New Results.
- [53] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.
- [54] M. Wolff and A. McPherson. Antibody-directed drug discovery. *Nature*, 345(6273):365–366, 1990.
- [55] D. Xu, Z. Yang, K. chun Wong, Z. Zhu, J. Li, and J. Ji. Reimagining target-aware molecular generation through retrieval-enhanced aligned diffusion, 2025. URL <https://arxiv.org/abs/2506.14488>.
- [56] J. Zhu, J. Zhu, A. Negri, D. Provasi, M. Filizola, B. S. Collier, and T. A. Springer. Closed headpiece of integrin $\alpha\text{IIB}\beta 3$ and its complex with an $\alpha\text{IIB}\beta 3$ -specific antagonist that does not induce opening. *Blood, The Journal of the American Society of Hematology*, 116(23): 5050–5059, 2010.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: See Section 1, Section 3, Section 4

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Section 6 discussed the limitations and future works.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper focuses on empirical methodology and experiments without presenting any formal theoretical results or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper provides all necessary details to reproduce the main experimental results, including descriptions of the model architecture(Section 3), training objectives(Section 3.3), datasets(Section 4), and evaluation metrics(Section 4.1, 4.2).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code link is at Appendix K.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See Section 4, Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper uses sufficiently large test samples to ensure minimal variance, with an analysis of error bars provided in Appendix J.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research fully complies with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses the potential positive societal impact in Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work has no foreseeable high-risk misuse. The models are trained on publicly available datasets and do not involve generative tasks applicable to harmful content creation.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: See Section 4.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: See Appendix K, the codes are documented with usage instructions and license details.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: This study does not involve any human subjects or crowdsourcing experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: This study does not involve any human subjects or crowdsourcing experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The research does not involve LLMs in the methodology or experiments.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Implementation of Interaction Site Match (ISM)

The **Interaction Site Match (ISM)** metric evaluates the fine-grained accuracy of predicted molecular interactions by assessing their strict consistency with the ground truth at the atomic level [2]. An interaction is considered a strict match only when the predicted interaction has the same interaction type (such as hydrogen bond or salt bridge), originates from the same residue on the target and terminates at the same residue as in the reference. To compute ISM, we iterate over all annotated residue-level interactions in the reference structure and search for corresponding interactions in the predicted structure. If a predicted interaction matches a reference one in both ligand atom and interaction type, it is counted as a match. The final ISM score is calculated as the ratio between the number of strictly matched interactions and the total number of interactions in the reference:

In cases where the reference structure contains no annotated interactions (i.e., the denominator is zero), the ISM score is defined as NaN and excluded from downstream averaging. This prevents distortion of the overall metric by non-informative cases and ensures that the evaluation focuses on structurally meaningful interaction sites. ISM thus captures the model's ability to recover detailed and chemically precise binding patterns at the atomic level.

B Implementation of Interaction Type Overlap (ITO)

The **Interaction Type Overlap (ITO)** metric measures the global agreement in interaction type distributions between the predicted and reference complexes, regardless of specific atom-level correspondences [2]. It focuses on comparing the overall frequency of each interaction type, rather than their precise locations.

To compute ITO, we count the number of occurrences of each interaction type in both the reference and predicted structures. For each interaction type, the overlap is defined as the minimum of the two counts. The ITO score is then calculated as the total overlap across all interaction types, normalized by the total number of interactions in the reference:

$$\text{ITO} = \frac{\sum_{\text{type}} \min(\text{count}_{\text{ref}}, \text{count}_{\text{pred}})}{\sum_{\text{type}} \text{count}_{\text{ref}}}$$

Similarly, if the reference structure contains no interactions, the ITO score is set to NaN and omitted from any aggregate analysis. This design ensures that only informative samples contribute to dataset-level evaluations. ITO captures the model's ability to reproduce the overall molecular interaction profile and is especially useful for assessing chemically meaningful binding preferences at a global scale.

C Definition of Diversity Metrics

Diversity quantifies the diversity of generated peptides as the ratio of unique clusters to total generations. Sequence and structure clustering thresholds are set at sequence identity above 40% and RMSD below 2 Å, respectively. For instance, a sequence diversity of 0.0593 for HCDR3 means that 100 generated sequences form approximately 6 distinct clusters. The diversity metrics for our model are presented in Table 6.

Table 6: Diversity of generated samples

Settings	Sequence Diversity
MEAN (HCDR3)	0.0100
RADiAnce (HCDR3)	0.0593
Pepflow	0.0745
RADiAnce (Peptide)	0.558

The results demonstrate that diffusion based methods achieve a higher degree of sequence diversity compared to non diffusion based methods.

D Ablations on Condition Integration Strategy

D.1 AdaLN-Zero Conditioning Mechanism

We implement the conditional noise prediction network $\epsilon_\theta(\mathcal{Z}_x^t, \mathcal{Z}_y, \mathbb{T}^v, t)$ by extending the Equivariant Full-Atom Transformer architecture [25] with AdaLN-Zero conditioning [44]. Each layer simultaneously updates the E(3)-invariant scalar latent features $\mathcal{Z}_t$ and vector latent features $\mathbf{v}_t$ through attention and feed-forward stages, all equipped with adaptive layer normalization and residual scaling.

Let $\mathbf{v}_i \in \mathbb{R}^{3 \times m \times h}$ denote the vector features corresponding to $\mathcal{Z}_i$, and let $\mathbf{T} \in \mathbb{R}^{n \times h}$ be the conditioning input. The update rules at layer l are:

$$\mathcal{Z}_t^{(l-0.5)}, \mathbf{v}_t^{(l)} = \mathcal{Z}_t^{(l-1)} + \alpha_{\text{attn}} \cdot \text{SelfAttn}(\text{AdaLN}(\mathcal{Z}_t^{(l-1)}, \mathbf{T}), \mathbf{v}_t^{(l-1)}), \quad (13)$$

$$\mathcal{Z}_t^{(l)} = \mathcal{Z}_t^{(l-0.5)} + \alpha_{\text{ffn}} \cdot \text{FFN}(\text{AdaLN}(\mathcal{Z}_t^{(l-0.5)}, \mathbf{T}), \mathbf{v}_t^{(l-1)}), \quad (14)$$

where $\text{AdaLN}(\cdot, \mathbf{T})$ applies zero-initialized, conditioning-aware layer normalization to the input, and $\alpha_{\text{attn}}, \alpha_{\text{ffn}} \in \mathbb{R}^{1 \times h}$ are scaling factors computed as:

$$\boldsymbol{\alpha} = f_{\text{scale}} \left(\frac{1}{n} \sum_{j=1}^n \mathbf{T}^{(j)} \right). \quad (15)$$

Here, $f_{\text{scale}} : \mathbb{R}^h \rightarrow \mathbb{R}^h$ is a learnable linear transformation initialized to zero, responsible for producing the residual scaling weights.

This conditioning scheme allows each layer to modulate feature updates based on template $\mathbf{T}$, while ensuring stable training via identity initialization and structured residuals.

D.2 Contextual Prompt Integration

To further enhance conditional control, we implement a context-aware integration layer within each transformer block. This design leverages prompt features $\mathbf{T}$ (e.g., interface representations retrieved from a database) to modulate the latent trajectory. The integration process involves attending to the prompt features to select the most relevant context and then fusing it with the latent representation. The update of scalar and vector features follows three stages:

Let $\mathcal{Z}_t^{(l-1)} \in \mathbb{R}^{m \times h}$ and $\mathbf{T} \in \mathbb{R}^{n \times h}$ denote the input features and prompt templates, respectively. At transformer layer l , the update rules are:

$$\mathbf{T}^{\text{sel}} = \text{Softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{h}} \right) \mathbf{T}, \quad \mathbf{Q} = \mathcal{Z}_t^{(l-1)} \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{T} \mathbf{W}_K, \quad (16)$$

$$\mathcal{Z}_t^{(l-0.5)} = \text{Fuse} \left(\mathcal{Z}_t^{(l-1)}, \mathbf{T}^{\text{sel}} \right), \quad (17)$$

$$\left[\mathcal{Z}_t^{(l)}, \mathbf{v}_t^{(l)} \right] = \text{FFN} \left(\text{SelfAttn} \left(\mathcal{Z}_t^{(l-0.5)}, \mathbf{v}_t^{(l-1)} \right) \right), \quad (18)$$

where $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{h \times h}$ are learnable projections. The attention mechanism computes similarity scores between latent tokens and prompt features to produce a soft selection $\mathbf{T}^{\text{sel}}$, enabling each latent position to selectively attend to the most informative prompts. The fusion function $\text{Fuse}(\cdot, \cdot)$ denotes a multilayer perceptron operating on the concatenation $[\mathcal{Z}_t^{(l-1)} || \mathbf{T}^{\text{sel}}]$, projecting it back to the original dimension h .

This mechanism enables fine-grained control over the generation process by allowing each latent token to dynamically attend to prompt embeddings and modulate its representation accordingly.

Table 7: Results for antibody CDR-H3 sequence-structure codesign.

Fusion method	AAR (%)	RMSD (Å)	$\Delta\Delta G$ (kJ/mol)	IMP (%)	ISM (%)
AdaLN-Zero	50.42	0.964	-7.047	68.33	73.69
In-context	<u>51.03</u>	1.019	-8.102	66.67	<u>72.57</u>
Cross Attention	54.66	0.9443	-6.236	71.67	71.64

D.3 Results

As shown in Table 7, the Cross Attention fusion mechanism demonstrates the most effective performance among all compared methods. It consistently achieves the best overall results across sequence accuracy, structural alignment, and interaction quality. This suggests that cross-attention enables more precise integration of contextual and structural information, leading to improved sequence-structure codesign for antibody CDRs. Its superior performance highlights the importance of flexible and fine-grained feature fusion in guiding generative antibody modeling.

E Implementation Details

E.1 Baselines

Peptide For **RFdiffusion**, since the training pipeline for custom datasets is not publicly available, we directly use the official pretrained weights and inference scripts for binder design. For **PepGLAD** and **PepFlow**, we adopt the official implementations and retrain them on our peptide dataset, using the default hyperparameters provided in their repositories. For **UniMoMo**, we retrain the model on the same cross-domain dataset as our method, using the hyperparameter settings provided in the original publication [31]. All models are evaluated on the same dataset to ensure fair comparison.

Antibody For **MEAN**, **DyMEAN**, **DiffAb**, **GeoAB-R**, and **GeoAB-D**, we use their official implementations and retrain the models on the same antibody dataset as our model, using the default hyperparameters specified in their repositories. For **UniMoMo**, we retrain the model on the same cross-domain dataset as our method, using the hyperparameter settings provided in the original publication [31]. To ensure consistency, we convert all antibody sequences from IMGT [33] to Chothia numbering [12], following the protocol of **DiffAb** [40]. The Chothia system provides stricter and more structure-consistent definitions of complementarity-determining regions (CDRs). This conversion may lead to a drop in amino acid recovery (AAR) compared to IMGT, as the Chothia system avoids overestimating recovery due to trivial unigram patterns [31]. For **GeoAB**, since the official implementation only supports HCDR3 prediction, we extend it to cover additional CDRs following the same processing strategy. Training samples that fail preprocessing are excluded. All models are evaluated on the same dataset to ensure fair comparison.

E.2 RADiAnce

We trained our model using 8 GPUs with 80 GB memory in parallel. The hyperparameter configurations for both Contrastive VAE and Diffusion models are summarized in Table 8.

F Comparison Between Generated Samples and Retrieved Exemplars

To quantitatively assess how well the generated interfaces recapitulate the interaction patterns of their retrieval-based conditioning, we report two key metrics: Interaction Type Overlap with the Reference (ITO-Gen) and Interaction Type Overlap with the Retrieved Exemplar (ITO-RAG), evaluated for both antibody HCDR3 and peptide cases (Figure 5).

Specifically, ITO-Gen measures the degree of overlap in interaction patterns between the generated samples and their corresponding reference complexes, reflecting the fidelity of binding mode reconstruction. In contrast, ITO-RAG quantifies the similarity between the generated samples and the retrieved exemplars, indicating the effectiveness of retrieval in guiding the generative process.

Table 8: Hyperparameters of **RADiAnce**

Name	Configuration	Description
<i>Contrastive VAE</i>		
Encoder / Decoder Type	EPT	Backbone architecture for encoder/decoder
latent_size	8	Dimension of latent state
hidden_size	512	Feature dimensionality for node and edge embeddings
edge_size	64	Size of edge type embeddings
n_layers	6	Transformer depth in encoder/decoder
n_heads	8	Number of heads for multihead self attention per layer
k_neighbors	9	Graph connectivity for spatial features
cutoff	10.0Å	Distance threshold for RBF kernels
KL_loss_weights	0.6 / 0.8	Weight for KL divergence of structure / sequence latents
atom_coord_loss_weights	1.0	Weight for atom coordinate loss
block_type_loss_weights	1.0	Weight for categorical loss on block (residue) types
contrastive_loss_weights	1.0	Weight for contrastive similarity loss
local_distance_loss_weights	0.5	Weight for intra-structure distance loss
bond_loss_weights	0.5	Weight for bond type classification loss
<i>Conditional Latent Diffusion</i>		
hidden_size	512	Dimension of hidden states
T	100	Diffusion steps
n_layers	6	Number of denoising layers
n_heads	8	Number of heads for multihead self and cross attention
n_rbf	64	Number of RBF kernels
cutoff	3.0Å	Cutoff distance for RBF kernels

The overlap with reference complexes reflects the model’s ability to reconstruct native binding modes, while the overlap with retrieved exemplars assesses how effectively information from retrieval guides generation. Our results indicate that the generated molecular interfaces capture a substantial proportion of the chemically meaningful interactions present in the reference structures—achieving, for example, 60.7% overlap for HCDR3 and 65.0% for peptide cases. At the same time, the retrieved exemplars themselves provide strong and relevant interaction patterns that align closely with the reference, offering reliable and informative guidance for the generative process.

These results collectively demonstrate that our retrieval-augmented framework, by leveraging structurally relevant examples, enables the effective transfer and reconstruction of key interaction motifs, thus enhancing the chemical rationality and interpretability of the designed interfaces.

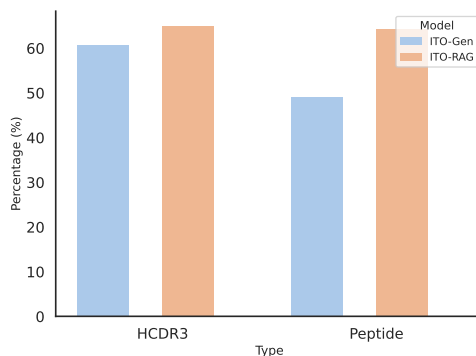


Figure 5: Interaction overlap analysis.

G Case Study: Qualitative Improvement Through Retrieval Conditioning

GPIIb/IIIa (PDB ID: 3NID) is a key platelet integrin that mediates adhesion and aggregation during thrombus formation [56]. In the 3NID structure, it adopts a closed headpiece conformation when bound to a specific antagonist that stabilizes the inactive state and prevents activation. As illustrated in Figure 6, we investigate the design of the HCDR3 loop for the GPIIb/IIIa binder under two settings: unconditional generation and retrieval-augmented generation. Without retrieval guidance, the model struggles to reconstruct the characteristic multi-hydrogen bond interactions observed in the reference structure. In contrast, retrieval from the database yields top-10 candidates that include two peptide complexes and one antibody-antigen complex, all of which exhibit similar interaction patterns to the target. Specifically, these retrieved structures share the distinctive formation of hydrogen bonds

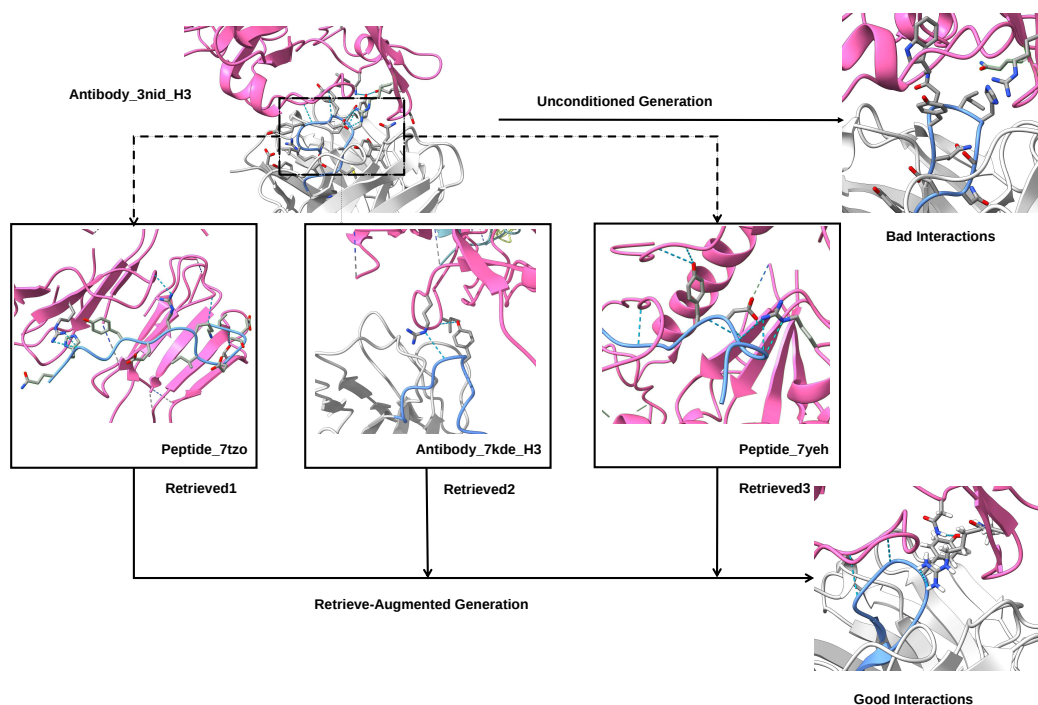


Figure 6: Case study of HCDR3 design for the GPIIb/IIIa(PDBID:3NID) binder. The binding site is shown in pink, binder key residues in blue, and hydrogen bonds in light blue dashed lines. Guided by retrieved exemplars, retrieval-augmented generation successfully preserves these key interaction modes.

between an arginine residue at the binding site and the binder, as well as between a tyrosine residue in the binder and the binding site. Consequently, the RAG-based generation inherits these critical interaction motifs, successfully recovering the hydrogen-bonding patterns mediated by arginine and tyrosine residues. This case exemplifies how our retrieval-augmented generation framework effectively leverages retrieved interaction patterns to guide and enhance molecular design.

H Cross-Domain Retrieval Ablation Studies

To further elucidate the impact of cross-domain retrieval on our model's generative performance, we conducted a series of ablation studies by selectively restricting the set of available templates during inference. Specifically, we compared the outcomes of conditioning the generation process using retrieval exemplars sourced from either a single domain (e.g., antibody-only, peptide-only, etc) versus those retrieved from the entire, cross-domain template pool.

The results in Table 9 and Table 10 reveal a consistent trend: models leveraging cross-domain retrieval demonstrate superior capability in reconstructing chemically and geometrically plausible interface interactions, as measured by both interaction recovery metrics and binding affinity scores. In contrast, restricting the retrieval to templates within the same domain as the query generally leads to diminished performance, manifesting as reduced interaction overlap and less robust generalization to diverse binding interfaces.

These results highlight the critical advantage of our approach: by comprehensively referencing interface templates across distinct molecular domains, our model can flexibly exploit shared principles of molecular recognition and binding site architecture, thereby enabling more rational and reliable binder generation. The empirical gains observed across all tested scenarios underscore the necessity of cross-domain template integration as a core design choice for retrieval-augmented generative frameworks.

Table 9: Results of recovery metrics for RADiAance under different template sources on antibody cdrs codesign task. Columns “Ab”, “Pep”, and “ProtFrag” indicate whether the respective template domain is included for retrieval during inference (✓ for included, ✗ for excluded).

CDR	Ab	Pep	ProtFrag	AAR (%)	RMSD (Å)	$\Delta\Delta G$ (kJ/mol)	IMP (%)	ISM (%)
H1	✓	✗	✗	90.54	0.295	-7.955	98.33	98.36
	✓	✓	✗	90.35	0.3007	-7.705	95.00	98.36
	✓	✓	✓	90.83	<u>0.2977</u>	-8.221	<u>96.67</u>	98.36
H2	✓	✗	✗	78.64	<u>0.2157</u>	-2.605	88.33	<u>87.83</u>
	✓	✓	✗	78.70	0.2158	-2.635	80.0	88.10
	✓	✓	✓	79.20	0.2135	-2.370	<u>83.33</u>	<u>87.83</u>
H3	✓	✗	✗	54.47	<u>0.9475</u>	-5.200	65.0	<u>71.96</u>
	✓	✓	✗	54.16	0.9557	-4.545	<u>66.7</u>	72.81
	✓	✓	✓	54.66	0.9443	-6.236	71.67	71.64
L1	✓	✗	✗	86.48	<u>0.2902</u>	-7.59	96.67	97.88
	✓	✓	✗	86.48	0.2887	-7.50	<u>95.00</u>	97.64
	✓	✓	✓	<u>86.20</u>	0.2907	-7.170	<u>95.00</u>	98.31
L2	✓	✗	✗	87.13	0.1550	-4.435	96.67	98.52
	✓	✓	✗	<u>86.99</u>	<u>0.1553</u>	-4.647	<u>95.00</u>	98.52
	✓	✓	✓	86.68	0.1558	-3.857	<u>95.00</u>	98.52
L3	✓	✗	✗	<u>76.71</u>	<u>0.4398</u>	-5.94	86.67	95.08
	✓	✓	✗	76.41	0.4372	-7.43	93.33	95.08
	✓	✓	✓	76.75	0.4478	-9.207	<u>90.00</u>	<u>94.82</u>

Table 10: Results of recovery metrics for RADiAance under different template sources on peptide codesign task. Columns “Ab”, “Pep”, and “ProtFrag” indicate whether the respective template domain is included for retrieval during inference (✓ for included, ✗ for excluded).

Ab	Pep	ProtFrag	AAR (%)	RMSD (Å)	$\Delta\Delta G$ (kJ/mol)	IMP (%)	ISM (%)
✗	✓	✗	38.97	2.27	3.424	<u>40.86</u>	<u>49.82</u>
✓	✓	✗	39.60	2.41	<u>2.423</u>	38.71	48.33
✓	✓	✓	<u>39.42</u>	<u>2.29</u>	1.963	41.94	52.15

I Ablation on Single-Domain Data

To ensure that the performance comparison is not biased by the use of multiple binder types, we conducted an ablation study where both the training and retrieval data were restricted to a single domain. Specifically, we evaluated **RADiAance** and UniMoMo under the antibody heavy chain CDR3 (AbH3) and peptide design tasks, each trained and retrieved using data from only that domain. This setup removes cross-type information and ensures a strictly matched comparison between the models.

Table 11: Ablation study on single-domain data. Both training and retrieval were performed within the same domain to ensure a fair comparison. Other single-domain results are omitted as they are the same result with those reported in the main text.

Task and models	AAR (%)	RMSD (Å)	$\Delta\Delta G$ (kJ/mol)	IMP (%)	ISM (%)
UniMoMo (AbH3)	48.78	1.39	-5.781	63.33	65.46
RADiAance (AbH3)	51.31	1.109	-5.994	68.33	69.71
UniMoMo (Peptide)	37.59	2.48	7.69	29.03	40.08
RADiAance (Peptide)	38.28	2.37	7.57	27.95	45.37

As shown in Table 11, even when trained and retrieved within a single domain, **RADiAance** consistently outperforms other models across evaluated metrics. This demonstrates that the performance gain of **RADiAance** is not solely due to multi-type data retrieval, but rather arises from its intrinsic ability to model contextually aligned binder–target interactions.

J Discussion

J.1 Reproducibility and Statistical Robustness

To demonstrate the robustness and reliability of our evaluation, we explicitly quantify the statistical variability of our model performance under different random initializations. Our main results are accompanied by standard deviations that measure model variability across independent runs with different random seeds. Specifically, we reran the entire inference and evaluation pipeline three times with distinct random seeds, and computed the mean and standard deviation across these runs. Given that each evaluation aggregates thousands of test instances, the per-run variation remains small, indicating robustness and reproducibility of our evaluation benchmark.

Table 12: Performance metrics with standard deviations across independent runs. The consistently small deviations confirm the stability of our evaluation.

Model	AAR (%)	RMSD (Å)	$\Delta\Delta G$ (kJ/mol)	IMP (%)	ISM (%)
HCDR3	54.66±0.0026	0.9443±0.0156	-6.236 ± 0.5862	71.67±0.0096	71.64±0.0085
Peptide	39.42±0.0014	2.29±0.0043	1.963 ± 0.2302	41.94±0.0108	52.15±0.0007

As shown in Table 12, the standard deviations across independent runs are consistently small across all metrics. This confirms that the evaluation is highly stable, with negligible sensitivity to random initialization or stochastic effects during inference.

J.2 Extended Evaluation Metrics

To provide a more comprehensive evaluation of our model, we further introduced several robust and widely used metrics, including DockQ, FoldX ΔG , and Binding Site Recovery. These metrics complement the previously reported ones by assessing structural and energetic aspects of the predicted complexes.

Table 13: Extended evaluation metrics for RADiAance and UniMoMo on antibody and peptide datasets.

Model	DockQ	FoldX ΔG (kJ/mol)	Binding Site Recovery
RADiAance (AbH3)	0.9582	-8.4250	0.9964
UniMoMo (AbH3)	0.9491	-7.9950	0.9962
RADiAance (Peptide)	0.7698	-7.9891	0.9857
UniMoMo (Peptide)	0.7592	-7.4901	0.9822

DockQ DockQ is a continuous quality score for protein–protein docking models ranging from 0 to 1. It integrates the fraction of native contacts (F_{nat}), ligand RMSD (LRMSD), and interface RMSD (iRMSD) to provide an overall measure of structural agreement with the native complex. Scores above 0.23, 0.49, and 0.80 correspond to acceptable, medium, and high-quality models, respectively [7].

FoldX ΔG FoldX calculates the binding free energy (ΔG) of a protein–protein complex as the difference between the Gibbs free energy of the complex and the sum of its unbound partners. More negative ΔG values indicate stronger and more stable interactions [9].

Binding Site Recovery Binding Site Recovery measures the proportion of true interface residues correctly identified by the model, representing the accuracy of interface residue prediction. Higher values indicate that the generated complex preserves the biologically relevant interaction sites.

As shown in Table 13, the newly introduced metrics are consistent with the previously reported performance trends. **RADiAance** outperforms UniMoMo across both antibody and peptide benchmarks, achieving higher DockQ scores and more favorable FoldX ΔG . These results confirm the robustness and reliability of our model across diverse evaluation metrics.

J.3 Dataset Analysis and Benchmark Reliability

Data leakage in retrieval-augmented generative frameworks can lead to overly optimistic results if test samples share high similarity with retrieved templates. To address this concern, we performed a comprehensive verification of our data split.

Established conventions and potential issues Following prior studies in antibody and peptide design, we adopt the RABD and LNR benchmarks, respectively, and enforce a 40% sequence-identity clustering to prevent data leakage across train, validation, and test partitions. However, as pointed out in prior discussions, sequence clustering alone may be insufficient, since structural similarity (e.g., TM-score) can still introduce hidden overlaps across sets.

Verification of data split integrity To rigorously verify the split, we computed both sequence identity and TM-score between each test sample and the entire training set (Table 14). The overall low similarity confirms that the test data are largely distinct from the training examples. Nonetheless, a few rare cases with **TM-score** > **0.50** were observed, indicating that minor structural overlaps may still occur under conventional data-splitting method.

Table 14: Verification of data-split integrity. All values represent mean $\pm$ standard deviation.

Test Set	TM-score with Train Set	SeqId with Train Set
RABD	0.2421 $\pm$ 0.053	0.1272 $\pm$ 0.021
LNR	0.2503 $\pm$ 0.069	0.0983 $\pm$ 0.017

Re-evaluation after removing overlapping cases To completely remove any potential confounding factor, we excluded any test sample that had at least one training neighbor with a TM-score > 0.50. On the LNR dataset, this excluded 43 samples, leaving 50 for re-evaluation. Table 15 summarizes the resulting performance across multiple baselines. All models experienced performance degradation, which may result from both inadvertent data overlap in previous works and the inherently more challenging or biased characteristics of the remaining test cases. Importantly, our model (**RADiAnce**) not only maintained the strongest overall performance after this correction but also exhibited the smallest performance drop among all methods, further demonstrating that its advantage does not stem from data leakage but from genuinely robust retrieval guided generation.

Table 15: Baseline performance drop after removing overlapping test cases (TM-score > 0.50). Numbers in parentheses show relative change; lower RMSD/ $\Delta\Delta G$ and higher AAR/ISM are better.

Model	AAR (%)	RMSD ($\text{\AA}$)	$\Delta\Delta G$ (kJ/mol)	ISM (%)
PepFlow	31.41 (-11.4%)	3.63 (+26.6%)	18.84 (+19.9%)	22.22 (-21.7%)
PepGLAD	33.85 (-12.4%)	3.43 (+25.2%)	18.66 (+22.3%)	27.67 (-15.2%)
UniMoMo	35.00 (-9.36%)	2.87 (+24.2%)	3.52 (+46.1%)	38.69 (-21.2%)
RADiAnce	35.79 (-9.20%)	2.80 (+22.3%)	2.21 (+12.7%)	42.26 (-19.0%)

Implications for future benchmarks This analysis reveals that even widely used benchmarks may inadvertently allow structural redundancy across data splits. Our findings suggest that future dataset construction should incorporate stricter filtering, combining both sequence and structure based similarity constraints (e.g., TM-score thresholds) to ensure fair evaluation. We encourage the community to establish standardized, publicly verifiable data splitting protocols to enhance reproducibility and avoid unintended information leakage in further research.

K Code Availability

The source code of the **RADiAnce** is available at <https://github.com/srhn225/RADiAnce>.