

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}^{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.

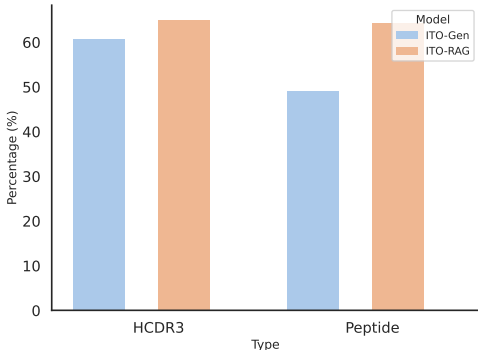


Figure 5: Interaction overlap analysis.

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

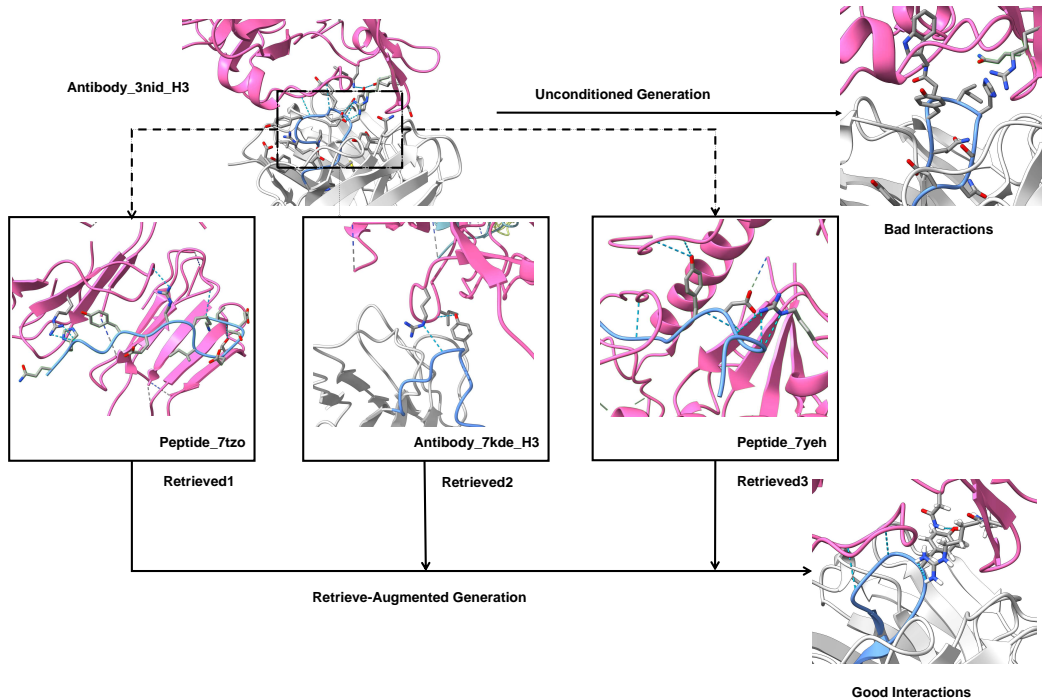


Figure 6: Case study of HCDR3 design for the GPIIb/IIIa(PDBID:3NID) binder. The binding site is shown in pink, binder key residues in blue, and hydrogen bonds in light blue dashed lines. Guided by retrieved exemplars, retrieval-augmented generation successfully preserves these key interaction modes.

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 RADiAnce 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 RADiAnce 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 **RADiAnce** 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
RADiAnce (AbH3)	51.31	1.109	-5.994	68.33	69.71
UniMoMo (Peptide)	37.59	2.48	7.69	29.03	40.08
RADiAnce (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, **RADiAnce** consistently outperforms other models across evaluated metrics. This demonstrates that the performance gain of **RADiAnce** 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 \pm 0.0026	0.9443 \pm 0.0156	-6.236 \pm 0.5862	71.67 \pm 0.0096	71.64 \pm 0.0085
Peptide	39.42 \pm 0.0014	2.29 \pm 0.0043	1.963 \pm 0.2302	41.94 \pm 0.0108	52.15 \pm 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 RADiAnce and UniMoMo on antibody and peptide datasets.

Model	DockQ	FoldX ΔG (kJ/mol)	Binding Site Recovery
RADiAnce (AbH3)	0.9582	-8.4250	0.9964
UniMoMo (AbH3)	0.9491	-7.9950	0.9962
RADiAnce (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. **RADiAnce** 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 (\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>.