RNACLIP: AUXILIARY PROTEIN–LIGAND DATA FOR POCKET-FREE RNA–LIGAND VIRTUAL SCREENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Small molecules targeting RNA offer therapeutic opportunities beyond conventional protein targets, yet current screening methods are constrained by the difficulty of locating binding pockets and the scarcity of validated RNA–ligand interactions. Here, we introduce RNAClip, a contrastive retrieval framework that brings abundant protein–ligand data into RNA virtual screening. By aligning receptors and ligands in a shared embedding space, the framework enables sequence-based screening without a predefined RNA pocket or binding pose. Auxiliary protein examples contribute both paired supervision and contrastive negatives, allowing RNA screening to draw on data beyond the RNA domain. We also construct a new evaluation set from recently released RNA–ligand structures in the Protein Data Bank(PDB). On this strictly independent test set, RNAClip outperforms all other evaluated models, achieving state-of-the-art(SOTA) screening performance. A series of ablation experiments further demonstrates that a small subset of auxiliary protein–ligand data can match the performance achieved with the full corpus, and that using auxiliary examples solely as contrastive negatives can preserve the gains from joint training.

1 INTRODUCTION

Targeting disease-associated RNAs with small molecules expands the therapeutic scope of drug discovery beyond conventional protein targets. Potential targets include transcripts encoding proteins that are difficult to drug, as well as noncoding RNAs implicated in disease (Ma et al., 2025; Childs-Disney et al., 2022). By modulating RNA splicing, translation, and processing, small molecules can regulate disease-relevant gene expression. The clinical use of the splicing modulator risdiplam for spinal muscular atrophy illustrates the therapeutic potential of this strategy (Fei et al., 2026). Identifying suitable RNA binders within the vast space of candidate compounds is therefore a central task in RNA-targeted drug discovery. Virtual screening can prioritize candidates for experimental testing, but accurate ligand ranking remains challenging given the limited availability of RNA structures and experimentally validated interaction data.

Two challenges limit current RNA virtual screening. First, identifying a suitable binding pocket is difficult before a ligand has been characterized. RNA adopts flexible conformations, and its binding surfaces can change upon ligand recognition; high-quality experimental structures are also scarce (Jiang et al., 2025). Structure-based methods therefore face an input bottleneck: docking requires a receptor structure and a search region, while direct scoring methods such as RNAmigos2(Carvajal-Patiño et al., 2025) rely on a predefined RNA binding site. Uncertainty in receptor structures and pocket localization can consequently propagate into ligand ranking. These limitations motivate *pocket-free, pose-free* screening, in which an RNA sequence and ligand chemical specifications suffice to rank candidates without a predefined binding pocket or a supplied receptor–ligand binding pose.

Second, eliminating the need for a known pocket does not resolve the scarcity of reliable training examples. Experimentally validated RNA–ligand interactions span a limited range of targets and compounds, restricting model generalization to unfamiliar RNA–ligand pairs (Carvajal-Patiño et al., 2025). Recent methods address this challenge through data perturbation and augmentation in RNAsmol (Ma et al., 2025), pretrained representations in DeepRNA-DTI (Bae & Nam, 2025), and RNA language modeling combined with contrastive learning and decoy enhancement in SMART-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Bind (Jiang et al., 2025). Although these approaches improve the use of available RNA data, the underlying supply of experimentally observed binding pairs remains limited. For example, there are only 2,366 RNA–ligand interfaces after filtering PDB structures released before September 30, 2021. This scarcity motivates the use of experimentally grounded supervision beyond the RNA domain.

Protein–ligand data offer a substantially larger source of experimentally grounded auxiliary examples. At the same time, contrastive learning provides a natural way to incorporate these data into RNA screening. As illustrated by DrugCLIP (Gao et al., 2023), virtual screening can be formulated as retrieval in a shared receptor–ligand embedding space, allowing models to learn from observed binding pairs without requiring measured affinities for every example. Within this formulation, an auxiliary protein–ligand pair contributes through two channels: its matched-pair loss provides *paired supervision*, while its receptor and ligand serve as competing candidates in the *contrastive objective*. The latter channel is particularly relevant to RNA screening, where selective target engagement must be considered alongside unintended interactions with proteins (Childs-Disney et al., 2022). Protein-associated ligands therefore provide a potentially useful source of competing candidates against which known RNA binders can be distinguished. Their role as negatives is defined by the training objective, however, and does not imply that protein-binding ligands cannot also bind RNA. Together, these two channels provide a rationale for exploring how protein–ligand data can support RNA screening.

To this end, we introduce RNAClip, a contrastive framework for pocket-free, pose-free RNA–ligand virtual screening that incorporates auxiliary protein–ligand data. We use frozen IntelliFold encoders (Qiao et al., 2025) to represent receptors and ligands separately, with RNA and protein sharing the receptor tower. Lightweight trainable heads project these representations into a common embedding space, where a symmetric contrastive objective learns from mixed RNA–ligand and protein–ligand examples. This design incorporates the larger protein corpus while preserving RNA binding pairs as direct supervision for the target task. At inference, candidate ligands are ranked by their cosine similarity to the RNA query, and ligand embeddings can be precomputed and reused across targets. The framework also enables us to investigate how much auxiliary data is needed and whether their benefits depend on the original protein–ligand pairings.

To assess generalization to recent experimental targets, we construct a recent PDB evaluation set from newly released RNA–ligand structures using quality filtering, structural redundancy filtering against historical PDB complexes, and property-matched decoy selection. The resulting benchmark contains ten RNA targets and 980 ligand candidates. On this benchmark, RNAClip outperforms all evaluated baselines, including DeepMIF, pocket-aware AutoDock Vina, and confidence-based AF3 scoring. Beyond overall ranking performance, improvements in early recognition further support the utility of auxiliary protein–ligand data for prioritizing RNA binders.

To understand how auxiliary protein data contribute to these gains, we conduct a series of ablation experiments examining both their quantity and their role in training. Experiments demonstrate that a small random auxiliary subset matches the observed screening performance obtained with the full corpus. Moreover, retaining auxiliary examples solely as contrastive negatives preserves the observed gains from joint training, with partner-shuffling experiments providing complementary evidence. These findings indicate that auxiliary protein–ligand data can benefit RNA screening even without intact protein–ligand paired supervision.

Our main contributions are summarized as follows:

- To the best of our knowledge, this is the first effort to incorporate auxiliary protein–ligand pairs into joint contrastive training for pocket-free RNA–ligand virtual screening, helping alleviate the scarcity of RNA–ligand training data.
- We develop RNAClip, a contrastive retrieval framework built on a pretrained biomolecular structure model. We demonstrate its screening effectiveness on a newly curated benchmark separated from the training data through temporal splitting and structural novelty filtering.
- Through ablation studies, we demonstrate the importance of auxiliary protein–ligand data and the pretrained structural model. We further find that the benefits of auxiliary data can be retained with a small subset or by using auxiliary examples solely as contrastive negatives, indicating that these gains do not require intact auxiliary paired supervision.

2 RELATED WORK

2.1 PROTEIN VIRTUAL SCREENING

Protein virtual screening ranks compounds using docking, learned interaction scores, or representation-based retrieval. AutoDock Vina searches ligand poses with an empirical scoring function (Trott & Olson, 2010), while GNINA integrates convolutional networks into pose scoring (McNutt et al., 2021). Large-scale docking has identified experimentally validated ligands, supporting the value of exploring extensive chemical libraries (Lyu et al., 2019). These approaches require receptor structures and candidate pose generation. Sequence-based affinity prediction offers an alternative: DeepDTA learns from protein sequences and molecular SMILES using quantitative binding labels (Öztürk et al., 2018). DrugCLIP instead learns separate protein-pocket and molecular encoders through contrastive training on binding pairs (Gao et al., 2023). Its shared embedding space supports cached compound representations and similarity-based retrieval without docking every candidate or requiring numerical affinity labels. Subsequent work extends this strategy to genome-wide screening with predicted protein structures (Jia et al., 2026). Thus, sequence-based affinity prediction avoids receptor-structure input, whereas pocket-based retrieval reduces screening-time computation while retaining a structural query.

2.2 RNA VIRTUAL SCREENING

RNA screening faces conformational flexibility, distinctive binding energetics, and limited experimental coverage (Sun et al., 2020; Carvajal-Patiño et al., 2025). HARIBOSS curates RNA–small-molecule complexes for structural analysis (Panei et al., 2022). Structure-based methods include rDock, which supports proteins and nucleic acids, and RLDOCK, which introduces RNA-specific scoring and searches alternative binding sites (Ruiz-Carmona et al., 2014; Sun et al., 2020). RLaffinity uses contrastive pretraining of voxelized RNA–ligand complexes followed by affinity regression (Sun & Gao, 2024). Graph-based methods reduce reliance on candidate pose generation: RNAmigos predicts ligand fingerprints from RNA binding-site graphs, while RNAmigos2 combines compatibility prediction from observed complexes with docking-derived supervision (Oliver et al., 2020; Carvajal-Patiño et al., 2025). Both retain a structural binding-site query, and RLaffinity requires a joint complex representation rather than independently encoded retrieval candidates.

Methods without RNA tertiary-structure input learn from sequence, secondary structure, and molecular features. DeepRSMA and DeepMIF use multiview fusion for affinity prediction (Huang et al., 2024; Song et al., 2026); RNAsmol studies sequence augmentation for interaction prediction, and DeepRNA-DTI jointly predicts interactions and nucleotide-level binding sites (Ma et al., 2025; Bae & Nam, 2025). SMRTnet combines RNA sequence and secondary structure with molecular representations and reports experimentally validated binders (Fei et al., 2026). These approaches differ in structural requirements and supervision, so avoiding tertiary structures does not imply identical screening settings. SMARTBind provides a sequence-based contrastive retrieval approach, mapping RNA language-model features and ligand fingerprints into a shared space (Jiang et al., 2025). Its ligand-specific decoy enhancement supplements known binding pairs with chemically selected alternatives, connecting RNA screening to the broader problem of negative-sample construction.

3 METHOD

3.1 OVERVIEW

Let r denote the sequence of a target RNA and $C = \{\ell_1, \ldots, \ell_N\}$ a library of candidate ligands. The goal of virtual screening is to identify the top k candidates most likely to bind the target. We assign each RNA–ligand pair a score $s(r, \ell_j)$ and rank the library in descending order of this score to select candidates for experimental testing. Our screening setting uses the RNA sequence and ligand chemical specifications, without a supplied binding pocket or experimental RNA–ligand pose.

We formulate screening as dense retrieval in a shared receptor–ligand embedding space, following contrastive protein–molecule screening (Gao et al., 2023). RNAClip learns this space jointly from the RNA–ligand collection $\mathcal{D}_{\mathrm{R}}$ and the auxiliary protein–ligand collection $\mathcal{D}_{\mathrm{P}}$. Figure 1 summarizes the two-tower training and screening workflow.

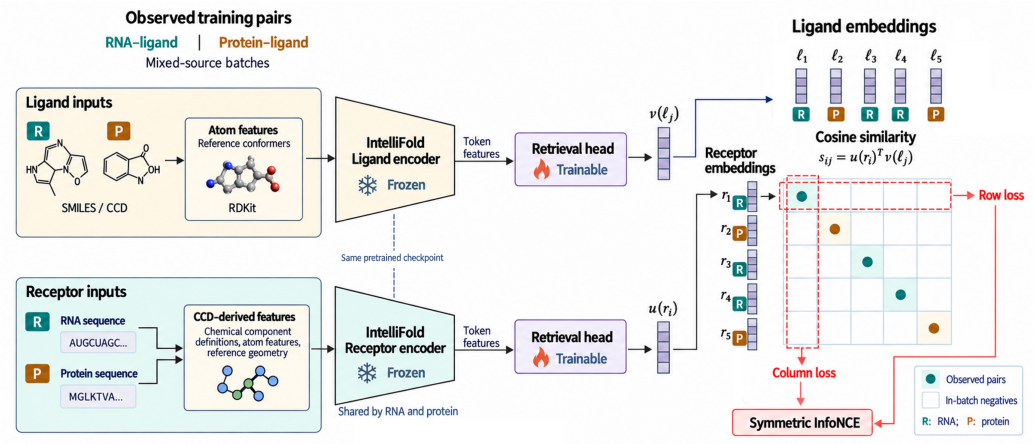


Figure 1: **Overview of RNAClip.** Frozen IntelliFold encoders and trainable heads embed receptors and ligands separately; RNA and protein share the receptor tower. The cosine-similarity matrix supplies observed pairs on the diagonal and in-batch negatives off the diagonal for a symmetric contrastive loss. R and P denote RNA and protein examples.

3.2 Receptor and Ligand Encoders

We adopt the representation trunk of IntelliFold(Qiao et al., 2025), which comprises an input embedding module, a multiple sequence alignment (MSA) processing module, a Pairformer stack, and a recycling mechanism. We initialize separate receptor and ligand encoders, E_r and E_ℓ, from the same pretrained checkpoint and freeze their parameters. RNA and protein share the receptor encoder.

First, both receptor and molecules are tokenized to atoms. Receptor sequences are mapped to residue definitions in the Chemical Component Dictionary (CCD), while ligands are specified by CCD entries or SMILES strings. From these chemical descriptions, the preprocessing pipeline derives atomic elements and generates reference conformers with RDKit. These constitute the atom-level inputs to the encoders. Each standard nucleotide or amino-acid residue forms one token, whereas each ligand heavy atom forms one token.

As described in AlphaFold3, the input embedding module applies local atom attention and aggregates atom features into token embeddings. These initialize a single representation $h_i \in \mathbb{R}^{H_s}$ for each token and a pair representation $p_{ij} \in \mathbb{R}^{H_p}$ for each token pair, incorporating relative-position and bond features into the latter. The MSA module updates the pair representations, and the Pairformer stack refines both representations through triangular pair updates and single-representation self-attention with pair biases. Recycling repeats this computation to produce the final single-representation matrix $F \in \mathbb{R}^{L \times H_s}$, where L is the number of tokens. We extract F without executing the diffusion structure-generation module. Detailed initialization and update equations are provided in Appendix B.1.

Finally, a trainable head maps F to a retrieval embedding $z \in \mathbb{R}^H$ through masked mean pooling, an affine projection, layer normalization, and ℓ_2 normalization:

$$\bar{h} = \frac{\sum_{t=1}^{L} a_t F_t}{\sum_{t=1}^{L} a_t}, \qquad g = \mathrm{LN}(W\bar{h} + b), \qquad z = \frac{g}{\|g\|_2}, \tag{1}$$

where $a_t \in \{0, 1\}$ masks padding, F_t is the representation of token t, $W \in \mathbb{R}^{H \times H_s}$ and $b \in \mathbb{R}^H$ are trainable parameters, and LN denotes layer normalization. In our implementation, $H_s = 384$, $H_p = 128$, and $H = 512$. We denote the receptor and ligand embeddings by $u(r), v(\ell) \in \mathbb{R}^H$, respectively. RNA and protein share the receptor head, while ligands use a separate head.

3.3 Contrastive Learning Objective

To learn a shared retrieval space, we first measure the similarity between a receptor and a ligand using the embeddings produced by the two towers. Since both embeddings are ℓ_2-normalized, their

inner product gives the cosine similarity:

$$s(r, \ell) = u(r)^\top v(\ell). \tag{2}$$

The training objective should assign an observed receptor–ligand pair a higher score than alternative pairings, so that its ligand can be retrieved from a candidate library.

Our structural data provide observed receptor–ligand pairs, but do not label the remaining possible pairings as binding or nonbinding. We therefore construct contrastive alternatives within each batch, following the in-batch strategy of CLIP (Radford et al., 2021). Given B paired examples $\{(r_i, \ell_i)\}_{i=1}^{B}$, we encode the receptors and ligands separately and compare every receptor with every ligand. This produces B^2 scores $s_{ij} = s(r_i, \ell_j)$: the B diagonal entries represent the observed pairs, and the $B(B-1)$ off-diagonal entries serve as negatives for the contrastive objective. These negatives are competing assignments, not experimentally established nonbinding pairs.

We optimize a symmetric InfoNCE objective (Oord et al., 2018; Radford et al., 2021). In the receptor-to-ligand direction, the model identifies the observed ligand among all B ligands for a given receptor. In the ligand-to-receptor direction, it identifies the corresponding receptor among all B receptors. The two losses for pair i are

$$\begin{aligned} \mathcal{L}_i^{r\to\ell} &= -\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ij}/\tau)}, \\ \mathcal{L}_i^{\ell\to r} &= -\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ji}/\tau)}, \end{aligned} \tag{3}$$

where the temperature τ controls the concentration of each softmax distribution and is fixed at 0.07. Averaging the two directions over the batch gives the joint-training loss. We introduce an anchor mask $m_i \in \{0, 1\}$ to also express the ablation that removes auxiliary supervision:

$$\mathcal{L}(m) = \frac{1}{2B} \sum_{i=1}^{B} m_i \big(\mathcal{L}_i^{r\to\ell} + \mathcal{L}_i^{\ell\to r}\big). \tag{4}$$

Joint training uses $m_i = 1$ for every RNA–ligand and protein–ligand pair.

An auxiliary protein–ligand example contributes to this objective through two channels. First, the ***paired-supervision channel*** learns the assigned protein–ligand correspondence through the example's two anchor losses. Second, the ***negative channel*** introduces its receptor and ligand embeddings as competing candidates for other anchors. For an RNA–ligand pair, the auxiliary ligand competes with the native ligand in the receptor-to-ligand loss, while the auxiliary protein receptor competes with the RNA receptor in the ligand-to-receptor loss. Adding protein examples therefore changes both the binding correspondences being learned and the candidates against which RNA pairs are distinguished, while keeping the local candidate-pool size fixed.

Our ablation study examines whether the benefits of auxiliary data depend on learning their original protein–ligand correspondences. To distinguish the contributions of the two channels, we introduce two interventions:

(i) **Negatives-only.** We set $m_i = 1$ for RNA pairs and $m_i = 0$ for protein pairs, removing both anchor-loss terms for each auxiliary pair while retaining all receptor and ligand embeddings in the denominators of the RNA losses. The normalization factor remains $1/(2B)$, preserving the coefficient of each retained RNA term relative to joint training.

(ii) **Pair shuffle.** We shuffle ligands among protein examples within each local batch, assigning each protein a ligand from another example. This preserves the RNA–ligand pairs and the set of molecules in the batch while altering the auxiliary protein–ligand correspondences. All pairs continue to contribute to the contrastive objective, with the shuffled protein–ligand pairs treated as positives.

The first intervention removes auxiliary paired supervision while preserving the negative channel; the second retains auxiliary paired supervision but alters the correspondences it teaches. Implementation details are provided in Appendix B.

3.4 Training and Inference

Constructing the training data. We use the RNA–ligand and protein–ligand subsets of the IntelliFold training set as our downstream training data (Qiao et al., 2025), which follows the AlphaFold 3 data-processing pipeline (Abramson et al., 2024). They clean PDB biological assemblies and use receptor–ligand interfaces as paired examples. Each receptor and ligand is processed separately, with their observed association providing the positive training pair. After filtering, the RNA collection $\mathcal{D}_{\mathrm{R}}$ contains 2,366 interfaces from 903 PDB entries, and the protein collection $\mathcal{D}_{\mathrm{P}}$ contains 431,972 interfaces from 87,531 entries. Preprocessing and interface-clustering details are given in Appendix A.1.

Sampling for contrastive training. We sample RNA–ligand and protein–ligand pairs at a $1:1$ ratio over each training epoch. The sampled pairs are shuffled into mixed batches for symmetric contrastive learning, with source proportions allowed to vary within each batch. We use the full protein corpus by default and hold source quotas and the training budget fixed when comparing auxiliary subsets. Sampling weights and optimization settings are specified in Appendix B.

Inference. We encode the ligand library once and reuse its embeddings across RNA queries. For a query embedding $u(r)$ and a matrix V whose rows are the candidate embeddings $v(\ell_j)$, the score vector is $Vu(r)$; candidates are ranked in descending order. Screening uses only the RNA sequence and ligand chemical specifications, without an experimental binding pocket or bound complex pose.

3.5 Construction of the recent PDB evaluation set

We construct the recent PDB evaluation set and a separate validation set in two steps: selecting RNA–ligand targets and assembling matched decoy pools.

Target selection. We collect RNA–ligand complexes released in the PDB after the September 30, 2021 structural training cutoff. Following FoldBench (Xu et al., 2025), we apply quality filtering, remove redundancy against historical structures using pocket-sequence identity and aligned ligand overlap, and cluster and deduplicate the retained interfaces. Cluster representatives undergo further screening and curation, leaving 17 targets after excluding ribosomal targets, invalid inputs, and redundant conformers. Appendix A details the reference set, structural comparison, and quality criteria.

Decoy selection. Following RNAsmol (Ma et al., 2025), we use its DecoyFinder script to search for 99 decoys for each target's observed ligand. Decoys are matched in molecular weight, hydrogen-bond donor and acceptor counts, logP, and rotatable bonds, while MACCS-fingerprint similarity is limited both to the observed ligand and between decoys. Selection thresholds and implementation details are given in Appendix A.2.2.

We reserve targets released in 2023 for validation and designate those released in 2024–2025 as the *recent PDB evaluation set* for testing. The validation set contains seven RNA targets and 657 ligand candidates, while the recent PDB evaluation set contains ten RNA targets and 980 candidates.

4 Experiments

4.1 Evaluation setup

Metrics. Our primary metric is mean per-target average precision (PT-AP), computed within each target's candidate pool and averaged with equal weight across the ten test targets. We also report hits at $k \in \{1, 2, 3, 5\}$, the number of targets whose native ligand ranks within the top k, expressed out of ten. To further assess early recognition, we report BEDROC with $\alpha = 20$ and 80.5 (Truchon & Bayly, 2007). BEDROC applies exponentially decreasing weight to later ranks; the larger α places greater emphasis on the earliest positions. We compute BEDROC using each target's candidate count and then average across targets. Higher values indicate better retrieval for all three metrics. Exact definitions and score-tie conventions are given in Appendix C.

Table 1: Results on the recent PDB evaluation set.

Task	Method	PT-AP ↑	Hits@1 ↑	BEDROC ↑	
				$\alpha = 80.5$	$\alpha = 20$
Docking score	AutoDock Vina	0.352	3/10	0.304	0.373
ipTM rank	AlphaFold 3	0.048	0/10	0.001	0.068
Affinity	BioLLMNet	0.054	0/10	0.000	0.060
	EMMPTNet	0.065	0/10	0.007	0.112
	DeepRSMA	0.167	1/10	0.102	0.197
	DeepMIF	0.374	3/10	0.317	0.419
Binary	RNAsmol	0.190	1/10	0.145	0.232
	SMARTBind	0.276	2/10	0.209	0.323
	DeepRNA-DTI	0.293	2/10	0.240	0.350
Ours	RNAClip (RNA only)	0.255	1/10	0.199	0.356
	RNAClip (RNA + protein)	**0.468**	**4/10**	**0.445**	**0.494**

Baseline. We compare RNAClip with DeepMIF (Song et al., 2026), SMARTBind (Jiang et al., 2025), DeepRNA-DTI (Bae & Nam, 2025), RNAsmol (Ma et al., 2025), DeepRSMA (Huang et al., 2024), EMMPTNet (Wu et al., 2025), BioLLMNet (Abir et al., 2025), AutoDock Vina (Trott & Olson, 2010), and AlphaFold 3 (Abramson et al., 2024). Vina is a pocket-aware reference using receptor coordinates and a native-ligand-defined search box; AF3 ranks candidates by the ipTM of a pLDDT-selected predicted complex. Implementation details are provided in Appendix C.3.

4.2 Screening performance

Table 1 summarizes the results on the recent PDB evaluation set. With the full auxiliary protein–ligand corpus, RNAClip achieves a PT-AP of 0.468, compared with 0.374 for DeepMIF, the strongest learned baseline. The improvement also appears in early recognition: RNAClip obtains BEDROC scores of 0.445 and 0.494 at $\alpha = 80.5$ and 20, respectively, compared with 0.317 and 0.419 for DeepMIF. The higher BEDROC scores indicate improved recovery of native ligands near the top of the ranking, a key requirement for virtual screening when only a small fraction of candidates can be tested experimentally. The advantage of RNAClip also extends to the evaluated structure-based approaches. AutoDock Vina achieves a PT-AP of 0.352, and $\text{BEDROC}_{80.5}$ of 0.304, falling below RNAClip in both overall ranking and early recognition. Under the fixed pLDDT-selection protocol, confidence-based AF3 scoring yields a PT-AP of 0.048, indicating limited ability to prioritize native binders in this setting.

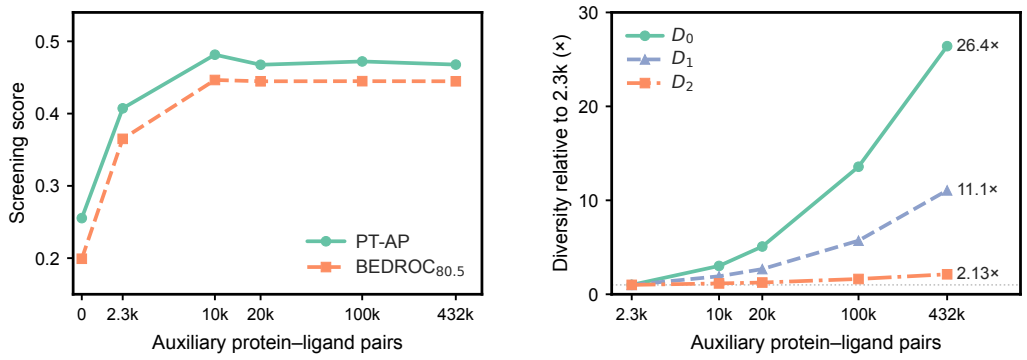
The RNA-only comparison in Table 1 assesses the contribution of auxiliary protein–ligand data. Adding this corpus increases PT-AP from 0.255 to 0.468 and BEDROC at $\alpha = 80.5$ from 0.199 to 0.445, while the number of targets with a rank-one native ligand increases from one to four. The gains in both average precision and early recognition support protein–ligand augmentation as an effective training strategy for RNA screening under the evaluated conditions.

4.3 Ablation Studies

We conduct a series of ablation studies to investigate how auxiliary protein–ligand data contribute to RNA–ligand virtual screening. Specifically, we examine the effects of auxiliary corpus size and selection, and test whether the screening gains depend on the original protein–ligand correspondences.

4.3.1 Auxiliary corpus size

We first investigate how the number of auxiliary protein–ligand pairs affects RNA screening. We compare RNA-only training with five random data sizes ranging from 2,366 pairs (matching the number of RNA–ligand training interfaces) to approximately 432k pairs (the full auxiliary dataset). The mixed configurations use the same sampling and optimization budget, allowing us to assess



(a) Screening performance

(b) Ligand diversity growth

Figure 2: Auxiliary corpus size, screening performance, and ligand diversity. (a) Screening performance. (b) Ligand diversity growth: Chemotype count (D_0), exponential Shannon diversity (D_1), and inverse-Simpson diversity (D_2)

the benefit of access to more auxiliary interfaces. Figure 2(a) shows that the smallest corpus raises PT-AP from 0.255 to 0.407 and BEDROC$_{80.5}$ from 0.199 to 0.365. Increasing the corpus to 10k pairs, only 2.3% of the full collection, further raises these scores to 0.481 and 0.447. Larger datasets yield PT-AP between 0.468 and 0.472 and BEDROC$_{80.5}$ of approximately 0.445, with no consistent additional gain. A small random subset therefore reaches the observed full-corpus screening level under the tested budget.

To verify that expanding the auxiliary corpus increases ligand chemical diversity, we quantify both chemotype coverage and effective diversity under the training sampler. We use chemotype count (D_0), exponential Shannon diversity (D_1), and inverse-Simpson diversity (D_2)(see Appendix D.2 for details). As shown in Figure 2(b), all three measures increase across the five data sizes. From the 2.3k subset to the full corpus, the number of chemotypes grows from 652 to 17,218 (26.4-fold), while D_1 and D_2 increase by 11.1-fold and 2.13-fold, respectively. These increases confirm that corpus expansion adds chemical diversity both in coverage and after accounting for sampling exposure.

Taken together, the two panels show that the limited screening gains beyond 10k pairs occur despite continued growth in ligand diversity. The absence of further improvement therefore cannot be attributed simply to a lack of new chemotypes: a chemically broader auxiliary corpus does not necessarily provide more useful training examples for RNA ranking under a fixed budget.

4.3.2 Auxiliary supervision and pretrained representations

The corpus-size study shows continued growth in chemical coverage without further screening gains beyond 10k pairs. We next fix the auxiliary corpus to Uniform-20k and ask whether its screening benefit requires explicit supervision of protein–ligand matches. We compare two interventions: negatives-only removes the auxiliary paired-supervision losses, whereas pair shuffle retains these losses but reassigns ligand partners among protein examples(see Section 3 for details).

Negatives-only masks both directional auxiliary anchor losses while retaining auxiliary embeddings in the RNA loss denominators, with RNA loss coefficients unchanged. As shown in Table 2, it achieves PT-AP 0.465, compared with 0.468 for joint training; both obtain four rank-one hits and BEDROC$_{80.5}$ 0.445 to three decimals. These closely aligned scores indicate that auxiliary matched-pair losses are not required to retain the observed screening level when auxiliary examples remain available as contrastive candidates.

Pair shuffle preserves the batch molecules and number of auxiliary loss terms but yields lower PT-AP (0.451) and BEDROC$_{80.5}$ (0.405), despite retaining four rank-one hits(Table 2). Thus, removing auxiliary paired supervision leaves the reported scores nearly unchanged, whereas imposing reassigned matches reduces early recognition. The negatives-only result therefore does not imply that arbitrary auxiliary pairings are equally effective.

Table 2: Ablation Studies: Corpus selection, auxiliary supervision, and training controls.

Category	Configuration	PT-AP ↑	Hits@1 ↑	BEDROC ↑	
				$\alpha = 80.5$	$\alpha = 20$
Corpus selection	Uniform-20k	**0.468**	**4/10**	**0.445**	**0.490**
	ProLIF-20k	0.371	2/10	0.311	0.488
	Ligand-diverse 20k	0.455	4/10	0.426	0.471
Auxiliary supervision	Negatives-only	0.465	4/10	0.445	0.489
	Pair shuffle	0.451	4/10	0.405	0.474
Training controls	No pretraining	0.265	2/10	0.220	0.294
	RNA-only	0.255	1/10	0.199	0.356

We then examine whether this performance depends on pretrained representations. With Uniform-20k and the joint objective unchanged, replacing the pretrained encoders with frozen random encoders reduces PT-AP to 0.265 and $\text{BEDROC}_{80.5}$ to 0.220(Table 2). This supports the importance of pretraining when only projection heads are learned.

4.3.3 Auxiliary corpus selection

We finally ask whether selecting different auxiliary interfaces improves screening at a fixed corpus size. We consider selection based on RNA-relevant interaction patterns and ligand chemotype coverage, using Uniform-20k as the size-matched reference.

We first replace uniform random sampling with interaction-fingerprint-based selection, yielding ProLIF-20k. Specifically, we use ProLIF fingerprints(Bouysset & Fiorucci, 2021) to characterize intermolecular contacts and favor protein–ligand interfaces with interaction patterns resembling those observed in RNA–ligand complexes (see Appendix D.2 for details). However, PT-AP decreases from 0.468 to 0.371, and $\text{BEDROC}_{80.5}$ from 0.445 to 0.311. Our analysis reveals that approximately 83% of training exposure is NAG-associated, while ligand D_2 drops from 37.9 to 2.8(Table 7). We hypothesize that the fingerprint-based selection preferentially captures NAG-associated interaction patterns, concentrating ligand exposure and potentially contributing to the performance decrease.

Motivated by this ligand concentration, we evaluate Ligand-diverse-20k, which explicitly prioritizes ligand chemotype coverage within the same 20k-pair budget. Comparison with Uniform-20k tests whether selecting for broader chemical coverage improves screening without increasing corpus size. As shown in Table 2, its PT-AP, hits@1, and $\text{BEDROC}_{80.5}$ are 0.455, 4, and 0.426, respectively. These results indicate that increasing ligand diversity alone does not yield consistent performance gains over Uniform-20k.

Taken together, these ablations support the importance of pretrained representations, while showing no further screening gains from increasing the auxiliary corpus beyond 10k pairs or from the evaluated interaction-based selection. Within the pretrained framework, the observed joint-training screening level can be retained when auxiliary examples contribute only as contrastive candidates and RNA–ligand pairs remain supervised. These results indicate that explicit supervision of auxiliary protein–ligand matches is not necessary for the screening performance observed in this setting.

5 Conclusion

We introduced RNAClip, a contrastive learning framework that adapts pretrained structure representations for pocket-free, pose-free RNA–ligand virtual screening with auxiliary protein–ligand data. We also constructed the recent PDB evaluation set through temporal splitting and structural novelty filtering; preliminary experiments on this set support the benefit of auxiliary protein–ligand data. Ablations show that a small random auxiliary subset reaches the observed full-corpus performance, while using auxiliary examples solely as contrastive negatives retains the observed joint-training performance. These findings highlight auxiliary contrastive negatives as a useful resource for RNA–ligand retrieval when experimentally validated RNA binding data are scarce.

AI USE STATEMENT

In this work, we used large language models solely for language polishing, including improving grammar, clarity, and readability of the manuscript. We did not use generative AI tools for any other research or writing tasks, including those requiring disclosure under the ICLR 2027 AI Policy for Authors. We reviewed all AI-assisted edits to ensure that they preserved the intended meaning and technical accuracy of the original text. We take full responsibility for the final content of this work, including all AI-assisted text.

REPRODUCIBILITY STATEMENT

Data construction, model implementation and training settings, and evaluation protocols are detailed in Appendices A, B, and C, respectively. The inference code and model weights are available in the anonymous repository at https://anonymous.4open.science/r/RNACLIP-FD60/.

REFERENCES

Abrar Rahman Abir, Md Toki Tahmid, and Md Shamsuzzoha Bayzid. Biollmnet: enhancing rna-interaction prediction with a specialized cross-llm transformation network. *Briefings in Bioinformatics*, 26(5):bbaf549, 2025.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Haelee Bae and Hojung Nam. Deeprna-dti: a deep learning approach for rna-compound interaction prediction with binding site interpretability. *Journal of Cheminformatics*, 18(1):3, 2025.

Cédric Bouysset and Sébastien Fiorucci. Prolif: a library to encode molecular interactions as fingerprints. *Journal of cheminformatics*, 13(1):72, 2021.

Juan G Carvajal-Patiño, Vincent Mallet, David Becerra, Luis Fernando Niño Vasquez, Carlos Oliver, and Jérôme Waldispühl. Rnamigos2: accelerated structure-based rna virtual screening with deep graph learning. *Nature Communications*, 16(1):2799, 2025.

Jessica L Childs-Disney, Xueyi Yang, Quentin MR Gibaut, Yuquan Tong, Robert T Batey, and Matthew D Disney. Targeting rna structures with small molecules. *Nature Reviews Drug Discovery*, 21(10):736–762, 2022.

Yuhan Fei, Pengfei Wang, Jiasheng Zhang, Xinyue Shan, Zilin Cai, Jianbo Ma, Yangming Wang, and Qiangfeng Cliff Zhang. Predicting small molecule–rna interactions without rna tertiary structures. *Nature Biotechnology*, pp. 1–12, 2026.

Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36:44595–44614, 2023.

Zhijian Huang, Yucheng Wang, Song Chen, Yaw Sing Tan, Lei Deng, and Min Wu. Deeprsma: a cross-fusion-based deep learning method for rna–small molecule binding affinity prediction. *Bioinformatics*, 40(12):btae678, 2024.

Yinjun Jia, Bowen Gao, Jiaxin Tan, Jiqing Zheng, Xin Hong, Wenyu Zhu, Haichuan Tan, Yuan Xiao, Liping Tan, Hongyi Cai, et al. Deep contrastive learning enables genome-wide virtual screening. *Science*, 391(6781):eads9530, 2026.

Shiyu Jiang, Amirhossein Taghavi, Tenghui Wang, Samantha M Meyer, Jessica L Childs-Disney, Chenglong Li, Mattew D Disney, and Yanjun Li. Small molecule approach to rna targeting binder discovery (smartbind) using deep learning without structural input. *bioRxiv*, 2025.

Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

Hongli Ma, Letian Gao, Yunfan Jin, Jianwei Ma, Yilan Bai, Xiaofan Liu, Pengfei Bao, Ke Liu, Zhenjiang Zech Xu, and Zhi John Lu. Rna–ligand interaction scoring via data perturbation and augmentation modeling. *Nature computational science*, 5(8):648–660, 2025.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

Carlos Oliver, Vincent Mallet, Roman Sarrazin Gendron, Vladimir Reinharz, William L Hamilton, Nicolas Moitessier, and Jérôme Waldispühl. Augmented base pairing networks encode rna-small molecule binding preferences. *Nucleic acids research*, 48(14):7690–7699, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Francesco P Panei, Rachel Torchet, Herve Menager, Paraskevi Gkeka, and Massimiliano Bonomi. Hariboss: a curated database of rna-small molecules structures to aid rational drug design. *Bioinformatics*, 38(17):4185–4193, 2022.

Leon Qiao, Wayne Bai, He Yan, Gary Liu, Nova Xi, Xiang Zhang, et al. Intfold: A controllable foundation model for general and specialized biomolecular structure prediction. *arXiv e-prints*, pp. arXiv–2507, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sergio Ruiz-Carmona, Daniel Alvarez-Garcia, Nicolas Foloppe, A Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E Hubbard, and S David Morley. rdock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS computational biology*, 10(4):e1003571, 2014.

Jinmiao Song, Annan Gao, Shengwei Tian, Qimeng Yang, Lei Deng, Qilin Feng, and Meitong Hou. Deepmif: A multiview interactive fusion-based deep learning method for rna–small molecule binding affinity prediction. *Journal of Chemical Information and Modeling*, 66(7):3575–3589, 2026.

Li-Zhen Sun, Yangwei Jiang, Yuanzhe Zhou, and Shi-Jie Chen. Rldock: a new method for predicting rna–ligand interactions. *Journal of chemical theory and computation*, 16(11):7173–7183, 2020.

Saisai Sun and Lin Gao. Contrastive pre-training and 3d convolution neural network for rna and small molecule binding affinity prediction. *Bioinformatics*, 40(4):btae155, 2024.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Jean-François Truchon and Christopher I Bayly. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of chemical information and modeling*, 47(2):488–508, 2007.

Zeyu Wu, Zhaohong Deng, Qunzhuo Wu, Yun Zuo, Xiaoyong Pan, Hongbin Shen, Xingze Fang, Yuxi Ge, and Shudong Hu. Explainable rna-small molecule binding affinity prediction based on multiview enhancement learning. *Journal of Chemical Information and Modeling*, 65(12): 6397–6416, 2025.

Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. Benchmarking all-atom biomolecular structure prediction with foldbench. *Nature Communications*, 17(1):442, 2025.

APPENDIX

A DATA CONSTRUCTION AND EVALUATION SETS

A.1 TRAINING DATA

We use the IntelliFold(Qiao et al., 2025) training datasets. It consists of 2,366 RNA–ligand interfaces from 903 PDB entries and 431,972 protein–ligand interfaces from 87,531 entries. Their 912 RNA and 67,011 protein interface clusters determine the sampling weights (Appendix B). All auxiliary-corpus experiments retain the same RNA–ligand training set; "Uniform" refers to random corpus selection, not uniform record exposure. The latest PDB releases are September 15, 2021 for RNA and September 29, 2021 for protein.

We define that each interface cluster corresponds to a unique combination of a receptor-cluster annotation and a ligand-chain CCD signature; interfaces sharing this combination belong to the same cluster. These interface clusters define the units balanced by the training sampler, with the sampling rule given in Appendix B.

A.2 RECENT PDB EVALUATION SET

A.2.1 TARGET SELECTION

The evaluation workflow adapts FoldBench to RNA–ligand targets (Xu et al., 2025), starting from a documented January 1, 2026 PDB/mmCIF snapshot. It indexes biological assembly 1, with limits of 300 MB per file and 12,000 biopolymer residues plus ligand atoms per assembly. These size limits do not themselves exclude every ribosomal entry.

Indexing removes entirely unknown polymer chains, chains with fewer than four resolved residues, implausible protein backbones, listed crystallization aids, and excluded ligands; the heavy-atom clash threshold is 0.10. A target must contain RNA and a ligand with heavy atoms and no recorded covalent connection. Its pocket comprises RNA residues within 7 Å of a ligand heavy atom, with at least three residues required. Pocket coordinates are used for curation, not as screening inputs.

Quality filters require resolution below 3.0 Å for X-ray diffraction or 4.5 Å for electron microscopy. X-ray ligand-validation records must additionally report occupancy 1, no stereochemical outliers or intermolecular clashes, and RSCC at least 0.6; these fields are not required for electron microscopy. The allowed ligand-contact fraction with non-primary polymer chains is at most 0.20. This assembly-based contact filter does not guarantee that an extracted RNA chain captures the entire binding environment.

To remove structural redundancy with historical PDB complexes, we use quality-filtered RNA–ligand structures released on or before January 13, 2023 as the reference set and assess candidates released after that date(we do not use the data between 2021-09-30 and 2023-01-13, because some baseline models use these as training). Global RNA sequence alignment identifies shared pocket positions; at least three mapped positions are required for a Kabsch fit. After ligand alignment, the overlap of occupied 2 Å voxel sets is

$$o(q, r) = \frac{|V_q \cap V_r|}{\min(|V_q|, |V_r|)}. \tag{5}$$

A candidate is removed if any reference has both pocket identity greater than 0.75 and ligand overlap greater than 0.55. These joint criteria permit repeated ligand identities and globally similar RNA sequences.

Retained candidates are clustered using the same thresholds, in order of better resolution, higher completeness, and earlier release. A candidate joins the qualifying centroid with the largest summed pocket identity and ligand overlap, or starts a new cluster.

Initial filtering and clustering yielded 109 representative RNA–ligand structures. We manually excluded all ribosome assemblies. Then, we excluded two representatives lacking mmCIF files, one whose ligand failed RDKit parsing, and seven without available decoys. Manual curation further removed four redundant conformers, seven iridium-hexammine phasing-reagent examples, and one

Table 3: Targets in the recent PDB evaluation set.

Target	Ligand	RNA (nt)	N	Cluster
8hze_A_K	NI4	35	100	LC03
8hzl_A_M	NI4	84	100	LC03
8u5k_A_G	VK0	36	100	LC05
8xzr_A_B	ANG	53	100	LC09
8yan_A_B	A1LYL	33	100	LC06
8zau_A_B	GUN	69	100	LC10
9bun_A_H	FH8	48	100	LC01
9fn3_A_C	A1IY6	58	80	LC07
9lkw_A_B	GUN	65	100	LC10
9mqt_A_B	A1BNU	332	100	LC02

incorrectly extracted small-polyamine ligand. The final dataset comprised 17 targets for validation and testing. Because RNAClip screens candidates using receptor sequences and ligand chemical specifications, we additionally assessed train–test overlap in the model's input space. The audit covered all 2,366 RNA–ligand pairs from 903 PDB entries in the downstream training corpus used for the released checkpoint. Receptor sequence similarity was defined as the number of matched nucleotides in a local alignment divided by the test receptor length. Ligand similarity was measured using ECFP4 fingerprints (Morgan radius 2, 2,048 bits) and Tanimoto similarity. For each test target, we defined joint pair similarity as the maximum, over all training pairs, of the minimum of receptor and ligand similarity. None of the ten test PDB entries appeared in this RNA–ligand training corpus, and all ten targets had joint pair similarity below 0.51.

A.2.2 Decoy selection

We draw decoys from the same filtered ZINC *in-vitro* bioactive compound subset used by RNAsmol (Ma et al., 2025), distributed through its official repository. The candidate library contains 306,179 records with distinct ZINC identifiers. Each record provides a SMILES string and a ZINC identifier.

Our input wrapper uses the DecoyFinder routines in RNAsmol's `modified_decoyfinder.py`. Relative to the native, candidates must lie within ± 25 Da molecular weight, ± 1 hydrogen-bond donor, ± 2 acceptors, ± 1 calculated logP, and ± 1 rotatable bond. MACCS-fingerprint Tanimoto similarity must be below 0.75 to the native and below 0.90 to every previously accepted decoy in that pool.

The search canonicalizes active SMILES, reuses searches for repeated actives, permits cross-active decoy reuse, and depends on library order. Exhausting the library can yield fewer decoys than requested. Our input-generation script reuses a saved active-to-decoy table with up to 100 decoys per active; finalized test pools contain 99 decoys for nine targets and 79 for `9fn3_A_C`. The scored identifiers define evaluation membership. A check of all 970 test decoys with RDKit 2026.03.1 confirms the five property tolerances and both similarity bounds. Maximum native–decoy and within-pool decoy–decoy similarities are 0.736 and 0.895.

A.2.3 Validation and test split

We split the 17 retained targets by PDB release year. The seven targets released in 2023 form the validation set, with 657 scored ligand candidates, including 57 for `7yci_B_C`. The ten targets released in 2024–2025 form the recent PDB evaluation set used for testing, with 980 candidates: one native ligand per target and 970 decoys in total. Nine test targets have 100 candidates each; `9fn3_A_C` has 80. Table 3 lists the test targets by PDB/receptor/ligand identifier, with CCD ligand codes, RNA lengths, scored-pool sizes N, and ligand-cluster assignments.

B Model implementation and training details

B.1 Encoder architecture and screening heads

We use IntelliFold version 0.1.0. The two independent encoder instances load the same IntelliFold checkpoint and return the final trunk single representation $\mathtt{s}$, with 384 features per token. The executed path contains a three-block atom-attention input embedder, a four-block MSA stack, recycling, and a 12-block Pairformer trunk; templates and diffusion structure generation are disabled. Our downstream task uses only the contrastive objective and trainable heads from Section 3.

Chemical inputs. Residues map to Chemical Component Dictionary (CCD) entries; ligand CCD entries or SMILES supply atom identities, elements, formal charges, atom names, and bonds. RDKit ETKDGv3 generates reference conformers. Failed generation falls back to CCD ideal coordinates, then eligible representative coordinates subject to the reference-date cutoff; remaining missing coordinates are zeroed. Each residue or ligand has its own reference frame, without a supplied receptor–ligand pose.

Initialization. Following AlphaFold 3 (Abramson et al., 2024), atom metadata and reference coordinates are linearly embedded. Reference offsets and inverse squared distances form atom-pair features for sequence-local attention. The outputs are projected, passed through ReLU, and averaged within each token. Concatenating residue-type, MSA-profile, and mean-deletion features yields x_i. Standard residues form one token each; ligand heavy atoms form individual tokens. Initial single and pair representations are

$$\begin{aligned} h_i^{(0)} &= W_s x_i, \\ p_{ij}^{(0)} &= W_{\text{left}} x_i + W_{\text{right}} x_j + e_{ij}^{\text{rel}} + e_{ij}^{\text{bond}}, \end{aligned} \tag{6}$$

with dimensions $H_s = 384$ and $H_p = 128$. The terms e_{ij}^{rel} and e_{ij}^{bond} encode relative positions and token bonds.

MSA, Pairformer, and recycling. The MSA stack updates pairs through outer product means, pair-weighted averaging, transitions, and triangular updates. Pairformer blocks apply outgoing and incoming triangular multiplication, starting- and ending-node triangular attention, and pair transitions. Updated pairs bias single-representation attention. For one head,

$$\alpha_{ij} = \text{softmax}_j\left(\frac{q_i^\top k_j}{\sqrt{d}} + b_{ij}\right), \qquad b_{ij} = w_b^\top \,\text{LN}(p_{ij}), \tag{7}$$

where q_i, k_j, v_j are query, key, and value vectors, and d is the query/key dimension. Gated, projected weighted values $\sum_j \alpha_{ij} v_j$ update the single representation residually, followed by a transition. Recycling layer-normalizes and projects the previous single and pair outputs, adds them to the initial representations, and repeats the trunk. The final single matrix $F \in \mathbb{R}^{L\times 384}$ enters Equation 1.

Each head has an affine $384 \to 512$ projection and layer normalization, for a total of

$$2\big(384 \cdot 512 + 512 + 2 \cdot 512\big) = 396{,}288$$

trainable parameters. Encoder gradients are disabled and the optimizer contains only head parameters. Temperature is fixed at 0.07. The complete normalized head is not a linear map.

B.2 Feature processing and optimization

Frozen encoders remain in training mode. The training token crop size is 384. For validation and testing, each encoder input is capped at 512 tokens: longer inputs are deterministically cropped to a contiguous window, while shorter inputs are retained in full. MSA depth is capped at 2,048; stochastic recycling gives one to three trunk passes in training and three at evaluation. Evaluated RNA inputs use query-only alignments, whereas training can read precomputed MSAs. The no-pretraining control omits checkpoint loading but still freezes both encoders, thus adapting heads to random features.

Each mixed epoch samples RNA and protein at a 1:1 ratio. Within each source $s \in \{\mathrm{R}, \mathrm{P}\}$, interfaces are sampled with replacement using inverse interface-cluster-size weights. Let $\mathcal{I}_s$ denote the selected source corpus, $g(i)$ the interface cluster of record i, $n_{g(i)}$ the number of records in that cluster within $\mathcal{I}_s$, and K_s the number of represented interface clusters. The unnormalized weight and source-conditional sampling probability are

$$w_i = \frac{1}{n_{g(i)}}, \qquad P(i \mid s) = \frac{w_i}{\sum_{j \in \mathcal{I}_s} w_j} = \frac{1}{K_s \, n_{g(i)}}. \tag{8}$$

Thus each interface cluster receives total probability $1/K_s$, with equal probability assigned to its member records. Cluster sizes and weights are recomputed within each selected auxiliary subset. The sampled RNA and protein records are shuffled into mixed batches, so the source ratio can vary within a batch.

Each GPU evaluates a local 12×12 score matrix without cross-device feature gathering. Eight GPUs and gradient accumulation 1 give 96 examples per optimizer update, but only 12 candidates per softmax. Distributed batch completion yields 48 updates per epoch and 4,800 over 100 epochs.

The principal mixed runs use bfloat16 precision on 8 NVIDIA H200 GPUs. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, and gradient clipping at norm 0.1. The learning rate warms up linearly to 10^{-4} over 1,000 optimizer updates and remains there through update 4,800. The scheduler advances per optimizer update; continuations restore its state without restarting warmup. We use a batch size of 12. The intended evaluation protocol selects the checkpoint with the highest PT-AP on the validation set and reports test-set metrics at that checkpoint.

B.3 Auxiliary-supervision interventions

Pair shuffle operates on protein-associated ligand positions at collation. With corruption probability one, it attempts a derangement up to eight times, then uses a nonzero cyclic shift if needed. RNA pairings and batch molecule sets remain unchanged. Repeated molecular identities can survive positional shuffling; reassigned pairs are not experimentally verified negatives.

C Evaluation protocol and detailed results

C.1 Targets and scored candidates

All methods use the same candidate identifiers and positive labels in the recent PDB evaluation set. The ten test targets and their 980 scored candidates are specified in Appendix A.2.3 and Table 3.

C.2 Metrics and ties

For one positive among N candidates, let a decoys score strictly above the positive and b tie with it. Average precision, midpoint rank, and hits are

$$\mathrm{AP} = \frac{1}{1 + a + b}, \qquad r = 1 + a + \frac{b}{2}, \qquad \mathrm{H@}\,k = \mathbb{1}[r \le k]. \tag{9}$$

PT-AP is the target mean of `average_precision_score`, named PT-AUPRC. Ties use exact equality of stored scores. AP includes the entire tied group; hits and reciprocal ranks use its midpoint. Thus AP equals $1/r$ for untied natives, and supplementary MRR averages $1/r$.

For early recognition (Truchon & Bayly, 2007), single-positive BEDROC is

$$\mathrm{BEDROC}_\alpha(r, N) = \frac{e^{-\alpha r/N} - e^{-\alpha}}{e^{-\alpha/N} - e^{-\alpha}}. \tag{10}$$

We use each target's actual N and midpoint r, then average over targets, with $\alpha = 20$ and 80.5. Midpoint-based hits and BEDROC are deterministic conventions, not expectations over random tie breaking. Supplementary AUROC uses standard score-tie handling and target averaging. All metrics increase with better ranking and summarize the same predictions.

C.3 BASELINE IMPLEMENTATIONS

DeepRNA-DTI, RNAsmol, and EMMPTNet use released task weights. We retrain SMARTBind, DeepMIF, DeepRSMA, and BioLLMNet. SMARTBind selects on its 133-complex HARIBOSS validation set; the other three select on our seven-target validation set. All four maximize PT-AP and retain the earlier epoch on ties. Training data and budgets remain method-specific.

DeepRNA-DTI. The 20 released checkpoints comprise five folds for each of unseen-pair, unseen-RNA, unseen-compound, and unseen-both scenarios (Bae & Nam, 2025). We average within each scenario, then equally across scenarios. Inputs use released RNA-FM and Mole-BERT encoders with RNA limited to 500 nucleotides.

RNAsmol. The main comparison uses released `pdb_bothaug_netperturbation` weights, molecular-graph preprocessing, and 500-token RNA padding/truncation (Ma et al., 2025). The RNA-perturbation checkpoint appears only as a supplementary diagnostic.

EMMPTNet. We average normalized affinity predictions from ten released folds (Wu et al., 2025). Inputs use 128-dimensional FastText RNA 5-mer embeddings, with normalization reconstructed from released training data and unseen 5-mers assigned its mean embedding. We reconstruct 17-dimensional atom features from SMILES because the released extractor returns only 14 dimensions, limiting exact replication.

DeepMIF. Released `model_pre20.pth` contains ligand-encoder pretraining weights, so we train the affinity model on 1,439 `All_sf_dataset_v1_2D.csv` examples spanning 341 RNA targets (Song et al., 2026). Training uses pretrained RNA-FM, supplied bpRNA `.st` annotations, batch size 32, seed 42, and 300 epochs; validation selects epoch 55. Evaluation RNA graphs use ViennaRNA-predicted secondary structures converted to module features.

DeepRSMA. Without released task-model weights, we train on 1,439 `All_sf` examples using pretrained RNA-FM, batch size 16, seed 1, and 1,500 epochs; validation selects epoch 242 (Huang et al., 2024). Evaluation contact graphs use SPOT-RNA-2D-Single predictions thresholded at 0.5. Both DeepRSMA and DeepMIF rank predicted affinities.

BioLLMNet. Six released models specialize by RNA type, which our evaluation inputs do not supply for checkpoint routing (Abir et al., 2025). We instead train one gated regression head on fixed mean-pooled BiRNA-BERT and Mole-BERT embeddings. Of 1,674 public rows, removing 215 with missing values or negative-label strings and 167 with invalid SMILES leaves 1,292. Training uses batch size 64, seed 42, and 300 epochs; validation selects epoch 168. This is a BiRNA-BERT/Mole-BERT reproduction, not a released ChemBERTa-based predictor.

SMARTBind. Released ten-fold weights use full HARIBOSS, which contains all seven of our validation targets (Jiang et al., 2025). We retrain the binding-score stage from pretrained RNA-FM and random projection heads. Of 1,246 complexes, 1,043 released before May 1, 2022 are used for training and 133 released in 2023 for validation; 33 from May–December 2022 and 37 from 2024 are excluded. A 100-epoch run uses batch size 24 and seed 2; HARIBOSS validation selects epoch 34. Ligands use Open Babel FP2 fingerprints and stored augmented decoys. This downstream date filter does not assert a common baseline pretraining cutoff.

AutoDock Vina. We evaluate AutoDock Vina docking scores (Trott & Olson, 2010) on exactly the 980 candidates used by the other methods. The original run contains 989 candidates; restricting it to the current panel removes nine obsolete `decoy_99` entries, with no missing predictions. Each ligand is docked with seeds 20260813–20260815; the screening score is the negative median of the three seed-best energies. Each seed uses one CPU, exhaustiveness 32, and up to 20 poses. Ligand conformers are generated from chemical specifications, without using the native pose as the starting conformation. Experimental RNA chains define the receptors. The search box is centered on the native ligand, with each dimension equal to its heavy-atom span plus 12 Å and a minimum of 20 Å. Seven local receptor repairs across `8u5k`, `8yan`, and `8zau` use aligned atoms from native-ligand

AF3 complexes; the unknown `8yan` residues retain their unknown base identity. Thus this reference uses structural and pocket information beyond the pocket-free inputs of RNAClip.

AlphaFold 3. We use a fixed AF3 predictor (Abramson et al., 2024) with seed 42 and five generated structures per RNA–ligand candidate. For each candidate, we select the structure with the highest mean atom pLDDT and use its ipTM as the screening score, with larger scores ranked first. This structure-selection rule does not use binding labels or select a checkpoint on the test set. All 980 test candidates have predictions. The stored ipTM values produce 30 decoys tied with their native ligand across nine targets; threshold AP is therefore 0.047553, whereas midpoint MRR is 0.052243. We retain these exact ties and apply the metric conventions above.

C.4 Early recognition and ligand-cluster sensitivity

Full-corpus RNAClip retrieves more native ligands within the shortest candidate lists than DeepMIF. At cutoffs of one, two, three, and five candidates per target, the counts are $4/10$, $5/10$, $5/10$, and $5/10$, respectively, compared with $3/10$, $3/10$, $4/10$, and $5/10$ for DeepMIF (Table 5). On this panel, RNAClip therefore reaches 50% target coverage with two candidates per target, whereas DeepMIF reaches the same coverage with five. The equal top-five counts locate the gain at the earliest positions.

We additionally assess whether the aggregate comparison changes when repeated native-ligand identities receive equal total weight. The ten targets form eight ligand clusters under the grouping defined in Appendix A: NI4 and GUN each occur in two targets, while the other six clusters contain one target each (Table 3). The primary target-macro AP therefore gives each of the two repeated-ligand clusters twice the weight of a singleton cluster. As a supplementary sensitivity analysis, we first average target AP within each cluster and then average over clusters. For the target set $\mathcal{T}$ and its partition $\mathcal{C}$ into ligand clusters,

$$\mathrm{AP}_{\mathrm{target}} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \mathrm{AP}_t, \qquad \mathrm{AP}_{\mathrm{cluster}} = \frac{1}{|\mathcal{C}|}\sum_{C\in\mathcal{C}} \frac{1}{|C|}\sum_{t\in C} \mathrm{AP}_t, \tag{11}$$

where $|\mathcal{T}| = 10$ and $|\mathcal{C}| = 8$. Both summaries retain the same target-level AP values, candidate pools, tie conventions, and checkpoints.

Table 4: Sensitivity to target and ligand-cluster weighting on the recent PDB evaluation set (higher is better).

Method	Target-macro AP	Ligand-cluster-macro AP
AutoDock Vina	0.352	0.306
AlphaFold 3	0.048	0.048
BioLLMNet	0.054	0.053
EMMPTNet	0.065	0.058
DeepRSMA	0.167	0.133
RNAsmol	0.190	0.228
SMARTBind	0.276	0.218
DeepRNA-DTI	0.293	0.334
DeepMIF	0.374	0.334
RNAClip (RNA only)	0.255	0.282
RNAClip (RNA + protein)	**0.468**	**0.546**

Table 4 applies both weighting schemes to every method in the main comparison. Under equal cluster weighting, full-corpus RNAClip has AP 0.546 and DeepMIF has AP 0.334, compared with target-macro values of 0.468 and 0.374. The ordering of these two methods is therefore preserved when repeated-ligand clusters receive the same total weight as singleton clusters. This reweighting describes sensitivity to the composition of the evaluation set; it does not add independent observations or replace target-macro AP as the primary metric.

Table 5: Supplementary retrieval metrics on the recent PDB evaluation set (higher is better).

Configuration	H@2	H@3	H@5	$BEDROC_{20}$	AUROC	MRR
RNAClip (full)	5/10	5/10	5/10	0.494	0.797	0.468
RNAClip (2.3k)	4/10	5/10	5/10	0.481	0.764	0.407
RNAClip (10k)	**5/10**	**5/10**	**5/10**	**0.528**	**0.849**	**0.481**
RNAClip (20k)	5/10	5/10	5/10	0.490	0.833	0.468
RNAClip (100k)	5/10	5/10	5/10	0.506	0.797	0.472
RNAClip (ProLIF)	4/10	5/10	5/10	0.488	0.890	0.371
RNAClip (shuffle)	4/10	4/10	5/10	0.474	0.844	0.451
RNAClip (negatives-only)	5/10	5/10	5/10	0.489	0.777	0.465
RNAClip (no pretraining)	2/10	3/10	3/10	0.294	0.765	0.265
RNAClip (RNA 2.3k)	3/10	3/10	4/10	0.356	0.769	0.255
DeepMIF	3/10	4/10	5/10	0.419	0.805	0.374
DeepRNA-DTI (overall)	2/10	4/10	4/10	0.350	0.784	0.293
SMARTBind (temporal)	2/10	2/10	3/10	0.323	0.851	0.276
RNAsmol (network)	2/10	2/10	2/10	0.232	0.588	0.190
DeepRSMA	1/10	1/10	1/10	0.197	0.808	0.167
RNAsmol (RNA)	0/10	0/10	0/10	0.057	0.625	0.049
EMMPTNet	0/10	0/10	1/10	0.112	0.566	0.065
BioLLMNet	0/10	0/10	0/10	0.060	0.669	0.054
AutoDock Vina	3/10	3/10	4/10	0.373	0.848	0.352
AlphaFold 3	0/10	0/10	0/10	0.068	0.556	0.052

Table 6: Full numerical results for the auxiliary corpus-size study.

Auxiliary corpus	Pairs	PT-AP ↑	Hits@1 ↑	BEDROC ↑	
				$\alpha = 80.5$	$\alpha = 20$
RNA-only	0	0.255	1/10	0.199	0.356
Uniform-2.3k	2,366	0.407	3/10	0.365	0.481
Uniform-10k	10,000	**0.481**	**4/10**	**0.447**	**0.528**
Uniform-20k	20,423	0.468	4/10	0.445	0.490
Uniform-100k	100,000	0.472	4/10	0.445	0.506
Uniform-full	431,972	0.468	4/10	0.445	0.494

Table 5 reports all four early cutoffs together with supplementary AUROC and MRR for the baselines and training configurations. Vina's three docking seeds are aggregated into one score per candidate.

D Supplementary analyses

D.1 Auxiliary corpus-size results

The detailed results are shown in Table 6. The performance panel includes the RNA-only control at zero auxiliary pairs; its horizontal axis is linear up to 2,366 pairs and logarithmic above. The diversity panel in Figure 2(b) covers only the five nonempty auxiliary corpora and uses a logarithmic horizontal axis. Its absolute values and normalization are given in Appendix D.2.

Table 7: Corpus composition. D_0 counts ligand chemotypes; D_1 and D_2 use expected exposure weights.

Corpus	Ligand diversity			Receptor D_2	Near RNA
	D_0	D_1	D_2		
Uniform-2.3k	652	154.8	30.3	809.5	0.496
Uniform-10k	1,967	296.6	35.0	1279.6	0.484
Uniform-20k	3,314	415.1	37.9	1462.2	0.477
Uniform-100k	8,851	883.5	49.3	1399.4	0.441
Uniform-full	17,218	1712.0	64.5	1102.1	0.393
ProLIF-20k	270	6.7	2.8	628.1	0.887
Ligand-diverse 20k	11,688	6866.9	1105.2	342.4	0.158

D.2 Exposure-weighted corpus composition

We distinguish the number of represented ligand chemotypes from their effective diversity under the training sampler. Within each auxiliary corpus, expected interface exposure is proportional to the inverse-cluster-size weight $w_i = 1/n_{g(i)}$ defined in Equation 8. The exposure share of chemotype c is

$$p_c = \frac{\sum_{i \in c} w_i}{\sum_i w_i}. \tag{12}$$

Writing $\mathcal{C}_+ = \{c : p_c > 0\}$, the three diversity measures are

$$D_0 = |\mathcal{C}_+|, \qquad D_1 = \exp\left(-\sum_{c \in \mathcal{C}_+} p_c \log p_c\right), \qquad D_2 = \left(\sum_{c \in \mathcal{C}_+} p_c^2\right)^{-1}. \tag{13}$$

The logarithm is natural. D_0 counts observed types regardless of exposure mass; D_1 and D_2 equal the number of equally exposed types with the same entropy or concentration, respectively. D_2 gives greater weight to common types. All three coincide for equal chemotype exposure.

Ligand chemotypes use deterministic leader/sphere-exclusion clustering of ECFP4 fingerprints at Tanimoto 0.55, with leaders ordered by full-corpus exposure mass. The same chemotype assignment is used across corpora; these groups differ from evaluation ligand clusters. Receptor diversity uses receptor-cluster identifiers. "Near RNA" is the exposure fraction with nearest training RNA-ligand similarity at least 0.4.

Table 7 reports the absolute diversity values. Figure 2(b) plots $D_q(N)/D_q(2{,}366)$ for $q \in \{0, 1, 2\}$, so each curve starts at one. The 2.3k baselines are $D_0 = 652$, $D_1 = 154.8411$, and $D_2 = 30.3133$; plotted ratios use the unrounded source values. In the full corpus, these become $17{,}218$, $1{,}712.0160$, and 64.4882, corresponding to 26.4-, 11.1-, and 2.13-fold growth, respectively, alongside a 182.6-fold increase in interface records. These deterministic corpus descriptors are separate from the screening measurements in panel (a).

The five Uniform corpora are nested. The additional chemotypes increase coverage substantially, while exposure-weighted effective diversity grows more slowly. Receptor D_2 also changes, increasing from 809.5 to $1{,}462.2$ at 20k before declining to $1{,}102.1$ in the full corpus; Near RNA falls from 0.496 to 0.393. Corpus expansion therefore changes several properties together.

ProLIF implementation. Interaction-based auxiliary selection compares the distribution of contact types in a protein–ligand interface with a reference RNA interaction profile. The pocket-based implementation uses ProLIF (Bouysset & Fiorucci, 2021) to detect contacts and aggregates them into a common vector representation across receptor types.

For each entry in the auxiliary protein–ligand inventory, the corresponding assembly mmCIF structure is parsed with Biopython, and the first structural model is used to identify the receptor and ligand chains. Chain roles are assigned from the interface-type annotations, requiring exactly one of the two chains to be annotated as a ligand. The protein pocket contains every residue with at

least one atom within 6 Å of any atom in the ligand chain. This neighborhood search uses the atoms present in the input structure, before hydrogen addition. The complete ligand chain and the selected protein residues are then serialized separately as PDB blocks and converted to RDKit molecules with sanitization disabled and existing hydrogens retained. Molecular preparation attempts a non-strict property-cache update, ring perception, and hydrogen addition with generated coordinates and residue annotations. Exceptions during these preparation operations are caught, and the resulting molecule is passed to the subsequent fingerprint calculation.

ProLIF evaluates seven interaction classes in a fixed order: van der Waals contacts, hydrophobic contacts, ligand hydrogen-bond donation, ligand hydrogen-bond acceptance, π–π stacking, ligand-cation–receptor-π contacts, and ligand-π–receptor-cation contacts. These correspond to `VdWContact`, `Hydrophobic`, `HBDonor`, `HBAcceptor`, `PiStacking`, `CationPi`, and `PiCation`, respectively. The fingerprint generator uses the default detection parameters for these classes. For interface i, let $\mathcal{P}_i$ denote the ligand–receptor residue pairs returned by ProLIF. The component for interaction class k is

$$c_{ik} = \sum_{(a,b)\in\mathcal{P}_i} \mathbf{1}\{\text{interaction class } k \text{ is detected between } a \text{ and } b\}. \tag{14}$$

Each residue pair contributes at most one count per class, even when multiple atom combinations satisfy its detection criteria. Summing over residue pairs yields a seven-dimensional vector $\mathbf{c}_i$ without retaining residue identities, allowing protein and RNA interfaces to be compared in the same interaction space.

We construct the reference vector $\boldsymbol{\mu}_{\mathrm{RNA}}$ by computing interaction-count vectors for all RNA–ligand interfaces in the training set and taking their arithmetic mean. The reference and candidate vectors are aligned by interaction name. Raw counts are averaged before cosine normalization; per-class standard deviations are not used. Each nonzero candidate vector receives the score

$$s_i = \frac{\mathbf{c}_i^{\top}\boldsymbol{\mu}_{\mathrm{RNA}}}{\|\mathbf{c}_i\|_2\,\|\boldsymbol{\mu}_{\mathrm{RNA}}\|_2}. \tag{15}$$

This score measures agreement in the relative abundance of interaction types with the mean reference profile.

The implementation retains an interface only when $s_i > 0.75$. Entries with missing structures or chains, empty ligand or pocket selections, failed molecular conversion or fingerprint calculation, or zero total interaction counts are excluded. Interfaces are processed independently in a process pool with a 60-second timeout per entry; timed-out entries are also excluded. Each retained record preserves the original interface annotations and adds its cosine score and interaction-count vector. Selection applies this threshold directly, without a fixed-size quota or additional ligand-diversity balancing.

Through analysis, we found that approximately 83% of exposure is NAG-associated. In an exposure-weighted sample of 200 examples, 45% have a ligand–standard-amino-acid distance below 1.8 Å, a proxy for covalent linkage rather than a full-corpus bond census. These diagnostics describe concentrated, heterogeneous contacts.

Ligand-diverse implementation. Ligand-diverse 20k selects $B = 20{,}423$ auxiliary protein–ligand interfaces, matching Uniform-20k, by combining ligand-composition diversity with ProLIF similarity ranking.

We scored all 431,972 interfaces in the auxiliary inventory (Appendix A) using the same 6 Å pocket, molecular preparation, seven interaction classes, and mean reference vector described above (Eqs. 14–15). Chains were resolved using mmCIF label identifiers, with author identifiers as fallback. Each input received a processing status under a 60-second timeout. Of these interfaces, 338,830 had nonzero interaction counts, 27,699 had zero counts, and 65,443 failed or timed out. Only successfully scored interfaces with nonzero counts were eligible; the score threshold was zero, with no additional positive cosine cutoff.

Eligible interfaces were partitioned into 27,742 groups using the sorted multiset of parenthesized Chemical Component Dictionary (CCD) codes in each ligand-chain cluster annotation. This key ignores component order but preserves multiplicity; annotations without parenthesized codes retain their full string as the key. These exact component-composition groups differ from the

ECFP4/Tanimoto chemotypes used in the exposure-weighted diversity analysis: chemically similar ligands with different CCD codes may belong to different groups.

Interfaces were stably sorted by descending cosine score, preserving scoring-table order for ties. Groups were ordered by their highest-scoring member, and interfaces were selected by round-robin traversal, taking the best remaining member of each group per pass. Because the number of eligible groups exceeded B, selection ended in the first pass:

$$i_g = \arg\max_{i \in g} s_i, \qquad \mathcal{S} = \{i_g : g \in \mathrm{TopB}_{g' \in \mathcal{G}}(s_{i_{g'}})\}, \tag{16}$$

where $\mathcal{G}$ denotes the eligible groups, TopB returns the B highest-scoring groups, and ties follow the stable ordering. Thus, each selected group contributed one interface.

Selected interfaces were mapped to the full inventory by assembly and chain identifiers, preserving the eight annotation fields and source-row order. The 2,366-interface RNA manifest from Uniform-20k was retained unchanged. The resulting auxiliary corpus contained 20,423 component-composition groups, 19,789 distinct CCD codes, 3,552 receptor clusters, and 20,423 interface clusters. Individual CCD codes can occur in multiple groups.

E LIMITATIONS

Our study still has some limitations. First, the test set contains only ten RNA targets with limited ligand diversity, restricting the precision of performance estimates and conclusions about generalization. Second, our ProLIF-based selection aggregates seven interaction types against a single mean reference profile, losing local spatial detail and potentially obscuring distinct binding modes. Finer-grained selection with diversity control warrants further study. Third, each target has one known positive ligand and property-matched decoys rather than experimentally confirmed nonbinders. This evaluation measures recovery of known ligands; ranking multiple active ligands and discovering new binders require broader benchmarks and prospective experimental validation.