First Comprehensive Benchmark for Tailored Small Molecule-Binding Aptamer Design

Mariia Eremeyeva **Nikita S. Serov**
Center for AI in Chemistry, ITMO University
St. Petersburg, Russia
eremeeva_maria@pish.itmo.ru, serov@pish.itmo.ru

Abstract

Aptamers are emerging as robust recognition elements for diagnostics and therapeutics, yet computational discovery pipelines remain limited to proteins, leaving small-molecule binding largely unexplored. To fill this gap, we present the first unified benchmark for aptamer–small molecule interactions, built from seven curated sources and comprising 2,210 annotated pairs, 1,430 unique DNA- and RNA-based aptamers, and 496 ligands spanning a broad chemical space. Over half of the entries include quantitative binding affinities, enabling both classification and regression tasks, while synthetic negatives generated via cross-pair sampling allow to rationally balance the dataset. Using this dataset, we conducted a systematic benchmarking study across multiple splitting and representation strategies for both aptamers and ligands. Our experiments covered discrete encodings, pretrained embeddings, and hybrid fusion schemes, evaluated with both shallow and deep learning (DL) models. This analysis establishes stable baselines for binding prediction and reveals the strengths and weaknesses of sequence- and embedding-based features. Beyond classification, we also provide the first regression baselines isolating the impact of aptamer-molecule compositional information on quantitative binding affinity estimation. This framework represents the next step toward scalable, data-driven aptamer discovery beyond SELEX-based single target-centered models and large scale computational screening using molecular docking.

1 Introduction

Aptamers are short single-stranded DNA or RNA oligonucleotides that can bind to ions, small molecules, proteins, and even cellular surfaces with high affinity [1]. Compared to antibodies, they are easier to synthesize, more stable, and easily modified, making them attractive for diagnostics, therapeutics, and detection [2].

Traditional discovery relies on target-specific SELEX (Systematic Evolution of Ligands by EXponential enrichment), a time- and resource-intensive experimental approach prone to biases such as limited success rates and the loss of rare but functional sequences during selection stage [3]. Despite numerous variants of SELEX exist, the method remains laborious and limited in scalability [4, 5].

In silico alternatives such as docking and molecular dynamics can model aptamer–target interactions, but are computationally demanding and unsuitable for large-scale design. Unlike classical drug discovery, the task here is inverse: identifying receptors (aptamers) for given targets, which greatly expands the search space.

Deep learning (DL) offers a third option, enabling prediction and generation of aptamers from accumulated data [6]. Yet, most research to date targets proteins, while small-molecule binders remain underexplored due to scarce datasets and limited generalization beyond specific case studies.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop AI4Science.

In this work, we:

- construct the first benchmark for classification and regression of aptamer–small molecule interactions across curated datasets;
- evaluate the impact of disjoint train–test splits and realistic class imbalance on model performance;
- compare numerous sequence representations, from simple one-hot and k-mer encodings to pretrained nucleotide embeddings with both shallow and DL models, highlighting their respective strengths and limitations.

2 Related Works

Current *in silico* aptamer design approaches fall into three groups: (1) optimization-based methods that pair predictive models with search algorithms, (2) target-specific generative models trained on SELEX data, and (3) structure-based approaches using spatial information from complexes.

Optimization-based methods use predictors (e.g. random forest, multi-layer perceptrons, transformer-based architectures) as scoring functions combined with Bayesian or evolutionary search [7]. Although applied to proteins they suffer from low sample efficiency and highly discrete search space.

SELEX data-based generative models (variational autoencoders, diffusion- and transformer-based models) have demonstrated the ability to capture the sequence distribution enriched during SELEX and to generate novel candidate aptamers. Representative examples of such approaches are AptaGPT [8], RaptGen [6], and AptaDiff [7]. However, they rely on large SELEX datasets, focus on RNA–protein systems, and rarely cover small molecules. Moreover, these strategies are tailored for a specific single target and lack generalizability.

Structure-based methods (e.g., AiDTA [9], RhoDesign [10]) attempt to guide sequence generation *via* docking or shape-conditioned learning. These approaches remain protein-centric, limited to rare well-characterized complexes, and not transferable to small molecules.

Despite progress, existing methods lack generalizable target-conditioned design for small-molecule binders. Our work addresses this gap by introducing the first comprehensive benchmark for aptamer–small molecule interactions.

3 Benchmarking Aptamer-Ligand Prediction

3.1 Datasets

Our benchmark dataset integrates data from seven curated sources, namely, RSAPred [11], AptamerBase [12], Apta-Index database (AptaGen), UTexas [13], RiboCentre, AptaDB [14], as well as manually curated dataset described in detail in the Appendix (Table 4, Table 5) which complements existing resources by adding missing target classes and affinity annotations. These databases comprise highly heterogeneous data with varying structural diversity of ligands, dominating type of aptamers (DNA or RNA, depending on the source), aptamer complexity, and annotation completeness, which complicates aptamer property modeling and conditional generation. Merge and unification of this data yields a total of 2,210 pairs spanning 1,430 unique aptamers and 496 ligands, with 50% sample coverage by dissociation constants K_d quantifying binding affinity, thereby forming a representative corpus for evaluating model generalization.

Closer data examination reveals heterogeneity stated above. t-SNE projections of unique aptamers and molecules (Figure 1) highlight distinct patterns of variability with some sources being conservative (e.g., RSAPred, AptaGen) or diverse (Manual, AptaDB) in terms of aptamers, while molecules from different sources largely overlap, reflecting shared chemical space with uneven coverage across scaffolds.

Overall, the final dataset provides a robust testbed for evaluating model generalization, sensitivity to sequence variability, and transferability across target types. Its diversity in molecular space, aptamer length, annotation type, and experimental origin highlights both the richness and the challenges of aptamer design for small molecule recognition.

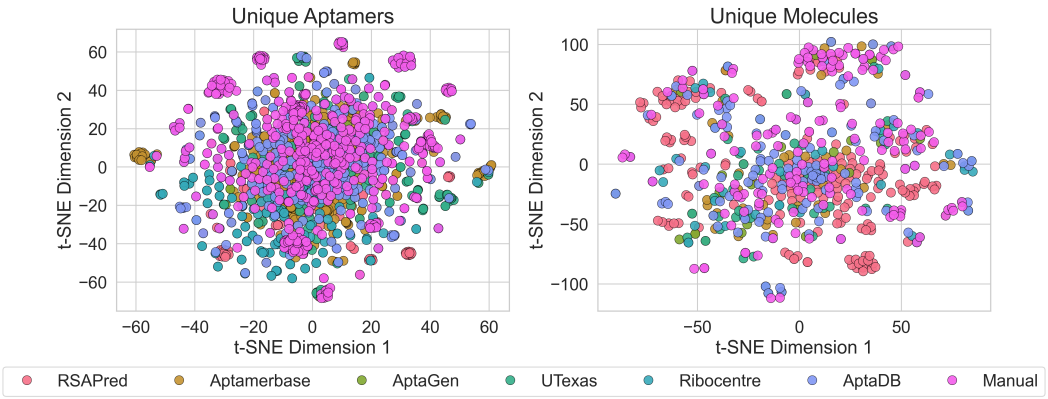

Figure 1: t-SNE projections of aptamer sequences (left) and target molecules (right), colored by origin. Each point represents an individual aptamer or molecule, embedded in 2-dimensional space using t-SNE from one-hot encoded sequences or Morgan molecular fingerprints, respectively. Coloring indicates the dataset source, revealing overlap or separation between the datasets. Compact clusters suggest internal redundancy, while dispersed patterns reflect broader diversity across sources.

3.2 Experiments

3.2.1 Negative sampling strategy

Negative samples were generated by cross-pairing aptamers and small molecules from the dataset for which no confirmed interactions are known. For each positive interaction, n negative pairs were created by pairing the aptamer with n different small molecules, effectively multiplying the dataset size by a factor of n. This approach, previously applied in protein–aptamer interaction studies, balances the training data by mixing entities to create presumed non-binding pairs [15]. However, it does not guarantee true absence of interaction, as some cross-paired samples may still bind but remain uncharacterized.

To explicitly assess the effect of negative sampling on model reliability, we evaluate multiple negative-to-positive ratios, including the absence of augmentation. As shown in Figure 2, models trained without negative augmentation exhibit low MCC, indicating limited robustness. Introducing synthetic negatives leads to a substantial and consistent improvement in MCC across all descriptor combinations. Performance peaks at moderate sampling ratios (2:1-3:1), while more aggressive augmentation yields substantial MCC drop. Based on this analysis, we adopt a ratio of 3:1 negatives-to-positives in all subsequent experiments.

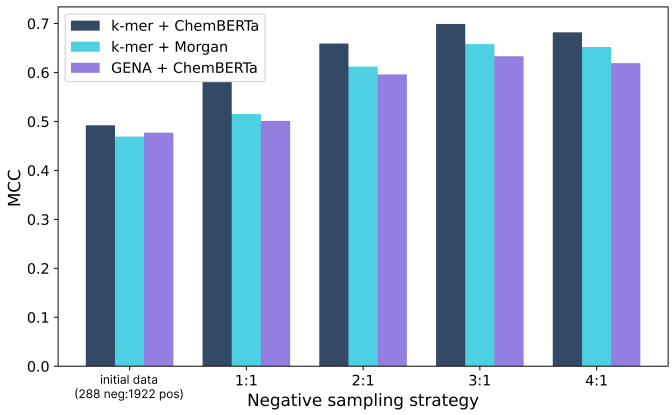

Figure 2: Effect of negative sampling strategies on MCC for different descriptor combinations using a LightGBM model under grouped cross-validation.

3.2.2 Splitting Protocols

To support reproducible evaluation, we define three complementary splitting protocols: (i) **stratified group splits** preserving label balance across folds, (ii) **aptamer-disjoint splits**, and (iii) **molecule-disjoint splits**.

In stratified group splits (Figure 3, center), individual aptamer–molecule pairs are assigned to training or test sets such that the overall distribution of binding affinities remains balanced. Importantly, this scheme does not enforce exclusivity over entities: the same aptamer or the same molecule may occur in both training and test sets, albeit in association with different partners. For example, an aptamer tested against two ligands may contribute one pair to training and the other to testing.

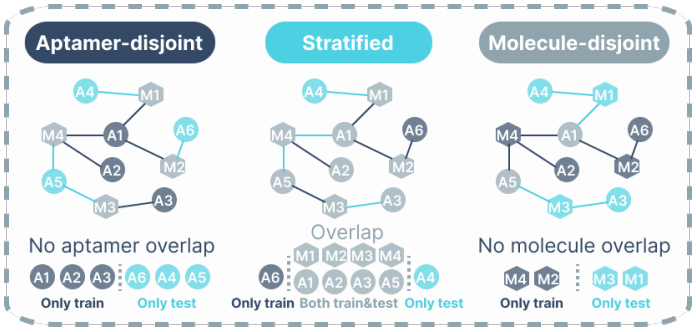

Figure 3: Illustration of the three data splitting strategies (stratified, aptamer-disjoint, and molecule-disjoint) represented as graphs of interactions between aptamers and molecules.

In contrast, in molecule-disjoint splits (Figure 3, right), exclusivity is applied to ligands rather than sequences. All binding interactions of a given molecule are confined to one partition, such that the model is forced to generalize across unseen ligands even if the training set contains overlapping aptamers. This configuration highlights the ability of the model to predict binding outcomes for entirely new molecular scaffolds.

Finally, aptamer-disjoint splits (Figure 3, left) group all ligand interactions associated with a given sequence together. If an aptamer binds to multiple molecules, all resulting pairs are assigned either entirely to training or entirely to testing. This ensures that no sequence representation seen during training reappears in validation or test evaluation, better reflecting the challenge of predicting binding for completely novel aptamers.

These distinct splitting strategies correspond to different practical scenarios and task complexities. Stratified splits simulate performance in settings where both aptamers and molecules are partially known but new interactions are being discovered. Aptamer-disjoint splits address the challenge of screening entirely novel sequences, such as in *de novo* aptamer design. Molecule-disjoint splits are probably the most indicative, since they evaluate generalization to new ligands, relevant for applications in biosensors design where predicting binding for novel molecular targets is critical. Together, these approaches provide a comprehensive framework for assessing model robustness across varied real-world tasks.

3.2.3 Aptamers and Molecules encoding

Aptamers were encoded as k-mers, one-hot encodings, or pretrained oligonucleotide embeddings [16] , while ligands were encoded by fingerprints (Morgan, MACCS), RDKit descriptors, or ChemBERTa embeddings [17] depending on specific setup. All shallow and DL models were compared with tuned hyperparameters, reporting ROC-AUC, F_1, and MCC for classification and RMSE, MAE, R^2, r_p, and r_s for regression.

3.2.4 Baseline Models

We benchmark shallow models (LightGBM, Random Forest, and MLP) by concatenating aptamer and ligand descriptors. Given the pronounced class imbalance in the dataset (1922 positive vs. 288 negative pairs), we evaluate all models under the fixed negative sampling regime selected in the previous analysis, based on cross-pairing aptamers and small molecules with no reported interactions.

A sanity check with randomized features under grouped cross-validation confirms that precision-oriented metrics such as PR-AUC and F1-score remain artificially inflated under imbalance, whereas ROC-AUC and MCC collapse to chance-level values. This sanity check was restricted to grouped splits, as it targets metric behavior rather than generalization.

Table 1 summarizes the performance of LightGBM baselines across different representation choices and evaluation splits under a fixed negative sampling ratio. Under grouped cross-validation, ChemBERTa-based ligand representations achieve higher ROC-AUC and MCC compared to classical fingerprints, while simple k-mer encodings outperform both one-hot and pretrained oligonucleotide embeddings. We attribute this behavior to the short length of aptamer sequences: pretrained models such as GENA-LM are optimized for substantially longer genomic contexts, whereas local k-mer statistics remain effective and robust in short-sequence regimes.

Across all evaluated settings, LightGBM achieves performance comparable to Random Forest baselines while offering improved training stability and substantially lower computational cost, making it a practical and reliable choice for large-scale benchmarking. Detailed comparisons with additional shallow models are reported in Appendix A.

To probe generalization beyond interpolation, we further evaluate the best-performing configurations under identity-disjoint splitting protocols. While holding out unseen aptamers results in a moderate degradation in performance, generalization to unseen molecules proves substantially more challenging, with MCC dropping sharply across all representations.

Table 1: Performance of LightGBM baselines across different representations and evaluation splits using a fixed negative sampling ratio of 3:1. Metrics are reported as mean values over 5-fold cross-validation.

Aptamer	Ligand	Grouped CV				Identity-disjoint MCC	
		ROC-AUC	MCC	F1	PR-AUC	Aptamer	Molecule
k-mer (k=4)	ChemBERTa	**0.90**	**0.70**	**0.77**	**0.86**	**0.63**	0.34
k-mer (k=4)	Morgan	0.88	0.66	0.74	0.83	0.61	**0.35**
GENA-LM	ChemBERTa	0.87	0.63	0.72	0.82	0.59	0.31
GENA-LM	Morgan	0.84	0.57	0.69	0.78	0.54	0.29
Random feat.	Random feat.	0.45	0.00	0.67	0.75	–	–

Feature analysis showed that aptamer k-mer and ChemBERTa features contribute nearly equally to the LGBM decision function (59% vs. 41%), confirming that both feature spaces are equally important for classification.

Having established robust tabular baselines, we next compare a set of deep learning architectures to assess whether end-to-end training on pretrained encoders can better capture cross-entity interaction patterns.

3.2.5 Pre-Trained Architectures

As shown in the previous section, tabular baselines based on k-mer and molecular fingerprints with LGBM already achieve strong and stable performance, but they rely on shallow concatenation and struggle with molecule-disjoint generalization. Motivated by prior success of pretrained encoders in aptamer-protein interaction modeling [18], we assess whether a similar paradigm can improve generalization for aptamer-small molecule prediction. Rather than proposing a new architecture, our goal is to systematically evaluate whether end-to-end deep learning models can close the performance gap observed for tabular baselines.

Setup. We employed **GENA-LM** as the aptamer encoder and **ChemBERTa** as the molecular encoder. We focus on GENA-LM as it supports variable-length oligonucleotides and has been shown to perform robustly across diverse regulatory DNA/RNA tasks; evaluation of alternative nucleotide language models is left for future work as well as testing of other molecular encoders. On top of frozen or partially fine-tuned embeddings, several architectural variants were evaluated, namely, *Identity* (direct projection with optional linear mapping), *CNN* (1D convolutions with global pooling),

LSTM (bidirectional LSTM with attention pooling), and *Transformer* blocks. For all configurations we used gated fusion to integrate aptamer and molecule embeddings, followed by an MLP head with dropout and normalization. Evaluation covered stratified group splits and identity disjoint protocols (aptamer-identity disjoint and molecule-identity disjoint).

Gated fusion. To combine aptamer and small-molecule representations, we use a gated fusion mechanism applied to the concatenation of the two embeddings. Given aptamer and ligand embeddings, they are first concatenated and passed through a lightweight gating network consisting of two linear layers with a ReLU nonlinearity and a sigmoid output. The resulting scalar gate is used as a stabilizing factor during fusion, allowing gradients to propagate through the gating network while preserving the original concatenated representation.

Results. Table 2 summarizes the top-performing configurations. Under stratified group splits, the three best models (Identity-LSTM, Identity-Transformer, Identity-CNN) achieve nearly identical performance (MCC ≈ 0.41), suggesting that once pretrained embeddings are available, the specific choice of top-layer encoder has only a secondary effect. All strong models relied on gated fusion, and partial unfreezing of the last two layers consistently yielded higher stability than full freeze.

Despite this consistency, deep learning models underperform compared to the tabular LightGBM baseline (MCC ≈ 0.70), indicating that end-to-end training with pretrained encoders does not yet close the performance gap in this setting. Notably, performance under aptamer-identity disjoint splits remains statistically comparable to that observed under grouped cross-validation (MCC ≈ 0.41), suggesting that GENA-LM representations generalize reliably to unseen aptamer sequences.

In contrast, molecule-identity disjoint evaluation reveals a pronounced degradation in performance, with MCC dropping to approximately 0.18. We attribute this behavior primarily to the limited coverage coupled with a high chemical diversity of small molecules in the current dataset, which constrains the ability of molecular encoders to extrapolate to unseen ligands. This result highlights data sparsity and limited coverage of chemical space, as well as the limited amount of training data available for deep learning, rather than architectural limitations, as the dominant bottleneck for generalization in aptamer-small molecule prediction.

Table 2: Top DL configurations across split types. Metrics are mean$\pm$std over 5 folds. All use gated fusion; aptamer encoder = GENA-LM, molecule encoder = ChemBERTa.

Split	Model	Partial Unfreeze	MCC	ROC-AUC
Grouped	Identity-LSTM	Last 2 layers	0.412 ± 0.034	0.726 ± 0.028
Grouped	Identity-Transformer	Last 2 layers	0.413 ± 0.029	0.722 ± 0.029
Grouped	Identity-CNN	Last 2 layers	0.408 ± 0.028	0.731 ± 0.021
Aptamer-disjoint	Identity-LSTM	Last 2 layers	0.407 ± 0.025	0.721 ± 0.023
Molecule-disjoint	Identity-LSTM	Last 2 layers	0.186 ± 0.051	0.585 ± 0.040

In summary, our results indicate that deep learning architectures, even when combined with strong pretrained encoders, do not yet outperform optimized LightGBM baselines in the considered setting. While aptamer representations derived from GENA-LM exhibit stable behavior under identity-disjoint evaluation, generalization across small molecules remains substantially more challenging. This effect is primarily driven by the higher structural diversity of chemical space, compounded by the limited number and coverage of small molecules available in the current dataset. Within this regime, the choice of top-layer architecture has only a minor influence on performance, with CNN, LSTM, and Transformer heads converging to similar results. All evaluated models employed gated fusion as a stabilizing integration mechanism, and partial unfreezing of encoder layers consistently improved training stability. Taken together, these findings suggest that in realistic discovery scenarios dominated by chemically diverse ligands, data coverage and representation of chemical space play a more critical role than architectural complexity. Interestingly, a similar pattern has been reported for aptamer-protein prediction, where APIPred [19], based on XGBoost and handcrafted features, outperformed more complex deep learning approaches.

3.2.6 Regression Task

In addition to binary binding prediction, we explore a regression setting aimed at estimating quantitative binding affinities. Following the classification results, we adopt the same representation—aptamer k-mer($k = 4$) descriptors combined with Morgan fingerprints for small molecules. An Optuna-tuned LightGBM regressor was trained and evaluated using cross-validation on samples with available affinity annotations.

Table 3: Cross-validation metrics for LightGBM regression with aptamer k-mer($k = 4$) and Morgan fingerprints (mean $\pm$ std over folds).

Model	RMSE $\downarrow$	MAE $\downarrow$	$R^2 \uparrow$	$r_P \uparrow$	$r_S \uparrow$
LGBMRegressor	2.42 ± 0.13	1.51 ± 0.10	0.458 ± 0.034	0.678 ± 0.025	0.624 ± 0.029

We emphasize that quantitative affinity prediction remains intrinsically challenging in this domain. Reported dissociation constants (K_d) are highly sensitive to experimental conditions, including buffer composition, ionic strength, pH, temperature, and the presence of cofactors, which introduces substantial heterogeneity into aggregated datasets. We therefore position these experiments as baseline estimates achievable from sequence- and structure-derived descriptors alone, rather than as precise predictors of absolute binding constants.

Despite these limitations, the obtained correlations indicate that even coarse regression models capture meaningful trends in binding strength. In particular, predicted affinities are sufficient to distinguish broad activity regimes—such as nano-, micro-, and millimolar binders—which is practically useful for prioritizing candidates and guiding experimental optimization. Feature importance analysis further shows that aptamer descriptors account for the majority of the predictive signal (71% of total gain), with molecular fingerprints contributing the remaining 29%, reflecting a combination of sequence-level motifs and chemical substructures. Overall, these results suggest that regression models can complement classification by enabling ranked screening under realistic data constraints, even when precise affinity estimation remains out of reach.

3.2.7 Future Directions

Our results highlight two main bottlenecks. First, generalization to unseen ligands remains a key challenge, driven by the high diversity of chemical space and the limited coverage of small molecules in current datasets. Future work will explore more explicit representations (e.g. GNN encoders) as more faithful descriptors of chemical structure. Second, our current gated fusion is limited; richer cross-entity integration, such as cross-attention layers, may better capture sequence-ligand dependencies. Finally, both classification and regression models are planned to be integrated into a reinforcement learning pipeline, where they will act as reward functions guiding sequence generation toward higher binding likelihood and affinity. This closes the loop from predictive modeling to generative design, paving the way for practical aptamer discovery workflows.

4 Conclusion

In this work we presented a comprehensive benchmark for aptamer–small molecule recognition, integrating seven curated datasets into a unified corpus with both classification and regression labels. Through systematic evaluation across shallow and deep models, we found that simple tabular approaches currently outperform end-to-end deep learning architectures, largely due to the high diversity of chemical space and limited ligand coverage in current datasets. These results establish strong baselines and highlight representation bottlenecks that must be addressed for further progress.

Taken together, this benchmark provides a standardized testbed for evaluating algorithms, clarifies current limitations of aptamer–ligand prediction, and suggests directions toward data-driven generative pipelines that may extend beyond SELEX and docking.

Acknowledgments and Disclosure of Funding

This research was supported by the ITMO University Research Projects in AI Initiative (RPAII), Project No. 640100. The authors declare no competing interests.

References

[1] Anthony D. Keefe, Supriya Pai, and Andrew Ellington. Aptamers as therapeutics. *Nature Reviews Drug Discovery*, 9(7):537–550, July 2010. Publisher: Nature Publishing Group.

[2] Varatharasa Thiviyanathan and David G. Gorenstein. Aptamers and the Next Generation of Diagnostic Reagents. *Proteomics. Clinical applications*, 6(0):563–573, December 2012.

[3] Andrew D. Ellington and Jack W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6287):818–822, August 1990. Publisher: Nature Publishing Group.

[4] Natalia Komarova and Alexander Kuznetsov. Inside the Black Box: What Makes SELEX Better? *Molecules*, 24(19):3598, October 2019.

[5] Michael Kohlberger and Gabriele Gadermaier. SELEX: Critical factors and optimization strategies for successful aptamer selection. *Biotechnology and Applied Biochemistry*, 69(5):1771–1792, October 2022.

[6] Generative aptamer discovery using RaptGen | Nature Computational Science.

[7] Zhen Wang, Ziqi Liu, Wei Zhang, Yanjun Li, Yizhen Feng, Shaokang Lv, Han Diao, Zhaofeng Luo, Pengju Yan, Min He, and Xiaolin Li. AptaDiff: de novo design and optimization of aptamers based on diffusion models. *Briefings in Bioinformatics*, 25(6):bbae517, September 2024.

[8] AptaGPT: Advancing aptamer design with a generative pre-trained language model | bioRxiv.

[9] Gaoxing Guo, Liangwei Guo, Jiaqiang Qian, Xiaoming He, Xinzhou Qian, Lei Wang, and Qiang Huang. De novo design of protein-binding aptamers through deep reinforcement learning assembly of nucleic acid fragments, June 2025. Pages: 2025.06.01.657174 Section: New Results.

[10] Deep generative design of RNA aptamers using structural predictions | Nature Computational Science.

[11] Sowmya R. Krishnan, Arijit Roy, and M. Michael Gromiha. Reliable method for predicting the binding affinity of RNA-small molecule interactions using machine learning. *Briefings in Bioinformatics*, 25(2):bbae002, January 2024.

[12] Jose Cruz-Toledo, Maureen McKeague, Xueru Zhang, Amanda Giamberardino, Erin McConnell, Tariq Francis, Maria C. DeRosa, and Michel Dumontier. Aptamer base: a collaborative knowledge base to describe aptamers and SELEX experiments. *Database: The Journal of Biological Databases and Curation*, 2012:bas006, February 2012.

[13] UTexas Aptamer Database: the collection and long-term preservation of aptamer sequence information | Nucleic Acids Research | Oxford Academic.

[14] (PDF) AptaDB: A Comprehensive Database Integrating Aptamer–Target Interactions.

[15] Bi-Qing Li, Yu-Chao Zhang, Guo-Hua Huang, Wei-Ren Cui, Ning Zhang, and Yu-Dong Cai. Prediction of Aptamer-Target Interacting Pairs with Pseudo-Amino Acid Composition. *PLOS ONE*, 9(1):e86729, 2014. Publisher: Public Library of Science.

[16] GENA-LM: a family of open-source foundational DNA language models for long sequences | Nucleic Acids Research | Oxford Academic.

[17] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction, October 2020. arXiv:2010.09885 [cs].

[18] Sawan Patel, Keith Fraser, Zhangzhi Peng, Adam D. Friedman, Owen Yao, Pranam Chatterjee, and Sherwood Yao. AptaBLE: A Deep Learning Platform for SELEX Optimization. December 2024.

[19] Zheng Fang, Zhongqi Wu, Xinbo Wu, Shixin Chen, Xing Wang, Saurabh Umrao, and Abhisek Dwivedy. APIPred: An XGBoost-Based Method for Predicting Aptamer–Protein Interactions. *Journal of Chemical Information and Modeling*, 64(7):2290–2301, April 2024. Publisher: American Chemical Society.

A Appendix

A.1 Dataset Descriptions

RSAPred. RSAPred contains 513 aptamer–molecule pairs, dominated by RNA sequences. It includes 139 unique aptamers and 195 unique small molecules, with relatively short sequences on average (33.1 ± 21.7 nt). All entries are annotated with quantitative binding constants, making this dataset a cornerstone for regression tasks. Its molecules are of moderate size (517.3 ± 227.1 Da), and activity labels are well balanced (455 positives vs 58 negatives).

AptamerBase. AptamerBase comprises 356 entries, with a stronger DNA representation. It includes 331 unique aptamers and 61 small molecules, featuring sequences of medium length (54.8 ± 24.1 nt). Out of all records, 164 carry pK_d annotations, making it partially suitable for regression. The ligands tend to be heavier and more variable (766.1 ± 1123.1 Da), and class labels are skewed toward positives (344 vs 12 negatives).

AptaGen. AptaGen is the smallest dataset, with only 44 rows and 43 unique aptamers. Despite its size, it offers dense quantitative annotation: 36 of its 44 entries have pK_d values. The sequences are short (38.3 ± 17.8 nt), while the 33 ligands show high variability in molecular weight (759.3 ± 1276.0 Da). Positives dominate the set (41 vs 3 negatives), making it useful for regression baselines but less diverse for classification.

UTexas. UTexas contains 188 entries with 181 unique aptamers, showing strong sequence diversity. It is moderately balanced between DNA and RNA and has long sequences (79.0 ± 27.0 nt). The dataset includes 63 unique small molecules of intermediate weight (647.6 ± 792.0 Da). A large fraction of entries (139) include quantitative pK_d values. The activity distribution is skewed toward negatives (49 positives vs 139 negatives).

RiboCentre. RiboCentre consists of 113 RNA-only entries, covering 112 unique aptamers and 48 distinct small molecules. Its sequences are among the longest in the benchmark (76.6 ± 36.4 nt), and molecules have moderate size (540.0 ± 549.7 Da). No quantitative binding constants are provided, making the dataset limited to classification. All sequences are labeled as active binders.

AptaDB. AptaDB contains 393 entries and contributes 341 unique aptamers and 126 ligands. DNA dominates this dataset. The sequences are moderately long (57.1 ± 23.1 nt), and molecules are chemically diverse (480.2 ± 664.4 Da). No regression labels are available, and all entries are labeled positive, restricting its use to binary classification or structure-based analysis.

Manual Curation. The manually curated dataset is the largest single source, with 680 rows. It contains 518 unique aptamers and 160 distinct molecules, with shorter average sequence lengths (45.0 ± 20.3 nt). Its molecules are lighter than in other datasets (368.8 ± 221.6 Da). More than half of its entries (373) include pK_d annotations, and it provides a relatively balanced class distribution (600 positives vs 80 negatives).

Table 4: Overview of curated aptamer-small molecule datasets. For each dataset we report: number of rows, DNA:RNA ratio, number of unique aptamers, mean aptamer length with standard deviation (len [nt]), number of unique target molecules, mean target molecular weight with standard deviation (MW [Da]), number of entries with quantitative binding constants (with pKd), and active vs inactive counts (pos:neg).

dataset	rows	DNA:RNA	uniq apt	len [nt]	uniq SM	MW [Da]	with pKd	pos:neg
RSAPred	513	22:491	139	33.1±21.7	195	517.3 ± 227.1	513	455:58
AptamerBase	356	217:139	331	54.8±24.1	61	766.1±1123.1	164	344:12
AptaGen	44	23:21	43	38.3±17.8	33	759.3±1276.0	36	41:3
UTexas	188	120:68	181	79.0±27.0	63	647.6 ± 792.0	139	49:139
Ribocentre	113	0:113	112	76.6±36.4	48	540.0 ± 549.7	0	113:0
AptaDB	393	293:100	341	57.1±23.1	126	480.2 ± 664.4	0	393:0
Manual	680	541:139	518	45.0±20.3	160	368.8 ± 221.6	373	600:80
Combined	2210	1173:1037	1430	49.8±26.9	496	525.1 ± 649.1	1217	1922:288

Combined Dataset. Merging all sources yields 2,210 entries, covering 1,430 unique aptamers and 496 unique ligands. The average sequence length is 49.8 ± 26.9 nt, and molecules span a wide chemical space (525.1 ± 649.1 Da). Overall, 1,217 entries carry regression labels, while binary activity is annotated for 2,210 entries (1,922 positives vs 288 negatives). This combined dataset forms the basis of our benchmark, offering both breadth and diversity across sequence and chemical spaces.

Table 5: Pairwise overlap between datasets. Each cell shows "unique aptamer–SM pairs / unique aptamers / unique SM". Diagonal entries correspond to dataset totals.

	RSAPred	AptamerBase	AptaGen	UTexas	Ribocentre	AptaDB	Manual	Combined
RSAPred	487/139/195	0/0/7	0/1/5	0/1/6	2/4/15	1/2/17	9/2/20	487/139/195
AptamerBase	0/0/7	331/331/61	4/4/11	26/28/24	4/5/12	19/20/27	3/16/11	331/331/61
AptaGen	0/1/5	4/4/11	44/43/33	3/3/12	5/6/8	21/22/23	5/6/11	44/43/33
UTexas	0/1/6	26/28/24	3/3/12	186/181/63	40/44/16	13/13/35	3/9/18	186/181/63
Ribocentre	2/4/15	4/5/12	5/6/8	40/44/16	113/112/48	10/10/23	1/3/11	113/112/48
AptaDB	1/2/17	19/20/27	21/22/23	13/13/35	10/10/23	385/341/126	27/76/21	385/341/126
Manual	9/2/20	3/16/11	5/6/11	3/9/18	1/3/11	27/76/21	646/518/160	646/518/160
Combined	487/139/195	331/331/61	44/43/33	186/181/63	113/112/48	385/341/126	646/518/160	2210/1430/496

A Experimental setup

A.1 Data splits

We implemented three complementary cross-validation protocols: (i) stratified group splits on aptamer–ligand pairs to preserve label proportions while preventing pair leakage, (ii) disjoint-aptamer splits (GroupKFold by sequence identity), (iii) disjoint-molecule splits (GroupKFold by canonical SMILES). Each used 5 folds and a fixed random seed.

A.2 Negative sampling

To mitigate class imbalance, we generated synthetic negatives by cross-pairing each unique aptamer with ligands it was never observed to bind. Each sequence was paired with up to n such ligands per positive pair, excluding known positives or duplicates. Synthetic rows were flagged and labeled as non-binders.

A.3 Feature representations

- Aptamers: k-mer frequency vectors ($k = 4, 6$) with DNA/RNA indicator; one-hot encodings truncated to a fixed length; embeddings from pretrained nucleotide transformer (GENA-LM).
- Ligands: Morgan fingerprints, MACCS keys with physicochemical RDKit descriptors, and ChemBERTa embeddings.

Continuous features were standardized within each fold.

A.4 Model training

We compared multilayer perceptrons and LightGBM classifiers. Feature selection was performed using LightGBM gain, and hyperparameters (e.g., `num_leaves`, learning rate, subsample) were tuned with Optuna on training folds.

A Evaluation metrics

For a binary classifier returning scores s_i for samples with labels $y_i \in \{0, 1\}$:

ROC-AUC. The area under the ROC curve is

$$\mathrm{AUC} = \int_0^1 TPR(FPR^{-1}(x))\,dx,$$

where $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$.

F_1 score. The harmonic mean of precision and recall:

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}.$$

Matthews correlation coefficient (MCC). A balanced measure accounting for all four entries of the confusion matrix:

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

PR-AUC. The area under the precision-recall curve (PR-AUC) summarizes the trade-off between precision and recall across decision thresholds. It is defined as

$$\text{PR-AUC} = \int_0^1 P(R^{-1}(x))\,dx,$$

where $\mathrm{Precision} = \frac{TP}{TP+FP}$ and $\mathrm{Recall} = \frac{TP}{TP+FN}$.

For regression models predicting continuous affinities $\hat{y}_i$ for targets y_i:

Root Mean Squared Error (RMSE).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Mean Absolute Error (MAE).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

Coefficient of Determination (R^2).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Pearson correlation (r_P). The linear correlation between predictions and targets:

$$r_P = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\,\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}.$$

Spearman correlation (r_S). The rank correlation between predictions and targets:

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

where d_i is the difference between the ranks of y_i and $\hat{y}_i$.

All metrics were computed per fold and are reported as mean across cross-validation runs.

A Additional baseline tables

Table 6: MCC (mean $\pm$ std) on the original dataset (no negative augmentation) under grouped CV. Results are reported for three models (LGBM, MLP, RF) across aptamer encodings and molecular descriptors.

LGBM

Aptamer	Morgan FP (1024)	Morgan FP (2048)	MACCS keys	RDKit descriptors	ChemBERTa
kmer(k=3)	0.462±0.035	0.474±0.035	0.379±0.058	0.471±0.054	0.482±0.038
kmer(k=4)	0.469±0.038	0.455±0.039	0.372±0.067	0.479±0.026	0.492±0.048
kmer(k=5)	0.487±0.046	0.480±0.048	0.424±0.051	0.496±0.053	0.507±0.052
onehot(L=216)	0.478±0.039	0.472±0.045	0.438±0.054	0.519±0.021	0.536±0.037
gena(mean,last)	0.417±0.055	0.382±0.041	0.286±0.050	0.388±0.038	0.477±0.035

MLP

Aptamer	Morgan FP (1024)	Morgan FP (2048)	MACCS keys	RDKit descriptors	ChemBERTa
kmer(k=3)	0.401±0.070	0.427±0.073	0.173±0.092	0.156±0.145	0.388±0.036
kmer(k=4)	0.353±0.106	0.273±0.100	0.168±0.044	0.200±0.133	0.261±0.133
kmer(k=5)	0.288±0.107	0.288±0.122	0.216±0.134	0.186±0.050	0.181±0.147
onehot(L=216)	0.421±0.079	0.464±0.045	0.177±0.071	0.129±0.133	0.351±0.039
gena(mean,last)	0.154±0.069	0.428±0.069	0.240±0.082	0.211±0.053	0.295±0.120

Random Forest

Aptamer	Morgan FP (1024)	Morgan FP (2048)	MACCS keys	RDKit descriptors	ChemBERTa
kmer(k=3)	0.471±0.033	0.471±0.034	0.353±0.058	0.470±0.031	0.493±0.033
kmer(k=4)	0.468±0.035	0.468±0.037	0.364±0.049	0.466±0.037	0.488±0.032
kmer(k=5)	0.476±0.032	0.478±0.034	0.378±0.058	0.481±0.029	0.496±0.031
onehot(L=216)	0.471±0.034	0.473±0.034	0.381±0.056	0.482±0.032	0.499±0.030
gena(mean,last)	0.348±0.056	0.348±0.058	0.273±0.043	0.335±0.059	0.402±0.044