AptaBLE: A Deep Learning Platform for de-novo Aptamer Generation and SELEX Optimization

Sawan Patel[1*] **Keith Fraser**[2,3,4*] **Zhangzhi Peng**[5*]

Owen Yao[1] **Adam Friedman**[1] **Pranam Chatterjee**[5,6,7†] **Sherwood Yao**[1†]

[1] Atom Bioworks
[2] Department of Biological Sciences, Rensselaer Polytechnic Institute
[3] Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute
[4] Future of Computing Institute, Rensselaer Polytechnic Institute
[5] Department of Biomedical Engineering, Duke University
[6] Department of Computer Science, Duke University
[7] Department of Biostatistics and Bioinformatics, Duke University

Abstract

Aptamers are single-stranded oligonucleotides that bind molecular targets with high affinity and specificity. However, their discovery and evolution remains constrained to conventional SELEX methods. Here we present an Aptamer Binding LanguagE (AptaBLE) model that addresses this challenge by combining pretrained protein and nucleic acid sequence encoders with a cross-attention architecture to capture the determinants of aptamer-protein binding, enabling robust prediction of binding interactions across diverse protein targets. The model employs a transformer-based architecture with multi-head cross-attention mechanisms, optimizing sequence-specific features and positional embeddings to learn the complex binding patterns between aptamers and their protein targets, while maintaining sequence-length diversity across diverse aptamer libraries. Our extensive evaluations across various benchmarks demonstrates AptaBLE's superiority over existing methods in recapitulating experimental binding profiles. AptaBLE exhibits strong/favorable generalizability to unseen proteins and generated aptamers. In real world applications, AptaBLE identified several experimentally validated CD117 ssDNA aptamers previously missed by traditional SELEX, and generated a novel ssDNA aptamer that shares a comparable binding profile with that of Apt62 to human CD4. These results showcase the ability of AptaBLE to capture molecular interactions that underpin aptamer-protein binding.

1 Introduction

Understanding and predicting molecular interactions between proteins and their binding partners is fundamental to molecular biology and drug discovery. Among the various molecules that can bind proteins, aptamers have emerged as particularly promising therapeutic and diagnostic agents due to their high specificity and affinity for target molecules. Aptamers are single-stranded DNA or RNA oligonucleotides that can form complex tertiary structures enabling them to bind molecular targets with antibody-like specificity but with several key advantages including greater stability under diverse

*Contributed equally.
†Corresponding author

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

conditions, lower immunogenicity, and more cost-effective production. Additionally, aptamers have emerged as promising alternatives to antibodies in analytical, diagnostic and therapeutic applications due to their unique ability to provide spatial complementarity for molecular targets through their tertiary structures.

The conventional approach for discovering aptamers that bind specific protein targets relies on Systematic Evolution of Ligands by EXponential enrichment (SELEX) [26]. The SELEX technique consists of five primary stages: library generation, binding, separation, amplification, and replication. While SELEX has been instrumental in aptamer discovery, it faces significant limitations including extended timeframes often spanning weeks to months per round, with multiple rounds typically required. Additionally, success rates can be modest, yielding only a limited number of candidate aptamers for subsequent characterization. Recent analysis has also revealed that the conventional SELEX approach may miss identifying potentially valuable aptamer sequences due to PCR bias during the amplification stage, where certain sequences can be preferentially amplified or lost due to differential PCR efficiency. Furthermore, analysis of SELEX data from intermediate rounds, not just final rounds, could provide valuable insights into the evolution of binding sequences, but this data is rarely available nor analyzed.

Aptamer design is inherently modifiable, and affinity maturation is a hallmark of aptamer design. This often requires the analysis of various truncations of a full-length aptamer, as well as single- or multi-point mutations, and motif identification. The truncation of an aptamer typically involves the removal of a contiguous subsequence within the full-length aptamer. The primary goal of truncation analysis is find truncants that preserve the full-length aptamer's binding kinetics with minimal sequence length so as to reduce cost of manufacturing. There are many instances where a truncated aptamer has maintained or even improved its binding affinity to its intended target over the full-length aptamer [18, 28, 12].

Another means of maturation concerns the insertion, deletion or substitution of particular bases within an aptamer sequence, often with the objective of introducing stability or inducing a structural motif (e.g. G-quadruplexes). Several groups have demonstrated this capability [32, 19, 3]. It is common for researchers to leverage structural tools to facilitate these analyses and preserve key structural motifs within the full-length aptamer [33], but such tools give no indication towards the functional consequences, e.g., impact on binding kinetics, of any user-imposed modifications.

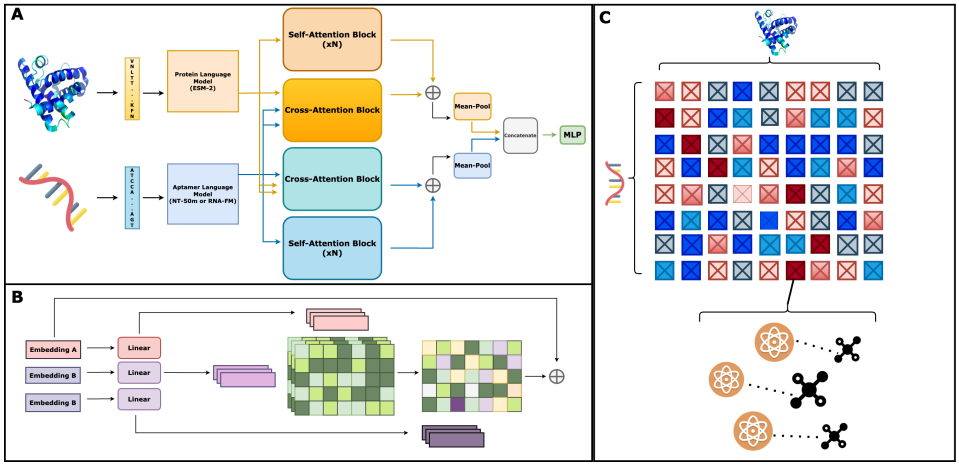

Figure 1: AptaBLE architecture. **A)** Protein and nucleic acids are represented via their constituent primary structure strings and are embedded via pretrained encoders (left). The fusion module combines embeddings into a shared representation, which is reduced into a prediction is made following several linear layers (right). **B)** The cross-attention module. Note that the original embeddings are added back to the cross-attended representation. **C** From our cross-attention formulation, we can extract and analyze attention maps to identify which amino acid residues or nucletide bases most inform the output binding score.

These limitations have motivated the development of computational approaches to predict and optimize aptamer-protein interactions. Recent advances in deep learning have shown promise in this

domain, with models demonstrating remarkable performance even with limited training data [23]. However, as highlighted by [4], many current computational approaches face challenges in generating physically plausible binding poses and struggle to generalize to proteins with low sequence similarity to training data. Early computational approaches focused primarily on sequence-based analysis, such as calculating frequency of appearance or enrichment between consecutive SELEX rounds [11]. However, these methods often overlooked the crucial role of nucleic acid secondary structure in determining binding interactions. More recent approaches have begun incorporating structural information, recognizing that aptamers bind their targets by forming tertiary structures that provide spatial complementarity, though with a limited amount of data [29].

The field of molecular docking faces two key computational challenges: accurately predicting how molecules orient themselves when binding (pose generation) and reliably evaluating the likelihood and strength of these predicted interactions (scoring). Traditional ligand docking methods employ heuristic or exhaustive search algorithms to explore potential ligand conformations [4, 25, 22, 27], but the vast configurational space makes comprehensive sampling computationally intractable, leading to overlooked feasible binding poses. Current scoring functions use simplified approximations of molecular forces to rank potential binding modes, but these approximations often fail to capture the full complexity of molecular recognition [4]. While deep learning methods have attempted to address these limitations, they frequently overlook the crucial fact that interactions occur at the residue-structure level rather than purely through sequence correlations.

Recent innovations have been proposed to address these limitations [5, 8, 17]. However, these methods were primarily developed for small molecule ligands and may not fully capture the unique characteristics of aptamer-protein interactions, including the greater conformational flexibility of nucleic acids and the extensive potential contact surfaces involved in aptamer binding. Further progress has been made in the peptide learning space [20, 21, 7], though many such approaches are not applicable with aptamer learning due to the scarcity of data. Recent structure-based methods show promise [10], though it is debated whether learning from aptamer crystal (or predicted) structures can capture functional capability due to the wide distribution of conformations that nucleic acids can take.

More recently, AptaTrans introduced a transformer-based method to model aptamer and protein sequences at the monomer level, utilizing pretrained encoders for representing monomers [23]. While AptaTrans demonstrated improved performance and efficiency over previous methods, its reliance on local interactions due to the convolution-centric method may not fully capture the complex three-dimensional interactions govern aptamer-protein binding. The model's architecture, which employs separate encoders for proteins and aptamers followed by a dot-product based interaction scoring mechanism, while effective for basic sequence-structure relationships, introduces a bottleneck and may oversimplify the complex spatial relationships involved in aptamer-protein binding. Additionally, the model's training only implicates the use of RNA sequences and an RNA encoder, which point to the need for more flexible, comprehensive approaches.

To address these limitations, we present AptaBLE, a novel deep learning framework that combines pretrained protein and nucleic acid sequence encoders with a cross-attention architecture. Our approach innovates beyond previous methods by explicitly modeling the determinants of aptamer-protein binding while maintaining sequence-length diversity. The transformer-based architecture with multi-head cross-attention mechanisms optimizes sequence-specific features and positional embeddings to learn complex binding patterns between aptamers and their protein targets. This design enables robust prediction of binding interactions across diverse protein targets while maintaining the ability to handle variable-length aptamer sequences.

The model successfully predicts the formation of a complex for both DNA and RNA aptamers with their corresponding protein targets and can generate novel sequences with conditioned binding targets. AptaBLE represents a significant advance in aptamer development, accelerating the discovery and optimization of therapeutic and diagnostic molecules. To democratize access to this technology, we have developed a freely available web application that enables researchers to predict aptamer-protein binding and generate optimized aptamer sequences without requiring specialized computational expertise or infrastructure. This platform represents a significant step toward making advanced aptamer design tools accessible to the broader scientific community: https://huggingface.co/spaces/binderhunt/aptaBLE

2 Results

2.1 Development and Architecture of AptaBLE

AptaBLE employs a symmetric architecture that processes both protein and aptamer sequences through specialized pretrained encoders (ESM2-650m for proteins, nucleotide-transformer-v2-50m or RNA-FM for aptamers), followed by a fusion mechanism for combining embeddings into a shared aptamer-protein representation (Figure 1A). The core of our model consists of symmetric cross-attention modules that enable bidirectional information flow between protein and aptamer representations, capturing complex interaction patterns without providing any structural information (Figure 1B). The fusion architecture comprises two cross-attention modules and two self-attention modules, the outputs of which are summed, pooled, and concatenated to arrive at a joint representation. This is ultimately processed by a multi-layer perceptron to generate binding predictions.

The cross attention module, adapted from [30], allows us to capture both global (long range) and local interactions across tokens. This demonstrates an improvement over prior work, where convolution mechanisms may be limited to capturing interactions within small communities of monomers. For generating an aptamer-attended protein representation, we pass the protein embedding as the queries Q and the aptamer embedding as key and value matrices K, V (Figure 1C). The converse is done to generate a protein-attended aptamer representation. We also process embeddings via self-attention. These self-attended representations are added to the cross-attended representations to generate an attention-enhanced representation of both the protein and aptamer. The aforementioned cross-attended representations yield attention maps connecting amino acids to bases, providing a glance into which interactions most inform the model output.

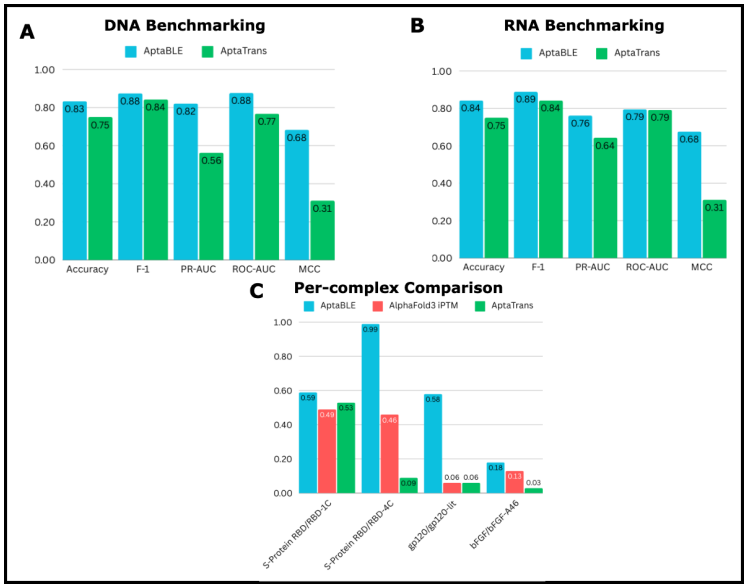

Figure 2: AptaBLE benchmarking. **A** Performance of baseline models across validated binding and non-binding ssDNA aptamer datasets. **B** Performance of baseline models across validated binding and non-binding RNA aptamer datasets. **C** Comparison of key model metrics across four well-studied aptamer-protein systems.

2.2 Benchmark Performance Against Existing Methods

AptaBLE was trained on validated, published aptamers, validated proprietary aptamers, and artificially generated pseudo-negative derivatives of those aptamers. Further details regarding our dataset and implementation can be found in the Supplementary section. We evaluated AptaBLE against existing methods using comprehensive benchmark datasets for both DNA and RNA aptamers (Figure 2A, B). Our model achieved superior performance across multiple metrics, including accuracy. Notably,

AptaBLE outperformed both the prevalent sequence-based method, AptaTrans, and the prevalent structure-based approach, AlphaFold3, in predicting stable aptamer-protein complexes (Figure 2C). Through the recall score (not depicted), we found that AptaBLE is far less susceptible to false positives as AptaTrans, which largely influences our improved metrics performance.

2.3 Attention Maps

To investigate the model's representation of both nucleic acids and amino acids, we studied attention maps generated during inference on two aptamer-protein complexes with known crystal structures. These aptamers, RBD-1C and RBD-4C, bind to the SARS-CoV-2 Spike protein receptor binding domain (RBD) [15]. The spike protein, or S-protein, is a highly-immunogenic glycoprotein which mediates viral entry into host cells by interacting with cell surface receptors. The S1 region contains the receptor-binding domain, which has been targeted before in aptamer discovery studies.

We see that one of our attention heads in the cross attention fusion module representing protein-attended nucleic acids reflects atomic interactions seen at the binding interface between the S-protein and the aptamer (Figure 3A, B). Here, the protein residues are the key and the constituent aptamer nucleic acids are the keys and values to produce the attention map. We did not explore the converse. Notably, we find that each nucleic acid implicated in the binding interface is 'more' attended to in both cases. However, we do not always see that the corresponding amino acid participating in the hydrogen bonding interaction also is relatively attended to within that column.

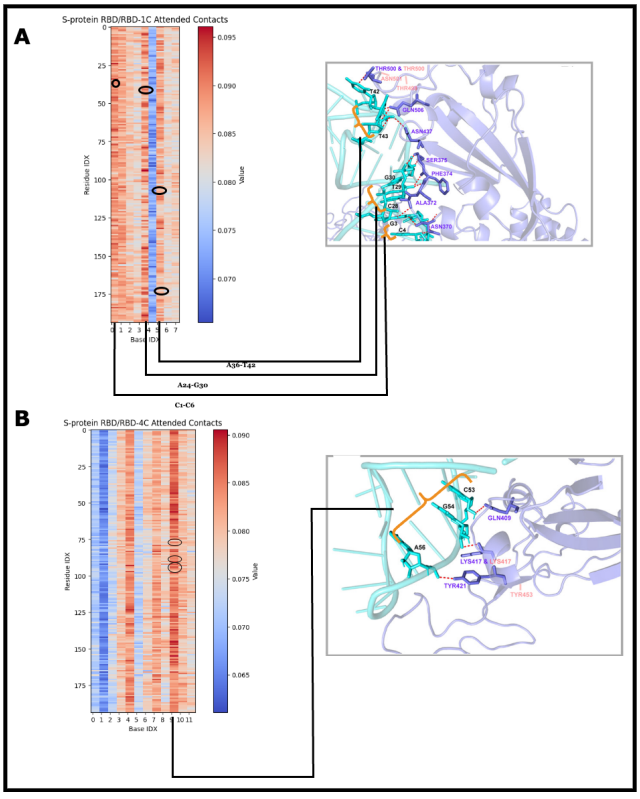

Figure 3: Connecting cross-attention maps to protein-aptamer binding interfaces. **A** A cross attention map from one attention head generated from predicting RBD-1C binding to SaRS CoV-2 S-Protein RBD and the experimentally-determined binding interface. **B** A cross attention map from one attention head generated from predicting RBD-4C binding to SaRS CoV-2 S-Protein RBD and the experimentally-determined binding interface.

We also note that each column in the attention map reflects a disjoint group of 6 nucleic acids, as we follow the nucleotide transformer-50m tokenization scheme.

2.4 CD117 Case Study

Focusing on the DNA aptamers, we investigate performance on a proprietary CD117 aptamer selection (Figure 2). We note here that neither CD117 nor any CD117 aptamers were included in our model training dataset.

2.4.1 CD117 Benchmarking

First, we evaluated model performance on a series of CD117-selective aptamers with known binding affinities. This diagnostic comprises of 13 binding aptamers, 5 validated non-binding aptamers and a series of scrambles of the aptamer with the highest binding affinity. The selection was performed via SELEX and affinities were validated via BLI. We show that the majority of binding aptamers fall above a 0.4 score, with the least-binding aptamer falling below (Figure 4A). Notably, all of the validated non-binding aptamers score highly, as well as a small subset of scrambles. However, this artifact disppears upon fine-tuning our model with a subset of CD117 aptamers (Figure 4B).

We also investigated to what degree these scores correlate with experimentally-determined binding affinities. We see that there is no observable correlation with experimental K_d's determined via Biolayer interferometry (Figure 4C). However, we observe an emergent correlation between and ordering of the AptaBLE scores for each aptamer according to their experimentally-determined Kd's after fine-tuning on a subset of three aptamers within this validation set (Figure 4D).

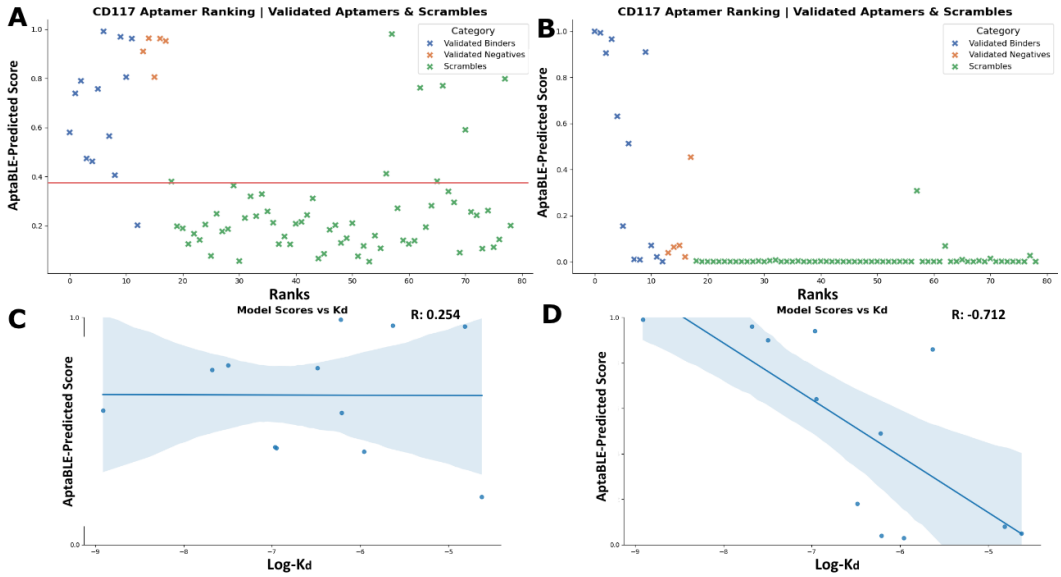

Figure 4: **Benchmarking against CD117 aptamers. A)** A series of aptamers targeting CD117, with K_d's validated via BLI assay, are scored by AptaBLE. Binding aptamers (blue) are ordered by their K_d. A set of non-binding aptamers are also shown (orange) in addition to 60 scrambles of the top-ranked aptamer (green). **B)** After fine-tuning on a subset of the binding aptamers and non-binding aptamers, an apparent ordering by K_d emerges. **C)** A scatterplot depicting the Pearson correlation between original model score from (**A**) and K_d for validated binders. One aptamer is omitted as a K_d was not fit for it. **D)** A scatterplot depicting the Pearson correlation between fine-tuned model score from (**B**) and K_d for validated binders. One aptamer is omitted as a K_d was not fit for it.

2.4.2 CD117 Reverse Analysis

We evolved ssDNA sequences binding to human CD117 through nine rounds of SELEX of increasingly stringent elution conditions. Originally, our work culminated in the identification of a number of clusters in the round 9 population via FastAptamer [1]. The most amplified, or 'fittest,' aptamer within each cluster was then tested in a binding assay to determine a dissociation constant for the aptamer in complex with CD117. This method of selection is quite common in SELEX studies,

where a large sequencing file at the end of some number of SELEX rounds is clustered based on edit distance to identify sub-populations of aptamers that are distinct at the sequence-level. This clustering procedure may, however, result in clusters that are strikingly similar at the structure-level. More alarmingly, the aptamers within the total population that are not clustered are disregarded completely, even if they possess a high number of copies relative to the clustered sequences.

In an effort to comb through the unclustered sequences, we utilized AptaBLE to efficiently parse through a population of unclustered sequences and identify high-affinity binders to CD117 that were previously dismissed due to the clustering methodology. Of the total round nine library, only 536 aptamers were clustered, where the library was clustered into groups all within an edit distance of 3 to the most copied aptamer within the group. 11 clusters were generated in total. This left approximately 35,000 sequences unclustered, of which several had been amplified to a fair degree via PCR in comparison to some of the clustered aptamers. The entire library was scored by AptaBLE (Figure S1).

To initially sift through the unclustered aptamers, we leveraged a combined approach utilizing AptaBLE scores and sequence copy number (Figure S1A). Setting the thresholds for AptaBLE score to 0.4 and a minimum copy number of 10, we identified a sub-population of unclustered aptamers that could be worthy binders specific and selective to CD117. We chose four aptamers in particular, according to their computed stabilities in testing buffer conditions (Figure S1B) and comparable structures to ABW02, a high-affinity CD117 binder according a hierarchical clustering analysis on their secondary structures (Figure S2). These aptamers were tested for binding to human CD117 via Biolayer interferometry, where we found that all four bound (Table 1). These unclustered aptamers, indicated as 'ABWUC' followed by an arbitrary number, showed comparable and even favorable binding kinetics than previously tested aptamers from our original SELEX selection methodology (ABW02-ABW07). Though none of the unclustered aptamers surpassed the K_d of ABW07, two of the apatamers finished in the top 3 overall. Additionally, 3 of the 4 unclustered aptamers we selected had favorable binding kinetics over previously-tested aptamers, indicating that our methodology can identify ssDNA aptamers that are omitted too early from the customary procedure. These aptamers were also structurally distal from the known-binding ABW02-ABW07 aptamer series (Figure S2) [16].

Aptamer	K_d **(M)**	**Clustered**	**Previously tested**
ABWUC1	1.99E-08	No	No
ABWUC2	3.26E-08	No	No
ABWUC3	ND	No	No
AGWUC4	6.02E-07	No	No
ABW02	2.10E-08	Yes	Yes
ABW03	6.18E-07	Yes	Yes
ABW04	1.10E-06	Yes	Yes
ABW05	3.29E-07	Yes	Yes
ABW06	1.53E-05	Yes	Yes
ABW07	1.22E-09	Yes	Yes

Table 1: **Aptamers identified from CD117 SELEX Round 9 pool by AptaBLE in comparison to pre-existing standard.** Note that we identified several aptamers that had been omitted from curve-fitting with the pre-existing method of analysis. All four of these aptamers, indicated by ABWUC* (in **bold**), bind with relatively high affinity in comparison to those which were progressed. In particular, the aptamers ABWUC1 and ABWUC2 were found to have among the best binding affinities to human CD117 in comparison to aptamers originally tested.

From the clustered aptamers, we are also interested in determining if the strategy of selecting the top-amplified aptamer within each cluster for binding is the most appropriate strategy for identifying aptamers with the strongest affinities. When using AptaBLE, the aptamer with the highest copy number is not always scored the highest within its cluster (Figure S3A). The overall distributions of scores for each aptamer within a cluster also varies widely across clusters (Figure S3B). Further investigation into the established leading clusters, such as cluster 2, 5 and 7, will be critical for elucidating how trustworthy the conventional raw copy number metric is for selecting a constituent aptamer from each cluster for binding analysis. Additionally, this work can also indicate which point

mutations within the leading aptamer can diminish binding affinity significantly or even destabilize the aptamer structure entirely.

2.5 Generation of Novel Aptamers

AptaBLE extends beyond prediction to enable target-specific aptamer generation. Using a classifier-guided optimization method implemented via monte carlo tree search [13], we generated novel aptamers targeting specific protein binding sites. This method navigates the exponentially large search space of possible nucleic acid sequences to output an aptamer predicted to bind to a particular protein target via iterative optimization (Figure 5A). While we are capable of generating both ssDNA and RNA aptamers via this method, we only focus on ssDNA aptamer generation here. We show several examples of ssDNAs generated by AptaBLE-MCTS predicted to bind to a variety of targets (Figure 5B), all of length 30.

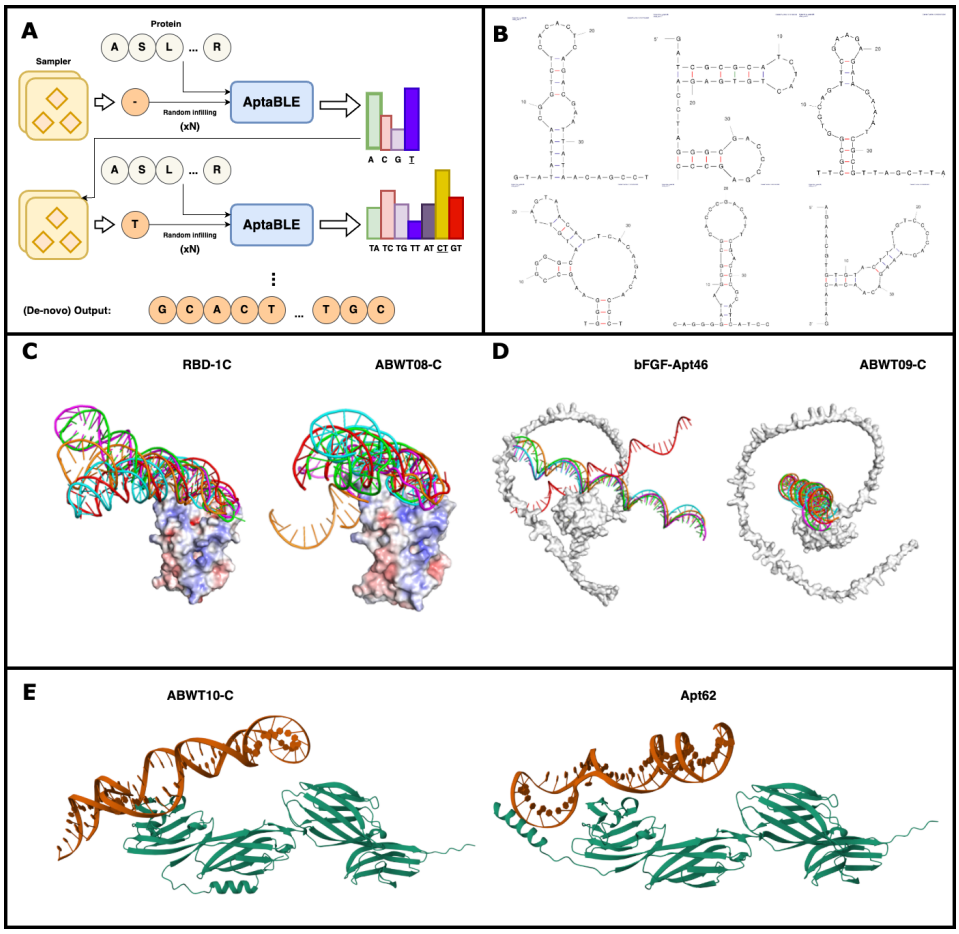

Figure 5: AptaBLE *de-novo* generation of protein target-binding aptamers. **A** An overview of the monte carlo tree search algorithm used for classifier-guided *de-novo*, conditional generation of binding aptamers. **B** Secondary structures of generated aptamers for a variety of targets predicted by mFold. **C** Docking of two AptaBLE-generated ssDNA aptamers in comparison to a validated binding aptamer. (Left) The predicted complexes of generated aptamer ABWT08-C and validated nanomolar-affinity aptamer RBD-1C in complex with the SaRS CoV-2 S-protein receptor binding domain, respectively. (Right) The predicted complexes of generated aptamer ABWT09-C and validated nanomolar-affinity aptamer bFGF-Apt46 with the bFGF protein, respectively.

We further tested this heuristic's generative capability by generating aptamers targeting the SARS-CoV-2 spike protein and basic fibroblast growth factor (bFGF), demonstrating comparable or improved binding properties compared to published sequences. bFGF is a signaling protein that

participates in a large variety of mitogenic and cell survival activities.

SELEX procedures have been conducted for both proteins, from which specifically-binding, selective, high-affinity aptamers have been discovered. We compare our generated aptamers to such published aptamers here, namely RBD-1C (K_d: 10 μ M) and bFGF-46 (K_d: 46 nM) (Figure 5C, D) [15, 9]. Our generated aptamer ABWT08-C showed binding to the same target site as the established RBD-1C aptamer while providing additional surface area coverage, suggesting potential improvements in binding efficiency.

Lastly, we also sought to improve an existing aptamer binding selectively to human CD4 (hCD4). This hCD4 aptamer, Apt62, has a reported binding affinity of 1.59 nm [31]. Via the aforementioned monte carlo tree search method, we generated a library of 1000 ssDNA aptamers predicted to bind to hCD4. We aimed to identify an aptamer with a similar structural profile as Apt62, the published hCD4 aptamer, and achieve comparable binding. Following an analysis implicating secondary structure clustering and docking, we were able to identify a candidate that met this criteria: ABWT10-C (Figure 5E). The predicted structures are generated via Chai-1 [6].

Via biolayer interferometry, we performed a head-to-head comparison of our aptamer, ABWT10-C, against Apt62. Though our preliminary kinetic assays were up against detection limits, we found that our aptamer had comparable affinity to Apt62, with no off-target binding to CD14 as a negative control (data not shown). We aim to reconfirm these results in the future.

2.6 Workflows

While still actively in development, AptaBLE is already capable of accelerating the aptamer discovery process. As explored above, SELEX experiments designed for aptamer discovery are time-consuming, expensive, and labor-intensive. In most cases, high-affinity aptamer discovery requires upwards of ten rounds of SELEX. Though next-gen sequencing has made throughput an issue of the past, it remains expensive and a critical rate-limiting step in the discovery procedure. Rather than requiring ten rounds of SELEX for identifying a target-binding library of aptamers that are not even guaranteed to be selective for the intended target, AptaBLE (in combination with other in-silico tools) could reduce the time to discovery by half, requiring half as many rounds of SELEX as typically required. An example of this workflow is illustrated in Figure 6A. Rather than repeat ten rounds of SELEX with increasingly stringent elution criteria, we propose the number of rounds to be cut in half. The library remaining after selection can then be tested with AptaBLE to deduce which fraction of the library is not only likely to bind to the target, but also less susceptible to off-target binding. Off-target binding is not typically controlled for during SELEX, making the use of AptaBLE particularly appealing.

Furthermore, AptaBLE can aid in selecting which aptamers will be tested in a binding assay. Typically, the final library following SELEX is clustered by sequence homology for some fixed edit distance. Within each cluster, the sequence with the largest number of reads is often progressed to a binding assay such as surface plasmon resonance (SPR) or biolayer interferometry (BLI). However, this methodology is susceptible to error as certain oligonucleotides may be more amplifiable via polymerase chain reaction (PCR). Oligonucleotide amplification occurs during every round of SELEX, which can encourage certain PCR-friendly oligonucleotides to persist for the entirety of selection and claim a large number of copies within the final pool. As such, oligonucleotides that bind more effectively to the intended target may be disregarded as a result of this greedy selection procedure. AptaBLE can control for this PCR-induced bias by verifying that each sequence progressed to a binding assay will bind, at the very least.

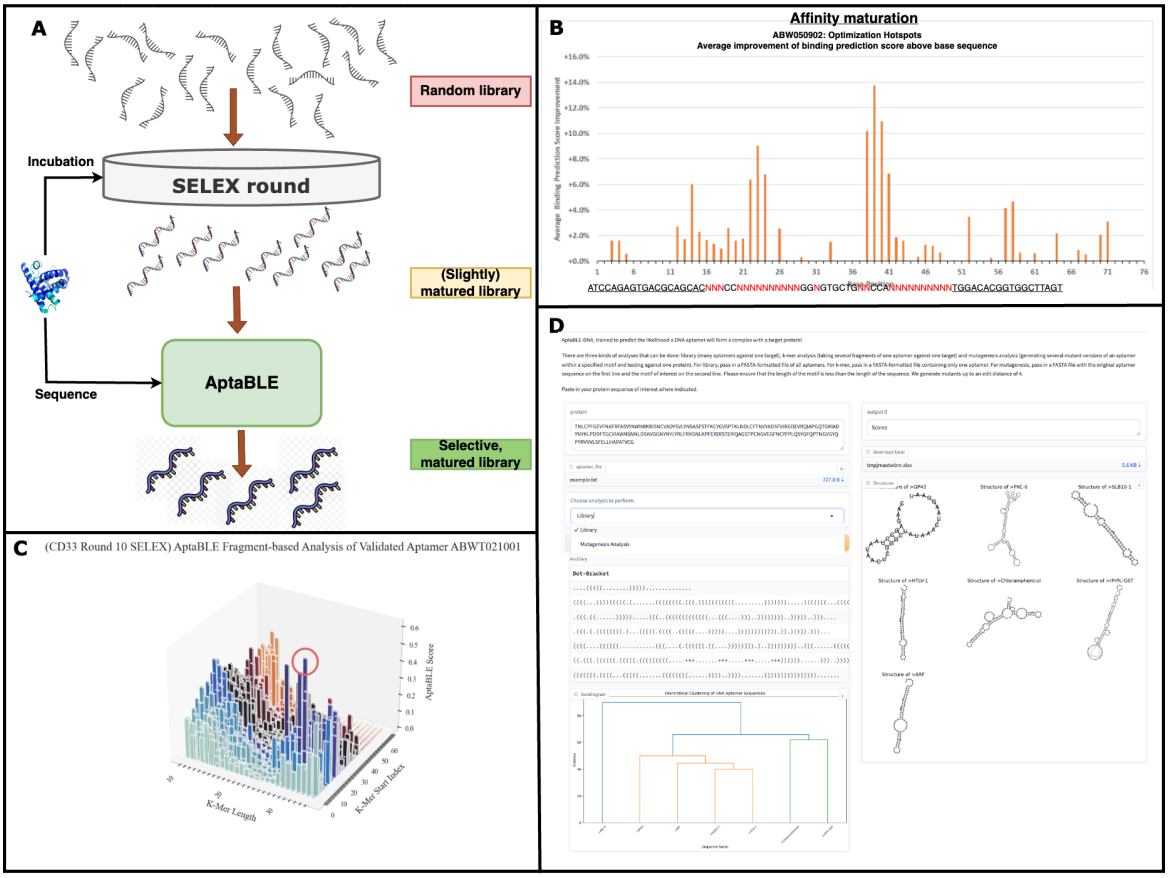

Figure 6: **Prospective workflows involving AptaBLE. A)** A condensed SELEX discovery procedure of novel target-specific, selective aptamers facilitated by AptaBLE. **B** K-mer truncation analysis of a known, binding aptamer for affinity optimization. **C** Mutagenesis of a known, binding aptamer for affinity optimization. **D** A graphical user interface serving AptaBLE and supplementary predictions to aptamer researchers, now available on Huggingface Spaces.

We also envision additional workflows to be made possible by AptaBLE in the near future. These pertain to aptamer affinity maturation, the optimization of an aptamer's design across a variety of metrics including binding affinity, binding specificity, cost to manufacture, size, etc. One such methodology is mutagenesis, which implicates the substitution of bases in the primary structure of an existing aptamer for improving its binding kinetics (Figure 6B). Additionally, k-mer analysis describes the procedure by which the shortest truncant or subsequence is identified that maintains (or in some cases, improves) the binding kinetics of the original aptamer (Figure 6C). These workflows have not yet been validated with AptaBLE, but we believe further optimization to our training data can yield precision at the single-base level.

We include a schematic of a graphical user interface which aptamer researchers can use to design and optimize their aptamers (Figure 6D).

3 Discussion

AptaBLE represents a significant advance in computational aptamer development, addressing key limitations of both traditional SELEX methods and existing computational approaches. By operating directly on sequence data, our model circumvents the scarcity of structural information while capturing the complex interaction patterns critical for aptamer-protein binding. The model's ability to identify overlooked high-affinity binders in SELEX data demonstrates its potential to enhance current aptamer

discovery workflows. The successful validation of AptaBLE's predictions through experimental binding assays, particularly in the hCD4 study and CD117 case study, highlights its practical utility in real-world applications. Our framework's capability to generate novel aptamer sequences with desired binding properties opens new possibilities for rational aptamer design, potentially reducing the time and resources required for aptamer development. We acknowledge certain limitations in our current implementation. The performance disparity between DNA and RNA aptamer prediction suggests room for improvement in RNA-specific modeling, particularly for modified nucleotides. Additionally, while our sequence-based approach has proven effective, integration with structural prediction tools could potentially enhance design capabilities further. Current challenges in RNA model performance may be attributed to existing RNA encoders not incorporating chemically-modified ribonucleic acids, which many RNA aptamers contain. Similarly, while we have not observed any drastic issues to this point, the need for an encoder trained specifically on ssDNA rather than genomic DNA could emerge as structures exhibited by both are not identical. Looking forward, we aim to elucidate the key interactions at the token level which contribute to the overall binding score by analyzing attention maps, in addition to refining how to interpret our model score. We believe that addressing these observations could bridge the gap between the wide-scale adoption of our model in aptamer design and generation. AptaBLE has the potential to significantly accelerate the development of aptamer-based therapeutics and diagnostics by streamlining the discovery process and enabling more rational design approaches.

References

[1] Khalid K Alam, Jonathan L Chang, and Donald H Burke. "FASTAptamer: a bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections". In: *Molecular Therapy-Nucleic Acids* 4 (2015).

[2] Ali Askari et al. "UTexas Aptamer Database: the collection and long-term preservation of aptamer sequence information". In: *Nucleic Acids Research* 52.D1 (2024), pp. D351–D359.

[3] Timothy L Bullock, Luke D Sherlin, and John J Perona. "Tertiary core rearrangements in a tight binding transfer RNA aptamer". In: *Nature structural biology* 7.6 (2000), pp. 497–504.

[4] Duanhua Cao et al. "Generic protein–ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling". In: *Nature Machine Intelligence* (2024), pp. 1–13.

[5] Duanhua Cao et al. "SurfDock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction". In: *Nature Methods* (2024), pp. 1–13.

[6] Chai Discovery. "Chai-1: Decoding the molecular interactions of life". In: *bioRxiv* (2024). DOI: 10.1101/2024.10.10.615955. eprint: https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955.full.pdf. URL: https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955.

[7] Tianlai Chen et al. "Pepmlm: Target sequence-conditioned generation of peptide binders via masked language modeling". In: *ArXiv* (2023).

[8] Gabriele Corso et al. "Diffdock: Diffusion steps, twists, and turns for molecular docking". In: *arXiv preprint arXiv:2210.01776* (2022).

[9] Akihiro Eguchi et al. "A DNA aptamer that inhibits the aberrant signaling of fibroblast growth factor receptor in cancer cells". In: *Jacs Au* 1.5 (2021), pp. 578–585.

[10] Tinglin Huang et al. "Protein-Nucleic Acid Complex Modeling with Frame Averaging Transformer". In: *arXiv preprint arXiv:2406.09586* (2024).

[11] Ryoga Ishida et al. "RaptRanker: in silico RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information". In: *Nucleic acids research* 48.14 (2020), e82–e82.

[12] Harleen Kaur and Lin-Yue Lanry Yung. "Probing high affinity sequences of DNA aptamer against VEGF165". In: *PloS one* 7.2 (2012), e31196.

[13] Gwangho Lee et al. "Predicting aptamer sequences that interact with target proteins using an aptamer-protein interaction classifier and a Monte Carlo tree search approach". In: *PloS one* 16.6 (2021), e0253760.

[14] Bi-Qing Li et al. "Prediction of aptamer-target interacting pairs with pseudo-amino acid composition". In: *PLoS One* 9.1 (2014), e86729.

[15] Yu-Chao Lin et al. “In-Silico Selection of Aptamer Targeting SARS-CoV-2 Spike Protein”. In: *International Journal of Molecular Sciences* 23.10 (2022), p. 5810.

[16] Ronny Lorenz et al. “ViennaRNA Package 2.0”. In: *Algorithms for molecular biology* 6 (2011), pp. 1–14.

[17] Wei Lu et al. “Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction”. In: *Advances in neural information processing systems* 35 (2022), pp. 7236–7249.

[18] Joanna Macdonald et al. “Truncation and mutation of a transferrin receptor aptamer enhances binding affinity”. In: *Nucleic acid therapeutics* 26.6 (2016), pp. 348–354.

[19] Yoshihiko Nonaka et al. “Affinity improvement of a VEGF aptamer by in silico maturation for a sensitive VEGF-detection system”. In: *Analytical chemistry* 85.2 (2013), pp. 1132–1137.

[20] Zhangzhi Peng, Benjamin Schussheim, and Pranam Chatterjee. “PTM-Mamba: A PTM-aware protein language model with bidirectional gated Mamba blocks”. In: *bioRxiv* (2024).

[21] Zhangzhi Peng et al. “Generative diffusion models for antibody design, docking, and optimization”. In: *bioRxiv* (2023), pp. 2023–09.

[22] Matthew P Repasky, Mee Shelley, and Richard A Friesner. “Flexible ligand docking with Glide”. In: *Current protocols in bioinformatics* 18.1 (2007), pp. 8–12.

[23] Incheol Shin et al. “AptaTrans: a deep neural network for predicting aptamer-protein interaction using pretrained encoders”. In: *BMC bioinformatics* 24.1 (2023), p. 447.

[24] Martin Steinegger and Johannes Söding. “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets”. In: *Nature biotechnology* 35.11 (2017), pp. 1026–1028.

[25] Oleg Trott and Arthur J Olson. “AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading”. In: *Journal of computational chemistry* 31.2 (2010), pp. 455–461.

[26] Craig Tuerk and Larry Gold. “Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase”. In: *science* 249.4968 (1990), pp. 505–510.

[27] Marcel L Verdonk et al. “Improved protein–ligand docking using GOLD”. In: *Proteins: Structure, Function, and Bioinformatics* 52.4 (2003), pp. 609–623.

[28] Cong Quang Vu et al. “Truncation of PDGF-BB aptamer by secondary structural analysis and immunoassay”. In: *International Journal of Pharma Medicine and Biological Sciences* 5.1 (2016), p. 86.

[29] Felix Wong et al. “Deep generative design of RNA aptamers using structural predictions”. In: *Nature Computational Science* (2024), pp. 1–11.

[30] Minghao Xu et al. “Protst: Multi-modality learning of protein sequences and biomedical texts”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 38749–38767.

[31] Nianxi Zhao et al. “Blocking interaction of viral gp120 and CD4-expressing T cells by single-stranded DNA aptamers”. In: *The international journal of biochemistry & cell biology* 51 (2014), pp. 10–18.

[32] X Zheng et al. “A saxitoxin-binding aptamer with higher affinity and inhibitory activity optimized by rational site-directed mutagenesis and truncation”. In: *Toxicon* 101 (2015), pp. 41–47.

[33] Michael Zuker. “Mfold web server for nucleic acid folding and hybridization prediction”. In: *Nucleic acids research* 31.13 (2003), pp. 3406–3415.

4 Supplementary

4.1 Datasets

We combine the training data from [2] and [14] in addition to proprietary aptamer binding data to form our ssDNA and RNA aptamer datasets. From [2] and our proprietary data, we take all aptamers with sub-micromolar K_d's as positive examples and all others as negatives. Additionally, we augment our training daataset by creating artificial negative examples through shuffling each positive aptamer's sequence to distort its secondary structure and pairing it with the same protein. We generate 4 artificial negatives per positive sample in this manner. Train-validation splits are performed via sequence homology clustering using the default 0.8 minimum sequence identity and 0.8 sequence alignment coverage [24].

Our benchmark datasets were constructed separately as largely non-overlapping with our training and validation datasets at both the aptamer and protein levels. Both the benchmark aptamesr and corresponding protein targets are not shown during training, though we do not explicitly control for sequence homology otherwise.

4.2 Environment

We train and perform our experiments on a single a2-ultra-gpu machine type available via the Google Cloud Platform. This machine has 170 GB RAM across 12 vCPUs, an Intel Cascade Lake processor, and an NVIDIA A100 80gb GPU. We implement our model using Python 3.9 and PyTorch 2.3.0.

However, we note that our model could be trained on any compute platform supporting a GPU with at least 48gb memory.

4.3 Training Details

We use the MultiStepLR Scheduler with an initial learning rate of $1e-5$ which decays by 0.1 every 5 timesteps. We use AdamW as our optimizer and binary cross-entropy as our loss function, implemented by PyTorch. Our batch size is fixed to 16 across both versions of our model. Training time takes approximately 4 hours until convergence, corresponding to 20 epochs. We performed a hyperparameter sweep over the initial learning rate, batch size and random seed for model configuration. We test several checkpoints across several trained models, of which we report the best-performing results here as determined purely through benchmark dataset performance.

We reproduced AptaTrans and trained their model as described in their methods, choosing the best-performing checkpoint for comparison as done for AptaBLE. We trained their model using their own dataset, as described in their method. Benchmark comparisons relied on converting DNA aptamers to their RNA homolog re AptaTrans inference by substituting thymine (T) for uracil (U), while a one-to-one comparison was done for the RNA aptamer inference.

4.4 Hyperparameters

We list our hyperparameters here.

1. **Cross-attention module**: The hidden size and number of heads are 512 and 8..
2. **Self-attention module**: The hidden size and number of heads are 512 and 8.
3. **Reshape layer**: We reshape ESM2 embeddings to have a hidden dim of either 512 or 640 from 1280 for compatability with Nucleotide-Transformer or RNA-FM, respectively.
4. **MLP**: We pass the fused representation through an MLP with linear layers mapping from either 1024 or 1280 (depending on DNA or RNA, respectively), to 512, to 256, to 1. Each layer is followed by a ReLU activation and 1-D Batchnorm.

4.5 Formalism

We take a vectorized protein $P \in \mathbb{W}^m$ of length m and aptamer $A \in \mathbb{W}^n$ of length n and map the pair to a score between 0 and 1.

$$(P, A) \rightarrow [0, 1] \tag{1}$$

To do this, we leverage two pretrained encoders to produce embeddings of the protein $P \rightarrow \hat{P}$ and the aptamer $A \rightarrow \hat{A}$. These are subsequently fused via a cross attention fusion module to generate a combined representation of the pair, U. For each cross attention fusion module, we learn three weight matrices: $\mathbb{W}_q, \mathbb{W}_k, \mathbb{W}_v \in \mathbb{R}^{hxh}$ which are used to generate the queries, keys and values, or $\mathbf{Q}, \mathbf{K}, \mathbf{V}$, for the module. We therefore generate the following matrices:

$$Q_p = \hat{P}W_q^p, K_p = \hat{P}W_k^p, V_p = \hat{P}W_v^p. \tag{2}$$

$$Q_a = \hat{A}W_a^p, K_a = \hat{A}W_k^a, V_a = \hat{A}W_v^a. \tag{3}$$

We also learn another set of weight matrices for the self-attention module:

$$Q_p^* = \hat{P}W_{q^*}^p, K_p^* = \hat{P}W_{k^*}^p, V_p^* = \hat{P}W_{v^*}^p. \tag{4}$$

$$Q_a^* = \hat{A}W_{q^*}^a, K_a^* = \hat{A}W_{k^*}^a, V_a^* = \hat{A}W_{v^*}^a. \tag{5}$$

From these matrices, we generate the fused representations as below:

$$P^* = \frac{1}{2}\left[MHA(Q_p, K_p, V_p) + MHA(Q_p, K_a, V_a)\right] + \hat{P} \tag{6}$$

$$A^* = \frac{1}{2}\left[MHA(Q_a, K_a, V_a) + MHA(Q_a, K_p, V_p)\right] + \hat{A} \tag{7}$$

To combine $P^* \in \mathbb{R}^{L_p,h}$ and $A^* \in \mathbb{R}^{L_a,h}$, we mean-pool each matrix and concatenate the result to yield a vector $V_0 \in \mathbb{R}^{2h}$. This vector is finally downsized by a series of linear-ReLU-batchnorm modules until a prediction score is produced via a sigmoid activation.

$$FF_i(V_i) = BatchNorm(ReLU(W_i V_i)) \text{ for i in [0, 3].} \tag{8}$$

where each $W_i \in \mathbb{R}^{h/i,h/2i}$.

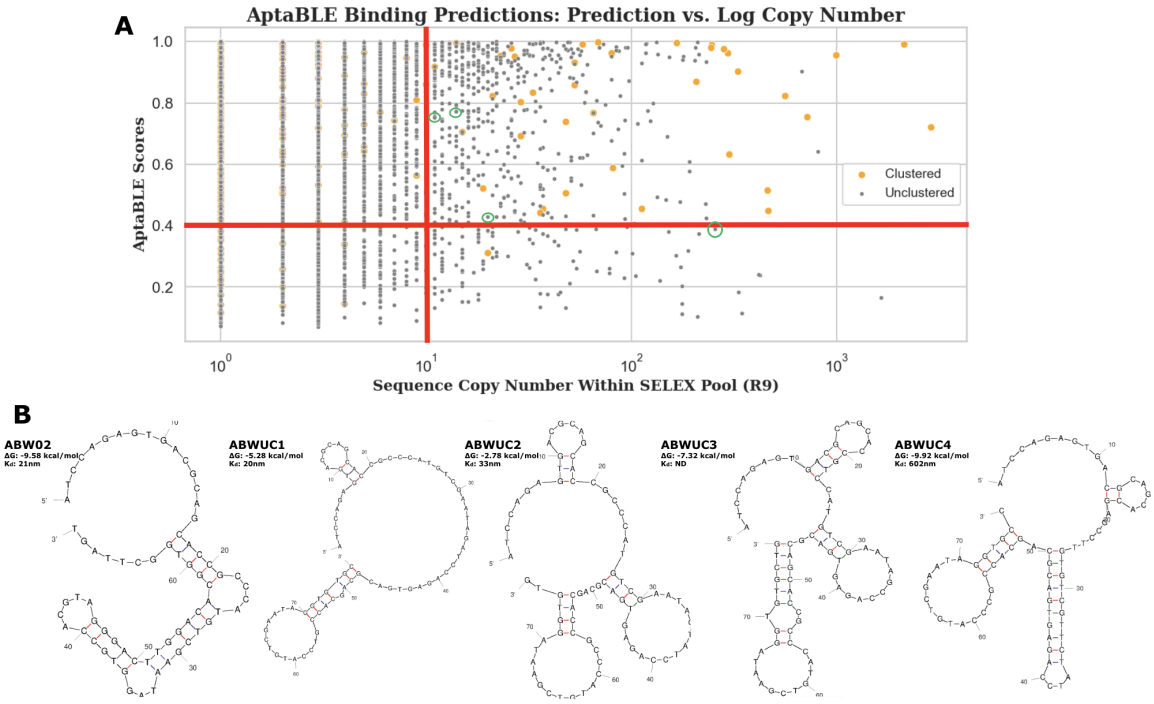

Figure S1: **ssDNA Aptamers sequenced following nine rounds of SELEX selection for targeting CD117. A)** CD117 aptamers from round 9 colored according to whether they were clustered or not. Clustering was sequence-based and performed via FastAptamer with a maximum edit distance set to 3. Red lines indicate selection thresholds used for identifying initial selection of unclustered aptamers for testing, with minimum AptaBLE score set to 0.4 and minimum copy number set to 10 reads. Encircled unclustered aptamers were chosen for binding experiments. **B)** Secondary structures for ABW02, a validated CD117 ssDNA aptamer with 20 nm K_d, and the four unclustered aptamers. Adjacent to each structure is the name of the aptamer and its computed free energy score from mFold [33].

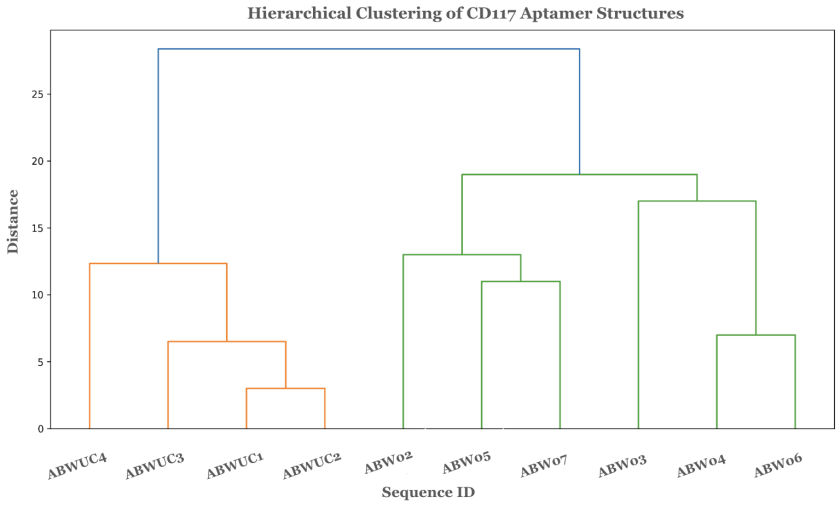

Figure S2: **Hierarchical clustering of all tested aptamers according to their dot-bracket representations.** The 4 unclustered aptamers advanced for binding experiments are labeled as 'ABWUC*'.

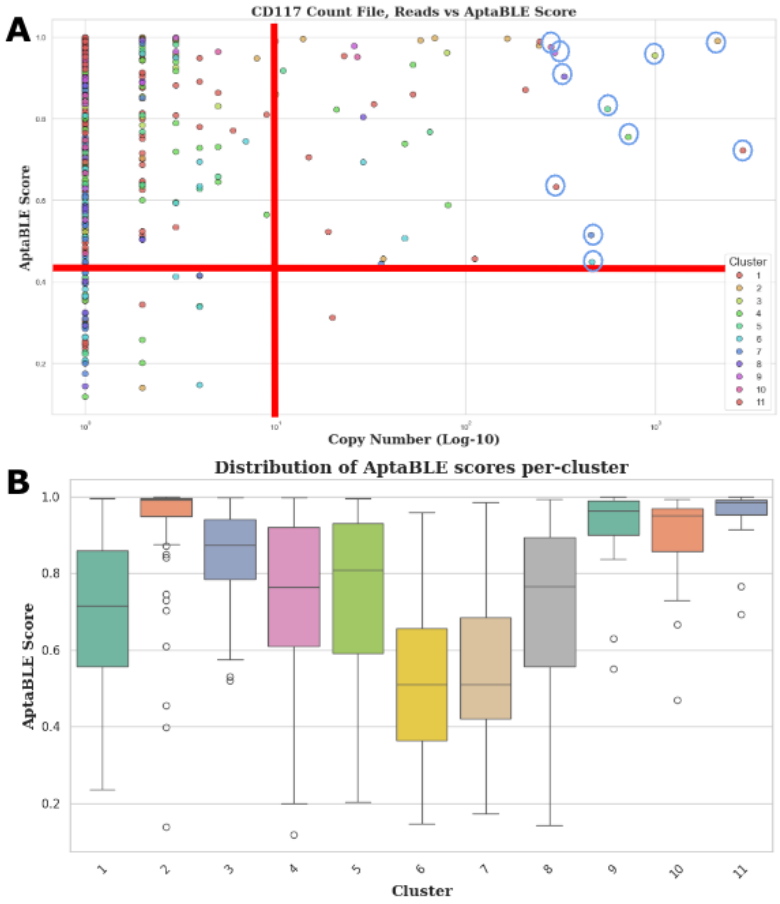

Figure S3: **ssDNA aptamers clustered by FastAptamer. A)** Only clustered aptamers from FastAptamer are shown, colored by their cluster. Circled aptamers were the most frequently observed within their respective cluster, according to the number of copies. **B)** Distribution of AptaBLE scores for each cluster.

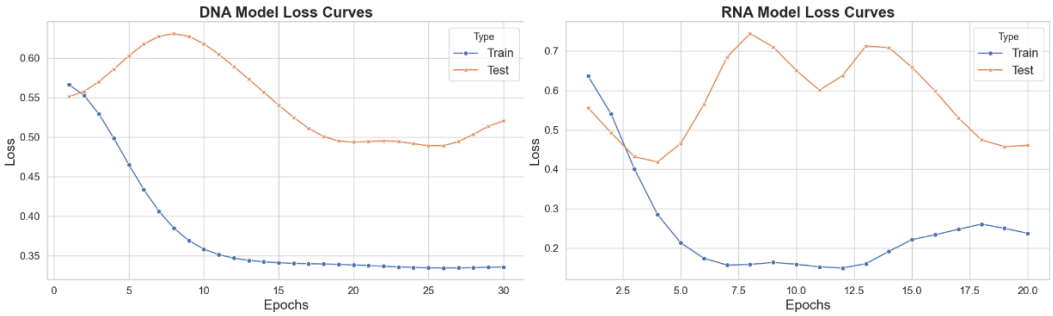

Figure S4: **Model loss curves.**