AptaBLE: A Deep Learning Platform for Aptamer Generation and Analysis

Sawan Patel,[1] Fred Zhangzhi Peng,[2] Keith Fraser,[3-5] Dhanush Gandavadi,[6] Abhisek Dwivedy,[6,7] Xing Wang,[6-10] Pranam Chatterjee,[11,12,†] Sherwood Yao[1,†]

1. Atom Bioworks, Cary, NC, USA
2. Department of Biomedical Engineering, Duke University, Durham, NC, USA
3. Department of Biological Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA
4. Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, USA
5. Future of Computing Institute, Rensselaer Polytechnic Institute, Troy, NY, USA
6. Department of Bioengineering, Grainger College of Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA
7. Holonayak Micro & Nanotechnology Lab, Grainger College of Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA
8. Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana−Champaign, Urbana, Illinois, USA
9. Department of Chemistry, University of Illinois at Urbana−Champaign, Urbana, Illinois, USA
10. Cancer Center at Illinois, University of Illinois at Urbana−Champaign, Urbana, Illinois, USA
11. Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA
12. Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA

[†]Corresponding authors: pranam@seas.upenn.edu, s.yao@atombioworks.com

Abstract

Aptamers are single-stranded oligonucleotides that bind molecular targets with high affinity and specificity, offering significant advantages over antibodies in therapeutic and diagnostic applications. However, their discovery through Systematic Evolution of Ligands by EXponential enrichment (SELEX) remains time-consuming, expensive, and susceptible to experimental biases. Here we present AptaBLE, a deep learning framework that combines pretrained protein and nucleic acid sequence encoders with a novel symmetric bidirectional cross-attention architecture to predict aptamer-protein binding interactions. This design enables robust generalization across diverse protein targets and ssDNA aptamer modalities while maintaining the ability to process variable-length sequences. We demonstrate that AptaBLE significantly outperforms existing methods, achieving superior accuracy in predicting aptamer-protein binding. Furthermore, we demonstrate two complementary de novo generation approaches that produce novel aptamers binding target proteins with desired specificity profiles and K_d's as low as 31nM to-date. AptaBLE represents a significant advance in computational aptamer design, providing an accessible, sequence-based platform for accelerating therapeutic and diagnostic aptamer development.

Introduction

Among therapeutic modalities, aptamers—single-stranded DNA or RNA oligonucleotides that fold into complex three-dimensional structures—have emerged as promising alternatives to antibodies with several key advantages. These include enhanced thermal and chemical stability, lower immunogenicity, cost-effective chemical synthesis, and the ability to target epitopes inaccessible to antibodies[1,2]. Despite these advantages, aptamer discovery remains constrained by the limitations of Systematic Evolution of Ligands by EXponential enrichment (SELEX)[3].

The SELEX process involves iterative rounds of selection, amplification, and enrichment from randomized oligonucleotide libraries containing between 10^{14} and 10^{15} unique sequences[3]. Each round requires careful optimization of binding, washing, and elution conditions, with typical campaigns requiring 10-15 rounds over several months. Furthermore, the polymerase chain reaction (PCR) amplification step in each round introduces sequence bias, where certain sequences amplify more efficiently regardless of their binding affinity, potentially causing high-affinity binders to be depleted from the pool[4,5]. Final selection of candidate aptamers typically relies on clustering by sequence homology and selecting the most abundant sequence from each cluster—a strategy that may overlook structurally distinct high-affinity binders present at lower copy numbers due to PCR bias[6,7].

These limitations have motivated the development of computational approaches to predict and design aptamer-protein interactions[8,9], but these approaches remain susceptible to PCR bias and provide limited mechanistic insight. Recent advances in deep learning offer new opportunities for aptamer design[13-15]. However, existing computational methods for aptamer-protein binding prediction face several limitations[16], such as a limited capacity for capturing long-range interactions and dataset contamination by training on exclusively one mode of nucleic acid. Structure-based deep learning methods[17-19], including AlphaFold3[20], require significant computational resources that limit its utility for large-scale screening applications. Furthermore, the scarcity of solved aptamer-protein structures, and the inherent flexibility of nucleic acid polymers, creates biases in structure-based training approaches that are reflected by comparably poor historical performance on any nucleic acid modality[20,21].

Here we present **AptaBLE** (Aptamer Binding LanguagE; **Figure 1A**), a deep learning framework capable of both predicting the binary interaction of aptamer-protein complexes and generating high-affinity aptamers *de novo*. AptaBLE is implemented as an accessible open-source platform with seamless integration into aptamer discovery workflows. By combining robust prediction with generative capabilities, AptaBLE provides a comprehensive platform for aptamer discovery and optimization that could substantially reduce the time and cost of developing aptamer-based therapeutics and diagnostics.

Results

AptaBLE retrospectively identifies lower-affinity aptamers to CD117 following SELEX

Focusing on the DNA aptamers, we investigate performance on retrospective analysis on a human CD117 (hCD117) aptamer SELEX library[25]. We note here that neither hCD117 nor any hCD117 aptamers were included in our model training dataset. Details regarding data collection are available in the Supplementary information.

Score Benchmarking

We compare metrics on the benchmark dataset using both AptaBLE and AptaTrans, showing that AptaBLE performs reliably better than AptaTrans on both DNA (**Figure S1A**) and RNA (**Figure S1B**) datasets. This is particularly seen when considering false negatives, i.e. correctly predicting when an aptamer does *not* bind to a protein. Additionally, we also highlight scores on the auxiliary benchmark dataset for four known aptamer-protein systems in which the aptamer is a low nanomolar binder to the paired protein target using AptaBLE, AptaTrans, and AlphaFold3 (**Figure S1C**).

Contact Prediction

Investigating the *SARS-CoV-2* Spike glycoprotein binding aptamers further, we ambitiously investigate if the cross-attention maps produced during a forward pass from AptaBLE recapitulate their known binding interfaces. RBD-1C makes contacts with the *SARS-CoV-2* Spike glycoprotein at three separate interfaces, all of which are slightly more active than the neighboring base regions (**Figure S2A**). However, we recognize that many amino acids on the Spike glycoprotein not implicated in the crystallized binding interface are predicted to participate by AptaBLE. The same is true for RBD-4C (**Figure S2B**).

CD117 SELEX Retrospective Analysis

First, we evaluated model performance on a series of hCD117-selective aptamers with known binding affinities. This diagnostic comprises 13 binding aptamers, 5 validated non-binding aptamers and a series of scrambles of the aptamer with the highest binding affinity. We show that the majority of binding aptamers fall above a 0.4 score, with the least-binding aptamer falling below the 0.4 threshold (**Figure 1B**). Notably, all of the validated non-binding aptamers score highly, as well as a small subset of scrambles. 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 1C**). Upon fine-tuning AptaBLE with a subset of two hCD117 aptamers, ABW02 and ABW04, the previously high-scoring scrambles sequences are marked as poor binders by the model (**Figure 1D**). We also observe a relatively stronger correlation between and ordering of the AptaBLE scores for each aptamer according to their experimentally-determined K_d's after fine-tuning (**Figure 1E**).

Investigating Unclustered Aptamers

We previously evolved ssDNA sequences binding to hCD117 through nine rounds of SELEX of increasingly stringent elution conditions[25]. Originally, our work culminated in the identification of a number of clusters in the round 9 population, with sequential edit distances of at least 3 between each other. 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 hCD117.

In an effort to comb through the unclustered sequences, we utilized AptaBLE to efficiently parse through a population of unclustered sequences and identify potential high-affinity binders to hCD117 that were previously dismissed due to the clustering methodology. Across the entire round nine library, only 536 out of 35,384 aptamers were clustered. The clustering methodology consists of producing groups all within a sequence edit distance of 3 to the most copied aptamer within its 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.

To initially sift through the unclustered aptamers, we leveraged a combined approach utilizing AptaBLE scores and sequence copy number (**Figure S3A**). While we focus on unclustered aptamers here, we also scored all clustered aptamers and the overall score distributions across clusters (**Figure S3B,C**). Setting the thresholds for AptaBLE score to 0.4 and a minimum copy number of 10, we identified a sub-population of the unclustered aptamers with potential as binders specific and selective to hCD117. We chose four aptamers, according to their predicted minimum free energies in testing buffer conditions and distinct structures to ABW02, our existing high-affinity hCD117 binder, according to a hierarchical clustering analysis on their predicted secondary structures (**Figure S4**). These unclustered aptamers were tested for binding to hCD117 via biolayer interferometry, where we found that all four bound to the target protein (**Figure 1F**). These aptamers, indicated as 'ABWUC' followed by an arbitrary number, showed comparable and in some cases 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, ABWUC1 and UC2 finished in the top 3 overall with sub-50 nM affinity to hCD117. Additionally, 3 of the 4 unclustered aptamers we selected had favorable binding kinetics over several previously-tested aptamers, indicating that our methodology can identify

ssDNA aptamers that are omitted too early from the customary procedure. These aptamers were generally structurally distal from the known-binding ABW02-ABW07 aptamer series.

De novo generated aptamers are selective to their target with low-nanomolar binding affinity

AptaBLE extends beyond prediction to enable target-specific aptamer generation. Using classifier-guided generation methods implemented either via Monte Carlo Tree Search[26] or Monte Carlo Tree Guidance[27], we generated novel aptamers targeting specific protein binding sites. Our two methods, named AptaBLE-MCTS (**Figure 2A**) and AptaBLE-MCTG (**Figure 2B**), navigate 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. In this work, we only focus on ssDNA aptamer generation.

Human CD25

With AptaBLE-MCTS, we designed 40N ssDNA aptamers validated to bind human CD25. In both cases, aptamers were generated conditioned to optimize binding human CD25 and avoid binding human albumin predicated on AptaBLE score as the metric. Aptamer 77 was found to bind CD25+ cells with low-nanomolar affinity (~31nM) and also exhibited poor binding with CD25- cells (**Figure 2C, top**) via a flow cytometry driven assay. We also identify Aptamer 64, also highly selective to CD25+ cells and a near-nanomolar affinity (~3 μM). Additionally, we investigated the nature of the interaction between Aptamer 77 and hCD25 using a docking procedure (**Figure 2D**). We confirm alignment between the experimental result and the *in silico* modeling, marked by a well-folded ssDNA aptamer positioned adjacent to a contiguous surface on the hCD25 protein dominated by positively-charged residues (e.g. arginine, lysine). We provide additional flow cytometry data, including for the hCD25- HEK293 cell line, in **Figure S5**.

Human TIGIT

Like above, we designed 40N ssDNA aptamers validated to bind human TIGIT and detarget human albumin (**Figure 2E)**. We highlight Aptamer 22 (~7 μM), designed by AptaBLE-MCTS, and Aptamer 27 (~13 μM), produced by AptaBLE-MCTG. Both aptamers are highly selective for the hTIGIT+ MM1.S cell line with near-nanomolar affinity. We provide additional flow cytometry data, including the hTIGIT- HEK293T cell line in **Figure S6**.

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 highlights its practical utility in real-world applications such as SELEX library analysis or *de novo* ssDNA aptamer design. 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.

Looking forward, we aim to further elucidate how to interpret AptaBLE scores and leverage the tool in more finer-grained workflows such as mutagenesis. Moreover, we anticipate the validation of AptaBLE-designed RNA aptamers in the near future. We believe that implementing new workflows with AptaBLE 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.

Materials and Methods

Architecture

AptaBLE employs a symmetric architecture that processes both protein and aptamer sequences through specialized pretrained encoders, followed by a fusion mechanism for combining embeddings into a shared aptamer-protein representation (**Figure 1A**). It 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. 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.

Protein encoder. We use ESM-2[13] with 150M parameters, pretrained on 250 million protein sequences from UniRef50. The encoder processes protein sequences through 33 transformer layers with 1280-dimensional hidden states, generating per-residue embeddings. We freeze all encoder parameters during training to preserve evolutionary information while reducing computational cost and overfitting risk.

Aptamer encoder - DNA. For DNA aptamers, we use Nucleotide Transformer v2-50M[14] (nucleotide-transformer-v2-50m-multi-species), pretrained on 850 billion nucleotides from reference genomes of 3200 species. The encoder uses 6-nucleotide tokenization (6-mers), processing sequences through 12 transformer layers with 512-dimensional hidden states. All parameters are frozen during training.

Aptamer encoder - RNA. For RNA aptamers, we use RNA-FM[15], pretrained on 23 million RNA sequences from RNAcentral. The encoder processes single nucleotides through 12 transformer layers with 640-dimensional hidden states. Parameters are frozen during training.

Dimension alignment. To enable cross-attention between protein (1280-dim) and aptamer (512-dim for DNA, 640-dim for RNA) embeddings, we apply learned linear projections to align dimensions. For DNA models, we project protein embeddings from 1280 to 512 dimensions. For RNA models, we project protein embeddings to 640 dimensions. These projections use standard linear layers without bias: W_align ∈ ℝ^(512×1280) for DNA, W_align ∈ ℝ^(640×1280) for RNA.

Fusion Mechanism

As shown in analogous work[28], we leverage a fusion mechanism to combine embeddings of the aptamer and protein modalities. We take a vectorized protein $P \in W^m$ of length *m* and aptamer $A \in W^n$ of length n and map the pair to a score between 0 and 1.

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

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: $W_q, W_k, W_v \in R^{hxh}$ which are used to generate the queries, keys and values, or Q, K, V, for the module. We therefore generate the following matrices:

$$Q_p = \hat{P}W_q^a,\ K_p = \hat{P}\ W_k^a,\ V_p = \hat{P}W_v^a$$
$$Q_a = \hat{P}W_q^p,\ K_a = \hat{P}\ W_k^p,\ V_a = \hat{P}W_v^p$$

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

$$Q^*_p = \hat{P}W_q^p,\ K_p = \hat{P}\ W_k^p,\ V_p = \hat{P}W_v^p$$
$$Q^*_a = \hat{P}W_q^a,\ K^*_a = \hat{P}\ W_k^a,\ V^*_a = \hat{P}W_v^a$$

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

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

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

De Novo Generation

To generate aptamers *de novo*, we adopt and extend two methods inspired by multi-objective conditional sequence design where AptaBLE is used to compute rewards, and navigate sequence-space exploration with either randomly-generated sample sequences or diffusion-generated sample sequences.

Monte Carlo Tree Search

Our first approach is a simple Monte Carlo Tree Search method[26] only requiring the use of the AptaBLE classifier. In addition to the below, the full algorithm is described in Algorithm 1.

<u>Initialization</u>

We take our final aptamer as a sequence of fixed length N, where initially the sequence is completely empty. At each iteration of the MCTS, a new token is either appended or prepended to the existing aptamer sequence based on the selection-expansion-rollout cycle followed by the algorithm.

<u>Selection</u>

From the root node, the algorithm progresses down the MCTS tree based on the UCT (upper confidence bounds applied to trees) score, defined as the following:

$$UCT_{i,s_i,n_i,N_i} = \frac{s_i}{n_i} + C\sqrt{\frac{ln(N_i)}{n_i}}$$

where i identifies the node, s_i is the accumulated exploitation reward for node i calculated during backpropagation, n_i is the number of visits to node i, N_i is the number of visits to the parent node of node i, and C is a tunable exploration parameter set to $\frac{1}{\sqrt{2}}$ according to prior work[35]. Ultimately, a leaf node is reached following the optimal UCT path.

<u>Expansion</u>

From the selected leaf node, a new child node is added. This reflects either the appending or prepending of a nucleotide base to the existing sequence represented by the currently selected node (i.e. _A, _C, _G, _T, A_, C_, G_, or T_, where _ denotes the current state of the node).

Rollout

For each child node x^{child_i}, the remainder of the aptamer sequence is completed by randomly appending nucleotides from the model alphabet until the sequence reaches length N. This random sequence generation from the child node denotes one sample from the child node state. Ultimately, each sequence generated from this method reflects the child node's construction and is passed into AptaBLE alongside the target protein for score computation. This reflects an individual play-out.

Backpropagation

For each child node x^{child_i}, we update its cumulative reward denoted by its UCT by populating the s_i value using the AptaBLE scores corresponding to that play-out. This procedure of exploring random play-outs is repeated for the indicated number of iterations in the MCTS. Ultimately, the procedure repeats from the selection step, with the child node having the greatest UCT score being added to the ensuing selection path.

Reward

The reward vector for a node consists of two scores which equally contribute to the s_i value in the UCT computation. The first is the AptaBLE score against the target protein and the second is the AptaBLE score against an off-target protein subtracted from 1 to function as a 'negative' selection. These scores are equally weighted and combined, then serving as the sequence's exploitation reward within its UCT.

Monte Carlo Tree Guidance

Our second approach is applying Monte Carlo Tree Guidance[27] to aptamer sequence design. This method requires the use of the AptaBLE classifier and an appropriate masked diffusion model to sample sequences from each explored state. In addition to the below, the full algorithm is described in Algorithm 2.

Here, *de novo* generation requires the use of a masked diffusion model (MDM) for sampling sequences at the rollout stage. We use a pretrained RNA masked diffusion model[29] for sampling sequences from each ssDNA aptamer state.

Initialization

We initialize a fully masked tokenized sequence with length L. This becomes the root node of the MCTS tree with timestep $t = 0$. Additionally, we construct a vector scoring function $f\colon V^L \rightarrow R^K$ that returns a score vector given a clean input sequence $x_0 \in V^L$ and initialize an empty Pareto-optimal set $P^* = \{\}$. We define the following hyperparameters: number of MCTS iterations N_{iter} and number of children nodes N_{child}. The root node is initialized as a fully-masked sequence of length L.

Selection

To determine the next unmasking step from a parent node with an already expanded set of child sequences $x_i^{child} \in children(x)$, we compute a selection score vector $U(x_i^{child}) = R(x_i^{child})/N_{visit}(x_i^{child})$ that divides the cumulative reward vector $R(x_i^{child})$ of a child node to the number of times it was visited $N_{visit}(x_i^{child})$. This encourages exploration of child nodes that have not been visited, in addition to high-reward nodes. Then, we select from the Pareto-optimal set of child nodes based on the normalized selection score vector.

$$\mathcal{P}^*_{select} = \{x_i^{child} \mid \nexists x_j^{child} \in children(x)\ s.t.\ U(x_j^{child}) \succ U(x_i^{child})$$

If we select a non-leaf node, we restart the selection process with the selected node as the new parent until we reach an expandable leaf node or a fully unmasked sequence. In the case of a fully unmasked sequence, we restart the selection from the root node.

Expansion

From the leaf node x_t at time t, we sample N_{child} distinct sequences by applying Gumbel noise to the denoising distribution $p_{0|t}^{\theta}(x_0|x_t)$

$$x_0^i \sim \log p_{0|t}^{\theta}(x_0|x_t) + g_i$$

where $g_i \sim Gumbel(0,1)$. Then, for each child node x_i^{child}, we randomly select k positions in x_t to unmask according to x_1^i.

Rollout

For each child node x_i^{child}, we use self-planned ancestral sampling to obtain a fully unmasked sequence $\tilde{x}_i^{child}$. This is in stark contrast to the aforementioned MCTS method, where sequences corresponding to each state are produced completely at random. Then, we feed each of the clean sequences into the vector score function $f(\tilde{x}_i^{child})$ to obtain a score vector. We then compute the reward vector $r(x_i^{child}) \in r^k$ of each child x_i^{child} where each element is the fraction of sequences in the Pareto-optimal set $x^* \in P^*$ that x_i^{child} dominates.

$$r_k(x_i^{child}) = \frac{1}{|\mathcal{P}^*|} \sum_{x^* \in \mathcal{P}^*} 1[f_k(x_i^{child}) \geq f_k(x^*)]$$

For all N_{child} children sequences, we add x_i^{child} to P^* if it is *non-dominated* by all sequences $x^* \in P^*$ and remove all sequences in P^* dominated by some x_i^{child}.

Backpropagation

For each child node x_i^{child}, we use its reward vector $r(x_i^{child})$ to initialize the cumulative reward $R(x_i^{child}) = r(x_i^{child})$ and the number of visits $N_{visit}(x_i^{child})$ is set to 1.

We then trace backward from x_i^{child} through its ancestors up to the root node x_0^{root}, updating each node x along the path by accumulating the child rewards and incrementing the visit count:

$$W(x) \leftarrow W(x) + \sum_{i=1}^{N_{child}} r(x_i^{child})$$
$$N_{visit}(x) \leftarrow N(x) + 1$$

These cumulative updates influence future selection, prioritizing unmasking paths that lead to higher-reward sequences for further expansion. After N_{iter} iterations of these four steps, we return the set of aptamer sequences optimized across the set of properties $x^* \in P^*$ and their corresponding score vectors $f(x^*)$.

Reward Vector

Similar to the MCTS method, the reward vector here reflects the AptaBLE scores for each aptamer against the target protein and an off-target protein serving as an *in silico* negative control (subtracted from 1). These scores are equally weighted and combined, forming the backpropagated reward.

Settings

All generated libraries consisted of 100 40N ssDNA aptamers.

MCTS

- Number of iterations: 100
- Rollout strategy: Random nucleotide completion from each leaf node
- Number of children: 8 (fixed)

MCTG

- Number of iterations: 128
- Rollout strategy: masked diffusion model sampling
- Number of children: 12

Retrospective Analysis

Aptamers from the described hCD117 SELEX were ultimately clustered by FastAptamer[9] and retrospectively scored via AptaBLE to identify which clusters are better predicted to serve as functional binders to hCD117 (**Figure S2C**). It is expected that the higher-scoring clusters would have been prioritized in the analysis rather than simply advancing sequences based on their number of copies following sequencing. As sequences from clusters 8-11 were not characterized for binding, focus was kept on clusters 1-7. From **Figure S2A**, sequences with the highest copy number are not always scored the highest by AptaBLE within their own cluster. Additionally, there is a tight correspondence between binding affinity from a representative aptamer within each cluster and the overall distribution of AptaBLE scores within a cluster (**Figure S2C**). From this analysis, sequences from clusters 2, 3, and 7 would have been prioritized for characterization using a workflow involving AptaBLE. Notably, the constituent aptamers from these clusters had the strongest binding affinity to hCD117 as determined via BLI.

Secondary structure prediction for all aptamers was done with ViennaRNA.

Biolayer Interferometry

Binding affinity measurements for all synthesized aptamers shown in **Figure 2A** were performed in a biolayer interferometry (BLI) assay with an Octet GatorPrime instrument. All assays were performed at 30°C in 96-well BLI 96-Flat Black Plates, with samples prepared in an assay buffer containing 50 mM HEPES, 150 mM NaCl, and 5 mM $MgCl_2$ (pH 7.4). For kinetic analysis of all free hCD117 aptamers (ABWUC1-4, ABW02-07), anti-VHH biosensors were used to immobilize hCD117 (Sino Biological, Inc.) at 50 µg/mL for 180 seconds, followed by capture of 6x-His-tagged recombinant human CD4 (rCD4; Sino Biological, Inc.) at 3 µg/mL for 180 seconds. All aptamers were then tested at concentrations of 20,000, 10,000, 5,000, 2,500, 1,250, and 625 nM, with association and dissociation phases of 900 seconds each.

Flow Cytometry

To assess the specificity and affinity of FAM-labeled aptamers (Aptamer 77 selected for CD25 using AptaBLE) toward Ramos cells ($CD25^+$), a comparative binding assay was performed using HEK293T cells ($CD25^-$) as a negative control. Aptamer-FAM was first dissolved in nuclease-free water to a final concentration of 1 mM. A 50 μM working solution was prepared using a folding buffer containing 20 mM HEPES (pH 7.4), 150 mM NaCl, and 5 mM $MgCl_2$. Aptamers were folded by heating at 95 °C for 5 min, cooling at 4 °C for 5 min, incubating at 25 °C for 5 min, and holding again at 4 °C for 5 min. The folded aptamers were serially diluted 4-fold in 1× selection buffer, ranging from 50 μM to 0.781 μM.

Ramos cells were harvested from a T-25 flask and transferred to a 15 mL centrifuge tube. The cells were pelleted by centrifugation at 400 × g for 10 min, after which the supernatant was completely removed. The pellet was gently resuspended in 1 mL of 1× selection buffer (20 mM HEPES (pH 7.4), 150 mM NaCl, 5 mM $MgCl_2$ and 10% FBS). Cell concentration and viability were determined using Trypan Blue. HEK293T cells were cultured and prepared in parallel, to serve as the negative control group.

A 96-well plate was prepared for the binding assay. Each well contained a total of 10,000 cells in a final volume of 100 μL. To achieve this, an appropriate volume of 1× selection buffer was first added to each well so that, after cell addition, the total volume equaled 90 μL. The calculated volume of cell suspension corresponding to 10,000 cells was added to each well. 10 μL of each concentration of folded aptamer was added to the designated wells with six replicates per concentration. Wells containing either Ramos or HEK293T cells without aptamer were used as fluorescence background controls. The plate was gently mixed and incubated at 37 °C in a CO_2 incubator for 40 min.

After incubation, the plate was centrifuged at 450 × g for 10 min at room temperature. The supernatant was removed using a multichannel pipette. Each well was resuspended in 200 μL of 1× selection buffer.

Flow cytometry was performed using an Attune cytometer equipped with a 488 nm laser and BL1 channel to detect FAM fluorescence. Standard gating was applied to select single, viable cells based on FSC-A vs SSC-A and FSC-H vs FSC-A parameters. A minimum of 5,000–10,000 events were collected per sample. Binding efficiency was quantified by comparing the median fluorescence intensity (MFI) in Ramos versus HEK293T populations.

All experiments were performed in six replicates, and K_d was determined by fitting the MFI values against aptamer concentration using a one-site specific binding model. Nonlinear regression analysis was carried out using GraphPad Prism to obtain best-fit parameters and standard errors.

Author Contributions

S.P. performed all non-docking *in silico* experiments and primarily wrote the manuscript. F.Z.P. advised on the model design and contributed to the manuscript. K.F. performed the docking results and contributed to the manuscript. D.G. and A.D. performed all *in vitro* experiments, including flow cytometry and functional assays. They also produced the bioinformatics analyses on the resulting data. X.W. advised on the *in vitro* experiments. S.Y. and P.C supervised the project.

Declarations

The authors acknowledge that this work was funded by Atom Bioworks, Inc.

Competing Interests Statement

The authors declare no competing interests.

Acknowledgements

The authors would like to thank Shrey Shah and Meenakshi Ranasinghe for consulting on the project.

References

1. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-510 (1990).
2. Zhou, J. & Rossi, J. Aptamers as targeted therapeutics: current potential and challenges. *Nat. Rev. Drug Discov.* **16**, 181-202 (2017).
3. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818-822 (1990).
4. Thiel, W. H., Bair, T., Wyatt Thiel, K., Dassie, J. P., Rockey, W. M., Howell, C. A., Liu, X. Y., Dupuy, A. J., Huang, L., Owczarzy, R., Behlke, M. A., McNamara, J. O. & Giangrande, P. H. Nucleotide bias observed with a short SELEX RNA aptamer library. *Nucleic Acid Ther.* **21**, 253-263 (2011).
5. Jarosch, F., Buchner, K. & Klussmann, S. In vitro selection using a dual RNA library that allows primerless selection. *Nucleic Acids Res.* **34**, e86 (2006).
6. Hoinka, J., Berezhnoy, A., Dao, P., Sauna, Z. E., Gilboa, E. & Przytycka, T. M. Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Res.* **43**, 5699-5707 (2015).
7. Dupuis, N. F., Holmstrom, E. D. & Nesbitt, D. J. Molecular-crowding effects on single-molecule RNA folding/unfolding thermodynamics and kinetics. *Proc. Natl Acad. Sci. USA* **111**, 8464-8469 (2014).
8. Ishida, R., Adachi, T., Yokota, A., Yoshihara, H., Aoki, K., Nakamura, Y. & Hamada, M. RaptRanker: in silico RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information. *Nucleic Acids Res.* **48**, e82 (2020).
9. Alam, K. K., Chang, J. L. & Burke, D. H. FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Mol. Ther. Nucleic Acids* **4**, e230 (2015).
10. Wong, F., He, D., Krishnan, A., Hong, L., Wang, A. Z., Wang, J., Hu, Z., Omori, S., Li, A., Rao, J., et al. Deep generative design of RNA aptamers using structural predictions. *Nat. Comput. Sci.* **4**, 863-872 (2024).
11. Woldeyes, R. A., Sivak, D. A. & Fraser, J. S. E pluribus unum, no more: from one crystal, many conformations. *Curr. Opin. Struct. Biol.* **28**, 56-62 (2014).
12. Hermann, T. & Patel, D. J. Adaptive recognition by nucleic acid aptamers. *Science* **287**, 820-825 (2000).
13. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. & Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* (2022). doi:10.1101/2022.07.20.500902
14. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., Sirelkhatim, H., Richard, G., et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv* (2023). doi:10.1101/2023.01.11.523679
15. Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., Zong, L., Hong, L., Xiao, J., King, I. & Zhang, Y. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv* preprint arXiv:2204.00300 (2022).
16. Shin, I., Kang, K., Kim, J., Sel, S., Choi, J., Lee, J.-W., Kang, H. Y. & Song, G. AptaTrans: a deep neural network for predicting aptamer-protein interaction using pretrained encoders. *BMC Bioinformatics* **24**, 447 (2023).
17. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. DiffDock: Diffusion steps, twists, and turns for molecular docking. *arXiv* preprint arXiv:2210.01776 (2022).
18. Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C. & Zheng, S. TANKBind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Adv. Neural Inf. Process. Syst.* **35**, 7236-7249 (2022).

19. Cao, D., Chen, G., Jiang, J., Yu, J., Zhang, R., Chen, M., Zhang, W., Chen, L., Zhong, F., Zhang, Y., et al. Generic protein-ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nat. Mach. Intell.* **6**, 1439-1453 (2024).
20. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024).
21. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
22. Askari, A., Kota, S., Ferrell, H., Swamy, S., Goodman, K. S., Okoro, C. C., Crenshaw, I. C. S., Hernandez, D. K., Oliphant, T. E., Badrayani, A. A., et al. UTExas Aptamer Database: the collection and long-term preservation of aptamer sequence information. *Nucleic Acids Res.* **52**, D351-D359 (2024).
23. Li, B.-Q., Zhang, Y.-C., Huang, G.-H., Cui, W.-R., Zhang, N. & Cai, Y.-D. Prediction of aptamer-target interacting pairs with pseudo-amino acid composition. *PLoS One* **9**, e86729 (2014).
24. Lin, Y.-C., Chen, W.-Y., Hwu, E.-T. & Hu, W.-P. In-silico selection of aptamer targeting SARS-CoV-2 spike protein. *Int. J. Mol. Sci.* **23**, 5810 (2022).
25. Dwivedy, A., Baskaran, D., Sharma, G., Hong, W., Gandavadi, D., Krissanaprasit, A., ... & Wang, X. (2025). Engineering Novel DNA Nanoarchitectures for Targeted Drug Delivery and Aptamer Mediated Apoptosis in Cancer Therapeutics. *Advanced Functional Materials*, *35*(22), 2425394.
26. Lee, G., Jang, G. H., Kang, H. Y. & Song, G. Predicting aptamer sequences that interact with target proteins using an aptamer-protein interaction classifier and a Monte Carlo tree search approach. *PLoS One* **16**, e0253760 (2021).
27. Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. Peptune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. *ArXiv*, 2412, (2025).
28. Xu, M., Yuan, X., Miret, S. & Tang, J. ProtST: Multi-modality learning of protein sequences and biomedical texts. *Proc. 40th Int. Conf. Mach. Learn.* **202**, 38749-38767 (2023).
29. Patel, S., Peng, F. Z., Fraser, K., Friedman, A. D., Chatterjee, P., & Yao, S. (2025). EvoFlow-RNA: Generating and Representing non-coding RNA with a Language Model. *bioRxiv*, 2025-02.
30. Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F. & Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
31. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026-1028 (2017).
32. Zhao, N., Pei, S. N., Parekh, P., Salazar, E., & Zu, Y. (2014). Blocking interaction of viral gp120 and CD4-expressing T cells by single-stranded DNA aptamers. *The international journal of biochemistry & cell biology*, *51*, 10-18.
33. Eguchi, A., Ueki, A., Hoshiyama, J., Kuwata, K., Chikaoka, Y., Kawamura, T., ... & Sando, S. (2021). A DNA aptamer that inhibits the aberrant signaling of fibroblast growth factor receptor in cancer cells. *Jacs Au*, *1*(5), 578-585.

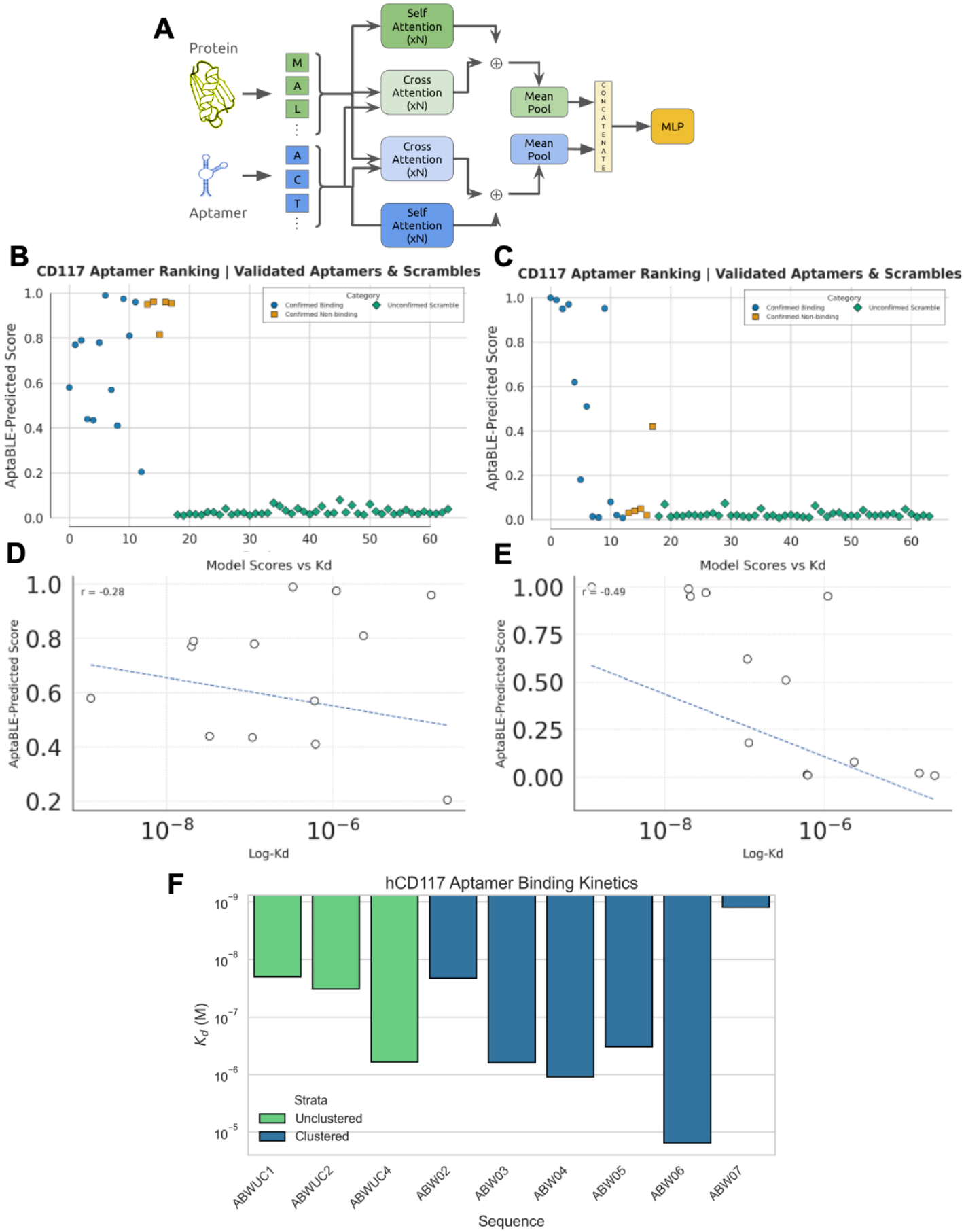
A
Protein
Aptamer
M
A
L
A
C
T
Self Attention (xN)
Cross Attention (xN)
Cross Attention (xN)
Self Attention (xN)
Mean Pool
Mean Pool
CONCATENATE
MLP
B
CD117 Aptamer Ranking | Validated Aptamers & Scrambles
AptaBLE-Predicted Score
Category
Confirmed Binding
Confirmed Non-binding
Unconfirmed Scramble
C
CD117 Aptamer Ranking | Validated Aptamers & Scrambles
AptaBLE-Predicted Score
Category
Confirmed Binding
Confirmed Non-binding
Unconfirmed Scramble
D
Model Scores vs Kd
r = -0.28
AptaBLE-Predicted Score
Log-Kd
E
Model Scores vs Kd
r = -0.49
AptaBLE-Predicted Score
Log-Kd
F
hCD117 Aptamer Binding Kinetics
K_d (M)
Strata
Unclustered
Clustered
ABWUC1
ABWUC2
ABWUC4
ABW02
ABW03
ABW04
ABW05
ABW06
ABW07
Sequence

Figure 1. AptaBLE construction and *in silico* results. 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)** A series of aptamers targeting human 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). **C)** After fine-tuning on a subset of the binding aptamers and non-binding aptamers, an apparent ordering by K_d emerges. **D)** A scatterplot depicting the Pearson correlation between original model scores from **B** and K_d for validated binders. One aptamer is omitted as a K_d was not fit for it. **E)** A scatterplot depicting the Pearson correlation between fine-tuned model scores from **C** and K_d for validated binders. One aptamer is omitted as a K_d was not fit for it. **F)** Three aptamers, previously ignored during standard SELEX library analysis, were highlighted by AptaBLE and showed high affinity to hCD117 (~20nM, ~33nM, ND). Two aptamers fall in the top four highest-affinity aptamers in the entire tested pool, including those initially selected from the original SELEX library analysis.

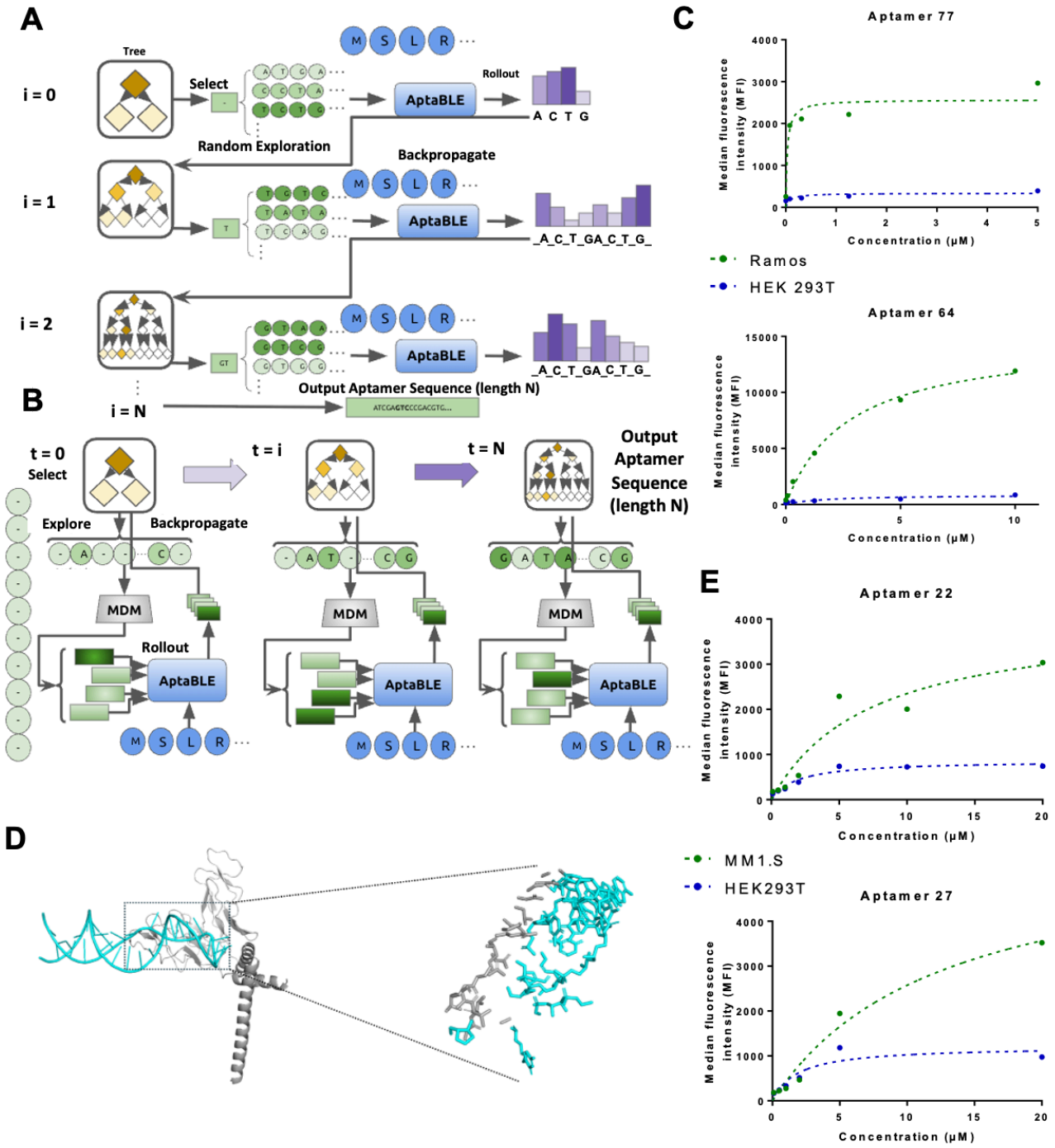

Figure 2. AptaBLE can discover selective, high-affinity binding aptamers. A) Algorithm schematic for AptaBLE-MCTS (see Algorithm 1 for details). **B)** Algorithm for AptaBLE-MCTG (see Algorithm 2 for details). **C)** Aptamers designed *de novo* for binding human CD25 show heightened selectivity to human CD25+ Ramos cells. Affinity is low/negligible to CD25- HEK293T cells. Binding curves are shown for Aptamer 77 (**top**; ~31nM) and Aptamer 64 (**bottom**; ~3 μM). **D)** Docking study of Aptamer 77 (cyan) with hCD25 (gray) complex using predicted structures (**left**) and the suggested binding interface within the complex (**right**). **E)** Aptamers designed *de novo* for binding human TIGIT show heightened selectivity to human TIGIT+ MM1.S cells. Affinity is low/negligible to TIGIT- HEK293T cells. Binding curves are shown for Aptamer 22 (**top**; ~7 μM) and Aptamer 27 (**bottom**; ~13 μM).

Appendix

AptaBLE performance exceeds competing methods

AptaBLE was trained on validated, published aptamers[22,23], validated proprietary aptamers, and artificially generated pseudo-negative derivatives of those aptamers. Further details regarding our training and benchmarking dataset and implementation can be found in the benchmark and auxiliary benchmark sections.

We benchmarked AptaBLE against two representative approaches: AptaTrans[16], the current state-of-the-art sequence-based method, and AlphaFold3,[20] leveraging structure-based prediction. AptaBLE significantly outperformed both baselines across all metrics for DNA aptamers (**Figure S1A**). The comparably high recall is particularly important for SELEX analysis, as false negatives represent missed therapeutic candidates. For RNA aptamers, AptaBLE also performed favorably over the competing models, though the margin over AptaTrans was smaller (**Figure S1B**). AptaBLE additionally outperformed both models in predicting stable ssDNA aptamer-protein complexes in a small auxiliary dataset (**Figure S1C**).

Runtime analysis demonstrated AptaBLE's practical efficiency. Scoring a single aptamer-protein pair required 48 milliseconds on average in batch mode on an A100 GPU,

Fusion attention maps suggest potential aptamer-protein binding interfaces

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)[24].

We observe one of our attention heads in the cross attention fusion module representing protein-attended nucleic acids reflecting atomic interactions seen at the binding interface between the S-protein and the aptamer (**Figure S2A,B**). 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. As such, it is inconclusive as to whether these cross attention maps indicate inter-residue interaction.

<u>Training Dataset</u>

Positive examples. We compiled aptamer-protein binding pairs from three sources:

1. UTexas Aptamer Database[22]: 847 DNA aptamers, 423 RNA aptamers with validated binding
2. Li et al. 2014 dataset[23]: 256 additional RNA aptamers
3. Proprietary data from Atom Bioworks

After removing duplicates and filtering for sequence quality (no ambiguous bases, length 20-100nt), the final positive dataset contained: 1,089 DNA aptamers binding 127 protein targets, 757 RNA aptamers binding 94 protein targets.

Negative examples. For each positive aptamer, we generated four negative examples by shuffling the sequence while preserving nucleotide composition and length. Shuffling was performed using Python's random.shuffle with different random seeds. We verified that shuffled sequences have different predicted secondary structures (>50% difference in base pairing) using ViennaRNA RNAfold[30].

This 4:1 negative:positive ratio was chosen based on preliminary experiments showing that higher ratios improved validation performance without requiring prohibitive computational resources. The shuffling strategy creates chemically plausible sequences (matching length and composition of real aptamers) while eliminating

binding, forcing the model to learn sequence-structure-function relationships rather than simple nucleotide frequencies.

Dataset splitting. To prevent data leakage and test generalization to novel proteins, we performed homology-based splitting:

1. Cluster all protein targets using MMseqs2[31] with parameters: --min-seq-id 0.8, --cov-mode 0, -c 0.8
2. Assign entire clusters to training (80%), validation (10%), or test (10%) sets
3. This ensures no test protein shares >80% sequence identity with any training protein

Benchmark Dataset

We collected an additional benchmark set of 16 ssDNA aptamers and 39 RNA aptamers to use for evaluating models capable of predicting aptamer-protein interactions. These aptamers and their protein targets are omitted from the training dataset. We evaluate three models, including AptaBLE, on each of these datasets in Figure S1A,B.

Auxiliary Benchmark Set

This dataset consists of four known-binding ssDNA aptamer-protein pairs which we highlight in this work – two aptamers binding *SARS-CoV-2* Spike glycoprotein[24], one aptamer binding gp120[32], and the final aptamer binding basic fibroblast growth factor[33]. Results on these pairs are reported in Figure S1C.

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.

Training

*Loss function***.** Binary cross-entropy loss:

$$L = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

where $y \in \{0,1\}$ is the true label and $\hat{y} \in [0,1]$ is the predicted probability.

*Optimization***.** AdamW optimizer with parameters: $\beta_1=0.9$, $\beta_2=0.999$, $\varepsilon=10^{-8}$, weight decay=0.01. Initial learning rate: 10^{-5}. Learning rate schedule: MultiStepLR with decay factor 0.1 at epochs [5, 10, 20].

Training hyperparameters.

- Batch size: 16 (limited by GPU memory for long sequences)
- Maximum epochs: 30 (DNA), 20 (RNA)
- Gradient clipping: max norm 1.0
- Mixed precision training: FP16 using torch.cuda.amp
- Early stopping: patience of 5 epochs on validation AUC-ROC

*Hardware and runtime***.** All training performed on Google Cloud Platform a2-ultragpu-1g instances with NVIDIA A100 80GB GPU, 12 vCPUs, 170GB RAM. Typical training time: 18 hours for DNA model, 14 hours for RNA model. Memory usage: 65GB GPU memory at peak (during backpropagation for longest sequences).

We use the MultiStepLR Scheduler with an initial learning rate of $1 * 10^{-5}$ 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. 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.

Baseline comparisons

AptaTrans. We implemented AptaTrans[16]. We retrained the model using their published code and dataset for fair comparison. For DNA aptamer evaluation, we converted sequences to RNA (T→U substitution) as required by their RNA-specific encoder. We used their best checkpoint (epoch 50) for all evaluations.

AlphaFold3. We ran AlphaFold3[20] structure prediction using the default server (version 2024.1) on a subset of 200 test examples due to computational cost (5 min/prediction). We used predicted local distance difference test (pLDDT) scores >70 as indicating binding. For aptamer-protein pairs with multiple predicted structures (5 models per prediction), we used the highest-confidence model.

Statistical testing. All performance comparisons used bootstrap resampling (1000 iterations) to compute 95% confidence intervals. Significance was assessed using permutation tests (1000 permutations) with α=0.05.

Attention analysis

Structure data. We obtained crystal structures for RBD-1C (PDB: 7KMD) and RBD-4C (PDB: 7KME) from Lin et al. 2022[24]. Structures were processed using BioPython to extract coordinates.

Contact definition. Protein-aptamer contacts were defined as residue-nucleotide pairs with any heavy atom within 5Å distance. This yielded 47 contacts for RBD-1C and 52 contacts for RBD-4C. We also note that each column in the attention map reflects a disjoint group of 6 nucleic acids, as we follow the aptamer encoder[14] tokenization scheme.

Attention extraction. For each aptamer-protein pair, we:

1. Performed forward pass through AptaBLE
2. Extracted attention weights from the protein-to-aptamer cross-attention module (8 heads, each Lp×La matrix)
3. Averaged across attention heads to yield single attention matrix A ∈ ℝ^(Lp×La)
4. For Nucleotide Transformer's 6-mer tokens, aggregated attention to nucleotide level by assigning each nucleotide the maximum attention from any overlapping token

<u>Architecture</u>
We list our model architecture configurations here.

Cross-attention module: The hidden size and number of heads are 512 and 8..
Self-attention module: The hidden size and number of heads are 512 and 8.
Reshape layer: We reshape ESM2 embeddings to have a hidden dim of either 512 or 640 from 1280 for compatibility with Nucleotide-Transformer or RNA-FM, respectively.

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.

Algorithms

Here, we provide algorithms for both aptamer generation methods described in this work.

Algorithm 1 Monte-Carlo Tree Search generation of single-stranded DNA aptamers.

Input: Score function $\mathbf{f}(x) : \mathcal{V}^L \to \mathbb{R}^K$, number of iterations N_{mcts}, number of children N_{child}
Output: A single ssDNA aptamer sequence $\S^*$
Initialize: $x_0 \leftarrow$ "",
for $i = 1$ to L **do**
  Selection:
  /* Traverse tree to expandable leaf */
  $x_{\text{leaf}} \leftarrow \textsc{Select}(x_0)$
  Expansion:
  for $j = 1$ to N_{child} **do**
    /* Sample from the categorical distribution of nucleotide bases to append/prepend. */
    Sample $x_1^{(j)} \sim C(\cdot|x_{\text{leaf}})$
    Create child $x_{\text{child}}^{(j)}$ by appending or prepending a nucleotide base to the parent node sequence
    Add $x_{\text{child}}^{(j)}$ to children(x_{leaf})
  Rollout:
  for $i = 1$ to N_{mcts} **do**
    /* Complete sequence using random sequence infilling */
    $\tilde{x}^{(j)} \leftarrow \textsc{RandomInfilling}(x_{\text{child}}^{(j)})$
    /* Evaluate properties using trained regressors */
    $\mathbf{s}^{(j)} \leftarrow \mathbf{f}(\tilde{x}^{(j)})$
    /* Compute UCT reward */
    $\mathbf{r}^{(j)} \leftarrow \text{UCT}(\mathbf{s}^{(j)})$
  Backpropagation:
  /* Update ancestor node statistics with child rewards */
  $\textsc{Backpropagate}(x_{\text{leaf}}, \{\mathbf{r}^{(j)}\})$
/* Return the final selected node */
return x^*

Algorithm 2 Monte-Carlo Tree Guidance generation of single-stranded DNA aptamers.

Input: Denoising model $p_\theta(x_1|x_t, t)$, score function $\mathbf{f}(x) : \mathcal{V}^L \to \mathbb{R}^K$, number of iterations N_{mcts}, number of children N_{child}
Output: Set of Pareto-optimal ssDNA aptamer sequences $\mathcal{P}^*$
Initialize: $x_0 \leftarrow [M]^L$, $\mathcal{P}^* \leftarrow \{\}$, $t \leftarrow 1$
for $i = 1$ to N_{mcts} **do**
  Selection:
  /* Traverse tree to expandable leaf */
  $x_{\text{leaf}} \leftarrow \textsc{Select}(x_0)$
  Expansion:
  for $j = 1$ to N_{child} **do**
    Sample $x_1^{(j)} \sim p_\theta(\cdot|x_{\text{leaf}}, t)$ with Gumbel noise
    Create child $x_{\text{child}}^{(j)}$ by unmasking k positions according to $x_1^{(j)}$
    Add $x_{\text{child}}^{(j)}$ to children(x_{leaf})
  Rollout:
  for each $x_{\text{child}}^{(j)} \in$ children(x_{leaf}) **do**
    /* Complete sequence using self-planned ancestral sampling */
    $\tilde{x}^{(j)} \leftarrow \textsc{SelfPlannedSampling}(x_{\text{child}}^{(j)})$
    /* Evaluate properties using trained regressors */
    $\mathbf{s}^{(j)} \leftarrow \mathbf{f}(\tilde{x}^{(j)})$
    Update $\mathcal{P}^*$ with $\tilde{x}^{(j)}$ if non-dominated
    /* Compute reward based on Pareto dominance */
    $\mathbf{r}^{(j)} \leftarrow \textsc{ComputeReward}(\mathbf{s}^{(j)}, \mathcal{P}^*)$
  Backpropagation:
  /* Update ancestor node statistics with child rewards */
  $\textsc{Backpropagate}(x_{\text{leaf}}, \{\mathbf{r}^{(j)}\})$
return $\mathcal{P}^*$

CD117 SELEX Data

SELEX Procedure

In vitro aptamer selection was performed using an N26 random DNA library, CoV2-RBD-1 AMSS[25]. The naïve library, in addition to enriched aptamer pools within each round of selection, were PCR-amplified using 0.5 mM of forward primer CoV2-RBD-1 Fwd and reverse primer CoV2-RBD-1 b-Rev via cycling at 95 °C, 56.5 °C, and 72 °C for 30 s, respectively, for an appropriate number of cycles. Subsequently, the PCR amplified products were immobilized for 15 minutes at room temperature on Streptavidin Mag Sepharose beads (Cytiva, Marlborough, MA) at a v/v of 5:1 PCR product to 10% bead slurry stock. Beads were washed with 2x5 column volumes (CVs) of 1x Selection Buffer (20mM HEPES, 150 mM NaCl, 5 mM MgCl2, pH 7.4), followed by 5 CVs of nuclease-free water. After removal of the nuclease-free water supernatant, sense ssDNA was eluted from the beads via two elutions, each using 4.35 CVs 10 mM NaOH was added to the beads and mixed for 5 minutes at room temperature. Sense ssDNA was then exchanged into 1x Selection Buffer via 7 kDa MWCO ZEBA desalting column (ThermoFisher, Waltham, MA).

In preparation for selection, individual populations of magnetic beads were immobilized with HSA (for negative selection; MP Biomedicals, Irvine, CA), 6xHis peptide (for negative selection; ApexBio, Houston, TX), or hCD117 (for positive selection; Sino Biological, Beijing, China). HSA was crosslinked to NHS Mag Sepharose beads (Cytiva, Marlborough, MA) via manufacturer instructions at a ratio of 2.5 µg BSA per CV beads. 6xHis peptide was immobilized on His Mag Sepharose Ni beads (Cytiva, Marlborough, MA) via manufacturer instructions at a ratio of 5 µg 6xHis peptide per CV beads.

Naïve library or regenerated aptamer pools were folded in 1x Selection Buffer via incubation at 95 °C for 5 min, 4 °C for 5 min, and 25 °C for 25 min. 500 pmol of naïve library was used for Round 1; subsequent rounds used between 40 to 80 pmol of the previous regenerated aptamer pool as required for selection stringency. Naïve library and aptamer pools were subjected to negative selection against 20 uL 10% His Mag Sepharose Ni beads (Cytiva, Marlborough, MA), BSA-crosslinked NHS Mag Sepharose beads (5 µg BSA; Cytiva, Marlborough, MA), and 6xHis peptide-immobilized His Mag Sepharose Ni beads (5 µg 6xHis peptide; Cytiva, Marlborough, MA). Naïve library and aptamer pools were incubated with negative selection beads for 30 minutes. Incubation time with, and the total amount of, hCD117 beads during positive selection were progressively shortened throughout the 9 round in vitro selection to increase selection stringency and encourage the selection of aptamers with faster kON. After incubation, hCD117 beads were washed with aliquots of 100 µL 1x Selection Buffer; the number of washes and their duration were progressively increased through the 9 round in vitro selection to increase selection stringency and encourage the selection of aptamers with slower kOFF. Aptamers were eluted from the hCD117 beads via incubation in 100 µL His Elution Buffer (20 mM HEPES, 150 mM NaCl, 5 mM MgCl2, 500 mM Imidazole, pH 7.4) for 15 minutes. ssDNA aptamers were amplified by PCR as described above for an optimized number of cycles. The aptamer pool from Rounds 9 was submitted for Next Generation Sequencing (NGS; Azenta Genewiz, Waltham, MA). Geneious Prime (Dotmatics, Boston, MA) was used for NGS paired end assembly and data processing. FASTAptamer Count and FASTAptamer Cluster in silico toolkits were used to identify dominant unique sequences and characterize them into clusters with an edit distance of 3 nucleotide differences. Clustered aptamers are shown in Figure S3A as gold points, where all points correspond to a sequence within the round 9 SELEX pool. Moreover, each cluster adopts a characteristic secondary structure that is uncommon to those which were unclustered by FastAptamer. Originally, dominant sequences representing the largest sequence clusters were selected for oligosynthesis and characterization. Structural homology across the most-amplified aptamers within each cluster and the unclustered aptamers retrospectively identified by AptaBLE as high-affinity aptamers is fairly limited, as shown by Figure S4 where we show a dendrogram following a hierarchical clustering of all sequences based on predicted secondary structures.

Additionally, we also show the distribution of AptaBLE scores for each of the 11 clusters produced by FastAptamer (Figure S3B,C).

Binding Assays

Binding affinity measurements for all synthesized aptamers shown in Figure 2A were performed in a biolayer interferometry (BLI) assay with an Octet GatorPrime instrument. All assays were performed at 30°C in 96-well BLI 96-Flat Black Plates, with samples prepared in an assay buffer containing 50 mM HEPES, 150 mM NaCl, and 5 mM $MgCl_2$ (pH 7.4). For kinetic analysis of all free hCD117 aptamers (ABWUC1-4, ABW02-07), anti-VHH biosensors were used to immobilize human CD117 (Sino Biological, Inc.) at 50 µg/mL for 180 seconds, followed by capture of 6x-His-tagged recombinant human CD4 (rCD4; Sino Biological, Inc.) at 3 µg/mL for 180 seconds. All aptamers were then tested at concentrations of 20,000, 10,000, 5,000, 2,500, 1,250, and 625 nM, with association and dissociation phases of 900 seconds each.

Retrospective analysis

Aptamers from the described hCD117 SELEX were ultimately clustered by FastAptamer[9] and retrospectively scored via AptaBLE to identify which clusters are better predicted to serve as functional binders to human CD117 (Figure S3B,C). Though a common practice, this procedure may result in clusters that are strikingly similar to each other 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.
It is expected that the higher-scoring clusters would have been prioritized in the analysis rather than simply advancing sequences based on their number of copies following sequencing. As sequences from clusters 8-11 were not characterized for binding, focus was kept on clusters 1-7. From Figure S3B, sequences with the highest copy number are not always scored the highest by AptaBLE within their own cluster. Additionally, there is a tight correspondence between binding affinity from a representative aptamer within each cluster and the overall distribution of AptaBLE scores within a cluster (Figure S3C). From this analysis, sequences from clusters 2, 3, and 7 would have been prioritized for characterization using a workflow involving AptaBLE. Notably, the constituent aptamers from these clusters had the strongest binding affinity to hCD117 as determined via BLI.

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 S2B). The overall distributions of scores for aptamers within a cluster also varies widely across clusters (Figure S2C). 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.

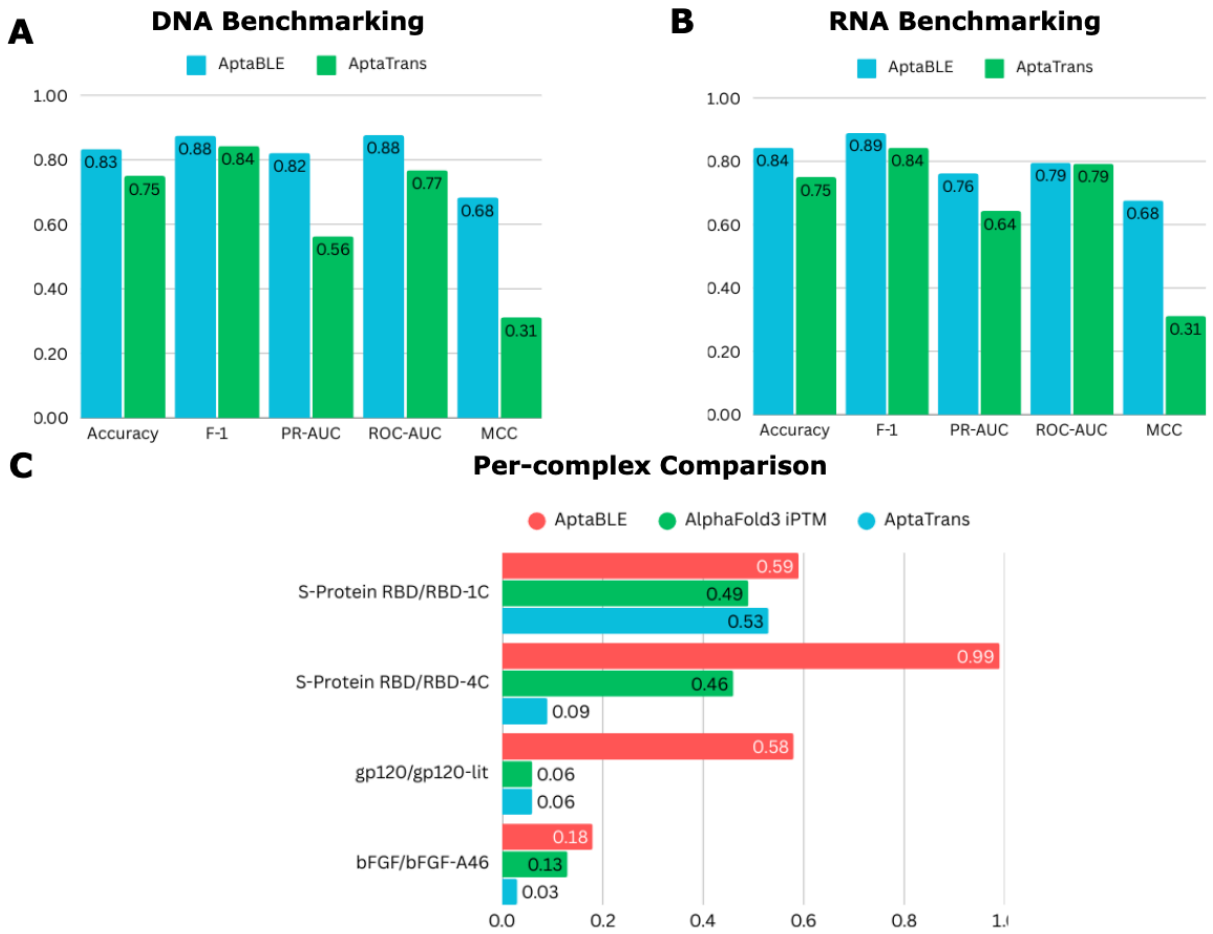

Figure S1. AptaBLE benchmarking. A) Performance of baseline models across validated binding and non-binding ssDNA aptamer datasets. **B)** Performance of baseline models across validated binding and non-binding RNA aptamer datasets. **C)** Comparison of key model metrics across four well-studied aptamer-protein systems.

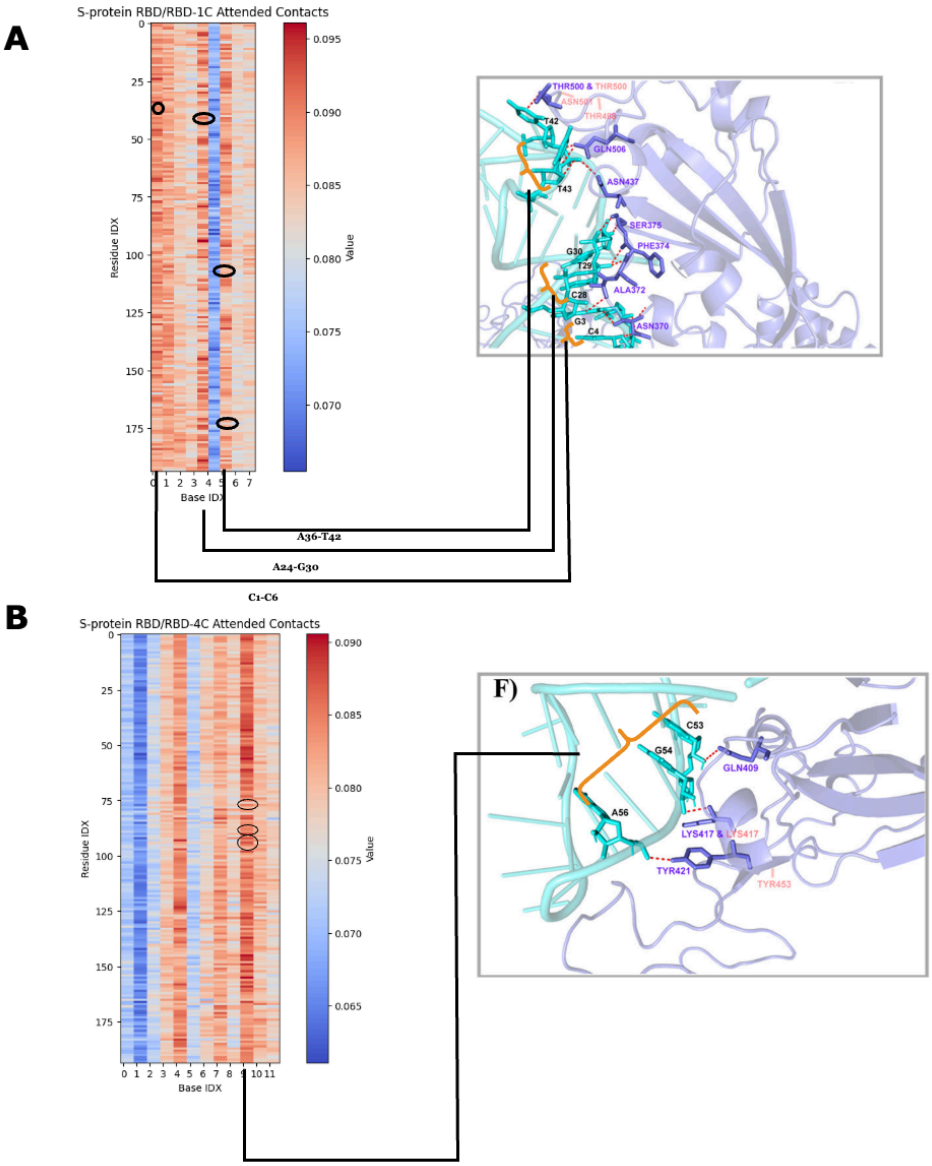

Figure S2. Connecting cross-attention maps to protein-aptamer binding interfaces. A) A cross attention map from one attention head generated from predicting RBD-1C binding to *SARS CoV-2* S-Protein RBD and the experimentally-determined binding interface. **B)** A cross attention map from one attention head generated from predicting RBD-4C binding to *SARS CoV-2* S-Protein RBD and the experimentally-determined binding interface.

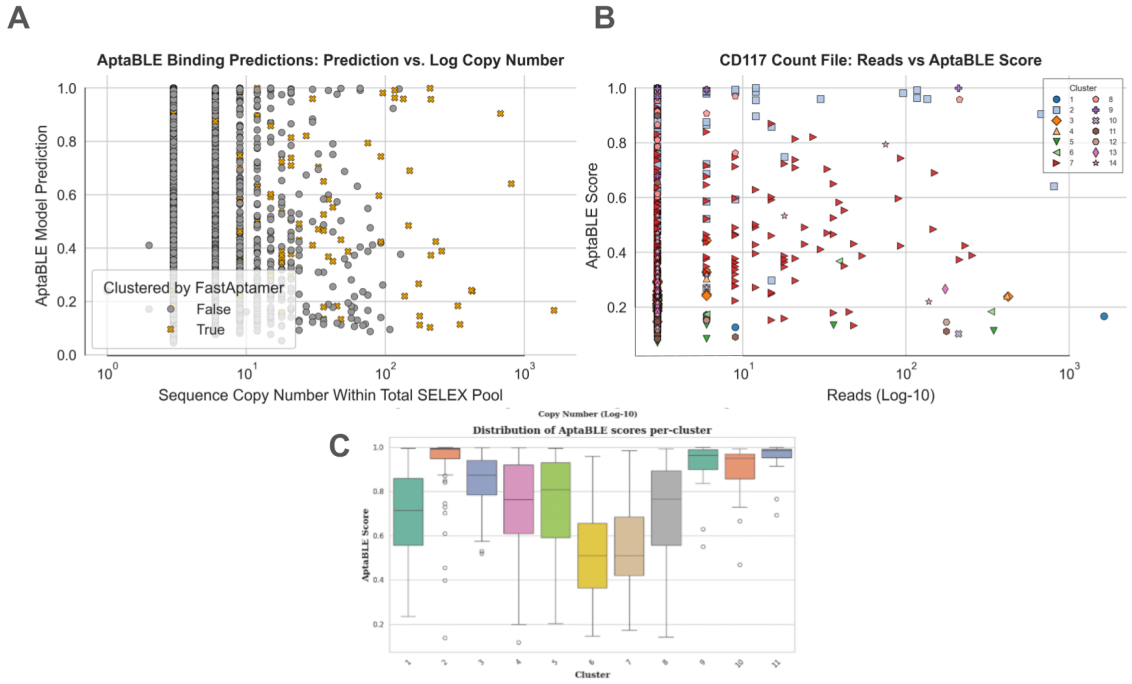

Figure S3. ssDNA Aptamers sequenced following nine rounds of SELEX selection for targeting hCD117. A) hCD117 aptamers from round 9 colored according to whether they were clustered or not. Clustering was sequence-based and performed via FastAptamer with a maximum edit distance set to 3. Red lines indicate 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)** 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. **C)** AptaBLE score distributions for each FastAptamer cluster.

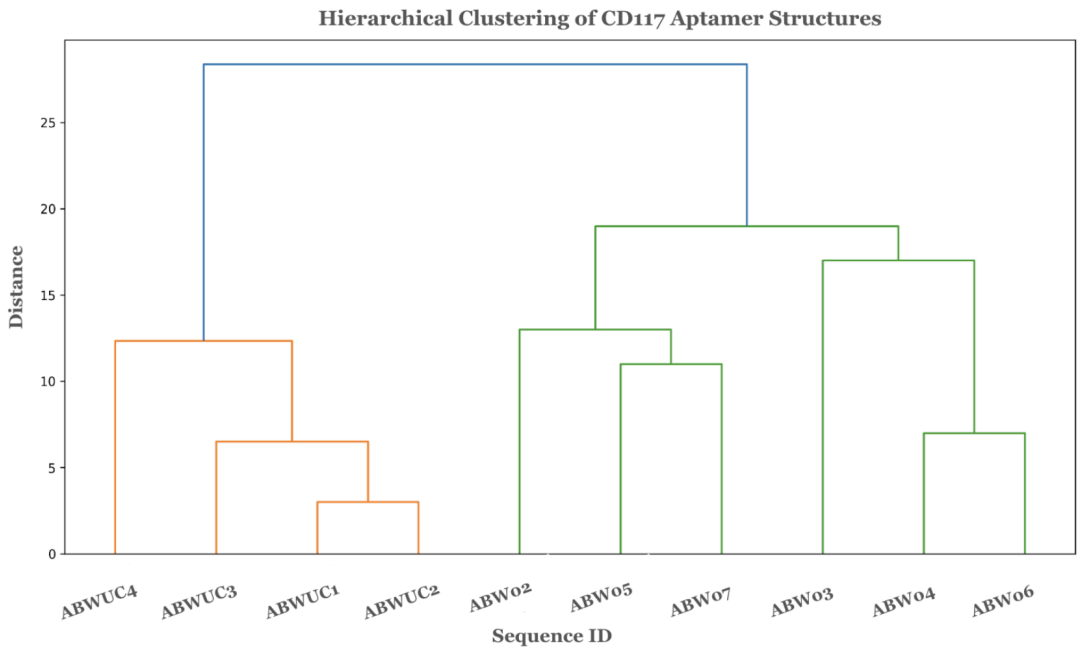

Figure S4. Hierarchical clustering of all tested aptamers according to their predicted secondary structures from dot-bracket representation. The 4 unclustered aptamers advanced for binding experiments are labeled as 'ABWUC*'.

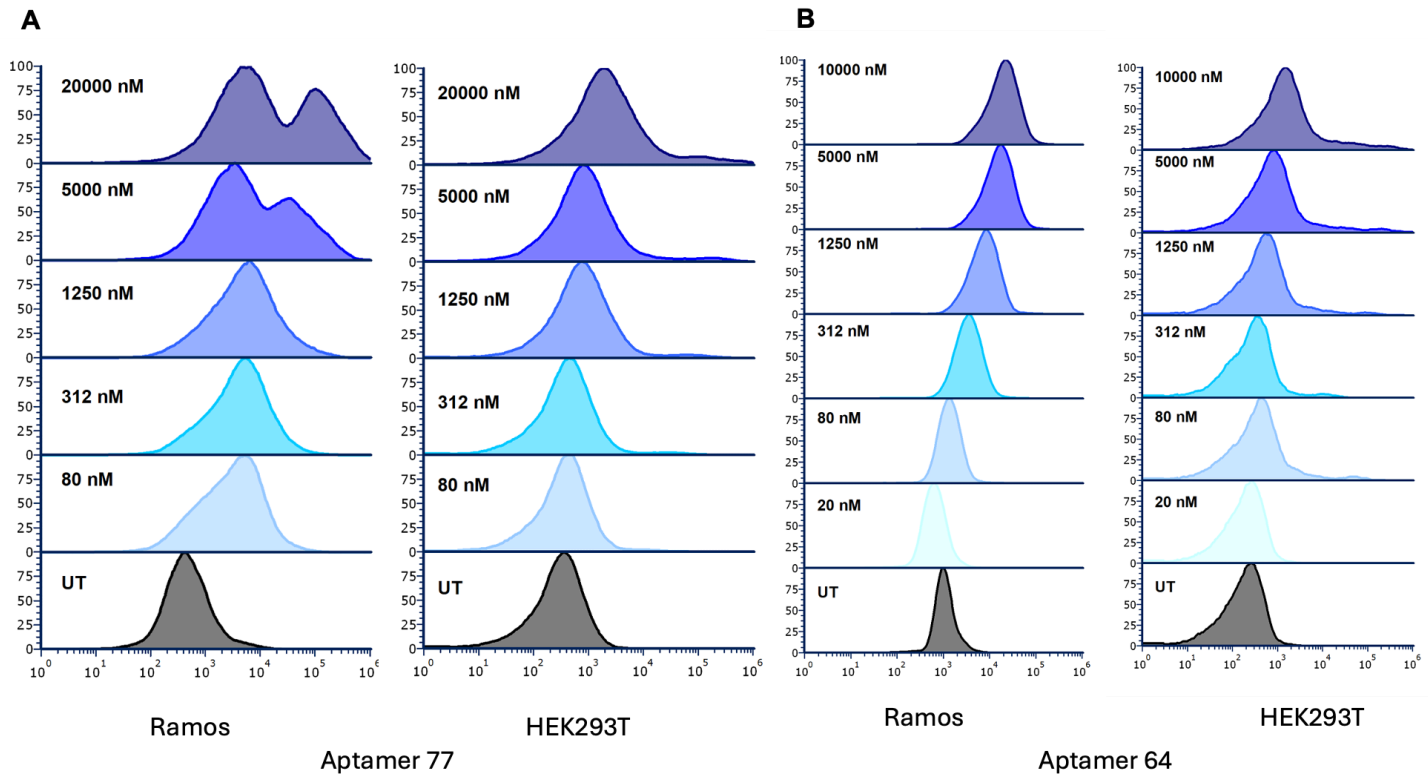

Figure S5. Concentration-dependent binding of Aptamers 77 and 64. Representative flow cytometry histograms showing fluorescence intensity distributions for **A)** Aptamer 77 and **B)** Aptamer 64 following incubation with Ramos (**left**) and HEK293T (**right**) cells at the indicated concentrations (80–20,000 nM). Untreated cells (UT) serve as negative controls.

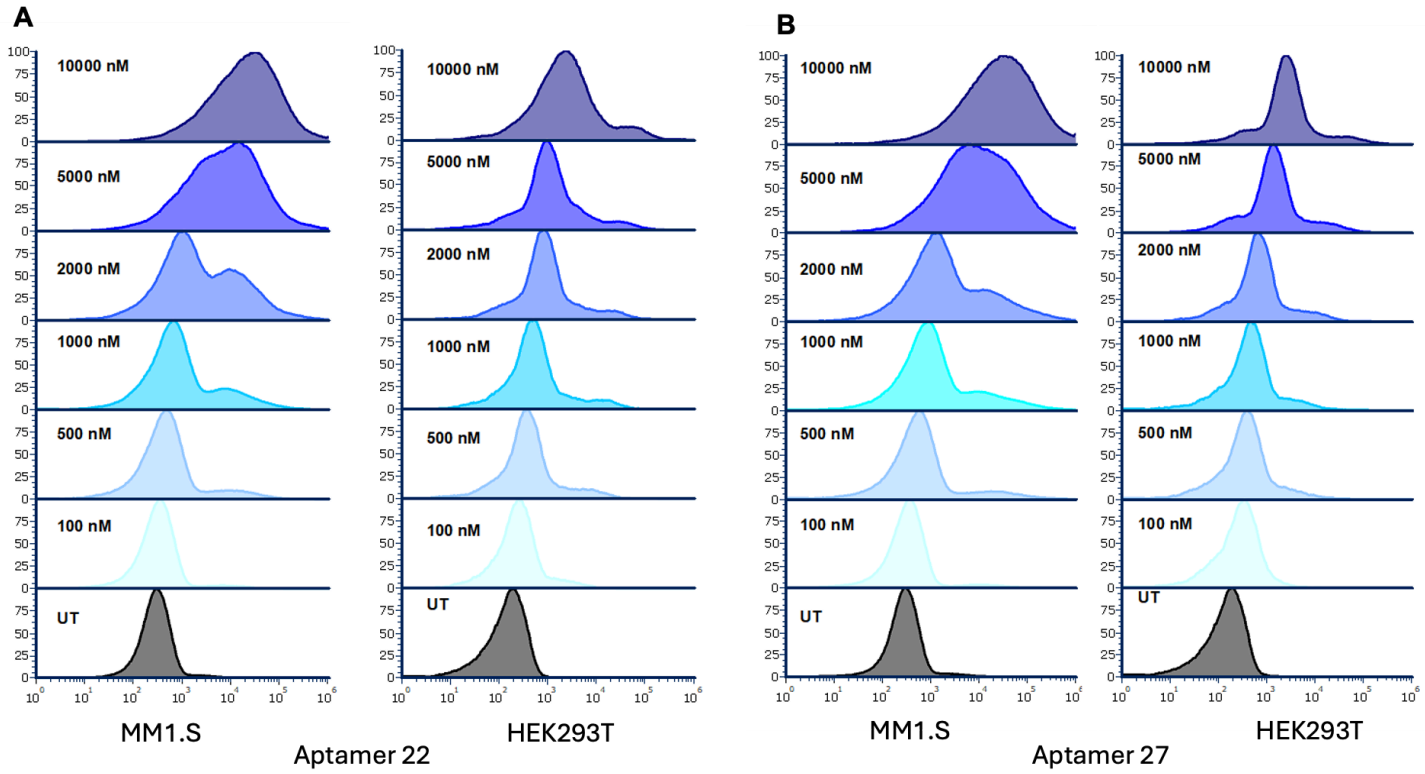

Figure S6. Concentration-dependent binding of Aptamers 22 and 27. Representative flow cytometry histograms showing fluorescence intensity distributions for **A)** Aptamer 22 and **B)** Aptamer 27 following incubation with MM1.S (**left**) and HEK293T (**right**) cells at the indicated concentrations (80–20,000 nM). Untreated cells (UT) serve as negative controls.