Molecular Therapy
Nucleic Acids
Original Article

DeepAptamer: Advancing high-affinity aptamer discovery with a hybrid deep learning model

Xin Yang,[1,2,3,4,9] Chi Ho Chan,[1,2,3,9] Shanshan Yao,[3,5] Hang Yin Chu,[3,5] Minchuan Lyu,[1,2,3] Ziqi Chen,[1,2,3] Huan Xiao,[3,5] Yuan Ma,[1,2,3] Sifan Yu,[1,2,3] Fangfei Li,[1,2,3] Jin Liu,[1,2,3] Luyao Wang,[1,2,3] Zongkang Zhang,[3,5] Bao-Ting Zhang,[3,5] Lu Zhang,[3,6] Aiping Lu,[1,2,3,7] Yaofeng Wang,[8] Ge Zhang,[1,2,3,7] and Yuanyuan Yu[1,2,3,7]

[1]Institute of Integrated Bioinformedicine and Translational Science, School of Chinese Medicine, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China; [2]Law Sau Fai Institute for Advancing Translational Medicine in Bone and Joint Diseases, School of Chinese Medicine, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China; [3]Guangdong-Hong Kong-Macao Greater Bay Area International Research Platform for Aptamer-based Translational Medicine and Drug Discovery, Kowloon, Hong Kong SAR, China; [4]Institute of Transdisciplinary Studies, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China; [5]School of Chinese Medicine, Chinese University of Hong Kong, New Territories, Hong Kong SAR, China; [6]Department of Computer Science, Faculty of Science, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China; [7]Institute of Systems Medicine and Health Sciences, School of Chinese Medicine, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China; [8]Centre for Regenerative Medicine and Health, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, New Territories, Hong Kong SAR, China

Oligonucleotide aptamers are typically identified through a rigorous and time-consuming process known as systematic evolution of ligands by exponential enrichment (SELEX), which requires 20 to 30 iterative rounds to eliminate non/weak binding sequences and enrich tight binding sequences with high affinity. Moreover, inherent experimental biases and non-specific interactions within SELEX could inadvertently exclude high-affinity candidates, leading to a high failure rate. To address these challenges, we proposed DeepAptamer for identifying high-affinity sequences from unenriched early SELEX rounds. As a hybrid neural network model combining convolutional neural networks and bidirectional long short-term memory, DeepAptamer integrated sequence composition and structural features to predict aptamer binding affinities and potential binding motifs. Trained on comprehensive SELEX data, DeepAptamer outperformed existing models in accuracy as substantiated by experimental evidence. More importantly, DeepAptamer effectively identified key nucleotides for target binding. DeepAptamer can efficiently identify high-affinity aptamers against various targets, enhancing its potential to discover promising sequences in initial screening stages and obviating the 20–30 iterative selection rounds required for full enrichment of selection pools. This represented a notable leap forward in aptamer technology, with broad implications for its application across a spectrum of selection targets.

INTRODUCTION

Aptamers are single-stranded oligonucleotides that bind to their targets with high sensitivity, selectivity, and affinity, rendering them ideal for diagnostic and therapeutic applications.[1,2] In the 1990s, the development of aptamers was accelerated by the introduction of the screening technology called systematic evolution of ligands by exponential enrichment (SELEX).[3] This screening method enabled the process of gradual accumulation and enrichment of the target-specific aptamers with repeated rounds of binding, partition, amplification, and regeneration. Subsequent sequencing of the SELEX pools using next generation sequencing (NGS) allowed for statistical analyses that unveil the dynamics within the aptamer populations across different SELEX stages, offering insights into the enrichment process of sequences (Figure S1).[4] Theoretically, sequences with relatively high enrichment are likely to be the high-affinity aptamers. These enriched sequences are then clustered by sequence similarity, with a few representative sequences chosen for experimental affinity characterizations, leaving the vast majority unexplored. Nonetheless, practical challenges such as sample loss, non-specific binding, amplification bias, and limitations in sequencing depth can affect the enrichment condition of high-affinity sequences. Consequently, it is desirable to establish a strategy that can swiftly and accurately identify high-affinity sequences within the extensive NGS datasets, regardless of whether they are enriched.

Aptamer-target binding relies on 3D conformational complementarity, making the prediction of aptamer affinity a complex endeavor

Received 2 August 2024; accepted 18 December 2024;
https://doi.org/10.1016/j.omtn.2024.102436.

[9]These authors contributed equally

Correspondence: Aiping Lu, Institute of Integrated Bioinformedicine and Translational Science, School of Chinese Medicine, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China.
E-mail: aipinglu@hkbu.edu.hk
Correspondence: Yaofeng Wang, Centre for Regenerative Medicine and Health, Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, New Territories, Hong Kong SAR, China.
E-mail: yaofeng.wang@hkisi-cas.org.hk
Correspondence: Ge Zhang, Institute of Integrated Bioinformedicine and Translational Science, School of Chinese Medicine, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China.
E-mail: zhangge@hkbu.edu.hk
Correspondence: Yuanyuan Yu, Institute of Integrated Bioinformedicine and Translational Science, School of Chinese Medicine, Hong Kong Baptist University, Kowloon, Hong Kong SAR, China.
E-mail: yuyuanyuan@hkbu.edu.hk

that must account for both the sequence and structural information of the aptamers. Current prediction approaches, including unsupervised clustering methods such as RaptRanker and SMART-Aptamer, clustered aptamers based on their predicted secondary structures, inferring affinity through comparison with high-affinity counterparts in database.[5–9] However, these methods were computationally intensive and may yield inaccuracies stemming from a preposition toward existing sequence data. On the other hand, supervised machine learning techniques utilized SELEX experiment data to construct models for affinity prediction. An improved SVM algorithm that combined sequence and structural features was proposed,[10] but it lacked the scalability required for large datasets. Deep learning approaches, recognized for their proficiency in extracting intricate features from voluminous data for sequence analysis are promising.[11–13] However, current deep learning models such as DeepBind and DeepSELEX, only focused on sequence information, neglecting the critical aspect of spatial conformation.[11,14] Therefore, to bridge this gap, it is essential to propose an advanced deep learning framework that can integrate both sequence and conformational information of aptamer sequences to improve the accuracy of aptamer affinity prediction.

In this paper, we introduce DeepAptamer, a hybrid deep learning framework that combined a convolutional neural network (CNN) and a bidirectional long short-term memory (BiLSTM) neural network. This architecture was engineered to harness the CNN's capability to extract distinctive features from aptamer sequence data and their conformational configurations, and BiLSTM was adept at detecting long-term associations within sequences. DeepAptamer was trained on an extensive dataset exceeding 300 GB of NGS data acquired from SELEX targeting three distinct targets. This model was then employed to identify potential high-affinity aptamers, evaluating both sequences enriched through SELEX and those overlooked in early SELEX rounds due to experimental bias. The efficacy of DeepAptamer was compared with existing models, with its efficacy substantiated through experimental validations. Furthermore, the key binding motifs were identified and validated. This research held the potential to considerably streamline the discovery process of potent aptamers by facilitating their identification in the early stages of SELEX, thereby obviating the need for fully enriched sequence pools. This approach could shorten the screening duration and increase the success rate of SELEX.

This study is crucial, as it addresses significant challenges in the discovery and characterization of high-affinity aptamers, which are essential for advancing diagnostic and therapeutic applications. By developing DeepAptamer, a hybrid deep learning framework that integrates both sequence and structural information, we aim to enhance the accuracy and efficiency of aptamer identification. This research not only provides a novel computational tool for researchers in molecular biology and bioinformatics but also benefits pharmaceutical companies and biotechnology firms that rely on aptamer technology for drug development and targeted therapies. Ultimately, this study has the potential to transform the aptamer discovery process, making it more efficient and accessible to a diverse range of stakeholders in science and medicine.

RESULTS

Architecture of DeepAptamer

DeepAptamer was a deep learning framework that integrated CNN and BiLSTM networks (Figure 1). DeepAptamer took advantage of both one-hot encoding and DNA shape features to capture the sequence and conformational information of aptamers. Additionally, DeepAptamer employed variational autoencoders (VAE) to identify the key nucleotides that contribute to aptamer-target binding. We applied DeepAptamer to datasets derived from multiple rounds of SELEX targeting an array of proteins. The protocol involved the following steps: (1) conducting SELEX to identify aptamers against a specific target protein, and sequencing the single-stranded DNA (ssDNA) pools from different SELEX rounds; (2) counting and analyzing the sequence enrichment across the rounds; (3) encoding the DNA sequences into matrices using one-hot coding and DNA conformational information; (4) curating the encoded datasets into training, test, and validation sets; (5) modeling the aptamer-target affinity using the framework combining CNN and BiLSTM to extract both sequence and conformation features and capture long-term dependencies within nucleotides; (6) outputting the aptamer affinity scores and integrating them into a fully concatenated layer; and (7) loading both data types into the VAE feature extraction program to identify the key nucleotides that are crucial for the aptamer-target binding. This protocol constructed a comprehensive pathway, from SELEX to the identification of high-affinity aptamer candidates, leveraging the advanced analytical prowess of DeepAptamer.

DeepAptamer consisted of input sequencing data analysis, model training, output data evaluation, and interpretability analysis (Figure 1). Following SELEX, the selection pool from each round was sequenced by NGS. The quality of the sequencing was assessed to exclude sequences with a high error rate. The enrichment levels of error-free sequences were statistically analyzed using reads per million (RPM), with those exhibiting high enrichment chosen as the input data. These sequences were then converted into matrices using one-hot encoding, while conformational information was extracted by DNA shape software using sliding window processing. The input data were randomly distributed into training, testing, and validation sets. The test set was used to refine and optimize the model during the training process. The validation set was used to evaluate the final performance post-training and turning. To ensure unbiased learning, the number of positive samples was equalized with that of negative samples. Upon completion of the model training by CNN and BiLSTM, all features were integrated within two fully connected layers for predicting the affinity values of the sequences. These one-hot vectors were then loaded into the variational autoencoder (VAE) module, for extracting distinctive features from DNA sequences and identifying key binding nucleotides in the sequences.[15] Subsequently, binding sites and binding patterns associated with the high-affinity sequences were then identified and characterized.

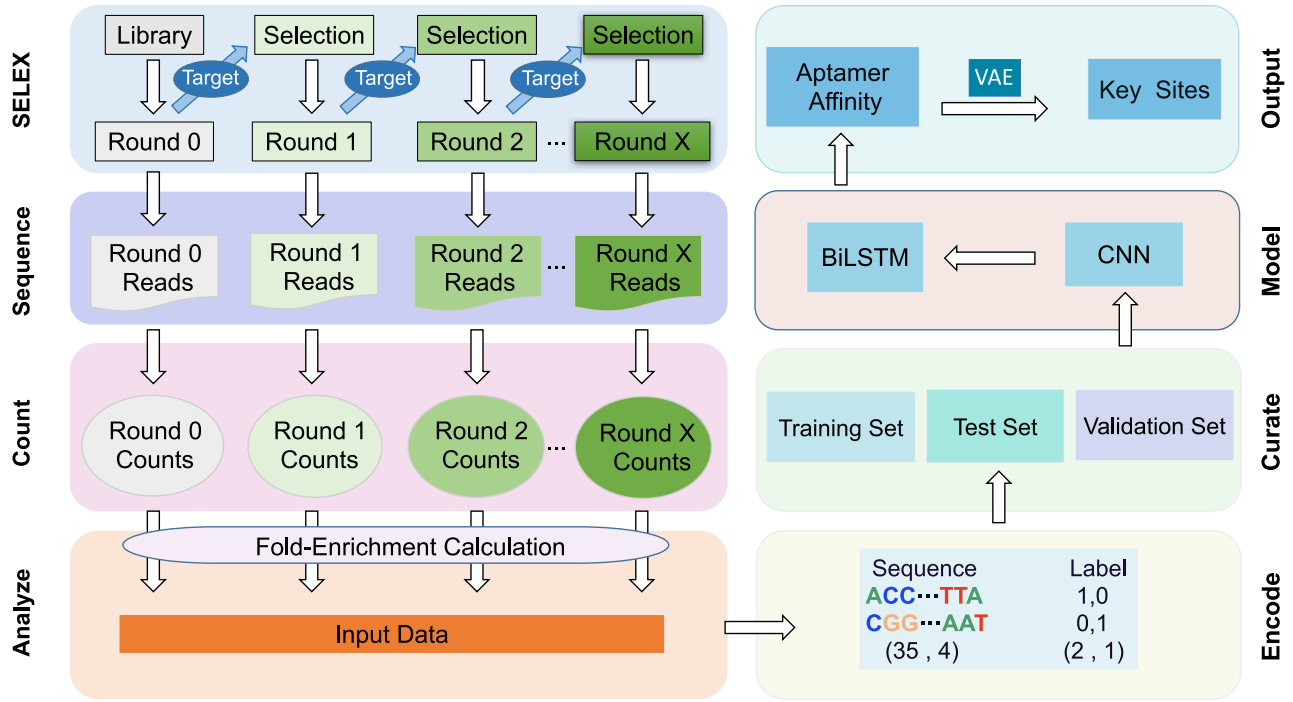


Figure 1. General architecture of DeepAptamer

The flowchart illustrates aptamer selection, sequencing, encoding, model training, and result prediction steps involved in the DeepAptamer pipeline.

Statistical analysis of NGS data after SELEX

DeepAptamer was trained with NGS datasets derived from SELEX targeting three distinct target proteins, each exhibiting a range of aptamer binding affinities: (1) B cell maturation antigen (BCMA), an emerging target for immunotherapy in multiple myeloma[16]; (2) connective tissue growth factor (CTGF), a member of the CCN protein family of secreted proteins with roles in cell proliferation, migration, adhesion, wound healing, and angiogenesis[17]; and (3) Dickkopf-1 (DKK1), a marker of poor prognosis in a variety of cancers.[18]

Each SELEX experiment yielded six to nine sequencing samples, corresponding to the ssDNA pools obtained from individual SELEX rounds (Figure 2). Following extraction of valid sequences from the raw NGS data, sequencing quality was assessed by calculating error rates and unmatched rates (Figure 2A). The error rate referred to the sequencing error by calibrating with the conserved regions in the initial SELEX library. The unmatched rate referred to the sequence differences identified from the two conserved primer regions. It was found that most SELEX pools had error rates at ∼7%, which was reasonable, while DK-26 (the 26th SELEX round against DKK1), DK-28, and DK-30 had relatively higher error rates (∼26%, ∼28%, and ∼34%, respectively), which might be attributed to the fact that the mutations in PCR accumulated through repeated SELEX rounds, thus resulting in a relatively high error rate. Similarly, the unmatched rates in the SELEX pools were relatively low (∼3%), except for DK-28 and DK-30 (∼18% and ∼20%, respectively). Sequences with these errors were then excluded. Despite these outliers, the overall sequencing quality was deemed sufficiently high for subsequent analysis.

The enrichment levels of the sequences within each SELEX pool were analyzed by the ratio of the number of different sequences in the total number of sequences (Figure 2B). The ratio in the initial ssDNA library before SELEX (IN_1) was 0.896. For CTGF-targeted SELEX, this ratio decreased gradually through SELEX rounds and dropped down to 0.155 at round 20, indicating an effective convergence of sequence types. For DKK1, the ratio remained almost unchanged until round 20, and dropped down to 0.020 at round 30. For BCMA, the ratio decreased significantly in the first several rounds, dropped down to 0.031 at round 6 and remained nearly unchanged after that. Therefore, considering the enrichment condition, the sequences from round 20 of CTGF, round 30 of DKK1, and round 6 of BCMA were chosen as the input data for subsequent model training.

The input sequences from round 20 of CTGF, round 30 of DKK1, and round 6 of BCMA were systematically categorized to form the positive and negative datasets. The occurrence of sequences was counted and ordered based on reads per million (RPM). Subsequently, the distribution of occurrence for the top 20,000 ranked sequences from each SELEX round was graphically represented (Figure 2C). According to the occurrence distribution patterns, the top 7,500 sequences for CTGF, 4,000 sequences for DKK1, and 2,000 sequences for

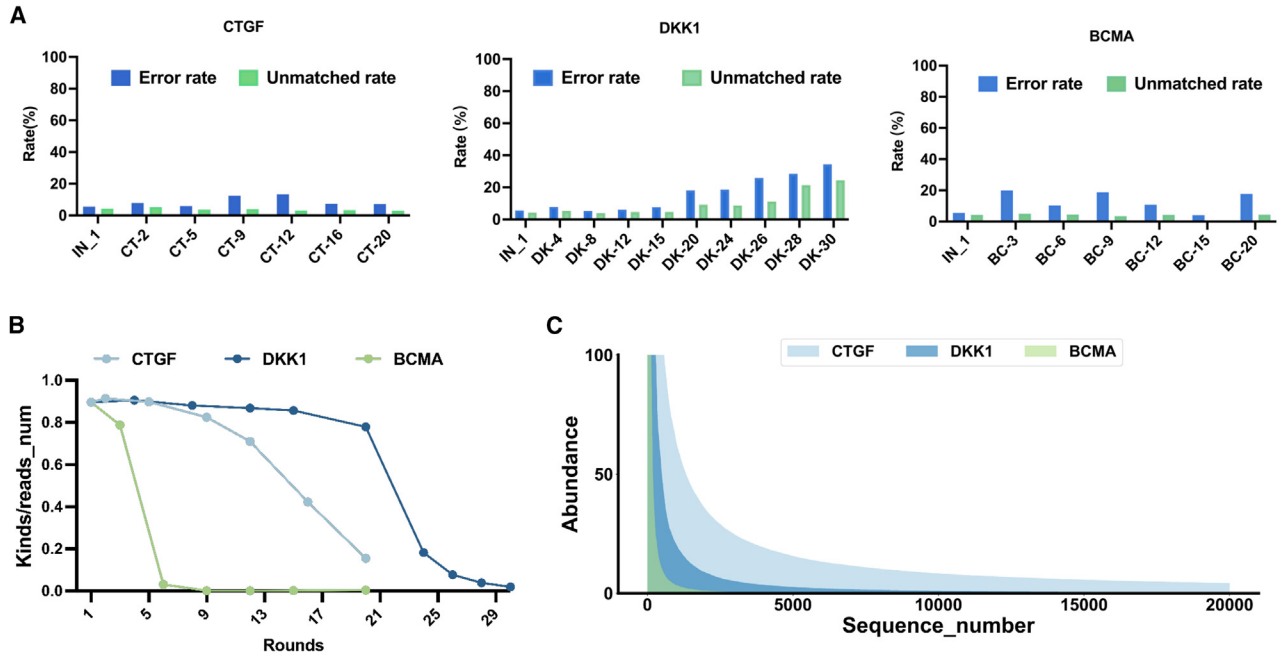


Figure 2. Sequence enrichment analysis from SELEX pools

(A) Sequencing quality assessment, including error and unmatched rates. (B) Enrichment analysis through SELEX rounds, showing sequence diversity reduction. (C) Occurrence distribution of the top 20,000 sequences ranked by reads per million (RPM).

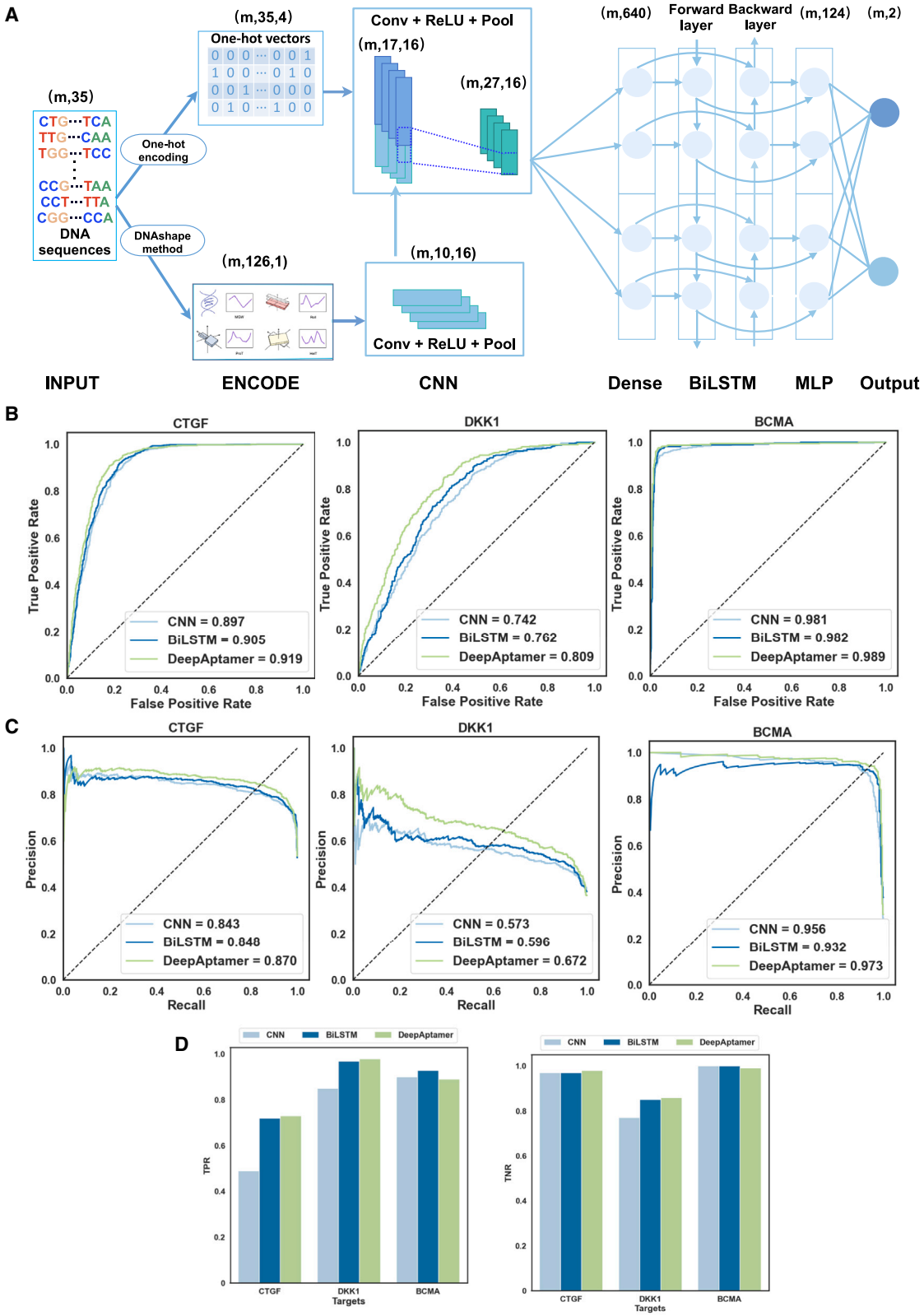
BCMA were designed as the positive dataset. For the negative dataset, 150,000 sequences in each SELEX were selected from those present in the initial library but absent in subsequent rounds.

The integrated DeepAptamer model demonstrated superior performance compared with the individual CNN or BiLSTM model

DeepAptamer was designed as a deep neural network model to predict the affinity of aptamer sequences from NGS data (Figure 3A). The model comprised four primary components: an encoding layer, a convolutional layer, a BiLSTM layer, and a multilayer perceptron layer. The encoding layer could transform the input DNA sequences (length $\leq$35 base pairs [bp]) into one-hot vectors and extract DNA shape features, such as minor groove width (MGW), propeller twist (ProT), helical twist (HelT), and roll, using the DNAshape method.[19] The convolutional layer applied two-dimensional convolutional filters and rectified linear units (ReLU) to the one-hot vectors and the DNA shape features separately, and then performed max-pooling to reduce the dimensionality of the inputs. The BiLSTM layer captured the long-term dependencies among the nucleotides by processing the outputs of the convolutional layer in both forward and backward directions. The multilayer perceptron layer integrated the features learned by the BiLSTM layer and output a two-dimensional vector that represented the predicted affinity of the aptamer sequence.

For both the positive and negative datasets, 72% of the sequences were randomly selected as the training dataset, 18% of the sequences were selected as the validation dataset, while the remaining 10% were used as the testing dataset. DeepAptamer was then trained with the training and validation subsets, providing four key metrics: train loss, validation loss, train accuracy, and validation accuracy. These metrics are crucial for adjusting the model's hyperparameters. Following training, DeepAptamer was utilized to predict the affinity of sequences within the corresponding testing dataset.

To comprehensively evaluate the performance of DeepAptamer, we calculated both the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) as primary metrics. The AUROC provides insight into the model's ability to distinguish between classes, while the AUPRC is particularly valuable for assessing performance with imbalanced datasets, focusing on the quality of positive predictions. We benchmarked DeepAptamer against individual CNN or BiLSTM models to evaluate its accuracy in affinity prediction. The AUROC values (Figure 3B) for CTGF were 0.897 for CNN, 0.905 for BiLSTM, and 0.919 for DeepAptamer. For DKK1, the AUROC values were 0.742 for CNN, 0.762 for BiLSTM, and 0.809 for DeepAptamer. For BCMA, the AUROC values were 0.981 for CNN, 0.982 for BiLSTM, and 0.989 for DeepAptamer. Similarly, the AUPRC values (Figure 3C) showed that DeepAptamer outperformed the individual models: for CTGF, AUPRC was 0.843 for CNN, 0.848 for BiLSTM, and 0.870 for DeepAptamer; for DKK1, AUPRC was 0.573 for CNN, 0.596 for BiLSTM, and 0.672 for DeepAptamer; for BCMA, AUPRC was 0.956 for CNN, 0.932 for BiLSTM, and 0.973 for DeepAptamer. These results confirmed that DeepAptamer consistently outperforms both CNN and BiLSTM models across different metrics, underscoring its enhanced predictive capabilities.



(legend on next page)

To further assess the performance of different modules, the confusion matrix was employed to determine the true positive rate (TPR) for the predicted results on the testing dataset (Figures 3D and S2). For CTGF, the TPRs were 0.49 for CNN, 0.72 for BiLSTM, and 0.73 for DeepAptamer. For DKK1, the TPRs were 0.85 for CNN, 0.97 for BiLSTM, and 0.98 for DeepAptamer. For BCMA, the TPRs were 0.90 for CNN, 0.93 for BiLSTM, and 0.89 for DeepAptamer. Furthermore, the true negative rate (TNR) of three modules were analyzed. For CTGF, the TNRs were 0.97 for CNN, 0.97 for BiLSTM, and 0.98 for DeepAptamer. For DKK1, the TNRs were 0.77 for CNN, 0.85 for BiLSTM, and 0.86 for DeepAptamer. For BCMA, the TNRs were 1.00 for CNN, 1.00 for BiLSTM, and 0.99 for DeepAptamer. The TNR and TPR data revealed no significant differences across the three modules.

As both local sequence features and structural conformations were incorporated into the training process, the combinational model DeepAptamer was deemed more appropriate for this study.

DeepApatamer demonstrated superior performance compared with current protein-DNA binding prediction models

DeepBind and DeepSELEX were deep learning models used for predicting the binding between transcription factors and genomic DNAs, and SVM was a machine learning model for predicting the binding between proteins and aptamers.[10,11,14] For comparison, SVM, DeepSELEX, and DeepBind were trained and tested with the same datasets (Figure 4). It was found that DeepAptamer exhibited higher AUROCs (0.963 for CTGF, 0.861 for DKK1, and 0.996 for BCMA) compared with SVM (0.431 for CTGF, 0.608 for DKK1, and 0.896 for BCMA), DeepBind (0.864 for CTGF, 0.761 for DKK1, and 0.984 for BCMA), and DeepSELEX (0.864 for CTGF, 0.649 for DKK1, and 0.996 for BCMA) (Figures 4A and 4B).

In addition, for the TPR and false positive case rate (FPR) values, DeepSELEX and DeepBind showed inferior performance compared with DeepAptamer for all three targets (TPRs were 0.00 and TNRs were 1.00) (Figure S3), suggesting that deep learning models used for predicting the binding of transcription factors and genomic double-stranded DNAs were not suitable for predicting aptamer-protein binding. We further validated the performance of DeepAptamer using external datasets of SELEX data targeting streptavidin.[20] F1 Scores was employed as the primary metric to reflect the model's ability to balance true positive predictions. It was found that DeepAptamer demonstrated superior performance compared with DeepBind and DeepSELEX models on this external dataset (Figure S4).

Pre-trained DeepAptamer could effectively identify the potential high-affinity aptamer sequences in early SELEX rounds

Theoretically, high-affinity sequences would be enriched through SELEX. However, due to limitations in experiments such as PCR bias, sequencing bias, and non-specific bindings, potential sequences with high affinity may not be enriched, leading to a high missing rate of potential candidates. To address this issue, the pre-trained DeepAptamer was used to predict and identify the potential high-affinity candidates from the sequences with low enrichment. Sequences that were present in early rounds but not enriched in the selected final input rounds (round 20 of CTGF, round 30 of DKK1, and round 6 of BCMA) were extracted. The binding affinity of these sequences was predicted by DeepAptamer, and the top 100 sequences were identified. The distribution of probability values of the top 100 sequences predicted by DeepAptamer from unenriched early rounds and top 100 sequences with high enrichment from the enriched rounds were plotted (Figure 5A). For CTGF, DKK1, and BCMA, the affinity of the top 100 sequences predicted by DeepAptamer from unenriched early rounds were higher (lower mean Kd values) than that of the top 100 sequences with high enrichment (higher mean Kd values).

To validate the predicted results, 10 sequences from the top 100 sequences predicted by DeepAptamer from unenriched early rounds and 10 sequences with high enrichment from the enriched rounds were randomly selected, and the binding affinity of these sequences was determined by Biolayer Interferometry (BLI) analysis. It was found that the mean Kd value of predicted sequences was significantly lower than that of the enriched sequences ($p = 0.020$), demonstrating a higher binding affinity of the predicted sequences (Figures 5B and 5C). These data suggested that the pre-trained DeepAptamer could effectively predict and identify potential high-affinity aptamers in the early SELEX rounds, without the requirement of enriched pools.

Pre-trained DeepAptamer could identify key nucleotides essential for target binding

DeepAptamer was further employed to identify the key nucleotides of the candidate sequences essential for target binding. The top 4,000 sequences, as ranked by high-affinity probability values predicted by DeepAptamer, served as the input data for VAE. The VAE reconstructed and processed these sequences to interpret important binding features (Figure 6A). The distribution of the reconstructed hidden layer for low-enrichment sequences was plotted, with different colored dots representing different degrees of high-affinity probability values (Figure 6B). For CTGF, the distribution of probability values for the input sequences was highly concentrated, predominantly centered at around 0.98, indicating the presence of a strong consensus within the high-affinity sequences for CTGF binding. In contrast, the distribution for DKK1 was broader, ranging from 0.7 to 0.95, suggesting a more diverse set of sequences contributing to high-affinity binding. For BCMA, the input sequences exhibited even broader range in probability value distribution, extending from 0.1 to 0.7, with most sequences exhibiting values above 0.6.

Figure 3. DeepAptamer model training

(A) Neural network structure combining CNN and BiLSTM to predict aptamer affinity. (B) AUROC values for different network configurations. (C) AUPRC values for different network confirgurations. (D) Quantitative TPR and TNR bar charts for DeepAptamer and individual models.

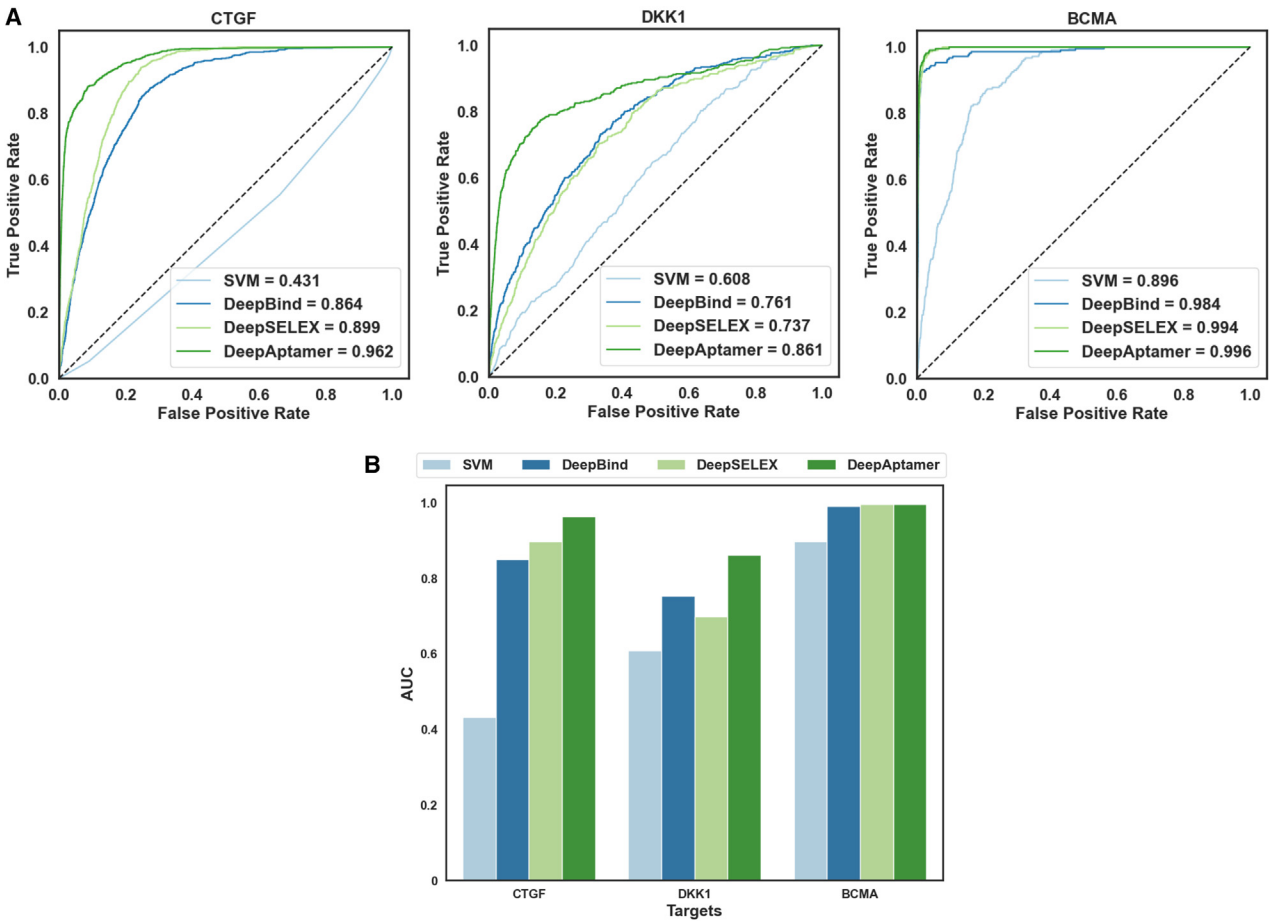


Figure 4. Performance comparison of DeepAptamer against SVM, DeepBind, and DeepSELEX on different targets
(A) AUROC scores for all models. (B) Quantitative AUC bar charts for each model.

This extensive distribution implied a more complicated binding landscape for BCMA.

Subsequently, the hidden layer features of these sequences were clustered using the Gaussian mixture model (GMM) (Figure 6C). It was observed that sequences with similar affinity probability values were grouped into the same clusters. Therefore, high-affinity sequences from the cluster with higher probability values were selected for analyzing the binding motif features. It was found that CATCA motif for CTGF, GGTTGG-NNN-GGTTGG motif for DKK1, and AATGCAG motif for BCMA were the common motif features among the high-affinity sequences (Figure 6D). Furthermore, experimental validation through BLI confirmed that the sequences containing the corresponding motifs had high binding affinities to the corresponding targets among the top-ranked sequences (Table S1). To investigate the role of the binding motifs, the identified nucleotides were mutated. The predicted secondary structures indicated that the structure was stabilized by the common binding motif, in conjugation with the two loops formed by the conserved primer region at both termini (Figure S5), while mutations on these nucleotides disrupted the original loops and changed the secondary structure. Furthermore, mutations within the core motif resulted in an increase in the minimum free energy (MFE) for the secondary structure, indicating a loss of structure stability (Figure S6). To further validate the significance of the predicted core motifs, we designed aptamers with altered nucleotide sequences in regions outside of the predicted motifs. Specifically, the core motifs identified by DeepAptamer were mutated to poly-T sequences, while sequences outside the motifs were mutated randomly to serve as control groups. BLI analysis revealed that mutations within the core motifs significantly reduced the binding affinity of the aptamers to their protein targets, whereas mutations outside the core motifs had no effect on binding (Figure S7). These results underscore the critical role of the predicted core motifs in aptamer-protein interactions and support the robustness of DeepAptamer's predictive capabilities.

Moreover, the biological activity of aptamer candidates, both with and without the core motif, was determined. Two DKK1 aptamer

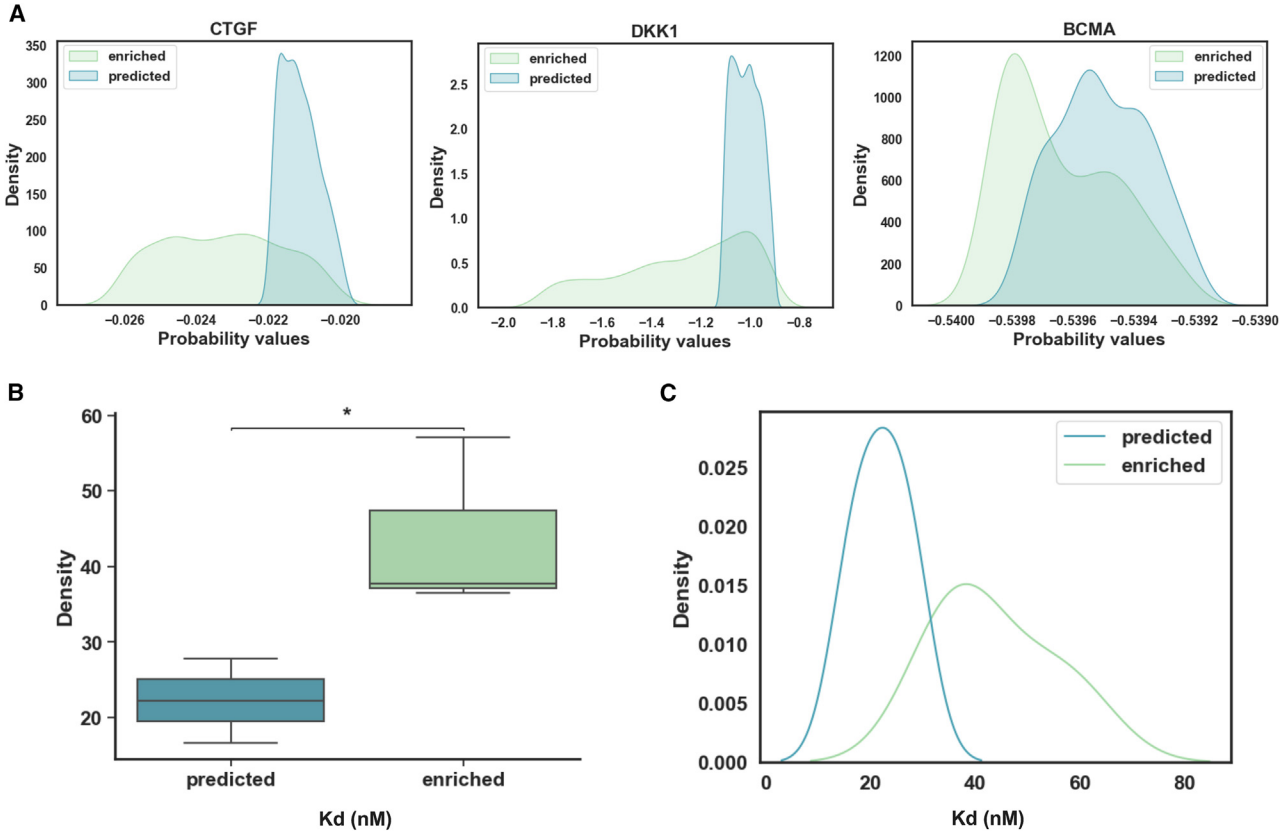


Figure 5. Validation of predicted high-affinity sequences

(A) Probability distribution of predicted vs. enriched sequences. (B) Binding affinity of predicted and enriched sequences via BLI. A paired *t*-test was performed to determine the difference between groups. $p<0.05$. (C) Affinity-level distribution between predicted and enriched sequences.

candidates were selected for this comparison. Candidate AptDK-20 was enriched in SELEX but lacked the core motif (Kd = 25.8 nM, Figure S8A). Candidate AptDK-5, which contained the core motif, was not enriched in the SELEX but predicted by DeepAptamer (Kd = 2.55 nM, Figure S8A). The results revealed that AptDK-5, which contained the core motif, exhibited significantly higher biological activity compared with AptDK-20, which lacked the core motif (Figure S8B). This finding implied that the presence of the core motif was intricately linked to the enhanced biological function of the aptamers.

Overall, these findings highlighted the importance of the identified core motifs by DeepAptamer in both binding affinity and structural stability. They provide essential insights for the subsequent modifications of these factors, which are pivotal for the advancement of drug discovery and development.

DISCUSSION

Recent advancements in computational methods for predicting aptamer-protein interactions have significantly enhanced the speed and efficiency of aptamer design. For example, the deep learning model AptaNet achieved an impressive accuracy of 91.38% in its predictions, effectively addressing the class imbalance commonly found in these datasets.[21] Concurrently, Zhang et al. demonstrated the effectiveness of ensemble classifiers that integrate multiple feature extraction methods, achieving a balanced sensitivity of 0.753 and specificity of 0.725.[22] The incorporation of *in silico* techniques into the SELEX process has notably accelerated aptamer selection, reducing both time and costs associated with traditional methodologies.[23] Additionally, Fang et al. highlighted the importance of diverse feature representations, such as k-mer and reverse complement k-mer frequencies, in enhancing the robustness of interaction predictions.[24] Collectively, these findings underscore the value of leveraging computational approaches, which not only streamline the development of aptamers but also broaden their potential applications in diagnostics and therapeutics.[25–27]

However, despite these advancements, existing models possess limitations that hinder their broader application in aptamer discovery. Currently, no models integrate SELEX and NGS technologies for deep learning-based affinity prediction, apart from earlier approaches utilizing SVM. Most available tools rely on small sample databases with limited aptamer counts (approximately 1,000) or focus on

generative modeling and data acquisition methods other than SELEX-seq technology. Given this context, we chose to compare DeepAptamer with DeepBind and DeepSELEX, both of which represent well-established algorithms in the realm of deep learning applied to NGS sequence data. Their proven methodologies in modeling biological interactions, particularly with DNA, provide a suitable benchmark for demonstrating the capabilities of DeepAptamer.

Building on these advancements, we trained the DeepAptamer model using SELEX-seq data targeting three specific targets. DeepAptamer integrates a sophisticated deep learning framework that utilizes multiple neural networks to mitigate noise interference inherent in raw data. It predicts binding affinity between aptamer sequences and proteins with high accuracy, while also effectively identifying key motifs essential for target binding.

The hybrid architecture of DeepAptamer, which synergizes CNN and BiLSTM networks, proved instrumental in achieving promising performance across all three targets. While CNNs are adept at identifying local patterns, they often struggle to capture long-range dependencies due to limited receptive fields.[28] Such dependencies are crucial in DNA sequences, where functional interactions may involve distant nucleotides. This limitation emphasizes the need for sequence-specific structural features beyond the capabilities of CNN alone. The BiLSTM component complements this by modeling long-term dependencies among nucleotides, thereby enhancing the insights garnered by the CNN's convolutional filters.[29] By leveraging the strengths of both CNN and BiLSTM, DeepAptamer effectively employs deep multimodal learning to integrate these distinct data modalities.

Across various datasets, DeepAptamer consistently outperformed alternative models, including DeepBind, DeepSELEX, and SVM. While most models demonstrated relatively strong results in terms of receiver operating characteristic (ROC) curves and TNR, DeepAptamer distinguished itself with a markedly superior TPR. This exceptional performance can be attributed to the intricate architecture of DeepAptamer, which merged CNN and BiLSTM to capture complex features indicative of nucleotide binding affinity. Other models lacked the sophistication to identify these features effectively. Although DeepSELEX performed commendably on the BCMA dataset, it still fell short compared with DeepAptamer's performance. These findings highlighted the effectiveness of DeepAptamer's multimodal deep learning strategy, which enabled superior predictive performance by mastering joint representations of various data modalities and identifying shape variants and key nucleotide binding sites. In contrast, other existing models, such as RaptGen and MLPD, did not account for DNA structural information during feature learning, which limited their ability to predict high-affinity aptamers effectively.[30,31] We also experimentally validated the binding affinities (Kd values) of the predicted aptamer sequences using BLI, confirming DeepAptamer's performance in classifying and predicting high-binding aptamer sequences.

Furthermore, we validated the performance of DeepAptamer using an independent third-party dataset, specifically the top 100 sequences from SELEX-seq targeting streptavidin, as reported by Tatjana et al.[20] This dataset comprises exclusively positive aptamer sequences exhibiting protein binding affinity. DeepAptamer achieved a commendable F1 score of 0.70 on this dataset, while both DeepBind and DeepSELEX underperformed in the same conditions. The superior performance of DeepAptamer can be attributed to its design, which incorporates strategies to address class imbalance between positive and negative samples during pre-training. In contrast, DeepBind and DeepSELEX were trained on datasets with significant class imbalances without addressing this issue, which likely contributed to their suboptimal performance. Consequently, DeepAptamer exhibits enhanced precision in identifying true positive sequences despite the inherent class imbalance in SELEX-seq datasets.

Another notable finding was DeepAptamer's ability to identify high-affinity sequences that were overlooked during SELEX enrichment but present in the unenriched early rounds. As validated by BLI, the binding affinities of these sequences were found to be higher than those of sequences identified with high enrichment during later rounds. This suggests that DeepAptamer can effectively identify candidates that were overlooked due to biases in the SELEX process, such as PCR bias, sequencing bias, and non-specific bindings. Consequently, DeepAptamer has the potential to significantly improve the success rate of SELEX while reducing the need for additional rounds, ultimately saving valuable time and resources.

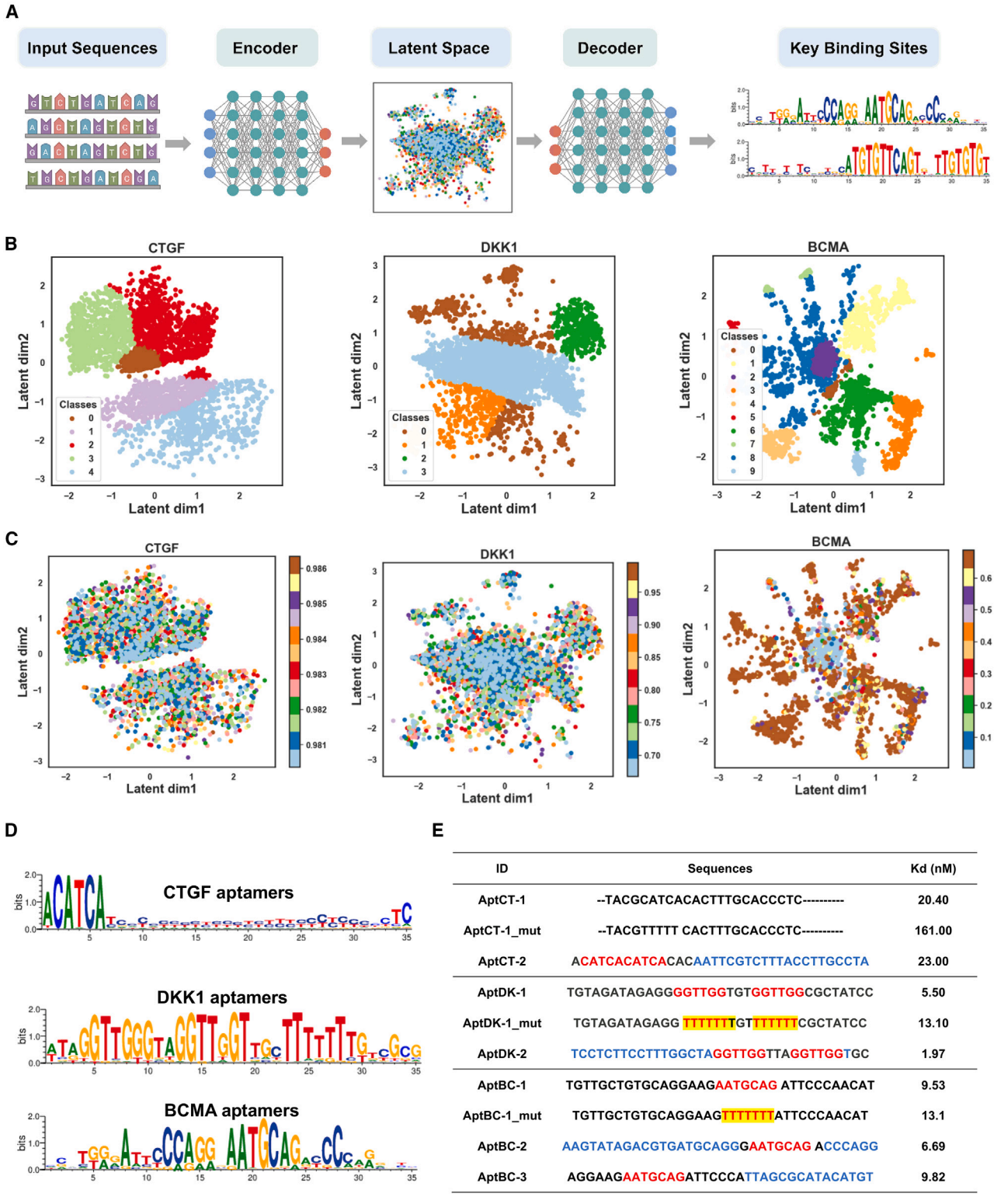
Furthermore, DeepAptamer demonstrated the capability to predict key nucleotides within core motifs that are crucial for target binding. These core motifs are often essential for high-affinity interactions, and alterations to these pivotal nucleotides have been linked to diminished binding affinity and disrupted structural integrity.

In conclusion, DeepAptamer represented an innovative resource for identifying aptamers with high binding affinity from early unenriched SELEX rounds. By circumventing experimental biases inherent in the SELEX process, DeepAptamer effectively uncovered high-affinity candidates that might otherwise be missed. This capability has significant implications for developing more efficient aptamer-based applications in therapeutic and diagnostic fields, enhancing the potential of aptamers as versatile molecular tools.

MATERIALS AND METHODS

Aptamer selection by SELEX

SELEX was performed according to our established methodologies.[32,33] His6-tagged proteins were immobilized on NTA magnetic beads at 4°C for 1 h. A HPLC purified ssDNA library, which contained a central random region (35 nt) and a conserved region for primer binding (18 nt) at each end and was denatured at 95°C for 10 min, rapidly cooled to 0°C for 10 min, and then incubated with immobilized proteins for 0.5–1 h at room temperature (RT) in a total incubation volume of 1 mL. BSA (10 μg for the first round and 1 μg for subsequent rounds) was added into the incubation system to avoid non-specific binding. After incubation, unbound sequences were removed with 500 μL washing buffer for three times. Bound



E

ID	Sequences	Kd (nM)
AptCT-1	--TACGCATCACACTTTGCACCCTC----------	20.40
AptCT-1_mut	--TACGTTTTT CACTTTGCACCCTC----------	161.00
AptCT-2	ACATCACATCACACAATTCGTCTTTACCTTGCCTA	23.00
AptDK-1	TGTAGATAGAGGGGTTGGTGTGGTTGGCGCTATCC	5.50
AptDK-1_mut	TGTAGATAGAGG TTTTTTTGTTTTTTCGCTATCC	13.10
AptDK-2	TCCTCTTCCTTTGGCTAGGTTGGTTAGGTTGGTGC	1.97
AptBC-1	TGTTGCTGTGCAGGAAGAATGCAG ATTCCCAACAT	9.53
AptBC-1_mut	TGTTGCTGTGCAGGAAGTTTTTTTATTCCCAACAT	13.1
AptBC-2	AAGTATAGACGTGATGCAGGGAATGCAG ACCCAGG	6.69
AptBC-3	AGGAAGAATGCAGATTCCCATTAGCGCATACATGT	9.82

Figure 6. Identification of key binding motifs

(A) Representation of ssDNA sequences using VAE embeddings. (B) Reconstructed hidden layer distribution for low-enrichment sequences. (C) GMM cluster visualization of the reconstructed hidden layer. (D) Sequence logos derived from VAE decoding. (E) Binding affinity validation of aptamer sequences with mutations outside predicted motifs.

sequences-proteins-beads were then collected for PCR amplification with forward primer and biotinylated reverse primer (step 1: 95°C for 1 min with initial denaturation; step 2: 95°C for 30 s, 56°C for 30 s, 72°C for 30 s, 12 cycles; step 3: 72°C for 2 min). The PCR products were applied to streptavidin magnetic beads by biotin-streptavidin binding according to the manufacturer's instructions. Single-stranded sequences were regenerated with 0.1 M NaOH. Negative selection was performed on blank NTA magnetic beads and other His6-labeled proteins. After 20 rounds of selection, the identified sequences were sent for NGS (Illumina MiSeq).

Statistical analysis of NGS data

Sequencing data from NGS was analyzed. The forward and reverse sequencing data from each round were pairwise checked to ensure that the bidirectional data matched. When the forward sequence was consistent with the corresponding reverse sequence after complementary base pairing, and if every base of the sequence met the minimum quality requirement, the forward sequence was then extracted from the raw data as a valid sequence. The total error rate occurring in each round and the frequency at which the forward sequence failed to pair with the reverse sequence were counted. The reads per million (RPM) of each sequence in each round was calculated to determine the enrichment level of the aptamer. This was done by dividing the number of occurrences of each sequence in each round by the total number of sequences in each round, and then multiplying by 1 million.

$$RPM = \frac{Number\ of\ reads\ mapped\ to\ an\ aptamer}{Total\ number\ of\ reads\ in\ the\ corresponding\ round} \times 10^6 \quad \text{(Equation 1)}$$

The top 20,000 high-frequency sequences were extracted for further analysis, and their corresponding frequency changes in each round were depicted.

Dataset dimensions

Dataset size

This study specifically utilized data derived from SELEX experiments targeting three distinct proteins, including CTGF, DKK1, and BCMA. Each SELEX experiment produced six to nine sequencing pools corresponding to different SELEX rounds, resulting in a robust dataset exceeding 300 GB of NGS data.

Selected specific rounds

For training purposes, we selected the following SELEX rounds for each target: CTGG: round 20; DKK1: round 30; BCMA: round 6.

Number of unique sequences

The dataset comprised high-frequency sequences computed from the samples: 7,500 unique sequences for CTGF; 4,000 unique sequences for DKK1; 2,000 unique sequences for BCMA. Additionally, we ensured a balanced representation of negative samples, selecting 150,000 sequences from the initial library that were absent in the subsequent rounds for each target.

Feature representation

One-hot encoding

To obtain each DNA sequence as inputting features, one-hot encoding was used to encode each base in the sequence as a four-dimensional vector.[34] Here, A, T, C, and G were represented by the 4-dimensional binary vectors $[1,0,0,0]^T$, $[0,0,1,0]^T$, $[0,1,0,0]^T$, $[0,0,0,1]^T$, respectively. These vectors are connected in the order of the sequence bases to form a matrix.

DNA shape features

Four pentameric DNA shape features, namely MGW, Roll, ProT, and HelT, were extracted from Monte Carlo simulation using sliding window.[19] To ensure consistency between the input DNA sequence set and the corresponding DNA shape properties, both sides of the matrix were filled with two zeros, the length was set to 1 + 4, and a sliding window was used to obtain a shape property matrix $n \times l$ (where n was the number of shapes and l was the matrix length). To eliminate bias due to different ranges of values for different shapes, zero-mean normalization was performed for each feature as follows:

$$x' = \frac{x - \mu}{\sigma} \quad \text{(Equation 2)}$$

where x was the original value, x' was the value after normalization; μ as the mean and σ was the standard deviation in the sample distribution.

Labels: response variables

Positive samples: Indicating that the protein bound to the sequence $(1, 0)^T$.

Negative samples: Indicating no binding occurred $(0,1)^T$.

Enrichment levels

The initial ssDNA library for the CTGF target had a ratio of approximately 0.896, which decreased steadily to 0.155 by round 20. DKK1 maintained a constant ratio until round 20, followed by a drop to 0.020 in round 30. For BCMA, the ratio significantly decreased in the earlier rounds, stabilizing at 0.031 by round 6.

Model training

The data were randomly divided into a training set (72%), a validation set (18%), and a test set (10%). The purpose of the validation set was to observe how the model performed on new data that was not used during training to facilitate the adjustment of appropriate hyperparameters. The test set was used to evaluate the final model performance.

A simple one-dimensional convolutional neural network (CNN) was employed, which was a widely used application of deep learning in functional genomics.[35] CNN could map an input sequence to a set of learned "filters" of a certain size, which typically recognize spatially invariant patterns. In the current dataset, the DNA sequences served as the filters. The CNN then combined these filters to determine more

detailed structures (such as presence or absence of transcription factor binding sites). For each layer of the convolutional network, the output was calculated by the following equation:

$$conv(X)_{ik} = f\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} W_{mn}^{k} X_{i+m,n} + b_{ik}\right) \quad \text{(Equation 3)}$$

where x was the input index, i was the output position index, k was the kernel size, w was the convolution weighting tensor interpreted as a 4-form pattern detector, and b was the bias term.

In this model, the activation function $f(x)$ was a nonlinear function derived from the ReLU element, which was widely used to reduce the gradient loss problem in backpropagation learning and facilitate good convergence.[36]

$$ReLU(x) = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases} \quad \text{(Equation 4)}$$

Then, by selecting the maximum value from the output of the convolutional layer in the maximum value accumulation layer, the dimensionality of the input was reduced, resulting in a computationally efficient model. The aggregation operation was defined as follows:

$$pooling(X)_{ik} = max\left\{X_{iM,k}, X_{(iM+1),k}, \ldots, X_{(iM+M-1),k}\right\} \quad \text{(Equation 5)}$$

After the maximum pooling layer, a BiLSTM layer followed, which captured the long-term orientation dependence and spatial distance between the model and the sequence.[37] BiLSTM was a sequential data processing alternative to RNN and was proposed to use special hidden units for long-term input storage. The key to BiLSTM was the unit state, which was controlled by structures called gates, including input gates, forget gates, and output gates, which were carefully controlled. In the first step, the forget gates decided which information to discard and which to keep. Next, it determined how much new information to add to the cell state. Finally, it decided which values to output. BiLSTM is a recurrent neural network primarily used for natural language processing.

$$f_t = sigmoid\left(W_f[h_{t-1}, x_t] + b_f\right) \quad \text{(Equation 6)}$$

$$i_t = sigmoid\left(W_i[h_{t-1}, x_t] + b_i\right) \quad \text{(Equation 7)}$$

$$C_t = \left(W_G[h_{t-1}, x_t] + b_G\right) \quad \text{(Equation 8)}$$

$$S_t = f_t S_{t-1} + i_t \odot C_t \quad \text{(Equation 9)}$$

$$O_t = sigmoid\left(W_o[h_{t-1}, x_t] + b_o\right) \quad \text{(Equation 10)}$$

$$h_t = \mathrm{O}_t \odot \tanh(S_t) \quad \text{(Equation 11)}$$

The above equations showed the weights of the input stream, where W was the weight matrix; b was the distortion; f_t, i_t, and O_t were the forgetting weights and input-output gate values; x_t, C_t, and h_t were the input vectors at time t, memory representation, and hidden layer state; and $\odot$ was the multiplication of the elements. To avoid overfitting by ignoring half of the symptom detectors and to improve the generality of the model, a rejection layer with probability 0.2 was added. The results of all DNA sequence decisions and shape information were then combined into a symptom vector and passed to the output stage.

Loss

We only learned from the reduced training set, but we plotted loss curves from both the training set and the validation set. Learning was deemed complete when the loss on the validation set was no longer improving or worsening during the training cycle. This could indicate that the model had reached its limit or was overfitting.

To train the proposed hybrid model, we minimized the focal loss function to address the challenges posed by class imbalance. The focal loss was defined as follows:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \quad \text{(Equation 12)}$$

where p_t was the predicted probability for the true class, α_t was the balancing parameter for class weights, and γ was the focusing parameter that adjusts the rate of down-weighting easy examples. This allowed the model to focus more on harder-to-classify samples.

Hyperparameters used in the optimized model

Learning Rate: Set to 0.001, allowing for effective convergence during training.

Batch Size: Set to 300 to ensure balanced updates while managing computational efficiency.

L2 Regularization Parameter (λ): A value of 0.01 was utilized to prevent overfitting by penalizing excessive weights.

Balancing Factor (α): Set to 0.25 to emphasize the minority class in the dataset.

Focusing Parameter (γ): Chosen as 2 to effectively down-weight easy-to-classify examples during training.

Optimization Algorithm: The AdaDelta optimizer was used, with the following parameters:

Decay Rate: Randomly selected from the range [0.2,0.5].

Momentum: Chosen from [0.9,0.99,0.999].

Delta Values: Selected from [1e−8,1e−6,1e−4].

Number of Epochs: Limited to 100 to prevent overfitting and ensure efficient training.

Early Stopping Criteria: Implemented to halt training when the validation loss did not improve for 10 consecutive epochs.

Accuracy

The accuracy of the neural network in the binary classification task was plotted. For this instance, binary precision was used to compute the fraction of predictions that correspond to the label or response variable.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{(Equation 13)}$$

Model evaluation

Confusion matrix

The confusion matrix was a two-dimensional matrix where the rows represent the actual categories, and the columns represent the predicted categories.[38] The four quadrants of this matrix represented true (true positive, TP), false negative (FN), true negative (TN), and false positive (FP).

True positive (TP): the model predicted positive categories as well as actual categories.

False negative (FN): the model predicted a negative category, but the actual category is positive.

True negative (TN): the model predicted negative categories as well as actual categories.

False positive (FP): the model predicted a positive category, but the actual category is negative.

ROC curve

ROC (receiver operating characteristic) curve, also known as receiver operating characteristic curve, was a tool used for evaluating classification models. The ROC curve showed the performance of the model at all possible classification thresholds. It created the curve by comparing the true case rate (TPR) with the false positive case rate (FPR). The true case rate (also known as sensitivity or recall) was the proportion of true cases (i.e., positive cases correctly predicted by the model) to all actual positive cases. The FP rate was the proportion of FP cases (i.e., positive cases incorrectly predicted by the model) to all actual negative cases.

$$TPR = \frac{TP}{TP+FN} \quad \text{(Equation 14)}$$

$$FPR = \frac{FP}{TN+FP} \quad \text{(Equation 15)}$$

The horizontal axis of the ROC curve was the FPR, and the vertical axis was the TPR. An ideal model would be as close as possible to the upper left corner, the point where the TPR was 1 and the FPR was 0, meaning that the model correctly predicted all the positive instances and did not incorrectly predict any of the negative instances as positive instances. A random-guess model appeared as a slanted line on the graph, as the model's ability to predict positive and negative instances was equal to the threshold that was changed, meaning that the TPR and FPR increase or decrease by the same amount. The area of the ROC curve was referred to as the AUROC (area under the ROC), which was used to quantify the overall performance of the model. The value of the AUC was between 0 and 1, where 1 indicated a perfect model, 0.5 indicated a random-guess model, and a value of less than 0.5 indicated that the model's prediction was worse than a random guess. ROC curve was robust to unbalanced datasets. It gave a reasonable assessment even when the number of positive and negative examples varied greatly. This was because the ROC curve did not directly consider the number of positive and negative cases, but instead focused on the model's ability in predicting positive and negative cases.

VAE

The generative model variational autoencoder (VAE) was used for learning a continuous representation of the data by learning it in the latent space and is able to generate new data. VAE first encoded the input data into two parameters, mean and variance, then sampled from the normal distribution defined by these two parameters to obtain the latent variable, and finally decoded this latent variable into the original data. This process could be represented by the following equation.

Encoding process:

$$\mu, \sigma = Encoder(x) \quad \text{(Equation 16)}$$

Sampling process:

$$z = \mu + \sigma \bullet \in \quad \text{(Equation 17)}$$

where $\in \sim N(0,1)$.

Decoding process:

$$x' = Decoder\ (z) \quad \text{(Equation 18)}$$

The goal of VAE was to maximize the marginal log likelihood of the data and to make the resulting latent variables obey a standard normal distribution by introducing KL divergence as a regular term. Therefore, the loss function (Loss Function) of VAE can be expressed as:

$$L = E[log\ P(X,Z)] - DKL(Q(Z,X)|P(Z)) \quad \text{(Equation 19)}$$

where $E[log\ P(X|Z)]$ was the reconstruction error, which represented the difference between the decoded data and the original data; $DKL(Q(Z|X)P(Z))$ was the KL divergence between the latent variable Z and the standard normal distribution, which was used to measure the similarity of the two distributions. $Q(Z|X)$ represented the distribution of the latent variable Z given the input data X; $P(Z)$

was the latent prior distribution of the variable Z, which was assumed as a standard normal distribution. In practice, we optimize this loss function by stochastic gradient descent (SGD) or other optimization algorithms to achieve the training of the VAE model.

Structure simulations

The secondary structures of the nucleic acid aptamers were predicted by RNAfold (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi) and the minimum free energy was calculated.[39] Docking of the aptamer sequences to target proteins was performed on HDOCK (http://hdock.phys.hust.edu.cn/).[40–43]

Biolayer interferometry

The binding affinity was measured using Fortebio Octet BLI Data Analysis 11.0.[44,45] A 200 μL amount of PBS was added in each well of A1-H1 for baseline 1. A 1,000 nM protein in 200 μL PBST was added in each well of A2-H2, A3-H3, A4-H4, and A5-H5 for loading. Different aptamer sequences were added into each well of A6-H6, A7-H7, A8-H8, and A9-H9 for association step. The concentration gradient was initiated at 1,000 nM, followed by six serial 2-fold dilutions in PBST. PBST in wells H6-H9 were served as the blank reference and PBST in wells A10-H10 were served as baseline 2 and dissociation. For regeneration, 5 M NaCl in PBS in wells A11-H11 and PBS in wells A12-H12 will be used. The assay protocol proceeded as follows: a 60-s initial baseline (baseline 1), a 300-s sample loading phase, a subsequent 300-s baseline measurement (baseline 2), a 420-s association phase, followed by a 420-s dissociation phase, and concluding with a 30-s regeneration step, all stages incorporating shaking at 1,000 rpm.

DATA AND CODE AVAILABILITY

The model weights, parameters, and figures are available on GitHub (https://github.com/TMBJ-lab/DeepAptamer).

ACKNOWLEDGMENTS

This study was supported by the Hong Kong General Research Fund from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 12102120, Y.Y.; Project No. 12102322, Y.Y.), Theme-based Research Scheme from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T12-201/20-R, A.L.), Health@InnoHK Program launched by Innovation Technology Commission of the Hong Kong Special Administrative Region, China (CRMH, Y.W.), Inter-institutional Collaborative Research Scheme from Hong Kong Baptist University (Project No. RC-ICRS/19-20/01, G.Z.), and Seed Funding for Collaborative Research Grants from Hong Kong Baptist University (Project No. RC-SFCRG/23-24/SCM/04, A.L.).

AUTHOR CONTRIBUTIONS

Y.Y., G.Z, Y.W., and A.L. designed and supervised the research. X.Y. and H.X. developed the DeepAptamer program and performed prediction. C.H.C, S.Y., H.Y.C., M.L., Z.C., Y.M., S.Y., F.L., J.L., L.W., Z.Z., B.-T.Z., and L.Z. performed experiments. X.Y., Y.W., and Y.Y. analyzed the data and wrote the paper. All authors read and provided comments on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests exist.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2024.102436.

REFERENCES

1. Ellington, A.D., and Szostak, J.W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature *346*, 818–822.
2. Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505–510.
3. Zhou, Q., Xia, X., Luo, Z., Liang, H., and Shakhnovich, E. (2015). Searching the Sequence Space for Potent Aptamers Using SELEX in Silico. J. Chem. Theor. Comput. *11*, 5939–5946.
4. Blind, M., and Blank, M. (2015). Aptamer Selection Technology and Recent Advances. Mol. Ther. Nucleic Acids *4*, e223.
5. Ishida, R., Adachi, T., Yokota, A., Yoshihara, H., Aoki, K., Nakamura, Y., and Hamada, M. (2020). RaptRanker: in silico RNA aptamer selection from HT-SELEX experiment based on local sequence and structure information. Nucleic Acids Res. *48*, e82.
6. Song, J., Zheng, Y., Huang, M., Wu, L., Wang, W., Zhu, Z., Song, Y., and Yang, C. (2020). A Sequential Multidimensional Analysis Algorithm for Aptamer Identification based on Structure Analysis and Machine Learning. Anal. Chem. *92*, 3307–3314.
7. Dao, P., Hoinka, J., Takahashi, M., Zhou, J., Ho, M., Wang, Y., Costa, F., Rossi, J.J., Backofen, R., Burnett, J., and Przytycka, T.M. (2016). AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments. Cell Syst. *3*, 62–70.
8. Hoinka, J., Berezhnoy, A., Sauna, Z.E., Gilboa, E., and Przytycka, T.M. (2014). AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application. Res. Comput. Mol. Biol. *8394*, 115–128.
9. Kramer, S.T., Gruenke, P.R., Alam, K.K., Xu, D., and Burke, D.H. (2022). FASTAptameR 2.0: A web tool for combinatorial sequence selections. Mol. Ther. Nucleic Acids *29*, 862–870.
10. Yu, X., Yu, Y., and Zeng, Q. (2014). Support vector machine classification of streptavidin-binding aptamers. PLoS One *9*, e99964.
11. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.
12. Blum, C.F., and Kollmann, M. (2019). Neural networks with circular filters enable data efficient inference of sequence motifs. Bioinformatics *35*, 3937–3943.
13. Chen, Z., Hu, L., Zhang, B.T., Lu, A., Wang, Y., Yu, Y., and Zhang, G. (2021). Artificial Intelligence in Aptamer-Target Binding Prediction. Int. J. Mol. Sci. *22*, 3605.
14. Asif, M., and Orenstein, Y. (2020). DeepSELEX: inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs. Bioinformatics *36*, i634–i642.
15. Kingma, D., and Welling, M. (2013). Auto-Encoding Variational Bayes (ICLR).
16. Shah, N., Chari, A., Scott, E., Mezzi, K., and Usmani, S.Z. (2020). B-cell maturation antigen (BCMA) in multiple myeloma: rationale for targeting and current therapeutic approaches. Leukemia *34*, 985–1005.
17. Ramazani, Y., Knops, N., Elmonem, M.A., Nguyen, T.Q., Arcolino, F.O., van den Heuvel, L., Levtchenko, E., Kuypers, D., and Goldschmeding, R. (2018). Connective tissue growth factor (CTGF) from basics to clinics. Matrix Biol. *68–69*, 44–66.
18. Tao, S.S., Cao, F., Sam, N.B., Li, H.M., Feng, Y.T., Ni, J., Wang, P., Li, X.M., and Pan, H.F. (2022). Dickkopf-1 as a promising therapeutic target for autoimmune diseases. Clin. Immunol. *245*, 109156.
19. Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. *41*, W56–W62.
20. Schütze, T., Wilhelm, B., Greiner, N., Braun, H., Peter, F., Mörl, M., Erdmann, V.A., Lehrach, H., Konthur, Z., Menger, M., et al. (2011). Probing the SELEX Process with Next-Generation Sequencing. PLoS One *6*, e29604.
21. Emami, N., and Ferdousi, R. (2021). AptaNet as a deep learning approach for aptamer-protein interaction prediction. Sci. Rep. *11*, 6074.
22. Zhang, L., Zhang, C., Gao, R., Yang, R., and Song, Q. (2016). Prediction of aptamer-protein interacting pairs using an ensemble classifier in combination with various protein sequence attributes. BMC Bioinf. *17*, 225.

23. Lee, S.J., Cho, J., Lee, B.-H., Hwang, D., and Park, J.-W. (2023). Design and Prediction of Aptamers Assisted by In Silico Methods. Biomedicines *11*, 356.

24. Fang, Z., Wu, Z., Wu, X., Chen, S., Wang, X., Umrao, S., and Dwivedy, A. (2024). APIPred: An XGBoost-Based Method for Predicting Aptamer–Protein Interactions. J. Chem. Inf. Model. *64*, 2290–2301.

25. Sun, D., Sun, M., Zhang, J., Lin, X., Zhang, Y., Lin, F., Zhang, P., Yang, C., and Song, J. (2022). Computational tools for aptamer identification and optimization. TrAC, Trends Anal. Chem. *157*, 116767.

26. Li, B.-Q., Zhang, Y.-C., Huang, G.-H., Cui, W.-R., Zhang, N., and Cai, Y.-D. (2014). Prediction of Aptamer-Target Interacting Pairs with Pseudo-Amino Acid Composition. PLoS One *9*, e86729.

27. Emami, N., Pakchin, P.S., and Ferdousi, R. (2020). Computational predictive approaches for interaction and structure of aptamers. J. Theor. Biol. *497*, 110268.

28. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., and Qiao, Y. (2023). Uniformer: Unifying convolution and self-attention for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. *45*, 12581–12600.

29. Teragawa, S., and Wang, L. (2023). ConF: A Deep Learning Model Based on BiLSTM, CNN, and Cross Multi-Head Attention Mechanism for Noncoding RNA Family Prediction. Biomolecules *13*, 1643.

30. Iwano, N., Adachi, T., Aoki, K., Nakamura, Y., and Hamada, M. (2022). Generative aptamer discovery using RaptGen. Nat. Comput. Sci. *2*, 378–386.

31. Bashir, A., Yang, Q., Wang, J., Hoyer, S., Chou, W., McLean, C., Davis, G., Gong, Q., Armstrong, Z., Jang, J., et al. (2021). Machine learning guided aptamer refinement and discovery. Nat. Commun. *12*, 2366.

32. Liang, C., Guo, B., Wu, H., Shao, N., Li, D., Liu, J., Dang, L., Wang, C., Li, H., Li, S., et al. (2015). Aptamer-functionalized lipid nanoparticles targeting osteoblasts as a novel RNA interference-based bone anabolic strategy. Nat. Med. *21*, 288–294.

33. Yu, Y., Wang, L., Ni, S., Li, D., Liu, J., Chu, H.Y., Zhang, N., Sun, M., Li, N., Ren, Q., et al. (2022). Targeting loop3 of sclerostin preserves its cardiovascular protective action and promotes bone formation. Nat. Commun. *13*, 4241.

34. Okada, S., Ohzeki, M., and Taguchi, S. (2019). Efficient partition of integer optimization problems with one-hot encoding. Sci. Rep. *9*, 13036.

35. Zhang, Y., Bao, W., Cao, Y., Cong, H., Chen, B., and Chen, Y. (2022). A survey on protein-DNA-binding sites in computational biology. Brief. Funct. Genomics *21*, 357–375.

36. Dittmer, S., King, E.J., and Maass, P. (2020). Singular Values for ReLU Layers. IEEE Transact. Neural Networks Learn. Syst. *31*, 3594–3605.

37. Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., Wang, H., Zhou, Y., Shi, L., Lan, F., and Wang, Y. (2019). Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nat. Commun. *10*, 4284.

38. Cabot, J.H., and Ross, E.G. (2023). Evaluating prediction model performance. Surgery *174*, 723–726.

39. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. (2008). The Vienna RNA websuite. Nucleic Acids Res. *36*, W70–W74.

40. Yan, Y., Tao, H., He, J., and Huang, S.Y. (2020). The HDOCK server for integrated protein-protein docking. Nat. Protoc. *15*, 1829–1852.

41. Yan, Y., Zhang, D., Zhou, P., Li, B., and Huang, S.Y. (2017). HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res. *45*, W365–W373.

42. Yan, Y., Wen, Z., Wang, X., and Huang, S.Y. (2017). Addressing recent docking challenges: A hybrid strategy to integrate template-based and free protein-protein docking. Proteins *85*, 497–512.

43. Huang, S.Y., and Zou, X. (2014). A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. Nucleic Acids Res. *42*, e55.

44. Kaur, H., Shorie, M., and Sabherwal, P. (2020). Biolayer interferometry-SELEX for Shiga toxin antigenic-peptide aptamers & detection via chitosan-WSe(2) aptasensor. Biosens. Bioelectron. *167*, 112498.

45. Mukherjee, M., Appaiah, P., Sistla, S., Bk, B., and Bhatt, P. (2022). Bio-Layer Interferometry-Based SELEX and Label-Free Detection of Patulin Using Generated Aptamer. J. Agric. Food Chem. *70*, 6239–6246.