Journal Title Here, 2024, pp. 1–19

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

AptaDiff: de novo design and optimization of aptamers based on diffusion models

Zhen Wang,[1,2] Ziqi Liu,[1,3] Wei Zhang,[1] Yanjun Li,[4] Yizhen Feng,[1,5] Shaokang Lv,[1,6] Han Diao,[1,6] Zhaofeng Luo,[1] Pengju Yan,[1,7,*] Min He[1,2,*] and Xiaolin Li[1,7,*]

[1]Hangzhou Institute of Medicine, Chinese Academy of Sciences, Hangzhou, 310018, Zhejiang, China, [2]College of Electrical and Information Engineering, Hunan University, Changsha, 410082, Hunan, China, [3]School of Molecular Medicine, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, 310024, Zhejiang, China, [4]Department of Medicinal Chemistry, Center for Natural Products, Drug Discovery and Development, University of Florida, Gainesville, 32610, FL, USA, [5]College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310014, Zhejiang, China, [6]Department of Chemical Biology, Zhejiang University of Technology, Huzhou, 313200, Zhejiang, China and [7]ElasticMind Inc, Hangzhou, 310018, Zhejiang, China
*Corresponding author. yanpengju@gmail.com, hemin607@163.com, xiaolinli@ieee.org

Abstract

Aptamers are single-stranded nucleic acid ligands, featuring high affinity and specificity to target molecules. Traditionally they are identified from large DNA/RNA libraries using in vitro methods, like Systematic Evolution of Ligands by Exponential Enrichment (SELEX). However, these libraries capture only a small fraction of theoretical sequence space, and various aptamer candidates are constrained by actual sequencing capabilities from the experiment. Addressing this, we proposed AptaDiff, the first in silico aptamer design and optimization method based on the diffusion model. Our Aptadiff can generate aptamers beyond the constraints of high-throughput sequencing data, leveraging motif-dependent latent embeddings from variational autoencoder, and can optimize aptamers by affinity-guided aptamer generation according to Bayesian optimization. Comparative evaluations revealed AptaDiff's superiority over existing aptamer generation methods in terms of quality and fidelity across four high-throughput screening data targeting distinct proteins. Moreover, Surface Plasmon Resonance (SPR) experiments were conducted to validate the binding affinity of aptamers generated through Bayesian optimization for two target proteins. The results unveiled a significant boost of 87.9% and 60.2% in RU values, along with a 3.6-fold and 2.4-fold decrease in KD values for the respective target proteins. Notably, the optimized aptamers demonstrated superior binding affinity compared to top experimental candidates selected through SELEX, underscoring the promising outcomes of our AptaDiff in accelerating the discovery of superior aptamers.

Key words: aptamer generation, discrete diffusion model, Bayesian optimization, deep learning

Introduction

Aptamers, single-stranded oligonucleotides, exhibit high affinity and specificity for specific targets due to their three-dimensional folding structure [1, 2, 3]. Like antibodies, aptamers have a wide range of applications, including targeted therapy [4, 5, 6], diagnostics [7, 8], biosensor technology [9] and biomarker discovery [10]. Aptamers offer several advantages over traditional protein antibodies. First, they possess a broad target range, encompassing metal ions, nucleotides, amino acids, growth factors, proteins, viruses, bacteria, living cells, and tissue samples [11, 12, 13, 14, 15, 16, 17]. This versatility enhances their applicability across various disciplines. Additionally, aptamers exhibit low immunogenicity and exceptional tissue penetration due to their oligonucleotide nature. Their small size and low molecular weight enable efficient tissue penetration and rapid in vivo target localization, surpassing antibodies [18, 19]. Moreover, aptamers are cost-effective and easily synthesized in vitro using DNA synthesis, PCR, or RNA transcription, ensuring minimal batch-to-batch variations and excellent repeatability. Lastly, aptamers demonstrate thermal stability, as their denaturation process is reversible, allowing them to recover their three-dimensional structure after exposure to high temperatures [20].

Aptamers are identified by the systematic evolution of ligands by exponential enrichment (SELEX). SELEX involves iterative cycles of affinity-based selection and amplification, wherein aptamer candidates demonstrating robust binding affinity to the target are successively enriched and isolated in each round. This iterative process ultimately results in the acquisition of aptamers with pronounced binding capacity for specific targets [21, 22]. In recent years, with the advancement of high-throughput sequencing technologies, it has become possible to screen a vast number of nucleic acid aptamer candidates using High-Throughput SELEX (HT-SELEX) techniques. However, compared to the vast potential sequence space, current findings suggest that aptamers with

© The Author 2024. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

desired properties for specific tasks are exceptionally rare [23]. The theoretical sequence space of a DNA aptamer library containing a 40-nt random region is on the order of 10^{25} ($\sim 4^{40}$), whereas the size of manually prepared SELEX input libraries currently stands at around 10^{15}, with sequencing throughput of approximately 10^{6}. Consequently, we can only assess a very small fraction of the theoretical diversity, making it extremely challenging to select aptamers that specifically bind to target proteins within this immense sequence space. Therefore, efficient computational methods for handling high-throughput sequencing data are crucial in aptamer development.

To effectively identify high-quality aptamers from high-throughput sequencing data, Singh et al. [24] used a three-dimensional non-fouling porous hydrogel with immobilized target proteins for high-affinity aptamer selection. In addition, some machine learning and deep learning models have been employed for aptamer assessment. AptaNet utilizes a multilayer perceptron (MLP) trained on an aptamer database to predict whether an aptamer binds to a protein [25]. Rube et al. [26] used interpretable machine learning to predict protein-ligand binding affinities from sequencing data. Ali Bashir et al. [27] partitioned aptamer sub-libraries by affinity level using Particle Display (PD) and trained machine learning models on these data to predict high-affinity DNA aptamers from experimental candidates. While these methods are valuable for analyzing existing sequences, they are limited to exploring the sequences that actually exist in the database. Simulation-based methods have been reported for sequence generation [28, 29, 30], but these methods require prior knowledge of sequence motifs. Thus, motif mining plays a crucial role in identifying protein-binding sites on DNA or RNA, aiding in understanding gene regulation and control mechanisms. Several machine learning-based methods have been developed for motif detection [31, 32, 33]. In addition to simulation-based methods, some approaches utilize experimental mutation information (nucleotide substitutions) for aptamer candidate discovery [29, 27]. While mutation information is advantageous for aptamer discovery, this approach comes with considerable uncertainty.

We focused on aptamer generation methods based on deep learning. Jinho IM et al. [34] employ LSTM networks to generate aptamer sequences similar to the training data by learning from high-throughput experimental data. Furthermore, other neural network-driven generative models like Generative Adversarial Networks (GANs) [35] and Variational Autoencoders (VAEs) [36] have been proposed for nucleic acid sequence generation [37, 38, 39]. Recently, denoising diffusion probability models (DDPMs) [40, 41, 42] have achieved significant successes in computer vision field for their superior generative capabilities and have been applied to domains like small molecule and protein design [43, 44, 45, 46, 47, 48]. Diffusion models operate by continually introducing Gaussian noise to perturb training data and subsequently learning to reverse the noise process for data recovery. After training, diffusion models can incorporate randomly sampled noise into the model and generate data by learning the denoising process.

In this paper, we introduce AptaDiff, a conditional discrete diffusion model for aptamer generation. It is the first computational model based on diffusion generative models in the field of aptamer generation. AptaDiff shares similarities with the multinomial diffusion [49] for categorical data but has been tailored specifically for aptamers. We enhance the aptamer sequence generative process by extending the diffusion model with motif-dependent latent representations as conditions (see Fig. 1). The latent space is obtained through a Variational Autoencoder (see Fig. 1b), where sequences form clusters based on motif structures (see Fig. 1d). This operation effectively equips the diffusion process with a latent space, enhancing the model's interpretability. Utilizing latent representations, we generated aptamers absent from high-throughput sequencing data and optimized them for greater affinity to target protein through affinity-guided Bayesian optimization (see Fig. 1f).

We demonstrate that the generated sequences exhibit characteristics similar to real HT-SELEX sequences across various metrics, and the optimized sequences show stronger affinity to the target protein. We believe that AptaDiff represents an effective technique for designing aptamer sequences as candidate therapeutics and opens up a promising new avenue for aptamer-based therapies.

Materials and Methods

Preliminaries

Variational Autoencoders

Variational autoencoders (VAE) comprise an encoder neural network and a decoder neural network, where the encoder $\mathcal{E}_\phi$ transforms input sequences ($\mathbf{x}$) into a latent distribution ($q_\phi(\mathbf{z}|\mathbf{x})$), and the decoder $\mathcal{D}_\theta$ learns to reconstruct input data from the latent representation ($\mathbf{z}$) by learning ($p_\theta(\mathbf{x}|\mathbf{z})$), where ϕ and θ represent model parameters. The whole framework can be trained by minimizing the reconstruction objective $\boldsymbol{d}(\mathcal{D}(\mathcal{E}(\mathbf{x}), \mathbf{x}))$. As a generative model, VAE's evaluation hinges on model evidence, yet computing the model evidence ($p_{\boldsymbol{\theta}}(\mathbf{X})$) for a dataset $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ proves computationally challenging. Alternatively, we turn to maximizing the Evidence Lower Bound (ELBO), denoted as $\mathcal{L}_{VAE}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X})$, to gauge the model's description of the dataset using Jensen's inequality. The whole VAE can be effectively optimized by:

$$\log p_{\boldsymbol{\theta}}(\mathbf{X}) \geq \mathcal{L}_{VAE}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}) = \sum_{i=1}^{N} \mathcal{L}_{VAE}\left(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}\right) \tag{1}$$

where

$$\begin{aligned} \mathcal{L}_{VAE}\left(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}\right) &= \mathcal{L}_{recon} + \mathcal{L}_{KL-reg}, \\ \mathcal{L}_{recon} &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_\theta\left(\mathbf{x}^{(i)} \mid \mathbf{z}\right)\right], \\ \mathcal{L}_{KL-reg} &= -D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}\left(\mathbf{z} \mid \mathbf{x}^{(i)}\right) \| p_{\boldsymbol{\theta}}(\mathbf{z})\right), \end{aligned} \tag{2}$$

which is a reconstruction loss $\mathcal{L}_{recon}$ combined with a regularization term $\mathcal{L}_{KL-reg}$. $D_{\mathrm{KL}}(p\|q)$ stands for the Kullback–Leibler divergence between distributions p and q. The reconstruction loss in practice is calculated as L_2 norm or cross-entropy loss. In this paper, we refer to the negative value of ELBO as the model loss.

Diffusion Probabilistic Models (DPMs)

Diffusion probabilistic models (DPMs) are a class of generative models based on a reversible, discrete-time diffusion process. Given data $\boldsymbol{x}_0$, a diffusion probabilistic model defines two Markov chains of diffusion processes. The forward diffusion process consists of predefined variational distributions $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ that gradually reduce the signal-to-noise ratio over time steps $t \in \{1, ..., T\}$ until the data distribution

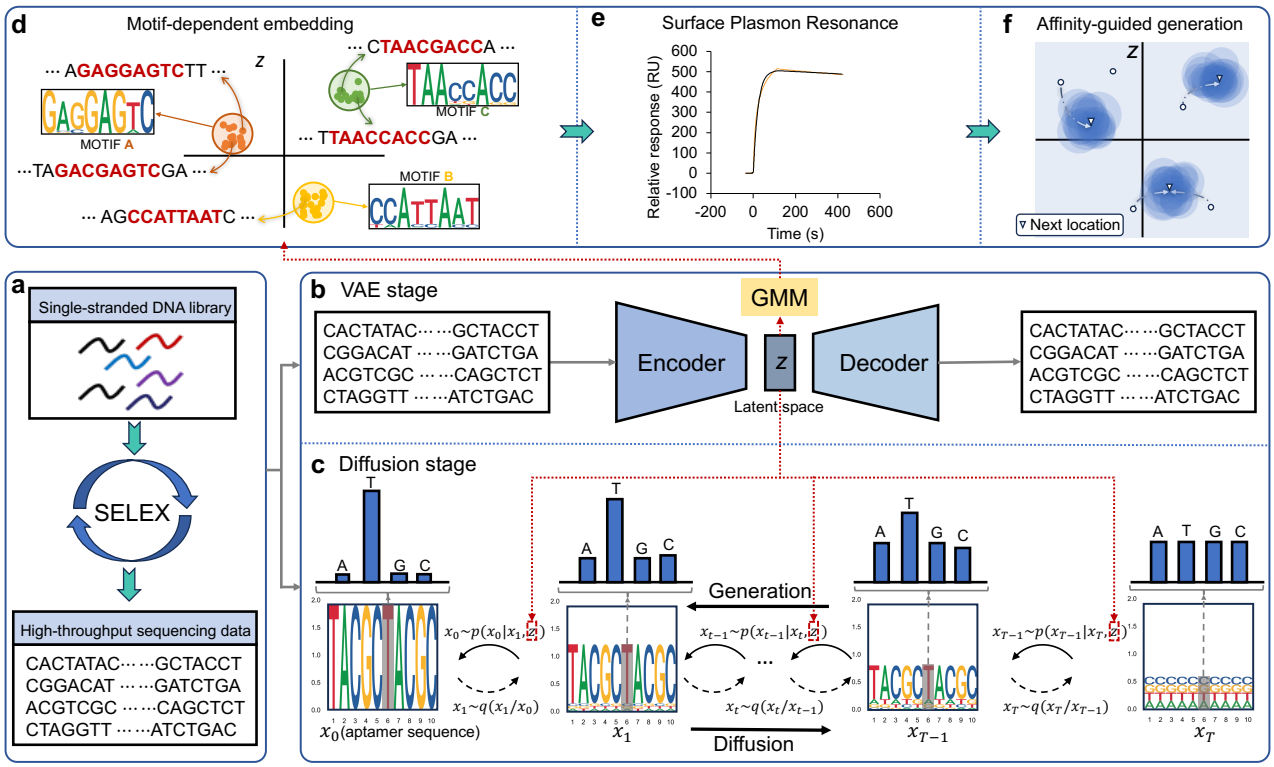

Fig. 1. Schematic view of the AptaDiff architecture, with **(a)**, **(b)**, and **(c)** representing the training process of AptaDiff. **(a)** Training data comes from high-throughput sequencing data from SELEX experiments. AptaDiff comprises a VAE phase and a diffusion phase. **(b)** In the VAE phase, a latent space with motif-dependent information can be obtained by a specific VAE (the latent space is shown in **z** in this figure). **(c)** During the diffusion stage, we model the training data using a Multinomial Diffusion conditioned on the motif-dependent embedding. The data used for diffusion is the aptamer sequence. In subfigure **c**, the diffusion process of the aptamer sequences is displayed below, while the diffusion state of individual nucleotides within the sequences is shown above. **(d)** The VAE constructs the latent space based on sequence similarity. **(e)** The binding affinity between the aptamer and the target was measured by surface plasmon resonance experiment. **(f)** AptaDiff can propose candidate aptamers through Bayesian optimization based on the binding affinity distribution by converting the latent representation into a probabilistic model.

approximately reaches the prior distribution. The generative diffusion process consists of learnable distributions $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ that start from the prior distribution and transform it to the desired distribution by progressively denoising noisy data.

Dataset

The training data used in this study comprises 4 independent HT-SELEX datasets, denoted as datasets A, B, C, and D. These datasets are directed toward four distinct target proteins. Specifically, datasets A and B were derived from HT-SELEX experiments conducted by ourselves (see Supplementary Section 2 for details), while datasets C and D were obtained from two public SELEX datasets provided by DNA Data Bank of Japan (DDBJ) [50, 51]. The original HT-SELEX data consists of variable random regions flanked by fixed primer regions. The target sequence lengths in datasets A, B, C, and D are 36 nt, 36 nt, 30 nt, and 40 nt, respectively. For detailed information about each dataset and their corresponding target proteins, please refer to Supplementary Section 2 and Table S9 for specific references and descriptions.

Overview of AptaDiff

Here, we propose AptaDiff, a novel conditioning diffusion-based generative model, to generate new aptamer sequences that are not included in the input SELEX dataset. The AptaDiff network comprises two main stages, as shown in Fig. 1. In the first stage, a latent space with motif-dependent information can be obtained by a specific VAE (see Fig. 1b). In the second stage, we can then model the training data using a Multinomial Diffusion [49] conditioned on the motif-dependent embedding (see Fig. 1c). This design effectively equips the diffusion process with a low-dimensional latent space. Using motif-dependent embedding as the condition can guide the generation of aptamers, thereby resulting in better generative performance than existing models.

In the proposed model, the initial stage is to train a VAE to learn the low-dimensional motif-dependent aptamer representation (Fig. 1d). Notably, motifs are neither explicitly generated by the model nor identified through direct comparison of aptamers. Instead, motifs are implicitly learned through a specific VAE. Following Iwano et al. [39], we constructed the VAE with a 1-D CNN-based encoder and a profile HMM decoder. Training this VAE yields a latent space with motif-dependency (a conclusion substantiated in reference [39]). The characteristics of motifs within this space are primarily captured by the profile HMM decoder. The encoder captures sequential motifs that are aligned in certain positions. Within this latent space, aptamer sequences can form clusters based on their motif structures, even though the specific details of these motifs remain unknown and implicit, as illustrated in Fig. 4a.

Details of the Diffusion Stage

Our approach to the aptamer sequence design problem builds on Multinomial Diffusion, a likelihood-based model for categorical data [49]. We follow the conventions and notation set by Hoogeboom et al. [49], which we review here.

The forward diffusion process for nucleotide (base) types is defined as follows:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \text{Multinomial}\left((1-\beta_t)\boldsymbol{x}_{t-1} + \frac{\beta_t}{4}\cdot\mathbf{1}\right) \quad (3)$$

where $\boldsymbol{x}_t$ will be represented in one-hot encoded format $\boldsymbol{x}_t \in \{0,1\}^K (K=4)$. Specifically, for category i, $x_i = 1$ and $x_j = 0$ for $i \neq j$. $\mathbf{1}$ is an all-one matrix. β_t is the probability of resampling another nucleotide over 4 types uniformly. When $t \to T$, β_t is set close to 1 and the distribution is closer to the uniform distribution. The following probability density provides an efficient way to perturb x_0 for timestep t during training [49].

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \text{Multinomial}\left(\bar{\alpha}_t\boldsymbol{x}_0 + \frac{1-\bar{\alpha}_t}{4}\cdot\mathbf{1}\right) \quad (4)$$

where $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{\tau=1}^{t}\alpha_\tau$. Intuitively, at each next time step, a small amount of uniform noise β_t is introduced over the 4 nucleotide classes, and there is a high probability $(1-\beta_t)$ that the previous value $\boldsymbol{x}_{t-1}$ is sampled.

The posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ can be derived from Eq.3 and Eq.4, that is:

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) = \text{Multinomial}\left(\widetilde{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{x}_0)/\sum_{k=1}^{4}\widetilde{\theta}_k\right) \quad (5)$$

where $\widetilde{\boldsymbol{\theta}}(\boldsymbol{x}_t,\boldsymbol{x}_0) = [\alpha_t\boldsymbol{x}_t + \frac{(1-\alpha_t)}{4}] \odot [\bar{\alpha}_{t-1}\boldsymbol{x}_0 + \frac{(1-\bar{\alpha}_{t-1})}{4}]$.

The generative diffusion process needs to approximate the posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ to denoise. Specifically, we predict the probability vector of $\hat{\boldsymbol{x}}_0$ from $\boldsymbol{x}_t$ and the latent embedding $\mathbf{z}$ corresponding to $\boldsymbol{x}_0$, and then parameterize $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ with the probability vector of $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\hat{\boldsymbol{x}}_0)$. The generative diffusion process is defined as :

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\hat{\boldsymbol{x}}_0) = \text{Multinomial}\left(\widetilde{\boldsymbol{\theta}}(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_0)/\sum_{k=1}^{4}\widetilde{\theta}_k\right) \quad (6)$$

where $\hat{\boldsymbol{x}}_0 = \mu(\boldsymbol{x}_t,\mathbf{z},t)$ represents a neural network model, specifically utilizing the Transformer architecture [52]. This model takes the latent embedding $\mathbf{z}$ corresponding to $\boldsymbol{x}_0$ and the aptamer state from the previous step as inputs, predicting the probability of nucleotide types at each position in the current state of the aptamer sequence. It's important to note that, unlike the forward diffusion process, the generative diffusion process relies on both the latent embedding $\mathbf{z}$ and the aptamer state from the previous step. The key distinction between these two processes lies in the fact that the forward diffusion process introduces noise to the data, rendering it independent of the data or latent embedding $\mathbf{z}$. On the other hand, the generative diffusion process is contingent upon the specified condition ($\mathbf{z}$) and full observation of the previous step.

The objective of training the generative diffusion process for aptamer sequence is to minimize the expected KL divergence between Eq.6 and posterior distribution Eq.5:

$$\mathcal{L}_t = \mathbb{E}_{\mathcal{R}^t\sim p}\left[D_{\text{KL}}\left(q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)\|p(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\hat{\boldsymbol{x}}_0)\right)\right] \quad (7)$$

Handling the Diffusion Model Stochasticity.

The proposed AptaDiff model consists of two types of latent representations: the low-dimensional VAE latent code $\mathbf{z}$ and the DDPM intermediate representations $\boldsymbol{x}_{1:T}$ associated with the DDPM reverse process (which are of the same size as the input sequence $\boldsymbol{x}_0$). The stochasticity in the generative diffusion process during the second stage of AptaDiff may lead to uncertainty in generated samples. To ensure that samples generated by AptaDiff are deterministic (i.e., controlled only by $\mathbf{z}$), we followed [53] to share all stochasticity during the reverse diffusion process across all generated samples. Since initializing the second stage Multinomial Diffusion in AptaDiff with different random noise during sampling might be interpreted as imparting different styles to samples, sharing this stochasticity in Multinomial Diffusion sampling across samples means using the same style for all generated samples. Through this straightforward technique, we can now achieve controlled generation using the low-dimensional VAE latent code, thereby enabling deterministic sampling (see Supplementary Section 4 for details).

Experimental Setup

In our study, we utilized one-hot encoding to encode the data, with each aptamer having dimensions of $(L, 4)$, where L represents the length of the aptamer, as mentioned earlier, and 4 corresponds to the length of the one-hot vector. During the training process, the batch size was set to 64. The one-hot encoding for aptamer nucleotides was as follows: Adenine (A): $(1,0,0,0)$, Cytosine (C): $(0,1,0,0)$, Guanine (G): $(0,0,1,0)$, Thymine (T): $(0,0,0,1)$.

The model was implemented using the PyTorch [54] library and Python. All sequences in the training set are carefully filtered. Sequences meeting the following three criteria are retained: (1) exact matching with primers (including forward and reverse primers), (2) precise alignment with the sequence design length, and (3) sequences read more than once. The sequences are split into training and test datasets in a 9 : 1 ratio. The model with the smallest test loss is selected through iterations. We conducted the training process using a single RTX 3090 GPU. The number of diffusion steps T was set to 1000. Adam was used as the training optimizer and all networks were trained up to 1000 epochs with early stopping when the test loss was not updated for 50 epochs.

Surface Plasmon Resonance Assay

After AptaDiff generates the sequence, we connect the obtained sequence to its fixed primer sequence, prepare it as a solid powder, and evaluate its binding affinity to the target protein through surface plasmon resonance (SPR) experiments. The instrument used for the SPR experiments (Fig. S3) was Biacore 8K (GE Healthcare). We mainly conducted SPR analysis on datasets A and B. The target proteins of datasets A and B were insulin-like growth factor-binding protein 3 (IGFBP3, UniProt: p17936) and Inactive tyrosine-protein kinase 7 (PTK7, UniProt: Q13308), respectively. Aptamers were prepared with fixed primer regions at both ends and a random region in the middle as follows: 5'-TCCAGCACTCCACGCATAAC-(36 nt random region)-GTTATGCGTGCTACCGTGAA-3' for dataset A and 5'-ATTGGCACTCCACGCATAGG-(36 nt random region)-CCTATGCGTGCTACCGTGAA-3' for dataset B. Aptamers screened by SELEX wet assay were used as positive controls. All evaluated sequences are listed in Supplementary Section 1 and

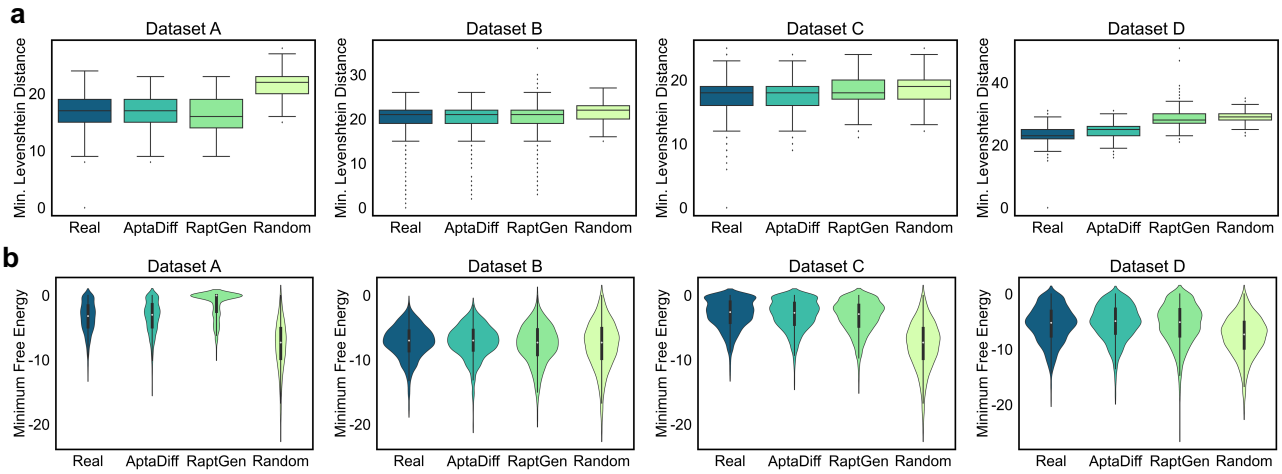

Fig. 2. Levenshtein distance distributions (a) and Minimum Free Energy distributions (b). (a) The Levenshtein distances measured between the generated sequences from AptaDiff, RaptGen, and random sequences to the nearest real sequences are shown in the distributions for four datasets (A, B, C, D). To quantify the real sequences, the distance to the nearest member with a non-zero distance was considered. Both random and generated sequences have a size of $2,000$, and all real sequences used for training are included here. Box plots characterize a sample using the 25th, 50th (also known as the median or Q2), and 75th percentiles—referred to as the lower quartile (Q1), median (m or Q2), and upper quartile (Q3)—and the interquartile range (IQR = Q3-Q1), covering the central 50% of the data. Whiskers extend outside the box up to 1.5 times the IQR. Any data points outside the whiskers are considered outliers and are plotted individually as dots. The results demonstrate that, across the four datasets, the sequences generated by our AptaDiff exhibit nearly identical distance distributions compared to the real sequences. **(b)**We compared the Minimum Free Energy (MFE) distribution among $2,000$ sequences generated by AptaDiff, $2,000$ sequences generated by RaptGen, $2,000$ randomly generated sequences, and the MFE distribution of the real HT-SELEX data used for training and validation. The results indicate that, when compared to real sequences, our AptaDiff-generated sequences exhibit nearly identical MFE distributions across all four datasets. In contrast, sequences generated by RaptGen show MFE distributions that closely resemble the original sequences in some datasets, while randomly generated samples exhibit different distributions.

Supplementary Table S1, S2, S4, S5, S6, S7 and S8). Aptamers are synthesized in vitro based on sequence information and prepared as a solid powder. The running buffer is 1xDPBS (1XDulbecco's Phosphate Buffered, PH 7.4-7.6) containing 5 mM magnesium ions (Mg^{2+}). The sensor chip is Biacore series S sensor chip CM5 (BR-1005-30, Cytiva), its gold film surface is covalently linked to the carboxymethyl dextran matrix, and the protein is covalently coupled to the carboxyl group on the surface of the sensor chip through the amino group (Fig. S3). Coupling the protein to the second channel of the CM5 chip as the experimental group, the specific steps are as follows:

- Mix equal volumes (1 : 1) of EDC (N-(3-Dimethylaminopropyl)-N'-ethylcarbodiimide hydrochloride, 0.4 M aqueous solution) and NHS (N-Hydroxy Succinyl, 0.1 M aqueous solution) to activate the carboxyl groups on the chip. The instrument parameters are injection time 420 s, and flow rate 10 μL/min.
- Dilute the proteins of dataset A and dataset B with 10 mM acetic acid-sodium acetate buffer solution at pH 4.5 for covalent coupling with the carboxyl groups on the surface of the chip. The instrument parameters are a flow rate of 10 μL/min and a binding time of 900 s.
- Use ethanolamine (1 M) to block the carboxyl group of the unbound protein on the chip, the instrument parameters are flow rate 10 μL/min, and blocking time 420 s.

The 9*His Peptide was coupled to the first channel of the CM5 chip as a control group, and the experimental steps were exactly the same as above. Afterward, the sequences of datasets A and B flowed through the sensor chip to measure their affinity with the protein (sequences were diluted with DPBS containing 5 mM magnesium ions at a concentration of 500 nM, the flow rate was 30 μL/min, and the binding time was 120 s, the dissociation time of A is 120 s, and the dissociation time of B is 300 s). In order to regenerate the sensor chip and completely remove the bound aptamers, a regeneration solution is required. The regeneration solution of protein A is 1.5 M sodium chloride with 5 mM NaoH (flow rate 30 μL/min, time 60 s), and the regeneration solution of protein B is 2 mM NaoH 1.5 M sodium chloride (flow rate 30 μL/min, time 30 s).

Results

AptaDiff shows the state-of-the-art generative performance, robustness, and generalization

We hope that the AptaDiff architecture described above can learn to generate aptamer sequences similar to real SELEX data. Specifically, we aim for the generated samples to align with the various attributes of the real SELEX data used for training and validation. Currently, there is a dearth of work focused on aptamer generation, and the evaluation of aptamer generation remains an open question. In this paper, we carefully set up our evaluation pipeline to reflect generative quality. We evaluate our model from multiple perspectives to comprehensively examine the quality of aptamer generation, including the distance between the generated sequence and the real sequence, the GC content, the minimum free energy (MFE) distribution, base distribution, the FID value, and motifs cluster analysis. We also compare our model with randomly generated sequences and the current state-of-the-art aptamer generation model—RaptGen [39]. Additionally, to validate the robustness and generalizability of our model, we conducted experiments on four independent HT-SELEX datasets (data sets A, B, C, and D, respectively for 4 different target proteins).

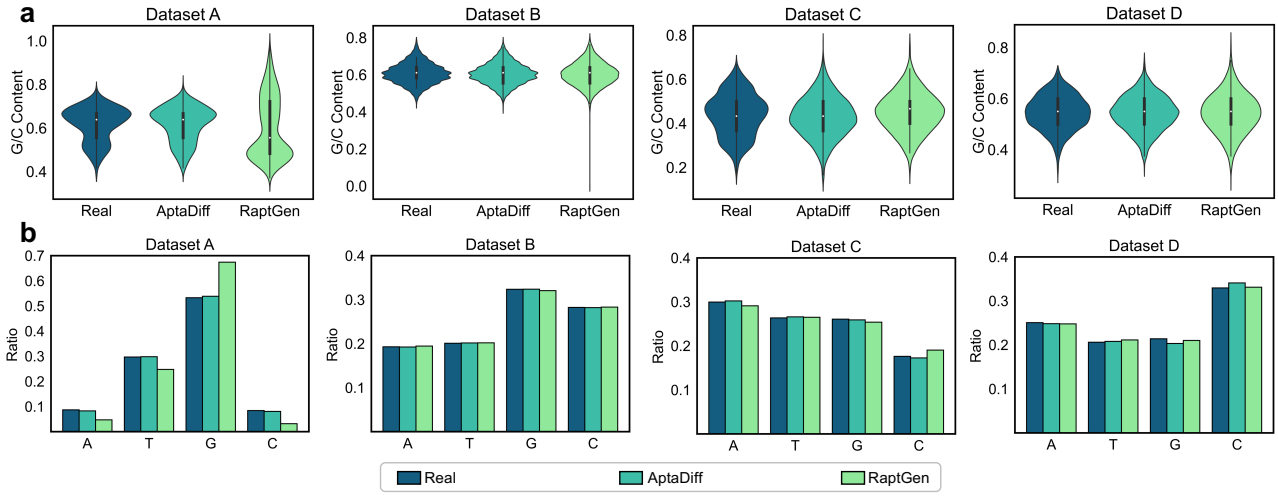

Fig. 3. GC content distributions (a) and base distributions (b). (a) We conducted an analysis of the GC content distribution for the generated sequence sets. We separately computed the GC content distribution for 2,000 sequences generated by AptaDiff and 2,000 sequences generated by RaptGen and compared their distributions to the GC content of 2,000 real sequences. The results demonstrate that, across the four datasets, our AptaDiff-generated sequences exhibit more similar mean values and distributions of GC content compared to the real sequences. **(b)** We conducted an analysis of the base distribution for the generated sequence sets. We separately computed the base distribution for 2,000 sequences generated by AptaDiff and 2,000 sequences generated by RaptGen and compared their distributions to the base distribution of 2,000 real sequences. The results indicate that, across the four datasets, the base distribution of sequences generated by AptaDiff is more similar to the real sequences compared to sequences generated by RaptGen.

Sequence Distance Evaluation

In order to assess the distance between the sets of generated sequences and random sequences in comparison to real SELEX data, we opted for the Levenshtein distance as our distance metric. The Levenshtein distance is a method for quantifying the dissimilarity between two strings by calculating the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another [55]. Following the research conducted by Ozden et al. [38], we determined the distance of each generated or random sequence to the nearest sequence in the real samples and utilized these sets of distances to represent the distance distributions of these sequence collections. The distance distribution of the real dataset was obtained by computing the minimum non-zero distance between each real sample and the entire real SELEX dataset. The distance distributions depicted in Fig. 2a indicate that, across the four datasets, our AptaDiff-generated sequences exhibit nearly identical distance distributions compared to real sequences, while randomly generated samples display different distributions. Compared with the sequences generated by RaptGen, the sequences generated by our model are closer to the distance distribution of the real sequences in the three data sets A, C, and D, and the performance on data set B is equivalent to RaptGen. This suggests that our AptaDiff model has more effectively learned the contextual characteristics of aptamer sequences.

Minimum Free Energy Distribution Evaluation

The minimum free energy (MFE) of nucleic acid aptamer sequences determines the stability of their secondary structure, and as such, we aim for the MFE distribution of the generated samples to align with that of real samples. We employed the ViennaRNA package [56] to calculate the MFE of aptamer sequences. As depicted in Fig. 2b, we evaluated the MFE distribution of 2,000 sequences each from AptaDiff, RaptGen, and randomly generated sequences against the MFE distribution of the real HT-SELEX data utilized for training and validation. The results indicate that, in comparison to real sequences, our AptaDiff-generated sequences exhibit nearly identical MFE distributions across all four datasets, while sequences from both RaptGen and random generation exhibit more distinct distributions. This suggests that AptaDiff effectively emulates real data in terms of minimum free energy.

GC Content and Base Distribution Evaluation

We conducted an investigation into the GC content and base distribution of the generated sequence sets. We separately computed the GC content and base distribution for 2,000 sequences generated by AptaDiff and 2,000 sequences generated by RaptGen. We then compared their distributions to those of 2,000 real sequences. The GC content distribution illustrated in Fig. 3a reveals that, across the four datasets, AptaDiff-generated sequences exhibit GC content distributions more similar to real sequences compared to sequences generated by RaptGen. This similarity is particularly pronounced when the datasets have higher GC content (e.g., datasets A and B). The base distribution shown in Fig. 3b illustrates that the base distribution of the sequences generated by AptaDiff is more similar to the real sequences compared with the sequences generated by RaptGen on the 4 datasets. It is worth noting that in our analysis, we did not include the GC content and base distribution of random sequences since all four bases have an equal probability of occurrence (25%) at every position in randomly generated sequences.

FID Evaluation

The Fréchet Inception Distance (FID) was originally introduced as a metric for evaluating the quality of images generated by generative models [57]. FID compares the distribution of generated images to that of a reference set of real images, often

Table 1. Quantitative Illustration of the generative performance on datasets A, B, C, and D. FID reported on 10k samples (Lower is better).

Source	Dataset	Target Protein	FID	
			RaptGen	AptaDiff
Ours	A	IGFBP3[1]	7.97	2.51
	B	PTK7[2]	6.61	1.67
Public	C	TGM2[3]	2.48	1.06
	D	ITGαVβ3[4]	2.65	1.58

1: IGFBP3, Insulin-like growth factor-binding protein 3.

2: PTK7, Inactive tyrosine-protein kinase 7.

3: TGM2, Human transglutaminase 2.

4: ITGαVβ3, Human integrin alpha V beta 3.

referred to as the "ground truth," and has become one of the most widely used metrics for assessing the quality of generative models.

In order to compare the distribution of aptamer sequences generated by our model to real data, we adapted the FID comparison for aptamers, following these steps: (1) We randomly selected $10,000$ sequences from both the generated sequences and real sequences. (2) We extracted features from these sequences using RNA-FM [58], resulting in feature vectors of dimension $(L, 640)$ for each sequence, where L represents the length of the aptamer sequence and 640 represents the dimension of each base embedding, thereby obtaining feature distributions for both datasets. (3) We computed the mean vectors and covariance matrices for the two distributions. (4) Finally, we calculated the Fréchet distance between the two distributions, yielding the FID score. As shown in Table 1, compared with RaptGen on 4 data sets, the FID values between the sequences generated by AptaDiff and the real sequences are smaller, indicating that the sequences we generated are more similar to the real sequences and of higher quality. Similar results were obtained by extracting features from the sequences using DNA-BERT2 [59] (see Supplementary Table S3).

Motif Clustering Evaluation

Since HT-SELEX sequences are amplified using specific binding motifs, and motifs play a crucial role in the aptamer-target binding process, the rationality of aptamer sampling is evidently important. We aim for our designs to incorporate motifs that are similar to those present in the training data. To assess the consistency of motifs contained in samples generated by AptaDiff and RaptGen relative to real data, we employed MEME [60] as a motif clustering analysis tool to evaluate their alignment with genuine aptamers. The motif clustering results presented in Fig. S1 and S2 demonstrate that, when compared to real sequences across the four datasets, AptaDiff-generated sequences contain similar motifs. Although their performance relative to RaptGen is not significantly superior, it still reaches a considerable level, indicating that our model effectively learns the motif information of aptamer sequences.

In summary, our proposed method consistently generates aptamer sequences across all four datasets that align with various attributes of real SELEX data. The performance surpasses that of the current state-of-the-art aptamer generation models, demonstrating the exceptional capabilities of our model in terms of generation quality, robustness, and generalization.

AptaDiff application in aptamer discovery for specific targets

Most existing papers in the field of biomolecular generation [61, 62, 63, 64] primarily focus on showcasing the capabilities of their models from an algorithmic design perspective but often neglect the evaluation of the effectiveness of these models in real-world drug design. Here, we verified the capability of our model in target-specific aptamer discovery through wet lab experiments. In order to provide a genuine assessment, we carefully selected two representative real-world targets (IGFBP3 and PTK7), prepared their corresponding protein crystal structures, and conducted an analysis of the binding affinities of the sequences we generated in the ensuing discussion. The original HT-SELEX data for IGFBP3 and PTK7 has 36 nt variable regions with fixed primer regions at both ends. In this study, we refer to Iwano et al. [39] to create a latent space using the variable region. According to the experiments and conclusions of Iwano et al. [39], when the dimension of the latent embedding is set to two, the VAE's ELBO loss is small. As the embedding dimension increases, the loss tends to increase. Additionally, a lower dimension is easier to visualize and more conducive to performing Bayesian optimization. Therefore, we also adopt a two-dimensional latent space for analysis. Fig. 4a shows the generated latent embeddings.

Due to the clustering of SELEX experiment sequences in latent space based on their motifs, we, similar to Iwano et al. [39], employed a Gaussian Mixture Model (GMM) to classify the distribution and selected eight distinct points representing the latent clustering centers of the GMM. Starting from these GMM centers, we attempted to generate corresponding sequences through AptaDiff for binding activity evaluation conditionally. The obtained aptamer sequences and the coordinates of each GMM center are listed in the Supplementary Table S4 and S6. After linking the obtained sequences with their fixed primer sequences, we prepared them as solid powders and assessed their binding affinity with the target protein through SPR experiments, the principle is shown in Fig. S3. Aptamers with the highest affinity selected through SELEX screening assay (see the Supplementary Tables S1 and S2) were used as positive controls. Although more than half of the candidate sequences exhibited weak or no binding affinity, a subset of sequences such as A-GMM-5, B-GMM-0, and B-GMM-4 demonstrated notable binding affinities (the Supplementary Table S4 and S6).

The VAE model within our AptaDiff framework associates each aptamer with a low-dimensional latent vector, resulting in a compact latent space (see Fig. 4a). To predict effective candidates from the vicinity of aptamers with high binding affinity, we employed a Bayesian optimization (BO) algorithm to learn an activity distribution in the latent space (see Fig. 4b) and generated aptamers based on the embeddings recommended after optimization. For Bayesian optimization, as shown in the flowchart in Supplementary Fig. S4, we initially embedded binding affinity data (RU values[1]) obtained from GMM centers into the latent space. Subsequently, we perform batch BO iterations with a Multi-point Expected Improvement heuristic method with local penalties to obtain the predicted

[1] RU values: Response Units, a measure of the change in refractive index on the sensor surface in SPR experiments, indicating the binding of the molecule to the target. Generally speaking, higher RU values indicate higher affinity.

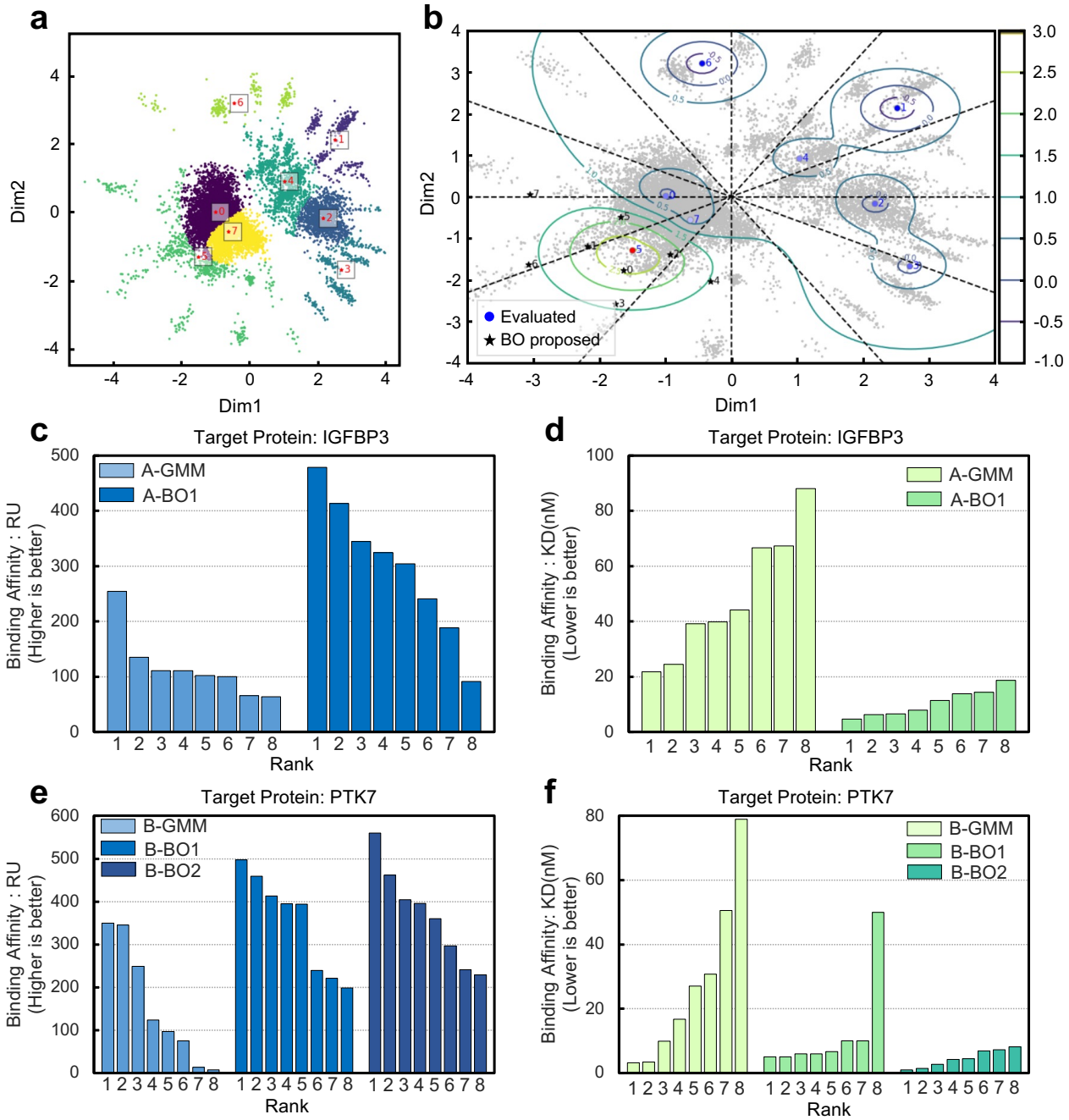

Fig. 4. Results of Gaussian Mixture Model (GMM) and Bayesian Optimization (BO). **(a)** Visualization of the latent space for Dataset A, with clustering estimated by GMM. The colored patches represent different clusters. Initial points for Bayesian optimization are obtained from each center of the Gaussian distribution. Their IDs, positions in the latent space, and corresponding affinity values with the target protein measured by SPR experiments are shown in Supplementary Table S4. **(b)** Affinity distribution for Dataset A and proposed Bayesian optimization (BO) points. The affinity data from Supplementary Table S4 are embedded into the latent space. Gray points represent the embedded points from Fig. 4a. Contours overlaid on the embedding represent predicted affinity levels, which serve as the acquisition function for Bayesian optimization. Bayesian optimization process suggests 8 points. Circles represent the re-embedding positions of GMM centers. Red and blue colors indicate high and low binding activities, respectively. Black stars represent the positions suggested by Bayesian optimization. **(c), (d), (e), (f).** Affinities of sequences proposed by different methods; A-GMM and B-GMM represent GMMs performed on the latent space of datasets A and B, respectively; A-BO1 and B-BO1 represent the first round of Bayesian optimization on datasets A and B, respectively; similarly, B-BO2 represents the second round of Bayesian optimization on dataset B. Rank represents the affinity ranking within the method (ranking of RU value (**c**) and KD value (**d**) of dataset A, and ranking of RU value (**e**) and KD value (**f**) of dataset B).

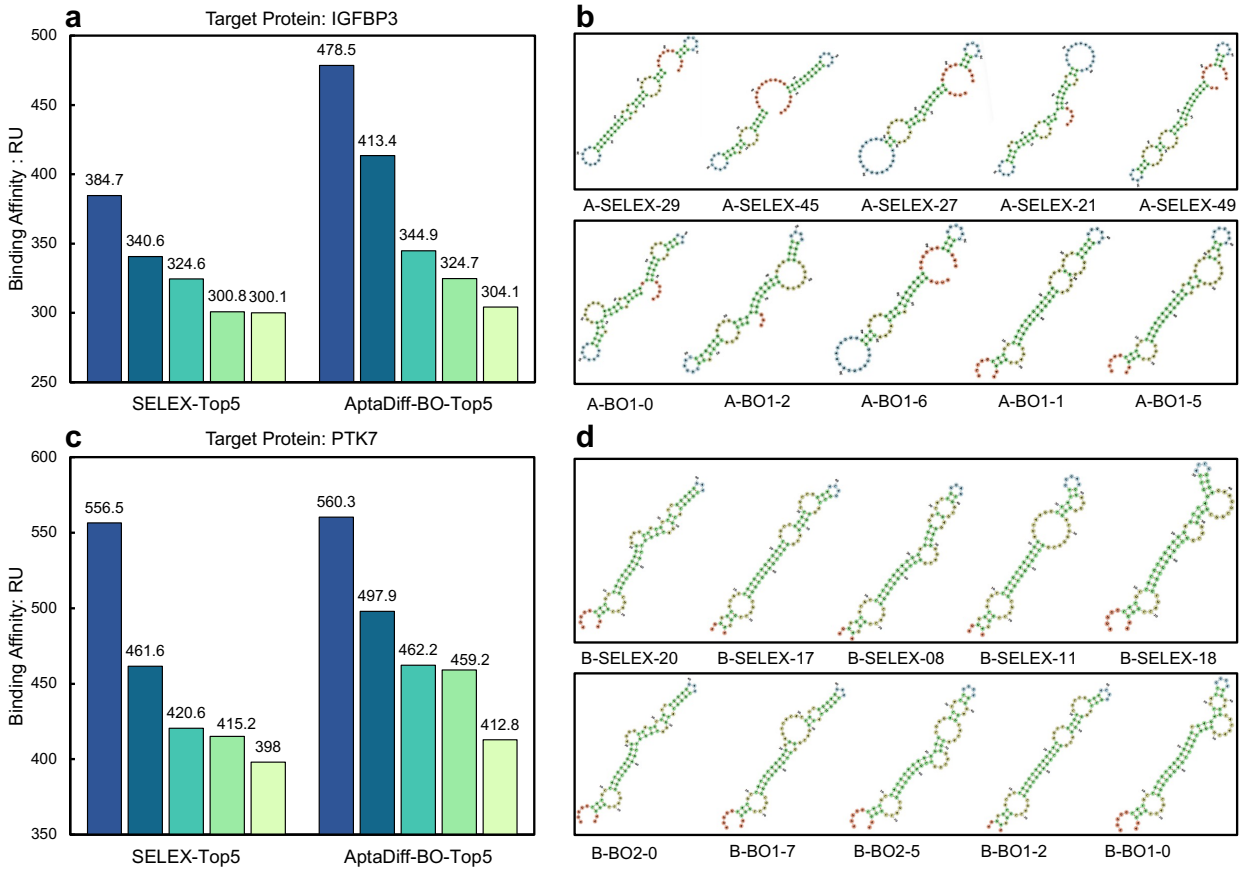

Fig. 5. Comparison of the affinity and secondary structure visualization of the top 5 aptamers obtained by SELEX and AptaDiff. (a), (b) Affinity comparison and secondary structure visualization of the top 5 aptamers obtained by SELEX and AptaDiff, respectively, for Dataset A. **(c), (d)** Affinity comparison and secondary structure visualization of the top 5 aptamers obtained by SELEX and AptaDiff, respectively, for Dataset B. The results indicate that, compared to the top 5 aptamers selected through traditional SELEX screening experiments, the aptamers generated after Bayesian optimization by AptaDiff exhibit overall higher affinity.

affinity distribution and propose multiple candidate positions in parallel, as shown in Fig. 4b. Finally, optimized sequences were generated by AptaDiff and their binding affinity was evaluated through SPR experiments. This process was iteratively repeated until the binding affinity of the optimized sequences reached or exceeded that of the positive control. We employed the GPyOpt package [65] to implement Bayesian optimization.

For the targets IGFBP3 (dataset A) and PTK7 (dataset B), we conducted one round and two rounds of Bayesian optimization, respectively. After generating the final eight new candidates, we achieved approximately 87.9% and 60.2% improvements in binding affinity (the RU value) for the aptamers (A-BO1-0, B-BO2-0, as shown in the Supplementary Table S5, S8 and Fig. 4c, 4e). Fig. 4d and 4f depict the KD values[2] of the aptamers' binding to the targets in datasets A and B before and after Bayesian optimization, which also reflects the affinity between the aptamer and the target. It is evident that after Bayesian optimization, the KD values of the obtained aptamers significantly improved. For both target proteins, the minimum KD values for the optimized sequences decreased by 3.6-fold and 2.4-fold, respectively, compared to before optimization (Supplementary Table S4-S8). These results demonstrate that AptaDiff can propose aptamer candidates in an affinity-guided manner and offers the opportunity to optimize their affinity.

Compared to the top 5 affinity aptamers selected through traditional SELEX screening experiments, the aptamers generated after Bayesian optimization exhibit higher overall affinity (see Fig. 5a and 5c, Supplementary Table S10 and S11). Surprisingly, in the experiment targeting the PTK7 protein, among the top 5 generated aptamers with affinity, the 1st-ranked sequence was the same as the 1st-ranked sequence screened by SELEX, and the 4th-ranked sequence was the same as the 2nd-ranked sequence screened by SELEX (sequence B-BO2-0 is the same as B-SELEX-20, B-BO1-5 is the same as B-SELEX-17). In addition, since secondary structure information is crucial for identifying aptamers with affinity, in this section, we used MXFold2 [66] webserver to conduct secondary structure analysis of the nucleic acid aptamers in Fig. 5a and 5c. The results are shown in Fig. 5b and 5d. The top 5 affinity aptamers we generated have similar secondary structures to the top 5 affinity aptamers screened by SELEX, indicating that

[2] KD value: Dissociation constants, the ratio of the protein sequence dissociation rate constant Kd to the protein sequence association rate constant Ka, a measure of the strength of binding between molecules in a biomolecular interaction. Generally speaking, lower KD values indicate higher affinity.

high-affinity aptamers tend to form similar structures. These findings indicate that our proposed AptaDiff-based aptamer generation pipeline can potentially replace traditional manual screening experiments, generating more high-affinity aptamers, reducing screening time, and contributing to significant savings in manpower, resources, and finances.

Discussion

Generating aptamers that effectively bind to target proteins is significant yet challenging. We introduce AptaDiff, a conditional diffusion model framework capable of generating high-fidelity and novel aptamers. By equipping AptaDiff with a motif-dependent latent space, we enhance interpretability and enable the generation of aptamers with desired binding characteristics through affinity-driven optimization. AptaDiff is the first deep learning framework to demonstrate the power of diffusion generative models on the discrete aptamer sequence space, generating sequences with attributes akin to SELEX-selected aptamers without exploring the challenging sequence space spanning up to $30 - 40$ nucleotides.

Presently, aptamer generation methods are broadly categorized into two types: those based on base mutations and those based on deep learning. Hoinka et al.[42] introduced several tools, including AptaCluster, AptaMut, and AptaSim, using mutation information from polymerase error rates for aptamer discovery. Bashir et al. [27] generate new aptamers by randomly mutating 4-nt bases and predict high-affinity aptamers using a trained machine learning model. This method only conducts mutations and evaluates them based on existing sequencing data, and is highly random. Jinho et al. [34] proposed an LSTM-based method using SELEX data but did not consider sequence clustering information. Iwano et al. [39] proposed a VAE-based method using HT-SELEX data, generating sequences not in the sequencing data but with lower fidelity.

In contrast, AptaDiff generates aptamers based on a diffusion model while considering sequence clustering information. It can generate sequences not present in the sequencing data (as shown in Fig. 2a, positive edit distances indicate sequences not present in the original SELEX data). Moreover, we have demonstrated that our generated sequences closely match the original data across multiple attribute metrics, outperforming existing state-of-the-art models on multiple datasets (Fig. 2a, 2b, 3a, and 3b, Table 1). This underscores the higher quality of sequences generated by AptaDiff relative to the aforementioned methods. Consequently, we consider AptaDiff a superior aptamer generation model compared to the methods described above, and we believe it will be a key tool for efficient aptamer discovery. However, AptaDiff does not support end-to-end training and requires a specialized VAE model to obtain a motif-dependent latent space before training the diffusion model. Future research should explore joint training of VAE and diffusion models.

In summary, we have presented a set of open-source conditional discrete diffusion models that provide a foundation for sequence-based nucleic acid aptamer engineering and design. The AptaDiff model can be directly employed for unconditional, affinity-guided aptamer sequence generation. We envision several possible avenues for extending AptaDiff. Firstly, on the model front, considering secondary structure features during the diffusion process would enhance AptaDiff's performance. One potential approach in this direction is to use diffusion models to co-design the sequence and secondary structure of aptamers. Secondly, in terms of data, employing aptamer sequences from earlier rounds of HT-SELEX experiments for training data could further expedite the SELEX screening process. Thirdly, in terms of application scenarios, AptaDiff could be used to generate a large pool of aptamer candidates for specific target proteins, forming an aptamer sequence library for SELEX screening, thereby increasing the likelihood of obtaining effective high-affinity aptamers.

As aptamers bind to target proteins through structural complementarity and binding data for aptamer-protein pairs are limited, assessing the binding affinity currently relies on wet-lab experiments, such as SPR. In the future, once a sufficient number of binding data for aptamer-protein pairs accumulates, it may become feasible to design *de novo* aptamers without wet-lab experiments in the loop. Furthermore, in the event of improved computational resources in the future, molecular dynamics simulations can be employed to model the binding conformations between aptamers and target proteins. The accumulation of a substantial amount of simulated data through such simulations could also significantly enhance computational aptamer design. These future prospects hold the potential to revolutionize the field of aptamer design, eliminating the need for extensive wet-lab experimentation and providing a more efficient and versatile approach to aptamer development.

Key Points

- We proposed AptaDiff, the first in silico aptamer design method based on the diffusion model. Aptadiff can generate aptamers beyond the constraints of high-throughput sequencing data.
- Aptadiff can optimize aptamers through affinity-guided generation via Bayesian optimization within a motif-dependent latent space, and the affinity of the optimized aptamers to the target protein is better than the best experimental candidate from traditional SELEX screening.
- Aptadiff consistently outperforms the current state-of-the-art method in terms of quality and fidelity across high-throughput screening data targeting distinct proteins.

Data availability

The source data used to train and evaluate our models are available at `https://github.com/wz-create/AptaDiff`. The HT-SELEX sequences of datasets A and B are available as DRA009383 and DRA009384 in DDBJ.

Code Availability

The code to reproduce our experiments is available at `https://github.com/wz-create/AptaDiff` under an MIT License.

Author Contribution

Z.W. and X.L. conceived the research project. X.L., M.H., and Y.P. supervised and advised the research project. Z.W. designed and implemented the AptaDiff framework. Z.W., Z.L., Y.L., and Y.F. conducted the computational analyses. Z.W., W.Z., L.S., D.H., and L.Z performed the wet experiments and analyses. Z.W., Z.L., Y.L., and X.L. wrote the manuscript. All the authors discussed the experimental results and commented on the manuscript.

Acknowledgments

We acknowledge helpful discussions with members of the AIM lab. The authors thank the anonymous reviewers for their valuable suggestions.

Funding

This work is supported in part by funds from the National Key Research and Development Program of China (2022YFC3600902) and the Zhejiang Province Soft Science Key Project (2022C25013).

Competing Interests Statement

The authors declare no competing interests.

References

1. Paromita Majumder, Katia N Gomes, and Henning Ulrich. Aptamers: from bench side research towards patented molecules with therapeutic applications. *Expert opinion on therapeutic patents*, 19(11):1603–1613, 2009.
2. Anthony D Keefe, Supriya Pai, and Andrew Ellington. Aptamers as therapeutics. *Nature reviews Drug discovery*, 9(7):537–550, 2010.
3. Maureen McKeague and Maria C DeRosa. Challenges and opportunities for small molecule aptamer development. *Journal of nucleic acids*, 2012, 2012.
4. Jiehua Zhou and John Rossi. Aptamers as targeted therapeutics: current potential and challenges. *Nature reviews Drug discovery*, 16(3):181–202, 2017.
5. Christopher P Rusconi, Elizabeth Scardino, Juliana Layzer, George A Pitoc, Thomas L Ortel, Dougald Monroe, and Bruce A Sullenger. Rna aptamers as reversible antagonists of coagulation factor ixa. *Nature*, 419(6902):90–94, 2002.
6. Christopher P Rusconi, Joseph D Roberts, George A Pitoc, Shahid M Nimjee, Rebekah R White, George Quick Jr, Elizabeth Scardino, William P Fay, and Bruce A Sullenger. Antidote-mediated control of an anticoagulant aptamer in vivo. *Nature biotechnology*, 22(11):1423–1428, 2004.
7. Edward N Brody and Larry Gold. Aptamers as therapeutic and diagnostic agents. *Reviews in Molecular Biotechnology*, 74(1):5–13, 2000.
8. Long Li, Shujuan Xu, He Yan, Xiaowei Li, Hoda Safari Yazd, Xiang Li, Tong Huang, Cheng Cui, Jianhui Jiang, and Weihong Tan. Nucleic acid aptamers for molecular diagnostics and therapeutics: advances and perspectives. *Angewandte Chemie International Edition*, 60(5):2221–2231, 2021.
9. Diana A Sousa, Mariana Carneiro, Débora Ferreira, Felismina TC Moreira, Maria Goreti F Sales, and Lígia R Rodrigues. Recent advances in the selection of cancer-specific aptamers for the development of biosensors. *Current Medicinal Chemistry*, 29(37):5850–5880, 2022.
10. Larry Gold, Deborah Ayers, Jennifer Bertino, Christopher Bock, Ashley Bock, Edward Brody, Jeff Carter, Virginia Cunningham, Andrew Dalby, Bruce Eaton, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *Nature Precedings*, pages 1–1, 2010.
11. Parag Parekh, Zhiwen Tang, Peter C Turner, Richard W Moyer, and Weihong Tan. Aptamers recognizing glycosylated hemagglutinin expressed on the surface of vaccinia virus-infected cells. *Analytical chemistry*, 82(20):8642–8649, 2010.
12. Kwame Sefah, ZW Tang, DH Shangguan, Hui Chen, Dalia Lopez-Colon, Ying Li, Parag Parekh, Jennifer Martin, Ling Meng, JA Phillips, et al. Molecular recognition of acute myeloid leukemia using aptamers. *Leukemia*, 23(2):235–244, 2009.
13. Abdullah Tahir Bayrac, Kwame Sefah, Parag Parekh, Ceren Bayrac, Basri Gulbakan, Huseyin Avni Oktem, and Weihong Tan. In vitro selection of dna aptamers to glioblastoma multiforme. *ACS chemical neuroscience*, 2(3):175–181, 2011.
14. John G Bruno and Johnathan L Kiel. In vitro selection of dna aptamers to anthrax spores with electrochemiluminescence detection. *Biosensors and Bioelectronics*, 14(5):457–464, 1999.
15. Romy Kirby, Eun Jeong Cho, Brian Gehrke, Travis Bayer, Yoon Sok Park, Dean P Neikirk, John T McDevitt, and Andrew D Ellington. Aptamer-based sensor arrays for the detection and quantitation of proteins. *Analytical Chemistry*, 76(14):4066–4075, 2004.
16. Dihua Shangguan, Ying Li, Zhiwen Tang, Zehui Charles Cao, Hui William Chen, Prabodhika Mallikaratchy, Kwame Sefah, Chaoyong James Yang, and Weihong Tan. Aptamers evolved from live cells as effective molecular probes for cancer study. *Proceedings of the National Academy of Sciences*, 103(32):11838–11843, 2006.
17. Zhiwen Tang, Parag Parekh, Pete Turner, Richard W Moyer, and Weihong Tan. Generating aptamers for recognition of virus-infected cells. *Clinical chemistry*, 55(4):813–822, 2009.
18. Eyetech Study Group et al. Preclinical and phase 1a clinical evaluation of an anti-vegf pegylated aptamer (eye001) for the treatment of exudative age-related macular degeneration. *Retina*, 22(2):143–152, 2002.
19. Eyetech Study Group et al. Anti-vascular endothelial growth factor therapy for subfoveal choroidal neovascularization secondary to age-related macular degeneration: phase ii study results. *Ophthalmology*, 110(5):979–986, 2003.
20. Hongguang Sun, Xun Zhu, Patrick Y Lu, Roberto R Rosato, Wen Tan, and Youli Zu. Oligonucleotide aptamers: new tools for targeted cancer therapy. *Molecular Therapy-Nucleic Acids*, 3:e182, 2014.
21. Craig Tuerk and Larry Gold. Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase. *science*, 249(4968):505–510, 1990.
22. Andrew D Ellington and Jack W Szostak. In vitro selection of rna molecules that bind specific ligands. *nature*, 346(6287):818–822, 1990.
23. Pardis C Sabeti, Peter J Unrau, and David P Bartel. Accessing rare activities from random rna sequences: the

importance of the length of molecules in the starting pool. *Chemistry & biology*, 4(10):767–774, 1997.
24. Naveen K Singh, Yixun Wang, Connie Wen, Brandon Davis, Xuelin Wang, Kyungsene Lee, and Yong Wang. High-affinity one-step aptamer selection using a non-fouling porous hydrogel. *Nature Biotechnology*, pages 1–8, 2023.
25. Neda Emami and Reza Ferdousi. Aptanet as a deep learning approach for aptamer–protein interaction prediction. *Scientific Reports*, 11(1):6074, 2021.
26. H Tomas Rube, Chaitanya Rastogi, Siqian Feng, Judith F Kribelbauer, Allyson Li, Basheer Becerra, Lucas AN Melo, Bach Viet Do, Xiaoting Li, Hammaad H Adam, et al. Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nature biotechnology*, 40(10):1520–1527, 2022.
27. Ali Bashir, Qin Yang, Jinpeng Wang, Stephan Hoyer, Wenchuan Chou, Cory McLean, Geoff Davis, Qiang Gong, Zan Armstrong, Junghoon Jang, et al. Machine learning guided aptamer refinement and discovery. *Nature Communications*, 12(1):2366, 2021.
28. Namhee Kim, Joseph A Izzo, Shereef Elmetwaly, Hin Hark Gan, and Tamar Schlick. Computational generation and screening of rna motifs in large nucleotide sequence pools. *Nucleic acids research*, 38(13):e139–e139, 2010.
29. Jan Hoinka, Alexey Berezhnoy, Phuong Dao, Zuben E Sauna, Eli Gilboa, and Teresa M Przytycka. Large scale analysis of the mutational landscape in ht-selex improves aptamer discovery. *Nucleic acids research*, 43(12):5699–5707, 2015.
30. Qingtong Zhou, Xiaole Xia, Zhaofeng Luo, Haojun Liang, and Eugene Shakhnovich. Searching the sequence space for potent aptamers using selex in silico. *Journal of chemical theory and computation*, 11(12):5939–5946, 2015.
31. Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.
32. Maciej Besta, Raphael Grob, Cesare Miglioli, Nicola Bernold, Grzegorz Kwasniewski, Gabriel Gjini, Raghavendra Kanakagiri, Saleh Ashkboos, Lukas Gianinazzi, Nikoli Dryden, et al. Motif prediction with graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 35–45, 2022.
33. Xiaoyong Pan, Peter Rijnbeek, Junchi Yan, and Hong-Bin Shen. Prediction of rna-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC genomics*, 19(1):1–11, 2018.
34. Jinho Im, Byungkyu Park, and Kyungsook Han. A generative model for constructing nucleic acid sequences binding to a protein. *BMC genomics*, 20(13):1–13, 2019.
35. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
36. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
37. Nathan Killoran, Leo J Lee, Andrew Delong, David Duvenaud, and Brendan J Frey. Generating and designing dna with deep generative models. *arXiv preprint arXiv:1712.06148*, 2017.
38. Furkan Ozden, Sina Barazandeh, Dogus Akboga, Urartu Ozgur Safak Seker, and A Ercument Cicek. Rnagen: A generative adversarial network-based model to generate synthetic rna sequences to target proteins. *bioRxiv*, pages 2023–07, 2023.
39. Natsuki Iwano, Tatsuo Adachi, Kazuteru Aoki, Yoshikazu Nakamura, and Michiaki Hamada. Generative aptamer discovery using raptgen. *Nature Computational Science*, 2(6):378–386, 2022.
40. Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
41. Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
42. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
43. Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.
44. Xingang Peng, Jiaqi Guan, Jian Peng, and Jianzhu Ma. Pocket-specific 3d molecule generation by fragment-based autoregressive diffusion models. 2023.
45. Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.
46. Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.
47. Sidney Lyayuga Lisanza, Jacob Merle Gershon, Sam Wayne Kenmore Tipps, Lucas Arnoldt, Samuel Hendel, Jeremiah Nelson Sims, Xinting Li, and David Baker. Joint generation of protein sequence and structure with rosettafold sequence space diffusion. *bioRxiv*, pages 2023–05, 2023.
48. Alexander E Chu, Lucy Cheng, Gina El Nesr, Minkai Xu, and Po-Ssu Huang. An all-atom protein generative model. *bioRxiv*, pages 2023–05, 2023.
49. Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.
50. Yoshio Tateno, Tadashi Imanishi, Satoru Miyazaki, Kaoru Fukami-Kobayashi, Naruya Saitou, Hideaki Sugawara, and Takashi Gojobori. Dna data bank of japan (ddbj) for genome scale research in life science. *Nucleic acids research*, 30(1):27–30, 2002.
51. Eli Kaminuma, Takehide Kosuge, Yuichi Kodama, Hideo Aono, Jun Mashima, Takashi Gojobori, Hideaki Sugawara, Osamu Ogasawara, Toshihisa Takagi, Kousaku Okubo, et al. Ddbj progress report. *Nucleic acids research*, 39(suppl_1):D22–D27, 2010.
52. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
53. Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable

and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*, 2022.

54. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
55. Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.
56. Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6:1–14, 2011.
57. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
58. Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *bioRxiv*, pages 2022–08, 2022.
59. Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.
60. Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.
61. Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science*, 13(9):2701–2713, 2022.
62. Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. Generating 3d molecules for target protein binding. *arXiv preprint arXiv:2204.09410*, 2022.
63. Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.
64. Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, pages 17644–17655. PMLR, 2022.
65. The GPyOpt authors. GPyOpt: A bayesian optimization framework in python. `http://github.com/SheffieldML/GPyOpt`, 2016.
66. Kengo Sato, Manato Akiyama, and Yasubumi Sakakibara. Rna secondary structure prediction using deep learning with thermodynamic integration. *Nature communications*, 12(1):941, 2021.

Supplementary Figures

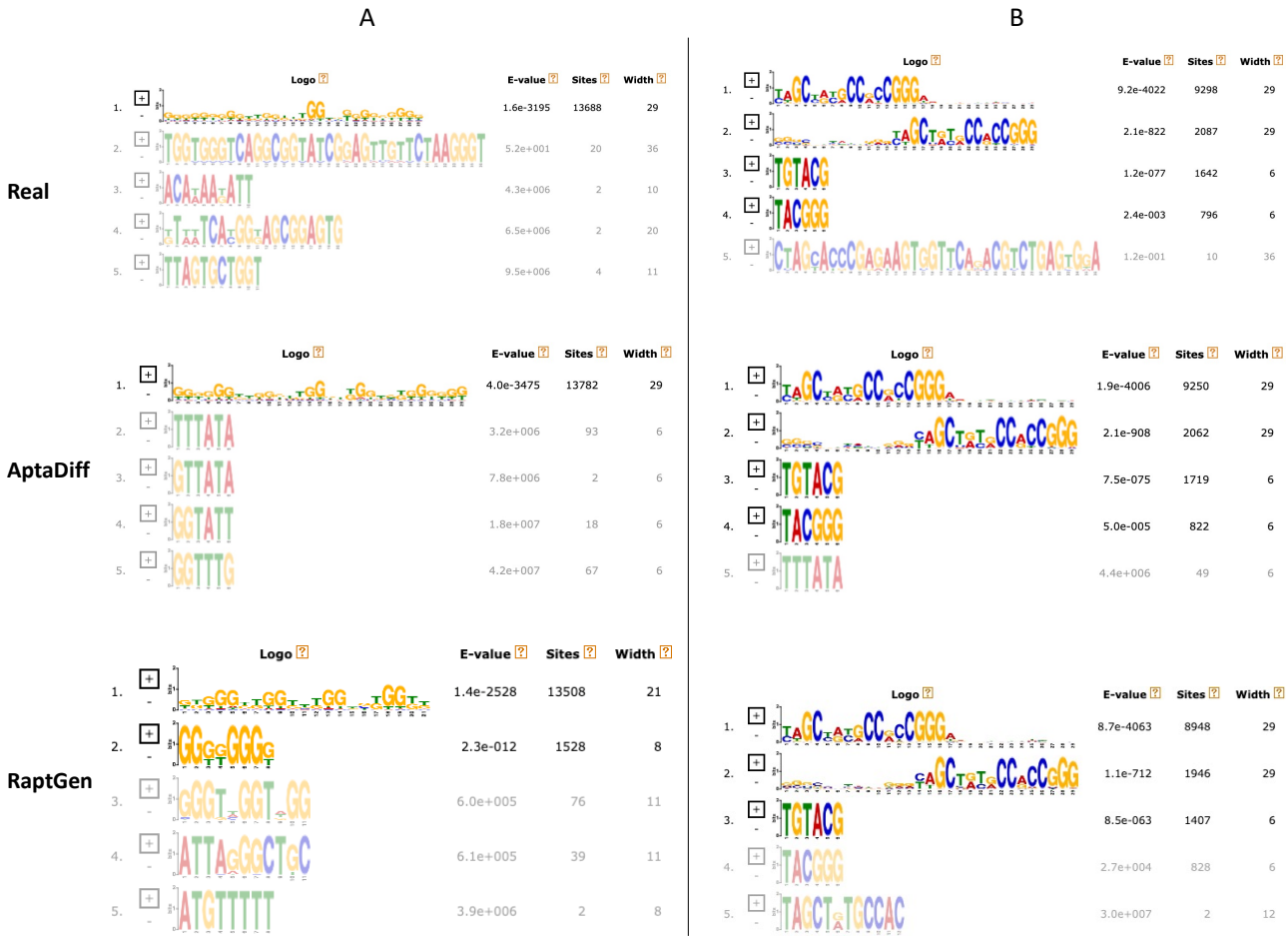

Fig. S1. Illustration of the motif clustering analysis by MEME. The left side of the figure presents the results on dataset A, while the right side shows the results on dataset B. "Real" denotes the real SELEX data, "AptaDiff" represents data generated by AptaDiff, and "RaptGen" represents data generated by RaptGen. The E-value indicates the clustering score obtained by MEME, where a smaller score suggests better clustering for motifs in that category. "Sites" denotes the number of sequences in that category, and "Width" indicates the maximum length of motifs in that category. The darkness of the sequence logo's colors is proportional to the "Sites" value, with darker colors indicating a higher number of sequences.

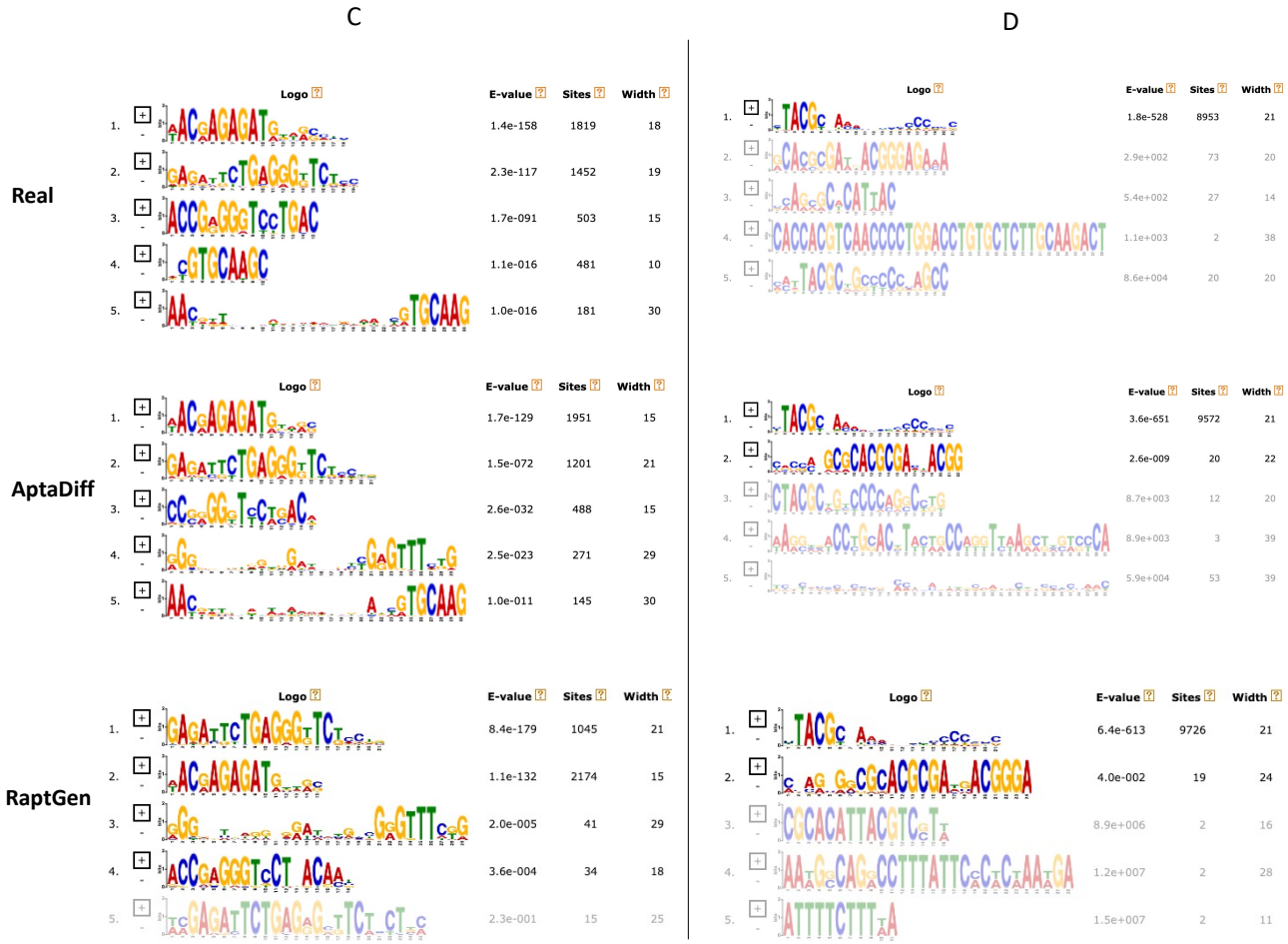

Fig. S2. Illustration of the motif clustering analysis by MEME. The left side of the figure presents the results on dataset C, while the right side shows the results on dataset D. "Real" denotes the real SELEX data, "AptaDiff" represents data generated by AptaDiff, and "RaptGen" represents data generated by RaptGen. The E-value indicates the clustering score obtained by MEME, where a smaller score suggests better clustering for motifs in that category. "Sites" denotes the number of sequences in that category, and "Width" indicates the maximum length of motifs in that category. The darkness of the sequence logo's colors is proportional to the "Sites" value, with darker colors indicating a higher number of sequences.

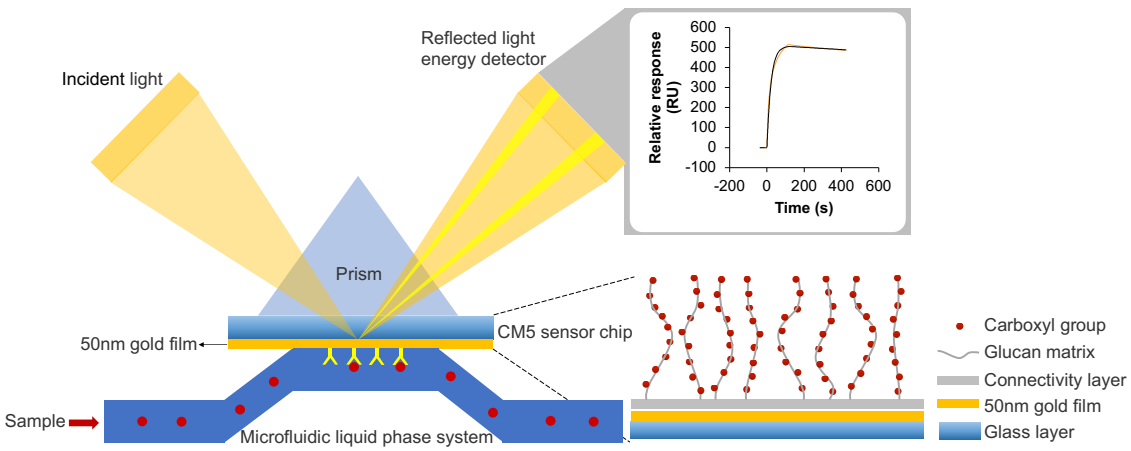

Fig. S3. Surface plasmon resonance (SPR) principle. In SPR experiments, when a sample solution (e.g., aptamers) flows over the surface of a chip coated with biomolecules (e.g., antibodies), interactions between molecules in the sample and those on the surface result in an increase in surface mass, consequently causing a change in the resonance angle frequency. RU values are employed to quantify the extent of this change in surface mass. Typically, 1 RU signifies the change in resonance angle frequency generated when biomolecules with a mass of 1 nanogram (ng) per square millimeter (mm^2) bind to the surface.

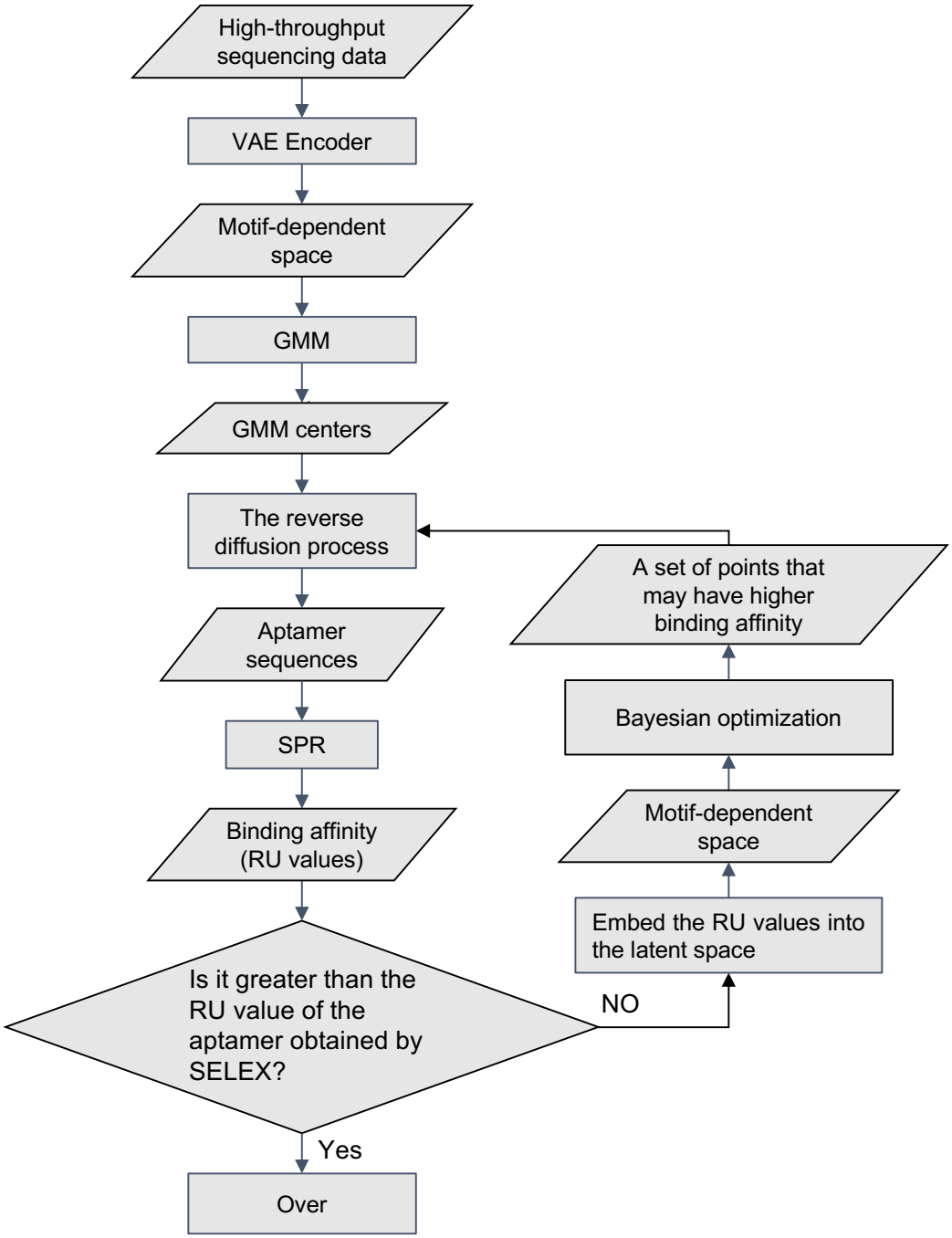

Fig. S4. Flowchart of the actual application of trained AptaDiff. Upon completing AptaDiff training, we employ a Gaussian Mixture Model (GMM) to classify points within the latent space, subsequently selecting the category center points of GMM as starting conditions to generate sequences through the reverse diffusion process. These sequences then undergo SPR experiments to measure their binding affinity (i.e., RU values) with target proteins. These RU values are subsequently embedded back into the latent space, after which Bayesian optimization is applied to identify a set of points likely to have higher binding affinity. These points, in turn, serve as conditions to generate sequences through the reverse diffusion process for further SPR experiments. This cycle continues until the candidate sequences' affinity surpasses those identified through traditional SELEX experiments.

Supplementary Tables

Table S1. Sequences hand-picked in SELEX experiments targeting the IGFBP3 protein. The aptamer sample concentration in the SPR experiment was 500 nM.

ID	sequences	length	RU
A-SELEX-01	TGGTGGGTCAGGCGGTATCGGAGTTGTTCTAAGGGT	36	276.6
A-SELEX-02	CGCATTTGGGTTGGTTTGGACTGGTTTTGCTGTTTG	36	138.1
A-SELEX-03	GCGGGTTGGTTGGTCTGGCTGGGGGTGGGTGGGACT	36	207.7
A-SELEX-04	GACGGTTGGTTGGTCTGGATTGGTTGGGTAGGTTGG	36	67.5
A-SELEX-05	TGGGTGGGTTGGATGGTCTGGCTGGGGGCGGGTTTG	36	190.2
A-SELEX-06	ATTAGGGCTGCAATTGGTGTTGGGTGGTTCGGGTAA	36	110.9
A-SELEX-07	GCATTTGGGTTGGTTTGGACTGGTTTTGCGCTCGTG	36	88.2
A-SELEX-08	GTTTATATGGCGATTGGATTGGTTTGGACTGGGTGG	36	40.4
A-SELEX-09	CGGGCGGGTGGGAATTTCAGAGTGCTGGTGGGTGTG	36	116.7
A-SELEX-10	GGGGGTTCGGGTTGGTTGGTCTGGTTTGGGGGTGGG	36	185.1
A-SELEX-11	TGGGCGGGTGGGTGGTTTGGTTGGCCTGGGGGTTAT	36	147.5
A-SELEX-12	GTGTCACATTTGGGTTGGTTTGGACTGGTTTTGTTA	36	75.4
A-SELEX-13	CGGGAAGTCTGGTGGCGGTGTGTGGTTTGGCTACGT	36	29.1
A-SELEX-14	GAGGGGCGAGGGTATGATTCGAGGGGTGATAGGGGG	36	57.1
A-SELEX-15	CTATTGGGGTCTGGGGTGTCTAGTGGGTTTGGCGGT	36	94.9
A-SELEX-16	GGGGGAGGTGGAGGTCTGTACTGGGTGGGCGGGTGT	36	257.8
A-SELEX-17	GGAGGGTGGGTCGGGTTGGTTGGTCTGGGTTGGGGG	36	139.5
A-SELEX-18	TCTCCGGTATGGGTTGGTTTGGCCTGGGGTGCGCGA	36	52.6
A-SELEX-19	TTCACGGTTTGTCAATTCACGGTAGGTGTGCTGGGT	36	23.7
A-SELEX-20	GCAGCAATTTGGGTTGGTTTGGACTGGTGTTTGTAG	36	54.9
A-SELEX-21	GGGCGGGTGGGGGAGGAGGTGGTTGCTCTGGGAATG	36	300.8
A-SELEX-22	CGGTGGTGGGATTCTTGGGTGGGAGGAGGGGGTATG	36	261.3
A-SELEX-23	GGAGGGAGGGAGGGGTATCCATGGTGGAGGTGGGGG	36	232.3
A-SELEX-24	TTGGATGGTCTGGTGGTTTGGTTGGTCTGGTGGGTG	36	62.9
A-SELEX-25	ATCGGTGGATTTGGGTTGGTTTGGACTGGTAGGGGT	36	46.9
A-SELEX-26	TGGCTTGGTTGGTCTGGTACTGGTTGGATTGGTTGG	36	35.1
A-SELEX-27	CAGGGAGGGTGGGTGAATTGGCGGAGGTGGGGGTGT	36	324.6
A-SELEX-28	GGGGGGACGGGTGGGGTTGGGTTAGGTGGTGTGTGA	36	261
A-SELEX-29	GGCGGTGGGTGGGTTGTTGTGGGAGGGGGAGGGGGA	36	384.7
A-SELEX-30	CGGGTGGGGGAGGAGGAGGGTATTGGGGTGGGTCTC	36	287.4
A-SELEX-31	CTGGCGGGTGGAAAATGGGAGGCGGGTGGGGGTTCG	36	197.5
A-SELEX-32	GGGTGGGTACGATTGGTTTGGTCTGGGGGGTGGGTT	36	193.4
A-SELEX-33	TCGGGTGGGACTGTTGGCGGAGGTGGGGGAGGGTAC	36	255.8
A-SELEX-34	GAGGGGCGAGGGTATGATTTGAGGGGTGATAGGGGG	36	68.6
A-SELEX-35	GGGGGAGGAGGAGGCAAGGATGGGTGGGCGGGTGTT	36	273.2
A-SELEX-36	GGTGGTGGGAGGGTCATATGGGAGGTGGTGTGGGGA	36	285.9
A-SELEX-37	TGGGCGGGTGGAATATGGGAGGTGGGAGGGGGTGTC	36	215.1
A-SELEX-38	GTAGGGGGGGTTTTGATCGGAGGCGGTGGGGGAGGG	36	298.4
A-SELEX-39	CTGGTGGTGGGTCATGGGAGGGTGGAGGGGGAATAG	36	261.6
A-SELEX-40	GCTCGAGATTTGGGTTGGTTTGGACTGGTGTTCTGC	36	56.9
A-SELEX-41	GTCGGGGGGGAGGGTCAGACTGGTGGCGGTGGGGGA	36	251.1
A-SELEX-42	GGGGTGCCCGTGTGGGTTGGTTTGGTCTGGTGGGTC	36	72.3
A-SELEX-43	GGGGGGGCATTGTTCTCGGAGGTGGTGGGGGAGGGT	36	213.7
A-SELEX-44	AGGGTGGGTTGGTTGGTCTGGTTGGGGGTCGGGTTC	36	181
A-SELEX-45	TTGGCGGGTTTCTGGGAGGGTGGCGGGGGTGGTATG	36	340.6
A-SELEX-46	GGCTGTGGTGTAGGGTTGGTTGGTCTGGTTGGGTGG	36	179.9
A-SELEX-47	TAGGTGGTGGGTAGACGGGTGGGAGGAGGGGGTCAT	36	238
A-SELEX-48	TGGAGTTTGGGTTGGTTTGGTCTGGTTGTTTGGGAA	36	145.2
A-SELEX-49	TGGGAGGTGGGCAATTGGGTGGGAGGTGGGGGTGCT	36	300.1
A-SELEX-50	GGGTTGGATTGGTTGGACTGGGGGTTGGGTGGGGGT	36	210
A-SELEX-51	TGGTTGTCTTCACGGTATTTCACGGTATGTGCTGGG	36	70.2

Table S2. Sequences hand-picked in SELEX experiments targeting the PTK7 protein. The aptamer sample concentration in the SPR experiment was 500 nM.

ID	sequences	length	RU
B-SELEX-01	TGTACGGCCCATTATTGGGTAGCTATACCACCGGGG	36	214.8
B-SELEX-02	CAGCTGTGCCATCGGGGAGAGTGAAACCTTGTACGA	36	382
B-SELEX-03	GATGGCCACACCGCCGGGAATCTGGTATGTACGGGA	36	291.4
B-SELEX-04	CAGCTGTGCCACCGGGAAATGGGTTCTCATGTACGG	36	297.1
B-SELEX-05	TAGCTATACCACCGGGAGGGAGTCACAACTGTACGG	36	364
B-SELEX-06	TGGTACGGGTACACTTACCTGCTGCGCCGCCGGGTT	36	178.8
B-SELEX-07	TTCGACAGCCGCGCCGCCGGGAACTTGTGTACGGTC	36	337.3
B-SELEX-08	TAGCTATGCCACCGGGAGCGTCTGATTAGTGTACGG	36	420.6
B-SELEX-09	CAGCTGTGCCATCGGGTAACGGGATCCTCCGAACGA	36	324.5
B-SELEX-10	ACACAGCTGTGCCACCGGGATATCAATGTACGGGGT	36	150.4
B-SELEX-11	TAGCAATGCCACCGGGAAGGATGTTGGACTGTACGG	36	415.2
B-SELEX-12	GAATGTACGGCCGCCTACGGTAGCTATACCACCGGG	36	296.9
B-SELEX-13	CGTAGCCATGCCACCGGGAGCTCTTGTGTACGGCGA	36	397
B-SELEX-14	GATAGCCACACCGCCGGGAATTACGGATGTACGGGG	36	357
B-SELEX-15	ACAGCTGTACCACCGGGAGGCAATCGACGTACGGTC	36	226.4
B-SELEX-16	GTGTACGGCTCGTCGAGAGCAGCTGTGCCACCGGGC	36	153.3
B-SELEX-17	ACAGCCGCGCCGCCGGGAACATTCCATGTGTACGGT	36	461.6
B-SELEX-18	TAGCCATGCCACCGGGTAGAAGTCTTCTGGTTACGG	36	398
B-SELEX-19	CAGCTGCGCCGTCGGGGGCGTTCTTCGGTGGTACGA	36	177
B-SELEX-20	ATTGCCACGCCGCCGGGAGTAACTTTCACGTACGGG	36	556.5
B-SELEX-21	CACAGCTGCGCCGTCGGGAAGTAGTTTCGTACGATG	36	105.7
B-SELEX-22	TACAGCTGTGCCACCGGGAGACTGATCGTACGGTAA	36	99.8
B-SELEX-23	GGGGTTGCCACGCCGCCGGGTATGTTGTACGGACTG	36	167.8
B-SELEX-24	AATAGCCACGCCGCCGGGTTGTGTATAATACGGGTC	36	89.5
B-SELEX-25	ATTGCTACGCCGCCGGGAGTAACTTTCACGTACGGG	36	381.9
B-SELEX-26	TTGCCATGCCACCGGGTTCGTCCTTGATAGATACGG	36	135
B-SELEX-27	GCAGCCGCGCCGCCGGGATTTGCTTAGAAGTACGGC	36	211.3
B-SELEX-28	TAGCGATGCCACCGGGATATTGTACTAAACTTACGG	36	133.5
B-SELEX-29	TAGCCATGCCACCGGGTAGAACTCTTCTGGTTACGG	36	277
B-SELEX-30	TCAGTACGGCTCCATGGAGTAGCTATGCCACCGGGT	36	242.1
B-SELEX-31	GAATGTACGGCCGCTTACGGTAGCTATACCACCGGG	36	224
B-SELEX-32	ATCCTGCTGTGCCACCGGGAGCTTGATGTACGGGAC	36	221.1
B-SELEX-33	ATATTGCTACGCCGCCGGGTTTAATTGGTACAGGAT	36	3.8
B-SELEX-34	CAGCTGCGCCGGCGGGGGCGTTCTTCGGTGGTACGA	36	7.1

Table S3. Quantitative Illustration of the generative performance on datasets A, B, C, and D. FID reported on 10k samples (Lower is better). Using DNA-BERT2 [59] as the feature extractor.

Source	Dataset	Target Protein	FID	
			RaptGen	AptaDiff
Ours	A	IGFBP3	1.2627	0.0222
	B	PTK7	0.0281	0.0067
Public	C	TGM2	0.0293	0.0097
	D	ITGαVβ3	0.0233	0.0137

Supplementary Sections

Supplementary Section 1: Sequences Obtained in the Present Study

The sequences obtained in the present study are shown in Supplementary Table S4-S8. The ID is named after a rule of dataID, the method to select the sequence, and the index of the sequence. The dim1 and dim2 represent the two dimensions of the two-dimensional latent space. RU and KD reflect the binding affinity of the target sequence as assessed by the SPR experiment. The sample concentration in the SPR experiment was 500 nM.

Supplementary Section 2: Statistics of the Dataset

The SELEX experiments for datasets A and B were conducted against target proteins IGFBP3 and PTK7, respectively. For each protein target, six rounds of SELEX screening were carried out, with each round encompassing the following steps: (1) positive selection where the ssDNA library was exposed to the target protein, (2) elution of nucleic acid molecules bound to the target protein followed by PCR amplification and enrichment of the library, and (3) high-throughput sequencing of the enriched library. Considering that candidate aptamer libraries enriched in later rounds of the SELEX experiments have a higher degree of enrichment, we selected the data obtained from the last round of experiments as the initial data for datasets A and B. Following this, we processed the data to get the datasets for model training. Specifically, all sequences in the training set were carefully filtered. Sequences meeting the following three criteria were retained: (1) exact matching with primers (including forward and reverse primers), (2) precise alignment with the sequence design length, and (3) sequences read more than once.
The statistics of the datasets are shown in Table S9. The column names are as follows:

- Source: Source of the datasets, where "Ours" indicates that the dataset comes from our own SELEX experiment, "Public" represents the public SELEX dataset obtained from DDBJ [50, 51].
- Target: The target protein.
- ID: The ID named after the rule; indicator of the dataset-round of the SELEX.
- Filtered unique: Number of unique sequences matched by primer match and target length.
- Target length: The designable target length is the length of the variable random region in the middle of the aptamer.

Note that datasets C and D are taken directly from the real training data used in Reference [39].

Supplementary Section 3: Comparative Analysis to Traditional SELEX Screening Experiments.

Compared with the top 5 affinity aptamers selected by traditional SELEX screening experiments, the overall affinity of the aptamers we generated after Bayesian optimization was higher (Table S10 and S11)

Supplementary Section 4: Details of Handling the Diffusion Model Stochasticity.

During the sample generation process following the training of AptaDiff, two main determinants influence the outcome: the latent representation ($\mathbf{z}$) from the VAE and the random noise point ($\boldsymbol{x}_T$) sampled from a multinomial distribution. Our design leverages the latent representation to steer the diffusion generation process. To secure the determinacy and reproducibility of the samples generated by AptaDiff, thereby ensuring that the generated samples are influenced solely by the latent representation, we adhere to the approach suggested in reference [53]. Specifically, we assign a fixed random seed to each latent representation ($\mathbf{z}$) to sample the nosie ($\boldsymbol{x}_T$) and share this stochasticity during the reverse diffusion process across all generated samples. By adopting this methodology, we ensure that each generated sample can be deterministically associated with a given latent representation, allowing for the reproducibility of results which is crucial for the verifiability and scientific rigor of our findings.

Table S4. The sequences obtained in the initial sequence selection for dataset A.

ID	dim1	dim2	AptaDiff-generated sequences	RU ↑	KD(nM) ↓
A-GMM-0	-0.98575	0.01455	AGGGGATTCAATGGGTGGCGGTGGCGGGGGCGGGCA	102.2	21.8
A-GMM-1	2.52046	2.12688	GACTGTTGGTTGGTCTGGATTGGTTGGGTAGGTTGG	64.0	44.2
A-GMM-2	2.18148	-0.16983	GTCTTGGGTTGGTTTGGTCTGGTTTTCTGTTGGTTG	111.0	66.6
A-GMM-3	2.71004	-1.67348	GCATTTGGGTTGGTTTGGACTGGTTTTGCGATCGTG	100.2	24.5
A-GMM-4	1.03976	0.91432	GAGACGCGGGTGTCGATCTTGGTTGGTCTGGTTTGG	110.7	67.3
A-GMM-5	-1.49014	-1.29256	TGGGGGGGAGGGTCAGTAGGTGGTGGTGGGGGTTTA	254.2	39.1
A-GMM-6	-0.43521	3.20198	TGGTGGGTCAGGCGGTATCGGACTCGGGGTGGGGGT	65.9	88.1
A-GMM-7	-0.61396	-0.55970	GGGAGGGGGTGGAGGGAGGATTAGACTGGGTGGGGT	135.1	39.8

Table S5. The sequences obtained after the first Bayesian optimization for dataset A.

ID	dim1	dim2	AptaDiff-generated sequences	RU ↑	KD(nM) ↓
A-BO1-0	-1.62510	-1.77080	GGGGCGGGTGGGATTTAGTGGTGGGTGGTGGGGGTC	**478.5**	4.72
A-BO1-1	-2.17348	-1.19782	CGGTGGTGGGATTCTTGGGTGGGAGGAGGGGGTATG	324.7	6.24
A-BO1-2	-0.91938	-1.38822	CGGGAGGGTGGGGGCGGTGGTTGGTATCTTGGGATG	**413.4**	6.56
A-BO1-3	-1.75079	-2.57572	CGGGCGGGTGGGAATGTCAGAGTGCTGGTGGGTGTG	188.6	11.4
A-BO1-4	-0.30913	-2.03224	TACGTGGTCTGGGGTTTGGGTTGGGTGGTTGGCGGT	91.6	13.8
A-BO1-5	-1.66651	-0.49218	TTGGTGGTGGGTCAATGGGAGGGTAGGTGGGGGTTG	304.1	14.4
A-BO1-6	-3.06663	-1.62786	TTGGCGGGTGGAAAATGGGAGGTGGGTGGGGGTGTC	344.9	7.87
A-BO1-7	-3.04112	0.05768	GGGGGAGGTGGAGGTCTGGACTGGGTGGGCGGGTGT	240.7	18.7

Table S6. The sequences obtained in the initial sequence selection for dataset B.

ID	dim1	dim2	AptaDiff-generated sequences	RU ↑	KD(nM) ↓
B-GMM-0	1.28170	1.18628	TGTACGGCCCATTATTGGGTAGCTATACCACCGGGG	349.7	9.95
B-GMM-1	-0.70062	-0.63145	GCAACAGCGCCGCCGGGAAAATTCCATGTGTACGGT	74.7	50.6
B-GMM-2	-0.26429	1.49547	ATTGCCACGACGCCGGGAGTAACTTTCACGTACGGG	96.6	78.9
B-GMM-3	1.51646	-1.10955	CCAGCCGCGCCGCCGGGAACATTCCATGTGTACGGT	123.5	30.8
B-GMM-4	0.38743	-0.66389	CAGCTGTGCCACCGGGAATTGCATTGCTATGTACGA	345.0	16.8
B-GMM-5	0.74827	0.01673	CGTTGCTATACCACCGGGAGGATTCCTCCTACGGAC	7.0	3.47
B-GMM-6	-0.42762	0.64653	GATAGCCACGCCGCCGGGTTTCTTTACATTACGGTG	248.7	3.17
B-GMM-7	-0.35786	-2.05457	TAGCTATGCCACCGGGAGCGTCTGATTAGTGTACGC	13.2	27

Table S7. The sequences obtained after the first Bayesian optimization for dataset B.

ID	dim1	dim2	AptaDiff-generated sequences	RU ↑	KD(nM) ↓
B-BO1-0	1.31427	1.13238	TTGCTATACCACCGGGAGGGAGTCACAACTGTACGG	412.8	6.7
B-BO1-1	0.39944	-0.72667	GAATATACGGCCGCCTGCGGTAGCTATACCACCGGG	221.4	10
B-BO1-2	-0.50163	0.59176	ACAGCCGCGCCGCCGGGAACATTCCATGTGTACGGT	459.2	5
B-BO1-3	0.21968	-0.59870	TGTACGGCACATTATTGGGTAGCTATACCACCGGGG	199.0	10
B-BO1-4	1.15773	1.15442	TAGACATGCCACCGGGTAGAAGTCTTCTGGTTACGG	394.7	6
B-BO1-5	0.57680	-0.48182	CATTGCTACGCCGCCGGGTCAAAAAGAGTGACGGTG	239.6	50
B-BO1-6	1.45701	1.37257	TAGCCATGCCACCGGGTAGAATTCCAATTGTTACGG	393.8	5
B-BO1-7	-0.35265	0.80801	GTAGCCACACCGCCGGGAGTAACTTTCACGTACGGG	497.9	6

Table S8. The sequences obtained after the second Bayesian optimization for dataset B.

ID	dim1	dim2	AptaDiff-generated sequences	RU ↑	KD(nM) ↓
B-BO2-0	-2.09899	-1.09744	ATTGCCACGCCGCCGGGAGTAACTTTCACGTACGGG	**560.3**	0.938
B-BO2-1	1.08411	-1.75004	CAGCTGTGCCACCGGGAAATGGGTTCTCATGTACGG	296.6	4.28
B-BO2-2	0.21863	-2.27909	TAGCTATGCCACCGGGAGGGAGGCTCAACTGTACGG	404.8	1.5
B-BO2-3	0.96820	-1.83689	CAGCCGTGCCACCGGGAAATGGGTTCTCATGTACGG	240.9	6.88
B-BO2-4	-0.07510	-0.30019	TAGCCATGCCACCGGGTAGAATTCCAATTGTTACGG	395.7	4.49
B-BO2-5	0.25922	-2.31897	TAGCTATACCACCGGGAGGGAGTCACAACTGTACGG	462.2	7.27
B-BO2-6	0.98799	-1.84701	CAGCTGTGCCACCGGGAAATGGGTTCTCATGTACGG	359.9	2.72
B-BO2-7	0.42106	-0.67002	CAGCTGTGCCACCGGGACGTTTGATACAAGGTACGG	229.3	8.17

Table S9. Statistics of the datasets used in our research.

Source	Dataset	Target	ID	Filtered unique	Target length
Ours	A	IGFBP3	A-6R	13802	36nt
	B	PTK7	B-6R	11532	36nt
Public	C	TGM2	C-4R	38613	30nt
	D	ITGαVβ3	D-4R	31358	40nt

Table S10. The sequences of the top 5 affinity aptamers we generated and the top 5 affinity aptamers selected through traditional SELEX screening for dataset A.

	ID	sequences	RU ↑	KD(nM) ↓
SELEX-TOP5	A-SELEX-29(PC)	GGCGGTGGGTGGGTTGTTGTGGGAGGGGGAGGGGGA	384.7	15.3
	A-SELEX-45	TTGGCGGGTTTCTGGGAGGGTGGCGGGGGTGGTATG	340.6	14.9
	A-SELEX-27	CAGGGAGGGTGGGTGAATTGGCGGAGGTGGGGGTGT	324.6	16.9
	A-SELEX-21	GGGCGGGTGGGGGAGGAGGTGGTTGCTCTGGGAATG	300.8	19.4
	A-SELEX-49	TGGGAGGTGGGCAATTGGGTGGGAGGTGGGGGTGCT	300.1	18.5
BO-TOP5	A-BO1-0	GGGGCGGGTGGGATTTAGTGGTGGGTGGTGGGGGTC	**478.5**	4.72
	A-BO1-2	CGGGAGGGTGGGGGCGGTGGTTGGTATCTTGGGATG	**413.4**	6.56
	A-BO1-6	TTGGCGGGTGGAAAATGGGAGGTGGGTGGGGGTGTC	344.9	7.87
	A-BO1-1	CGGTGGTGGGATTCTTGGGTGGGAGGAGGGGGTATG	324.7	6.24
	A-BO1-5	TTGGTGGTGGGTCAATGGGAGGGTAGGTGGGGGTTG	304.1	14.4

Table S11. The sequences of the top 5 affinity aptamers we generated and the top 5 affinity aptamers selected through traditional SELEX screening for dataset B.

	ID	sequences	RU ↑	KD(nM) ↓
SELEX-TOP5	B-SELEX-20(PC)	ATTGCCACGCCGCCGGGAGTAACTTTCACGTACGGG	556.5	0.997
	B-SELEX-17	ACAGCCGCGCCGCCGGGAACATTCCATGTGTACGGT	461.6	4.72
	B-SELEX-08	TAGCTATGCCACCGGGAGCGTCTGATTAGTGTACGG	420.6	1.72
	B-SELEX-11	TAGCAATGCCACCGGGAAGGATGTTGGACTGTACGG	415.2	9.03
	B-SELEX-18	TAGCCATGCCACCGGGTAGAAGTCTTCTGGTTACGG	398	1.16
BO-TOP5	B-BO2-0	ATTGCCACGCCGCCGGGAGTAACTTTCACGTACGGG	**560.3**	0.938
	B-BO1-7	GTAGCCACACCGCCGGGAGTAACTTTCACGTACGGG	497.9	6
	B-BO2-5	TAGCTATACCACCGGGAGGGAGTCACAACTGTACGG	462.2	7.27
	B-BO1-2	ACAGCCGCGCCGCCGGGAACATTCCATGTGTACGGT	459.2	5
	B-BO1-0	TTGCTATACCACCGGGAGGGAGTCACAACTGTACGG	412.8	6.7